2025 • 29 • 7

Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

VAVILOV JOURNAL OF GENETICS AND BREEDING

Основан в 1997 г.
Периодичность 8 выпусков в год
doi 10.18699/vjgb-25-120

2025 • 29 • 7

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Publication frequency: 8 issues a year
doi 10.18699/vjgb-25-120

Founders

Siberian Branch of the Russian Academy of Sciences

Federal Research Center Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

СОДЕРЖАНИЕ • 2025 • 29 • 7

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII • VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2025 • 29 • 7

Vavilovskii Zhurnal Genetiki i Selektsii. 2025;29(7):911-912
doi 10.18699/vjgb-25-97

N.A. Kolchanov

T.A. Bukharina

Dear colleagues,

We introduce the current issue of the *Vavilov Journal of Genetics and Breeding*, dedicated to computational biology.

Methods for genome sequencing have been rapidly developed over the past two decades. Sequencing has become cheaper by almost five orders of magnitude: for instance, from $100,000 to $500 for a personal human genome. Great progress has been made in transcriptomics, proteomics, metabolomics, and other omics technologies. We witness a new generation of techniques for biological object visualization on the genome, cellular, tissue, and organismal levels of living system organization. This informational explosion makes genetics the main source of huge bodies of data. Genetics outruns not only other fields of knowledge but global social media in the rate of information accumulation. Indeed, up to 40 exabytes of data are produced in life sciences annually, whereas the largest social platform YouTube produces only 2 exabytes, 20 times less.

Analysis of big genetic data has given rise to a new paradigm of modern genetics. It is focused on gene networks: groups of orchestrated genes that interact via their products: RNA, proteins, and metabolites. Gene networks are responsible for the formation of molecular, biochemical, cellular, physiological, morphological, behavioral, and other traits of the body on the base of information encoded in the genome. The regulation of gene networks is enormously complicated. The complexity is evident from the fact that the operation of a particular gene network element can be controlled by tens and hundreds of elementary regulatory processes. This is true for gene transcription regulation, mediated by tens of transcription factors, which interact with binding sites in gene promoters, and for proteins, whose activity is modulated by interaction with numerous ligands, acting as allosteric regulators. The same is true for metabolic pathways, where the number of elementary regulatory processes sometimes exceeds the number of biochemical reactions by an order of magnitude. Another fundamental property of living systems found in big data analysis is the extremely high level of genetic variability in populations of humans, animals, plants, and microorganisms.

Analysis of big genetic data requires the development of a new generation of methods to process very large bodies of information. This generation includes bioinformatics methods for the reconstruction, analysis, and modeling of structural organization and molecular mechanisms of the functioning of genomes, genes, and genetic macromolecules encoded by them: RNA and proteins. It also includes novel methods of computational systems biology for the reconstruction, analysis, and modeling of genetics systems operating on the levels of cells, tissues, organs, and entire organisms.

The new epoch of big genetic data, including life sciences, demands transformation of key approaches in bioinformatics and computational systems biology. What are fundamental trends in this field? First, it is the integration of conventional

methods in bioinformatics and computational systems biology with artificial intelligence and deep machine learning. Second, employment of the results as grounds for the development of a new generation of software and data support for interpreting big genetic data, and, most importantly, for planning experiments to verify the results of computer-aided predictions from big data analysis. Progress in this direction would mark a fundamental transformation of the basic paradigm in modern research: Science directed by hypotheses is complemented by new science directed by big data analysis.

This progress occurs in all sciences, but just bioinformatics and computational systems biology are at the forefront.

Science Editors:

N.A. Kolchanov,
Full Member of the Russian Academy of Sciences,
Dr. Sci. (Biology), Research Advisor
of the Institute of Cytology and Genetics, Novosibirsk, Russia

T.A. Bukharina,
Academic Secretary of the Department of Systems Biology,
Institute of Cytology and Genetics, Novosibirsk, Russia

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):913-924

doi 10.18699/vjgb-25-98

DNA damage reflected in the evolution of G-runs in genomes

I.R. Grin [1], D.O. Zharkov [1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
dzharkov@niboch.nsc.ru

Abstract. DNA oxidation is one of the main types of damage to the genetic material of living organisms. Of the many dozens of oxidative lesions, the most abundant is 8-oxoguanine (8-oxoG), a premutagenic base that leads to G→T transversions during replication. Double-stranded DNA can conduct holes through the π system of stacked nucleobases. Such electron vacancies are ultimately localized at the 5′-terminal nucleotides of polyguanine runs (G-runs), making these positions characteristic sites of 8-oxoG formation. While such properties of G-runs have been studied *in vitro* at the level of chemical reactivity, the extent to which they can influence mutagenesis spectra *in vivo* remains unclear. Here, we have analyzed the nucleotide context of G-runs in a representative set of 62 high-quality prokaryotic genomes and in the human telomere-to-telomere genome. G-runs were, on average, shorter than polyadenine runs (A-runs), and the probability of a G-run being elongated by one nucleotide is lower than in the case of A-runs. The representation of T in the position 5′-flanking G-runs is increased, especially in organisms with aerobic metabolism, which is consistent with the model of preferential G→T substitutions at the 5′-position with 8-oxoG as a precursor. Conversely, the frequency of G and C is increased and the frequency of T is decreased in the position 5′-flanking A-runs. A biphasic pattern of G-run expansion is observed in the human genome: the probability of sequences longer than 8–9 nucleotides being elongated by one nucleotide increases significantly. An increased representation of C in the 5′-flanking position to long G-runs was found, together with an elevated frequency of 5′-G→A substitutions in telomere repeats. This may indicate the existence of mutagenic processes whose mechanism has not yet been characterized but may be associated with DNA polymerase errors during replication of the products of further oxidation of 8-oxoG.

Key words: DNA damage; mutagenesis; 8-oxogianine; G-runs; telomeres

For citation: Grin I.R., Zharkov D.O. DNA damage reflected in the evolution of G-runs in genomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):913-924. doi 10.18699/vjgb-25-98

Funding. This study was supported by the Russian Science Foundation (project 24-14-00285, human genome analysis). The prokaryotic genome analysis part was supported by the Russian Ministry of Science and Higher Education (project 125012300657-2).

Отражение процессов повреждения ДНК в эволюции G-трактов в геномах

И.Р. Грин [1], Д.О. Жарков [1, 2]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
dzharkov@niboch.nsc.ru

Аннотация. Окисление ДНК представляет собой один из главных видов повреждения генетического материала живых организмов. Из многих десятков продуктов окислительного повреждения ДНК в наибольшем количестве встречается 8-оксогуанин (8-oxoG) – предмутагенное основание, приводящее при репликации к трансверсиям G→T. Двуцепочечная ДНК обладает способностью к проводимости положительных зарядов, связанных с дефицитом электронов в π-системе азотистых оснований. Такие заряды в конечном итоге локализуются на 5′-концевом нуклеотиде полигуаниновых трактов (G-трактов). В связи с этим 5′-концевые нуклеотиды G-трактов служат характерными местами образования 8-oxoG. Эти свойства G-трактов хорошо изучены *in vitro* на уровне реакционной способности, но остается неясным, насколько они могут отражаться в спектрах мутагенеза *in vivo*. В работе проанализирован нуклеотидный контекст G-трактов в репрезентативном наборе из 62 полных геномов прокариот и в геноме человека с покрытием «от теломеры до теломеры». Показано, что G-тракты в среднем короче полиадениновых трактов (A-трактов) и вероятность удлинения G-трактов на один нуклеотид ниже, чем в случае A-трактов. Установлено, что представленность T в положении, примыкающем к G-трактам с 5′-стороны, повышена, в особенности у организмов с аэробным метаболизмом, что согласуется с моделью преимущественных мутаций G→T в 5′-положении с 8-oxoG как предшественником. В то же время в положении, примыкающем

к A-трактам, повышена частота встречаемости G и C и снижена частота встречаемости T. В геноме человека наблюдается двухфазный характер разрастания G-трактов: начиная с длины 8–9 нуклеотидов вероятность их удлинения на один нуклеотид заметно увеличивается. Выявлена повышенная представленность C с 5′-стороны от длинных G-трактов и A при заменах в теломерных повторах, что может свидетельствовать о существовании мутагенных процессов, механизм которых пока не охарактеризован, но может быть связан с ошибками ДНК-полимераз при репликации продуктов дальнейшего окисления 8-oxoG.
Ключевые слова: повреждение ДНК; мутагенез; 8-оксогуанин; G-тракты; теломеры

Introduction

Oxidative DNA damage is an inevitable consequence of respiration, which relies on the oxidation of organic compounds with molecular oxygen and has been the basis of energy metabolism in the vast majority of living organisms for over two billion years (Prorok et al., 2021). Damaged nucleotides are generally quickly repaired; however, some of them may remain in DNA until replication, which is one of the main sources of mutations (Liu et al., 2016; Chatterjee, Walker, 2017; Tubbs, Nussenzweig, 2017). Based on our understanding of the molecular mechanisms of DNA polymerase errors, it has now become possible to identify characteristic patterns of mutations caused by various types of genotoxic stress or even by specific damaged bases (Alexandrov et al., 2013; Koh et al., 2021).

Of all DNA structural elements, the guanine base has the lowest redox potential (Cadet et al., 2008, 2017; Fleming, Burrows, 2022). The most common product of its oxidation, 7,8-dihydro-8-oxoguanine (8-oxoG), occurs in DNA at the background level of ~$1/10^6$ guanines, and this level increases significantly under oxidative stress of various origins (ESCODD et al., 2005; Dizdaroglu et al., 2015; Chiorcea-Paquim, 2022; Fig. 1*a*, *b*). The presence of an oxygen atom at C8 in 8-oxoG sterically hinders the regular *anti* conformation of its nucleoside, 8-oxo-2′-deoxyguanosine (8-oxodG), and the *syn* conformation becomes energetically favorable (Cho et al., 1990; Fig. 1*c*, *d*). Consequently, in the absence of Watson–Crick bonds with cytosine, which additionally stabilize the *anti* conformation, 8-oxodG preferentially adopts the *syn* conformation, in which it can form a Hoogsteen-type pair with adenine (Kouchakdjian et al., 1991; McAuley-Hecht et al., 1994; Lipscomb et al., 1995). Because of this, DNA polymerases incorporate dAMP opposite 8-oxoG in the DNA template with high frequency (Shibutani et al., 1991; Miller, Grollman, 1997; Maga et al., 2007; Yudkina et al., 2019).

In the living cell, the outcome of primary DNA oxidation events can be influenced by numerous additional factors and DNA repair systems that remove damaged bases from the genome. Even so, 8-oxoG exhibits relatively high mutagenicity *in vivo*, characterized by a spectrum dominated by G→T transversions mostly independent of the surrounding nucleotide context (Wood et al., 1992; Moriya, 1993). Such mutations are frequently found in human tumors and form the basis of the SBS18 and SBS36 mutational signatures (Alexandrov et al., 2013; Pilati et al., 2017; Viel et al., 2017; Kucab et al., 2019). Guanidinohydantoin and spiroiminodihydantoin, the products of further oxidation of 8-oxoG, also significantly contribute to mutagenesis, predominantly causing G→C transversions (Fleming, Burrows, 2017; Kino et al., 2020).

The stacked π system of DNA has considerable hole conductivity (Giese, 2002; Genereux, Barton, 2010). Numerous experiments and quantum mechanical calculations show that a positive charge resulting from one-electron oxidation of DNA can migrate along the π system over significant distances, and its final acceptors are the G bases, which are mainly oxidized to 8-oxoG. In this case, the G bases located in the first 5′-position in runs of several Gs are especially sensitive to oxidation (Sugiyama, Saito, 1996; Saito et al., 1998; Kurbanyan et al., 2003; Adhikary et al., 2009).

Although the mechanism of positive charge migration and preferential oxidation of guanines at the 5′-end of G-runs is generally accepted today, all experimental data supporting it

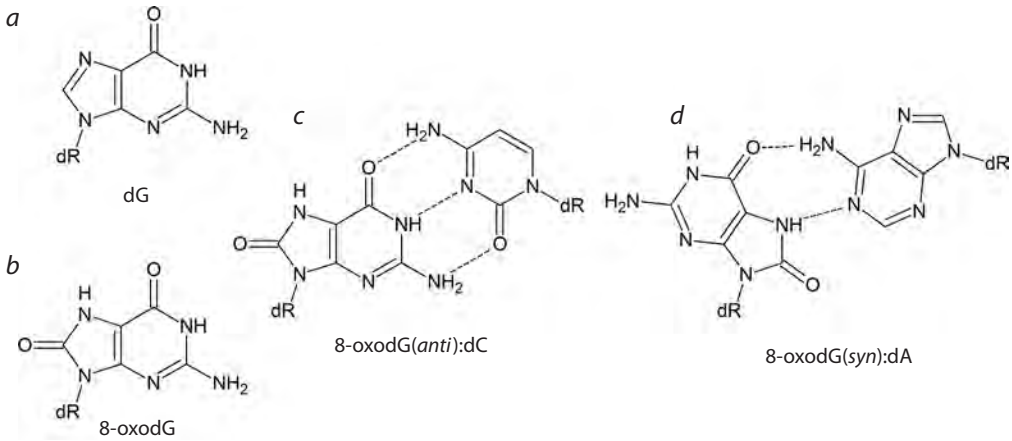


Fig. 1. Structures of 2′-deoxyguanosine (*a*), 8-oxo-2′-deoxyguanosine (*b*), Watson–Crick 8-oxodG(*anti*):dC pair (*c*) and Hoogsteen 8-oxodG(*syn*):dA pair (*d*).

were obtained in relatively simple *in vitro* systems. The mutagenesis spectra caused by the appearance of 8-oxoG in this context have not yet been studied. If preferential conversion of G to 8-oxoG does indeed occur at the 5′-end of G-runs, it can be expected that the mutagenic properties of 8-oxoG at these positions will result in an increased frequency of G→T mutations, which should be reflected in an increased frequency of T before G-runs. In this study, to test this hypothesis, we analyzed the occurrence of nucleotides flanking G-runs from the 5′-side in prokaryotic and human genomes.

Materials and methods

The T2T-CHM13v2.0 human genome assembly, which includes full-length telomeres and highly repetitive regions (Nurk et al., 2022), and the prokaryotic genomes listed in Table 1 were used for the analysis.

UGENE v37.0 software package (Okonechnikov et al., 2012) and custom-written bash scripts were used to extract nucleotide frequencies at given positions. The expected frequency of nucleotides in the flanking positions before and after G_n (or A_n) runs in prokaryotic genomes was calculated based on the total number of A, C, and T (or C, G, and T) in a given genome as $p_A = N_A/(N_A+N_C+N_T)$, where p_A is the expected representation (in this case, for A), and N_A, N_C, and N_T are the numbers of A, C, and T in both strands of the genome, respectively. For the human genome, due to the well-known underrepresentation of the CG dinucleotide, the expected frequency was calculated in a similar way but based on the number of AG, CG, and TG dinucleotides. Statistical analysis was performed using SigmaPlot v11.0 (Grafiti, USA), DATAPLOT (National Institute of Standards and Technology, USA), and RStudio v1.2 (Posit PBC, USA). Dunn's correction was used for all multiple comparisons and test series to adjust the significance level.

Results and discussion

To analyze the nucleotide distribution in prokaryotic genomes, a sample of 54 bacterial and 8 archaeal genomes was compiled, maximally reflecting the taxonomic diversity in these domains of life (Table 1). Only high-quality genomes classified in the RefSeq database (O'Leary et al., 2016) as reference genomes were included. The sample taxonomic representation was one genome per phylum, with the exception of Methanobacteriota and Thermoproteota for Archaea, and Actinomycetota, Bacteroidota, and Thermodesulfobacteriota for Bacteria with a representation of 2 genomes from different orders per phylum, as well as Bacillota and Pseudomonadota (3 genomes from different orders per phylum). The G+C content in the studied genomes ranged from 23.5 to 69 % (Table 1). The parameters of archaeal genomes did not differ significantly from those of bacterial ones, so the representatives of both domains were considered as a single group of prokaryotes.

Since the prokaryotic genomes mostly consist of protein-coding sequences, mutations in which can be subject to natural selection, we have first assessed the possible impact of all 16 potential amino acid substitutions resulting from G→A, G→C and G→T nucleotide substitutions in the first position of G-runs (codon changes HHG→HHH, HGG→HHG, GGG→HGG, where H is A, C or T). Two independent metrics were used for this purpose: the conservation index C_n, calculated on the basis of partition distances in a set of physicochemical properties of amino acid residues (Taylor, 1986; Livingstone, Barton, 1993), and the weights of amino acid substitutions in the BLOSUM62 matrix, compiled from several hundred groups of homologous proteins (Henikoff S., Henikoff J.G., 1992). Although G→A substitutions generally caused smaller changes in the properties and occurrence of amino acid residues, as expected for class-conserving point mutations, the difference from G→C and G→T substitutions was not statistically significant (Kruskal–Wallis test with Dunn's correction for multiple comparisons, $p > 0.05$).

All genomic sequences were searched for the HG_nH and BA_nB runs and the corresponding complementary-strand DC_nD and VT_nV runs (H = A, C or T; B = C, G or T; D = A, G or T; V = A, C or G) with the length $n \geq 2$. The frequency of polypurine runs in the genomes was higher than that expected from a random nucleotide distribution with the same G+C composition (one-sample Wilcoxon test, $p < 0.001$), indicating the functional importance of such sequences. An increased frequency of substitutions at the first position of G-runs should gradually lead to their shortening. Indeed, when comparing the lengths of G-runs and A-runs in prokaryotic genomes, adjusted for the content of the respective purine nucleotides, it turned out that G-runs are, on average, shorter (Fig. 2*a*). In this case, HGG trinucleotides were more common than BAA, but in longer repeats, the frequency of A-runs was higher (Fig. 2*b*).

For a more detailed analysis of the run length distribution, we have studied the variability of their lengths in each genome. The number of G-runs and A-runs in each genome decreased almost strictly exponentially in the length range from 2 to 5–6. At $n > 5$–6, deviations in either direction were observed in some cases due to the small number of such runs, especially in small genomes (Fig. 3*a*, *b*). Using the linear portion of the relationship between the log of the number of repeats and run length, one can determine the increment coefficient k_{inc}, which indicates how easily a run can be extended by one nucleotide in a genome with a given nucleotide composition: the higher the k_{inc}, the greater the proportion of longer runs in the genome. When comparing the dependence of k_{inc} for G-runs and A-runs in genomes of different composition, we have found that G-runs grow more slowly with increasing G+C content than A-runs grow with increasing A+T content (Fig. 3*c*). Thus, in prokaryotic genomes, the balance of G-run elongation and shortening, determined by many factors, is shifted towards shortening compared to A-runs.

The lengths of polypurine runs can change in either direction due to DNA polymerase slippage during DNA synthesis (Kunkel, Bebenek, 2000) or selection based on the physicochemical properties of polypurine regions (Bansal et al., 2022), but these processes are independent of the nucleotides surrounding the run. In contrast, shortening of G-runs due to damage to the 5′-terminal base should be accompanied by a characteristic mutational spectrum determined by the properties of replicative DNA polymerases. Therefore, it was of interest to determine the extent to which the frequencies of 5′-flanking nucleotides differ from each other and from their overall abundance in the genome. To quantitatively characterize these differences, we have introduced the Δrep parameter representing the difference between the observed

Table 1. Prokaryotic genomes used for the analysis

Species	Phylum	Genome assembly	G+C, %	O_2
Archaea domain				
Methanobacterium formicicum	Methanobacteriota	GCF_001458655.1	41.0	AN
Methanosarcina barkeri		GCF_000970025.1	39.0	AN
Nanobdella aerobiophila	Nanobdellota	GCF_023169545.1	24.5	A
Nitrososphaera viennensis	Nitrososphaerota	GCF_000698785.1	52.5	A
Promethearchaeum syntrophicum	Promethearchaeota	GCF_008000775.2	31.0	AN
Sulfolobus acidocaldarius	Thermoproteota	GCF_000012285.1	36.5	A
Thermoproteus tenax		GCF_000253055.1	55.0	AN
Cand. *Nanohalobium constans*	Cand. Nanohalarchaeota	GCF_009617975.1	43.0	A
Bacteria domain				
Acidobacterium capsulatum	Acidobacteriota	GCF_000022565.1	60.5	A
Bifidobacterium longum	Actinomycetota	GCF_000196555.1	60.5	AN
Mycobacterium tuberculosis		GCF_000195955.2	65.5	A
Aquifex aeolicus	Aquificota	GCF_000008625.1	43.5	A
Fimbriimonas ginsengisoli	Armatimonadota	GCF_000724625.1	61.0	A
Atribacter laminatus	Atribacterota	GCF_015775515.1	38.5	AN
Bacillus subtilis	Bacillota	GCF_000009045.1	43.5	A
Clostridioides difficile		GCF_018885085.1	28.5	AN
Lactococcus lactis		GCF_003176835.1	35.0	A
Bacteroides fragilis	Bacteroidota	GCF_000025985.1	43.0	AN
Saprospira grandis		GCF_000250635.1	46.5	A
Cyclonatronum proteinivorum	Balneolota	GCF_003353065.1	51.5	A
Bdellovibrio bacteriovorus	Bdellovibrionota	GCF_000196175.1	50.5	A
Caldisericum exile	Caldisericota	GCF_000284335.1	35.5	AN
Calditrix abyssi	Calditrichota	GCF_001886815.1	45.0	AN
Campylobacter jejuni	Campylobacterota	GCF_000009085.1	30.5	A
Chlamydia trachomatis	Chlamydiota	GCF_000008725.1	41.5	AN
Chlorobium limicola	Chlorobiota	GCF_000020465.1	51.5	AN
Chloroflexus aurantiacus	Chloroflexota	GCF_000018865.1	56.5	A
Desulfurispirillum indicum	Chrysiogenota	GCF_000177635.2	56.0	AN
Coprothermobacter proteolyticus	Coprothermobacterota	GCF_000020945.1	45.0	AN
Synechococcus elongatus	Cyanobacteriota	GCF_022984195.1	55.5	A
Deferribacter thermophilus	Deferribacterota	GCF_049472675.1	30.5	AN
Deinococcus radiodurans	Deinococcota	GCF_020546685.1	66.5	A
Dictyoglomus thermophilum	Dictyoglomota	GCF_000020965.1	33.5	AN
Elusimicrobium minutum	Elusimicrobiota	GCF_000020145.1	40.0	AN
Fibrobacter succinogenes	Fibrobacterota	GCF_000146505.1	48.0	AN
Fidelibacter multiformis	Fidelibacterota	GCF_041154365.1	45.5	AN
Fusobacterium nucleatum	Fusobacteriota	GCF_003019295.1	27.0	AN
Gemmatimonas aurantiaca	Gemmatimonadota	GCF_000010305.1	64.5	A
Ignavibacterium album	Ignavibacteriota	GCF_000258405.1	34.0	A
Kiritimatiella glycovorans	Kiritimatiellota	GCF_001017655.1	63.5	AN
Lentisphaera profundi	Lentisphaerota	GCF_028728065.1	40.5	A
Mycoplasma mycoides	Mycoplasmatota	GCF_018389705.1	23.5	A
Myxococcus xanthus	Myxococcota	GCF_000012685.1	69.0	A
Nitrospina watsonii	Nitrospinota	GCF_946900835.1	57.0	A
Nitrospira moscoviensis	Nitrospirota	GCF_001273775.1	62.0	A
Planctopirus limnophila	Planctomycetota	GCF_000092105.1	53.5	A

Table 1 (end)

Species	Phylum	Genome assembly	G+C, %	O_2
Escherichia coli	Pseudomonadota	GCF_000005845.2	51.0	A
Pseudomonas aeruginosa		GCF_000006765.1	66.5	A
Sphingomonas paucimobilis		GCF_016027095.1	65.5	A
Rhodothermus marinus	Rhodothermota	GCF_000024845.1	64.5	A
Spirochaeta thermophila	Spirochaetota	GCF_000184345.1	61.0	AN
Thermanaerovibrio acidaminovorans	Synergistota	GCF_000024905.1	64.0	AN
Desulfovibrio desulfuricans	Thermodesulfobacteriota	GCF_017815575.1	57.0	AN
Thermodesulfobacterium commune		GCF_000734015.1	37.0	AN
Thermodesulfobium narugense	Thermodesulfobiota	GCF_000212395.1	34.0	AN
Thermomicrobium roseum	Thermomicrobiota	GCF_000021685.1	64.5	A
Thermosulfidibacter takaii	Thermosulfidibacterota	GCF_001547735.1	43.0	AN
Thermotoga maritima	Thermotogota	GCF_000230655.2	46.0	AN
Verrucomicrobium spinosum	Verrucomicrobiota	GCF_000172155.1	60.5	A
Vulcanimicrobium alpinum	Vulcanimicrobiota	GCF_027923555.1	68.5	A
Cand. *Cloacimonas acidaminovorans*	Cand. Cloacimonadota	GCF_000146065.2	38.0	AN
Cand. *Velamenicoccus archaeovorus*	Cand. Omnitrophota	GCF_004102945.1	53.0	AN

Note. Assembly ID in the RefSeq database (O'Leary et al., 2016). A, aerobes and facultative anaerobes; AN, anaerobes.

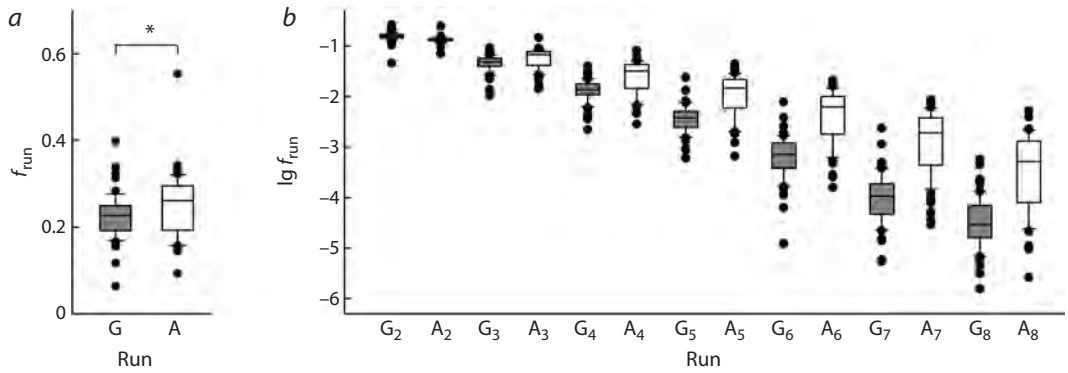


Fig. 2. Length of polypurine runs in prokaryotic genomes. *a*, the total fraction of G or A in runs of any length in the respective purine nucleotide content in the genome (f_{run}). * $p < 0.05$ (Mann–Whitney test). *b*, the fraction of G or A in the runs 2 to 8 nucleotides long in the respective purine nucleotide content in the genome. In all cases, the difference between G-runs and A-runs is significant at $p < 0.001$ (Mann–Whitney test).

Here and below, the line in the box marks the median, the boundaries of the box correspond to the first and third quartiles, the whiskers, to the 10th and 90th percentiles, and the dots are outliers.

and expected frequency of each nucleotide. The frequency of T in the first position before G-runs was statistically significantly higher than expected and than the frequency of A and C (Fig. 4*a*). The frequency of A and C nucleotides in this position was slightly lower than expected, but this difference did not reach significance; their representation also did not differ from each other. T was more frequent than either A or C nucleotide at any G-run length, and its representation was higher than expected before G_2, G_4, G_5, and G_6 runs (Fig. 4*b*). A was underrepresented in this position only before G_4 runs, and C was underrepresented before G_5 and longer G-runs. In contrast, T was underrepresented both at the 3′-side of G-runs and at the second position from their 5′-side (Fig. 4*a*). Overall, these data support a model of preferential oxidation of the first G in the runs to 8-oxoG followed by G→T transversions.

Quite unexpectedly, the nucleotide distribution before A-runs was even more uneven than before G-runs. At this position, T was underrepresented, while C and G were overrepresented (Fig. 4*c*). For C, this deviation was explained primarily by overrepresentation of CAA trinucleotides, while for G, an increased frequency of occurrence was observed up to a run length of 6 nucleotides (Fig. 4*d*). A decrease in the fraction of T also occurred in runs of any length (Fig. 4*d*). After A-runs, the occurrence of C and T was lower than expected, while G was higher than expected (Fig. 4*c*). It is possible that

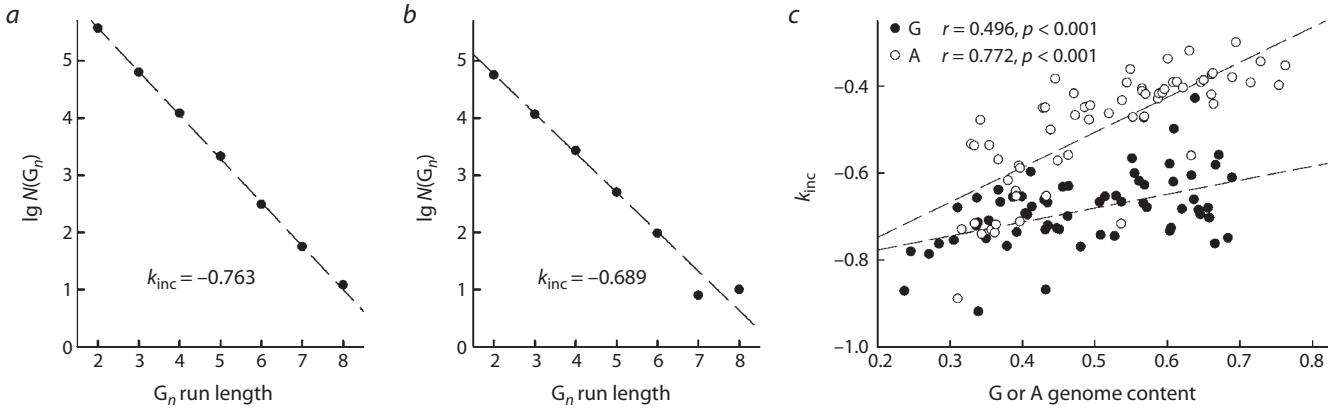


Fig. 3. Dependence of the number of polypurine runs in prokaryotic genomes on the run length and the nucleotide composition of the genome. *a*, *b*, examples of the dependence of the number of G-runs $N(G_n)$ on their length for the genomes of *E. coli* (*a*; genome size 4.64×10^6 bp, G+C content 51.0 %) and *Ch. trachomatis* (*b*; genome size 1.04×10^6 bp, G+C content 41.5 %). *c*, dependence of k_{inc} on the nucleotide composition of the genome (G+C content for G-runs, A+T content for A-runs). Black dots, G-runs, white dots, A-runs; dashed lines show linear regressions with the regression coefficients indicated on the plot.

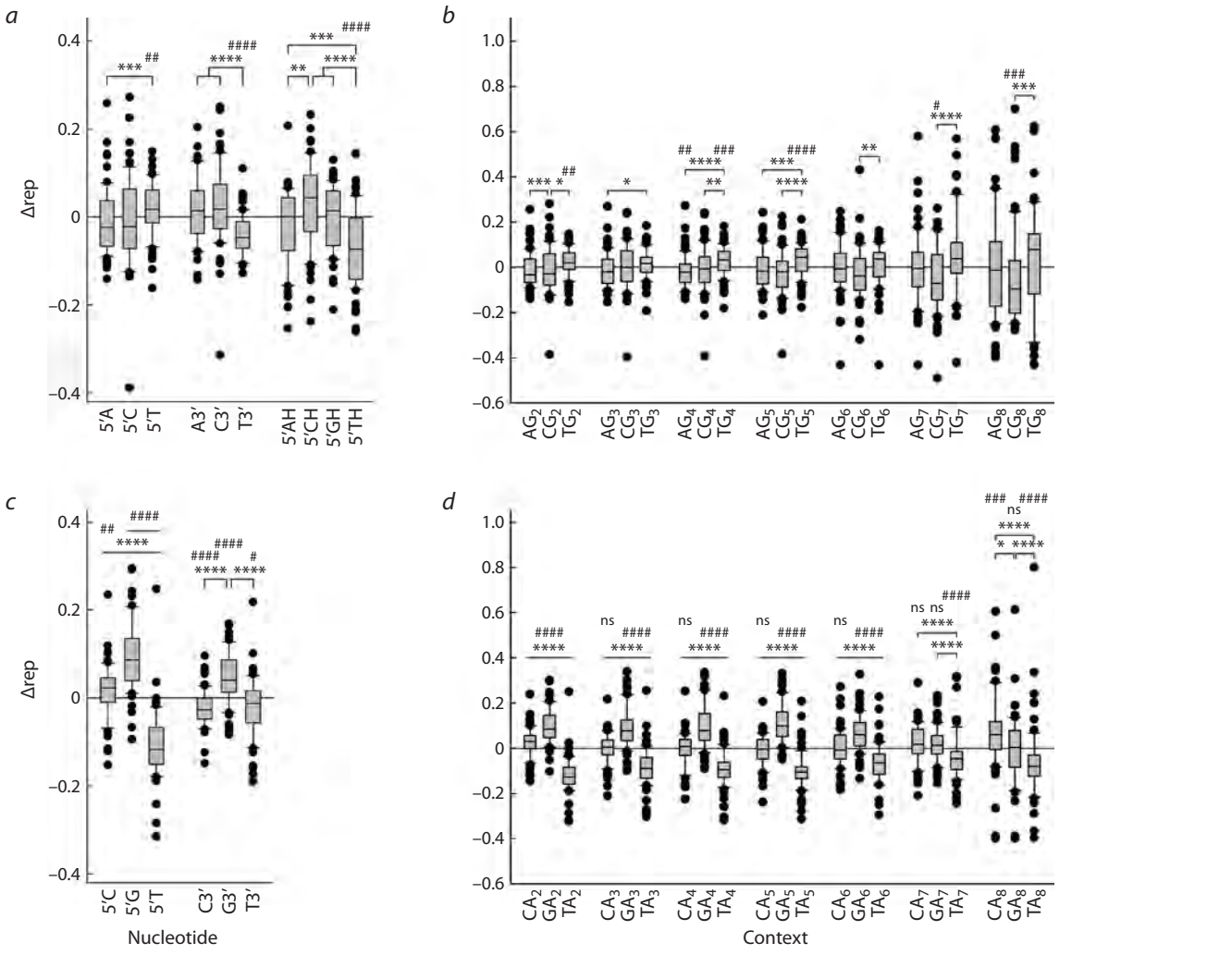


Fig. 4. Representation of different 5′- and 3′-flanking nucleotides in polypurine runs. *a*, *c*, deviation from the frequency of 5′- and 3′-flanking nucleotides for G-runs (*a*) and A-runs (*c*) of any length expected on the basis of the content of the respective nucleotide in the genome. *b*, *d*, deviation from the frequency of 5′-flanking nucleotides in G-runs (*b*) and A-runs (*d*) 2–8 nucleotides long.

Difference from expected: # $p < 0.05$, ## $p < 0.01$, ### $p < 0.005$, #### $p < 0.001$ (one-sample Wilcoxon test with Dunn's correction for multiple comparisons); ns, no significant difference. Differences between groups: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$ (Kruskal–Wallis test with Dunn's correction for multiple comparisons).

these deviations can also be explained by DNA damage and subsequent DNA polymerases errors; however, the mechanistic reasons underlying such events remain unclear at present.

The amount of 8-oxoG generated in the genome directly depends on the presence of reactive oxygen species in the intracellular environment (Halliwell, Gutteridge, 2015). Prokaryotes are exceptionally diverse in their energy metabolism pathways: some follow a strictly anaerobic lifestyle, while others are obligate aerobes or facultative anaerobes and are subject to more intense oxidative stress. We have compared the statistics of the occurrence of 5′-flanking nucleotides of G-runs in the genomes of these two groups (Table 1). In aerobic prokaryotes, T was found at this position with an increased frequency compared to the expected, and A, with a decreased frequency (Fig. 5). For anaerobic microorganisms, no significant difference in the occurrence of 5′-flanking nucleotides was found (Fig. 5). However, when comparing the abundance of A, C and T directly between the aerobic and anaerobic groups, the differences did not reach statistical significance, which is most likely due to insufficient sample size. For A-runs, the difference in the occurrence of 5′-flanking nucleotides in the genomes of aerobes and anaerobes was the same as in the combined group (compare Fig. 4*c* and Fig. 5). Thus, the reduced level of oxidative stress in anaerobic microorganisms may be associated with a less pronounced predominance of T at the position flanking the 5′-side of G-runs; however, further research is required to answer this question more definitively.

Unlike those of prokaryotes, eukaryotic genomes are characterized by a large number of repetitive elements such as transposons, satellite and telomeric DNA, the precise sequences of which are inaccessible to traditional high-throughput sequencing methods (Richard et al., 2008; Liao et al., 2023). The advent of ultra-long sequencing (Oxford Nanopore, PacBio HiFi) has made it possible to fill these gaps. The recently published human genome read using a combination of methods with telomere-to-telomere (T2T) coverage and high quality (estimated telomeric error rate of $\sim 4\times10^{-8}$) (Nurk et al., 2022), provides the opportunity to analyze the context of G-runs without the distortions caused by a higher representation of unique sequences.

The significantly larger size of the human genome compared to prokaryotic ones allowed us to identify interesting patterns in the distribution of G_n runs size. For n = 2–8, their number decreased exponentially and was described by an increment coefficient $k_{inc} = -0.674$, which is very close to the center of the distribution of k_{inc} values for G-runs in prokaryotes (compare Fig. 6*a* and Fig. 3*c*; z = 0.141). For n = 9–16, the exponential dependence was preserved, but the rate of decrease in the number of runs decelerated: the k_{inc} value increased to −0.198, which lies far outside the range of k_{inc} values for prokaryotic genomes (compare Fig. 6*a* and Fig. 3*c*; z = 5.97). Runs of this size were absent in prokaryotic genomes or were present in a handful of cases, so it was impossible to detect this transition. Further increase in the length of G-runs was accompanied by an even greater deceleration of the rate of decrease in their number (Fig. 6*a*). Obviously, around n = 8–9 (the breakpoint value determined by the piecewise regression method: $n = 8.72 \pm 0.04$), the balance of G-run

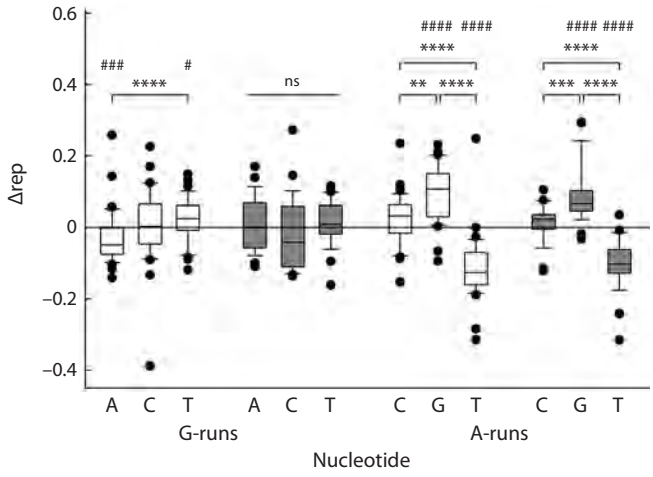


Fig. 5. Representation of 5′-flanking nucleotides in polypurine runs in the genomes of aerobic (white) and anaerobic (gray) microorganisms.

Deviation from the expected representation is shown for G-runs and A-runs of any length. Difference from expected: # $p < 0.05$, ### $p < 0.005$, #### $p < 0.001$ (one-sample Wilcoxon test with Dunn's correction for multiple comparisons). Differences between groups: ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$ (Kruskal–Wallis test with Dunn's correction for multiple comparisons); ns, no significant difference.

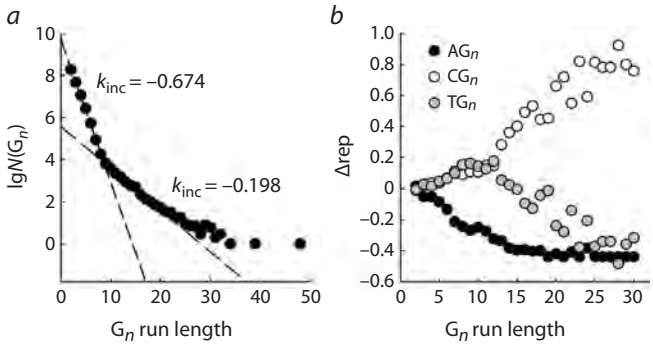


Fig. 6. Dependence of the number of G-runs and the representation of the 5′-flanking nucleotides in the human genome on the run length. *a*, dependence of the number of G-runs $N(G_n)$ on their length. The dashed lines correspond to the linear regression; the k_{inc} values for n = 2–8 and n = 9–16 are shown on the plot. *b*, The dependence of Δrep of the 5′-flanking nucleotide (black, A, white, C, gray, T) on the run length.

shortening and elongation is shifted in favor of the latter; run growth due to DNA polymerase slippage during replication or repair becomes self-sustaining, as in the well-studied case of trinucleotide repeat runs (Mirkin, 2007; McMurray, 2010).

An even more unexpected pattern emerged from the analysis of the frequency of 5′-flanking nucleotides. Since it is well known that the number of CG dinucleotides in the human genome is reduced due to their role in epigenetic regulation (Fazzari, Greally, 2004), the expected frequency was calculated based on the dinucleotide rather than the total nucleotide frequency. At n = 2, the nucleotide frequency closely matched the expected value, but then the Δrep values for A steadily decreased, while the representation of C and T, in contrast, increased at virtually the same rate (Fig. 6*b*). However, starting from n = 8–11 (the breakpoint value for Δrep(C)−Δrep(T), de-

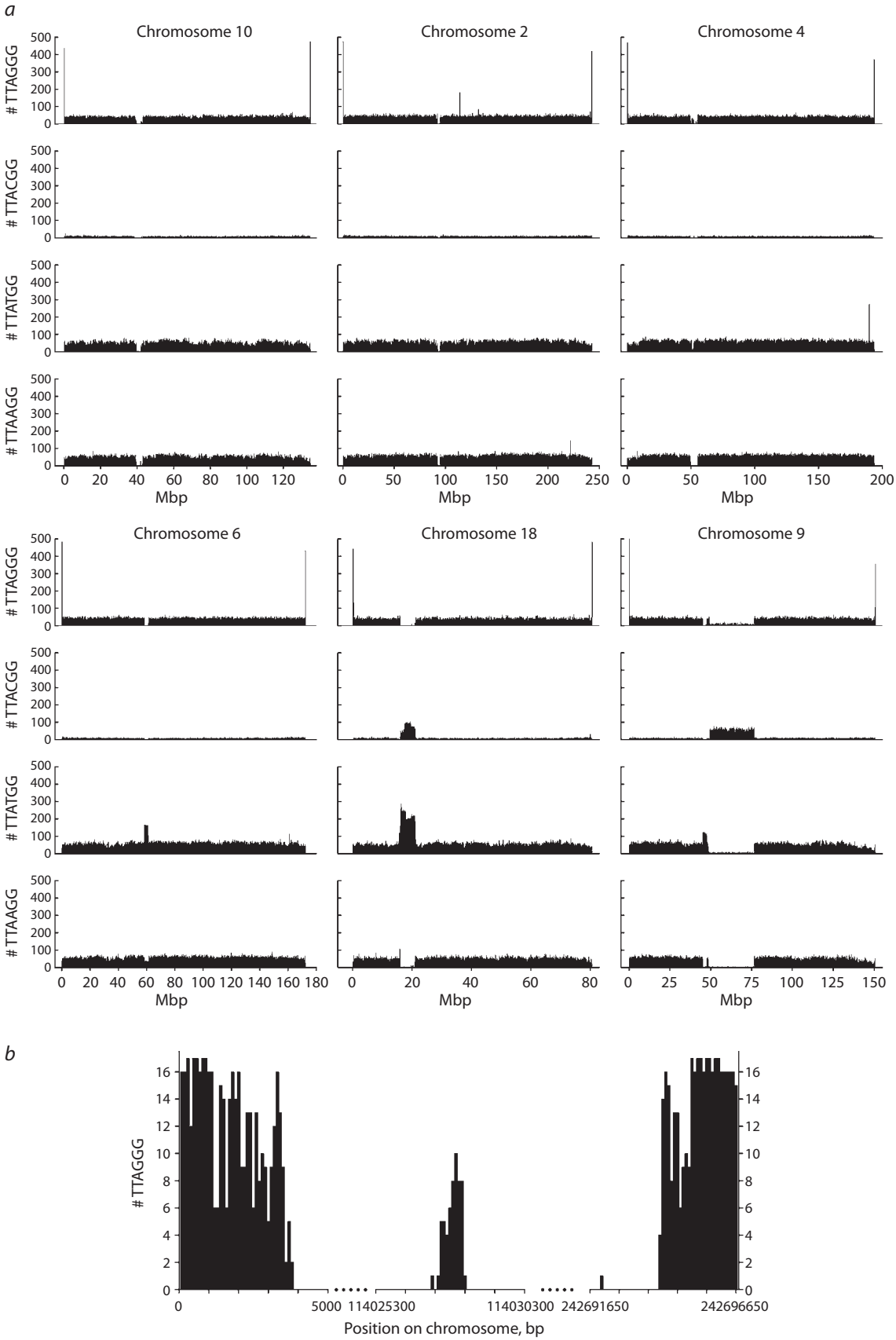


Fig. 7. Examples of the distribution of TTAGGG, TTACGG, TTATGG, and TTAAGG repeats in human chromosomes. *a*, distribution of the repeats along the entire length of chromosomes 10, 2, 4, 6, 18, and 9. The number of repeats is calculated in 100-kb bins. *b*, distribution of TTAGGG repeats in telomeric regions and in the region of fusion of the ancestral telomeres on chromosome 2. The number of repeats is calculated in 100-bp bins.

termined by the piecewise regression method: $n = 9.28 \pm 1.10$), the dependencies for C and T diverged sharply: the representation of T decreased, while the representation of C increased. One possible explanation for this phenomenon may be that longer G-runs serve as more effective traps for holes migrating along the DNA duplex leading to hyperoxidation of the 5′-terminal 8-oxoG to guanidinohydantoin and spiroiminodihydantoin with a corresponding switch in the preferential nucleotide substitutions from G→T to G→C.

Telomeric DNA is a distinct class of highly repetitive DNA in eukaryotic genomes, represented in humans by multiple copies of the TTAGGG hexanucleotide. Telomeric repeats are known to serve as hotspots for DNA oxidation to form 8-oxoG (Billard, Poncet, 2019; Opresko et al., 2025). Telomere ends in germline cells are elongated by telomerase, a specialized DNA polymerase that uses telomerase RNA as a template, so changes in these regions are not associated with damage to genomic DNA. However, even in the presence of active telomerase, the bulk of telomere length is replicated by the usual semiconservative mechanism (Pfeiffer, Lingner, 2013; Higa et al., 2017; Bonnell et al., 2021), which can lead to the accumulation of mutations in them. Thus, the telomere sequence in human somatic cells (in the case of the T2T genome, the immortalized telomerase-expressing CHM13hTERT chorionic cell line) reflects both their recent elongation by telomerase in germline cells and mutagenesis events in past generations and in individual development.

The distribution of TTAGGG repeats in chromosomes (calculated for both DNA strands) had a fairly expected pattern, with frequency peaks at the ends of the chromosomes and a dip in the pericentromeric region (Fig. 7*a*). The only exception was chromosome 8, for which, on the contrary, a slight increase in the number of these repeats was observed in the centromere region. On chromosome 2, a peak in the frequency of telomeric repeats was clearly visible in the region of the fusion of two ancestral hominid chromosomes that formed the evolutionarily young human chromosome 2 (Ijdo et al., 1991; Fig. 7*a*). However, a more detailed analysis of this region shows that it has already significantly degraded, keeping far fewer TTAGGG repeats than in true telomeres (Fig. 7*b*). Interestingly, similar peaks were found on chromosomes 15 and 22 in the introns of the active protein-coding genes *ATP10A* and *MICAL3*; they may represent remnants of translocated telomere fragments.

TTAAGG, TTACGG, and TTATGG repeats were distributed across chromosomes without telomeric peaks. The overall frequency of TTACGG repeats was significantly lower than that of TTAAGG and TTATGG, consistent with the reduced abundance of CG dinucleotides in the human genome (Fig. 7*a*). Separate peaks in repeat frequency were observed on chromosome 2 for TTAAGG, chromosomes 8, 12, 17, and Y for TTACGG, and chromosomes 4 and 22 for TTATGG (Fig. 7*a*). A characteristic pattern of repeat distribution in the pericentromeric region with gaps in all TTANGG variants was observed for chromosomes 1–5, 7, 10–12, 16, 19, and 21. In other cases, one repeat type predominated in the centromere region, while others were depleted, with their combined deficiency compensating for the excess of the predominant type, as shown in Fig. 7*a* for chromosome 6. In chromosomes 6, 13–15, 22, and X, TTATGG was the predominant repeat; in chromosome 8, it was TTAGGG, and in chromosome 17, TTACGG. Chromosome 18 was distinguished by coinciding peaks in the distribution of two repeats, TTACGG and TTATGG (Fig. 7*a*). In the long arm of chromosome 9, in the region of constitutive heterochromatin adjacent to the pericentromeric region with an excess of TTATGG, there was a long stretch with a predominance of TTACGG.

Obviously, the cases of co-localization or oppositely phased localization of TTANGG repeats in non-telomeric regions are not due to point mutations in the TTAGGG repeat but reflect the presence of repeating elements containing one or two of these hexanucleotides in these loci. In contrast, point mutations in the first position of the G_3-run of the telomeric repeat should be most obvious in the regions consisting mainly of TTAGGG, that is, in the telomeres proper and intrachromosomal blocks of telomere-like repeats. To analyze the frequency of substitutions in such regions, we have singled out the telomeric regions and intrachromosomal blocks where at least 100 copies of the TTAGGG repeat were found in 100-kb bins. They were divided into shorter 100-bp bins. A bin filled with only TTAGGG repeats corresponds to 16 or 17 copies (depending on the position of the first complete hexanucleotide in the bin). The bins containing at least 9 TTAGGG copies, accounting for more than half the bin length, were selected for analysis.

Counting the occurrence of TTAAGG, TTACGG, and TTATGG in the studied regions revealed clear significant enrichment of G→A substitutions at the first position of the G_3-run compared to G→C and G→T substitutions (Table 2). In comparison with G→A, the total number of G→C and G→T changes was fivefold lower, and their frequencies did not differ significantly from each other. Thus, although telomeric repeats serve as preferential sites of guanine oxidation, this is not reflected in the increased frequency of G→T point mutations. The difference between the representation of A and C+T at the 5′-flanking position of GG dinucleotides between telomeric repeats and the rest of the genome may indicate the existence of a mutational process in telomeres that is distinct from G oxidation at the 5′-position of GGG.

Conclusion

In conclusion, the analysis of the nucleotide context of G-runs in a set of 62 complete prokaryotic genomes and in the human T2T genome revealed that the representation of T at the position adjacent to G-runs is generally increased, which is consistent with the model of G oxidation at the 5′-position of the runs followed by G→T mutations. Other patterns in the distribution of 5′-flanking nucleotides were also identified: uneven nucleotide frequency at the position adjacent to A-runs, increased representation of C at the 5′-side of long G-runs in the human genome, and the predominance of G→A substitutions at the 5′-position in telomeric repeats. The hypothesis that G-run elongation may lead to a shift in the specificity of single-nucleotide mutations from G→T to G→C due to a change in the nature of the precursor lesion can be tested experimentally. The characteristic mutation spectrum in telomeric repeats may be caused by their tendency to fold into G-quadruplex structures, which hinder the movement of DNA polymerases (Pfeiffer, Lingner, 2013; Higa et al., 2017;

Table 2. Representation of TTANGG in telomeres and intrachromosomal blocks of telomere-like repeats

Number of TTAGGG in 100-nt bins	TTAGGG	TTAAGG	TTACGG	TTATGG	$\chi^2_{(A=C=T)}$*	$\chi^2_{(A=T)}$**
Telomeres						
17	7361	0	0	0	–	–
16	6512	3	0	0	0.0498	0.0833
15	1935	9	1	1	0.00297	0.0114
14	980	26	0	3	8.12×10^{-10}	1.95×10^{-5}
13	715	23	1	2	1.85×10^{-8}	2.67×10^{-5}
12	468	14	2	2	3.35×10^{-4}	0.00270
11	341	6	0	0	0.00248	0.0143
10	490	16	1	0	6.97×10^{-7}	6.33×10^{-5}
9	327	8	0	0	3.35×10^{-4}	0.00468
Intrachromosomal blocks						
16–17	16	0	0	0	–	–
15	45	0	1	0	0.368	–
11–14	221	0	0	0	–	–
10	90	2	1	2	0.818	1.000
9	99	10	0	4	0.00439	0.109
Combined						
17	7361	0	0	0	–	–
16	6528	3	0	0	0.0498	0.0833
15	1980	9	2	1	0.00865	0.0114
14	1022	26	0	3	8.12×10^{-10}	1.95×10^{-5}
13	767	23	1	2	1.85×10^{-8}	2.67×10^{-5}
12	540	14	2	2	3.35×10^{-4}	0.00270
11	396	6	0	0	0.00248	0.0143
10	580	18	2	2	8.84×10^{-6}	3.35×10^{-4}
9	423	18	0	4	5.12×10^{-6}	0.00284

* χ^2 values for the null hypothesis of equal representation of A, C and T.
** χ^2 values for the null hypothesis of equal representation of A and T.

Bonnell et al., 2021), but this proposal requires a detailed study of the fidelity of human replicative DNA polymerases on intact and damaged templates of this structure. For A-runs, the existence of preferential sites of DNA damage is not known; given that A-runs are longer than G-runs (Fig. 2), the difference in the relative representation of C, G, and T in the 5′-flanking position may not be associated with the mutational process. The explanation of all these identified patterns requires further research.

References

Adhikary A., Khanduri D., Sevilla M.D. Direct observation of the hole protonation state and hole localization site in DNA-oligomers. *J Am Chem Soc*. 2009;131(24):8614-8619. doi 10.1021/ja9014869

Alexandrov L.B., Nik-Zainal S., Wedge D.C., Aparicio S.A.J.R., Behjati S., Biankin A.V., Bignell G.R., … Campo E., Shibata T., Pfister S.M., Campbell P.J., Stratton M.R. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421. doi 10.1038/nature12477

Bansal A., Kaushik S., Kukreti S. Non-canonical DNA structures: diversity and disease association. *Front Genet*. 2022;13:959258. doi 10.3389/fgene.2022.959258

Billard P., Poncet D.A. Replication stress at telomeric and mitochondrial DNA: common origins and consequences on ageing. *Int J Mol Sci*. 2019;20(19):4959. doi 10.3390/ijms20194959

Bonnell E., Pasquier E., Wellinger R.J. Telomere replication: solving multiple end replication problems. *Front Cell Dev Biol*. 2021;9:668171. doi 10.3389/fcell.2021.668171

Cadet J., Douki T., Ravanat J.-L. Oxidatively generated damage to the guanine moiety of DNA: mechanistic aspects and formation in cells. *Acc Chem Res*. 2008;41(8):1075-1083. doi 10.1021/ar700245e

Cadet J., Davies K.J.A., Medeiros M.H.G., Di Mascio P., Wagner J.R. Formation and repair of oxidatively generated damage in cellular DNA. *Free Radic Biol Med*. 2017;107:13-34. doi 10.1016/j.freeradbiomed.2016.12.049

Chatterjee N., Walker G.C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen.* 2017;58(5):235-263. doi 10.1002/em.22087

Chiorcea-Paquim A.-M. 8-oxoguanine and 8-oxodeoxyguanosine biomarkers of oxidative DNA damage: a review on HPLC–ECD determination. *Molecules.* 2022;27(5):1620. doi 10.3390/molecules27051620

Cho B.P., Kadlubar F.F., Culp S.J., Evans F.E. ^{15}N nuclear magnetic resonance studies on the tautomerism of 8-hydroxy-2′-deoxyguanosine, 8-hydroxyguanosine, and other C8-substituted guanine nucleosides. *Chem Res Toxicol.* 1990;3(5):445-452. doi 10.1021/tx00017a010

Dizdaroglu M., Coskun E., Jaruga P. Measurement of oxidatively induced DNA damage and its repair, by mass spectrometric techniques. *Free Radic Res.* 2015;49(5):525-548. doi 10.3109/10715762.2015.1014814

ESCODD (European Standards Committee on Oxidative DNA Damage), Gedik C.M., Collins A. Establishing the background level of base oxidation in human lymphocyte DNA: results of an interlaboratory validation study. *FASEB J.* 2005;19(1):82-84. doi 10.1096/fj.04-1767fje

Fazzari M.J., Greally J.M. Epigenomics: beyond CpG islands. *Nat Rev Genet.* 2004;5(6):446-455. doi 10.1038/nrg1349

Fleming A.M., Burrows C.J. Formation and processing of DNA damage substrates for the hNEIL enzymes. *Free Radic Biol Med.* 2017;107:35-52. doi 10.1016/j.freeradbiomed.2016.11.030

Fleming A.M., Burrows C.J. Chemistry of ROS-mediated oxidation to the guanine base in DNA and its biological consequences. *Int J Radiat Biol.* 2022;98(3):452-460. doi 10.1080/09553002.2021.2003464

Genereux J.C., Barton J.K. Mechanisms for DNA charge transport. *Chem Rev.* 2010;110(3):1642-1662. doi 10.1021/cr900228f

Giese B. Long-distance electron transfer through DNA. *Annu Rev Biochem.* 2002;71:51-70. doi 10.1146/annurev.biochem.71.083101.134037

Halliwell B., Gutteridge J.M.C. Free Radicals in Biology and Medicine. Oxford Univ. Press, 2015

Henikoff S., Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(22):10915-10919. doi 10.1073/pnas.89.22.10915

Higa M., Fujita M., Yoshida K. DNA replication origins and fork progression at mammalian telomeres. *Genes.* 2017;8(4):112. doi 10.3390/genes8040112

Ijdo J.W., Baldini A., Ward D.C., Reeders S.T., Wells R.A. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc Natl Acad Sci USA.* 1991;88(20):9051-9055. doi 10.1073/pnas.88.20.9051

Kino K., Kawada T., Hirao-Suzuki M., Morikawa M., Miyazawa H. Products of oxidative guanine damage form base pairs with guanine. *Int J Mol Sci.* 2020;21(20):7645. doi 10.3390/ijms21207645

Koh G., Degasperi A., Zou X., Momen S., Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer.* 2021;21(10):619-637. doi 10.1038/s41568-021-00377-7

Kouchakdjian M., Bodepudi V., Shibutani S., Eisenberg M., Johnson F., Grollman A.P., Patel D.J. NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7*H*-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7*H*-dG(*syn*)·dA(*anti*) alignment at lesion site. *Biochemistry.* 1991;30(5):1403-1412. doi 10.1021/bi00219a034

Kucab J.E., Zou X., Morganella S., Joel M., Nanda A.S., Nagy E., Gomez C., Degasperi A., Harris R., Jackson S.P., Arlt V.M., Phillips D.H., Nik-Zainal S. A compendium of mutational signatures of environmental agents. *Cell.* 2019;177(4):821-836.e816. doi 10.1016/j.cell.2019.03.001

Kunkel T.A., Bebenek K. DNA replication fidelity. *Annu Rev Biochem.* 2000;69:497-529. doi 10.1146/annurev.biochem.69.1.497

Kurbanyan K., Nguyen K.L., To P., Rivas E.V., Lueras A.M.K., Kosinski C., Steryo M., González A., Mah D.A., Stemp E.D.A. DNA-protein cross-linking via guanine oxidation: dependence upon protein and photosensitizer. *Biochemistry.* 2003;42(34):10269-10281. doi 10.1021/bi020713p

Liao X., Zhu W., Zhou J., Li H., Xu X., Zhang B., Gao X. Repetitive DNA sequence detection and its role in the human genome. *Commun Biol.* 2023;6:954. doi 10.1038/s42003-023-05322-y

Lipscomb L.A., Peek M.E., Morningstar M.L., Verghis S.M., Miller E.M., Rich A., Essigmann J.M., Williams L.D. X-ray structure of a DNA decamer containing 7,8-dihydro-8-oxoguanine. *Proc Natl Acad Sci USA.* 1995;92(3):719-723. doi 10.1073/pnas.92.3.719

Liu B., Xue Q., Tang Y., Cao J., Guengerich F.P., Zhang H. Mechanisms of mutagenesis: DNA replication in the presence of DNA damage. *Mutat Res.* 2016;768:53-67. doi 10.1016/j.mrrev.2016.03.006

Livingstone C.D., Barton G.J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci.* 1993;9(6):745-756. doi 10.1093/bioinformatics/9.6.745

Maga G., Villani G., Crespan E., Wimmer U., Ferrari E., Bertocci B., Hübscher U. 8-oxo-guanine bypass by human DNA polymerases in the presence of auxiliary proteins. *Nature.* 2007;447(7144):606-608. doi 10.1038/nature05843

McAuley-Hecht K.E., Leonard G.A., Gibson N.J., Thomson J.B., Watson W.P., Hunter W.N., Brown T. Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs. *Biochemistry.* 1994;33(34):10266-10270. doi 10.1021/bi00200a006

McMurray C.T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet.* 2010;11(11):786-799. doi 10.1038/nrg2828

Miller H., Grollman A.P. Kinetics of DNA polymerase I (Klenow fragment exo$^-$) activity on damaged DNA templates: effect of proximal and distal template damage on DNA synthesis. *Biochemistry.* 1997;36(49):15336-15342. doi 10.1021/bi971927n

Mirkin S.M. Expandable DNA repeats and human disease. *Nature.* 2007;447(7147):932-940. doi 10.1038/nature05977

Moriya M. Single-stranded shuttle phagemid for mutagenesis studies in mammalian cells: 8-oxoguanine in DNA induces targeted G·C→T·A transversions in simian kidney cells. *Proc Natl Acad Sci USA.* 1993;90(3):1122-1126. doi 10.1073/pnas.90.3.1122

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., … Zook J.M., Schatz M.C., Eichler E.E., Miga K.H., Phillippy A.M. The complete sequence of a human genome. *Science.* 2022;376(6588):44-53. doi 10.1126/science.abj6987

Okonechnikov K., Golosova O., Fursov M.; UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics.* 2012;28(8):1166-1167. doi 10.1093/bioinformatics/bts091

O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., … Tatusova T., DiCuccio M., Kitts P., Murphy T.D., Pruitt K.D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44(D1):D733-D745. doi 10.1093/nar/gkv1189

Opresko P.L., Sanford S.L., De Rosa M. Oxidative stress and DNA damage at telomeres. *Cold Spring Harb Perspect Biol.* 2025;17(6):a041707. doi 10.1101/cshperspect.a041707

Pfeiffer V., Lingner J. Replication of telomeres and the regulation of telomerase. *Cold Spring Harb Perspect Biol.* 2013;5(5):a010405. doi 10.1101/cshperspect.a010405

Pilati C., Shinde J., Alexandrov L.B., Assié G., André T., Hélias-Rodzewicz Z., Ducoudray R., Le Corre D., Zucman-Rossi J., Emile J.-F., Bertherat J., Letouzé E., Laurent-Puig P. Mutational signature analysis identifies *MUTYH* deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol.* 2017;242(1):10-15. doi 10.1002/path.4880

Prorok P., Grin I.R., Matkarimov B.T., Ishchenko A.A., Laval J., Zharkov D.O., Saparbaev M. Evolutionary origins of DNA repair pathways: role of oxygen catastrophe in the emergence of DNA glycosylases. *Cells*. 2021;10(7):1591. doi 10.3390/cells10071591

Richard G.-F., Kerrest A., Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev*. 2008;72(4):686-727. doi 10.1128/MMBR.00011-08

Saito I., Nakamura T., Nakatani K., Yoshioka Y., Yamaguchi K., Sugiyama H. Mapping of the hot spots for DNA damage by one-electron oxidation: efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. *J Am Chem Soc*. 1998;120(48):12686-12687. doi 10.1021/ja981888i

Shibutani S., Takeshita M., Grollman A.P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*. 1991;349(6308):431-434. doi 10.1038/349431a0

Sugiyama H., Saito I. Theoretical studies of GG-specific photocleavage of DNA via electron transfer: significant lowering of ionization potential and 5′-localization of HOMO of stacked GG bases in B-form DNA. *J Am Chem Soc*. 1996;118(30):7063-7068. doi 10.1021/ja9609821

Taylor W.R. The classification of amino acid conservation. *J Theor Biol*. 1986;119(2):205-218. doi 10.1016/S0022-5193(86)80075-3

Tubbs A., Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell*. 2017;168(4):644-656. doi 10.1016/j.cell.2017.01.002

Viel A., Bruselles A., Meccia E., Fornasarig M., Quaia M., Canzonieri V., Policicchio E., … Maestro R., Giannini G., Tartaglia M., Alexandrov L.B., Bignami M. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine*. 2017;20:39-49. doi 10.1016/j.ebiom.2017.04.022

Wood M.L., Esteve A., Morningstar M.L., Kuziemko G.M., Essigmann J.M. Genetic effects of oxidative DNA damage: comparative mutagenesis of 7,8-dihydro-8-oxoguanine and 7,8-dihydro-8-oxoadenine in *Escherichia coli*. *Nucleic Acids Res*. 1992;20(22):6023-6032. doi 10.1093/nar/20.22.6023

Yudkina A.V., Shilkin E.S., Endutkin A.V., Makarova A.V., Zharkov D.O. Reading and misreading 8-oxoguanine, a paradigmatic ambiguous nucleobase. *Crystals*. 2019;9(5):269. doi 10.3390/cryst9050269

Conflict of interest. The authors declare no conflict of interest.
Received August 3, 2025. Revised October 1, 2025. Accepted October 17, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):925-939

doi 10.18699/vjgb-25-99

Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs

V.G. Levitsky [1, 2], T.Yu. Vatolina[2], V.V. Raditsa[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
levitsky@bionet.nsc.ru

Abstract. *De novo* motif search is the main approach for determining the nucleotide specificity of binding of the key regulators of gene transcription, transcription factors (TFs), based on data from massive genome-wide sequencing of their binding site regions *in vivo*, such as ChIP-seq. The number of motifs of known TF binding sites (TFBSs) has increased several times in recent years. Due to the similarity in the structure of the DNA-binding domains of TFs, many structurally cognate TFs have similar and sometimes almost indistinguishable binding site motifs. The classification of TFs by the structure of the DNA-binding domains from the TFClass database defines the top levels of the hierarchy (superclasses and classes of TFs) by the structure of these domains, and the next levels (families and subfamilies of TFs) by the alignments of amino acid sequences of domains. However, this classification does not take into account the similarity of TFBS motifs, whereas identification of valid TFs from massive sequencing data of TFBSs, such as ChIP-seq, requires working with TFBS motifs rather than TFs themselves. Therefore, in this study we extracted from the Hocomoco and Jaspar databases the TFBS motifs for human and fruit fly *Drosophila melanogaster*, and considered the pairwise similarity of binding site motifs of cognate TFs according to their classification from the TFClass database. We have shown that the common tree of the TF hierarchy by the structure of DNA-binding domains can be split into separate branches representing non-overlapping sets of TFs. Within each branch, the majority of TF pairs have significantly similar binding site motifs. Each branch can include one or more sister elementary units of the hierarchy and all its/their lower levels: one or more TFs of the same subfamily, or the whole subfamily, one or several subfamilies of the same family, an entire family, etc., up to the entire class. Analysis of the seven largest human and two largest Drosophila TF classes showed that the similarity of TFs in terms of TFBS motifs for different corresponding levels (classes, families) is noticeably different. Supplementing the hierarchical classification of TFs with branches combining significantly similar motifs of TFBSs can increase the efficiency of identifying involved TFs through enriched motifs detected by *de novo* motif search for massive sequencing data of TFBSs from the ChIP-seq technology.

Key words: *de novo* motif search; motifs of transcription factor binding sites; structural variants of motifs of transcription factor binding sites; similarity of motifs of transcription factor binding sites; cooperative action of transcription factors; massive whole-genome sequencing of transcription factor binding sites

For citation: Levitsky V.G., Vatolina T.Yu., Raditsa V.V. Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):925-939. doi 10.18699/vjgb-25-99

Funding. The work was supported by the Russian Science Foundation, grant No. 24-14-00133.

Acknowledgements. The software package development and data analysis was performed in part on the equipment of the Bioinformatics Shared Access Center within the framework of State Assignment Kurchatov Genomic Center of ICG SB RAS and supported by the budget project No. FWNR-2022-0020.

Связь иерархической классификации транскрипционных факторов по структуре ДНК-связывающего домена и вариабельности мотивов сайтов связывания этих факторов

В.Г. Левицкий [1, 2], Т.Ю. Ватолина[2], В.В. Радица[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
levitsky@bionet.nsc.ru

Аннотация. Поиск мотивов *de novo* – базовый подход определения нуклеотидной специфичности связывания важнейших регуляторов транскрипции генов, транскрипционных факторов (ТФ), на основе данных массового полногеномного секвенирования районов их сайтов связывания *in vivo*, таких как ChIP-seq. Количество известных мотивов сайтов связывания ТФ (ССТФ) возросло в несколько раз в последние годы. Из-за сходства структуры ДНК-связывающих доменов ТФ многие структурно родственные ТФ имеют сходные или даже неразличимые мотивы сайтов связывания. Классификация ТФ по структуре ДНК-связывающих доменов из базы данных TFClass определяет верхние уровни иерархии (суперклассы и классы ТФ) по структуре этих доменов, а следующие уровни (семейства и подсемейства ТФ) по выравниваниям аминокислотных последовательностей доменов. Однако эта классификация не учитывает сходство мотивов ССТФ, а для идентификации действующих ТФ по данным массового секвенирования ССТФ ChIP-seq приходится иметь дело с мотивами ССТФ, а не с самими ТФ. Поэтому в данной работе мы взяли из баз данных Hocomoco/Jaspar мотивы ССТФ человека/плодовой мушки *Drosophila melanogaster* и рассмотрели сходство мотивов сайтов связывания в парах родственных ТФ согласно их классификации в базе данных TFClass. Показано, что общее дерево иерархии ТФ по структуре ДНК-связывающих доменов можно разделить на отдельные неперекрывающиеся множества ТФ – ветви. В пределах каждой ветви большинство пар ТФ имеет значимо похожие мотивы сайтов связывания. Каждая ветвь включает одну или несколько сестринских элементарных единиц иерархии и все более низкие ее/их уровни: один или несколько ТФ одного подсемейства или целое подсемейство, одно или несколько подсемейств одного семейства, целое семейство и т.д. до целого класса. Анализ семи крупнейших классов ТФ человека и двух плодовой мушки показал, что сходство ТФ по мотивам ССТФ для разных соответствующих уровней (классов, семейств) заметно отличается. Дополнение иерархической классификации ТФ ветвями, объединяющими значимо сходные мотивы ССТФ, может повысить эффективность идентификации ТФ, вовлеченных в регуляцию транскрипции, по результатам *de novo* поиска обогащенных мотивов для данных массового секвенирования ССТФ с помощью технологии ChIP-seq.

Ключевые слова: *de novo* поиск мотивов; мотивы сайтов связывания транскрипционных факторов; структурные варианты мотивов сайтов связывания транскрипционных факторов; сходство мотивов сайтов связывания транскрипционных факторов; кооперативное действие транскрипционных факторов; массовое полногеномное секвенирование сайтов связывания транскрипционных факторов

Introduction

The study of the regulation mechanisms of eukaryotic genes transcription is necessary for understanding molecular genetic processes in the cell. Gene transcription is carried out under the control of special proteins, transcription factors (TFs), which regulate it specifically by the nucleotide context by binding to genomic DNA (Lambert et al., 2018). This specificity is due to nucleotide sequences of binding sites being recognized by individual TFs (TFBSs). The variability of binding sites reflects the ability of each TF to bind to different DNA sequences; therefore, the set of similar binding site sequences interacting with a TF is called the motif of its binding sites (D'haeseleer, 2006). The length of the region of genomic DNA directly interacting with an individual TF, as well as the length of the TFBS motif, usually vary from 6 to 20 base pairs (bp) (Spitz, Furlong, 2012; Zambelli et al., 2013; Vorontsov et al., 2024). One TF may have several distinct motifs of binding sites. The most popular model of the TFBS motif is the positional weight matrix (PWM). To build a model of the PWM motif, it is necessary to calculate the nucleotide frequencies at all positions using this alignment of the TFBSs representing this motif, and calculate the contributions (or weights) to the total estimate of affinity using these frequencies for each of the four nucleotides at each position. The total estimate of affinity for a potential site in a DNA sequence is equal to the sum of the weights corresponding to the nucleotides encountered, for all its positions (Wasserman, Sandelin, 2004).

Experimental ChIP-seq technology is based on chromatin immunoprecipitation (ChIP), i.e. application of antibodies to the target protein under study, for example, a TF. This technology is used to identify interactions of target proteins with genomic DNA *in vivo*. The essence of this technology is to perform chromatin immunoprecipitation and subsequently to map the genomic loci of the interaction between a target protein and genomic DNA. TFs *in vivo*, as a rule, act as part of multiprotein complexes formed by protein-protein interactions of several TFs, which allows them to regulate gene transcription together, even without direct connections of each of the TFs with genomic DNA. Therefore, *in vivo* TFs can bind to DNA in a variety of ways:

- directly, there is a binding site of the target TF in DNA;
- with another "partner" TF, binding sites for both target and partner TFs co-occur in DNA, they are found with a spacer or an overlap (Levitsky et al., 2019);
- indirectly, there is a binding site for a partner TF in DNA, and that for the target TF is absent (Slattery et al., 2014).

The individual genomic loci mapped in a ChIP-seq experiment are called peaks and range in length from several hundreds to thousands of bp (Johnson et al., 2007, Nakato, Shirahige, 2017; Lloyd, Bao, 2019). Each of the peaks does not necessarily contain the binding site of the target TF, direct binding can be performed by one of the possible partner TFs. Massive application of other *in vivo* experimental sequencing technologies besides ChIP-seq, e.g. CUT&RUN (Sken, Henikoff, 2017), as well as *in vitro* technologies (PBM, HT-SELEX) (Stormo, Zhao, 2010; Jolma et al., 2013; Franco-Zorrilla et al., 2014) allowed to accumulate data on the nucleotide specificity of binding sites of hundreds of TFs for the main model eukaryotic species. Several databases (DBs) performed uniform primary processing of massive genome-wide TFBS sequencing data, including ChIP-seq data (GTRD, Kolmykov et al., 2021; ReMap, Hammal et al., 2022; Cistrome DB, Taing et al., 2024).

Enrichment analysis of TFBS motifs, in particular the *de novo* motif search (Zambelli et al., 2013; Liu et al., 2018; Bailey, 2021), was initially used only to confirm the vali-

dity of the results of ChIP-seq experiments (sets of DNA sequences or peaks). Then, the *de novo* motif search became the standard approach for analysis of peak sets, allowing to determine enriched motifs, presumably corresponding to the motifs of the binding sites of the target TF and several partner TFs, cooperatively acting in the regulation of gene transcription (Spitz, Furlong, 2012; Slattery et al., 2014; Morgunova, Taipale, 2017).

To date, for several hundred TFs of the main eukaryotic taxa, such as mammals, insects and higher plants, TFBS motifs of the PWM model (nucleotide frequency matrices) are compiled in a number of DBs, JASPAR (Rauluseviciute et al., 2024), Hocomoco (Vorontsov et al., 2024) and Cis-BP (Weirauch et al., 2014). For example, the Hocomoco DB (version 12, Vorontsov et al., 2024) amounts to 1,443 binding site motifs for 949 human TFs. The analysis pipeline used by the Hocomoco DB for human and mouse TFBS motifs allowed identifying more than one structural type of motif for several hundred annotated TFs.

For a single TF, both the number of different binding site motifs and the structure and variability of each of the motifs are determined by the structure of the DNA binding domain (DBD) of this TF (Wingender, 1997, 2013). Based on the analysis of the similarity of the structure of DBDs of TFs and the alignment of the amino acid sequences of DBDs of TFs, a hierarchical classification of TFClass was developed, first for human TFs, and then for their orthologs in rodents and mammals (Wingender et al., 2013, 2015, 2018). This classification has six hierarchy levels. The upper levels of the hierarchy, superclass and class are defined according to the general topology and structural features of the DBDs of TFs. The next levels of the family and subfamily are deduced by the similarity of amino acid sequences of DBDs of TFs based on their alignments. The lower levels are the TF gene and the structural variant of its protein. In total, mammals have nine superclasses. Analysis of the structure of DBDs of TFs in plants did not reveal additional superclasses, however, about half of the TF classes turned out to be plant-specific (Plant-TFClass DB, Blanc-Mathieu et al., 2024).

The most important function of TFs *in vivo* is their ability to bind DNA specifically. However, the TFClass classification does not take into account the similarity of TFBS motifs at certain hierarchy levels, in specific classes, families, etc. The similarity of TFBS motifs can vary greatly in different classes of TFs. For example, the largest class of mammalian TFs, C2H2 zinc finger factors {2.3}, has the most noticeable variability in TFBS motifs (Najafabadi et al., 2015; Lambert et al., 2018). Hereinafter, numbers in curly brackets denote the TF classification nomenclature from the TFClass (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018). For example, TF JUN belongs to the superclass Basic domains {1}, the class Basic leucine zipper factors (bZIP) {1.1}, the Jun-related family {1.1.1}, and the Jun subfamily {1.1.1.1}. To determine a functioning TF by a given enriched motif of its binding sites as a result of a *de novo* motif search, we can apply not only the classification of TFs by the structure of their DBDs but also the classification of TFs by the similarity of TFBS motifs.

An important step in the analysis of the results of *de novo* enriched motif search applied for ChIP-seq data is the most precise determination of the motifs of binding sites of target and partner TFs based on the enriched motifs obtained. A common way to limit the list of putative TFs for each enriched motif is to assess the significance of its similarity to the TFBS motifs of known TFs from the DBs (Weirauch et al., 2014; Rauluseviciute et al., 2024; Vorontsov et al., 2024). Standard tools such as TomTom (Gupta et al., 2007) can be used to assess similarity in the pairs of motifs of the PWM model.

The estimate of the total number of human TFs is 1,659 (Shen et al., 2023); however, both the number of structurally different DBDs of TFs and the number of TFs with distinct binding site motifs are much smaller, since the TFs with similar DBDs usually have similar binding site motifs (Ambrosini et al., 2020). The most obvious exception to this general rule is the TF class C2H2 zinc finger {2.3} (Lambert et al., 2018).

The presence of two or more structurally distinct binding site motifs for a single TF is widespread across various TF classes (Vorontsov et al., 2024). This is explained by the ability of certain TFs to bind only as dimers of related TFs (for example, TF pairs from the classes Basic helix-loop-helix factors (bHLH) {1.2}, or Basic leucine zipper factors (bZIP) {1.1}), or as a dimer or monomer (for example, TFs from the class Nuclear receptors with C4 zinc fingers {2.1}) (Amoutzias et al., 2008). Commonly, TFBS motifs of related TFs from the same class or family exhibit a high to moderate degree of similarity depending on the position of the class, family, or subfamily in the TFClass/Plant-TFClass hierarchy. However, even among the TFBS motifs of the same TF, a certain variety of structural variants can be observed. For example, for TF CDX2 (Homeo domain factors {3.1} class) and THB (Nuclear receptors with C4 zinc fingers {2.1} class), there are two and four motifs in Hocomoco (version 12), respectively. The two TFBS motifs of CDX2 TF are not significantly similar (p-value > 0.001, Gupta et al., 2007) (Fig. 1*a*), significant similarity is also absent in three of the six possible pairs of the four THB binding site motifs (Fig. 1*b*, *c*). It can be assumed that more often families or subfamilies, rather than TF classes, represent significantly similar motifs (Nagy G., Nagy L., 2020; de Martin et al., 2021; Zenker et al., 2025). We study this issue in more detail in this work.

The most important step in the analysis of ChIP-seq data, *de novo* motif search, reveals a list of enriched motifs for ChIP-seq peaks. For the PWM motif model, each motif is a matrix of nucleotide frequencies, and it is necessary to determine a list of known TFs from DBs, such as Jaspar (Rauluseviciute et al., 2024), Hocomoco (Vorontsov et al., 2024) or Cis-BP (Weirauch et al., 2014), having significantly similar motifs of binding sites of known TFs. However, in addition to the dependence of the number of binding site motifs on the DBD structure of a TF, TFs are extremely unevenly distributed in superclasses, classes, and even families. In the most complete human/mouse DB of TFBS motifs (Hocomoco, version 12, Vorontsov et al., 2024), the five largest TF classes represent about 75 % of all motifs (1,082 of 1,443): C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1}, Basic

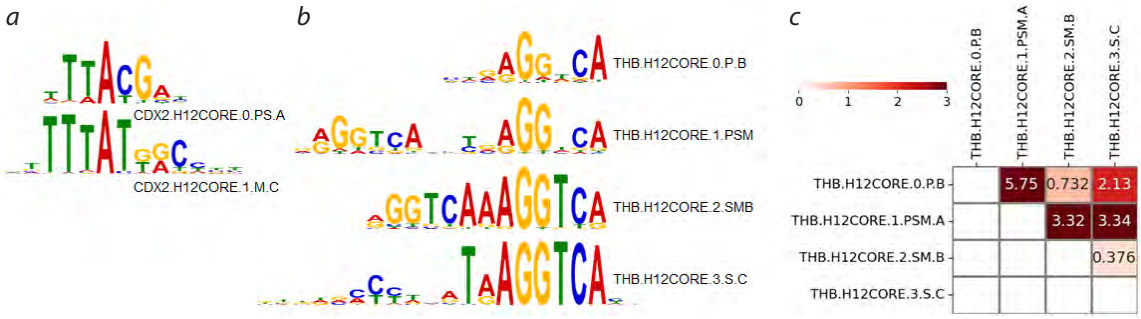


Fig. 1. Similarity of different binding site motifs representing individual TFs.

a, *b* – two/four binding site motifs of CDX2 / THB TFs from the Homeo domain factors {3.1} / Nuclear receptors with C4 zinc fingers {2.1} classes. For each motif, the Hocomoco DB identifier is indicated (Vorontsov et al., 2024). The PWM motif model logo represents nucleotide frequencies at positions as letter heights (Schneider, Stephens, 1990); *c* – motif similarity estimates calculated by the TomTom tool (Gupta et al., 2007) for four TFBS motifs of THB TF, the color reflects the significance of the similarity, $-\mathrm{Log}_{10}[p\text{-value}]$.

helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1}, and Basic leucine zipper factors (bZIP) {1.1}. The ten largest classes comprise about 90 % of all motifs (1,303 out of 1,443). The eight largest TF families from a total of four classes represent more than 51 % (742 out of 1,443) of all TFBS motifs: More than 3 adjacent zinc fingers {2.3.3}, HOX-related {3.1.1}, Multiple dispersed zinc fingers {2.3.4}, Paired-related HD {3.1.3}, NK-related {3.1.2}, Three-zinc finger Kruppel-related {2.3.1}, Tal-related {1.2.3}, and Ets-related {3.5.2}. A recent analysis of 1,725 TFs of the model plant *Arabidopsis thaliana* revealed about 40 % of them (686) with available TFBS motifs; the inclusion of TFBS motifs for 92 TFs from other plants showed an extremely limited vocabulary of only 74 distinct plant TFBS motifs (Zenker et al., 2025).

Very often, an enriched motif from the results of a *de novo* motif search has a high similarity to the TFBS motifs of known TFs from one or more families of the same class, or even an entire class falls into the list of TF candidates. The result is a list of several dozen TFs, and choosing a specific TF among them is not an easy task. Such long lists of TF candidates may complicate the identification of TFs most likely associated with enriched motifs. However, this complexity can be reduced by the systematic analysis of the similarity of the binding site motifs of TFs classified by the hierarchy levels from the TFClass DB. To date, for cognate TFs of a given structure of a DBD (class, family and subfamily), it has not been determined which of these levels is sufficient to identify a set of TFs with significantly similar binding site motifs. To solve this issue, one needs to find a set of certain arrays (or branches) of several consecutive levels of the TFClass hierarchical classification, for which the TFBS motifs are significantly similar. This approach is able to further systematize the hierarchical classification of TFs, adapt it to apply to the results of a *de novo* motif search. The resulting refined TF hierarchy will reflect the similarity of DBDs of TFs and the similarity of TFBS motifs.

We propose to include the annotation of the branches of similar binding site motifs of known TFs in a standard protocol of *de novo* motif search applied to the results of genome-wide mapping of TFBS *in vivo*, for example, using ChIP-seq technology. The application of branches can notably simplify the analysis of enriched TFBS motifs. The TF branches connect the generally accepted units of the hierarchical classification of TFs by DBDs, namely superclasses, classes, families, subfamilies (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018) to the similarity of TFBS motifs (Gupta et al., 2007).

Materials and methods

Input data and parameters. The input data are sets of TFBS motifs; each motif is represented by a nucleotide frequency matrix, an identifier and a TF name; for each TF, its superclass, class, family and subfamily (if any) are indicated, according to the TFClass DB (Wingender et al., 2013, 2015, 2018). TFBS motifs for human *Homo sapiens* and fruit fly *Drosophila melanogaster* were extracted from Hocomoco (version 12, https://hocomoco.autosome.org/) (Vorontsov et al., 2024) and Jaspar https://jaspar.elixir.no/ (Rauluseviciute et al., 2024). Both DBs construct TFBS motifs based on *in vivo* massive sequencing data (e. g. ChIP-seq), and *in vitro* ones (e. g. HT-SELEX). TFBS motifs are nucleotide frequency matrices consistent with the traditional PWM model. In both DBs, TF classification is applied according the DBD structure by hierarchy levels of superclass, class, family, subfamily and TF (TFClass DB, Wingender, 2013; Wingender et al., 2013, 2015, 2018). We selected for analysis the classes amounting to at least 50 TFBS motifs: seven / two classes for human / *Drosophila* TFs, see theTable.

Similarity metric of two TFs. We applied the TomTom tool (Gupta et al., 2007) to assess the significance of similarity (*p*-value) in pairs of TFBS motifs, the parameter of the motif comparison function was the Pearson correlation coefficient. Two TFBS motifs were considered similar if the significance level reached the threshold, $-\mathrm{Log}_{10}[p\text{-value}] > \mathrm{Thr} = 3$.

We define the similarity metric for a pair of TFs based on their binding site motifs according to the distribution of similarity in all possible pairs of binding site motifs of one and another TF, since TFs can have one or more binding site motifs. Let two TFs X/Y have N_X/N_Y motifs, $\{M_i\}$, $1 \le i \le N_X$ and $\{M_j\}$, $1 \le j \le N_Y$, correspondingly. The distribution of similarity estimates in a pair of these TFs based on their binding site motifs includes $N_X \times N_Y$ pairs of motifs. Let the similarity Score(M_i, M_j) of motifs M_i and M_j be given

TFBS motif sets from the Hocomoco and Jaspar DBs used in analysis

Taxon: species	TF class	Number of motifs	Number of TFs
Mammals: *H. sapiens*	Basic leucine zipper factors (bZIP) {1.1}	86	47
	Basic helix-loop-helix factors (bHLH) {1.2}	115	76
	Nuclear receptors with C4 zinc fingers {2.1}	93	44
	C2H2 zinc finger factors {2.3}	479	373
	Homeo domain factors {3.1}	309	184
	Fork head/winged helix factors {3.3}	65	43
	Tryptophan cluster factors {3.5}	67	38
	Total	1,214	805
Insects: *D. melanogaster*	C2H2 zinc finger factors {2.3}	79	57
	Homeo domain factors {3.1}	106	90
	Total	185	147

by TomTom (Gupta et al., 2007) as the logarithm of the significance *p*-value:

$$\mathrm{Score}(M_i, M_j) = -\mathrm{Log}_{10}[p\text{-value}(M_i, M_j)]. \quad (1)$$

Then for two TFs X and Y, the similarity metrics $\mathrm{Score}_{X,Y}$ will be defined as follows:

$$\mathrm{Score}_{X,Y} = \max_{1\le i\le N_X,\ 1\le j\le N_Y} \{\mathrm{Score}(M_i, M_j)\}. \quad (2)$$

If this metrics $\mathrm{Score}_{X,Y}$ (2) exceeds the pre-defined threshold Thr, then TFs X and Y can be considered significantly similar in their binding site motifs. For one TF, the heterogeneity of binding site motifs is estimated as the median (the second quartile, Q2) of the distribution over all possible pairs of binding site motifs of that TF:

$$\mathrm{Score}_X = \mathrm{Median}_{1\le i<N_X,\ i<j\le N_X} \{\mathrm{Score}(M_i, M_j)\}. \quad (3)$$

Similarity metric of two sets of TFs. Let a class have a family A with N_A TFs. The distribution of all possible TF pairs in this family includes $N_A \times (N_A - 1)/2$ variants. Let a family B from the same class have N_B TFs. The distribution of all possible TF pairs of families A and B includes $N_A \times N_B$ variants. For both the intra-family and inter-family cases, for all TF pairs, the similarity estimates are calculated by the formula (2). Likewise, pairs of subfamilies in the same family and pairs of classes in the same superclass are considered.

For the obtained distribution of similarity estimates, it is possible to calculate five similarity metrics for two sets of TFs: minimum (Min), quartiles Q1, Q2 (median) and Q3, and maximum (Max). Min/Max metrics indicate the choice of the minimum/maximum values, and quartile metrics indicate the value of the corresponding fraction of the entire distribution. For example, the Q2 (median) metric for two sets of TFs reflects a level of similarity of 50 % of all possible TF pairs from these sets. Let the first {X} and second {Y} sets have K and T TFs, $1\le k\le K$, $1\le t\le T$, then based on the distributions of the similarity values in TF pairs calculated by the formula (2) $\{\mathrm{Score}_{X(k),Y(t)}\}$, the similarity metric $\mathrm{Score}_{\{X\},\{Y\}}$ of the two TF sets is calculated as follows:

$$\mathrm{Score}_{\{X\},\{Y\}} = \mathrm{Median}_{1\le k\le K,\ 1\le t\le T} \{\mathrm{Score}_{X(k),Y(t)}\}. \quad (4)$$

Definition of the branch in the TF hierarchical classification. If the similarity score of two sets of TFs based on their binding site motifs exceeds the predetermined threshold Thr, then these TFs can be referred to the same branch. Next, consider the median metric (4). For example, an entire class can belong to the same branch if more than half of all its possible TF pairs are similar in terms of binding site motifs. Although it is possible that certain families of a class do not show significant similarity, with a probability of more than 50 %, an arbitrary pair of TFs from this class shows the significant similarity of binding site motifs.

To perform cluster analysis and construct trees reflecting the similarity of TFs based on the TFBS of the sister classes of the same superclass, the sister families of the same class, etc., we used the UPGMA algorithm scheme (unweighted pair group method with arithmetic mean) (Sokal, Michener, 1958). During the classification, we applied the median metric (Q2, formula (4)) described above to evaluate any pair of objects.

To search for branches, the analysis starts at the superclass level, and continues at lower levels of the hierarchy: the class, family, subfamily, or TF. First, the TF similarity metric is calculated within a given hierarchy level, for example, a class, as well as for all families of this class. This gives a list of families with similarities exceeding the threshold Thr. All such families initially refer to different branches; to analyze the remaining families, we need to go to a lower level. Then the TF similarity metrics are calculated for all possible pairs of the sister families of this class. This gives the similarity matrix for families of the class. The diagonal values of the matrix show the similarities within each family and those above the diagonal provide the similarities for all pairs of different families. Next, we select a pair of families with the highest similarity. If this similarity exceeds the threshold, then a pair of such families (branches) are joined into one branch. After that, the similarities in all pairs of updated branches are recalculated. Calculations continue as long as there are pairs of branches that allow joining based on their similarity. In such a way we can gradually descend to the lower levels and reach the level of TF.

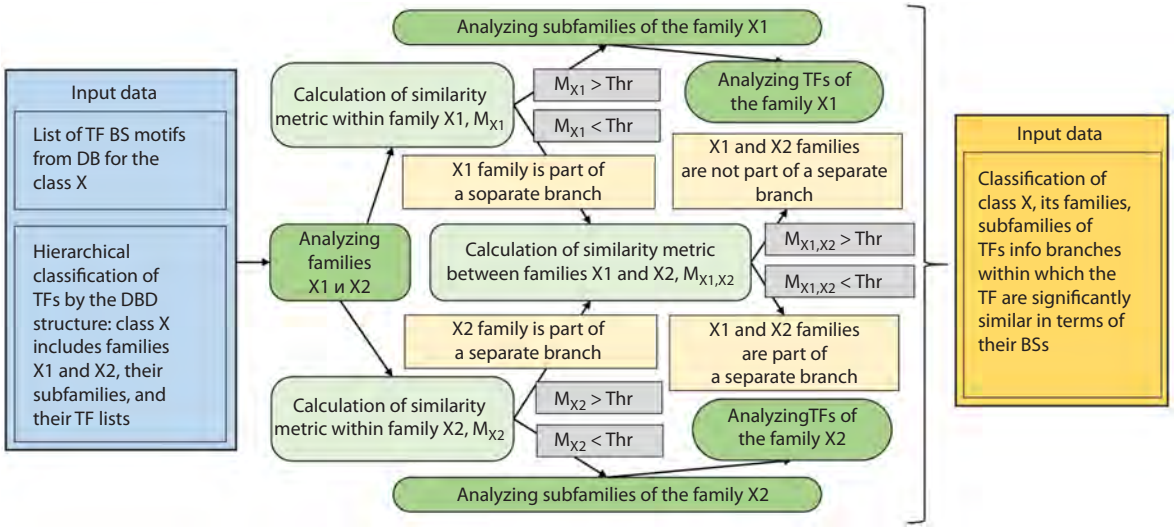


Fig. 2. Scheme of analysis to determine branches of similar motifs of TFBS. The scheme shows in detail the stage of analysis of one class X consisting of two families X1 and X2. The blue color shows the input data, dark green – analysis stages, light green – similarity metric calculations, gray – verification of similarity conditions for motifs, light yellow – intermediate results, dark yellow – final results. The scheme discloses the analysis of two families X1 and X2 of class X. The analysis of subfamilies of these families and the analysis of TFs in each of the subfamilies are performed similarly to the analysis of families X1 and X2, as described in the text.

The similarity of the binding site motifs of single TFs is analyzed separately (see formula (3)), although, obviously, this analysis takes place inside one branch, since according to formulas (2) and (4), each branch for any TF contains all its binding site motifs, and we can only note TFs (Fig. 1) having significantly different binding site motifs.

The purpose of the whole analysis is to sequentially find such sets of TFs (for example, for a class, this is a list of family clusters), for which the metric (4) exceeds the given threshold Thr, and the list for each of the branches includes as many elementary classification units as possible.

TF superclasses are heterogeneous enough in the similarity of binding site motifs since each superclass splits into multiple branches. A branch in the TFClass hierarchy is defined as the maximum possible set of TFs from the highest class level to the lowest level (in practice, this is a class, family, subfamily, TF), such that in this set for the majority of TF pairs there is a significant similarity of TFs based on their binding site motifs, according to the similarity metric (4).

A branch may include one or more sister classification units:
- a whole class,
- one or more families of the same class,
- one or more subfamilies of the same family,
- one or more TFs of the same subfamily.

The final result of the analysis is the determination of the set of all branches, within each of the branches, the metric (4) indicates significant similarity of TFs based on TFBS motifs. Figure 2 is a scheme of the analysis used in the work.

Results

Similarity of TFs in sister subfamilies of the same families

In order to start a massive analysis of different degrees of similarity of binding site motifs to cognate TFs according to the TFClass hierarchical classification, we test the TFBS motif similarity for subfamilies of individual families belonging to various TF classes. Figure 3 shows the fraction of similar TFs based on the binding site motifs within subfamilies of different families, using the five metrics Min, Q1, Q2, Q3, and Max. The Q2 metric (median) is calculated according to the formula (4), others metrics are computed likewise. By construction, among these metrics from Min to Max, the fraction of the similar TFBS motifs is growing. However, regardless of the metric choice, some subfamilies show a lower similarity or even a complete lack of similar TFBS motifs, compared to other subfamilies. For example, for the three subfamilies of the Fox {3.3.1} family, the values of the Q2 metric are close to 100 % (Fig. 3*f*), and for the subfamilies TWIST {1.2.3.2}/MEIS {3.1.4.2} of the families Tal-related {1.2.3}/TALE-type HD {3.1.4}, respectively, these values are less than 50 % (Fig. 3*b*, *d*).

Thus, the similarity of TFs based on binding site motifs can vary significantly across the subfamilies of the same families. Obviously, the same conclusion can be drawn for the families of the same classes. Further, in the analysis, the median metric (Q2) (4) was used to assess the similarity of the two sets of TFs, since the meaning of its application is the most transparent compared to the Min, Q1, Q3, and Max metrics. Hereinafter, the value of the Q2 metric is called "similarity".

Similarity analysis of human TFs

Figure 4 shows the human TF similarity trees based on binding site motifs for the main classes of the three largest superclasses: Basic domain {1}, Zinc-coordinating DNA-binding domains {2} and Helix-turn-helix domains {3}. Of all the classes, only one class Tryptophan cluster factors {3.5} shows the significant similarity of TFs based on their binding sites motifs (similarity 3.68). The classes Basic leucine zipper factors (bZIP) {1.1} and Nuclear receptors with C4 zinc fingers {2.1} reach the similarity values of 2.51 and 2.68,

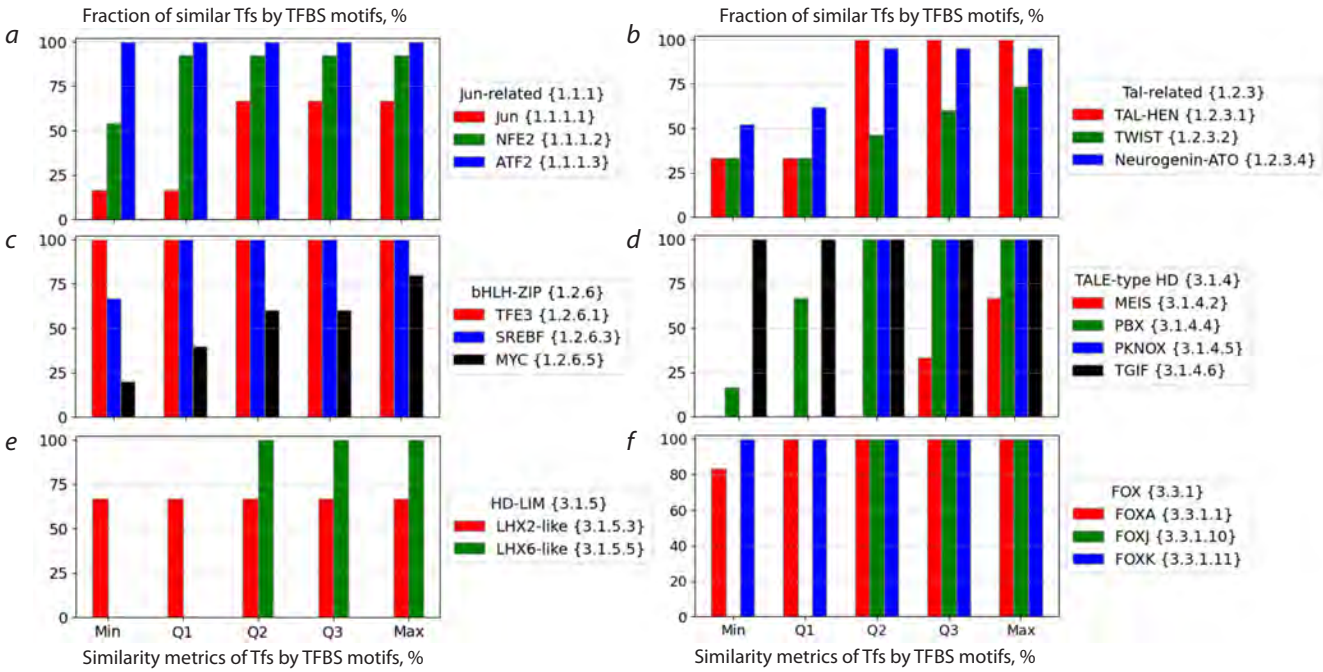


Fig. 3. Fraction of significantly similar TFs based on the binding site motifs for subfamilies of different families using the five similarity metrics: Min, Q1, Q2, Q3, and Max.

a–*e*, and *f* – Jun-related {1.1.1}, Tal-related {1.2.3}, bHLH-ZIP {1.2.6}, TALE-type HD {3.1.4}, HD-LIM {3.1.5}, and FOX {3.3.1} families, respectively. Color marks subfamilies. The X axis lists TF similarity metrics; the Y axis shows the fraction of significantly similar TFs based on the binding site motifs in the subfamily. Significant similarity requires the criterion $-\mathrm{Log}_{10}[p\text{-value}] > 3$ (Tomtom tool, Gupta et al., 2007).

respectively, indicating a trend towards significant similarity. The classes Fork head/winged helix factors {3.3}, Homeo domain factors {3.1} and Basic helix-loop-helix factors (bHLH) {1.2} show lower similarity values of 1.14, 1.42 and 1.47. The lowest similarity of TFs based on the binding site motifs is found for the class C2H2 class zinc finger factors {2.3} (0.44); this class is the largest in human, allowing the greatest variability in the structure of TFs (Najafabadi et al., 2015; Lambert et al., 2018, 2019).

Therefore, to identify branches within all classes except the class Tryptophan cluster factors {3.5}, it is necessary to proceed to the analysis of their families. Next, we will separately consider each of the three superclasses in more detail.

The first superclass has two large classes, Basic leucine zipper factors (bZIP) {1.1} and Basic helix-loop-helix factors (bHLH) {1.2}; the similarity of TFs based on binding site motifs between these classes is very low (0.523, Fig. 5*a*). The similarity of TFs within each class is noticeably higher, but the Basic leucine zipper factors (bZIP) {1.1} class has distinctly more similar TFs (2.51) than the Basic helix-loop-helix factors (bHLH) {1.2} class (1.47).

There are eight families in the Basic leucine zipper factors (bZIP) {1.1} class (Fig. 5*b*, *e*): from Jun-related {1.1.1} to C/EBP-related {1.1.8}. Each family of the class has one or more other families with significantly similar TFs based on binding site motifs. As a result, all families fall into four branches (Fig. 5*e*); there are two branches of two families (XBP1-related {1.1.5} and CREB-related {1.1.7}, ATF4-related {1.1.6} and C/EBP-related {1.1.8}), and the branches

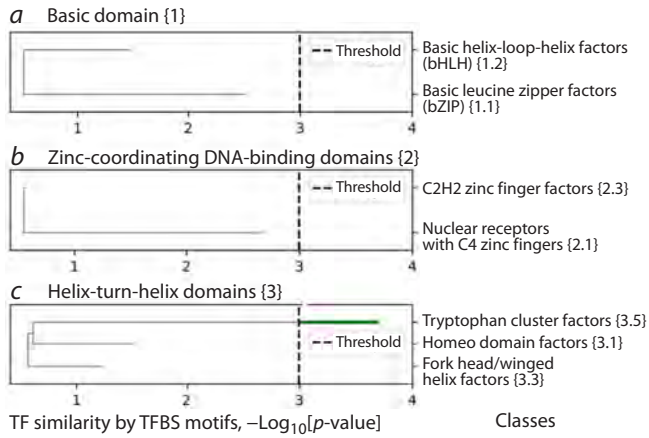


Fig. 4. Similarity of TFs based on binding site motifs in the largest classes of the three largest human superclasses.

a, *b*, and *c* – class TF trees for the superclasses Basic domain {1}, Zinc-coordinating DNA-binding domains {2}, and Helix-turn-helix domains {3}. The X axis reflects the value of the Q2 metric, the dash line shows its threshold value 3. The green color shows the class Tryptophan cluster factors {3.5}, which forms a separate branch, and the gray color indicates paths, the Q2 metric values of which are less than the threshold. Horizontal line break marks the value of the Q2 metric.

of one (Maf-related {1.1.3}) and three families (Jun-related {1.1.1}, Fos-related {1.1.2}, B-ATF-related {1.1.4}).

In the Basic helix-loop-helix factors (bHLH) {1.2} class, within each of the families, with the exception of one (PAS {1.2.5}), TFs have significant similarities based on the binding

site motifs (Fig. 5*b*, values on the diagonal), but there are no significant similarities between TF families based on the binding site motifs. Therefore, each of the families, with the exception of the PAS {1.2.5} family, forms a separate branch (Fig. 5*f*). The PAS family {1.2.5} is divided into four branches {1.2.5.1}, {1.2.5.2}, {1.2.5.3} and {1.2.5.4} by four subfamilies (Fig. 5*d*).

The second superclass has two large classes Nuclear receptors with C4 zinc fingers {2.1} and C2H2 zinc finger factors {2.3}, the similarity of TFs based on binding site motifs between these classes is very low (0.554, Fig. 6*a*). In the Nuclear receptors with C4 zinc fingers {2.1} class, TFs have the similarity only slightly below the threshold (2.68), and the TF similarity in the class C2H2 zinc finger factors {2.3} is very low (0.443).

In the class Nuclear receptors with C4 zinc fingers {2.1} (Fig. 6*b*), only one family, Steroid hormone receptors {2.1.1}, has a similarity of TFs 2.39 below the threshold. This family is divided into two branches according to the two subfamilies: GR-like (NR3C) {2.1.1.1} and ER-like (NR3A) {2.1.1.2} (Fig. 6*c*). The similarity of TFs between these subfamilies is low (0.822), and within each subfamily, it is high (6.41 and 3.59). TFBS motifs from these related subfamilies have a similar structure: TFs of both subfamilies can bind DNA as

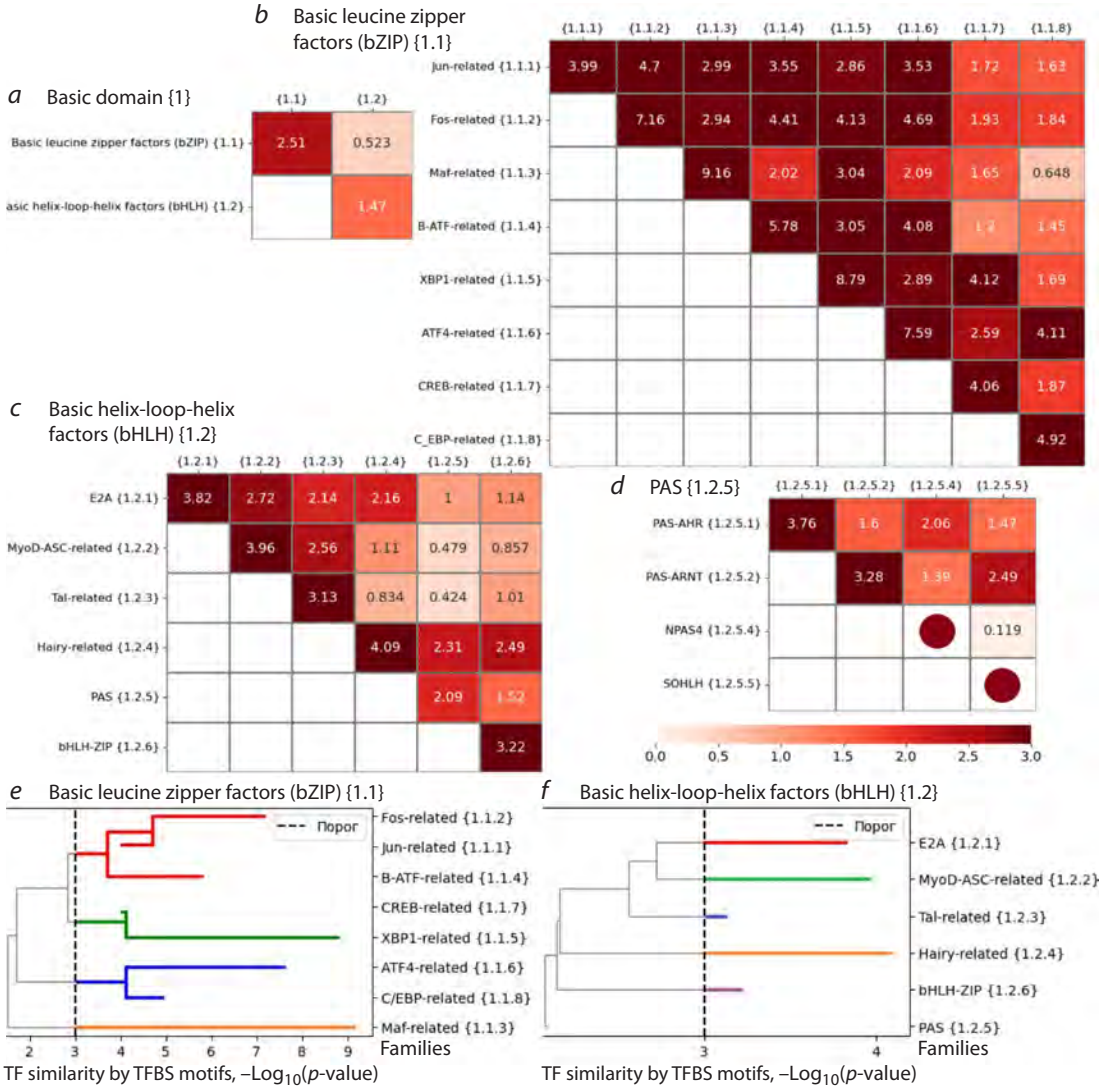


Fig. 5. TF similarity based on binding site motifs for the Basic domain {1} superclass.

a–d – heatmaps for classes of the superclass, for families of the Basic leucine zipper factors (bZIP) {1.1}/Basic helix-loop-helix factors (bHLH) {1.2} classes and for subfamilies of the PAS {1.2.5} family of the Basic helix-loop-helix factors (bHLH) {1.2} class. A brown circle on the heatmap diagonal means that the subfamily has only one TF with one TFBS motif. The color reflects the value of the Q2 similarity metric. Here and further to the right of each heatmap are the names of classes/families/subfamilies along with their numerals, and above are only numerals; *e* and *f* – family trees for the classes Basic leucine zipper factors (bZIP) {1.1} and Basic helix-loop-helix factors (bHLH) {1.2}. The Y axis reflects the value of the Q2 metric, the dash line shows its threshold value 3. All colors except gray reflect individual branches, and gray highlights paths, the Q2 metric value of which is less than the threshold. A horizontal line break marks the value of the Q2 metric for the family. The Jun-related {1.1.1} family (*e*) has a lower similarity of 3.99 (*b*) than the similarity of the union of Jun-related {1.1.1} and Fos-related {1.1.2} families, so the direction of the path of the Jun-related {1.1.1} family from the junction point of these two families changes to the opposite.

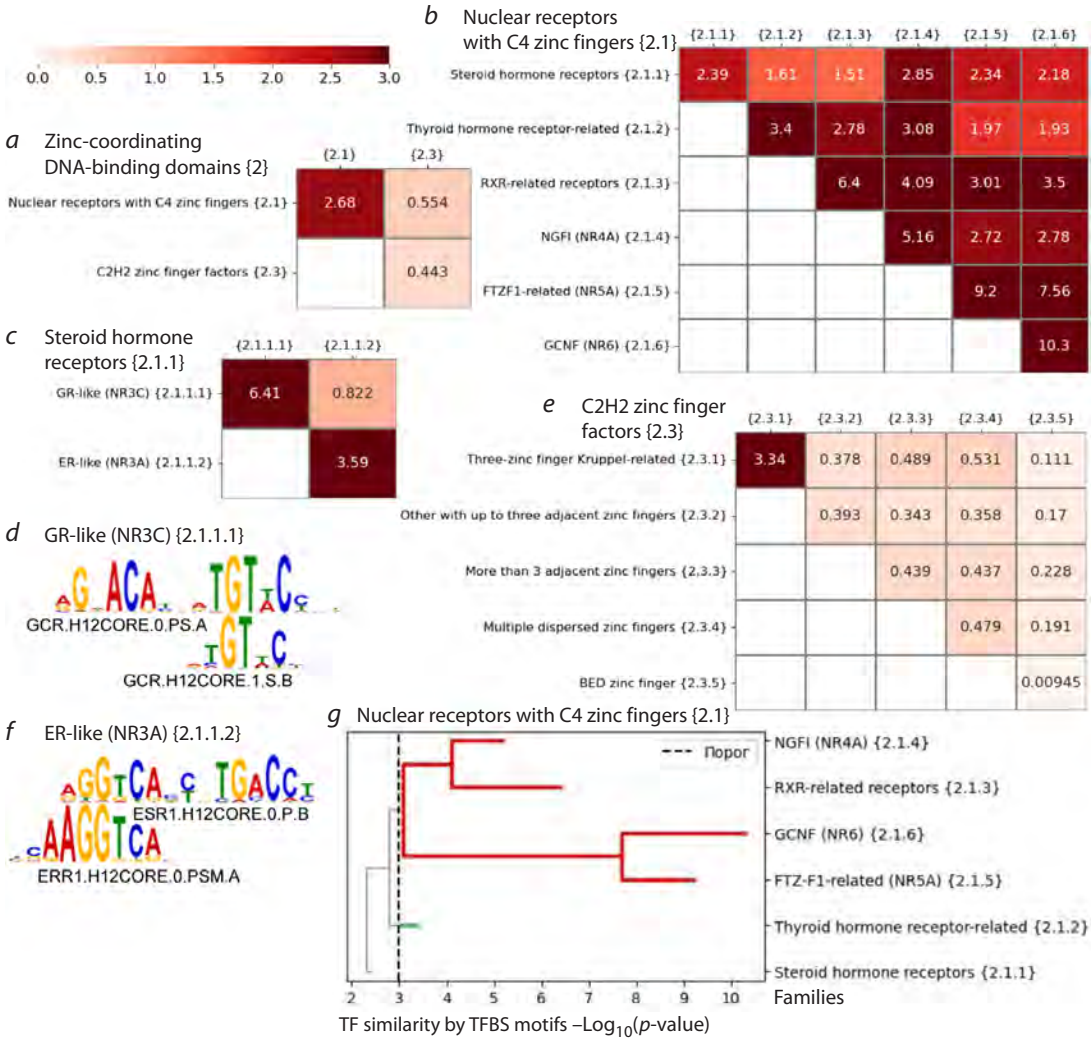


Fig. 6. Similarity of TFs based on binding site motifs for the superclass Zinc-coordinating DNA-binding domains {2}.

a, *b*, *c* and *f* – heatmaps for classes of the superclass, for families of the class Nuclear receptors with C4 zinc fingers {2.1}, for subfamilies of the family Steroid hormone receptors {2.1.1} of the class Nuclear receptors with C4 zinc fingers {2.1} and for families of the class C2H2 zinc finger factors {2.3}; *d*, *e* – examples of TF binding site motifs from the GR-like (NR3C) {2.1.1.1}/ER-like (NR3A) {2.1.1.2} subfamilies of the family Steroid hormone receptors {2.1.1}; *g* – family tree for the Nuclear receptors with C4 zinc fingers {2.1} class. The Y axis implies the value of the Q2 metric, the dash line means the threshold value 3. Red and green colors reflect separate branches, and paths are highlighted in gray, if the respective value of the Q2 metric is less than the branch threshold. Horizontal line break marks the value of the Q2 metric.

monomers or as dimers formed by an inverted repeat (Nagy G., Nagy L., 2020), but regardless of this, the monomeric subunits in TFBS motifs of the GR-like (NR3C) {2.1.1.1} (Fig. 6*d*) and ER-like (NR3A) {2.1.1.2} subfamilies (Fig. 6*e*) are clearly distinct. The Thyroid hormone receptor-related {2.1.2} family forms a separate branch, since the similarity of its TFs with the TFs of four of the five other families is below the threshold 3 (Fig. 6*b*, *g*). Four families from the RXR-related receptors {2.1.3} to GCNF (NR6) {2.1.6} form one branch: Figure 6*f* shows the tree dividing the Nuclear receptors with the C4 zinc fingers {2.1} class into branches by families.

In the C2H2 zinc finger factors {2.3} class (Fig. 6*f*), only one family, Three-zinc finger Kruppel-related {2.3.1}, forms a separate branch. To determine the branches of the other four families of the class, we need to go down to the levels of subfamilies or TFs, see the list of all branches of the C2H2 zinc finger factors {2.3} class in Table S1[1].

The third superclass includes three large classes Homeo domain factors {3.1}, Fork head/winged helix factors {3.3}, and Tryptophan cluster factors {3.5}. The similarity between TFs of different classes based on binding site motifs is very low in all three possible pairs of classes (Fig. 7*a*, cells above the diagonal). Similarity of TFs within each of the classes Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} is medium, 1.42 and 1.12. The class Tryptophan cluster factors {3.5} forms one branch (Fig. 4).

In the class Fork head/winged helix factors {3.3}, two families E2F {3.3.2} and RFX {3.3.3} represent two separate

[1] Supplementary Table S1 and Figures S1 and S2 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Levitsky_Engl_29_7.pdf

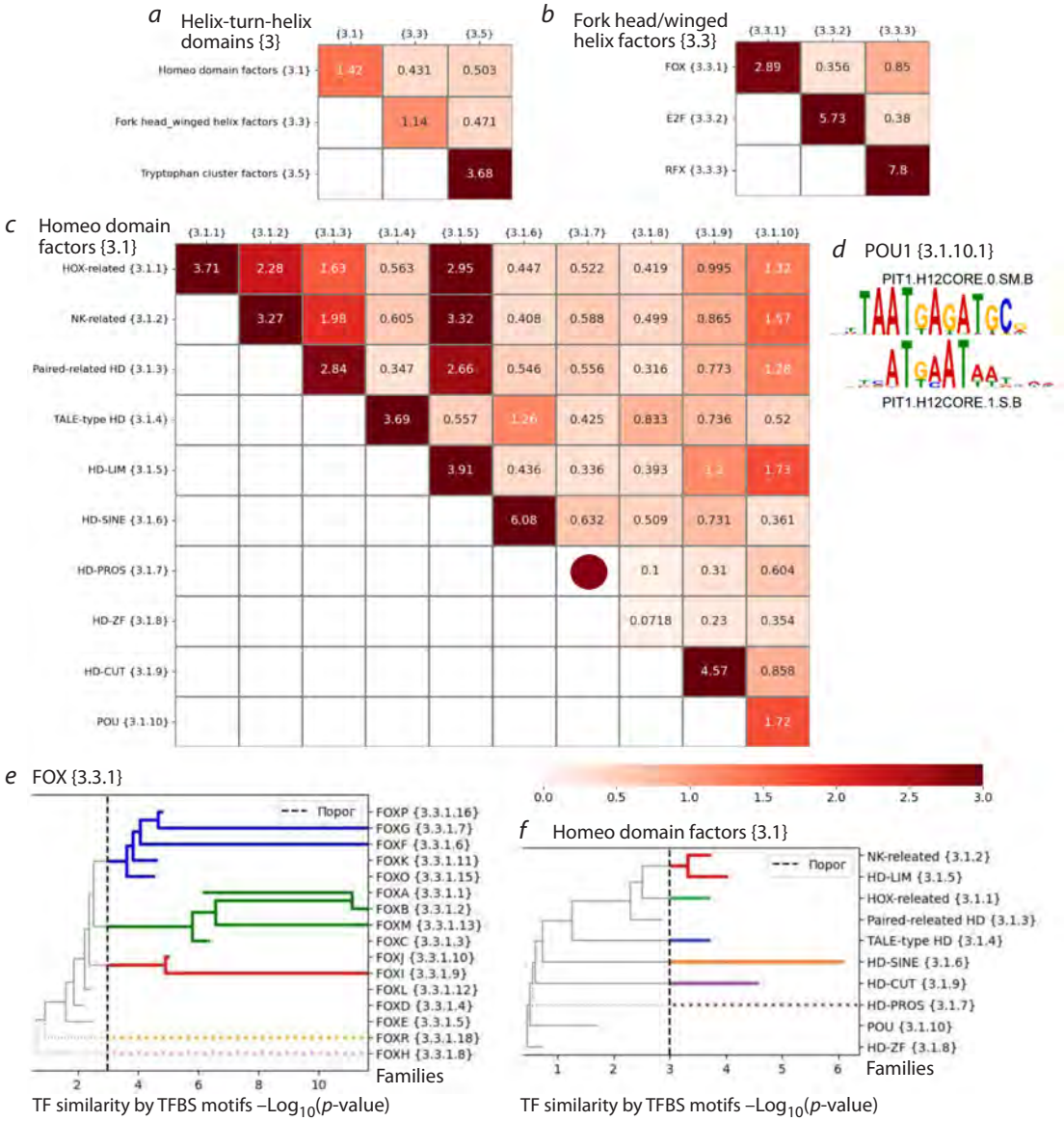


Fig. 7. Similarity of TFs based on binding site motifs for the superclass Helix-turn-helix domains {3}.

a–c – heatmaps for classes of the superclass, for families of the classes Fork head/winged helix factors {3.3} and Homeo domain factors {3.1}. The brown circle on the heatmap diagonal means that the family has only one TF with one binding site motif. The color reflects the value of the Q2 similarity metric; *d* – logo of two binding site motifs of TF PIT1 from the subfamily POU1 {3.1.10.1}; *e* and *f* – trees for subfamilies of the FOX {3.3.1} family and for families of the Homeo domain factors {3.1} class. The Y axis reflects the value of the Q2 metric, the dash line shows its threshold value 3. Dotted lines mean a single TF with one binding site motif in the current family or subfamily. All colors except gray reflect individual branches, and gray indicates paths, the Q2 metric value of which is less than the branch threshold. Horizontal line break marks the value of the Q2 metric. The subfamily FOXA {3.3.1.1} (*e*) has a lower similarity of 6.22 (Fig. S1) than the similarity of the union of the subfamilies FOXA {3.3.1.1} and FOXB {3.3.1.2}, so the direction of the path of the subfamily FOXA {3.3.1.1} from the junction point of these two subfamilies changes to the opposite.

branches, and the similarity of TFs of the FOX family {3.3.1} almost reaches the threshold (similarity value 2.89, Fig. 7*b*). A vivid illustration of the correctness of the division of the Fork head/winged helix factors {3.3} class into three families (Fig. 7*b*) is a noticeable excess of the similarity of TFs within families (three values on the diagonal) in relation to the similarity of TFs between families (three values above the diagonal).

Among the 16 subfamilies of the FOX family {3.3.1} (Fig. 7*e*), only three subfamilies FOXD {3.3.1.4}, FOXH {3.3.1.5} and FOXL {3.3.1.12} achieved TF similarity below the threshold 3: 2.19, 2.48 and 2.17, respectively. Four, five and two subfamilies form separate branches (Fig. 7*e*). There are two subfamilies, FOXH {3.3.1.8} and FOXR {3.3.1.18}, with low similarity of TFs based on binding site motifs with other subfamilies and between themselves (Fig. S1).

Two families (NK-related {3.1.2} and HD-LIM {3.1.5}) of the Homeo domain factors {3.1} class merge into one branch; each of five HOX-related {3.1.1}, TALE-type HD {3.1.4}, HD-SINE {3.1.6}, HD-PROS {3.1.7} and HD-CUT {3.1.9}

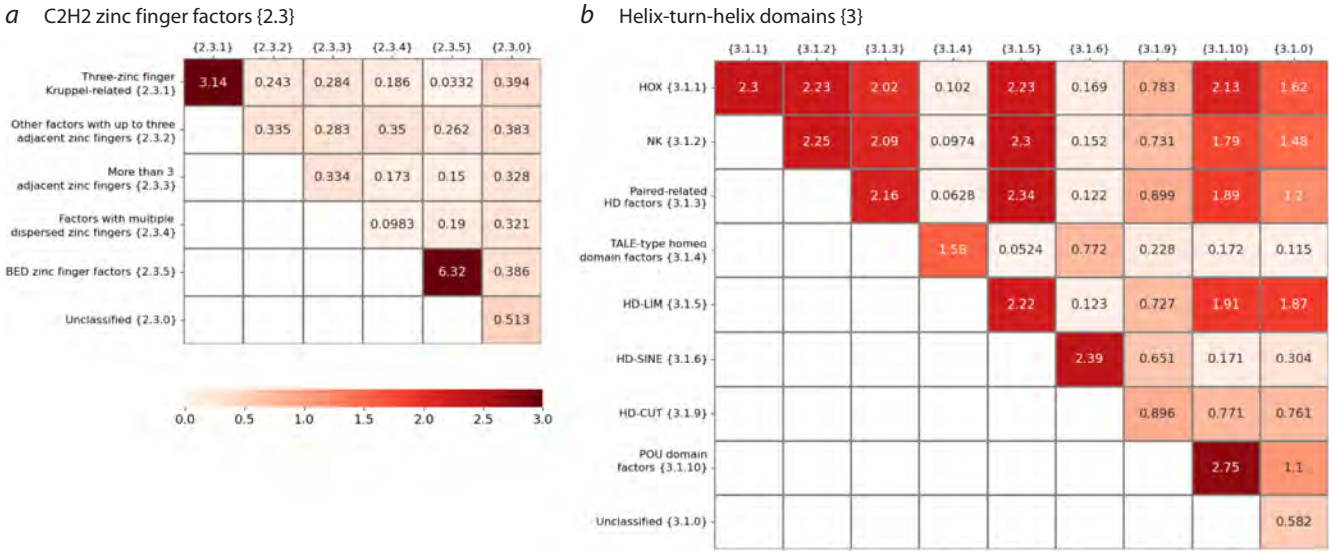


Fig. 8. Similarity of *Drosophila* TFs from the two large classes based on binding site motifs.
a and *b* – heatmaps for families of the classes C2H2 zinc finger factors {2.3} and Homeo domain factors {3.1}. The color reflects the value of the Q2 similarity metric.

families represents a separate branch (Fig. 7*c*, *f*). To find branches for the remaining families Paired-related HD {3.1.3}, HD-ZF {3.1.8} and POU {3.1.10}, it is necessary to proceed to the subfamily level (Fig. S2, Table S1). The Paired-related HD {3.1.3} family is divided into two separate branches, combining 12 and 6 subfamilies (Fig. S2*a*, Table S1). The HD-ZF {3.1.8} family is divided into two branches according to two subfamilies, ZEB {3.1.8.3} and ZHX {3.1.8.5} (Fig. S2*b*). Three subfamilies POU2 {3.1.10.2}, POU3 {3.1.10.3} and POU5 {3.1.10.5} merge into one branch. The subfamily POU1 {3.1.10.1} is represented by one TF PIT1 with two significantly dissimilar TFBS motifs PIT1.H12CORE.0.SM.B and PIT1.H12CORE.1.S.B (Fig. 7*d*). The remaining three subfamilies POU4 {3.1.10.4}, POU6 {3.1.10.6} and HNF1-like {3.1.10.7} of the family POU {3.1.10} form separate branches (Fig. S2*c*).

The full list of branches for the seven largest TF classes Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1}, Homeo domain factors {3.1}, Fork head / winged helix factors {3.3} and Tryptophan cluster factors {3.5} is given in Table S1.

In general, based on the results presented in Figures 5–7 and in Figures S1, S2 and Table S1, we can conclude that often TFs of the same family already have dissimilar binding site motifs. However, this general trend is broken for some classes and families. It is most clearly violated for the largest class of human TFs C2H2 zinc finger factors {2.3} (Fig. 6*f*), for which it is necessary to descend to the level of subfamilies or even to the level of TFs to determine branches.

Similarity analysis of Drosophila TFs

To determine how the discovered patterns of similarity in different classes of TFs depend on the choice of taxon, we conducted an analysis analogous to that carried out above for the insect taxon sufficiently distant from the mammalian taxon. According to the Jaspar DB, there are only two classes of insect TFs with more than 50 binding site motifs (see the Table). All these TFs belong to the species *D. melanogaster*. The results obtained for insect TFs from these two classes, C2H2 zinc finger factors {2.3} and Homeo domain factors {3.1}, are in good agreement with the results obtained above for human TFs from seven classes (Fig. 4–7).

In the *Drosophila* C2H2 zinc finger factors {2.3} class (Fig. 8*a*), as well as in the same class in human (Fig. 6*f*), only one family, Three-zinc finger Kruppel-related {2.3.1}, has significantly similar TFs based on binding site motifs. Only TFs of one other family, BED zinc finger {2.3.5}, have very different similarity of binding site motifs (human 0.001, *Drosophila* 6.32). However, this family is very small: in *Drosophila*, it contains two almost indistinguishable binding site motifs of one TF Dref; and in human, two TFs ZBED1 and ZBED5 have clearly dissimilar to each other motifs of binding sites. The other three common families in both taxa, Other factors with up to three adjacent zinc fingers {2.3.2}, More than 3 adjacent zinc fingers {2.3.3}, Multiple dispersed zinc fingers {2.3.4}, as well as all remaining *Drosophila* TFs with unspecified families, assigned to the family Unclassified {2.3.0}, show extremely low similarity of TFs based on binding site motifs. In general, for both human and *Drosophila* TFs, the class C2H2 zinc finger factors {2.3} has TFs with very low similarity of binding site motifs (Fig. 6*f*, 8*a*).

Drosophila TFs from the Homeo domain factors {3.1} class (Fig. 8*b*) show slightly less similarity in terms of binding site motifs than TFs from the same human class (Fig. 7*c*). However, in each of these two taxa, among the eight common families, families with greater and lesser similarity of TFs based on binding site motifs are distinguished. Namely, in both taxa, TFs from four families – HOX-related {3.1.1},

NK-related {3.1.2}, Paired-related HD {3.1.3} and HD-LIM {3.1.5} – have the greatest similarity, both within and between families (Fig. 7*c*, *f*); however, the similarity itself exceeds the value 2 for *Drosophila* TFs, but does not reach the threshold 3 (Fig. 8*b*). The remaining families have TFs that are not similar both to each other and to TFs of the above families of the class. In general, much smaller similarity in the binding site motifs of *Drosophila* TFs of the Homeo domain factors class {3.1} (Fig. 8*b*) compared with the human TFs of the same class (Fig. 8*c*) can be explained by the noticeably smaller number of available massive sequencing data for *Drosophila* TFBSs (see the Table). Another explanation is the difference in the methods for obtaining TFBS motifs in the Hocomoco and Jaspar DBs.

Discussion

We propose a new systematic approach to refine the hierarchical classification of TFs according to the structure of DBDs by a set of branches combining TFs with similar motifs of binding sites. The similarity of the binding site motifs of known TFs can now be evaluated with various experimental massive sequencing technologies, including *in vitro* HT-SELEX and *in vivo* ChIP-seq data, for example, experimental results for different tissue conditions and developmental stages.

Estimates of the total numbers of human/*Drosophila* TFs are 1,659/651 (AnimalTFDB, Shen et al., 2023). The Hocomoco DB (version 12) for human and the Jaspar DB for *Drosophila* annotated 1,443 TFBS motifs for 949 TFs and 334 TFBS motifs for 273 TFs. Hence, although the ratios of the number of TFs with known binding site motifs to the estimates of the total numbers of TFs for human and *Drosophila* are close (57 and 51 %), on average, one TF accounts for 1.52/1.22 annotated binding site motifs for human (Hocomoco)/*Drosophila* (Jaspar). In accordance with this, the GTRD (Kolmykov et al., 2021) provides data on 21988/3027 ChIP-seq experiments for 1,531/595 human/*Drosophila* TFs. Therefore, the diversity of structural types of TFBS motifs has already been studied markedly better in human than in *Drosophila*.

The possible correspondence of the enriched motifs from the results of a *de novo* motif search to binding sites of target or partner TFs complicates the task of analyzing TF binding data *in vivo*. *In vitro* massive sequencing data, such as HT-SELEX or DAP-seq, reflect only the direct binding of target TFs, and completely exclude the cooperative binding of target TFs to any partner TFs and indirect binding of target TFs. Therefore, the nucleotide binding specificity of target TFs *in vitro* can determine only a fraction of their binding loci *in vivo*. *In vivo* TFBS sequencing data reflect the main cooperative mechanism of target TF binding to genomic DNA, including its interactions with various partner TFs (Morgunova, Taipale, 2017). This complicates the connection of enriched *de novo* motifs to specific partner TFs.

The variability of TFBS motifs derived from the systematization of their modern massive sequencing data reflects the diversity of the structure of TF DBDs. DBDs of TFs are important for the function of the direct binding of target and partner TFs. For example, only TFs from certain classes have the ability to function as dimers of closely related TFs (Amoutzias et al., 2008). Among the ones studied here (see the Table), those are TF classes Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2} and Nuclear receptors with C4 zinc fingers {2.1}. The main function of a TF, its ability to interact with genomic DNA, depends on the place of this TF in the general hierarchy of the structure of the DBDs of all TFs, that is, on a superclass, class, family and subfamily of this TF. Previously, these levels of hierarchical classification of TFs were defined by the structure of their DBDs and the alignments of amino acid sequences of DBDs of TFs (TFClass DB, Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018); notably, the similarity of TFBS motifs was not taken into account to define the hierarchy. A systematic analysis of the similarity of TFBS motifs can make the classification of TFs more efficient for the practical application at the stage of interpreting enriched motifs, the results of a *de novo* motif search based on massive mapping of TFBS *in vivo*, such as ChIP-seq.

Deducing the general topology of the branches of significantly similar TFBS motifs consists in selecting for each TF such a level of hierarchy among options of one class, one or more sister families (or subfamilies), or individual TF, so that for the TFs of the entire branch, most TF pairs have significantly similar binding site motifs. To determine the list of branches, we need the following: the hierarchical classification of TFs according to the structure of their DBDs from the TFClass/Plant-TFclass DBs; TFBS motif sets from DBs; the formula for calculating similarity in a pair of TF sets based on their binding site motifs (4). Identifying all branches along the TFClass/Plant-TFclass hierarchy will help avoid excessive detail in the output data of a *de novo* motif search. These misleading data and excessive information arise since for any of the individual classification units, such as a specific class, or family/subfamily, there is the variability of the TFBS motifs similarity not restricted. Initially, there were no such restrictions for DBD TFs, too (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018).

We include TF classes with more than 50 TFBS motifs in the analysis (see the Table). Of the seven largest human classes (Fig. 4), only one, the Tryptophan cluster factors {3.5} class, shows significant similarity of TFBS motifs. For the classes Basic leucine zipper factors (bZIP) {1.1} and Nuclear receptors with C4 zinc fingers {2.1}, similarity is below the significance threshold (value 3), but is still noticeable (values between 2 and 3). Even the classes Basic helix-loop-helix factors classes (bHLH) {1.2}, Homeo domain factors {3.1} and Fork head/winged helix factors {3.3} have lower similarity (values ranging from 1 to 2). However, for the C2H2 zinc finger factors {2.3} class, the similarity value is less than 1. This low value reflects the presence of a majority of TF pairs with completely different binding site motifs in this class; approximately the same similarity values are observed between binding site motifs in any pair of TFs from different classes of the same superclass (see values in cells above the diagonal in Fig. 5*a*, 6*a*, 7*a*). Similar discrepancies are observed at a lower level of TF families.

For each of the classes Basic leucine zipper factors (bZIP) {1.1} and Nuclear receptors with C4 zinc fingers {2.1}, in

most cases, several sister families are joined into one branch (Fig. 5*e*, 6*g*). For the classes Basic helix-loop-helix factors classes (bHLH) {1.2}, Homeo domain factors {3.1} and Fork head/winged helix factors {3.3} (Fig. 5*f*, 7*b*, *f*), partitioning into branches is closer to the level of families. The level of families is clearly not enough to distinguish branches in the C2H2 zinc finger factors {2.3} class (Fig. 6*f*). So, our analysis confirms clear differences in the variability of binding site motifs for the largest classes of human TFs (Fig. 4–7) (Lambert et al., 2018; Ambrosini et al., 2020). A concordant trend is also observed for the motifs of binding sites from the two largest classes of insect TFs (Fig. 8). This conclusion is in good agreement with the results of a massive comparison of the nucleotide specificity of orthologous human and *Drosophila* TFs, where it was found that, in general, human and *Drosophila* TFBS motifs showed a high level of conservation (Nitta et al., 2015). Later, a detailed analysis refined this finding. The analysis of similarity of binding site motifs of TFs from various classes in different eukaryotic taxa in lines of multicellular animals and higher plants showed that conservation in both animal and plant lineages is highly dependent on the TF class (Lambert et al., 2019). For example, almost half of the dissimilar binding site motifs of orthologous human and *Drosophila* TFs belonged to the C2H2 zinc finger factors {2.3} class, which is consistent with the results of our analysis (Fig. 6*f*, 8*a*). The analysis (Lambert et al., 2019) also showed that for some orthologous TFs of *Drosophila* and human, the similarity extended even to the level of subtle dinucleotide frequency preferences in the TFBS motifs.

We have also concluded that among the large classes of TFs, the class C2H2 zinc finger factors {2.3} has TFs with the most variable binding site motifs in human and *Drosophila* (Fig. 6*f*, 8*a*). Compared to the class C2H2 zinc finger factors {2.3}, both taxa have less variable TFBS motifs in the class Homeo domain factors {3.1}. However, for TFs of the class Homeo domain factors {3.1}, a greater variability of binding site motifs is found in *Drosophila* compared to human (Fig. 7*c*, 8*b*). This result may reflect differences in the TFBS motifs processing pipelines in the Hocomoco and Jaspar DBs.

In the Hocomoco DB, binding site motifs for each individual TF reflect data from several massive sequencing experiments for this TF (Kolmykov et al., 2021; Vorontsov et al., 2024), such as ChIP-seq and HT-SELEX; for example, often even available data of human and mouse species are combined. The goal of the analysis in the Hocomoco DB is to integrate all available data on the binding sites of individual TFs. This allows identifying as much as possible different structural types of motifs of the binding sites of each TF. The Jaspar DB has a simpler way of presenting each of the motifs with a separate experiment, which can be considered justified since there is still only a small amount of data on individual TFs. For insect TFBS motifs, an analysis similar to that carried out to obtain Hocomoco DB TFBS motifs has not yet been carried out, which is partly due to the significantly smaller pool of massive sequencing data available (Kolmykov et al., 2021; Rauluseviciute et al., 2024). It can be assumed that the approach of the Hocomoco DB compared to that of the Jaspar DB most likely reflects a greater number of minor motifs of binding sites for each of the TFs, which may contribute to a greater similarity of motifs deduced in our study, according to the formulas (2) and (4). Nevertheless, regular updates and an increase in the amount of data on known TFBS motifs in both Hocomoco and Jaspar DBs in recent years (Vorontsov et al., 2024; Rauluseviciute et al., 2024) indicate that the classification of TFBS motifs may be refined in the near future.

In general, based on our results, we can conclude that for both taxa, mammals and insects, marked differences in the similarity of binding site motifs of TFs from large classes and their families make it difficult to use the standard TFClass DB terminology, which includes TF classes, families and subfamilies, to describe the variability of TFBS motifs. Therefore, a more efficient detection of functionally involved TFs by massive sequencing of TFBS *in vivo* requires a systematic analysis of the similarity of binding site motifs of known TFs in order to define the variability of TFBS motifs within different elementary classification units from classes to individual TFs.

In the future, a more extensive analysis of the similarity of binding site motifs within all classes, families, subfamilies of TFs and individual TFs in model species of mammals, insects and higher plants can be a solid basis for more efficient definition of TFBS motifs from ChIP-seq massive sequencing data. Based on the performed massive analysis, we suggest that the results of a *de novo* motif search, for the detected enriched motifs, should indicate not only the names of TFs with the names of the class/family/subfamily attached to them, but also the branches of the hierarchical classification of TFs defined in our study. These branches are composite classification units that integrate several consecutive hierarchy levels. Each branch represents, within the framework of united multi-level classification of TFs by similarity and DBD alignment, a set of TFs with significantly similar binding site motifs.

Conclusion

In this work, we present the approach for a systematic analysis of the similarity of the motifs of binding sites of known TFs based on a multi-level hierarchy of TFs according to the structure of DBDs from the TFClass DB, which includes the levels of superclasses, classes, families, subfamilies and individual TFs. In the general hierarchy, we determined for the large classes of mammalian (human) and insect (fruit fly) TFs the common trees of branches with TFs significantly similar in motifs of binding sites. Our analysis included seven mammalian TF classes, Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1}, C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} and Tryptophan cluster factors {3.5}, and two classes of insect TFs, C2H2 zinc finger factors {2.3} and Homeo domain factors {3.1}. We have shown that both for the taxon of mammals and for the taxon of insects, the similarity of the binding site motifs is noticeably different among TFs from distinct classes. A systematic analysis of the similarity of the binding site motifs of structurally related TFs, determined according to the hierarchical classification, allowed to determine the levels of the hierarchy (classes, families, subfamilies, TFs), starting from which and lower in the hierarchy the bind-

ing site motifs of known TFs become significantly similar. In addition to improving the identification of involved TFs from the results of a *de novo* motif search, leading to more efficient identification of gene regulation mechanisms, our results may refine the hierarchical classification of TFs by their DBDs. We do not redefine the classification of TFs by elementary units from the class, family and lower in the hierarchy; we provide additional information about the similarity of the TFBS motifs, which reflects the main function of TFs, the function of specific binding to the DNA sequence, which, of course, should more accurately distinguish different TFs.

References

Ambrosini G., Vorontsov I., Penzar D., Groux R., Fornes O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Bucher P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.* 2020;21(1):114. doi 10.1186/s13059-020-01996-3

Amoutzias G.D., Robertson D.L., Van de Peer Y., Oliver S.G. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci.* 2008;33(5):220-229. doi 10.1016/j.tibs.2008.02.002

Bailey T.L. STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* 2021;37(18):2834-2840. doi 10.1093/bioinformatics/btab203

Blanc-Mathieu R., Dumas R., Turchi L., Lucas J., Parcy F. Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.* 2024;29(1):40-51. doi 10.1016/j.tplants.2023.06.023

D'haeseleer P. What are DNA sequence motifs? *Nat Biotechnol.* 2006; 24(4):423-425. doi 10.1038/nbt0406-423

de Martin X., Sodaei R., Santpere G. Mechanisms of binding specificity among bHLH transcription factors. *Int J Mol Sci.* 2021;22(17): 9150. doi 10.3390/ijms22179150

Franco-Zorrilla J.M., López-Vidriero I., Carrasco J.L., Godoy M., Vera P., Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci USA.* 2014;111(6):2367-2372. doi 10.1073/pnas.1316278111

Gupta S., Stamatoyannopolous J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol.* 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24

Hammal F., de Langen P., Bergon A., Lopez F., Ballester B. ReMap 2022: A database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* 2022;50(D1):D316-D325. doi 10.1093/nar/gkab996

Johnson D.S., Mortazavi A., Myers R.M., Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science.* 2007;316(5830): 1497-1502. doi 10.1126/science.1141319

Jolma A., Yan J., Whitington T., Toivonen J., Nitta K.R., Rastas P., Morgunova E., … Hughes T.R., Lemaire P., Ukkonen E., Kivioja T., Taipale J. DNA-binding specificities of human transcription factors. *Cell.* 2013;152(1-2):327-339. doi 10.1016/j.cell.2012.12.009

Kolmykov S., Yevshin I., Kulyashov M., Sharipov R., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., Kolpakov F. GTRD: An integrated view of transcription regulation. *Nucleic Acids Res.* 2021; 49(D1):D104-D111. doi 10.1093/nar/gkaa1057

Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell.* 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029

Lambert S.A., Yan A.W.H., Sasse A., Cowley G., Albu M., Caddick M.X., Morris Q.D., Weirauch M.T., Hughes T.R. Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet.* 2019;51(6):981-989. doi 10.1038/s41588-019-0411-1

Levitsky V., Zemlyanskaya E., Oshchepkov D., Podkolodnaya O., Ignatieva E., Grosse I., Mironova V., Merkulova T. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res.* 2019;47(21):e139. doi 10.1093/nar/gkz800

Liu B., Yang J., Li Y., McDermaid A., Ma Q. An algorithmic perspective of *de novo* cis-regulatory motif finding based on ChIP-seq data. *Brief Bioinform.* 2018;19(5):1069-1081. doi 10.1093/bib/bbx026

Lloyd S.M., Bao X. Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr Protoc Cell Biol.* 2019;84(1):e89. doi 10.1002/cpcb.89

Morgunova E., Taipale J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol.* 2017;47:1-8. doi 10.1016/j.sbi.2017.03.006

Nagy G., Nagy L. Motif grammar: The basis of the language of gene expression. *Comput Struct Biotechnol J.* 2020;18:2026-2032. doi 10.1016/j.csbj.2020.07.007

Najafabadi H.S., Mnaimneh S., Schmitges F.W., Garton M., Lam K.N., Yang A., Albu M., Weirauch M.T., Radovani E., Kim P.M., Greenblatt J., Frey B.J., Hughes T.R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol.* 2015;33(5): 555-562. doi 10.1038/nbt.3128

Nakato R., Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform.* 2017;18(2):279-290. doi 10.1093/bib/bbw023

Nitta K.R., Jolma A., Yin Y., Morgunova E., Kivioja T., Akhtar J., Hens K., Toivonen J., Deplancke B., Furlong E.E., Taipale J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife.* 2015;4:e04837. doi 10.7554/eLife.04837

Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., … Lenhard B., Sandelin A., Wasserman W.W., Parcy F., Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059

Schneider T.D., Stephens R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990;18(20):6097-6100. doi 10.1093/nar/18.20.6097

Shen W.K., Chen S.Y., Gan Z.Q., Zhang Y.Z., Yue T., Chen M.M., Xue Y., Hu H., Guo A.Y. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* 2023;51(D1):D39-D45. doi 10.1093/nar/gkac907

Skene P.J., Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife.* 2017;6:e21856. doi 10.7554/eLife.21856

Slattery M., Zhou T., Yang L., Dantas Machado A.C., Gordân R., Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.* 2014;39(9):381-399. doi 10.1016/j.tibs.2014.07.002

Sokal R.R., Michener C.D. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull.* 1958;38:1409-1438. Available: https://archive.org/details/cbarchive_33927_astatisticalmethodforevaluatin1902/page/n1/mode/2up

Spitz F., Furlong E.E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet.* 2012;13(9):613-626. doi 10.1038/nrg3207

Stormo G.D., Zhao Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet.* 2010;11(11):751-760. doi 10.1038/nrg2845

Taing L., Dandawate A., L'Yi S., Gehlenborg N., Brown M., Meyer C.A. Cistrome Data Browser: integrated search, analysis and visualization of chromatin data. *Nucleic Acids Res.* 2024;52(D1):D61-D66. doi 10.1093/nar/gkad1069

Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A., Kamenets V., … Medvedeva Y.A., Jolma A., Kolpakov F., Makeev V.J., Kulakovskiy I.V. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* 2024;52(D1):D154-D163. doi 10.1093/nar/gkad1077

Wasserman W.W., Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet.* 2004;5(4):276-287. doi 10.1038/nrg1315

Weirauch M.T., Yang A., Albu M., Cote A.G., Montenegro-Monter A., Drewe P., Najafabadi H.S., … Bouget F.Y., Ratsch G., Larrondo L.F., Ecker J.R., Hughes T.R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* 2014;158(6):1431-1443. doi 10.1016/j.cell.2014.08.009

Wingender E. Classification scheme of eukaryotic transcription factors. *Mol Biol.* 1997:31(4):483-497. (translated from Вингендер Э. Классификация транскрипционных факторов эукариот. *Молекулярная биология*. 1997;31(4):584-600. Russian)

Wingender E. Criteria for an updated classification of human transcription factor DNA-binding domains. *J Bioinform Comput Biol.* 2013;11(1):1340007. doi 10.1142/S0219720013400076

Wingender E., Schoeps T., Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 2013;41(D1):D165-D170. doi 10.1093/nar/gks1123

Wingender E., Schoeps T., Haubrock M., Dönitz J. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* 2015;43(D1):D97-D102. doi 10.1093/nar/gku1064

Wingender E., Schoeps T., Haubrock M., Krull M., Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* 2018;46(D1):D343-D347. doi 10.1093/nar/gkx987

Zambelli F., Pesole G., Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform.* 2013;14(2):225-237. doi 10.1093/bib/bbs016

Zenker S., Wulf D., Meierhenrich A., Viehöver P., Becker S., Eisenhut M., Stracke R., Weisshaar B., Bräutigam A. Many transcription factor families have evolutionarily conserved binding motifs in plants. *Plant Physiol.* 2025;198(2):kiaf205. doi 10.1093/plphys/kiaf205

Conflict of interest. The authors declare no conflict of interest.
Received July 9, 2025. Revised September 9, 2025. Accepted September 10, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):940-951

doi 10.18699/vjgb-25-100

PlantReg 1.1 identifies the mutual arrangement of transcription factor binding sites in the target promoters for the elucidation of molecular mechanisms within regulatory networks

V.V. Lavrekha [1, 2#], N.A. Omelyanchuk [1#], A.G. Bogomolov [1], Y.A. Ryabov [1], P.K. Mukebenova [1, 2], E.V. Zemlyanskaya [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
ezemlyanskaya@bionet.nsc.ru

Abstract. The development of high-throughput sequencing has expanded the possibilities for studying the regulation of gene expression, including the reconstruction of gene regulatory networks and transcription factor regulatory networks (TFRNs). Identifying the molecular aspects for regulation of biological processes via these networks remains a challenge. Solving this problem for plants will significantly advance the understanding of the mechanisms shaping agronomically important traits. Previously, we developed the PlantReg program to reconstruct the transcriptional regulation of biological processes in the model species *Arabidopsis thaliana* L. The links established by this program between TFRNs and the genes regulating biological processes specify the type of regulation (activation/suppression). However, the program does not determine whether activation/suppression of the target gene is due to the cooperative or competitive interaction of transcription factors (TFs). We assumed that using information on the mutual arrangement of TF binding sites (BSs) in the target gene promoter as well as data on the activity type of TF effector domains would help to identify the cooperative/competitive action of TFs. We improved the program and created PlantReg 1.1, which enables precise localization of TF BSs in extended TF binding regions identified from genome-wide DAP-seq profiles (https://plamorph.sysbio.ru/fannotf/). To demonstrate the capabilities of the program, we used it to investigate the regulation of target genes in previously reconstructed TFRNs for auxin response and early reaction to salt stress in *A. thaliana*. The study focused on genes encoding proteins involved in chlorophyll and lignin biosynthesis, ribosome biogenesis, and abscisic acid (ABA) signaling. We revealed that the frequency of competitive regulation under the influence of auxin or salt stress could be quite high (approximately 30 %). We demonstrated that competition between bZIP family TFs for common BS is a significant mechanism of transcriptional repression in response to auxin, and that auxin and salt stress can engage common competitive regulatory mechanisms to modulate the expression of some genes in the ABA signaling pathway.

Key words: gene ontology; biological processes; gene regulatory networks; binding site; transcription factor; *Arabidopsis thaliana*

For citation: Lavrekha V.V., Omelyanchuk N.A., Bogomolov A.G., Ryabov Y.A., Mukebenova P.K., Zemlyanskaya E.V. PlantReg 1.1 identifies the mutual arrangement of transcription factor binding sites in the target promoters for the elucidation of molecular mechanisms within regulatory networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):940-951. doi 10.18699/vjgb-25-100

Funding. The work was funded by the budget project FWNR-2022-0020.

Программа PlantReg 1.1: анализ взаимного расположения сайтов связывания транскрипционных факторов в промоторах генов-мишеней для уточнения молекулярных механизмов их активности в регуляторных сетях

В.В. Лавреха [1, 2#], Н.А. Омельянчук [1#], А.Г. Богомолов [1], Ю.А. Рябов [1], П.К. Мукебенова [1, 2], Е.В. Землянская [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
ezemlyanskaya@bionet.nsc.ru

[#] Equal contribution.

Аннотация. Развитие высокопроизводительного секвенирования расширило возможности изучения регуляции экспрессии генов, в том числе для реконструкции генных регуляторных сетей и регуляторных сетей транскрипционных факторов (РСТФ). Актуальной задачей остается выявление молекулярных аспектов регуляции данными сетями биологических процессов. Решение этой задачи для растений позволит существенно продвинуться в понимании механизмов формирования хозяйственно важных признаков. Ранее мы разработали программу PlantReg для реконструкции транскрипционной регуляции биологических процессов у модельного вида *Arabidopsis thaliana* L. Воспроизводимые этой программой связи между РСТФ и генами, обеспечивающими протекание биологических процессов, охарактеризованы по типу регуляции (активация/подавление). Однако программа не позволяла определять, в каких случаях активация/подавление экспрессии гена-мишени обусловлены кооперативным или конкурентным взаимодействием транскрипционных факторов (ТФ). Мы предложили использовать информацию о взаимном расположении сайтов связывания (СС) ТФ в промоторе гена-мишени, а также данные о типе активности трансактивационных доменов ТФ для выявления кооперативного/конкурентного действия ТФ. Мы усовершенствовали программу, создав версию PlantReg 1.1, где обеспечили возможность точной локализации СС ТФ в протяженных районах связывания ТФ, устанавливаемых на основании полногеномных профилей DAP-seq (https://plamorph.sysbio.ru/fannotf/). Для демонстрации возможностей программы была исследована регуляция генов-мишеней ранее реконструированных нами РСТФ ответа на ауксин и солевой стресс у *A. thaliana*. В фокусе изучения были гены, кодирующие белки, участвующие в процессах биосинтеза хлорофилла и лигнина, биогенеза рибосом и в передаче сигнала абсцизовой кислоты. В данной работе установлено, что частота случаев конкурентной регуляции под влиянием ауксина и солевого стресса может быть достаточно высока (около 30 %). Показано, что конкуренция ТФ семейства bZIP за общие СС является значимым механизмом подавления транскрипции в ответ на ауксин, и что ауксин и солевой стресс могут задействовать общие механизмы конкурентной регуляции для модуляции экспрессии некоторых генов сигнального пути абсцизовой кислоты.
Ключевые слова: генная онтология; биологические процессы; генные регуляторные сети; сайт связывания; транскрипционный фактор; *Arabidopsis thaliana*

Introduction

Development of genome-wide analysis techniques (such as RNA-seq (Deshpande et al., 2023), ChIP-seq (Park, 2009), and DAP-seq (O'Malley et al., 2016)) has opened up wide opportunities for systems biological research on mechanisms that ensure transcriptional regulation of biological processes and the formation of phenotypes (Marshall-Colón, Kliebenstein, 2019; Zemlyanskaya et al., 2021). Based on the analysis of genomic and transcriptomic data, the community is actively developing approaches to infer gene regulatory networks and TFRNs (Ko, Brandizzi, 2020; Rybakov et al., 2024). A TFRN is a set of regulatory interactions (links) between TF-coding genes, represented as a graph. The graph nodes correspond to the genes, and the directed edges reflect the regulatory interactions of a TF, encoded by one gene, with another gene. TFRN inference and identification of relationships between these networks and biological processes (or phenotypes) are essential to understanding the core regulatory circuits that drive biological processes, and to developing predictive models for these regulations (Huang et al., 2025; Leong et al., 2025; Sun Y. et al., 2025).

Several software tools for TFRN inference in various species are currently available to researchers. For example, the NetAct R package (Su et al., 2022) allows reconstructing mammalian TFRNs based on transcriptomic data and a database of TF target genes curated by the authors. Previously, we developed the CisCross-FindTFnet program for TFRN inference in the model plant species *Arabidopsis thaliana* (Omelyanchuk et al., 2024) and the PlantReg program for establishing regulatory links between TFRNs and genes that mediate the biological processes under the TFRN control (Lavrekha et al., 2024). Both programs integrate data from transcriptomic experiments and a representative collection of genome-wide DAP-seq TF binding profiles, with PlantReg employing the results of CisCross-FindTFnet as input data.

An important step in TFRN inference is to determine the mode of regulation exerted by a TF within the network (activators or repressors), since this characteristic shapes the network topology and dynamics (Dhatterwal et al., 2024). Large-scale determination of the activity of transcriptional effector domains in more than 400 *A. thaliana* TFs (Hummel et al., 2023) contributed to solving this problem. However, this is not sufficient for the correct classification of links within the network, since many TFs can function both as activators and suppressors, depending on the cell type, conditions, TF isoforms, specific promoters, and other factors (Boyle, Després, 2010; Martínez et al., 2018; Nagahage et al., 2018; Wang et al., 2020). This is why, when reconstructing the TFRN from transcriptomic data, the modes of regulation exerted by TFs are usually inferred from the profiles of their targets among differentially expressed genes (DEGs) (Su et al., 2022; Omelyanchuk et al., 2024).

Previously, we reconstructed two TFRNs in *A. thaliana*: the first, TFRN-A, controls the transcriptional response to auxin, the second, TFRN-S, controls the early response to salt stress (Lavrekha et al., 2024; Omelyanchuk et al., 2024). Using the PlantReg algorithm, we demonstrated how TFRN-A is involved in regulation of four different biological processes by auxin (activation of ribosome biogenesis and suppression of response to ABA, as well as chlorophyll and lignin biosynthesis), and how TFRN-S enhances ABA response during early salt stress. In these networks, TFs were divided into four classes: upregulated activator (UA), upregulated suppressor (US), downregulated activator (DA), and downregulated suppressor (DS). DAs and DSs form an R subnetwork (normally active before stimulus application, repressed due to stimulus action), UAs and USs set up an A subnetwork (activated by the stimulus).

An important role of transcriptional repression has been identified in transcriptional responses to both auxin and salt

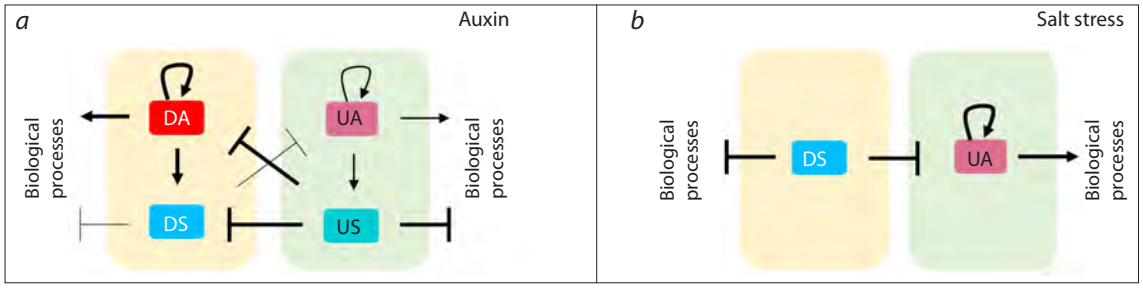


Fig. 1. Principles of regulation of biological processes by TFs from TFRN-A (*a*) and TFRN-S (*b*).

Yellow and green rectangles represent the repressed and activated subnetworks of TFRNs. Arrow thickness reflects the number of corresponding links in TFRNs. UA – upregulated activator; US – upregulated suppressor; DA – downregulated activator; DS – downregulated suppressor.

stress. The auxin response is characterized by extensive reprogramming of the large R subnetwork, which was active before hormone treatment, through its suppression by US-type TFs from the A subnetwork (Fig. 1*a*) (Omelyanchuk et al., 2024). In contrast, the salt stress response activates the wide A subnetwork, partly through the inhibition of its DS-type suppressors from the R subnetwork (Fig. 1*b*) (Lavrekha et al., 2024).

The majority of the suppressors from both TFRNs are also involved in the regulation of the above-mentioned biological processes, affected by auxin and salt stress (Lavrekha et al., 2024; Omelyanchuk et al., 2024). However, according to the literature, most of the predicted suppressors in both TFRNs possess an activator-type transcriptional effector domain (Hummel et al., 2023; Omelyanchuk et al., 2024). Suppression of targets by these TFs may occur due to their cooperative or competitive interactions with other TFs. The PlantReg program enables establishing regulatory links between TFs and genes that mediate biological processes, but it does not detect cooperation or competition among TFs. At the same time, it is crucial to understand the mechanisms of TF interactions in transcriptional regulation to effectively use TFRNs and their relations to biological processes in plant bioengineering.

Information on the mutual arrangement of the TF binding sites (BSs) in the target promoter, coupled with data on the activity of the TF effector domains (Hummel et al., 2023), can be used to identify and characterize the cooperative or competitive action of TFs. For example, if the BSs of two predicted suppressors, operating within the same subnetwork, are close to each other, and only one has a transcriptional effector domain exhibiting suppressor activity, while the other TF is a transcriptional activator, it is plausible to assume that a cooperative interaction between TFs converts an activator TF into a repressor (Fig. 2*a*). Such examples are widespread and described in detail in the literature (Hanna-Rose, Hansen, 1996; Ahn et al., 2006; Veerabagu et al., 2014; Martínez et al., 2018; Wang et al., 2020).

Similarly, if the BSs of a predicted activator from one subnetwork and a predicted repressor from another subnetwork overlap in the promoter of a target gene, and the predicted activity of one of the TFs does not match the established activity of its transcriptional effector domain, we can assume that TFs may compete for the common BS, and the replacement of a strong activator with a weaker one manifests itself as suppression of the target gene, while the replacement of a strong repressor with a weaker one manifests itself as activation of the target gene (Fig. 2*b*). A decrease in promoter activity with an increase in the concentration of a weak activator compared to a strong one, as well as the reverse transition, have been shown in a number of experiments (Tamura et al., 2004; Zhang et al., 2006; Chupreta et al., 2007; Selvaraj et al., 2015; Ren et al., 2015; Brackmann et al., 2018).

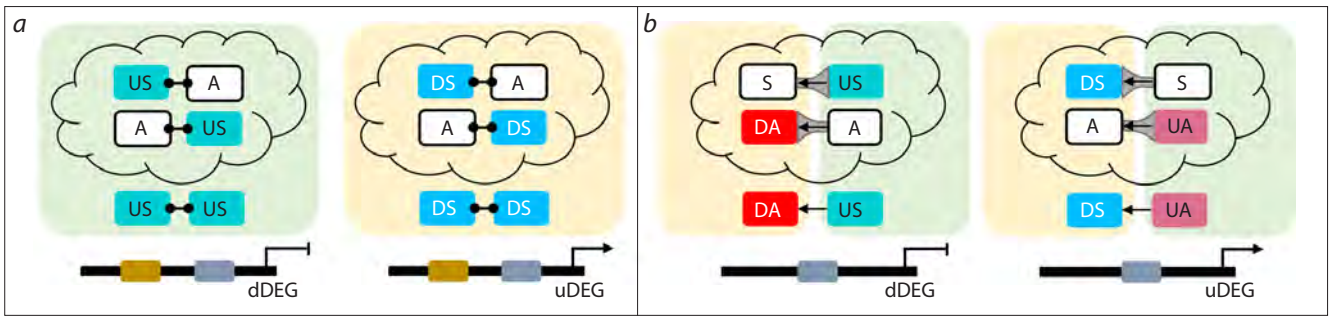


Fig. 2. Cooperative (*a*) and competitive (*b*) regulation of a target gene by a pair of TFs from a TFRN.

Yellow and green rectangles represent the repressed and activated TFRN subnetworks. Predicted TF modes of regulation are shown at the bottom, while possible alternative modes are shown at the top (in the cloud). The connected dots between TFs in (*a*) denote protein interactions; in (*b*) arrows represent the substitution of one TF with another after stimulus application and gray funnels designate the ratio of TF activities (larger bases correspond to higher activity); uDEG – upregulated DEGs; dDEG – downregulated DEGs; UA – upregulated activator; US – upregulated suppressor; DA – downregulated activator; DS – downregulated suppressor.

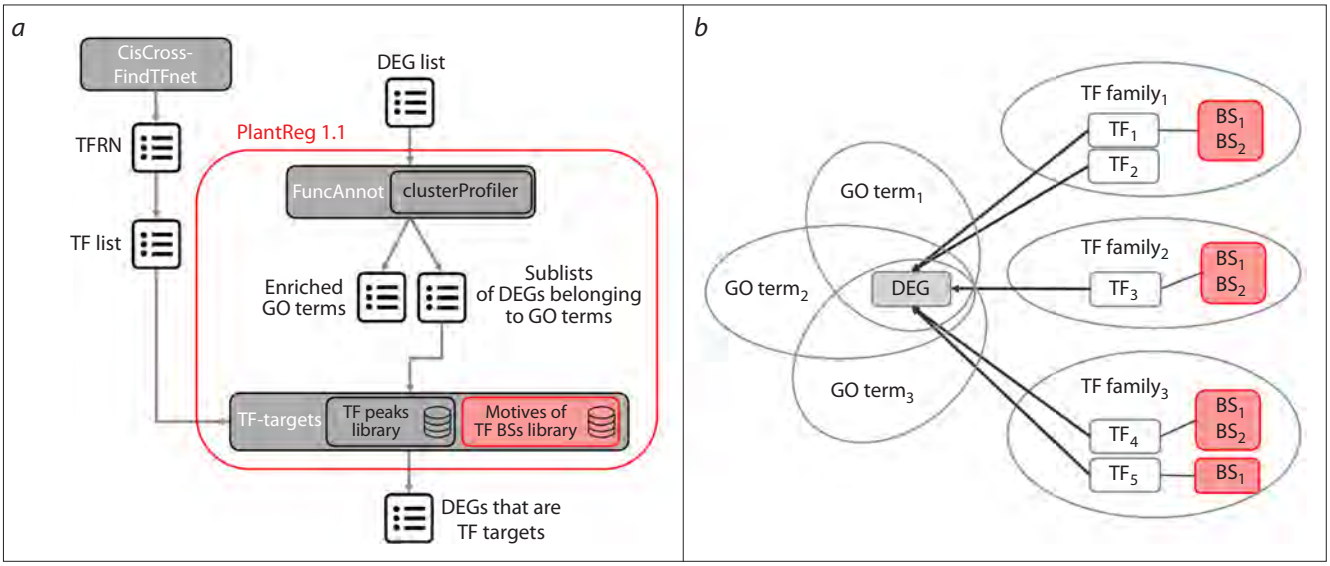


Fig. 3. PlantReg 1.1 workflow (*a*) and output structure (block 1) (*b*).
Updates in PlantReg 1.1 compared to the original version are highlighted in pink.

To identify TF targets, the PlantReg program recruits DAP-seq peaks. However, this does not enable precise localization of TF BSs, since the peak size (over 150 bp) significantly exceeds the length of the sequences recognized by TFs (below 20 bp). In this study, we improved the program by creating PlantReg version 1.1, which enables precise localization of TF BSs in extended TF binding regions from genome-wide DAP-seq profiles (https://plamorph.sysbio.ru/fannotf/). We used PlantReg 1.1 to identify genes involved in chlorophyll and lignin biosynthesis, ribosome biogenesis, and ABA signaling, the expression of which can be suppressed under TFRN-A or TFRN-S control due to competition between TF activators for common BSs.

The analysis revealed that the frequency of competitive regulation under auxin and salt stress exposure can be quite high. Furthermore, we demonstrated that competition between bZIP family TFs for common BSs is an essential mechanism for transcription repression in *A. thaliana* auxin response, and that auxin and salt stress can utilize common competitive regulation to modulate the expression of some genes in ABA signaling.

Materials and methods

Integration of data on TF BSs in 5′-regulatory regions into PlantReg 1.1. The original PlantReg version (Lavrekha et al., 2024) was designed to reconstruct the mechanisms underlying transcriptional regulation of biological processes in *A. thaliana* based on the analysis of a DEG list and a list of TFs – known or putative transcriptional regulators of these DEGs. PlantReg performs gene ontology (GO) enrichment analysis of the input DEG list, and identifies potential TF targets among DEGs associated with enriched biological processes, recruiting genome-wide TF binding profiles available in the web version of the program (Fig. 3*a*). The output of PlantReg is presented in five blocks, which reflect the relationships between biological processes, DEGs, and TFs that regulate the expression of these DEGs.

The basic workflow of the updated PlantReg 1.1 version is shown in Figure 3. In addition to the original functionality, it includes data on recognized TF BSs in the 5′-regulatory regions (Fig. 3*a*), which are added to output blocks 1 and 4 to enable investigation on the mutual arrangement of BSs in promoters. The output block 1 in the original PlantReg version presents a sublist of DEGs associated with enriched biological processes (Fig. 3*b*). Each gene in the sublist is characterized by a set of associated GO terms (biological processes) with evidence codes, the number of GO terms, a list of potential transcriptional regulators with an indication of their TF families, and the number of TFs.

In output block 4, the same information is presented in an alternative format with the GO terms and transcriptional regulators for each gene listed line by line. In PlantReg 1.1, the nucleotide sequence of the TF BS recognized in the corresponding TF binding region, the genomic coordinates of the TF BS (block 4) or the coordinates of the TF BS relative to the transcription start site (block 1), and the DNA strand harboring the TF BS were added to the description of each gene (Fig. 3*b*). Information on the TF BS localization is available only when the CisCross-MACS2 genome-wide profile collection is selected as a parameter.

Recognition of TF BSs in the 5′-regulatory regions of *A. thaliana* genes. Position frequency matrices describing the BSs of *A. thaliana* TFs were generated by *de novo* motif search in DAP-seq peaks from the CisCross-MACS2 collection available in the web version of the PlantReg program (Lavrekha et al., 2024). The CisCross-MACS2 peak set collection was compiled previously (Lavrekha et al., 2022) by processing raw data from genome-wide DAP-seq profiling of BSs for 403 *A. thaliana* TFs (O'Malley et al., 2016). In each peak set, the top 2,000 peaks were selected by height and used

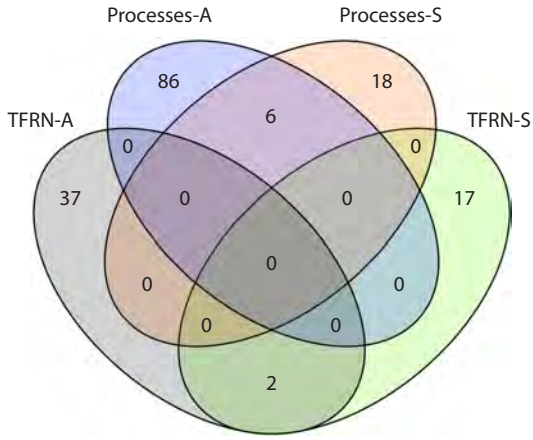


Fig. 4. The number of genes encoding TFs in TFRN-A and TFRN-S, as well as TF target genes, which mediate biological processes affected by auxin (Processes-A) or salt stress (Processes-S).

The "Processes-A" set includes genes for chlorophyll biosynthesis, lignin biosynthesis, ribosome biogenesis and ABA transport, conjugation, and the signaling pathways. The "Processes-S" set includes only genes for ABA transport, conjugation, and the signaling pathway.

for *de novo* motif search employing the STREME program (Bailey et al., 2021).

A background set was generated by the AntiNoise program (Raditsa et al., 2024). The motif with the highest enrichment significance (with a *p*-value below 0.05) was assumed to describe the BS for TF of interest. To test this assumption, the identified motifs were juxtaposed to known TF BSs by comparing with motifs from the JASPAR2024 CORE (Rauluseviciute et al., 2024), CisBP (Weirauch et al., 2014), and ArabidopsisDAPv1 (O'Malley et al., 2016) databases using the Tomtom program (Gupta et al., 2007).

The search for potential TF BSs in the 5′-regulatory regions *of A. thaliana* genes ([–2500; +1) relative to the transcription start site) was performed using the position weight matrix method with the scan_sequence function of the universalmotif R-package (Tremblay, 2024). To extract the nucleotide sequences of the 5′-regulatory regions, the *A. thaliana* TAIR10 genome version (Lamesch et al., 2012) and the Araport11 genomic annotation (Cheng et al., 2017) were used.

Search for genes, the transcription of which is regulated by competitive suppression or activation. Regulatory links between components of the TFRN-A/S (Lavrekha et al., 2024; Omelyanchuk et al., 2024) and genes involved in biological processes affected by auxin and salt stress, as well as competitive gene suppression or activation under auxin and salt stress exposure were identified using PlantReg 1.1. As input, we used the lists of TFs that constituted TFRN-A (39 elements) and TFRN-S (19 elements) (the lists are designated as "TFRN-A" and "TFRN-S" in Fig. 4) (Table S1)[1], as well as the lists of DEGs upregulated (uDEGs) and downregulated (dDEGs) by auxin (5,201 uDEGs and 6,704 dDEGs) or salt stress (1,476 uDEGs and 944 dDEGs), which were used previously to reconstruct the TFRNs (De Rybel et al., 2012; Wu et al., 2021; Omelyanchuk et al., 2024).

[1] Supplementary Tables S1–S7 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Lavrekha_Engl_29_7.xlsx

The lists of uDEGs and dDEGs from the two transcriptomic experiments were separately fed into the PlantReg 1.1 program along with the corresponding list of TFs from the TFRN-A or TFRN-S. The threshold for GO terms enrichment was set at 0.001. To localize the TF binding regions, the CisCross-MACS2 collection of genome-wide TF binding profiles and a 5′-regulatory region length of 1,000 bp were selected. This analysis resulted in "TF-regulator–target gene" pairs, where the TFs from the TFRN-A or TFRN-S were TF-regulators, and the uDEGs and dDEGs from the corresponding transcriptomic experiment were the target genes.

The DAP-seq data, recruited by PlantReg 1.1 to map TF binding regions in the *A. thaliana* genome, contain two types of peak sets: in the first case ("col" peak sets), native genomic DNA from leaves was used to prepare libraries; in the second case ("colamp" peak sets), genomic DNA with methylcytosine epigenetic marks removed by PCR amplification was used. TF-regulator–target gene pairs reconstructed using "col" peak sets were selected from the PlantReg 1.1 output. Next, among the target genes regulated by TFRN-A, we chose the genes annotated with GO terms related to chlorophyll biosynthesis (16 genes), lignin biosynthesis (14 genes), ABA signaling (34 genes), and ribosome biogenesis (28 genes); these processes were previously considered in (Omelyanchuk et al., 2024).

Among the target genes regulated by TFRN-S, we selected genes annotated with GO terms related to ABA signaling (24 genes), which was previously discussed in (Lavrekha et al., 2024). As a result, 110 genes were chosen (designated as "Processes-A" and "Processes-S" in Fig. 4) (Tables S1–S3).

To identify among these genes the ones potentially regulated by competitive suppression or activation, we selected the genes that met the following requirements: a) more than one TF was involved in the regulation of the gene, b) the BSs of these TFs considerably overlapped (over 80 %), and c) the genes encoding these TFs changed their expression in opposite directions in the transcriptomic experiment.

Results

A collection of the predicted TF BSs in 5′-regulatory regions of *A. thaliana* genes, integrated into PlantReg 1.1

To enable prediction of cooperative and competitive interactions of TFs in the transcriptional regulation of biological processes, automatic localization of TF BSs in 5′-regulatory regions was implemented in PlantReg 1.1. For this purpose, the results of TF BS recognition in promoters using the position weight matrices (see the "Materials and methods" section) were systematized and integrated into PlantReg 1.1. For 300 TFs (74 %), the motif identified *de novo* in at least one peak set ("col" or "colamp") was similar to a known BS for this TF available in the JASPAR, CisBP, or ArabidopsisDAPv1 databases (Fig. 5*a*). The proportion of TFs with BSs recognized in more than 90 % of peaks mapped to 5′-regulatory regions was quite high and varied from 42 (for 500 bp-long 5′-regulatory regions) to 74 % (for 2,000 bp-long 5′-regulatory regions) (Fig. 5*b*).

In the following sections, we illustrate the potential of using the new functionality of PlantReg 1.1 to solve specific biological challenges.

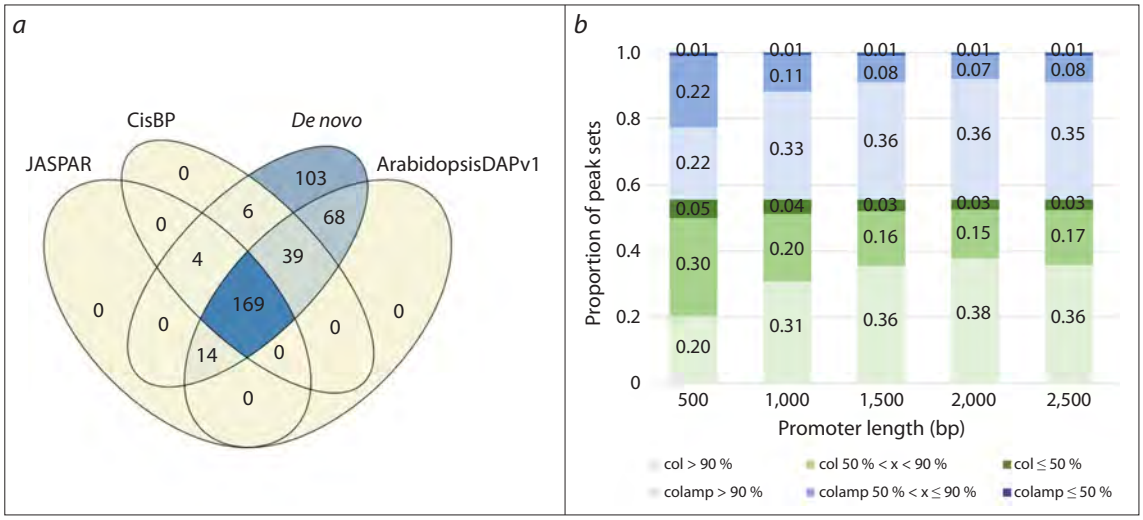


Fig. 5. Characteristics of the collection of predicted TF BSs in the 5′-regulatory regions of *A. thaliana* genes integrated into PlantReg 1.1.

a – comparison of motifs recognized de novo in DAP-seq peaks with known TF BSs in the JASPAR, CisBP, and ArabidopsisDAPv1 databases; *b* – proportions of DAP-seq peak sets mapped to the 5′-regulatory regions (col – shades of green/colamp – shades of blue) with the motifs recognized in more than 90 % of peaks (light shade), in 50–90 % of peaks, and in less than 50 % of peaks (dark shade).

Competitive regulation of gene expression in response to auxin and salt stress in *A. thaliana*

We assumed above that the suppression of target gene transcription with an increase in the level of US-type TFs or activation due to a decrease in the level of DS-type TFs in response to auxin and under salt stress may occur through competitive regulation of their expression by a pair of activator TFs. To test this hypothesis, we identified regulatory links between TFRN-A/S and genes involved in chlorophyll and lignin biosynthesis, ribosome biogenesis, and ABA signaling using PlantReg 1.1. Fourteen genes were picked as potential targets for competitive regulation by TFs from TFRN-A and TFRN-S (Tables S1, S6 and S7).

Additionally, 11 genes encoding TFs from TFRN-A and TFRN-S were also found as potential targets for competitive regulation (Tables S4 and S5). All 25 selected genes (12 dDEGs, 10 uDEGs, and three genes, *ABCG25* (*ATP-binding cassette family G25*), *GBF3* (*G-box binding factor 3*), and *PYL7/RCAR2* (*PYR1-like 7/Regulatory components of ABA receptor 2*), the expression of which changed in opposite directions under auxin and salt stress) made up as much as 32 % of the total number of genes regulated by suppressors (79 genes) (Tables S1, S6 and S7). Thus, the competitive regulation of the target genes by TFRNs may be a frequent event.

TFs are grouped into families, classes, and superclasses based on the similarity of their DNA-binding domains (Blanc-Mathieu et al., 2024). TFs from the same family often recognize similar DNA sequences and, therefore, can compete for the binding sites. In the 5′-regulatory regions of 25 selected genes, TFs can compete within the following six families: AP2/ERF (APETALA2/ETHYLENE RESPONSIVE FACTOR), bZIP (BASIC LEUCINE-ZIPPER), BZR1/BES1 (BRASSINAZOLE RESISTANT 1/BRI1 EMS SUPPRESSOR 1), HD-ZIP (HOMEODOMAIN LEUCINE ZIPPER), MYB (V-MYB AVIAN MYELOBLASTOSIS VIRAL ONCOGENE HOMOLOG), WRKY (Table S6). In addition, we identified possible competition between TFs from different families of the same superclass, namely: “basic domains” and “Helix-Turn-Helix domains” (Table S6).

Moreover, in the promoters of uDEGs *MAPKKK18* (*Mitogen-activated protein kinase kinase kinase 18*) and *RRP47* (*Sas10/Utp3/C1D family*), the same BS can be occupied by TFs from the families belonging to two different superclasses: AP2/ERF (“Beta-hairpin exposed by an alpha/beta-scaffold” superclass) and bZIP (“basic domains” superclass) in the first case, and AP2/ERF and LBD (“Zinc-coordinating DNA binding domains” superclass) in the second case (Table S6). In the distal promoter of dDEG *GBF3*, TFs from the families of two other superclasses, bZIP (“basic domains”) and MYB (“Helix-Turn-Helix domains”), can compete for the common BS. We also detected a possible competition for the common BS among TFs from different families belonging to two (in the promoters of *AFP1* (*ABI five binding protein*), *MYB73* and *PYL7*) and even three different superclasses (in the distal promoter of *GBF3*) (Table S6).

Competition of bZIP family TFs in promoters of genes regulated by TFRN-A

To identify combinations of activator TFs systematically recruited by TFRN-A or TFRN-S to suppress target gene expression, we conducted a comparative analysis of TF-regulator–target gene pairs determined with PlantReg 1.1. Three DA-type TFs (bZIP3, bZIP68, and GBF3) and a US-type TF (bZIP53) share common BSs in the promoters of several genes regulated by TFRN-A. These include *CHLG* (*Chlorophyll G*) (Fig. 6*a*, *b*), *HEME2* (*AT5G14220*), and *CH1* (*Chlorina 1*), which encode chlorophyll biosynthetic enzymes, as well as *ABCG25*, encoding ABA exporter that transports ABA across the plasma membrane (Tables S6 and S7).

Fig. 6. Overlapping TF BSs in target gene promoters revealed with PlantReg 1.1.

a – distal *CHLG* promoter with overlapping BSs for bZIP family TFs (bZIP3, bZIP53, bZIP68, and GBF3); *c* – distal and core *GBF3* promoters with overlapping BSs for bZIP family TFs (bZIP3, bZIP53, bZIP68, TGA4/5/9, and GBF3), TF color coding in (*c*) is the same as in (*a*); *e*, *g* – proximal *UGT72B1* and *BEH2* promoters, respectively, with overlapping BSs for BEH2 (red fill color) and BMY2 (blue fill color). *b*, *d*, *f*, *h* – transcriptional regulation of the *CHLG*, *GBF3*, *UGT72B1*, and *BEH2* genes, respectively. Underlined BSs lie on the antisense strand with regard to the gene body strand. Coordinates are given relative to the transcription start site. US – upregulated suppressor; DA – downregulated activator.

Since bZIP53 was described in the literature as a transcriptional activator (Alonso et al., 2009; Weltmeier et al., 2009), it is logical to assume that the suppression of the above-mentioned genes may be a consequence of competition among bZIP family TFs for common BSs in promoters, resulting in replacement of a strong activator by a weaker one. Indeed, the activity of the transactivation domains of these TFs was previously investigated and it was shown that bZIP53 is a transcriptional activator, but a much weaker one than representatives of the same family bZIP3, bZIP68, and GBF3 (Hummel et al., 2023).

It is noteworthy that a similar combination of transcriptional regulators competing for a common BS (bZIP3, bZIP68, and GBF3 as DA, bZIP53 as US) was identified in the promoters of dDEGs *ERF15*, *GBF3* (Fig. 6*c*, *d*), and *AT1G19000* encoding TFs from the TFRN-A (Tables S6 and S7). Thus, competition between the bZIP family TFs for a common BS is likely to be an essential mechanism of transcriptional repression in auxin response.

We also found a potential replacement of the bZIP3, bZIP68, and GBF3 activators with a weaker one, bZIP53, in the promoter of *GBF3*, which itself encodes a TF involved in its competitive regulation (Fig. 6*c*, *d*). A similar situation was observed for BEH2 (BES1/BZR1 HOMOLOG2) (DA) and BMY2 (BETA-AMYLASE 2, also known as BETA-AMYLASE 8/BAM8) (US), both belonging to the BZR1/BES1 family. These TFs regulate not only the expression of *DFB* and *UGT72B1*, the genes that control lignin levels, but also *BEH2* (Fig. 6*e–h*).

Theoretically, such feedback could act as a "trigger" for more intensive competitive suppression of common targets by a pair of activator TFs: an increase in the abundance of a weaker transcriptional activator leads to competitive suppression of the gene encoding a stronger transcriptional activator (which is a common target for both TFs including that stronger one), and thereby the inhibitory effect on other common targets will increase. Some DA-type TFs can compete with each other for a common BS prior to auxin treatment, when R subnetwork is active (Fig. 6*a–d*), providing additional regulatory flexibility to TFRN-A.

Competitive regulation of ABA signaling genes by TFRN-A and TFRN-S

Both auxin and salt stress modulate response to ABA: in the first case, it is attenuated, and in the second case, it is enhanced (Lavrekha et al., 2024; Omelyanchuk et al., 2024). Comparison of the regulatory links inferred based on data from different experiments enables a deeper understanding of transcription

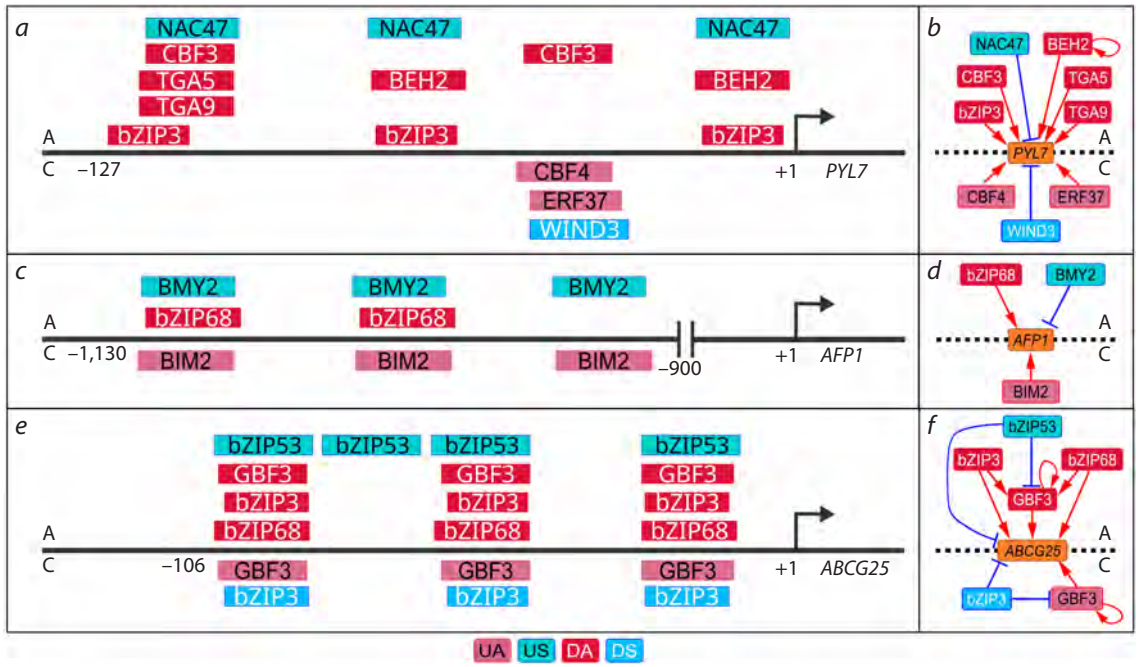


Fig. 7. Overlap of TF BSs in target promoters under auxin treatment and early salt stress, revealed using PlantReg 1.1.

a – proximal *PYL7* promoter; *c* – distal *AFP1* promoter; *e* – proximal *ABCG25* promoter. *b*, *d*, *f* – transcriptional regulation of the *PYL7*, *ATF1*, and *ABCG25* genes, respectively. For each panel, the details of regulation in response to auxin (A) and early salt stress (S) are located at the top and bottom, respectively. TF BSs are represented by rectangles according to the color coding of the regulation type: UA – upregulated activator; US – upregulated suppressor; DA – downregulated activator; DS – downregulated suppressor.

regulation. Using PlantReg 1.1, we found that three genes involved in ABA signaling (*PYL7*, *AFP1*, and *ABCG25*) are under the control of both TFRNs.

Downregulation of *PYL7* by auxin and its upregulation by salt stress is carried out by TF sets specific for each stimulus. These TF sets bind to different sites in the *PYL7* promoter (Fig. 7*a*, *b*). Apparently, auxin and salt stress utilize distinct molecular mechanisms for competitive modulation of *PYL7* expression. In contrast, both stimuli can engage the same set of competing activator TFs to regulate *AFP1* and *ABCG25*, but in different ways. *AFP1* gene expression is mediated by bZIP68 under normal conditions. After auxin treatment, bZIP68 is replaced by BMY2 (which is likely a weaker activator); under salt stress, on the contrary, bZIP68 is replaced by BIM2 (BES1-interacting Myc-like protein 2), which is a stronger activator according to (Hummel et al., 2023) (Fig. 7*c*, *d*). In the *ABCG25* promoter, auxin induces replacement of activator TFs from the bZIP family with a weak activator bZIP53 that results in a decrease in *ABCG25* transcripts (Fig. 7*e*, *f*). Salt stress modulates the relocation of a similar set of activators within the same set of BSs, but in this case, downregulation of *bZIP3* expression is accompanied by accumulation of *GBF3* transcripts.

Interestingly, a similar pattern was observed in the promoter of *GBF3* encoding a TF involved in both TFRNs. Under salt stress, which activates *GBF3*, GBF3 TF replaces bZIP3 at three BSs in the proximal *GBF3* promoter (–116; +1) (Fig. 8*a*, *b*), and at seven BSs in the distal promoter (–1,312; –701) (Fig. 8*c*–*f*), thereby apparently enhancing its self-activation. After auxin treatment, another redistribution of bZIP family TFs occurs at the same sites (Fig. 8). These results are consistent with an important role of competition for BSs between bZIP family TFs in modulation of gene expression (Schindler et al., 1992; Foster et al., 1994; Ko, Brandizzi, 2022). At the same time, auxin response recruits some specific mechanisms for *GBF3* regulation that are not involved in the response to salt stress. Thus, MYB3R1 can replace MYB70 and MYB73 at the common site after auxin treatment.

Discussion

In this work, we collected and systematized information on potential TF BSs in *A. thaliana* promoters to integrate it into the PlantReg 1.1 program. Along with the data on TF effector domain activity (Hummel et al., 2023), this allows to predict the cooperative and competitive interaction of TFs within the TFRNs in the transcriptional regulation of biological processes. Previously, we reconstructed two TFRNs that control the responses to salt stress and auxin in *A. thaliana* and showed that transcriptional repression plays an important role in both cases (Lavrekha et al., 2024; Omelyanchuk et al., 2024). However, according to the literature, the overwhelming majority of predicted suppressors in the TFRNs have activator-type effector domains (Hummel et al., 2023; Omelyanchuk et al., 2024). We used PlantReg 1.1 to identify the molecular mechanisms underlying the possible transformation of activator TFs into transcriptional repressors.

We found that more than one-third of the targets of TFs that were predicted as suppressors could be competitively regulated by a pair of TFs, one of which is a strong transcriptional activator and the other is a weak one. Thus, competitive

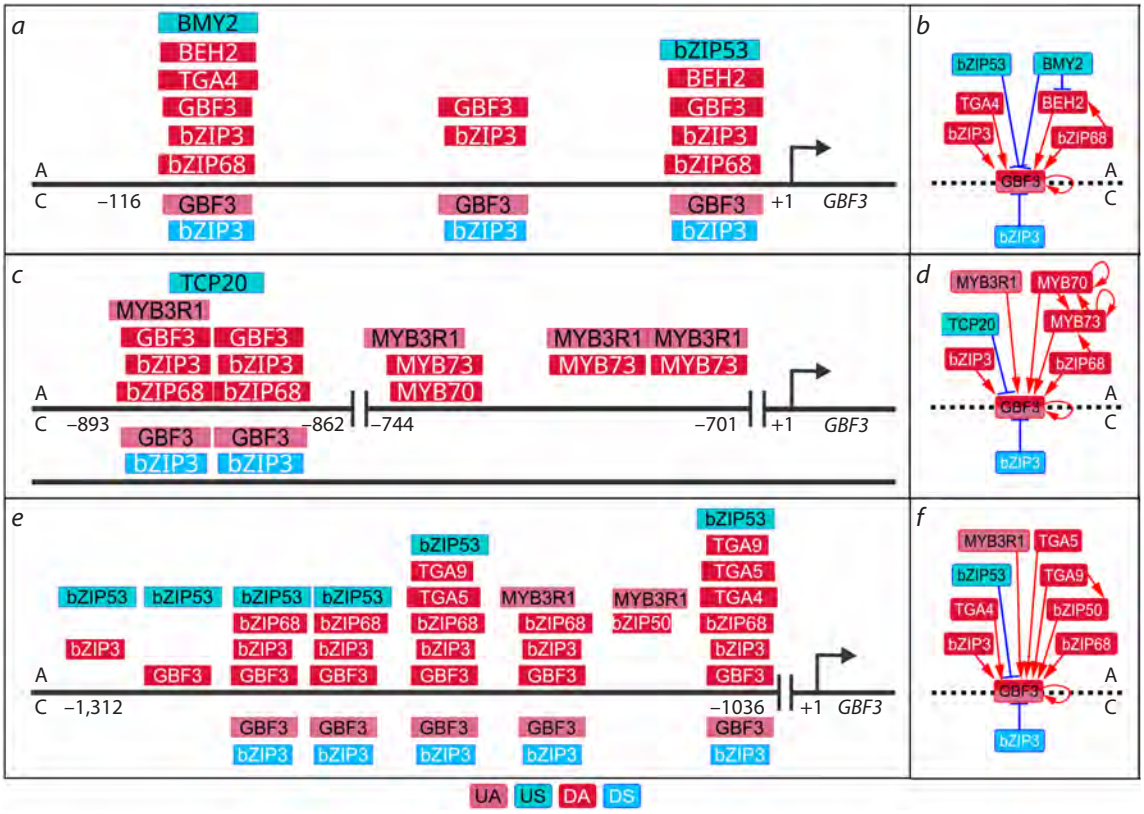


Fig. 8. Overlapping BS TFs in the *GBF3* promoter under auxin treatment and early salt stress, revealed using PlantReg 1.1.

a, *c*, *e* – proximal (–116; +1) and distal (–1,312; –701) *GBF3* promoters with overlapping BS TFs; *b*, *d*, *f* – diagrams of *GBF3* transcriptional regulation in the proximal and distal promoters. For each panel, the regulations in response to auxin (A) and early salt stress (S) are located at the top and bottom, respectively. TF BSs are represented by rectangles according to the color coding of the regulator type: UA – upregulated activator, US – upregulated suppressor, DA – downregulated activator, DS – downregulated suppressor.

regulation of gene expression is likely a universal mechanism allowing modulation of gene expression during responses to salt stress and auxin in *A. thaliana*.

Auxin is a key regulator of most plant processes involved in switching between developmental programs (Vanneste et al., 2025). The most standard concept of switching is replacement of a repressor with an activator, such as the substitution of the E2F TF repressor complex with E2F activators before the onset of the cell cycle in the promoters of many plant and animal genes (van den Heuvel, Dyson, 2008; Sánchez-Camargo et al., 2021), or, conversely, replacement of a transcriptional activator with a repressor (Berlow et al., 2017). However, the results obtained with PlantReg 1.1 indicate that in the auxin response, instead of the canonical activator–repressor switch, substitution of a strong activator with a weaker one can be actively used to suppress transcription.

At least some of the cases when a strong activator is substituted with a weaker one, predicted by PlantReg 1.1, are supported by previously published data. These include, for example, the replacement of three activators, bZIP3, bZIP68, and GBF3, by a weaker activator bZIP53 during auxin-induced suppression of chlorophyll biosynthesis genes *CHLG*, *HEME2*, and *CH1* (Hummel et al., 2023). Competition between bZIP family TFs for a common binding site and its influence on target gene expression has been previously described for many TFs from this family (Schindler et al., 1992; Foster et al., 1994; Ko, Brandizzi, 2022). It is also known that a number of bZIP family TFs redundantly regulate chlorophyll biosynthesis in a complex manner. In particular, chlorophyll biosynthesis is impaired in the *gbf1 gbf2 gbf3* triple mutant, demonstrating the important role of GBFs in this process (Sun T. et al., 2025). Overexpression of another family member, *bZIP1*, results in decreased chlorophyll levels, while the *bzip1 bzip53* double mutant demonstrates a less pronounced decrease in chlorophyll levels and attenuated *CHLG* expression compared to the wild type (p_{adj} = 0.03) (Wildenhain et al., 2025).

The plant-specific BZR1/BES1 TF family mediates transcriptional response to brassinosteroids (plant steroid hormones). In addition to BZR1 and BES1, this family also includes four other TFs, called BES1 homologues: BEH1, BEH2, BEH3, and BEH4 (Shi et al., 2022). Recently, the BZR1/BES1 family has been supplemented with two unusual TFs, BAM7 and BMY2, which are similar to β-amylases but also exhibit very weak homology to BES1 (Reinhold et al., 2011). These TFs lack amylase catalytic activity but possess BZR1-like domains that bind to the sequences recognized by TFs from this family. BMY2 is a transcriptional activator, while BAM7 regulates its activity.

It has been previously suggested that BMY2 controls the transcription of target genes by competing with the other BZR1/BES1 TFs for BSs (Reinhold et al., 2011). According

to the results obtained with PlantReg 1.1, this may take place in the promoters of some genes downregulated by auxin, including *UGT72B1* (*UDP-glucose-dependent glucosyltransferase 72 B1*), which encodes a monolignol-conjugating enzyme. In the *UGT72B1* promoter, BMY2 (which is likely a weaker activator) competes with BEH2 (Fig. 6*e*, *f*; Tables S6 and S7).

A more detailed analysis of the BEH2 and BMY2 targets predicted using PlantReg 1.1 provides several important details to auxin regulation of lignin biosynthesis. Auxin, by activating *BMY2*, inhibits *BEH2* self-activation replacing BEH2 with the less active BMY2 TF at their common BSs (Fig. 5*g*, *h*). This leads to a decrease in BEH2 levels in the nucleus, which in turn facilitates the replacement of this TF at its sites in the *UGT72B1* promoter with a weaker activator BMY2 (Fig. 5*e*, *f*) and, consequently, causes a decrease in the *UGT72B1* transcript level. Activation of *UGT72B1* by BMY2 is supported by an increase in *UGT72B1* transcript level upon BMY2 overexpression and downregulation of this gene in the *bmy2 bam7* double mutant (Reinhold et al., 2011).

Notably, auxin suppresses the transcription of most genes encoding lignin biosynthetic enzymes (Omelyanchuk et al., 2024), thereby reducing monolignol levels. At the same time, auxin downregulates *UGT72B1* expression and as a consequence inhibits monolignol conjugation, partially compensating for the decrease in monolignol levels. Interestingly, brassinosteroids also modulate lignin levels through BEH2. Brassinosteroids enhance lignin biosynthesis by activating most of the enzymes involved in this process (Percio et al., 2025). They simultaneously suppress BEH2 via both GSK3 (GLYCOGEN SYNTHASE KINASE 3)-like kinases and BES1 (Otani et al., 2022). Since BEH2 activates *UGT72B1*, which conjugates monolignols, brassinosteroids restrict the withdrawal of monolignols from lignin biosynthesis, thereby further increasing the lignin level.

The data obtained using PlantReg 1.1 allow formulating specific hypotheses for planning further experimental studies. It is worth emphasizing, however, that these predictions may contain false-positive results. For example, in the pair of TFs HB21 (DA) and HB40 (US) from TFRN-A, which bind the same sites in the promoter of the auxin-repressed gene *bZIP50*, HB40 is a more potent activator. This means that competition for BS with HB21 cannot explain the suppression of target gene expression with HB40 increase. It is possible that *HB21* and *HB40* are expressed in different tissues or at different developmental stages. To explain why HB40, which is an activator by nature, can function as a repressor, we need to explore how this TF recruits corepressors.

Conclusion

PlantReg 1.1 is designed to reveal regulatory relationships between TFRNs and genes that mediate the biological processes controlled by these networks. The updated version of the program includes functionality for precise localization of TF BSs in target promoters. Due to this, it becomes possible to analyze the mutual arrangement of TF BSs and, using data on the effector TF domains, to identify potential cooperative or competitive TF action in the promoter of a particular gene.

PlantReg 1.1 was successfully applied to reconstruct the transcriptional mechanisms regulating chlorophyll and lignin biosynthesis, ribosome biogenesis, and ABA response under auxin and salt stress. Analysis of the mutual arrangement of TF BSs revealed that the activity of a number of genes regulating these processes can be suppressed as a result of competition between a pair of activator TFs for a common BS, with a weaker activator replacing a stronger one. Some of the obtained results were supported by literature data.

Thus, the results obtained using PlantReg 1.1 allow formulating specific hypotheses for planning further experimental studies. It is worth emphasizing, however, that the predictions may contain false-positive results. Reducing their incidence is one possible direction for further development of the program.

References

Ahn J.H., Miller D., Winter V.J., Banfield M.J., Jeong H.L., So Y.Y., Henz S.R., Brady R.L., Weigel D. A divergent external loop confers antagonistic activity on floral regulators FT and TFL1. *EMBO J.* 2006;25(3):605-614. doi 10.1038/sj.emboj.7600950

Alonso R., Oñate-Sánehez L., Weltmeier F., Ehlert A., Diaz I., Dietrich K., Vicente-Carbajosa J., Dröge-Laser W. A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of *Arabidopsis* seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell.* 2009;21(6): 1747-1761. doi 10.1105/tpc.108.062968

Bailey T.L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 2021;37(18):2834-2840. doi 10.1093/bioinformatics/btab203

Berlow R.B., Dyson H.J., Wright P.E. Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature*. 2017; 543(7645):447-451. doi 10.1038/nature21705

Blanc-Mathieu R., Dumas R., Turchi L., Lucas J., Parcy F. Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.* 2024;29(1):40-51. doi 10.1016/j.tplants.2023.06.023

Boyle P., Després C. Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. *Plant Signal Behav.* 2010;5(6):629-634. doi 10.4161/psb.5.6.11570

Brackmann K., Qi J., Gebert M., Jouannet V., Schlamp T., Grünwald K., Wallner E.-S., Novikova D.D., Levitsky V.G., Agustí J., Sanchez P., Lohmann J.U., Greb T. Spatial specificity of auxin responses coordinates wood formation. *Nat Commun.* 2018;9(1):875. doi 10.1038/s41467-018-03256-2

Cheng C.Y., Krishnakumar V., Chan A.P., Thibaud-Nissen F., Schobel S., Town C.D. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 2017;89(4):789-804. doi 10.1111/tpj.13415

Chupreta S., Brevig H., Bai L., Merchant J.L., Iñiguez-Lluhí J.A. Sumoylation-dependent control of homotypic and heterotypic synergy by the Krüppel-type zinc finger protein ZBP-89. *J Biol Chem.* 2007;282(50):36155-36166. doi 10.1074/jbc.M708130200

De Rybel B., Audenaert D., Xuan W., Overvoorde P., Strader L.C., Kepinski S., Hoye R., Brisbois R., Parizot B., Vanneste S., Liu X. A role for the root cap in root branching revealed by the non-auxin probe naxillin. *Nat Chem Biol*. 2012;8(9):798-805. doi 10.1038/nchembio.1044

Deshpande D., Chhugani K., Chang Y., Karlsberg A., Loeffler C., Zhang J., Muszyńska A., … Eskin E., Zhao F., Mohammadi P., Łabaj P., Mangul S. RNA-seq data science: from raw data to effective interpretation. *Front Genet.* 2023;14:997383. doi 10.3389/fgene.2023.997383

Dhatterwal P., Sharma N., Prasad M. Decoding the functionality of plant transcription factors. *J Exp Bot.* 2024;75(16):4745-4759. doi 10.1093/jxb/erae231

Foster R., Izawa T., Chua N.H. Plant bZIP proteins gather at ACGT elements. *FASEB J.* 1994;8(2):192-200. doi 10.1096/fasebj.8.2.8119490

Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24

Hanna-Rose W., Hansen U. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* 1996;12(6):229-234. doi 10.1016/0168-9525(96)10022-6

Huang W., Quan M., Qi W., Xiao L., Fang Y., Zhou J., Ren T., Li P., Chen Y., El-Kassaby Y.A., Du F., Zhang D. Phylostratigraphic analysis revealed that ancient ohnologue *PtoWRKY53* innovated a vascular transcription regulatory network in *Populus*. *New Phytol*. 2025;248:2295-2315. doi 10.1111/nph.70403

Hummel N.F.C., Zhou A., Li B., Markel K., Ornelas I.J., Shih P.M. The trans-regulatory landscape of gene networks in plants. *Cell Syst.* 2023;14(6):501-511.e4. doi 10.1016/j.cels.2023.05.002

Ko D.K., Brandizzi F. Network-based approaches for understanding gene regulation and function in plants. *Plant J.* 2020;104(2):302-317. doi 10.1111/tpj.14940

Ko D.K., Brandizzi F. Transcriptional competition shapes proteotoxic ER stress resolution. *Nat Plants.* 2022;8(5):481-490. doi 10.1038/s41477-022-01150-w

Lamesch P., Berardini T.Z., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., … Nelson W.D., Ploetz L., Singh S., Wensel A., Huala E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 2012;40(D1):D1202-D1210. doi 10.1093/nar/gkr1090

Lavrekha V.V., Levitsky V.G., Tsukanov A.V., Bogomolov A.G., Grigorovich D.A., Omelyanchuk N., Ubogoeva E.V., Zemlyanskaya E.V., Mironova V. CisCross: a gene list enrichment analysis to predict upstream regulators in *Arabidopsis thaliana*. *Front Plant Sci*. 2022;13:942710. doi 10.3389/fpls.2022.942710

Lavrekha V.V., Omelyanchuk N.A., Bogomolov A.G., Zemlyanskaya E.V. PlantReg: the reconstruction of links between transcription factor regulatory networks and biological processes under their control. *Vavilov J Genet Breed*. 2024;28(8):950-959. doi 10.18699/vjgb-24-102

Leong R., He X., Beijen B.S., Sakai T., Goncalves J., Ding P. Unlocking gene regulatory networks for crop resilience and sustainable agriculture. *Nat Biotechnol.* 2025;43(8):1254-1265. doi 10.1038/s41587-025-02727-4

Marshall-Colón A., Kliebenstein D.J. Plant networks as traits and hypotheses: moving beyond description. *Trends Plant Sci.* 2019;24(9):840-852. doi 10.1016/j.tplants.2019.06.003

Martínez C., Espinosa-Ruíz A., de Lucas M., Bernardo-García S., Franco-Zorrilla J.M., Prat S. PIF 4-induced BR synthesis is critical to diurnal and thermomorphogenic growth. *EMBO J.* 2018;37(23):e99552. doi 10.15252/embj.201899552

Nagahage I.S.P., Sakamoto S., Nagano M., Ishikawa T., Kawai-Yamada M., Mitsuda N., Yamaguchi M. An NAC domain transcription factor ATAF2 acts as transcriptional activator or repressor dependent on promoter context. *Plant Biotechnol.* 2018;35(3):285-289. doi 10.5511/plantbiotechnology.18.0507a

O'Malley R.C., Huang S.C., Song L., Lewsey M.G., Bartlett A., Nery J.R., Galli M., Gallavotti A., Ecker J.R. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*. 2016;165(5):1280-1292. doi 10.1016/j.cell.2016.04.038

Omelyanchuk N.A., Lavrekha V.V., Bogomolov A.G., Dolgikh V.A., Sidorenko A.D., Zemlyanskaya E.V. Computational reconstruction of the transcription factor regulatory network induced by auxin in *Arabidopsis thaliana* L. *Plants*. 2024;13(14):1905. doi 10.3390/plants13141905

Otani Y., Kawanishi M., Kamimura M., Sasaki A., Nakamura Y., Nakamura T., Okamoto S. Behavior and possible function of *Arabidopsis* BES1/BZR1 homolog 2 in brassinosteroid signaling. *Plant Signal Behav.* 2022;17(1):2084277. doi 10.1080/15592324.2022.2084277

Park P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet.* 2009;10(10):669-680. doi 10.1038/nrg2641

Percio F., Rubio L., Amorim-Silva V., Botella M.A. Crucial roles of brassinosteroids in cell wall composition and structure across species: new insights and biotechnological applications. *Plant Cell Environ.* 2025;48(3):1751-1767. doi 10.1111/pce.15258

Raditsa V.V., Tsukanov A.V., Bogomolov A.G., Levitsky V.G. Genomic background sequences systematically outperform synthetic ones in de novo motif discovery for ChIP-seq data. *NAR Genomics Bioinform*. 2024;6(3):lqae090. doi 10.1093/nargab/lqae090

Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., … Lenhard B., Sandelin A., Wasserman W.W., Parcy F., Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059

Reinhold H., Soyk S., Simková K., Hostettler C., Marafino J., Mainiero S., Vaughan C.K., Monroe J.D., Zeeman S.C. β-amylase-like proteins function as transcription factors in *Arabidopsis*, controlling shoot growth and development. *Plant Cell.* 2011;23(4):1391-1403. doi 10.1105/tpc.110.081950

Ren G., Cui K., Zhang Z., Zhao K. Division of labor between IRF1 and IRF2 in regulating different stages of transcriptional activation in cellular antiviral activities. *Cell Biosci.* 2015;5(1):17. doi 10.1186/s13578-015-0007-0

Rybakov M.A., Omelyanchuk N.A., Zemlyanskaya E.V. Reconstruction of gene regulatory networks from single cell transcriptomic data. *Vavilov J Genet Breed*. 2024;28(8):974-981. doi 10.18699/vjgb-24-104

Sánchez-Camargo V.A., Romero-Rodríguez S., Vázquez-Ramos J.M. Non-canonical functions of the E2F/DP pathway with emphasis in plants. *Phyton.* 2021;90(2):307-330. doi 10.32604/phyton.2021.014967

Schindler U., Beckmann H., Cashmore A.R. TGA1 and G-box binding factors: two distinct classes of Arabidopsis leucine zipper proteins compete for the G-box-like element TGACGTGG. *Plant Cell.* 1992;4(10):1309-1319. doi 10.1105/tpc.4.10.1309

Selvaraj N., Budka J.A., Ferris M.W., Plotnik J.P., Hollenhorst P.C. Extracellular signal-regulated kinase signaling regulates the opposing roles of JUN family transcription factors at ETS/AP-1 sites and in cell migration. *Mol Cell Biol.* 2015;35(1):88-100. doi 10.1128/mcb.00982-14

Shi H., Li X., Lv M., Li J. BES1/BZR1 family transcription factors regulate plant development via brassinosteroid-dependent and independent pathways. *Int J Mol Sci.* 2022;23(17):10149. doi 10.3390/ijms231710149

Su K., Katebi A., Kohar V., Clauss B., Gordin D., Qin Z.S., Karuturi R.K.M., Li S., Lu M. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. *Genome Biol.* 2022;23(1):270. doi 10.1186/s13059-022-02835-3

Sun T., Hazra A., Lui A., Zeng S., Wang X., Rao S., Owens L.A., Fei Z., Zhao Y., Mazourek M., Giovannoni J.G., Li L. GLKs directly regulate carotenoid biosynthesis via interacting with GBFs in plants. *New Phytol.* 2025;246(2):645-665. doi 10.1111/nph.20457

Sun Y., Li J., Huang J., Li S., Li Y., Lu B., Deng X. Architecture of genome-wide transcriptional regulatory network reveals dynamic functions and evolutionary trajectories in *Pseudomonas syringae*. *eLife.* 2025;13:RP96172. doi 10.7554/eLife.96172.3

Tamura T., Sakata T., Igarashi H., Okumura K. Transcription factor HUB1 represses SP1-mediated gene expression through the CACCC box of HTLV-I U5RE but not the GC box. *J Health Sci*. 2004;50(4):417-422. doi 10.1248/jhs.50.417

Tremblay B.J.M. universalmotif: an R package for biological motif analysis. *J Open Source Software*. 2024;9(100):7012. doi 10.21105/joss.07012

van den Heuvel S., Dyson N.J. Conserved functions of the pRB and E2F families. *Nat Rev Mol Cell Biol.* 2008;9(9):713-724. doi 10.1038/nrm2469

Vanneste S., Pei Y., Friml J. Mechanisms of auxin action in plant growth and development. *Nat Rev Mol Cell Biol*. 2025;26(9):648-666. doi 10.1038/s41580-025-00851-2

Veerabagu M., Kirchler T., Elgass K., Stadelhofer B., Stahl M., Harter K., Mira-Rodado V., Chaban C. The interaction of the *Arabidopsis* response regulator ARR18 with bZIP63 mediates the regulation of *PROLINE DEHYDROGENASE* expression. *Mol Plant*. 2014;7(10):1560-1577. doi 10.1093/mp/ssu074

Wang L., Ko E.E., Tran J., Qiao H. TREE1-EIN3-mediated transcriptional repression inhibits shoot growth in response to ethylene. *Proc Natl Acad Sci USA*. 2020;117(46):29178-29189. doi 10.1073/pnas.2018735117

Weirauch M.T., Yang A., Albu M., Cote A.G., Montenegro-Montero A., Drewe P., Najafabadi H.S., … Bouget F.Y., Ratsch G., Larrondo L.F., Ecker J.R., Hughes T.R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(6):1431-1443. doi 10.1016/j.cell.2014.08.009

Weltmeier F., Rahmani F., Ehlert A., Dietrich K., Schütze K., Wang X., Chaban C., Hanson J., Teige M., Harter K., Vicente-Carbajosa J., Smeekens S., Dröge-Laser W. Expression patterns within the *Arabidopsis* C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development. *Plant Mol Biol*. 2009;69(1):107-119. doi 10.1007/s11103-008-9410-9

Wildenhain T., Smaczniak C., Marsell A., Draken J., Maag D., Kreisz P., Krischke M., Müller M.J., Kaufmann K., Weiste C., Dröge-Laser W. A subset of group S_1 bZIP transcription factors controls resource management during starvation and recovery in Arabidopsis. *Plant Cell*. 2025;37(7):koaf149. doi 10.1093/plcell/koaf149

Wu T., Goh H., Azodi C.B., Krishnamoorthi S., Liu M.J., Urano D. Evolutionarily conserved hierarchical gene regulatory networks for plant salt stress response. *Nat Plants*. 2021;7(6):787-799. doi 10.1038/s41477-021-00929-7

Zhang X., Li L., Fourie J., Davie J.R., Guarcello V., Diasio R.B. The role of Sp1 and Sp3 in the constitutive *DPYD* gene expression. *Biochim Biophys Acta*. 2006;1759(5):247-256. doi 10.1016/j.bbaexp.2006.05.001

Zemlyanskaya E.V., Dolgikh V.A., Levitsky V.G., Mironova V. Transcriptional regulation in plants: using omics data to crack the cis-regulatory code. *Curr Opin Plant Biol*. 2021;63:102058. doi 10.1016/j.pbi.2021.102058

Conflict of interest. The authors declare no conflict of interest.
Received August 29, 2025. Revised September 23, 2025. Accepted September 23, 2025.

COMPUTATIONAL GENOMICS
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):952-962

doi 10.18699/vjgb-25-101

FlyDEGdb knowledge base on differentially expressed genes of *Drosophila melanogaster*, a model object in biomedicine

O.A. Podkolodnaya [1$], M.A. Deryuzhenko [1$], N.N. Tverdokhleb [1], K.A. Zolotareva [1], Yu.V. Makovka [1], N.L. Podkolodny [1, 2], V.V. Suslov [1], I.V. Chadaeva [1], L.A. Fedoseeva [1], A.A. Seryapina [1], D.Yu. Oshchepkov [1], A.G. Bogomolov [1], E.Yu. Kondratyuk [1, 3], O.E. Redina [1], A.L. Markel [1, 4], N.E. Gruntenko [1], M.P. Ponomarenko [1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, Russia
[4] Novosibirsk State University, Novosibirsk, Russia

pon@bionet.nsc.ru

Abstract. Since the work of Nobel Prize winner Thomas Morgan in 1909, the fruit fly *Drosophila melanogaster* has been one of the most popular model animals in genetics. Research using this fly was honored with the Nobel Prize many times: in 1946 (Muller, X-ray mutagenesis), in 1995 (Lewis, Nüsslein-Volhard, Wieschaus, genetic control of embryogenesis), in 2004 (Axel and Buck, the olfactory system), in 2011 (Steinman, dendritic cells in adaptive immunity; Beutler and Hoffman, activation of innate immunity), and in 2017 (Hall, Rosbash and Young, the molecular mechanism of the circadian rhythm). The prominent role of Drosophila in genetics is due to its key features: short life cycle, frequent generational turnover, ease of maintenance, high fertility, small size, transparent embryos, simple larval structure, the possibility to observe visually chromosomal rearrangements due to the presence of polytene chromosomes, and accessibility to molecular genetic manipulation. Furthermore, the highly conserved nature of several signaling pathways and gene networks in Drosophila and their similarity to those of mammals and humans, taken together with the development of high-throughput genomic sequencing, motivated the use of *D. melanogaster* as a model organism in biomedical fields of inquiry: pharmacology, toxicology, cardiology, oncology, immunology, gerontology, and radiobiology. These studies add to the understanding of the genetic and epigenetic basis of the pathogenesis of human diseases. This paper describes our curated knowledge base, FlyDEGdb (https://www.sysbio.ru/FlyDEGdb), which stores information on differentially expressed genes (DEGs) in Drosophila. This information was extracted from 50 scientific articles containing experimental data on changes in the expression of 20,058 genes (80 %) out of the 25,079 Drosophila genes stored in the NCBI Gene database. The changes were induced by 52 stress factors, including heat and cold exposure, dehydration, heavy metals, radiation, starvation, household chemicals, drugs, fertilizers, insecticides, pesticides, herbicides, and other toxicants. The FlyDEGdb knowledge base is illustrated using the example of the *dysf* (*dysfusion*) Drosophila gene, which had been identified as a DEG under cold shock and in toxicity tests of the herbicide paraquat, the solvent toluene, the drug menadione, and the food additive E923. FlyDEGdb stores information on changes in the expression of the *dysf* gene and its homologues: (a) the *Clk*, *cyc*, and *per* genes in Drosophila, and (b) the *NPAS4*, *CLOCK*, *BMAL1*, *PER1*, and *PER2* genes in humans. These data are supplemented with information on the biological processes in which these genes are involved: oocyte maturation (oogenesis), regulation of stress response and circadian rhythm, carcinogenesis, aging, etc. Therefore, FlyDEGdb, containing information on the widely used model organism, Drosophila, can be helpful for researchers working in the molecular biology and genetics of humans and animals, physiology, translational medicine, pharmacology, dietetics, agricultural chemistry, radiobiology, toxicology, and bioinformatics.

Key words: human; disease; biomedicine; model animal; fruit fly *Drosophila melanogaster*; differentially expressed genes (DEGs); RNA-Seq; qPCR; microarray; knowledge base

For citation: Podkolodnaya O.A., Deryuzhenko M.A., Tverdokhleb N.N., Zolotareva K.A., Makovka Yu.V., Podkolodny N.L., Suslov V.V., Chadaeva I.V., Fedoseeva L.A., Seryapina A.A., Oshchepkov D.Yu., Bogomolov A.G., Kondratyuk E.Yu., Redina O.E., Markel A.L., Gruntenko N.E., Ponomarenko M.P. FlyDEGdb knowledge base on differentially expressed genes of *Drosophila melanogaster*, a model object in biomedicine. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):952-962. doi 10.18699/vjgb-25-101

Funding. This work was supported by budget project FWNR-2022-0019.

Acknowledgements. We are grateful to the Bioinformatics Shared Access Center for providing computational resources.

$ Equal contribution.

База знаний FlyDEGdb по дифференциально экспрессирующимся генам *Drosophila melanogaster* – модельного объекта биомедицины

О.А. Подколодная [1$], М.А. Дерюженко [1$], Н.Н. Твердохлеб[1], К.А. Золотарева [1], Ю.В. Маковка [1], Н.Л. Подколодный [1, 2], В.В. Суслов [1], И.В. Чадаева [1], Л.А. Федосеева [1], А.А. Серяпина [1], Д.Ю. Ощепков [1], А.Г. Богомолов [1], Е.Ю. Кондратюк [1, 3], О.Е. Редина [1], А.Л. Маркель [1, 4], Н.Е. Грунтенко [1], М.П. Пономаренко [1] ✉

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Сибирский федеральный научный центр агробиотехнологий Российской академии наук, р. п. Краснообск, Новосибирская область, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ pon@bionet.nsc.ru

Аннотация. С 1909 г. благодаря исследованиям нобелевского лауреата Моргана дрозофила *Drosophila melanogaster* стала одним из самых популярных модельных животных в генетике. Фундаментальные исследования с дрозофилой в качестве модельного объекта неоднократно были отмечены Нобелевской премией: в 1946 г. (Мёллер, мутагенез при рентгеновском излучении), в 1995 (Льюис, Нюсслейн-Фольхард, Вишаус, генетический контроль эмбриогенеза), в 2004 (Эксел и Бак, обонятельная система), в 2011 (Стайнман, дендритные клетки в адаптивном иммунитете; Бётлер и Офман, активация врожденного иммунитета) и в 2017 г. (Холл, Росбаш и Янг, молекулярный механизм циркадного ритма). Столь яркая роль дрозофилы в генетике обусловлена рядом ее ключевых признаков: кратким жизненным циклом, частой сменой поколений, легкостью в содержании, высокой плодовитостью, малым размером, прозрачностью эмбриона, простым строением личинки, возможностью визуальных наблюдений хромосомных перестроек за счет наличия политенных хромосом, доступностью для молекулярно-генетических манипуляций. Кроме того, благодаря высокой консервативности ряда сигнальных путей и генных сетей дрозофилы и их сходству с таковыми у млекопитающих и человека в совокупности с техническим развитием геномного секвенирования стало возможно использование *D. melanogaster* как модельного объекта в биомедицинских исследованиях в области фармакологии, токсикологии, кардиологии, онкологии, иммунологии, геронтологии и радиобиологии для поиска генетической и эпигенетической основ патогенеза болезней человека. В настоящей статье описана созданная нами курируемая база знаний FlyDEGdb (https://www.sysbio.ru/FlyDEGdb), в которой представлена информация о дифференциально экспрессирующихся генах (ДЭГ) дрозофилы, экстрагированная из 50 научных статей с экспериментальными данными об изменении экспрессии 20 058 генов (80 %) из числа всех 25 079 генов дрозофилы согласно базе данных NCBI Gene под действием 52 стрессовых факторов, включая высокую и низкую температуры, обезвоживание, тяжелые металлы, радиацию, голод, яды, бытовую химию, лекарства, удобрения, инсектициды, пестициды и гербициды. Содержание базы знаний FlyDEGdb проиллюстрировано на примере гена *dysf* (*dysfusion*) дрозофилы, который был идентифицирован в качестве ДЭГ при множестве стрессовых воздействий: холодовом шоке и в испытаниях на токсичность гербицида параквата, растворителя толуола, лекарственного препарата менадиона, пищевой добавки E923. В FlyDEGdb представлена информация об изменениях экспрессии гена *dysf* и его гомологов *Clk*, *cyc*, *per* у дрозофилы и генов *NPAS4*, *CLOCK*, *BMAL1*, *PER1* и *PER2* человека, а также информация о биологических процессах, в которые вовлечены эти гены: созревание ооцитов (оогенез), регуляция стресс-ответа и циркадного ритма, канцерогенез, старение и др. Поэтому FlyDEGdb, содержащая информацию о таком модельном организме, как дрозофила, может быть полезна для исследователей, работающих в области молекулярной биологии и генетики человека и животных, физиологии, трансляционной медицины, фармакологии, диетологии, агрохимии, радиобиологии, токсикологии и биоинформатики.

Ключевые слова: человек; заболевание; биомедицина; модельное животное; дрозофила; *Drosophila melanogaster*; дифференциально экспрессирующиеся гены (ДЭГ); RNA-seq; qPCR; микрочип; база знаний

Introduction

Animal models are broadly employed in biomedical studies of the physiological, genetic, and epigenetic mechanisms regulating evolutionarily fixed phenotypic human traits in health and disease, as well as in response to external and internal stress factors (Mukherjee et al., 2022). Their use is based on strict criteria of the correspondence between the human phenotypic features under study and their counterparts in model animals (Gryksa et al., 2023). Over a century ago, Thomas Hunt Morgan (1910), Professor of Experimental Zoology in the Columbia University, laid the foundation of a series of discoveries in heredity in a then new biological object, *Drosophila melanogaster*. His results were honored with the Nobel Prize “For his discoveries concerning the role played by the chromosome in heredity” in 1933. Later genetic studies using *Drosophila* were honored with the Nobel Prize five times more. In 1946, it was awarded to Hermann Muller “For the discovery of the production of mutations by means of X-ray irradiation”; in 1995, to Edward Lewis, Christiane Nüsslein-Volhard, and Eric Wieschaus “For their discoveries concern-

ing the genetic control of early embryonic development"; in 2004, to Richard Axel and Linda Buck "For their discoveries of odorant receptors and the organization of the olfactory system"; in 2011, to Ralph Steinman "For his discovery of the dendritic cell and its role in adaptive immunity" together with Jules Hoffman and Bruce Beutler "For their discoveries concerning the activation of innate immunity"; and in 2017, to Jeffrey Hall, Michael Rosbash, and Michael Young "For their discoveries of molecular mechanisms controlling the circadian rhythm" (Lakhotia, 2025).

This great significance of *Drosophila* for research is determined by the low cost of their maintenance, high fertility, frequent generation turnover, small size, optical transparency of embryos, simple larva structure, short life cycle, availability of numerous natural strains adapted to various ecoclimatic conditions (Telonis-Scott et al., 2013; Chen et al., 2015; von Heckel et al., 2016; Mikucki et al., 2024), relatively small genome, and ease of molecular genetic manipulations. It is of special importance that many signaling pathways and gene networks of *Drosophila* are similar to those of the human (Yu et al., 2022). Owing to this fact, many results in translational medicine, pharmacology, toxicology, immunology, gerontology, etc. obtained with *Drosophila* can be transferred to humans (De Gregorio et al., 2001; Chatterjee, Perrimon, 2021; Wu K. et al., 2021; Ali et al., 2022; Rand et al., 2023).

Within this line of inquiry, scientists of the Institute of Cytology and Genetics (ICG) of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, have investigated features of stress response in rats (Markel, 1985; Oshchepkov et al., 2024) and mice (Chadaeva et al., 2019; Avgustinovich et al., 2025) for over 40 years. The results, reported in many publications, present valuable data on changes in gene expression induced by various experimental procedures. Huge volumes of genome-wide data (Big Data) on DEGs in rats and mice have been obtained and documented in our knowledge bases RatDEGdb (Chadaeva et al., 2023) and MiceDEGdb (Podkolodnaya et al., 2024), respectively.

D. melanogaster is another model species, in which experiments on stress in animals have been conducted at ICG for over 25 years (Gruntenko et al., 1999; 2023). The effort on developing the FlyDEGdb knowledge base, which stores information on *Drosophila* DEGs, is the continuation of our works in biomedical knowledge bases RatDEGdb and MiceDEGdb. The pilot version of FlyDEGdb v.0.1 is freely available at https://www.sysbio.ru/FlyDEGdb. It stores experimental data on the expression of 80 % of *Drosophila* genes: 20,058 of the 25,079 annotated in NCBI Gene (Brown et al., 2015). The information presented in FlyDEGd was extracted from 50 papers reporting experimental data on the action of 52 stress factors on 31 *D. melanogaster* strains. The factors included heat and cold, dehydration, heavy metals, radiation, starvation, household chemicals, drugs, fertilizers, insecticides, pesticides, herbicides, and other toxicants. The informational content of FlyDEGdb v0.01 is illustrated by the *Drosophila dysf* (*dysfusion*) gene, which was identified as a DEG in cold shock and in tests of the herbicide paraquat, solvent toluene, drug menadione, and food additive E923. FlyDEGdb presents data on changes in the expression of *dysf* itself and its homologs: *Clk*, *cyc*, and *per* in *Drosophila* and *NPAS4*, *CLOCK*, *BMAL1*, *PER1*, and *PER2* in the human. In addition, FlyDEGdb provides information on the biologic processes involving these genes: oogenesis, regulation of stress response and circadian rhythms, carcinogenesis, aging, etc.

We also compare data on stress-induced *Drosophila* DEGs presented in FlyDEGdb with data on changes in the expression of DEGs of the hypothalamus of rat strains WAG and ISIAH in response to restriction stress, reported by D.Y. Oshchepkov et al. (2024) and presented in RatDEGdb. The responses of rats and *Drosophila* to stresses reveal a common molecular event: reduction in the expression of large gene groups involved in the formation of the plasma membrane. The FlyDEGdb knowledge base, storing information on the model species *Drosophila*, can be a useful tool for students of the molecular biology and genetics of the human and animals, physiology, translational medicine, pharmacology, nutrition science, agricultural chemistry, radiobiology, toxicology, and bioinformatics.

Materials and methods

Stress-inducible *Drosophila* DEGs. Experimentally detected *Drosophila* DEGs were sought in the PubMed database (Lu, 2011) with queries composed from various combinations of key words "Drosophila melanogaster", "differentially expressed gene", "stress response", "drying", "heat shock", "radiation", "cold shock", "oxidative stress", "continuous lighting", "toxin", "diet", "heavy metal", "drug", "herbicide", "pesticide", "insecticide", "RNA-seq", "microarray", and "qPCR".

Only DEGs with reported $\log_2(\text{DEG}) = \log_2([\text{DEG expression in } \textit{Drosophila} \text{ under a particular stress factor}] / [\text{normal DEG expression}])$ values and P_{ADJ} estimates of statistical significance with correction for multiple comparisons for the stress-induced expression of the DEG were added to FlyDEGdb. In addition, we eliminated those in which the $\log_2(\text{DEG})$ values ranged from –0.46 to 0.46. This range corresponds to statistically insignificant ($p \geq 0.05$, Fisher's Z-test) differences in DEG expression before and after the exposure to stress with ±5 % accuracy of expression measurements.

FlyDEGdb knowledge base. Figure 1 illustrates the informational structure of the FlyDEGdb knowledge base. It includes five relational tables. The first of them, named "FlyDEGs" (Fig. 1A), stores experimental data on a particular *Drosophila* DEG, which is assigned a unique number (field "FlyDEGid"). Field "FlyStrain" of the table indicates the *Drosophila* strain in which the DEG has been found in experiments. Field "FlyBioSample" indicates the tissue sample studied in the experiment. Field "PhenomenonFlyModel" indicates the corresponding stress factor. Fields "FlyModel Subject" and "FlyNormalSubject" indicate the model and control individuals, respectively, used in the experiment. The experiment type, "RNA-seq", "Microarray", or "RT-qPCR", is shown in field "ExperimentType". Field "FlyGeneSymbol" contains the identifier of the *Drosophila* DEG according to the NCBI Gene database (Brown et al., 2015). Fields "Log2(Model/Norm)" and "Padj" contain the quantity of the stress-induced change in DEG expression as compared to the

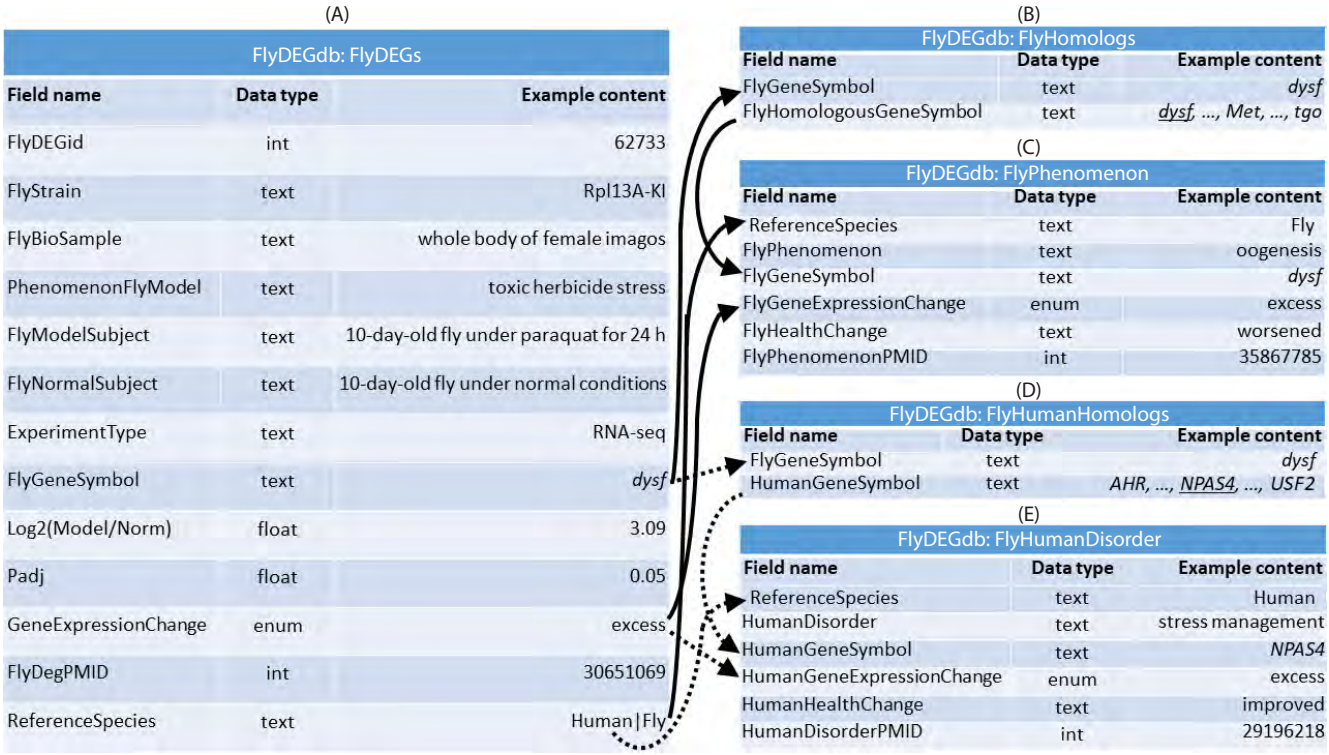


Fig. 1. The informational structure of the FlyDEGdb knowledge base on differentially expressed genes (DEGs) of *Drosophila melanogaster*. Relational tables: (A) FlyDEGs – experimental data on DEGs in Drosophila tissue samples in response to a stress factor relative to the norm according to the paper cited; (B) FlyHomologs – lists of *Drosophila* genes homologous to particular *Drosophila* genes according to the FlyBase database (Ozturk-Colak et al., 2024); (C) FlyPhenomenon – phenotypic traits associated with deviations in the expression of *Drosophila* genes relative to the norm according to the paper cited; (D) FlyHumanHomologs – human genes homologous to a particular *Drosophila* gene according to FlyBase (Ozturk-Colak et al., 2024); (E) HumanDisorder – human diseases associated with deviations in particular human genes relative to the norm according to the paper cited.

Names of relational tables and their fields were chosen following the guideline on the construction of friendly interfaces (Wade, 1984). Data types: int – integer number; float – real number; enum – binary indicator; text – character string; PMID – identifier of the referred paper in PubMed (Lu, 2011). Arrows (→) – relational links pointing to the annotation of experimental data on *Drosophila* DEGs (relational table FlyDEGs) on the one side and, on the other side, data on ipsidirectional changes in the expression of homologous *Drosophila* (solid lines) or human (dotted lines) DEGs indicated in the FlyHomologs and FlyHumanHomologs tables, obtained in independent experiments referred to in relational tables FlyPhenomenon and HumanDisorder.

norm and its significance level with correction for multiple comparisons, respectively, as they are reported. The source is indicated in field "FlyDegPMID" as its identifier in PubMed (Lu, 2011).

Finally, field "ReferenceSpecies" indicates the reference biologic species ("Fly" for *Drosophila* or "Human" for the human in the pilot version FlyDEGdb v0.1), the experimental data on which are used in the annotation of a particular DEG. Absence of such annotation is indicated as "ND".

Here we apply the term "annotation" to the supplementation of experimental data on stress-induced changes in the expression of a particular *Drosophila* DEG reported in a particular paper with experimental data from independent sources on phenotypic manifestations of ipsidirectional changes in the expression of homologous human and *Drosophila* genes. Supplementary Table S1[1] provides details of the annotation procedure.

To conclude the description of the informational structure of FlyDEGdb (Fig. 1), we indicate the data types used: int, integer number; float, real number; enum, binary indicator; text, character string.

[1] Supplementary Tables S1–S3 and Figure S1 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Podkol_Engl_29_7.pdf

The relational tables FlyDEGs, FlyHomologs, FlyPhenomenon, FlyHumanHomologs, and HumanDisorder were integrated to the FlyDEGdb knowledge base (https://www.sysbio.ru/FlyDEGdb) by using the MySQL-compatible database management studio MariaDB 10.2.12 (MariaDB Corp AB, Finland).

Statistical methods. The statistical analysis of *Drosophila* DEGs was conducted with Past v.4.04 application (Hammer et al., 2001) and the STATISTICA package (Statsoft™, United States).

Results and discussion

FlyDEGdb knowledge base

We sought papers on *Drosophila* DEGs in PubMed (Lu, 2011) with keywords listed in section "Materials and methods" to populate FlyDEGdb. We found 51 articles describing 287 experiments on 31 *Drosophila* strains originating from various geographical areas and their transgenic modifications. The articles described over 190,000 stress-inducible *Drosophila* DEGs. The results of the search are shown in Tables S1–S3. The articles cover a wide range of *Drosophila* studies concerning age-related human diseases, *Drosophila* tests of drugs,

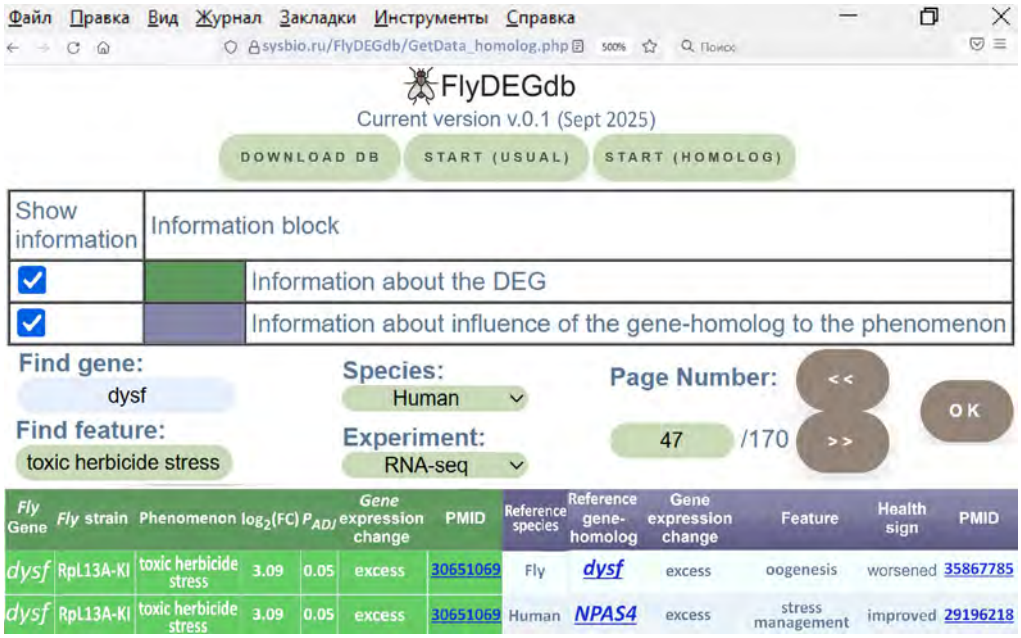


Fig. 2. The interface of the FlyDEGdb knowledge base on *D. melanogaster* DEGs supports the real-time dialogue for user access to the informational content.

Interface commands: DOWNLOAD DB – download the entire body of information of the current version FlyDEGdb v0.1 as a text file in an Excel-compatible format; START (HOMOLOG) – access to *Drosophila* DEGs annotated with the use of independent experimental data on the phenotypic manifestation of ipsidirectional expression changes relative to normal values in homologous genes in reference biologic species: *Drosophila* and the human; START (USUAL) – access to *Drosophila* DEGs omitting annotation. Left half of the table with information on *Drosophila* DEG (green background): experimental data on the *Drosophila* DEG considered; right half (lilac background): annotation of the DEG on the grounds of independent data on the phenotypic manifestation of ipsidirectional expression changes in homologous genes in reference biologic species: *Drosophila* and the human.

and tests for toxicity of household chemicals, fertilizers, insecticides, pesticides, herbicides, etc.

Figure 2 illustrates user access to the information stored in the pilot FlyDEGdb version.

Three buttons at the top of the FlyDEGdb interface provide access to the information:

- "DOWNLOAD DB" allows downloading all information from the current version FlyDEGdb v0.1 as a text file in an Excel-compatible format.
- "START (USUAL)" provides access to experimental data on stress-inducible *Drosophila* DEGs described in the main relational Table "FlyDEGs" (Fig. 1A).
- "START (HOMOLOG)" provides access to the annotations of *Drosophila* DEGs as described in section "Materials and methods".

Below there are interface fields for choosing the needed type of information: experimental data on *Drosophila* DEGs and/or annotation of *Drosophila* DEGs. The "Page Number" field allows alphabetical navigation over all DEGs stored in the knowledge base.

The bottom part of the interface outputs tabulated information on DEGs obtained by the user according to the specified query. Its description is provided in section "Materials and methods". Their storage in FlyDEGdb is shown in Figure 1.

Table 1 provides a detailed description of the *Drosophila dysf* DEG in response to various stress factors, as well as information on homologous DEGs in *Drosophila* and the human. Seven columns on the left contain experimental data on *Drosophila dysf* in response to the toxic effect of the herbicide paraquat, which increases the expression, and toluene, which decreases it. The expression changes are characterized by $\log_2$(DEG) values and significance levels P_{ADJ}. Column PMID indicates information sources.

Six columns on the right in Table 1 contain the results of annotation of the *Drosophila dysf* gene compared to the homologous *Clk* gene of the same species and to homologous human genes *NPAS4* and *CLOCK* on the base of four independent PMID papers. It is apparent that (a) the *dysf* upregulation (excess) is associated with *Drosophila* oogenesis impairments; (b) the downregulation (deficit) of the *Clk* gene, homologous to *Drosophila dysf*, disrupts the circadian rhythm; (c) the upregulation of the human *NPAS4* gene, homologous to *Drosophila dysf*, improves the efficiency of the stress response; (d) the downregulation of the human *CLOCK* gene, homologous to *Drosophila dysf*, disrupts the circadian rhythm. Similar examples are shown in rows 5–11.

Comparison of stress-induced homologous rat and *Drosophila* genes on the grounds of information from FlyDEGdb and RatDEGdb

Table 2 presents information on DEGs detected in a restriction stress experiment in the hypothalamus or WAG and ISIAH rats (Oshchepkov et al., 2024) in comparison with homologous *Drosophila* DEGs described in FlyDEGdb.

Table 1. Examples of the presentation of experimental data on *Drosophila* DEGs in the FlyDEGdb knowledge base and their annotations related to data on the phenotypic manifestations of ipsidirectional expression changes of homologous DEGs genes in *Drosophila* and the human

Experimental data on *Drosophila* DEGS available in FlyDEGdb								Annotation of the DEGS related to homologs in *Drosophila* and the human					
No.	*Drosophila* DEG (Fly DEG)	Fly strain	Stress factor (Phenomenon)	log_2(DEG)	P_{ADJ}	Gene expression change	PMID	Reference species	Reference gene-homolog	Gene expression change	Phenotypic trait/ associated biologic process/disease (Feature)	Effect on the trait (Health sign)	PMID
1	*dysf*	RpL13A-KI	herbicide paraquat	3.09	0.05	excess	30651069	*Drosophila*	*dysf*	excess	fertility	impairment	35867785
2	*dysf*	RpL13A-KI	herbicide paraquat	3.09	0.05	excess	30651069	human	*NPAS4*	excess	stress response	improvement	29196218
3	*dysf*	Canton-S	solvent toluene	–1.22	0.005	deficit	33484011	*Drosophila*	*Clk*	deficit	circadian rhythm	impairment	36809369
4	*dysf*	Canton-S	solvent toluene	–1.22	0.005	deficit	33484011	human	*CLOCK*	deficit	circadian rhythm	impairment	15950223
5	*dysf*	WT mix #1	drug menadione	0.57	0.001	excess	34747443	*Drosophila*	*Clk*	excess	senescence	improvement	35100266
6	*dysf*	WT mix #1	drug menadione	0.57	0.001	excess	34747443	human	*BMAL1*	excess	carcinogenesis	complication	22510946
7	*dysf*	WT mix #2	cold shock	0.59	0.001	excess	27502401	*Drosophila*	*cyc*	excess	progeny viability	improvement	33482172
8	*dysf*	WT mix #2	cold shock	0.59	0.001	excess	27502401	human	*PER1*	excess	apoptosis	improvement	16678109
9	*dysf*	white-KO	food additive E923	0.95	0.05	excess	38249074	*Drosophila*	*per*	excess	survival	improvement	25165772
10	*dysf*	white-KO	food additive E923	0.95	0.05	excess	38249074	human	*PER2*	excess	Q fever	predisposition	22984121
11	*dysf*	w1118	food additive E923	0.98	0.05	excess	38249074	human	*PER1*	excess	carcinogenesis	aggravation	24144995

Note. *Drosophila* strains: RpL13A-KI, transgenic strain with upregulation of the *RpL13A* gene, encoding ribosomal protein L13A of the 40S ribosome subunit; WT mix #1: mix of accessions collected in Germany, on Cyprus, in Malaysia, and Zambia; WT mix #2: mix of accessions collected in Sweden, Denmark, Zambia, and Zimbabwe; white-KO: transgenic *Drosophila* strain with double knockout of the *white* (*w*) gene; food additive E923: ammonium persulfate (APS). PMID: 15950223 – (Oishi et al., 2005); 16678109 – (Gery et al., 2006); 22510946 – (Elshazley et al., 2012); 22984121 – (Mehraj et al., 2012); 24144995 – (Wang et al., 2013); 25165772 – (Goda et al., 2014); 27502401 – (von Heckel et al., 2016); 29196218 – (Drouet et al., 2018); 30651069 – (Huang et al., 2019); 33482172 – (Fyie et al., 2021); 33484011 – (de Oliveira et al., 2021); 34747443 – (Ramnarine et al., 2022); 35100266 – (Jauregui-Lozano et al., 2022); 35867785 – (Wu J.W. et al., 2022); 36809369 – (Tabuloc et al., 2013); 38249074 – (Balakireva et al., 2024).

Table 2. Using the FlyDEGdb knowledge base to the analysis of DEG expression in the hypothalamus of WAG and ISIAH rats in response to restriction stress (Oshchepkov et al., 2024)

(Oshchepkov et al., 2024)				FlyDEGdb (this paper)			PC1, 65 %	PC2, 33 %
No.	rat DEG	$\log_2$(stress/norm)		*Drosophila* DEG	$\log_2$(stress/norm)	N_{FlyDEG}	overall stress response	interspecies difference
		WAG	ISIAH					
	I	II	III	IV	V	VI	VII	VIII
1	*Acr*	–0.75	–0.76	*Jon74E*	–6.28	28	–1.17	–3.56
2	*Alox12**	–0.72	–0.78			0		
3	*Atp2b4*	–1.00	–0.61	*PMCA*	–0.48	7	–1.04	–0.26
4	*Cd180*	–1.19	–1.61	*Toll–7*	–3.60	43	–2.07	–1.96
5	*Cdkn1a*	0.93	1.35	*dap*	–0.79	12	2.01	–0.64
6	*Chrna7*	–0.95	–0.68	*nAChRa6*	–2.23	21	–1.12	–1.25
7	*Creb5*	–0.63	–0.66	*Atf-2*	0.28	3	–0.74	0.16
8	*Cryab*	0.67	0.82	*l(2)efl*	–1.96	22	1.35	–1.26
9	*Cyp26b1*	–0.76	–0.80	*Cyp313a2*	–5.80	714	–1.19	–3.29
10	*Ddit4*	0.61	0.65	*scyl*	–0.81	31	1.22	–0.59
11	*Dhrs9*	–1.60	–1.40	*CG8888*	–1.84	13	–2.18	–0.96
12	*Evi2b**	–0.62	–0.68			0		
13	*Fkbp5*	0.79	1.38	*Fkbp59*	–1.99	28	1.87	–1.32
14	*Flvcr2**	–0.70	–0.59			0		
15	*Fmo2*	0.69	0.96	*Fmo-2*	–2.30	56	1.46	–1.47
16	*Fosb*	1.75	1.23	*kay*	–1.12	29	2.58	–0.85
17	*Fosl1*	1.23	1.70	*kay*	–1.12	22	2.51	–0.86
18	*Fosl2*	0.68	0.74	*kay*	–1.12	22	1.33	–0.78
19	*Gpd1*	1.02	1.68	*Gpdh1*	–1.08	9	2.32	–0.83
20	*Hpd*	0.69	0.62	*Hpd*	–2.97	21	1.18	–1.82
21	*Hspa1b*	2.88	1.01	*Hsc70-4*	–0.56	14	3.37	–0.56
22	*Il17rd**	–0.76	–0.65			0		
23	*Il21r**	–0.60	–0.66			0		
24	*Lims2*	0.59	1.02	*stck*	–1.15	20	1.47	–0.81
25	*Lmod2*	1.07	1.00	*tmod*	–3.56	26	1.75	–2.20
26	*Maff*	0.59	0.82	*maf-S*	–1.72	20	1.29	–1.12
27	*Map3k6*	1.21	1.18	*Ask1*	–0.51	9	2.12	–0.48
28	*Mt2A*	0.65	0.59	*MtnA*	–8.97	156	0.89	–5.24
29	*Npas4*	1.09	–0.67	*dysf*	–1.22	6	0.61	–0.76
30	*P2ry4*	–1.14	–0.76	*PK2-R1*	–4.04	51	–2.07	–2.27
31	*Pcdh11x**	–0.61	–0.69			0		
32	*Pik3ap1*	–0.63	–0.95	*stumps*	–2.81	27	–1.08	–1.58
33	*Pla2g3*	0.78	1.86	*GIIIspla2*	–2.41	24	2.21	–1.59
34	*Plek*	–0.61	–0.90	*kmr*	–2.83	27	–1.03	–1.59
35	*Ptch1*	–0.64	–0.82	*ptc*	–1.75	17	–0.95	–0.98
36	*Rasgrp3*	–0.66	–0.62	*Sos*	–1.04	18	–0.79	–0.59
37	*Rin3*	0.59	1.07	*spri*	–1.13	14	1.51	–0.80
38	*Scrt2*	0.65	0.64	*scrt*	–1.65	24	1.21	–1.07
39	*Tmc7*	–0.87	–0.72	*Tmc*	–2.16	4	–1.08	–1.22
40	*Tnfrsf11a**	0.82	1.18			0		
41	*Ttll10*	–0.68	–0.86	*TTLL1B*	–9.99	27	–1.33	–5.67
42	*Zbtb16*	1.23	2.05	*CG43120*	1.44	36	2.87	0.57
Overall number of *Drosophila* DEGs homologous to rat DEGs						1,601		

Note. N_{FlyDEG} – number of *Drosophila* DEGs homologous to the rat DEG according to FlyBase (Ozturk-Colak et al., 2024). * Rat genes (*Flvcr2*, *Alox12*, *Evi2b*, *Il17rd*, *Il21r*, *Pcdh11x*, *Tnfrsf11a*), for which no homologous *Drosophila* genes are found in FlyDEGdb v0.1 (N_{FlyDEG} = 0).

Consider the representation of this information by the example of the first row of the table. It describes the rat *Acr* DEG. Column I indicates the gene name; columns II and III, stress-induces changes in its expression in rats of the WAG and ISIAH strains, respectively. Column IV indicates the name of the homologous *Drosophila Jon74E* gene; column V, the magnitude of its expression change; and column VI shows the total number of such *Drosophila* DEGs homologous to *Acr*.

Columns VII and VIII show the values of the first (PC1) and second (PC2) principal components revealed in the analysis of the above-described experimental data on the magnitude of stress-induced change in *Drosophila* DEG expression from FlyDEGdb and homologous rat genes from RatDEGdb (Oshchepkov et al., 2024). The analysis was conducted with Past v.4.04 software (Hammer et al., 2001).

The first principal component (PC1) is the weighted-mean estimate of the overall stress-induced change in the expression of homologous *Drosophila* (DEG_{FLY}) and rat (DEG_{ISIAH} and DEG_{WAG}) genes:

$$PC1 = 0.1\log_2(DEG_{FLY}) + \\ + [\log_2(DEG_{ISIAH}) + \log_2(DEG_{WAG})]. \quad (1)$$

Principal component PC1 explains 65 % of the variance in the entire set of the considered experimental data on homologous rat and *Drosophila* DEGs.

Principal component PC2 is the weighted-mean estimate of the interspecies difference between *Drosophila* and rat in stress-induces changes in the expression of DEGs and their homologs:

$$PC2 = \log_2(DEG_{FLY}) - \\ - 0.1\,[\log_2(DEG_{ISIAH}) + \log_2(DEG_{WAG})]. \quad (2)$$

Principal component PC2 explains 33 % of the variance in the considered experimental data.

Thus, we were first to find that two-thirds (65 %) of the variance in gene expression change in the rat and *Drosophila* exposed to stress were determined by common mechanisms of response to stress (PC1), and one-third (33 %) reflects interspecies difference between the rat and *Drosophila* (PC2).

The statistical significance ($p < 0.05$) of principal components PC1 and PC2 found in our study was deduced from 1,000 bootstrap samples with a special module of Past v.4.04 software (Hammer et al., 2001) (Fig. S1).

The numerical values of PC1 and PC2 are shown in columns VII and VIII of Table 2 and in Figure 3. For example, the PC1 and PC2 values for the rat *Acr* gene and the homologous *Drosophila Jon74E* gene, described in the first row of Table 2, are –1.17 and –3.56, respectively.

Figure 3 presents the results of the correlation analysis between principal components PC1 and PC2 on the grounds of experimental data on pairs of homologous rat and *Drosophila* DEGs (Table 1). Each point in the figure corresponds to the PC values calculated for a pair of DEGs: *Drosophila* gene from FlyDEGdb and the homologous rat gene from RatDEGdb. The PC1 and PC2 values are plotted along the Y and X axes, respectively. We see that the red dash-dotted line PC1 = 0 divides all DEGs into two disjoint groups: (1) group of DEGs with PC1 < 0, indicating stress-induced downregulation in both rats and *Drosophila*, and (2) group with PC1 > 0, indicating stress-induced upregulation in both species, according to Equation (1).

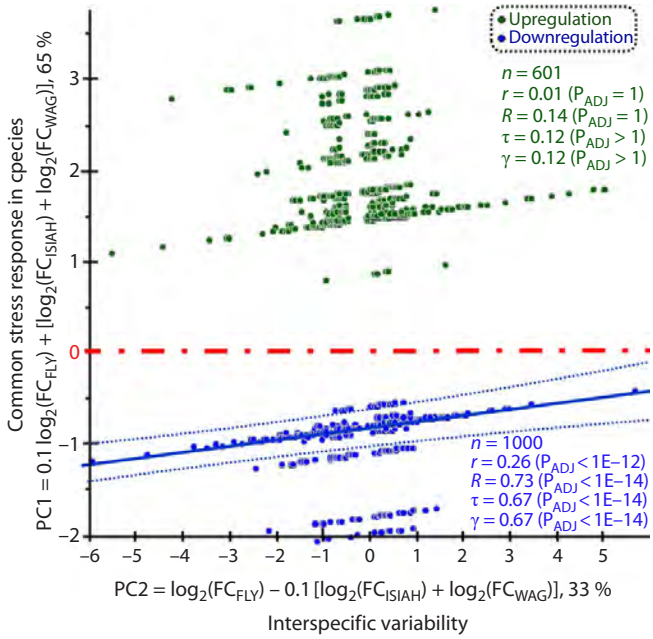


Fig. 3. Results of the correlation analysis between principal components PC1 and PC2 for experimental data on pairs of DEGs homologous between the rat and *Drosophila* (Table 1).

Principal components: PC1, Y axis; PC2, X axis. Each point corresponds to the values calculated for a certain pair of DEGs: *Drosophila* gene from FlyDEGdb and its homolog from RatDEGdb. The red dash-dotted line is the boundary between figure areas for stress-induced downregulation (blue) and upregulation (green) according to the PC1 estimate by Equation (1); the solid line reflects the linear correlation between PC1 and PC2 at PC1 < 0; the dotted lines border the 95 % confidence range for the correlation; alphabetical designations r, γ, R, τ, and PADJ are correlation coefficients, respectively: linear correlation, Goodman–Kruskal generalization; Spearman–Kendal rank correlation, and their statistical significance levels with Bonferroni correction for multiple comparisons, as calculated with Statistica software (Statsoft™, United States).

We can see a qualitative difference between the two DEG groups (above and below the red line) found in our comparison of stress-induced changes in the expression of homologous *Drosophila* and rat genes. The DEG group with stress-induced downregulation (blue) demonstrates a highly significant ($p < 10^{-12}$) positive correlation between PC1 and PC2. By contrast, no correlation between PC1 and PC2 is observed in the second DEG group with stress-induced upregulation (green).

Unexpectedly, our results on the rat and *Drosophila* coincided with independent observations by D.Yu. Oshchepkov et al. (2025). They analyzed changes in the expression of homologous genes of the rat and human induced by stFress and hypertension, respectively. In both cases, a significant correlation between the first and second principal components was noted only in the stress-induced downregulation of homologous genes.

The correlation between PC1 and PC2 in the PC1 < 0 area, which corresponds to stress-induced downregulation in the human, rat, and *Drosophila*, implies that the species may share common molecular mechanisms for gene inhibition under stress conditions of different sorts.

Table 3. Assessments of gene ontology term enrichment in the group of Drosophila genes with stress-induced downregulation

Gene Ontology (GO)			Enrichment		P_{ADJ}
No.	Gene Ontology identifier, GO:ID	Gene Ontology term	Share of *Drosophila* DEGs with stress-induced downregulation	Share of GO:ID	
1	GO:0005887	integral component of plasma membrane	12 of 56	12 of 520	0.0025
2	GO:0005892	acetylcholine-gated channel complex	3 of 56	3 of 7	0.005
3	GO:0005886	plasma membrane	18 of 56	18 of 1485	0.005
4	GO:0120025	plasma membrane-bounded cell projection	11 of 56	11 of 717	0.05
5	GO:0005929	cilium	6 of 56	6 of 188	0.05

The commonly known molecular mechanisms for gene expression downregulation under stress include the arrest of pre-mRNA splicing in eukaryotes (Yost, Lindquist, 1986; Cuesta et al., 2000) and translation inhibition (Bresson et al., 2020).

We used the STRING software (Szklarczyk et al., 2021) to assess the Gene Ontology (GO) term enrichment in the group of *Drosophila* genes with stress-induced downregulation (blue dots in Figure 3). The results are shown in Table 3.

The analysis revealed five GO terms in which the list of *Drosophila* genes with stress-induced downregulation is significantly ($p < 0.05$) enriched. Four of the five (GO:0005887, GO:0005892, GO:0005886, GO:0120025, and GO:0005929) are directly related to components of the plasma membrane. The fifth term (GO:0005929, cilium), also belongs to this group, as cilia are specific organelles on the outer surface of eukaryotic cell membranes. This fact implies that the plasma membrane of *Drosophila* cells is one of the universal targets of stress factors described in FlyDEGdb. In this regard, note that stress-induced downregulation of *Drosophila* genes encoding components of plasma membranes in cells can slower their growth under stress. Our assumption agrees with the results presented in (Kassahn et al., 2009), where mechanisms of animal response to stress factors are reviewed. It should also be mentioned that M.F. Haque et al. (2025) detected an inhibition of *Escherichia coli* cell growth under stress.

To conclude, we note that the year 2023 marked the 80th anniversary of the famous maxim by Hans Selye "Stress is the spice of life" (Rochette et al., 2023). Our work once more illustrates the fundamental significance of the stress issue in life sciences.

Conclusion

We developed the FlyDEGdb knowledge base, which is a body of experimental data on differentially expressed genes (DEGs) of *Drosophila* and their response to a broad range of stressing factors: cold, heat, dehydration, heavy metals, ionizing radiation, starvation, household chemicals, drugs, agricultural fertilizers, insecticides, pesticides, herbicides, and other toxicants. The knowledge base, storing information on the commonly used model species, *D. melanogaster*, can be employed by students of translational molecular biology and genetics of the human and animals, physiology, translational medicine, pharmacology, nutrition science, agricultural chemistry, radiation biology, toxicology, and bioinformatics.

References

Ali M.Z., Anushree, Bilgrami A.L., Ahsan J. *Drosophila melanogaster* chemosensory pathways as potential targets to curb the insect menace. *Insects*. 2022;13(2):142. doi 10.3390/insects13020142

Avgustinovich D.F., Chadaeva I.V., Kizimenko A.V., Kovner A.V., Bazovkina D.V., Ponomarev D.V., Evseenko V.I., Naprimerov V.A., Lvova M.N. The liver-brain axis under the influence of chronic *Opisthorchis felineus* infection combined with prolonged alcoholization in mice. *Vavilov J Genet Breed*. 2025;29(1):92-107. doi 10.18699/vjgb-25-11

Balakireva Y., Nikitina M., Makhnovskii P., Kukushkina I., Kuzmin I., Kim A., Nefedova L. The lifespan of *D. melanogaster* depends on the function of the *Gagr* gene, a domesticated *gag* gene of Drosophila LTR retrotransposons. *Insects*. 2024;15(1):68. doi 10.3390/insects15010068

Bresson S., Shchepachev V., Spanos C., Turowski T.W., Rappsilber J., Tollervey D. Stress-induced translation inhibition through rapid displacement of scanning initiation factors. *Mol Cell*. 2020;80(3):470-484.e8. doi 10.1016/j.molcel.2020.09.021

Brown G.R., Hem V., Katz K.S., Ovetsky M., Wallin C., Ermolaeva O., Tolstoy I., Tatusova T., Pruitt K.D., Maglott D.R., Murphy T.D. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015;43(D1):D36-D42. doi 10.1093/nar/gku1055

Chadaeva I., Ponomarenko P., Rasskazov D., Sharypova E., Kashina E., Kleshchev M., Ponomarenko M., Naumenko V., Savinkova L., Kolchanov N., Osadchuk L., Osadchuk A. Natural selection equally supports the human tendencies in subordination and domination: a genome-wide study with *in silico* confirmation and *in vivo* validation in mice. *Front Genet*. 2019;10:73. doi 10.3389/fgene.2019.00073

Chadaeva I.V., Filonov S.V., Zolotareva K.A., Khandaev B.M., Ershov N.I., Podkolodnyy N.L., Kozhemyakina R.V., … Stefanova N.A., Kolosova N.G., Markel A.L., Ponomarenko M.P., Oshchepkov D.Y. RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):794-806. doi 10.18699/VJGB-23-92

Chatterjee N., Perrimon N. What fuels the fly: energy metabolism in *Drosophila* and its application to the study of obesity and diabetes. *Sci Adv*. 2021;7(24):eabg4336. doi 10.1126/sciadv.abg4336

Chen J., Nolte V., Schlotterer C. Temperature stress mediates decanalization and dominance of gene expression in *Drosophila melanogaster*. *PLoS Genet*. 2015;11(2):e1004883. doi 10.1371/journal.pgen.1004883

Cuesta R., Laroia G., Schneider R.J. Chaperone Hsp27 inhibits translation during heat shock by binding eIF4G and facilitating dissociation of cap-initiation complexes. *Genes Dev*. 2000;14(12):1460-1470. doi 10.1101/gad.14.12.1460

De Gregorio E., Spellman P.T., Rubin G.M., Lemaitre B. Genome-wide analysis of the *Drosophila* immune response by using oligonucle-

otide microarrays. *Proc Natl Acad Sci USA*. 2001;98(22):12590-12595. doi 10.1073/pnas.221458698

de Oliveira D.S., Rosa M.T., Vieira C., Loreto E.L.S. Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. *J Evol Biol*. 2021;34(4):628-638. doi 10.1111/jeb.13762

Drouet J.B., Peinnequin A., Faure P., Denis J., Fidier N., Maury R., Buguet A., Cespuglio R., Canini F. Stress-induced hippocampus Npas4 mRNA expression relates to specific psychophysiological patterns of stress response. *Brain Res*. 2018;1679:75-83. doi 10.1016/j.brainres.2017.11.024

Elshazley M., Sato M., Hase T., Yamashita R., Yoshida K., Toyokuni S., Ishiguro F., Osada H., Sekido Y., Yokoi K., Usami N., Shames D.S., Kondo M., Gazdar A.F., Minna J.D., Hasegawa Y. The circadian clock gene *BMAL1* is a novel therapeutic target for malignant pleural mesothelioma. *Int J Cancer*. 2012;131(12):2820-2831. doi 10.1002/ijc.27598

Fyie L.R., Gardiner M.M., Meuti M.E. Artificial light at night alters the seasonal responses of biting mosquitoes. *J Insect Physiol*. 2021;129:104194. doi 10.1016/j.jinsphys.2021.104194

Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43(D1):D1049-D1056. doi 10.1093/nar/gku1179

Gery S., Komatsu N., Baldjyan L., Yu A., Koo D., Koeffler H.P. The circadian gene Per1 plays an important role in cell growth and DNA damage control in human cancer cells. *Mol Cell*. 2006;22(3):375-382. doi 10.1016/j.molcel.2006.03.038

Goda T., Sharp B., Wijnen H. Temperature-dependent resetting of the molecular circadian oscillator in Drosophila. *Proc Biol Sci*. 2014;281(1793):20141714. doi 10.1098/rspb.2014.1714

Gruntenko N.E., Khlebodarova T.M., Sukhanova M.Jh., Vasenkova I.A., Kaidanov L.Z., Rauschenbach I.Yu. Prolonged negative selection of *Drosophila melanogaster* for a character of adaptive significance disturbs stress reactivity. *Insect Biochem Mol Biol*. 1999;29(5):445-452. doi 10.1016/s0965-1748(99)00021-1

Gruntenko N.E., Deryuzhenko M.A., Andreenkova O.V., Shishkina O.D., Bobrovskikh M.A., Shatskaya N.V., Vasiliev G.V. *Drosophila melanogaster* transcriptome response to different *Wolbachia* strains. *Int J Mol Sci*. 2023;24(24):17411. doi 10.3390/ijms242417411

Gryksa K., Schmidtner A.K., Masís-Calvo M., Rodriguez-Villagra O.A., Havasi A., Wirobski G., Maloumby R., Jägle H., Bosch O.J., Slattery D.A., Neumann I.D. Selective breeding of rats for high (HAB) and low (LAB) anxiety-related behaviour: a unique model for comorbid depression and social dysfunctions. *Neurosci Biobehav Rev*. 2023;152:105292. doi 10.1016/j.neubiorev.2023.105292

Hammer O., Harper D.A.T., Ryan P.D. PAST: PAleontological STatistics software package for education and data analysis. *Palaeontol Electron*. 2001;4(1):1-9

Haque M.F., Tarusawa T., Ushida C., Ito S., Himeno H. cAMP-CRP-activated *E. coli* causes growth arrest under stress conditions. *Front Microbiol*. 2025;16:1597530. doi 10.3389/fmicb.2025.1597530

Huang K., Chen W., Zhu F., Li P.W., Kapahi P., Bai H. RiboTag translatomic profiling of *Drosophila* oenocytes under aging and induced oxidative stress. *BMC Genomics*. 2019;20(1):50. doi 10.1186/s12864-018-5404-4

Jauregui-Lozano J., Hall H., Stanhope S.C., Bakhle K., Marlin M.M., Weake V.M. The Clock:Cycle complex is a major transcriptional regulator of *Drosophila* photoreceptors that protects the eye from retinal degeneration and oxidative stress. *PLoS Genet*. 2022;18(1):e1010021. doi 10.1371/journal.pgen.1010021

Kassahn K.S., Crozier R.H., Portner H.O., Caley M.J. Animal performance and stress: responses and tolerance limits at different levels of biological organisation. *Biol Rev Camb Philos Soc*. 2009;84(2):277-292. doi 10.1111/j.1469-185X.2008.00073.x

Lakhotia S.C. Tracing the roots of molecular biology. Part III: Morgan's Fly Room and emergence of modern genetics. *Reson*. 2025;29(3):379-409. doi 10.1007/s12045-024-1669-x

Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* (*Oxford*). 2011;2011:baq036. doi 10.1093/database/baq036

Markel A.L. Genetic model of stress-induced arterial hypertension. *Izvestiya AN SSSR*. 1985;3:466-469 (in Russian)

Mehraj V., Textoris J., Capo C., Raoult D., Leone M., Mege J.L. Overexpression of the *Per2* gene in male patients with acute Q fever. *J Infect Dis*. 2012;206(11):1768-1770. doi 10.1093/infdis/jis600

Mikucki E.E., O'Leary T.S., Lockwood B.L. Heat tolerance, oxidative stress response tuning and robust gene activation in early-stage *Drosophila melanogaster* embryos. *Proc R Soc B*. 2024;291(2029):20240973. doi 10.1098/rspb.2024.0973

Morgan T.H. Sex limited inheritance in *Drosophila*. *Science*. 1910;32(812):120-122. doi 10.1126/science.32.812.120

Mukherjee P., Roy S., Ghosh D., Nandi S.K. Role of animal models in biomedical research: a review. *Lab Anim Res*. 2022;38(1):18. doi 10.1186/s42826-022-00128-1

Oishi K., Ohkura N., Amagai N., Ishida N. Involvement of circadian clock gene *Clock* in diabetes-induced circadian augmentation of plasminogen activator inhibitor-1 (*PAI-1*) expression in the mouse heart. *FEBS Lett*. 2005;579(17):3555-3559. doi 10.1016/j.febslet.2005.05.027

Oshchepkov D.Y., Makovka Y.V., Fedoseeva L.A., Seryapina A.A., Markel A.L., Redina O.E. Effect of short-term restraint stress on the hypothalamic transcriptome profiles of rats with Inherited Stress-Induced Arterial Hypertension (ISIAH) and normotensive Wistar Albino Glaxo (WAG) rats. *Int J Mol Sci*. 2024;25(12):6680. doi 10.3390/ijms25126680

Oshchepkov D.Yu., Makovka Yu.V., Chadaeva I.V., Bogomolov A.G., Fedoseeva L.A., Seryapina A.A., Ponomarenko M.P., Markel A.L., Redina O.E. Genes representing a stress-dependent component in the development of arterial hypertension. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2025;29(8). doi 10.18699/vjgb-25-139

Ozturk-Colak A., Marygold S.J., Antonazzo G., Attrill H., Goutte-Gattat D., Jenkins V.K., Matthews B.B., Millburn G., Dos Santos G., Tabone C.J.; FlyBase Consortium. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024;227(1):iyad211. doi 10.1093/genetics/iyad211

Podkolodnaya O.A., Chadaeva I.V., Filonov S.V., Podkolodnyy N.L., Rasskazov D.A., Tverdokhleb N.N., Zolotareva K.A., Bogomolov A.G., Kondratyuk E.Y., Oshchepkov D.Y., Ponomarenko M.P. MiceDEGdb: a knowledge base on differentially expressed mouse genes as a model object in biomedical research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2025;29(1):153-161. doi 10.18699/vjgb-25-18

Ramnarine T.J.S., Grath S., Parsch J. Natural variation in the transcriptional response of *Drosophila melanogaster* to oxidative stress. *G3* (*Bethesda*). 2022;12(1):jkab366. doi 10.1093/g3journal/jkab366

Rand M.D., Tennessen J.M., Mackay T.F.C., Anholt R.R.H. Perspectives on the *Drosophila melanogaster* model for advances in toxicological science. *Curr Protoc*. 2023;3(8):e870. doi 10.1002/cpz1.870

Rochette L., Dogon G., Vergely C. Stress: eight decades after its definition by Hans Selye: "Stress is the spice of life". *Brain Sci*. 2023;13(2):310. doi 10.3390/brainsci13020310

Szklarczyk D., Gable A.L., Nastou K.C., Lyon D., Kirsch R., Pyysalo S., Doncheva N.T., Legeay M., Fang T., Bork P., Jensen L.J., von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605-D612. doi 10.1093/nar/gkaa1074

Tabuloc C.A., Cai Y.D., Kwok R.S., Chan E.C., Hidalgo S., Chiu J.C. CLOCK and TIMELESS regulate rhythmic occupancy of the BRAHMA chromatin-remodeling protein at clock gene promoters. *PLoS Genet*. 2023;19(2):e1010649. doi 10.1371/journal.pgen.1010649

Telonis-Scott M., van Heerwaarden B., Johnson T.K., Hoffmann A.A., Sgro C.M. New levels of transcriptome complexity at upper thermal limits in wild *Drosophila* revealed by exon expression analysis. *Genetics*. 2013;195(3):809-830. doi 10.1534/genetics.113.156224

von Heckel K., Stephan W., Hutter S. Canalization of gene expression is a major signature of regulatory cold adaptation in temperate *Drosophila melanogaster*. *BMC Genomics*. 2016;17:574. doi 10.1186/s12864-016-2866-0

Wade J. Practical guidelines for a user-friendly interface. *ACM SIGAPL APL Quote Quad*. 1984;14(4):365-371. doi 10.1145/384283.801122

Wang T., Yang P., Zhan Y., Xia L., Hua Z., Zhang J. Deletion of circadian gene *Per1* alleviates acute ethanol-induced hepatotoxicity in mice. *Toxicology*. 2013;314(2-3):193-201. doi 10.1016/j.tox.2013.09.009

Wu J.W., Wang C.W., Chen R.Y., Hung L.Y., Tsai Y.C., Chan Y.T., Chang Y.C., Jang A.C. Spatiotemporal gating of Stat nuclear influx by *Drosophila* Npas4 in collective cell migration. *Sci Adv*. 2022;8(29):eabm2411. doi 10.1126/sciadv.abm2411

Wu K., Tang Y., Zhang Q., Zhuo Z., Sheng X., Huang J., Ye J., Li X., Liu Z., Chen H. Aging-related upregulation of the homeobox gene *caudal* represses intestinal stem cell differentiation in *Drosophila*. *PLoS Genet*. 2021;17(7):e1009649. doi 10.1371/journal.pgen.1009649

Yost H.J., Lindquist S. RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis. *Cell*. 1986;45(2):185-193. doi 10.1016/0092-8674(86)90382-x

Yu S., Luo F., Xu Y., Zhang Y., Jin L.H. *Drosophila* innate immunity involves multiple signaling pathways and coordinated communication between different tissues. *Front Immunol*. 2022;13:905370. doi 10.3389/fimmu.2022.905370

Conflict of interest. The authors declare no conflict of interest.

Received October 3, 2025. Revised October 14, 2025. Accepted October 15, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):963-977

doi 10.18699/vjgb-25-102

Reconstruction and analysis of the gene network regulating apoptosis in hepatocellular carcinoma based on scRNA-seq data and the ANDSystem knowledge base

A.V. Adamovskaya [1, 2] ✉, I.V. Yatsyk [1, 2], M.A. Kleshchev [1, 2], P.S. Demenkov [1, 2], T.V. Ivanisenko [1, 2], V.A. Ivanisenko [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ adamovskayaav@bionet.nsc.ru

Abstract. Hepatocellular Carcinoma (HCC) is the most common primary liver cancer characterized by rapid progression, high mortality rate and therapy resistance. One of the key areas in studying the molecular mechanisms of HCC development is the analysis of disturbances in apoptosis processes in hepatocytes. Throughout life apoptosis ensures the elimination of old and defective cells while the attenuation of this process serves as one of the leading factors in carcinogenesis. In this study we reconstructed and analyzed the gene network regulating hepatocyte apoptosis in humans based on single-cell transcriptome sequencing (scRNA-seq) data and the ANDSystem knowledge base which employs artificial intelligence and computational systems biology methods. Comparative analysis of gene expression revealed weakened transcription of genes involved in the regulation of inflammatory processes and apoptosis in tumor hepatocytes compared to hepatocytes of normal liver tissue. The reconstructed network included 116 differentially expressed genes annotated in Gene Ontology as genes involved in the apoptotic process (apoptotic process GO:0006915), along with their 116 corresponding protein products. It also included 16 additional proteins that, while lacking GO apoptosis annotation, were differentially expressed in HCC and interacting with genes and proteins participating in the apoptosis process. Computational analysis of the gene network identified several key protein products encoded by the genes *NFKB1*, *MMP9*, *BCL2*, *A4*, *CDKN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*. These proteins exhibited both a high degree of connectivity with other network objects and differential expression in HCC. Of particular interest are proteins CDKN1A, ERBB2, IL8, and EGR1, which are not annotated in Gene Ontology as apoptosis participants but have a statistically significant number of interactions with genes involved in apoptosis. This indicates their role in regulating programmed cell death. The obtained results can guide the design of new experiments studying the role of apoptosis in carcinogenesis and aid in the search for novel therapeutic targets and approaches for HCC therapy using apoptosis modulation in malignant hepatocytes. Furthermore, the proposed approach to reconstructing and analyzing the apoptosis regulation gene network in hepatocellular carcinoma can be applied to analyze other tumor forms providing a systemic understanding of disturbances in key regulatory processes in oncogenesis and potential therapy targets.

Key words: hepatocellular carcinoma; single cell transcriptomics; apoptosis; gene networks; cognitive system ANDSystem

For citation: Adamovskaya A.V., Yatsyk I.V., Kleshchev M.A., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Reconstruction and analysis of the gene network regulating apoptosis in hepatocellular carcinoma based on scRNA-seq data and the ANDSystem knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):963-977. doi 10.18699/vjgb-25-102

Funding. The study was supported by budget project FWNR-2022-0020.

Реконструкция и анализ генной сети регуляции апоптоза при гепатоцеллюлярной карциноме на основе данных scRNA-seq и базы знаний ANDSystem

А.В. Адамовская [1, 2] ✉, И.В. Яцык [1, 2], М.А. Клещев [1, 2], П.С. Деменков [1, 2], Т.В. Иванисенко [1, 2], В.А. Иванисенко [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ adamovskayaav@bionet.nsc.ru

Аннотация. Гепатоцеллюлярная карцинома – наиболее распространенный первичный рак печени, характеризующийся быстрым прогрессированием, высокой летальностью и устойчивостью к терапии. Одним

из ключевых направлений в изучении молекулярных механизмов развития гепатоцеллюлярной карциномы является анализ нарушений процессов апоптоза в гепатоцитах. На протяжении всей жизни благодаря апоптозу происходит элиминация старых и дефектных клеток, тогда как ослабление апоптотической гибели служит одним из ведущих факторов канцерогенеза. В настоящем исследовании выполнены реконструкция и анализ генной сети регуляции апоптоза гепатоцитов у человека на основе данных секвенирования транскриптома одиночных клеток (scRNA-seq) и базы знаний ANDSystem, использующей методы искусственного интеллекта и компьютерной системной биологии. Сравнительный анализ экспрессии генов показал ослабление транскрипции генов, вовлеченных в регуляцию воспалительных процессов и апоптоза, в опухолевых гепатоцитах по сравнению с гепатоцитами нормальной ткани печени. Реконструированная сеть включала 116 дифференциально экспрессирующихся генов, аннотированных в Gene Ontology как гены, вовлеченные в процесс апоптоза (apoptotic process GO:0006915), 116 соответствующих белков, а также 16 дополнительных белков, не имеющих GO-аннотации, но дифференциально экспрессируемых при гепатоцеллюлярной карциноме и вовлеченных во взаимодействия с генами и белками, участвующими в процессе апоптоза. Компьютерный анализ генной сети выявил ряд ключевых белков – продуктов генов *NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*, демонстрирующих как высокое число связей с другими объектами сети, так и дифференциальную экспрессию при гепатоцеллюлярной карциноме. Особый интерес представляют белки CDKN1A, ERBB2, IL8 и EGR1, не аннотированные в Gene Ontology как участники апоптоза, но обладающие статистически значимым числом взаимодействий с генами, вовлеченными в апоптоз, что указывает на их роль в регуляции программируемой клеточной гибели. Полученные результаты могут найти применение для планирования новых экспериментов по изучению роли апоптоза в канцерогенезе и поиска новых мишеней и подходов для терапии гепатоцеллюлярной карциномы, основанных на модуляции апоптоза в злокачественных гепатоцитах. Предложенный подход к реконструкции и анализу генной сети регуляции апоптоза при гепатоцеллюлярной карциноме может быть использован для анализа других форм опухолей и дает системное представление о нарушениях ключевых регуляторных процессов в онкогенезе и потенциальных мишенях для терапии.

Ключевые слова: гепатоцеллюлярная карцинома; транскриптомика одиночных клеток; апоптоз; генные сети; когнитивная система ANDSystem

Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer arising from the malignant transformation of hepatocytes. Approximately 750,000 people die from this disease worldwide each year (Ganesan, Kulik, 2023). This malignancy is characterized by marked resistance to anticancer drugs and a high rate of recurrence (Zou et al., 2025), underscoring the relevance of investigating both the molecular mechanisms of tumorigenesis and the development of tumor resistance – and, on this basis, identifying targets for anticancer therapy. The principal risk factors for HCC include chronic infection with hepatitis B and C viruses, alcoholic cirrhosis, and non-alcoholic steatohepatitis; other established risk factors comprise obesity, type 2 diabetes mellitus, and tobacco smoking (Ogunwobi et al., 2019).

Viral infections and/or adverse environmental factors (exposure to hepatotoxic agents) induce alterations in the functioning of a number of signaling pathways in hepatocytes, leading to their malignant transformation and the development of HCC. It has been established that the hepatitis B virus X protein (HBx) suppresses the activity of the pro-apoptotic protein p53, impairs DNA repair, and activates several signaling cascades (STAT, NF-κB, AP-1, etc.) involved in cell proliferation and survival, thereby promoting HCC progression (Jiang Y. et al., 2019). The pathogenesis of HCC involves changes in: (a) growth factor signaling pathways such as insulin-like growth factor (IGF), epidermal growth factor (EGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), and hepatocyte growth factor (HGF/MET); (b) signaling pathways related to cell differentiation, including WNT, Hedgehog, and Notch; and (c) angiogenesis-related pathways driven by vascular endothelial growth factor (VEGF) and FGF (Dhanasekaran et al., 2016). In addition, disruption of apoptosis – programmed cell death – makes a crucial contribution to HCC progression (Fabregat, 2009). Chronic liver inflammation resulting from hepatitis B or C virus infection or exposure to adverse environmental factors leads to hepatocyte apoptosis accompanied by a compensatory increase in their proliferation, which, under conditions of high oxidative stress caused by inflammation, results in the accumulation of DNA mutations and an increased likelihood of malignant transformation of hepatocytes (Yang et al., 2019). Moreover, apoptosis plays a key role in eliminating malignant cells; therefore, activation of apoptosis is one of the mechanisms of action of anticancer drugs in HCC (Hajizadeh et al., 2023). It has been shown that suppression of the extrinsic and intrinsic apoptosis pathways – particularly by regulatory microRNAs – may be associated with the development of HCC and poor clinical outcomes (Khlebodarova et al., 2023). It has also been established that the hepatitis B virus HBx protein suppresses the activity of the pro-apoptotic protein p53, contributing to the initiation and progression of HCC (Jiang Y. et al., 2019). Available data indicate that disruption of the balance between pro-apoptotic and anti-apoptotic proteins in hepatocytes is one of the factors underlying HCC development and the emergence of drug resistance (Ladd et al., 2024; Wu et al., 2024). This necessitates investigating the mechanisms by which apoptotic pathways in hepatocytes are perturbed during HCC development and identifying key regulatory nodes of apoptosis, the expres-

sion of which differs between healthy and tumor hepatocytes.

It is well known that disturbances in the interactions among tumor cells, the stroma, and immune cells play an important role in disease progression, fostering HCC development, the emergence of drug resistance, and recurrence (Xue et al., 2022). Notably, HCC exhibits a high degree of cellular heterogeneity, which highlights the importance of methods that probe the molecular processes of HCC development at the single-cell level (Li X. et al., 2022).

One such method – single-cell transcriptome sequencing – provides valuable information on gene expression features across different cell types within tumor tissue. This is particularly relevant when comparing malignantly transformed hepatocytes within the tumor to normal hepatocytes from histologically unaltered liver tissue (Zhang et al., 2022). However, differential expression analysis alone is insufficient to elucidate the mechanisms of tumor transformation. Based on such experimental data, it is necessary to reconstruct gene networks – ensembles of coordinately functioning genes – which provide valuable insights into dysregulated molecular mechanisms of gene–gene interactions responsible for the development of pathological processes (Saik et al., 2019; Ivanisenko V.A. et al., 2022; Antropova et al., 2023; Butikova et al., 2025).

The aim of our study was to reconstruct and analyze the gene network regulating apoptosis in hepatocytes in human hepatocellular carcinoma using an integrated approach that combines single-cell transcriptomic data with the ANDSystem software-information platform designed for gene network reconstruction based on automated analysis of scientific publications and biomedical factual databases (Demenkov et al., 2011; Ivanisenko V.A. et al., 2015, 2019). The system employs artificial intelligence methods and an ontological description of the domain, ensuring high coverage and accuracy in knowledge extraction from diverse sources of experimental information (Ivanisenko T.V. et al., 2020, 2022, 2024).

By comparing scRNA-seq transcriptomic data for normal hepatocytes and hepatocytes malignantly transformed in HCC, we identified 1,853 differentially expressed genes (DEGs). Using ANDSystem, we reconstructed an interaction network between the DEGs and genes annotated in Gene Ontology as involved in apoptosis (GO:0006915). Analysis of the resulting gene network highlighted several DEGs, the products of which (including BCL2, NFKB1, FOXO1, MCL1, CDKN1A, ERBB2, IL8, and EGR1) exhibit significant connectivity with components of the apoptosis network. Notably, some of these proteins (CDKN1A, ERBB2, IL8, EGR1) were not annotated in Gene Ontology as apoptosis participants, underscoring their potential novelty and importance for understanding the mechanisms of programmed cell death in HCC. In addition, based on scRNA-seq data, we observed decreased expression of key inhibitors of apoptosis in hepatocellular carcinoma cells. This finding suggests that evasion of apoptosis in HCC may be driven not by the enhancement of anti-apoptotic mechanisms but, on the contrary, by disruption of pro-apoptotic signaling pathways. The results obtained may be useful for planning further experimental studies aimed at elucidating the mechanisms of apoptosis regulation in hepatocytes in HCC and are also of interest for developing targeted therapeutic strategies aimed at modulating apoptotic processes in tumor cells of the liver.

Material and methods

GEO database. For the analysis, we used single-cell transcriptome sequencing data from primary hepatocellular carcinoma (HCC) specimens and paired histologically normal liver tissues, available in the NCBI Gene Expression Omnibus (GEO) under accession GSE149614. Data from eight patients were analyzed (patients 3, 4, 5, 6, 7, 8, 9, and 10).

Transcriptome data analysis. Single-cell RNA-sequencing (scRNA-seq) data processing and downstream analyses were performed in Python using the Scanpy package (v1.9.3) (Wolf et al., 2018). Initial filtering included: (1) removing cells with detected expression for fewer than 100 genes, and (2) removing genes detected in fewer than 3 cells. Normalization was carried out with scanpy.pp.normalize_total(), followed by a log1p transformation. Cell clustering was performed using the Leiden algorithm (Traag et al., 2019). Differentially expressed (marker) genes for each identified cluster were determined with scanpy.tl.rank_genes_groups(), employing the Wilcoxon rank-sum test.

Based on the expression of known hepatocyte marker genes (ALB, HNF4A, SERPINA1, CYP3A4, TAT, TF) (Si-Tayeb et al., 2010) and the clustering results, cells classified as hepatocytes were selected. For subsequent comparative analyses between tumor and normal hepatocytes, pseudobulk samples (Squair et al., 2021) were generated for each patient by aggregating expression values across all cells separately for tumor and normal tissue.

Statistically significant differences in gene expression between the pseudobulk tumor group and the pseudobulk normal hepatocyte group were identified in R using DESeq2 (v1.42.0) (Love et al., 2014). Differentially expressed genes were defined by thresholds of p-value < 0.05 and $|\log FC| > 0.5$.

Reconstruction of gene networks. Reconstruction and analysis of the gene network regulating hepatocyte apoptosis in human hepatocellular carcinoma were performed using the ANDSystem software-information platform (Demenkov et al., 2011; Ivanisenko V.A. et al., 2015, 2019). The effectiveness of ANDSystem has been demonstrated in a number of studies, including reconstruction of the endothelial apoptosis regulatory network in lymphedema (Saik et al., 2019) and investigations of molecular mechanisms associated with hepatocellular carcinoma (Demenkov et al., 2023; Khlebodarova et al., 2023). The system has also been applied to the interpretation of omics data – metabolomics (Ivanisenko V.A. et al., 2022, 2024) and proteomics (Momynaliev et al., 2010; Larina et al., 2015) – demonstrating its versatility and applicability to diverse data types and diseases.

The network reconstruction comprised several stages. First, using the Query Wizard of the ANDVisio software

module (Demenkov et al., 2011), a graphical user interface within ANDSystem, we reconstructed an associative gene network that included genes and their protein products involved in apoptosis. The list of human protein-coding genes participating in apoptosis was obtained from The Gene Ontology Resource (https://geneontology.org/) for the term GO:0006915 "apoptotic process".

At the second stage, we searched for novel proteins involved in the regulation of apoptosis in hepatocytes during HCC development. We considered as candidates those proteins that are not annotated in The Gene Ontology Resource as apoptosis participants but regulate the expression of the initial genes involved in apoptosis.

To identify such proteins, using the Pathway Wizard in ANDVisio, we retrieved all direct relationships of the types Expression regulation, Expression upregulation, Expression downregulation, and Interaction from the protein products of all DEGs identified in the experiment to the DEGs involved in apoptosis according to Gene Ontology.

We then assessed the statistical significance of the specificity of the linkage between the identified proteins and the baseline apoptosis gene network constructed in stage 1. The specificity metric was defined as the proportion of a protein's interactions that connect to genes in the network relative to the total number of that protein's genome-wide interactions. The statistical significance of the deviation between the observed number of a given protein's interactions with network genes and the number expected by chance was evaluated using the hypergeometric distribution:

$$P(X \geq x) = \sum_{k=x}^{\min(N,\, n)} \frac{\left(\frac{n}{k}\right)\left(\frac{M-n}{N-k}\right)}{\frac{M}{N}},$$

where M is the total number of protein-coding genes in the database, n is the number of genes in the analyzed gene network, N is the total number of human genes that interact with the protein under study, and x is the number of network genes that interact with the protein under study.

P-values were calculated using the Python library (scipy.stats.hypergeom). To correct for multiple testing, the Bonferroni adjustment (Narkevich et al., 2020) was applied, under which DEGs were considered statistically significant if their Bonferroni-adjusted p-value satisfied $p < 0.05$. All computations were performed using statsmodels and other standard Python tools.

Thus, the final gene network regulating apoptosis during HCC development included both the DEGs and their products annotated in Gene Ontology as participating in the apoptotic process, and the protein products of DEGs that were statistically significantly linked to this apoptosis network but not annotated as apoptosis participants in Gene Ontology.

Gene network analysis. For each network component (gene or protein), ANDSystem computed the Network Connectivity metric, defined as the number of other network objects (nodes) to which the component is connected (i.e., its degree). Network hubs were defined as proteins and genes, Network Connectivity of which exceeded the critical value (quantile) corresponding to a p-value of 0.05. The quantile was calculated from the empirical distribution of Network Connectivity across all nodes of the gene network. Thus, the number of connections for hub nodes was statistically significant at $p < 0.05$.

Phylostratigraphic analysis of gene networks. The evolutionary age of genes was determined using the GenOrigin database (http://chenzxlab.hzau.edu.cn/) (Tong et al., 2021), which provides gene age annotations across species inferred by phylostratigraphic analysis. To assess the statistical significance of differences in the distribution of gene ages between the full set of human protein-coding genes and the genes in the reconstructed apoptosis network of hepatocytes in HCC, we applied a hypergeometric test. The probability of observing m or more genes from a given age interval among M network genes was calculated using the hypergeom.pmf function from SciPy. The analysis was performed for the 20 age intervals represented in GenOrigin. The following parameters were used in the calculations: N – the total number of human protein-coding genes; n – the number of human protein-coding genes in a given age interval; M – the number of genes in the reconstructed network; m – the number of network genes within the interval under analysis. Differences were considered statistically significant at $p < 0.05$.

Functional annotation of gene sets. Functional annotation of the genes represented in the network was performed using the web-based Database for Annotation, Visualization and Integrated Discovery (DAVID 2021) (https://david.ncifcrf.gov/; Sherman et al., 2022) with default settings. Over-representation analysis of Gene Ontology terms describing biological processes, molecular functions, and cellular components, as well as KEGG pathways (i.e., enrichment analysis of gene sets to identify key biological processes associated with the genes under study), was carried out for (i) the complete set of DEGs identified from the hepatocyte transcriptome analysis and (ii) the subset of DEGs included in the hepatocyte apoptosis regulatory gene network. In DAVID, over-representation of GO terms and KEGG pathways was evaluated using Fisher's exact test (Sherman et al., 2022). Statistical significance of enrichment was defined as a Bonferroni–Šidák-adjusted p-value < 0.05 (Šidák, 1967).

Results

Analysis of differential gene expression in HCC

As a result of comparing single-cell transcriptomes (malignantly transformed tumor hepatocytes vs. hepatocytes from histologically normal liver tissue), 1,853 differentially expressed genes (DEGs) were identified. The data for these DEGs are provided in Table S1[1]. Among them, 964 genes showed increased expression and 889 genes showed decreased expression in tumor hepatocytes compared with normal liver cells. The results of the functional annotation

[1] Tables S1–S7 and Figs S1 and S2 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Adam_Engl_29_7.xlsx

Table 1. Overrepresented Gene Ontology terms for genes with increased and decreased expression in tumor hepatocytes compared with hepatocytes from histologically normal liver tissue in HCC

Genes with increased expression				Genes with reduced expression			
#	Gene Ontology term	%*	*p*-value**	#	Gene Ontology term	%*	*p*-value**
1	GO:0051301~cell division	7.7	0.000	1	GO:0007165~signal transduction	12.4	0.000
2	GO:0007059~chromosome segregation	3.6	0.000	2	GO:0035556~intracellular signal transduction	5.5	0.000
3	GO:0006325~chromatin organization	3.1	0.020	3	GO:0045944~positive regulation of transcription by RNA polymerase II	10.9	0.000
4	GO:0006281~DNA repair	3.7	0.005	4	GO:0000122~negative regulation of transcription by RNA polymerase II	9.4	0.000
5	GO:0006260~DNA replication	2.8	0.000	5	GO:0045893~positive regulation of DNA-templated transcription	6.6	0.021
6	GO:0000398~mRNA splicing, via spliceosome	2.8	0.004	6	GO:0006915~apoptotic process	7.0	0.000
7	GO:0006364~rRNA processing	2.7	0.000	7	GO:0043065~positive regulation of apoptotic process	3.9	0.012
8	GO:0006412~translation	3.4	0.000	8	GO:0043066~negative regulation of apoptotic process	5.3	0.017
9	GO:0032543~mitochondrial translation	2.4	0.000	9	GO:0006954~inflammatory response	4.7	0.019
10	GO:0006457~protein folding	3.3	0.000	10	GO:0016477~cell migration	4.0	0.000

* Proportion of genes associated with the given term relative to the total number of up- or downregulated genes; ** *p*-value for the statistical significance of Gene Ontology term over-representation with the Bonferroni–Šidák correction. The table reports the ten most significant terms (those with the highest proportion of DEGs associated with the term relative to the total number of DEGs) describing biological processes for the upregulated and downregulated gene sets.

of DEGs using the DAVID web resource – namely, the lists of significantly overrepresented Gene Ontology terms and KEGG pathways – are presented in Tables S2 and S3. The ten most significant biological process terms (those with the highest proportion of DEGs associated with the term relative to the total number of DEGs) for the upregulated and downregulated gene sets are shown in Table 1.

For the genes with increased expression in malignantly transformed cells, significantly overrepresented terms were related to cell division (#1, #2 in Table 1), chromatin organization (#3 in Table 1), DNA repair and replication (#4, #5 in Table 1), mRNA splicing (#6 in Table 1), rRNA processing (#7 in Table 1), protein translation (#8, #9 in Table 1), and protein folding (#10 in Table 1). For the upregulated genes, KEGG pathways related to oxidative phosphorylation (hsa00190: Oxidative phosphorylation) and DNA replication (hsa03030: DNA replication) were significantly overrepresented (Table S2).

For the genes with decreased expression, significantly overrepresented terms described intracellular signal transduction (#1, #2 in Table 1), transcriptional regulation (#3–5 in Table 1), positive and negative regulation of apoptosis (#6–8 in Table 1), inflammation (#9 in Table 1), cell migration (#10 in Table 1), T-cell receptor signaling pathways (#10 in Table S3), and receptor tyrosine kinases (#11 in Table S3).

For the genes with increased expression, significantly overrepresented KEGG pathways included the MAPK signaling pathway (hsa04010), NF-κB signaling (hsa04064), chemokine signaling (hsa04062), and T-cell receptor signaling (hsa04660) (Table S3).

Gene network of DEGs involved in the apoptosis process according to Gene Ontology data

As described in the "Materials and methods" section, reconstruction of the gene network regulating apoptosis in hepatocytes during HCC development was carried out in two stages. Given the well-established importance of apoptosis in HCC (Hajizadeh et al., 2023; Ladd et al., 2024; Wu et al., 2024), as well as the over-representation of apoptosis-related processes among downregulated genes identified in our study (Table 1) in malignantly transformed hepatocytes, the first

stage incorporated into the gene network those genes and their protein products that, according to Gene Ontology, are involved in apoptosis and the expression of which in tumor hepatocytes differs from that in hepatocytes from histologically normal liver tissue. Of the 746 protein-coding genes (Table S4) annotated in The Gene Ontology Resource under the term "apoptotic process" (GO:0006915), 116 (16 % of all genes annotated to this term) were differentially expressed in malignantly transformed hepatocytes. Of these, 49 genes were upregulated and 67 genes were downregulated in tumor hepatocytes compared with healthy liver cells, accounting for 42.2 and 57.8 %, respectively, of the 116 apoptosis-related DEGs. The associative gene network reconstructed using ANDSystem (Fig. S1) comprised the 116 DEGs involved in apoptosis and their 116 protein products. Characteristics of this network are presented in Table 2 (column "Gene network, stage 1"); its visualization is shown in Fig. S1, and the full list of components (proteins and genes) is provided in Table S5.

At the second stage, to identify novel protein regulators of apoptosis during the malignant transformation of hepatocytes, the network reconstructed in stage one was expanded by adding the protein products of all DEGs revealed by the comparative analysis of transcriptomes from malignantly transformed hepatocytes and hepatocytes of histologically normal liver tissue. In expanding the network, we selected relationship types pertaining to gene expression regulation – expression regulation, expression upregulation, expression downregulation, and interaction. We found that, of the 116 apoptosis-related DEGs, the expression of 68 genes (59 %) is regulated by 223 proteins encoded by genes that are differentially expressed in tumor hepatocytes relative to normal liver tissue, but are not annotated in Gene Ontology as participating in apoptosis. The list of these genes is provided in Table S6. Of them, 102 genes were upregulated and 121 genes were downregulated.

According to functional annotation, the downregulated genes were significantly overrepresented (Bonferroni-adjusted p-value < 0.05) for biological processes including leukocyte cell–cell adhesion (GO:0007159), neutrophil chemotaxis (GO:0030593), cell division (GO:0051301), and positive regulation of the PI3K/Akt signaling pathway (GO:0051897).

Next, for the 223 candidate proteins potentially involved in regulating hepatocyte apoptosis during HCC development, we assessed the statistical significance of their specificity of association with the apoptosis regulatory gene network. For each protein, we calculated the probability that the observed fraction of its interactions with network genes relative to its total interactions with human protein-coding genes could arise by chance. As a result, 16 DEGs (11 downregulated and 5 upregulated) were identified as significantly associated (Bonferroni-adjusted p-value < 0.05) with 43 apoptosis genes (Table 3). As seen in Table 3, the products of *IL8*, *ERBB2*, *EGR1*, *TGFB2*, and *CDKN1A* have the highest numbers of links to DEGs already annotated in Gene Ontology as apoptosis participants. Proteins encoded by *CDN1A*, *ETS2*, *EGR1*, *BACH2*, *KLF5*, and *FEN1* are transcription factors according to The Human Transcription Factors database (Lambert et al., 2018; https://humantfs.ccbr.utoronto.ca/).

The final gene network of hepatocyte apoptosis in HCC is shown in Fig. S2, and its characteristics are presented in Table 2 (column "Gene network, stage 2"). The complete list of proteins and genes in the network is provided in Table S7. As seen in Table 2, upon expanding the initial apoptosis gene network with proteins that regulate the expression of apoptosis genes, the number of links of all types increased, with the exception of downregulation. Network hubs – that is, the nodes (genes or proteins), Network Connectivity (the number of other nodes connected to a given node) of which exceeded the critical (quantile) threshold corresponding to a p-value of 0.05 (see "Materials and methods") – are listed in Table 4.

A total of 11 network hubs were identified (Table 4), 10 of which are proteins, and one is the gene *MMP9*, the product of which also appears among the network hubs. According to scRNA-seq data (Table S1), the expression of genes encoding three proteins (CDK1, MMP9, G3P) was increased in malignantly transformed hepatocytes compared with hepatocytes from histologically normal liver tissue, whereas the expression of genes encoding the remaining seven proteins (NFKB1, BCL2, A4, CDKN1A, ERBB2, MCL1, FOXO1) was decreased. The genes encoding two network hubs – CDKN1A and ERBB2 – had not previously been annotated in Gene Ontology as participants in the apoptotic process.

Network of gene expression regulation involved in hepatocyte apoptosis during the development of hepatocellular carcinoma

Taking into account the scRNA-seq-identified changes in the expression of genes, the products of which are involved in hepatocyte apoptosis during HCC development, we analyzed gene expression regulation within the final apoptosis network. To this end, we filtered the edges of the reconstructed network, retaining only those proteins that either enhance (edge type "expression upregulation," Fig. 1) or suppress (edge type "expression downregulation," Fig. 2) the expression of genes comprising the final apoptosis regulatory network.

The expression-activation network (Fig. 1) comprised 38 proteins that activate the expression of 40 gene components of the apoptosis network. According to ANDSystem, NFKB1 activates the expression of 15 genes (including *BCL2*, *MCL1*, *CFLAR*, etc.), IL-8 activates 5 genes, ERBB2 activates 4 genes, and EGR1, SDF1, and TGFB2 each activate 3 genes; the remaining proteins in the expression-activation network regulate fewer than three apoptotic genes. In our scRNA-seq analysis (Table S1), both these regulators and their target genes exhibited decreased expression in malignantly transformed hepatocytes compared with hepatocytes from histologically normal liver tissue. By contrast, the matrix metalloproteinase gene *MMP9*, which was upregulated, is activated, according to ANDSystem, by

Table 2. Characteristics of associative networks of genes and proteins involved in apoptosis of hepatocytes in HCC

Parameter	Gene network	
	Stage 1	Stage 2
Number of network components	238	248
genes	116	116
proteins	116	132
Number of interactions	1,512	1,933
Of these, the following types of interactions:		
Gene expression		
Expression	116	116
Differential expression	2	2
Coexpression	7	7
Protein interactions		
Interaction	259	385
Catalyze	21	29
Cleavage	2	5
Modification	34	50
Regulatory interactions		
Regulation	85	95
Upregulation	69	79
Downregulation	24	24
Expression downregulation	122	134
Expression regulation	309	385
Expression upregulation	165	213
Activity downregulation	35	47
Activity regulation	60	79
Activity upregulation	25	35
Modification downregulation	15	17
Modification regulation	54	64
Modification upregulation	51	60
Degradation downregulation	9	12
Degradation regulation	17	34
Degradation upregulation	7	21
Transport regulation	24	40

five proteins (MEIN1, PPIA, TRIB3, CHK1, FEN1), the expression of which was also increased in tumor hepatocytes. In addition, CDC20, FEN1, KLF5, and their target genes showed increased expression.

The expression-repression network (Fig. 2) of genes involved in apoptosis in HCC comprised 15 proteins connected by "expression downregulation" type of interactions to 9 genes. According to ANDSystem, the expression of *MMP9* can be suppressed by five proteins (NFKB1, GELS, NR4A1, FOXO1, EGR1), the expression of which is reduced in malignantly transformed hepatocytes according to scRNA-seq, which may account for the elevated MMP9 expression observed in the scRNA-seq analysis. The expression of *BCL2*, which is decreased in tumor hepatocytes, can be suppressed by four proteins (CDK1, VDAC1, MMP9, CYC), the expression of which is increased in malignant

Table 3. List of proteins encoded by DEGs of malignantly transformed hepatocytes that are involved in the regulation of apoptosis in HCC but are not annotated in Gene Ontology as participants in apoptosis (GO:0006915, apoptotic process)

No.	Protein	Name of the protein	Number of interactions		Expression	*p*-value
			DEG	Total		
1	IL8	C-X-C motif chemokine ligand 8	10	25	Decreased	0.00000
2	ERBB2	Erb-b2 receptor tyrosine kinase 2	9	32		0.00025
3	**EGR1**	Early growth response 1	7	23		0.01576
4	**CDKN1A**	Cyclin dependent kinase inhibitor 1A	6	35		0.00004
5	TGFB2	Transforming growth factor beta 2	6	14		0.00861
6	**ETS2**	ETS proto-oncogene 2 transcription factor	5	8		0.00018
7	KLF5	KLF transcription factor 5	5	13	Increased	0.00196
8	SDF1	C-X-C motif chemokine ligand 12	5	15	Decreased	0.04080
9	GELS	Gelsolin	4	14		0.0012
10	K2C7	Keratin 7	3	3		0.00071
11	IMA1	Karyopherin subunit alpha 2	3	12	Increased	0.00198
12	**FEN1**	Flap structure-specific endonuclease 1	3	8		0.00576
13	NEP	Neprilysin	3	9	Decreased	0.00765
14	CDC20	Cell division cycle 20	3	19	Increased	0.01919
15	NEUT	Neurotensin	3	4		0.02776
16	**BACH2**	BTB domain and CNC homolog 2	3	7	Decreased	0.03851

Note. Number of interactions to apoptosis DEGs – the number of expression-regulatory links from the protein to genes involved in apoptosis according to Gene Ontology; Total number of links – the number of links from the protein to all components of the final gene network (genes and proteins); Expression – direction of the gene's expression change in tumor hepatocytes relative to normal cells (increased; decreased); *p*-value – statistical significance of the protein's association with apoptosis genes, computed using the hypergeometric test with the Bonferroni correction. Proteins are sorted in descending order of the significance of their association with the apoptosis network. Transcription factors are shown in bold, according to The Human Transcription Factors database (Lambert et al., 2018; https://humantfs.ccbr.utoronto.ca/).

Table 4. Hubs of the apoptosis gene network in hepatocytes in human hepatocellular carcinoma

No.	Object type in the network	Name of the object in the network	Protein/gene name	Number of linked network objects	*p*-value	Expression
1	Protein	NFKB1	Nuclear factor kappa B subunit 1	87	0.004	Decreased
2	Gene	MMP9	Matrix metallopeptidase 9	48	0.008	Increased
3	Protein	BCL2	BCL2 apoptosis regulator	46	0.012	Decreased
4	Protein	A4	Amyloid beta precursor protein	43	0.016	
5	Protein	**CDN1A**	Cyclin dependent kinase inhibitor 1A	35	0.020	
6	Protein	CDK1	Cyclin dependent kinase 1	33	0.024	Increased
7	Protein	**ERBB2**	Erb-b2 receptor tyrosine kinase 2	31	0.028	Decreased
8	Protein	MMP9	Matrix metallopeptidase 9	29	0.036	Increased
9	Protein	G3P	Glyceraldehyde-3-phosphate dehydrogenase	29	0.036	
10	Protein	MCL1	MCL1 apoptosis regulator, BCL2 family member	27	0.044	Decreased
11	Protein	FOXO1	Forkhead box O1	27	0.044	

Note. *p*-value – the critical threshold (quantile) calculated from the observed distribution of Network Connectivity across all nodes of the gene network. Proteins not previously annotated in Gene Ontology as participants in the apoptotic process are shown in bold.

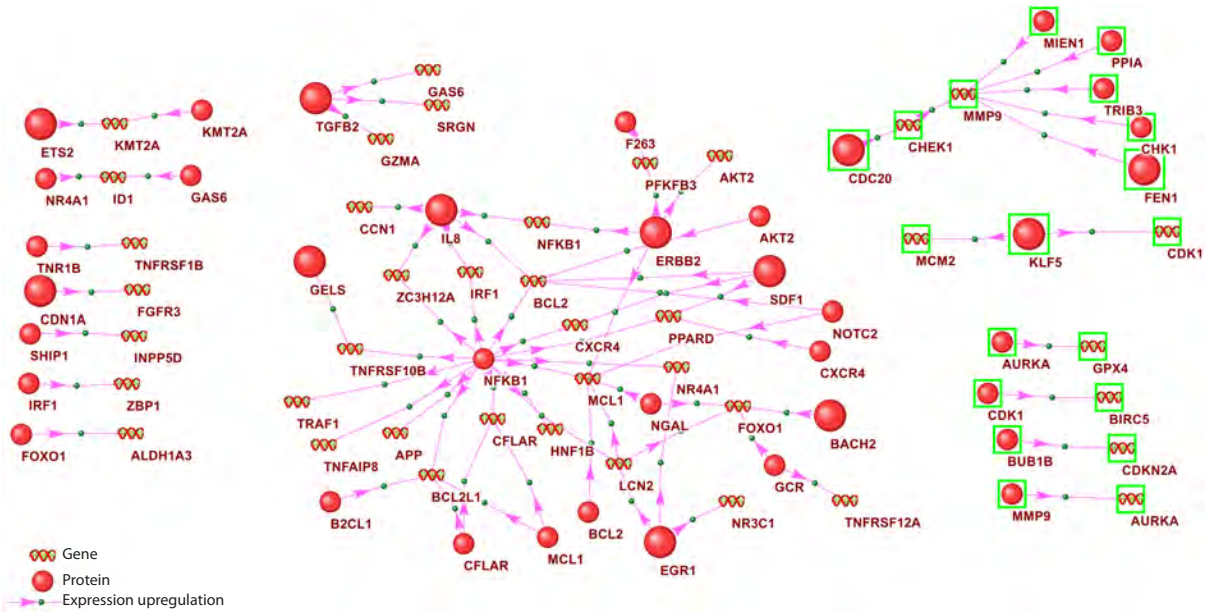


Fig. 1. Gene network of expression activation for gene components of the apoptosis regulatory network during HCC development.

Proteins and genes with increased expression are outlined in green; those with decreased expression are not outlined. Proteins that had not previously been annotated in Gene Ontology as participants in apoptosis are shown as larger circles. Shown are only the protein components of the hepatocyte apoptosis regulatory network in HCC (see Fig. S2) that activate (type of interaction – expression upregulation) the expression of gene components of the same network.

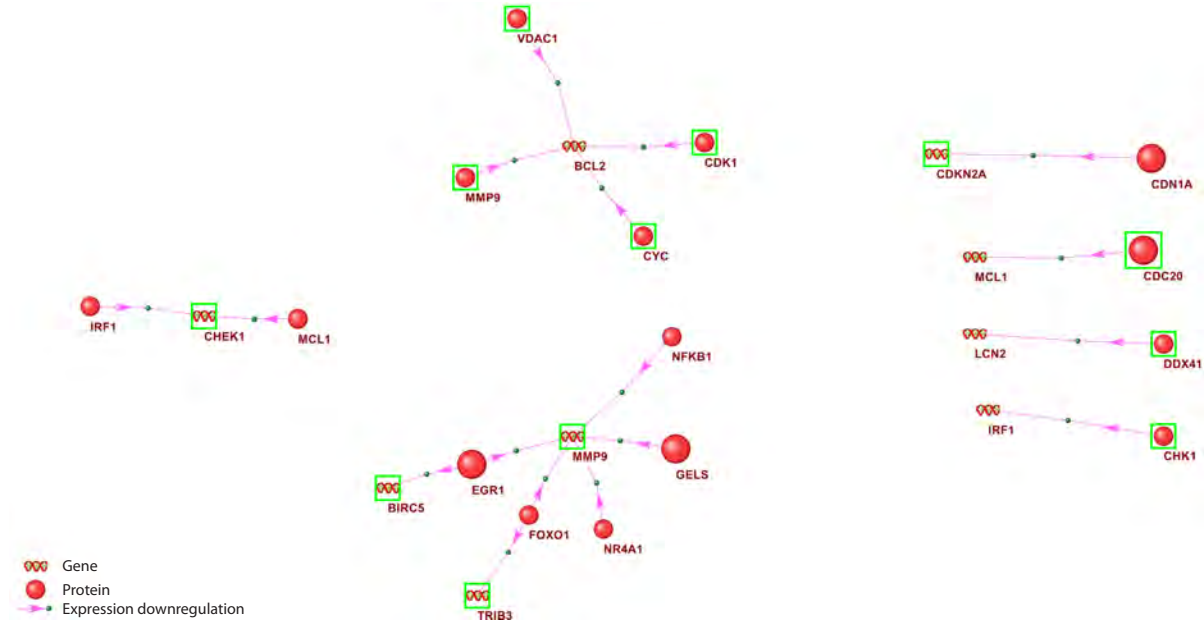


Fig. 2. Gene network of expression repression for gene components of the apoptosis regulatory network during HCC development.

Proteins and genes with increased expression are outlined in green; those with decreased expression are not outlined. Proteins not previously annotated in Gene Ontology as participants in apoptosis are shown as larger circles. Shown are only the protein components of the hepatocyte apoptosis regulatory network in HCC (see Fig. S2) that suppress (type of interaction – expression downregulation) the expression of gene components of the same network.

hepatocytes compared with hepatocytes from healthy liver tissue. Among the proteins involved in apoptosis regulation in HCC but not annotated in Gene Ontology as participants in this process, the expression-repression network included EGR1, CDN1A, GELS, and CDC20.

Phylostratigraphic analysis of the gene network

The analysis of the evolutionary age distribution of genes in the reconstructed apoptosis network in HCC is presented in Figure 3. The proportion of genes in the reconstructed

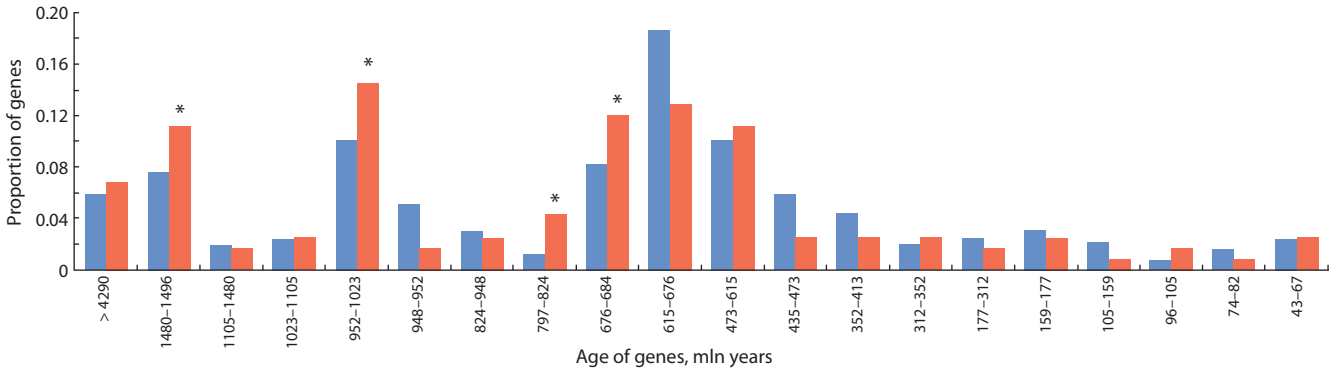


Fig. 3. Distribution of the evolutionary age of genes in the reconstructed hepatocyte apoptosis network during HCC development.

The X-axis shows gene age intervals (million years) according to the GenOrigin database; the Y-axis shows the proportion of genes in each interval. Blue bars indicate the distribution for the full set of human protein-coding genes; red bars indicate the distribution for genes in the reconstructed hepatocyte apoptosis network in HCC. * – denotes statistical significance of the difference in gene representation for a given age interval between the full set of human protein-coding genes and the reconstructed network.

apoptosis network was significantly higher ($p < 0.05$, hypergeometric test) than that among all human protein-coding genes in the following age intervals: (1) 1,480–1,496 million years, 13 genes; (2) 952–1,023 million years, 17 genes; (3) 797–824 million years, 5 genes; (4) 676–684 million years, 14 genes.

Discussion

Apoptosis is a tightly regulated and evolutionarily conserved program of cell death that performs key functions in normal physiological processes such as embryogenesis and tissue homeostasis in the adult organism. Resistance to apoptosis is a well-known hallmark of cancer cells that supports their survival and tumor growth (Kashyap et al., 2021). However, the literature also reports that apoptotic processes can be activated in tumor cells, especially at late stages of neoplasm development. Thus, although evasion of apoptosis is a well-established oncogenic mechanism (Moyer et al., 2025), tumor cell populations cannot continuously suppress the apoptotic program across all cells within a tumor (reviewed in Morana et al., 2022). This indicates specific features of apoptosis regulation during malignant progression that depend on tumor stage, tissue of origin, and cell type, given the well-known cellular heterogeneity of tumors (Li C. et al., 2020). Therefore, detailed investigation of the molecular genetic mechanisms of apoptosis in different types of malignancies – particularly HCC – at the single-cell level is required.

In the present study, using publicly available scRNA-seq data, we performed a comparative analysis of the transcriptomes of malignantly transformed hepatocytes and hepatocytes from histologically normal liver tissue, and we reconstructed the gene network regulating apoptosis in hepatocytes during human hepatocellular carcinoma. Analysis of the scRNA-seq data and gene expression regulation within the reconstructed network showed that expression of genes *NFKB1*, *BCL2*, and *MCL1* – network hubs (Table 4) – is reduced in malignant hepatocytes compared with healthy cells. The BCL2 and MCL1 proteins are known key inhibitors of apoptosis, as they prevent activation of BAX/BAK, which is required to increase mitochondrial membrane permeability and subsequently activate effector caspases (Newton et al., 2024). Upregulation of *BCL2* expression is considered one of the major mechanisms by which cells acquire resistance to apoptosis during malignant transformation (Moyer et al., 2025). However, in our study we observed decreased expression of *BCL2* and *MCL1* in HCC hepatocytes, which – according to analysis of the apoptosis regulatory network – may be due both to reduced expression of proteins that activate *BCL2* and *MCL1* expression (such as NF-κB, SDF1, ERBB, IL-8; Fig. 1) and to increased expression of proteins that suppress *BCL2* expression (Fig. 2).

It is noteworthy that NFKB1 is the principal hub of the hepatocyte apoptosis network in HCC (Table 4) and a key protein in the network that activates expression of genes involved in hepatocyte apoptosis (Fig. 2), which, according to ANDSystem, can activate a number of anti-apoptotic genes, including *BCL2* and *MCL1*. In tumors, activation of the NF-κB signaling pathway promotes survival by inhibiting apoptosis (Gupta et al., 2023); therefore, the decreased *NFKB1* expression found in our study (Tables S1 and 4) may plausibly increase hepatocyte susceptibility to apoptosis. On the other hand, activation of NFKB1 is reported to be necessary for apoptosis via the extrinsic pathway induced by chemokines – particularly IL1b (Wang P. et al., 2023) – and mediated by the TNFR1 receptor (Moyer et al., 2025). Thus, reduced *NFKB1* expression in malignantly transformed hepatocytes could, on the one hand, facilitate apoptosis of malignant hepatocytes by weakening expression of apoptosis inhibitors, but on the other hand hinder induction of extrinsic apoptosis, which requires NF-κB activation. In addition, our scRNA-seq analysis (Table S1) showed increased expression of genes encoding pro-apoptotic proteins in tumor hepatocytes, such as BID – a BAX/BAK activator (Moyer et al., 2025) – and FADD (FAS-associated death domain protein), a key component of the extrinsic apoptotic pathway (Nagata et al., 2017; Kashyap et al., 2021). One of the apoptosis network hubs, cyclin-dependent kinase 1 (CDK1), also shows increased gene expression in malignant hepatocytes

(Table S1). G. Massacci et al. (2023) demonstrated that CDK1 phosphorylates BCL2L1, BCL2, and MCL1, thereby suppressing their anti-apoptotic functions. However, that study also emphasized that the role of CDK1 in apoptosis regulation may depend on experimental context and cell-specific features.

Overall, the scRNA-seq data indicate decreased expression of key anti-apoptotic genes and increased expression of important pro-apoptotic genes in malignant hepatocytes compared with healthy hepatocytes. Our results suggest that, in the context of HCC, a reduction in anti-apoptotic protein levels is insufficient to trigger apoptosis. This, in turn, suggests that evasion of apoptosis by upregulating inhibitors of apoptosis is not the predominant mechanism of HCC progression, which may instead be driven by other causes likely related to the hepatocyte microenvironment – particularly dysregulation of inflammatory processes – as supported by scRNA-seq studies (Lu et al., 2022; Jiang S. et al., 2024). We also believe that activating pro-apoptotic effectors, such as caspases, should be a key therapeutic objective.

It is well known that NF-κB proteins are major regulators of inflammation, and increased expression stimulates the inflammatory response (Wang P. et al., 2023). Therefore, the reduced expression of the *NFKB1* gene, which encodes one member of this family, NFKB1, is consistent with the attenuated expression of genes involved in the inflammatory response in malignant hepatocytes, as indicated by the functional annotation of DEGs (Table 1).

A search for regulatory links between the DEGs controlling hepatocyte apoptosis in HCC and proteins – the products of other DEGs identified by scRNA-seq – allowed us to identify more than 200 proteins (Table S6) that could potentially modulate the expression of genes governing hepatocyte apoptosis during HCC, even though they are not annotated in Gene Ontology as regulators of this process. Notably, functional annotation of the genes encoding these proteins revealed in tumor cells a reduced expression of genes, the products of which support leukocyte migration and adhesion – chemokines (*CCL5*, *CXCL2*, *CXCL8*, *CXCL1*), transforming growth factor-β2 (*TGFB2*), the tyrosine kinase *SYK*, and integrin *ITGA4*. However, according to ANDSystem, these same proteins can regulate key nodes of the hepatocyte apoptosis regulatory network. In particular, CCL5 induces expression of matrix metalloproteinase 9 (*MMP9*) (Sevenich, Joyce, 2014), which is one of the principal hubs of the reconstructed apoptosis regulatory network in HCC hepatocytes. MMP9 is a member of the multifunctional family of zinc-dependent endopeptidases and is activated during inflammation and in certain cancers. Matrix metalloproteinases cleave extracellular matrix proteins and play crucial roles in cellular apoptosis, angiogenesis, tumor growth, and metastasis (Verma et al., 2015). MMP9 is known to be capable of inducing apoptosis (Liang et al., 2019). These findings indicate that reduced expression of genes encoding key immune defense components may promote tumor progression not only by weakening the immune response to transformed cells but also by influencing apoptotic processes within them.

At the same time, the previously proposed statistical approach (Yatsyk et al., 2025) for assessing the significance of a given protein's or gene's association with a network of interest (in this case, apoptosis), together with analysis of the reconstructed network, enabled us to prioritize several proteins – potential participants in the regulation of the apoptotic process in hepatocytes – the altered expression of which is likely to disrupt apoptosis regulation in hepatocytes and thereby contribute to the onset and progression of HCC. These proteins (ERBB2, CDN1A, IL8, EGR1) are significantly associated with the hepatocyte apoptosis regulatory network in HCC and act as central regulators (hubs) influencing a large number (>20) of its nodes.

The ERBB family of erythroblastic leukemia viral oncogene homologs, which includes the epidermal growth factor receptor (EGFR) and ERBB2, ERBB3, and ERBB4, regulates a broad range of essential cellular functions, such as survival, growth, and migration of tumor cells, and has therefore attracted attention as a therapeutic target in cancer (Chen et al., 2024). ERBB2, a member of this family, the expression of which was reduced in malignant hepatocytes according to scRNA-seq, has not previously been annotated as involved in apoptosis regulation, yet it emerged as a statistically significant hub of the reconstructed apoptosis regulatory network (Table 4). Elevated *ERBB2* expression is associated with breast tumor growth, and suppression of *ERBB2* and *ERBB3* induces apoptosis in breast cancer cells (Xiang et al., 2010). Although there are no data on the role of ERBB2 in apoptosis induction in HCC, our network analysis indicates that this protein regulates several apoptosis-related proteins and genes in HCC, including NFKB1, AKT2, CDK1, MCL1, and FOXO1. In particular, ERBB2 has been shown to phosphorylate cyclin-dependent kinase CDK1, increasing the resistance of cancer cells to apoptosis induced by the cytostatic anticancer drug paclitaxel (Vahedi et al., 2015). ERBB2 also appears to activate expression of the anti-apoptotic genes *NFKB1*, *AKT2*, and *MCL1* (Fig. 1), which are downregulated in malignant hepatocytes according to our scRNA-seq data. Thus, ERBB2 is an important potential node in the regulation of apoptosis in hepatocytes, and changes in its expression may contribute to HCC development.

IL-8, also known as CXCL8, is a pro-inflammatory chemokine of the CXC family. Elevated IL-8 levels are associated with poor prognosis across various cancers, including hepatocellular carcinoma. In HCC, increased *IL-8* expression is also linked to enhanced metastatic potential of tumor cells (Han et al., 2023). Choi et al. (2016) showed that *IL-8* knockdown promoted apoptosis in HCC cells.

CDN1A (also known as CDKN1A), cyclin-dependent kinase inhibitor 1A encoded by the *CDKN1A* gene, has not previously been annotated in Gene Ontology as a protein involved in apoptosis; however, its role in apoptosis during HCC development has been discussed in the literature (Thangavelu et al., 2024). Reports emphasize that the role of CDN1A in regulating apoptosis during tumorigenesis is context-dependent, as CDKN1A can both suppress and promote apoptosis (Manu et al., 2019). Experimental data indicate

that CDKN1A is a p53 target and can stimulate apoptosis in tumor cells by activating the TNF receptor or the pro-apoptotic protein BAX, or by modulating the intrinsic apoptotic pathway via changes in mitochondrial membrane permeability (Abbas, Dutta, 2009). The natural compound N-trans-feruloyloctopamine can enhance apoptosis of HCC cells through its interaction with CDKN1A (Ma et al., 2021).

ANDSystem data indicate that this protein is one of the central nodes of the apoptosis regulatory network in hepatocytes during HCC development. It interacts with other network hubs, in particular with well-known apoptosis regulators such as NFKB1, BCL2, and CDK1. However, scRNA-seq analysis showed that *CDKN1A* expression was reduced in tumor hepatocytes compared with normal liver cells (Table 4). These findings suggest that attenuation of CDKN1A expression in hepatocytes may represent an important link in HCC pathogenesis, facilitating tumor-cell evasion of apoptosis; nevertheless, its role in hepatocyte apoptosis regulation in HCC requires further experimental investigation.

Early growth response protein 1 (EGR1) suppresses proliferation and enhances apoptosis of malignantly transformed cells in many tissues and organs, including the liver (reviewed in Wang B. et al., 2021). It has also been shown that EGR1 can inhibit HCC growth by repressing transcription of *PFKL* (phosphofructokinase-1, liver type) and by inhibiting aerobic glycolysis in tumor cells (Pan et al., 2024). In our study, EGR1, the expression of which is reduced, acts as an activator of genes (*LCN2*, *NR3C1*, *NR4A1*; Fig. 1) involved in apoptosis control, the expression of which is likewise reduced in malignant hepatocytes. Our results suggest that decreased *EGR1* expression may be one of the mechanisms underlying weakened apoptosis during malignant transformation.

The use of phylostratigraphic analysis to assess gene evolutionary age is important for studying the evolution of gene networks and identifying their key components (Mustafin et al., 2021). Notably, most genes in the hepatocyte apoptosis network and those in the overrepresented age intervals are older than 600 million years (Fig. 3), whereas relatively young genes are scarce, indicating evolutionary conservation of the network genes and their importance for cellular viability. In particular, the overrepresented group of genes aged 1,480–1,496 million years corresponds to the period of mitochondrial–eukaryotic cell symbiosis (Raval et al., 2023). During these stages of symbiosis, many genes responsible for mitochondrial programmed cell death evolved, including key factors regulating cytochrome *c* release and oxidative stress control – early adaptations that maintained symbiotic balance (Zmasek, Godzik, 2013). Moreover, we found a statistically significant excess of genes in the hepatocyte apoptosis network, relative to the human genome as a whole, within the 952–1,023-million-year interval. This interval includes, in particular, proteins such as BCL2 – a network hub – and BCL2L1. These proteins are well-known key inhibitors of apoptosis (Moyer et al., 2025). Orthologs of BCL2 family genes are found in sponges (*Porifera*), placozoans (*Placozoa*), and hydras (*Hydra*) (Banjara et al., 2020), i.e., at a relatively early stage of metazoan evolution. The critical role of apoptosis in innate and adaptive immunity suggests that this function arose early in the evolution of multicellularity and likely preceded the adaptation of apoptosis to other processes – such as development, homeostasis, and removal of damaged cells in *Metazoa* – laying the groundwork for complex multicellular life (Suraweera et al., 2022). Thus, changes in hepatocyte gene expression during HCC involve highly conserved genes – including the network hub *BCL2* – that, beyond apoptosis, may regulate other cellular processes, underscoring the complexity of regulatory interactions during malignant transformation.

Accordingly, our study – using an integrated approach that included hepatocyte transcriptome analysis and reconstruction/analysis of a DEG network involved in apoptosis – provides new insights into the regulation of hepatocyte apoptosis during human HCC development. Our findings, which show decreased expression of key apoptosis inhibitor genes, support the view that evasion of apoptosis is not invariably characteristic of cancer cells and that the role of apoptosis in tumor development depends on the cell type, tissue context, and tumor microenvironment (Morana et al., 2022). In addition, reduced expression in malignant hepatocytes of genes involved in inflammatory control, together with decreased NFKB1 – a central regulator of inflammation (Wang P. et al., 2023) – points to an important role for interactions between hepatocytes and the immune system in HCC development, warranting further experimental and theoretical investigation. The identified network hubs (*NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*) may serve as useful targets for modulating apoptosis in hepatocytes in HCC therapy, an increasingly promising direction (Ladd et al., 2024; Wu et al., 2024).

Conclusion

Analysis of scRNA-seq data from normal and malignantly transformed hepatocytes revealed changes in the expression of genes involved in the control of hepatocyte apoptosis in HCC. In malignant hepatocytes, expression of the key apoptosis inhibitors *BCL2* and *MCL1* was decreased, as was the expression of genes involved in the inflammatory response. These findings indicate that evasion of apoptosis by upregulating key apoptosis inhibitors does not appear to be a characteristic feature of hepatocytes during HCC development. Reconstruction and analysis of the hepatocyte apoptosis – regulatory network in HCC showed that reduced expression of *NFKB1* may be an important factor underlying the decreased expression of a range of apoptosis-related genes, including *BCL2* and *MCL1*. In addition, network reconstruction and analysis identified several key genes (*NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*) that both display differential expression in malignant versus healthy hepatocytes and function as hubs of the hepatocyte apoptosis network in HCC. Dysregulated expression of these genes may lead to apoptosis dysregulation in tumor cells.

Among the DEGs, we also identified genes (*CDKN1A*, *ERBB2*, *IL8*, *EGR1*) that, although not annotated in Gene Ontology as apoptosis participants, exhibited numbers of regulatory interactions of their products with apoptosis genes that significantly exceeded chance expectations according to a hypergeometric test. This suggests that the proteins encoded by these genes play specific roles in regulating hepatocyte apoptosis in HCC and represent promising candidates for further investigation.

The results obtained can be used to guide future experimental studies on the regulation of hepatocyte apoptosis in HCC. The hypotheses proposed may facilitate the development of targeted therapeutic strategies aimed at modulating programmed cell death in malignant liver cells.

References

Abbas T., Dutta A. p21 in cancer: intricate networks and multiple activities. *Nat Rev Cancer*. 2009;9(6):400-414. doi 10.1038/nrc2657

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Volianskaia A.R., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Kolchanov N.A., Lavrik I.N., Ivanisenko V.A. Reconstruction of the regulatory hypermethylation network controlling hepatocellular carcinoma development during hepatitis C viral infection. *J Integr Bioinform*. 2023;20(3): 20230013. doi 10.1515/jib-2023-0013

Banjara S., Suraweera C.D., Hinds M.G., Kvansakul M. The Bcl-2 family: Ancient origins, conserved structures, and divergent mechanisms. *Biomolecules*. 2020;10(1):128. doi 10.3390/biom10010128

Butikova E.A., Basov N.V., Rogachev A.D., Gaisler E.V., Ivanisenko V.A., Demenkov P.S., Makarova A.A., … Pokrovsky A.G., Vinokurov N.A., Kanygin V.V., Popik V.M., Shevchenko O.A. Metabolomic and gene networks approaches reveal the role of mitochondrial membrane proteins in response of human melanoma cells to THz radiation. *Biochim Biophys Acta Mol Cell Biol Lipids*. 2025; 1870(2):159595. doi 10.1016/j.bbalip.2025.159595

Chen Y., Lu A., Hu Z., Li J., Lu J. ERBB3 targeting: A promising approach to overcoming cancer therapeutic resistance. *Cancer Lett*. 2024;599:217146. doi 10.1016/j.canlet.2024.217146

Choi S.H., Park J.Y., Kang W., Kim S.U., Kim Y., Ahn S.H., Ro S.W., Han K.H. Knockdown of HIF-1α and IL-8 induced apoptosis of hepatocellular carcinoma triggers apoptosis of vascular endothelial cells. *Apoptosis*. 2016;21(1):85-95. doi 10.1007/s10495-015-1185-2

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2011;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Demenkov P.S., Antropova E.A., Adamovskaya A.V., Mishchenko E.L., Khlebodarova T.M., Ivanisenko T.V., Ivanisenko N.V., Venzel A.S., Lavrik I.N., Ivanisenko V.A. Prioritization of potential pharmacological targets for the development of anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway: the reconstruction and analysis of associative gene networks help. *Vavilov J Genet Breed*. 2023;27(7):784-793. doi 10.18699/VJGB-23-91

Dhanasekaran R., Bandoh S., Roberts L.R. Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances. *F1000Res*. 2016;5:879. doi 10.12688/f1000research.6946.1

Fabregat I. Dysregulation of apoptosis in hepatocellular carcinoma cells. *World J Gastroenterol*. 2009;15(5):513-520. doi 10.3748/wjg.15.513

Ganesan P., Kulik L.M. Hepatocellular carcinoma: New developments. *Clin Liver Dis*. 2023;27(1):85-102. doi 10.1016/j.cld.2022.08.004

Gupta R., Kadhim M.M., Turki Jalil A., Obayes A.M., Aminov Z., Alsaikhan F., Ramírez-Coronel A.A., Ramaiah P., Tayyib N.A., Luo X. Multifaceted role of NF-κB in hepatocellular carcinoma therapy: Molecular landscape, therapeutic compounds and nanomaterial approaches. *Environ Res*. 2023;228:115767. doi 10.1016/j.envres.2023.115767

Hajizadeh M., Hajizadeh F., Ghaffarei S., Amin Doustvandi M., Hajizadeh K., Yaghoubi S.M., Mohammadnejad F., Khiabani N.A., Mousavi P., Baradaran B. MicroRNAs and their vital role in apoptosis in hepatocellular carcinoma: miRNA-based diagnostic and treatment methods. *Gene*. 2023;888:147803. doi 10.1016/j.gene.2023.147803

Han X., Wu J., Sha Z., Lai R., Shi J., Mi L., Yin F., Guo Z. Dicer suppresses hepatocellular carcinoma via interleukin-8 pathway. *Clin Med Insights Oncol*. 2023;17:11795549231161212. doi 10.1177/11795549231161212

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved ai-based short names recognition. *Int J Mol Sci*. 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019; 20(Suppl. 1):34. doi 10.1186/S12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci Rep*. 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci*. 2024;25(21): 11847. doi 10.3390/ijms252111847

Jiang S., Lu H., Pan Y., Yang A., Aikemu A., Li H., Hao R., Huang Q., Qi X., Tao Z., Wu Y., Quan C., Zhou G., Lu Y. Characterization of the distinct immune microenvironments between hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *Cancer Lett*. 2024; 588:216799. doi 10.1016/j.canlet.2024.216799

Jiang Y., Han Q.J., Zhang J. Hepatocellular carcinoma: Mechanisms of progression and immunotherapy. *World J Gastroenterol*. 2019; 25(25):3151-3167. doi 10.3748/wjg.v25.i25.3151

Kashyap D., Garg V.K., Goel N. Intrinsic and extrinsic pathways of apoptosis: Role in cancer development and prognosis. *Adv Protein Chem Struct Biol*. 2021;125:73-120. doi 10.1016/bs.apcsb.2021.01.003

Khlebodarova T.M., Demenkov P.S., Ivanisenko T.V., Antropova E.A., Lavrik I.N., Ivanisenko V.A. Primary and secondary micro-RNA modulation the extrinsic pathway of apoptosis in hepatocellular carcinoma. *Mol Biol*. 2023;57(2):165-175. doi 10.1134/S0026893323020103

Ladd A.D., Duarte S., Sahin I., Zarrinpar A. Mechanisms of drug resistance in HCC. *Hepatology*. 2024;79(4):926-940. doi 10.1097/HEP.0000000000000237

Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell*. 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029

Larina I.M., Pastushkova L.Kh., Tiys E.S., Kireev K.S., Kononikhin A.S., Starodubtseva N.L., Popov I.A., Custaud M.A., Dobrokhotov I.V., Nikolaev E.N., Kolchanov N.A., Ivanisenko V.A. Permanent proteins in the urine of healthy humans during the Mars-500 experiment. *J Bioinform Comput Biol*. 2015;13(1):1540001. doi 10.1142/S0219720015400016

Li C., Xu J. Identification of potentially therapeutic target genes of hepatocellular carcinoma. *Int J Environ Res Public Health*. 2020;17(3):1053. doi 10.3390/ijerph17031053

Li X.Y., Shen Y., Zhang L., Guo X., Wu J. Understanding initiation and progression of hepatocellular carcinoma through single cell sequencing. *Biochim Biophys Acta Rev Cancer.* 2022;1877(3):188720. doi 10.1016/j.bbcan.2022.188720

Liang Y., Yang C., Lin Y., Parviz Y., Sun K., Wang W., Ren M., Yan L. Matrix metalloproteinase 9 induces keratinocyte apoptosis through FasL/Fas pathway in diabetic wound. *Apoptosis*. 2019;24(7-8):542-551. doi 10.1007/s10495-019-01536-w

Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi 10.1186/s13059-014-0550-8

Lu Y., Yang A., Quan C., Pan Y., Zhang H., Li Y., Gao C., Lu H., Wang X., Cao P., Chen H., Lu S., Zhou G. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat Commun*. 2022;13(1):4594. doi 10.1038/s41467-022-32283-3

Ma B., Li J., Yang W.K., Zhang M.G., Xie X.D., Bai Z.T. N-*trans*-feruloyloctopamine wakes up BBC3, DDIT3, CDKN1A, and NOXA signals to accelerate HCC cell apoptosis. *Anal Cell Pathol* (*Amst*). 2021;2021:1560307. doi 10.1155/2021/1560307

Manu K.A., Cao P.H.A., Chai T.F., Casey P.J., Wang M. p21cip1/waf1 coordinate autophagy, proliferation and apoptosis in response to metabolic stress. *Cancers* (*Basel*). 2019;11(8):1112. doi 10.3390/cancers11081112

Massacci G., Perfetto L., Sacco F. The Cyclin-dependent kinase 1: more than a cell cycle regulator. *Br J Cancer.* 2023;129(11):1707-1716. doi 10.1038/s41416-023-02468-8

Momynaliev K.T., Kashin S.V., Chelysheva V.V., Selezneva O.V., Demina I.A., Serebryakova M.V., Alexeev D., Ivanisenko V.A., Aman E., Govorun V.M. Functional divergence of *Helicobacter pylori* related to early gastric cancer. *J Proteome Res*. 2010;9(1):254-267. doi 10.1021/pr900586w

Morana O., Wood W., Gregory C.D. The apoptosis paradox in cancer. *Int J Mol Sci*. 2022;23(3):1328. doi 10.3390/ijms23031328

Moyer A., Tanaka K., Cheng E.H. Apoptosis in cancer biology and therapy. *Annu Rev Pathol*. 2025;20(1):303-328. doi 10.1146/annurev-pathmechdis-051222-115023

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nagata S., Tanaka M. Programmed cell death and the immune system. *Nat Rev Immunol*. 2017;17(5):333-340. doi 10.1038/nri.2016.153

Narkevich A.N., Vinogradov K.A., Grjibovski A.M. Multiple comparisons in biomedical research: the problem and its solutions. *Hum Ecol*. 2020;10:55-64. doi 10.33396/1728-0869-2020-10-55-64 (in Russian)

Newton K., Strasser A., Kayagaki N., Dixit V.M. Cell death. *Cell*. 2024;187(2):235-256. doi 10.1016/j.cell.2023.11.044

Ogunwobi O.O., Harricharran T., Huaman J., Galuza A., Odumuwagun O., Tan Y., Ma G.X., Nguyen M.T. Mechanisms of hepatocellular carcinoma progression. *World J Gastroenterol*. 2019;25(19):2279-2293. doi 10.3748/wjg.v25.i19.2279

Pan M., Luo M., Liu L., Chen Y., Cheng Z., Wang K., Huang L., Tang N., Qiu J., Huang A., Xia J. EGR1 suppresses HCC growth and aerobic glycolysis by transcriptionally downregulating PFKL. *J Exp Clin Cancer Res*. 2024;43(1):35. doi 10.1186/s13046-024-02957-5

Raval P.K., Martin W.F., Gould S.B. Mitochondrial evolution: Gene shuffling, endosymbiosis, and signaling. *Sci Adv*. 2023;9(32):p.eadj4493. doi 10.1126/sciadv.adj4493

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics*. 2019;12(Suppl. 2):47. doi 10.1186/s12920-019-0492-9

Sevenich L., Joyce J.A. Pericellular proteolysis in cancer. *Genes Dev*. 2014;28(21):2331-2347. doi 10.1101/gad.250647.114

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Si-Tayeb K., Noto F.K., Nagaoka M., Nagaoka M., Li J., Battle M.A., Duris C., North P.E., Dalton S., Duncan S.A. Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology*. 2010;51(1):297-305. doi 10.1002/hep.23354

Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*. 1967;62(318):626-633. doi 10.2307/2283989

Squair J.W., Gautier M., Kathe C., Anderson M.A., James N.D., Hutson T.H., Hudelle R., … Barraud Q., Levine A.J., La Manno G., Skinnider M.A., Courtine G. Confronting false discoveries in single-cell differential expression. *Nat Commun*. 2021;12(1):5692. doi 10.1038/s41467-021-25960-2

Suraweera C.D., Banjara S., Hinds M.G., Kvansakul M. Metazoans and intrinsic apoptosis: An evolutionary analysis of the Bcl-2 family. *Int J Mol Sci*. 2022;23(7):3691. doi 10.3390/ijms23073691

Thangavelu L., Altamimi A.S.A., Ghaboura N., Babu M.A., Roopashree R., Sharma P., Pal P., Choudhary C., Prasad G.V.S., Sinha A., Balaraman A.K., Rawat S. Targeting the p53-p21 axis in liver cancer: Linking cellular senescence to tumor suppression and progression. *Pathol Res Pract*. 2024;263:155652. doi 10.1016/j.prp.2024.155652

Tong Y.-B., Shi M.-W., Qian S.H., Chen Y.-J., Luo Z.-H., Tu Y.-X., Xiong Y.-L., Geng Y.-J., Chen C., Chen Z.-X. GenOrigin: a comprehensive protein-coding gene origination database on the evolutionary timescale of life. *J Genet Genomics.* 2021;48(12):1122-1129. doi 10.1016/j.jgg.2021.03.018

Traag V.A., Waltman L., van Eck N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9:5233. doi 10.1038/s41598-019-41695-z

Vahedi S., Chueh F.Y., Dutta S., Chandran B., Yu C.L. Nuclear lymphocyte-specific protein tyrosine kinase and its interaction with CR6-interacting factor 1 promote the survival of human leukemic T cells. *Oncol Rep*. 2015;34(1):43-50. doi 10.3892/or.2015.3990

Verma S., Kesh K., Gupta A., Swarnakar S. An overview of matrix metalloproteinase 9 polymorphism and gastric cancer risk. *Asian Pac J Cancer Prev*. 2015;16(17):7393-7400. doi 10.7314/apjcp.2015.16.17.7393

Wang B., Guo H., Yu H., Chen Y., Xu H., Zhao G. The role of the transcription factor EGR1 in cancer. *Front Oncol*. 2021;11:642547. doi 10.3389/fonc.2021.642547

Wang P., Qian H., Xiao M., Lv J. Role of signal transduction pathways in IL-1β-induced apoptosis: Pathological and therapeutic aspects. *Immun Inflamm Dis*. 2023;11(1):e762. doi 10.1002/iid3.762

Wolf F., Angerer P., Theis F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi 10.1186/s13059-017-1382-0

Wu X., Cao J., Wan X., Du S. Programmed cell death in hepatocellular carcinoma: mechanisms and therapeutic prospects. *Cell Death Discov.* 2024;10(1):356. doi 10.1038/s41420-024-02116-x

Xiang S., Sun Z., He Q., Yan F., Wang Y., Zhang J. Aspirin inhibits ErbB2 to induce apoptosis in cervical cancer cells. *Med Oncol*. 2010;27(2):379-387. doi 10.1007/s12032-009-9221-0

Xue R., Zhang Q., Cao Q., Kong R., Xiang X., Liu H., Feng M., … Zhan Q., Deng M., Zhu J., Zhang Z., Zhang N. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. *Nature*. 2022;612(7938):141-147. doi 10.1038/s41586-022-05400-x

Yang Y.M., Kim S.Y., Seki E. Inflammation and liver cancer: molecular mechanisms and therapeutic targets. *Semin Liver Dis*. 2019;39(1):26-42. doi 10.1055/s-0038-1676806

Yatsyk I.V., Volyanskaya A.R, Kleshchev M.A., Antropova E.A., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Reconstruction and computational analysis of the insulin response gene network in human type 2 diabetes using transcriptomic data. *Gene Expr*. 2025 (в печати)

Zhang Q.Y., Ho D.W., Tsui Y.M., Ng I.O. Single-cell transcriptomics of liver cancer: hype or insights? *Cell Mol Gastroenterol Hepatol*. 2022;14(3):513-525. doi 10.1016/j.jcmgh.2022.04.014

Zmasek C.M., Godzik A. Evolution of the animal apoptosis network. *Cold Spring Harb Perspect Biol*. 2013;5(3):a008649. doi 10.1101/cshperspect.a008649

Zou Y., Wan X., Zhou Q., Zhu G., Lin S., Tang Q., Yang X., Wang S. Mechanisms of drug resistance in hepatocellular carcinoma. *Biol Proced Online*. 2025;27(1):19. doi 10.1186/s12575-025-00281-6

Conflict of interest. The authors declare no conflict of interest.
Received July 16, 2025. Revised September 13, 2025. Accepted September 15, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):978-989

doi 10.18699/vjgb-25-103

Hedgehog signaling in humans: the HH_Signal pathway_db knowledge base

T.A. Bukharina [1, 2], A.M. Bondarenko[2], D.P. Furman[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
bukharina@bionet.nsc.ru; furman@bionet.nsc.ru

Abstract. The rapid advancement of omics technologies (genomics, transcriptomics, proteomics, metabolomics) and other high-throughput methods for experimental studies of molecular genetic systems and processes has led to the generation of an unprecedentedly vast amount of heterogeneous and complex biological data. Effective use of this information resource requires systematic approaches to its analysis. One such approach involves the creation of domain-specific knowledge/data repositories that integrate information from multiple sources. This not only enables the storage and structuring of heterogeneous data distributed across various resources but also facilitates the acquisition of new insights into biological systems and processes. A systematic approach is also critical to solving the fundamental problem of biology – clarifying the regularities of morphogenesis. Morphogenesis is regulated through evolutionarily conserved signaling pathways (Hedgehog, Wnt, Notch, etc.). The Hedgehog (HH) pathway plays a key role in this process, as it begins functioning earlier than others in ontogenesis and determines the progression of every stage of an organism's life cycle: from structuring embryonic primordia, histo- and organogenesis, to maintaining tissue homeostasis and regeneration in adults. Our work presents HH_Signal_pathway_db, a knowledge base that integrates curated data on the molecular components and functional roles of the human Hedgehog (HH) signaling pathway. The first release of the database (available upon request at bukharina@bionet.nsc.ru) contains information on 56 genes, their protein products, the regulatory interaction network, and established associations with pathological conditions in humans. HH_Signal_pathway_db provides researchers with a tool for gaining new knowledge about the role of the Hedgehog pathway in health and disease, and its potential applications in developmental biology and translational medicine.

Key words: knowledge base; Hedgehog signaling pathway; morphogenesis; evolution; gene networks; regulatory circuits

For citation: Bukharina T.A., Bondarenko A.M., Furman D.P. Hedgehog signaling in humans: description in the HH_Signal_pathway_db knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):978-989. doi 10.18699/vjgb-25-103

Funding. This work was supported by the budget project FWNR-2022-0020.

Acknowledgements. The authors express their sincere gratitude to Academician N.A. Kolchanov for his interest in the work and fruitful discussion, to R.A. Ivanov and D. Sci. S.A. Lashin for their assistance in determining the evolutionary characteristics of the Hedgehog signaling pathway genes; to N.L. Podkolodny for providing data on the affinity of the TBP protein for gene promoters; and to PhD V.A. Ivanisenko and his colleagues I.V. Yatsyk, and A.V. Adamovskaya for their help with the gene network construction software training.

Сигнальный путь Hedgehog у человека: описание в базе знаний HH_Signal_pathway_db

Т.А. Бухарина [1, 2], А.М. Бондаренко[2], Д.П. Фурман[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
bukharina@bionet.nsc.ru; furman@bionet.nsc.ru

Аннотация. Стремительное развитие омиксных технологий (геномики, транскриптомики, протеомики, метаболомики) и других высокопроизводительных методов экспериментального исследования молекулярно-генетических систем и процессов привело к генерации беспрецедентно огромных объемов разнородных и сложных биологических данных. Эффективное использование этого информационного ресурса требует системных подходов к их анализу. Один из подходов состоит в создании предметно-ориентированных баз знаний/данных – репозиториев, интегрирующих информацию из множества источников, что позволяет не только хранить и структурировать распределенные по различным источникам гетерогенные данные, но и получать новые сведения о биологических системах и процессах. Критически важен системный подход и к решению фундаментальной

задачи биологии – выяснению закономерностей морфогенеза. Регуляция морфогенеза осуществляется через эволюционно консервативные сигнальные пути (Hedgehog, Wnt, Notch и др.). Ключевая роль в этом процессе принадлежит пути Hedgehog (HH), поскольку в онтогенезе он начинает функционировать ранее других и детерминирует реализацию каждого этапа индивидуального развития организма: от структурирования эмбриональных зачатков, гисто- и органогенеза до поддержания тканевого гомеостаза и процесса регенерации у взрослых особей. Нами создана база знаний HH_Signal_pathway_db, в которую сведена информация о компонентах и функциях HH сигнального пути у человека. Первый релиз базы (доступен по запросу bukharina@bionet.nsc.ru) содержит информацию о входящих в него 56 генах, их белковых продуктах, сети регуляторных взаимодействий, а также об установленных связях с некоторыми патологическими состояниями человека. HH_Signal_pathway_db предоставляет исследователям инструмент для получения новых знаний о роли пути Hedgehog в норме и при патологии и возможностях применения их в области биологии развития и трансляционной медицины.
Ключевые слова: база знаний; сигнальный путь Hedgehog; морфогенез; эволюция; генные сети; регуляторные контуры

Introduction

Modern molecular-genetic and biomedical studies using advanced techniques generate vast amounts of heterogeneous information (Regev et al., 2017; Schermelleh et al., 2019, Kenneth, 2022). This includes data obtained during investigations of various aspects of morphogenesis – a fundamental process leading to the formation of intricate organism architecture. Understanding the mechanisms underlying morphogenesis is essential not only for answering one of biology's most profound questions – how a single cell gives rise to a highly complex, spatially organized multicellular organism – but also for explaining the mechanisms of tissue regeneration, the causes of congenital anomalies, and pathological conditions of various etiologies, including oncological diseases.

Numerous genes, proteins, miRNAs, and signaling molecules are involved in regulating morphogenesis (ENCODE Project Consortium, 2012; Briscoe, Thérond, 2013; Bartel, 2018; Ghafouri-Fard et al., 2022; McIntyre et al., 2024). Some of these components belong to specific signaling pathways.

Signaling pathways (signal transduction) act as transmitter of signals received at the external cell membrane into the nucleus. Cascades of intermolecular interactions involving ligands, receptors recognizing those ligands, intracellular signal transducers of both protein and non-protein nature, transcription factors and co-regulators, etc., mediate pathways. The outcome of pathways' activity is alteration of target gene expression and corresponding protein levels, which ultimately leads to changes in the functional state of the cell.

Signaling pathways in animals and humans are evolutionarily conserved, and their roles are similar across different taxonomic groups. The pathways constitute complex networks characterized by crosstalk, and the development of a fully-functional organism requires the precise coordination of their activities. Signaling pathways are critically important for normal ontogenesis, mutations or alterations in gene expression within these pathways can lead to severe developmental disorders (Artavanis-Tsakonas et al., 1999; Ingham, McMahon, 2001; Logan, Nusse, 2004; Rubin, 2007; Perrimon et al., 2012; Briscoe, Thérond, 2013; Huttlin et al., 2017).

The Hedgehog (HH) signaling pathway, which owes its name to the discovery of the *hedgehog* (*hh*) gene in *Drosophila melanogaster* in the early 1980s, plays a substantial role in controlling morphogenesis. The larvae of flies mutant for this gene are covered with spines, giving them a hedgehog-like appearance (Nüsslein-Volhard, Wieschaus, 1980).

The Hedgehog signaling pathway is not merely one of the pathways orchestrating organismal development, but a central regulator of morphogenesis. It determines the anterior-posterior and dorso-ventral body axes and segmentation of embryonic primordia in animals, histo- and organogenesis, and the maintenance of stem cell pools in adult tissues, among other processes. Dysfunction of this signaling pathway is associated with numerous congenital anomalies and human diseases, including cancer of various organs (Ingham, McMahon, 2001; Spinella-Jaegle et al., 2001; Varjosalo, Taipale, 2007; Briscoe, Thérond, 2013; Wu et al., 2017; Skoda et al., 2018; Jamieson et al., 2020; Fitzsimons et al., 2022; Ingham, 2022; Dutta et al., 2023; Jing et al., 2023). It is exactly the reason, that there continues to be unrelenting interest in comprehensive investigation of the molecular-genetic organization and functioning mechanisms of the HH pathway. The general scheme of the Hedgehog signaling pathway is shown in Figure 1.

For the transmission of the HH signal, the recipient cell must contain a specific set of core proteins involved in the process, which must be in certain functional states. These proteins include: the transmembrane receptors Patched1 and Patched2 (PTCH1/2), the inactive form of the transmembrane protein Smoothened (SMO), complexes formed by transcription factors GLI1/3 and scaffold protein Suppressor of fused homolog (SUFU), active protein kinase A (PKA), which is responsible for generating the repressive form of the transcription factor GLI3 (GLI3R).

When the signaling pathway is inactive due to absence of HH ligands (Fig. 1*a*), PTCH1/2 receptors are localized on the primary cilium – a specialized external organelle of the cell that acts as a sensor for outside signals (Ingham, McMahon, 2001; Eggenschwiler, Anderson, 2007; Oro, 2007; Carballo et al., 2018).

PTCH1/2 block the migration of the SMO protein, which is located in the intracellular space, to the ciliary membrane, and SMO cannot interact with protein kinase A (PKA) to inhibit its activity. As a result, PKA phosphorylates the GLI3/SUFU complex, the complex dissociates, and GLI3 undergoes proteolytic cleavage to form the repressor GLI3R, which then enters the nucleus and suppresses the transcription of its target genes, including some genes of the HH pathway itself (Gorojankina, 2016; Dilower et al, 2023).

Signal transduction activation occurs when extracellular ligands – proteins belonging to the Hedgehog family (three types exist in humans: Sonic Hedgehog (SHH), Indian Hedge-

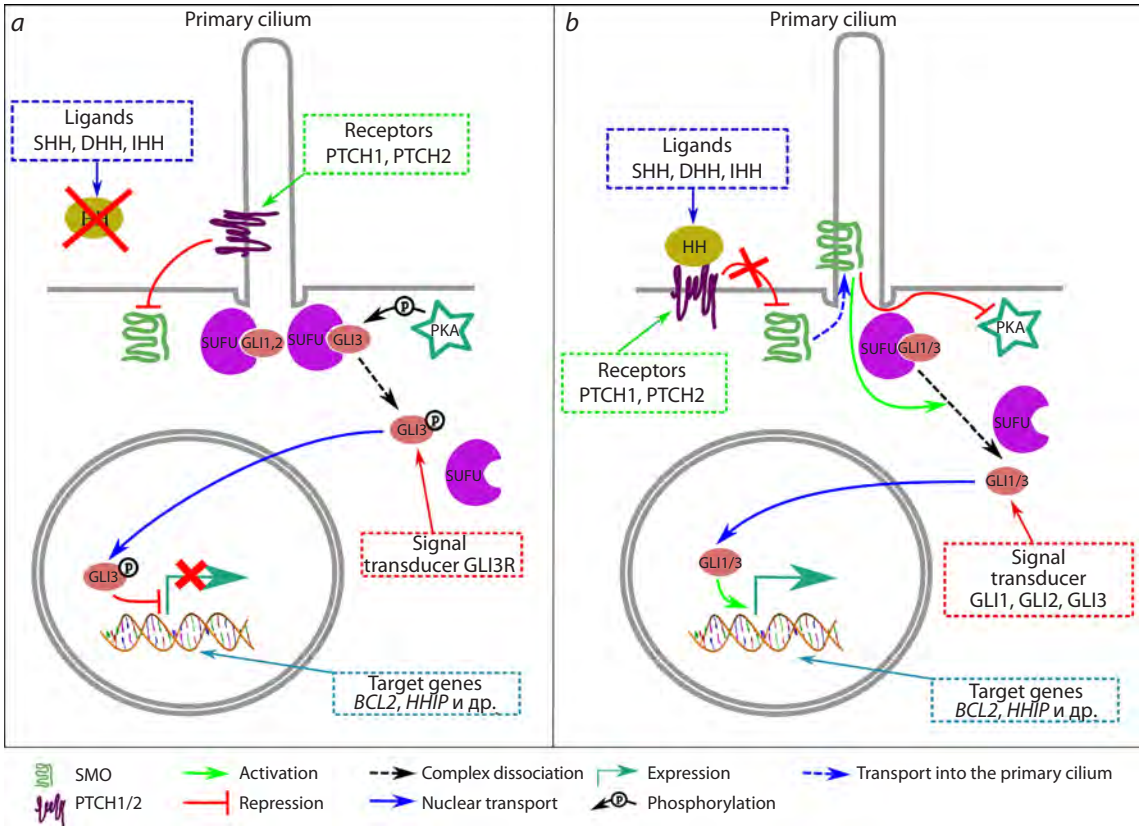


Fig. 1. General scheme of the human Hedgehog signaling pathway.

a – the mechanism of action when no HH ligand is present; *b* – the mechanism when PTCH1/2 receptors bind to HH ligands (details explained in text).

hog (IHH), and Desert Hedgehog (DHH)) – bind to PTCH1/2. The ligand/receptor complex is then removed from the ciliary membrane and transported to the intracellular space, where it is degraded in the lysosome. The position of PTCH1/2 is taken by SMO, which suppresses the activity of protein kinase A, thereby preventing the phosphorylation of the SUFU/GLI3 complex and the formation of GLI3R. Subsequently, within the cilium, the SUFU/GLI1/3 complexes are degraded, and the active forms of GLI1/3 are generated. These enter the nucleus and activate the transcription of target genes, ensuring signal transmission (Ingham, McMahon, 2001; Varjosalo, Taipale, 2007; Briscoe, Therond, 2013; Gorojankina, 2016) (Fig. 1*b*).

There are two variants of the HH pathway – the canonical one, shown in Figure 1, and the non-canonical one, in which the activation of the GLI1/3 transcription factors occurs without the involvement of SMO, thereby altering the signal transduction route (Brennan et al., 2012; Briscoe, Thérond, 2013; Carballo et al., 2018).

Currently, information concerning the HH pathway in humans is scattered across a vast number of sources (at the time of writing, on request "Hedgehog signaling" in PubMed alone returns 15,247 publications: https://pubmed.ncbi.nlm.nih.gov/?term=hedgehog+signaling), and this body of literature is continually expanding. Despite the extensive growth in the number of studies in this field, a complete and thorough understanding of the evolution, structure, and mechanisms of the HH pathway has not yet been achieved (Ingham et al., 2011; Briscoe; Thérond, 2013; Breeze, 2022).

To integrate, structure, and analyze existing data, the authors are creating a specialized knowledge base HH_Signal_pathways_db. The database is curated with diverse information related to all aspects of the organization and functioning of the Hedgehog pathway, which enables a systematic approach to its study.

Bioinformatic analysis of the structural and functional organization of the HH pathway opens up opportunities for deeper insight into the molecular-genetic basis of morphogenesis, mechanisms of organ and tissue regeneration, the aging process, the emergence of pathologies of various etiologies, as well as for developing methods for their diagnosis and pharmacotherapy.

As part of this work, new results have been obtained, including reconstruction of the associative gene network of the HH signaling pathway, identification of regulatory circuits, and acquisition of data regarding the evolution of genes involved in the pathway.

Materials and methods

Structure and content of the HH_Signal_pathway_db knowledge base. Figure 2 shows a block diagram of the database format developed by the authors.

The list of genes included in the human HH pathway (Table 1) was extracted from the KEGG database (https://www.genome.jp/kegg/) by querying (Environmental Information Processing→Signal Transduction→Hedgehog Signaling Pathway).

To fill the "gene information" and "gene product information" blocks, data were retrieved from the NCBI Gene (https://www.ncbi.nlm.nih.gov/gene), UniProt (https://www.uniprot.org), TRRUST (https://www.grnpedia.org/trrust/) databases.

Data for the "TPB affinity to the promoter" block (TBP, the TATA-binding protein, is a key regulator of transcription initiation in eukaryotic genes) was taken from the Human_SNP_TATAdb database (Filonev et al., 2023).

The "evolutionary characteristics" block was filled using Orthoweb, a specialized software package developed to calculate two evolutionary indices: the phylostratigraphic age index (PAI) and the divergence index (DI) (Mustafin et al., 2021; Ivanov et al., 2024).

The PAI index reflects the distance of a taxon from the root of the phylogenetic tree and is calculated as the distance from the root to the node where the divergence of the species under study from the most distant related taxon occurred: the higher the PAI, the "younger" the gene in question. For human genes, PAI values range from 0 (Cellular Organisms, the root of the tree) to 15 (*Homo sapiens*).

The gene evolutionary variability index (DI – Divergence Index) estimates the ratio between non-synonymous substitutions (which alter the encoded amino acid) in the sequences of the analyzed gene and its ortholog (dN), and synonymous substitutions (which do not change the encoded amino acid) (dS) in the nucleotide sequences of genes and their orthologs:

$$DI = \frac{\sum_{i=1}^{n} dnds_i}{n},$$

where $dnds_i$ is the dN/dS value for the gene and its *i*-th ortholog, and *n* is the number of orthologous genes.

The DI allows for determining the type of selection pressure acting on a given gene. DI values <1 and >1 are interpreted as evidence of stabilizing and positive selection, respectively, while DI = 1 indicates neutral evolution (Jeffares et al., 2015; Spielman, Wilke, 2015).

To construct the associative gene network and identify regulatory circuits (lower-dimensionality gene networks), the cognitive software and information system ANDSystem was used. This platform employs artificial intelligence methods to automatically extract knowledge from scientific publications and factual databases and, via the ANDVisio module, visualizes the results as a graph (Demenkov et al., 2011; Ivanisenko et al., 2015, 2019, 2022).

The gene network was reconstructed for 56 genes of the Hedgehog signaling pathway. It reflects associations with proteins encoded by these genes ("expression"), with transcription factors regulating gene expression ("expression regulation"), with proteins regulating protein transport ("transport regulation"), and with miRNAs involved in post-transcriptional regulation of protein expression ("miRNA regulation").

Functional annotation of genes was performed using the DAVID web resource (https://davidbioinformatics.nih.gov/) (Sherman et al., 2022). This tool identifies biological processes that are statistically overrepresented in the analyzed gene set. The false discovery rate (FDR), calculated using the Benjamini-Hochberg correction, was used as the significance criterion. Only processes with an FDR < 0.05 were considered.

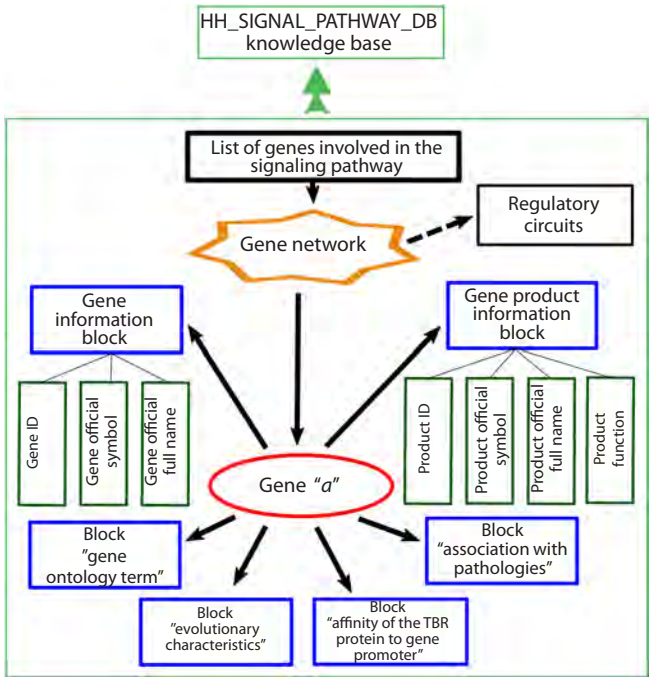


Fig. 2. Block diagram of the HH_Signal pathway_db knowledge base.

Results and discussion

The HH_Signal_pathway_db knowledge base

The current version of the HH_Signal_pathway_db contains structured information on 56 human genes related to the HH pathway (Table 1). The first release of the database contains the following blocks: 1) a list of HH signaling pathway genes with links to literary sources from the PubMed database; 2) lists of proteins encoded by HH signaling pathway genes and their functions; 3) Gene Ontology terms; 4) values of gene evolutionary age indices (PAI); 5) values of gene evolutionary variability indices (DI); 6) values of TBP binding affinity to gene promoters, a key determinant of transcription intensity; 7) lists of pathologies associated with each gene; 8) a reconstructed associative gene network and the regulatory circuits identified within it. A sample of filled database blocks for a specific gene, using the *SMURF2* gene as an example, is shown in Figure 3.

Below are some results of bioinformatic analysis of the information presented in the HH_Signal_pathway_db.

Functional annotation of HH signaling pathway genes

Analysis of biological process terms in Gene Ontology (GO) for the 56 genes performed using the DAVID resource, revealed 221 biological processes statistically significantly associated with the signaling pathway. Generally, these processes can be conditionally grouped into three main categories: morphogenesis (94), intracellular processes (60), and intercellular communication (67). Table 2. For all processes listed FDR < 0.05.

Morphogenesis

- GO:0042733~embryonic digit morphogenesis
- GO:0042475~odontogenesis of dentin-containing tooth

Table 1. Genes of the Hedgehog signaling pathway (according to the KEGG database

No.	Gene symbol	Gene ID	Gene full name
1	*ARRB1*	408	arrestin beta 1
2	*ARRB2*	409	arrestin beta 2
3	*BCL2*	596	BCL2 apoptosis regulator
4	*BOC*	91653	BOC cell adhesion associated, oncogene regulated
5	*BTRC*	8945	beta-transducin repeat containing E3 ubiquitin protein ligase
6	*CCND1*	595	cyclin D1
7	*CCND2*	894	cyclin D2
8	*CDON*	50937	cell adhesion associated, oncogene regulated
9	*CSNK1A1*	1452	casein kinase 1 alpha 1
10	*CSNK1A1L*	122011	casein kinase 1 alpha 1 like
11	*CSNK1D*	1453	casein kinase 1 delta
12	*CSNK1E*	1454	casein kinase 1 epsilon
13	*CSNK1G1*	53944	casein kinase 1 gamma 1
14	*CSNK1G2*	1455	casein kinase 1 gamma 2
15	*CSNK1G3*	1456	casein kinase 1 gamma 3
16	*CUL1*	8454	cullin 1
17	*CUL3*	8452	cullin 3
18	*DHH*	50846	desert hedgehog signaling molecule
19	*DISP1*	84976	dispatched RND transporter family member 1
20	*EFCAB7*	84455	EF-hand calcium binding domain 7
21	*EVC*	2121	EvC ciliary complex subunit 1
22	*EVC2*	132884	EvC ciliary complex subunit 2
23	*FBXW11*	23291	F-box and WD repeat domain containing 11
24	*GAS1*	2619	growth arrest specific 1
25	*GLI1*	2735	GLI family zinc finger 1
26	*GLI2*	2736	GLI family zinc finger 2
27	*GLI3*	2737	GLI family zinc finger 3
28	*GPR161*	23432	G protein-coupled receptor 161
29	*GRK2*	156	G protein-coupled receptor kinase 2
30	*GRK3*	157	G protein-coupled receptor kinase 3
31	*GSK3B*	2932	glycogen synthase kinase 3 beta
32	*HHAT*	55733	hedgehog acyltransferase
33	*HHATL*	57467	hedgehog acyltransferase like
34	*HHIP*	64399	hedgehog interacting protein
35	*IHH*	3549	Indian hedgehog signaling molecule
36	*IQCE*	23288	IQ motif containing E
37	*KIF3A*	11127	kinesin family member 3A
38	*KIF7*	374654	kinesin family member 7
39	*LRP2*	4036	LDL receptor related protein 2
40	*MEGF8*	1954	multiple EGF like domains 8
41	*MGRN1*	23295	mahogunin ring finger 1
42	*MOSMO*	730094	modulator of smoothened
43	*PRKACA*	5566	protein kinase cAMP-activated catalytic subunit alpha
44	*PRKACB*	5567	protein kinase cAMP-activated catalytic subunit beta
45	*PRKACG*	5568	protein kinase cAMP-activated catalytic subunit gamma
46	*PTCH1*	5727	patched 1
47	*PTCH2*	8643	patched 2
48	*SCUBE2*	57758	signal peptide, CUB domain and EGF like domain containing 2
49	*SHH*	6469	sonic hedgehog signaling molecule
50	*SMO*	6608	smoothened, frizzled class receptor
51	*SMURF1*	57154	SMAD specific E3 ubiquitin protein ligase 1
52	*SMURF2*	64750	SMAD specific E3 ubiquitin protein ligase 2
53	*SPOP*	8405	speckle type BTB/POZ protein
54	*SPOPL*	339745	speckle type BTB/POZ protein like
55	*SUFU*	51684	SUFU negative regulator of hedgehog signaling
56	*TPTEP2-CSNK1E*	102800317	TPTEP2-CSNK1E readthrough

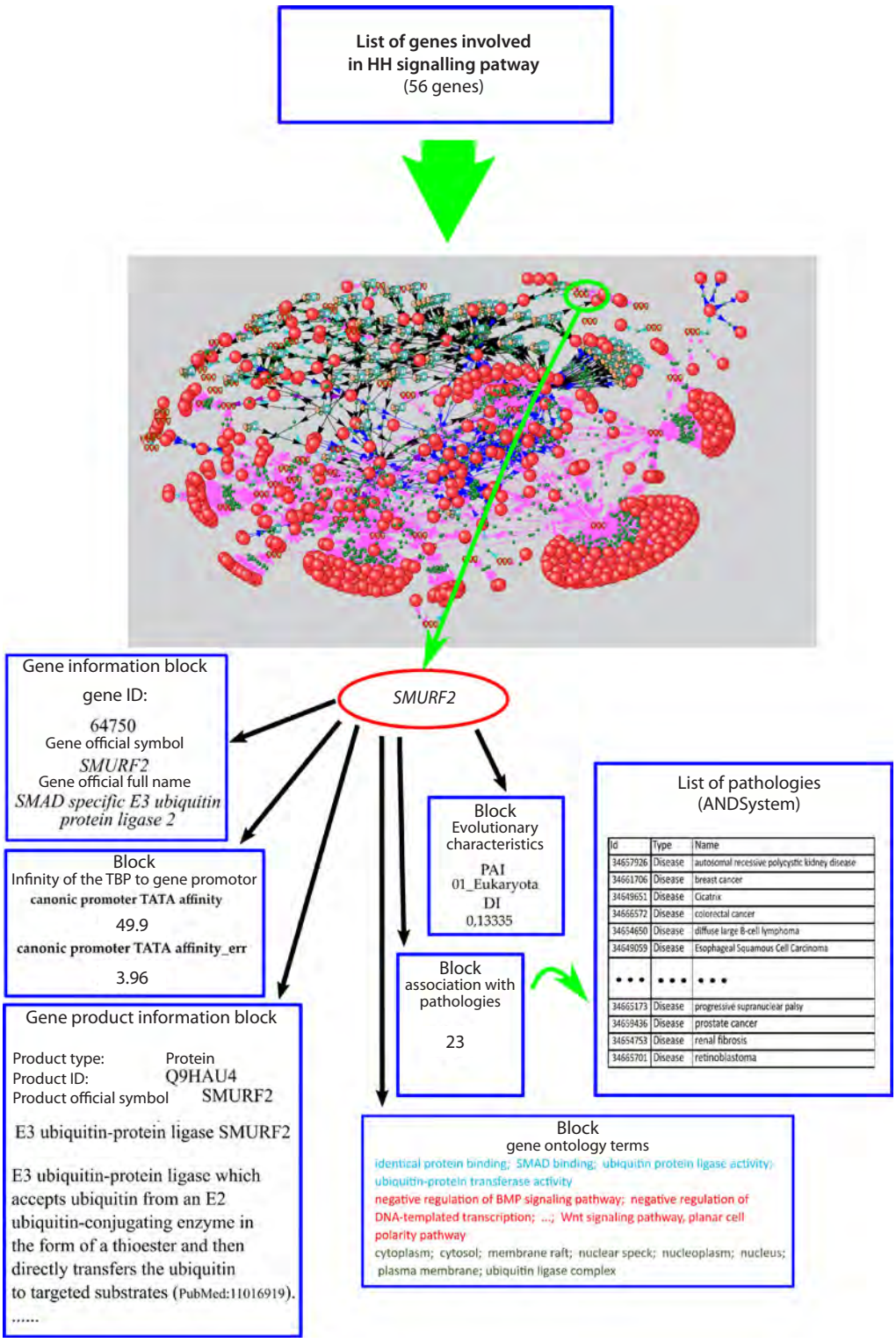


Fig. 3. An example of filling out the HH_Signal_pathway_db knowledge base block for the *SMURF2* gene.

- GO:0007507~heart development
- GO:0001658~branching involved in ureteric bud morphogenesis
- GO:0003151~outflow tract morphogenesis
- GO:0030324~lung development
- GO:0003180~aortic valve morphogenesis
- GO:0045766~positive regulation of angiogenesis
- GO:0001501~skeletal system development
- GO:0001942~hair follicle development
- GO:0021983~pituitary gland development
- GO:0001822~kidney development
- GO:0001525~angiogenesis
- GO:0042060~wound healing
- GO:0001889~liver development
- GO:0072091~regulation of stem cell proliferation

etc.

Intracellular processes

Regulation of transcription

- GO:1902895~positive regulation of miRNA transcription
- GO:1902894~negative regulation of miRNA transcription
- GO:0006357~regulation of transcription by RNA polymerase II
- GO:0006338~chromatin remodeling
- GO:0006355~regulation of DNA-templated transcription
- GO:0010468~regulation of gene expression

Response to stress

- GO:0071456~cellular response to hypoxia
- GO:0034599~cellular response to oxidative stress
- GO:0071466~cellular response to xenobiotic stimulus
- GO:0034644~cellular response to UV
- GO:0006974~DNA damage response

Regulation of cyclic processes

- GO:0048511~rhythmic process
- GO:0051726~regulation of cell cycle

Apoptosis

- GO:0043066~negative regulation of apoptotic process
- GO:0043065~positive regulation of apoptotic process

Intercellular communication

- GO:0042127~regulation of cell population proliferation
- GO:0050673~epithelial cell proliferation
- GO:0010595~positive regulation of endothelial cell migration
- GO:0001938~positive regulation of endothelial cell proliferation
- GO:0042127~regulation of cell population proliferation
- GO:0072089~stem cell proliferation

etc.

Involvement in signaling pathways

- GO:0038084~vascular endothelial growth factor signaling pathway
- GO:0007173~epidermal growth factor receptor signaling pathway
- GO:0008543~fibroblast growth factor receptor signaling pathway
- GO:0007224~smoothened signaling pathway
- GO:0060070~canonical Wnt signaling pathway
- GO:0030509~BMP signaling pathway
- GO:0000165~MAPK cascade
- GO:0007219~Notch signaling pathway
- GO:0070371~ERK1 and ERK2 cascade

etc.

A significant role of the Hedgehog signaling pathway is its participation in the morphogenetic processes of embryogenesis, histogenesis, and organogenesis. The pathway genes are involved in the formation of the nervous system, the development of cartilage and skeletal tissue, angiogenesis, and the development of kidneys, liver, lungs, heart, the endocrine pancreas, and genitals (Ingham, McMahon, 2001; Roy, Ingham, 2002; Fitzsimons et al., 2022; Ingham, 2022; Dilower et al., 2023).

Among the fundamental intracellular processes regulated by HH pathway genes are transcription (Gao Y. et al., 2023), response to stress stimuli (Chung et al., 2022), and maintenance of genomic stability (Ingham, McMahon, 2001). Furthermore, the signaling pathway modulates the cellular response to hypoxia, oxidative stress, and other adverse factors, which can be critical for cell survival (Kim, Lee, 2023; van der Weele et al., 2024). The involvement of Hedgehog signaling pathway elements in DNA repair (Gao Q. et al., 2019), apoptosis (Harris et al., 2011; Rimkus et al., 2016), and cell cycle regulation confirms its role in controlling cell proliferation and differentiation (Roy, Ingham, 2002).

According to available data, the HH pathway acts as a mediator of intercellular communication not only by itself; its components, in particular beta-arrestins (ARRB1/2), kinases (CCND1, CSNK1A1, CSNK1E, CSNK1A1L, GSK3B, PRKACA, PRKACB, PRKACG, TPTEP2-CSNK1E), ubiquitination proteins (BTRC, CUL1, FBXW11), and others, are involved in other signaling cascades, including MAPK/ERK, Wnt, Notch, and VEGF. The participation of HH pathway proteins in other signaling pathways has also been demonstrated by other authors (Rubin, 2007; Butí et al., 2014; Edeling et al, 2016; Luo, 2017; Fang et al., 2023).

Associative gene network of the Hedgehog signaling pathway

The network reconstructed with ANDSystem contains information on 56 genes, 504 proteins, 126 miRNAs, and 1,412 interactions of various types between its elements. A general view of the network is presented in Figure 4.

Analysis of the gene network revealed certain patterns pertaining to intra-network interactions. Specifically, it was shown that there are at least seven regulatory circuits within the network (Fig. 5, 6). These can be tentatively divided into two groups.

The circuits of the first group mediate the auto-regulation of the signaling pathway as a whole. The second group regulates the interaction of some components within the signaling pathway itself. The first group comprises four circuits – three with positive feedback loops, implementing pathway auto-activation (Fig. 5*a*–*c*), and one with a negative feedback loop, mediating autorepression of the pathway (Fig. 5*d*). The auto-activation circuits include the membrane proteins GAS1, BOC, CDON, which participate in the interaction of the PTCH1/2 receptor with its HH ligand, thereby facilitating signal transduction. The expression of the genes encoding these membrane proteins is controlled by the GLI1/3 transcription factors (Allen et al., 2007; Song et al., 2015; Echevarría-Andino et al., 2023).

The main component of the fourth circuit is the HHIP protein, which prevents the binding of PTCH1/2 to HH, thereby prohibiting signal propagation. The *HHIP* gene is a target of GLI1/3 transcription factors (Chuang, McMahon, 1999; Falkenstein, Vokes, 2014).

The second group, defining the character of certain interactions within the HH pathway, is formed by three circuits. The first controls the interaction between PTCH1 and SMO via a positive feedback loop (Fig. 6*a*). The second is a mutual regulation circuit of the genes encoding the GLI1/3 transcription factors (Fig. 6*b*). It can exist in two states depending on the functional status of the pathway. In the presence of the HH signal, the circuit operates in a mode of mutual gene activation via positive feedback loops. In the absence of the signal, the repressor form GLI3R suppresses the transcription of the

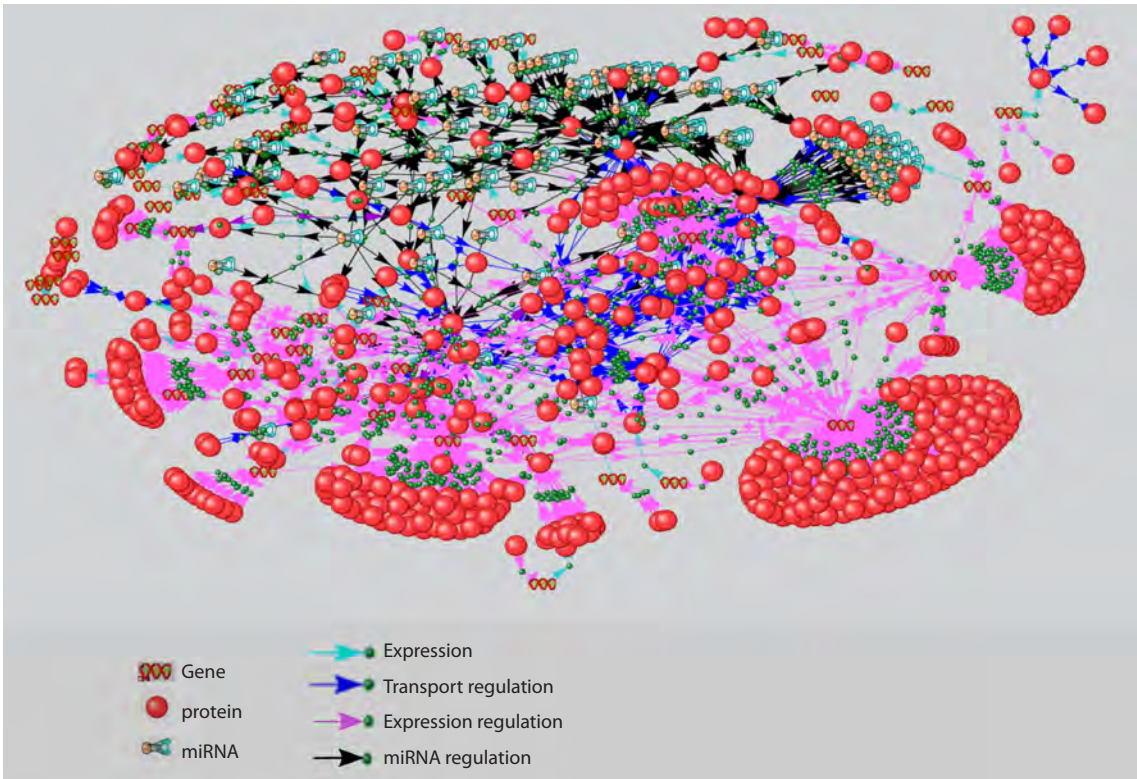


Fig. 4. A reconstruction of the associative gene network for the human Hedgehog signaling pathway, generated by the ANDSystem tool.

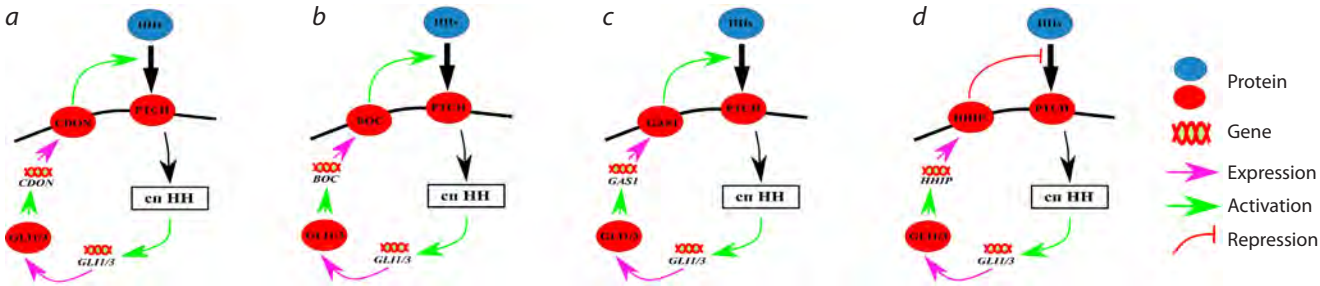


Fig. 5. Auto-regulation of the HH signaling pathway.

a–c – regulatory circuits with positive feedback; *d* – regulatory circuit with negative feedback; SP – signaling pathway.

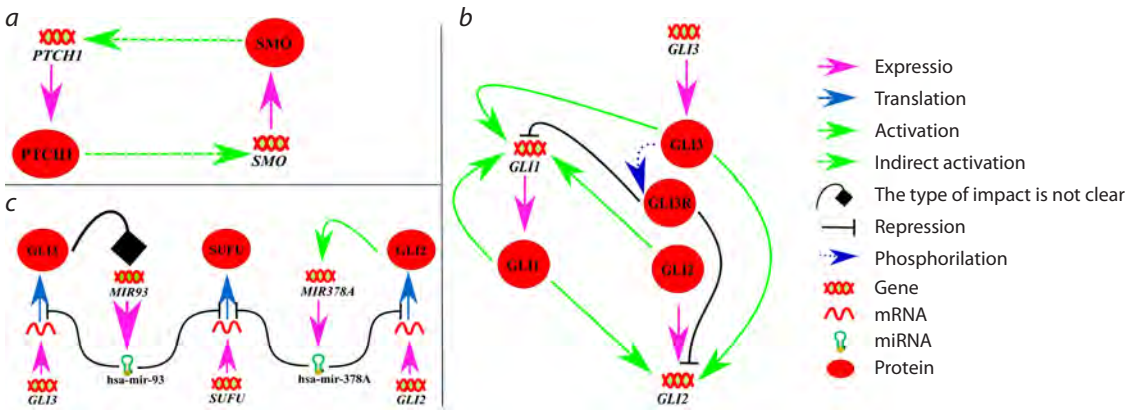


Fig. 6. Schemes of mutual regulation of components in three regulatory circuits of the HH signaling pathway.

a – regulation of PTCH1 and SMO; *b* – auto-regulation of GLI1/3; *c* – regulation of GLI2/3 and SUFU.

GLI1/2 genes and turns off the auto-activation. Thus, the balance between the activator and repressor forms of GLI is maintained (Wang et al., 2000; Vokes et al., 2007; Briscoe, Thérond, 2013). The third circuit of the group functions with the participation of two miRNAs – hsa-mir-93 and hsa-mir-378A, regulating the levels of GLI2/3 and SUFU via negative feedback loops (Fig. 6*c*). The involvement of miRNAs, including hsa-mir-93 and hsa-mir-378A, in regulating the expression of HH pathway proteins was established by A. Helwak et al. (2013). Analysis of the reconstructed HH signaling pathway gene network revealed that the genes encoding these miRNAs are targets for the GLI2/3 transcription factors.

Evolutionary characteristics of human Hedgehog signaling pathway genes:

The distribution of genes by values of their phylostratigraphic indecies PAI is presented in Table 2 and Figure 7.

The vast majority of pathway genes are characterized by indices of PAI = 01 (35 genes) and PAI = 02 (18 genes), indicating their emergence at the level of the first unicellular eukaryotes and the first multicellular animals. Two genes – *BCL2* and *SUFU* – originated significantly earlier – at the cellular level of biological organization (their PAI = 00). Both of these genes control the cell pool – *BCL2* as an apoptosis regulator, and *SUFU* as an inhibitor of tumor growth, i. e., uncontrolled cell proliferation (Willis et al., 2003; Cheng, Yue, 2008).

Only one gene, *HHIP*, originated during the formation of chordates, has a PAI value of 03. The eponymous protein inhibits the signaling cascade already at its initial stage by binding to the PTCH1 receptor and preventing the ligand–receptor interaction.

Previously, independent data on the emergence time of certain components of the human Hedgehog (HH) signaling pathway prior to vertebrate divergence had been obtained for all HH ligands (Kumar et al., 1996) and for the GLI transcription factors (Shimeld et al., 2007), and these findings are consistent with the results presented.

A comparison of the PAI value distribution between HH cascade genes and all human protein-coding genes (Fig. 7) showed a statistically significant bias towards more ancient values in HH pathway genes ($p < 0.05$, Mann–Whitney test). This aligns with the fact that this pathway is activated earlier than others in ontogeny, suggesting that its core components therefore had to emerged at early stages of multicellular organisms evolution. Indeed, all forms of HH, GLI, PTCH, and SMO proteins, which play the main role in signal transduction, are characterized by PAI = 01–02, and their functional analogs are present even in invertebrate animals (Ingham, McMahon, 2001; Wilson, Chuang, 2010). Notably, all genes of the regulatory circuits except *HHIP*, have ancient origin, at that *HHIP* is the only gene included in the regulatory circuit with negative feedback.

Figure 8 shows the distribution of DI index values for HH pathway genes. Given that this pathway orchestrates the implementation of fundamental cellular processes involved in morphogenesis, including division, differentiation, and apoptosis, it is unsurprising that 89 % of its genes (50) have a DI index <0.5, with 12 of them (≈21 %) having an index below 0.1. This fact confirms that the signaling pathway, and the genes of the regulatory circuits governing its function, are under stabilizing selection which limits the accumulation of genomic changes.

In the analyzed set of 56 genes, only two have DI > 1 – these are *CSNK1A1L* (1.213) and *EFCAB7* (1.051). This finding, within the framework of the applied method, suggests that these genes may be under positive selection. The kinase CSNK1A1L phosphorylates GLI1/3 proteins. According to KEGG database data (hsa04340), in the human HH signaling pathway, several other kinases (CSNK1A1, CSNK1D, CSNK1E, CSNK1G1, CSNK1G2, CSNK1G3, TPTEP2-CSNK1E), encoded by genes of the same name, also participate in this process. All of them fall into the group with PAI = 02_Eukaryota, however, the DI values for them range from 0.0361 for *CSNK1A1* to 0.264 for *CSNK1D*, indicating the action of stabilizing selection on them. It can be assumed that *CSNK1A1L* might have "incorporated" into the signaling pathway later in evolution than the other kinase genes, and therefore may currently be experiencing the influence of positive, rather than stabilizing, selection.

The EFCAB7 protein, together with EVC, EVC2, and IQCE proteins, is involved in anchoring SMO to the primary cilium of mammalian cells, which distinguishes the signal transduction mechanism from the analogous process in *Drosophila*, whose cells do not possess primary cilia (Chen et al., 2009; Gorojankina, 2017). Probably, the weak pressure of positive selection on the *EFCAB7* gene, reflected in its DI value close to one, is related precisely to the later emergence of the mechanism involving primary cilia in the signal transduction process compared to other pathway components performing the same function – the *EVC*, *EVC2*, and *IQCE* genes (Chen

Table 2. Distribution of 56 human Hedgehog signaling pathway genes according to phylostratigraphic index (PAI) values

PAI Index_Taxon	Genes
00_Cellular Organisms	***SUFU***, *BCL2*
01_Eucaryota	*ARRB1*, *ARRB2*, *BTRC*, *CCND1*, *CCND2*, *CSNK1A1*, *CSNK1A1L*, *CSNK1D*, *CSNK1E*, *CSNK1G1*, *CSNK1G2*, *CSNK1G3*, *CUL1*, *CUL3*, ***DHH***, *DISP1*, *EFCAB7*, *FBXW11*, *GRK2*, *GRK3*, *GSK3B*, ***IHH***, *KIF3A*, *KIF7*, *MOSMO*, *PRKACA*, *PRKACB*, *PRKACG*, ***PTCH1***, ***PTCH2***, *SMURF1*, *SMURF2*, *SPOP*, *SPOPL*, *TPTEP2-CSNK1E*
02_Metazoa	***BOC***, ***CDON***, *EVC*, *EVC2*, ***GAS1***, ***GLI1***, ***GLI2***, ***GLI3***, *GPR161*, *HHAT*, *HHATL*, *IQCE*, *LRP2*, *MEGF8*, *MGRN1*, *SCUBE2*, ***SHH***, ***SMO***
03_Chordata	***HHIP***

Note. Gene names belonging to regulatory circuits with feedback are highlighted in bold.

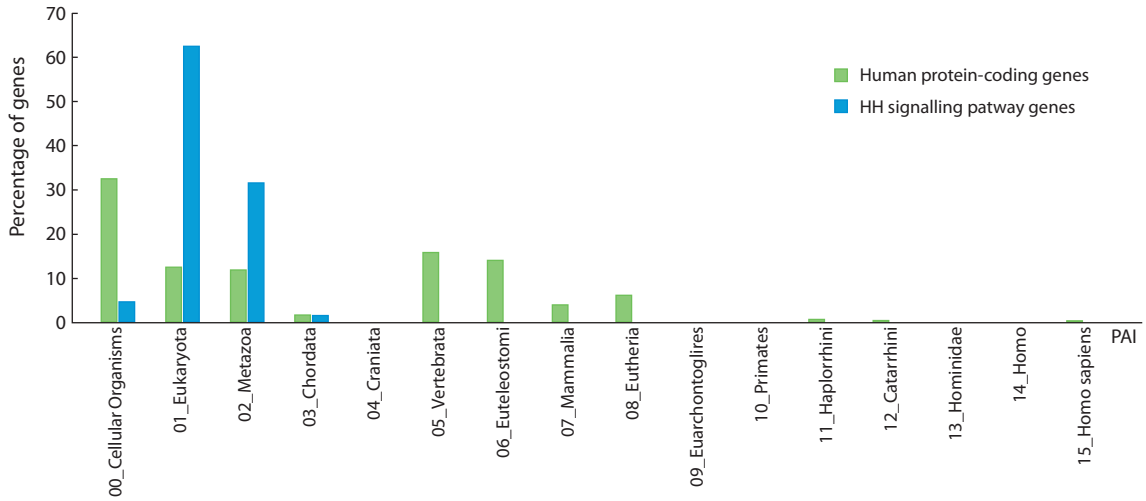


Fig. 7. Distribution of PAI values among genes of the Hedgehog signaling pathway (56 genes) and all human protein-coding genes (19,491 genes).

The differences in values are statistically significant at $p < 0.05$ according to the Mann–Whitney test.

et al., 2009; Wilson, Chuang 2010), which are evidently under stabilizing selection, as indicated by their DI values of 0.298, 0.421, and 0.679, respectively.

Thus, the overwhelming majority of Hedgehog signaling pathway genes can be characterized as ancient, subject to stabilizing selection, preventing the accumulation of genetic variability and promoting functional stability of the genes. Their conservatism confirms the critical role of the HH pathway in regulating fundamental ontogenetic processes.

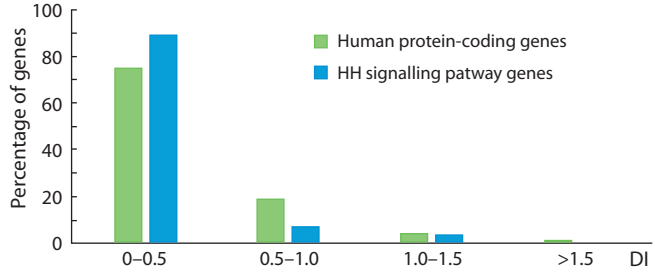


Fig. 8. Distribution of DI values for Hedgehog signaling pathway genes compared to all human protein-coding genes.

Conclusion

A prototype of the HH_Signal_pathway_db knowledge base has been developed. It accumulates information on the structural and functional organization of the evolutionarily conserved Hedgehog (HH) signaling pathway in humans, integrating data from KEGG, NCBI Gene, UniProt, and other sources. The database systematizes fragmented data on the HH signaling pathway in humans and can serve as a tool for systematic analysis of its role in ontogenesis, maintaining homeostasis, and pathology development.

The bioinformatic analysis of some data from the base, in particular, showed that: 1) according to functional annotation, the pathway's genes are associated with three categories of processes: intracellular, organ morphogenesis, and intercellular communication, including interaction with other signaling cascades; 2) the vast majority of the pathway's genes are of ancient origin and subject to stabilizing selection; 3) the reconstructed associative gene network of the HH pathway contains 56 genes, 504 proteins, 126 miRNAs, and establishes 1,412 interactions among them; 4) the network's functioning is regulated by seven regulatory circuits – five with positive and two with negative feedback. One of the negative feedback circuits involve two miRNAs.

References

Allen B.L., Tenzen T., McMahon A.P. The Hedgehog-binding proteins Gas1 and Cdo cooperate to positively regulate Shh signaling during mouse development. *Genes Dev*. 2007;21(10):1244-1257. doi 10.1101/gad.1543607

Artavanis-Tsakonas S., Rand M.D., Lake R.J. Notch signaling: cell fate control and signal integration in development. *Science*. 1999; 284(5415):770-776. doi 10.1126/science.284.5415.770

Bartel D.P. Metazoan microRNAs. *Cell*. 2018;173(1):20-51. doi 10.1016/j.cell.2018.03.006

Breeze E. Role of Hedgehog signalling pathway in the maintenance and regeneration of adult tissues. *J Cell Signal*. 2022;7:281. doi 10.35248/2576-1471.22.7.281

Brennan D., Chen X., Cheng L., Mahoney M., Riobo N.A. Noncanonical Hedgehog signaling. *Vitam Horm*. 2012;88:55-72. doi 10.1016/B978-0-12-394622-5.00003-1

Briscoe J., Thérond P.P. The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat Rev Mol Cell Biol*. 2013; 14(7):416-429. doi 10.1038/nrm3598

Butí E., Mesquita D., Araújo S.J. Hedgehog is a positive regulator of FGF signalling during embryonic tracheal cell migration. *PLoS One*. 2014;9(3):e92682. doi 10.1371/journal.pone.0092682

Carballo G.B., Honorato J.R., de Lopes G.P.F., Spohr T.C.L.S.E. A highlight on Sonic hedgehog pathway. *Cell Commun Signal*. 2018;16(1):11. doi 10.1186/s12964-018-0220-7

Chen M.H., Wilson C.W., Li Y.J., Law K.K., Lu C.S., Gacayan R., Zhang X., Hui C.C., Chuang P.T. Cilium-independent regulation of Gli protein function by Sufu in Hedgehog signaling is evolutionarily conserved. *Genes Dev*. 2009;23(16):1910-1928. doi 10.1101/gad.1794109

Cheng S.Y., Yue S. Role and regulation of human tumor suppressor SUFU in Hedgehog signaling. *Adv Cancer Res*. 2008;101:29-43. doi 10.1016/S0065-230X(08)00402-8

Chuang P.T., McMahon A.P. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature*. 1999; 397(6720):617-621. doi 10.1038/17611

Chung K.M., Kim H., Roque C.G., McCurdy E.P., Nguyen T.T.T., Siegelin M.D., Hwang J.Y., Hengst U. A systemic cell stress signal confers neuronal resilience toward oxidative stress in a Hedgehog-dependent manner. *Cell Rep*. 2022;41(3):111488. doi 10.1016/j.celrep.2022.111488

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2011;11(3):149-161. doi 10.3233/ISB-2012-0449

Dilower I., Niloy A.J., Kumar V., Kothari A., Lee E.B., Rumi M.A.K. Hedgehog signaling in gonadal development and function. *Cells*. 2023;12(3):358. doi 10.3390/cells12030358

Dutta R.K., Jun J., Du K., Diehl A.M. Hedgehog signaling: implications in liver pathophysiology. *Semin Liver Dis*. 2023;43(4):418-428. doi 10.1055/a-2187-3382

Echevarría-Andino M.L., Franks N.E., Schrader H.E., Hong M., Krauss R.S., Allen B.L. CDON contributes to Hedgehog-dependent patterning and growth of the developing limb. *Dev Biol*. 2023;493: 1-11. doi 10.1016/j.ydbio.2022.09.011

Edeling M., Ragi G., Huang S., Pavenstädt H., Susztak K. Developmental signalling pathways in renal fibrosis: the roles of Notch, Wnt and Hedgehog. *Nat Rev Nephrol*. 2016;12(7):426-439. doi 10.1038/nrneph.2016.54

Eggenschwiler J.T., Anderson K.V. Cilia and developmental signaling. *Annu Rev Cell Dev Biol*. 2007;23:345-373. doi 10.1146/annurev.cellbio.23.090506.123249

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi 10.1038/nature11247

Falkenstein K.N., Vokes S.A. Transcriptional regulation of graded Hedgehog signaling. *Semin Cell Dev Biol*. 2014;33:73-80. 10.1016/j.semcdb.2014.05.010

Fang Z., Meng Q., Xu J., Wang W., Zhang B., Liu J., Liang C., Hua J., Zhao Y., Yu X., Shi S. Signaling pathways in cancer-associated fibroblasts: recent advances and future perspectives. *Cancer Commun* (*Lond*). 2023;43(1):3-41. doi 10.1002/cac2.12392

Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85

Fitzsimons L.A., Brewer V.L., Tucker K.L. Hedgehog morphogens act as growth factors critical to pre- and postnatal cardiac development and maturation: how primary cilia mediate their signal transduction. *Cells*. 2022;11(12):1879. doi 10.3390/cells11121879

Gao Q., Zhou G., Lin S.J., Paus R., Yue Z. How chemotherapy and radiotherapy damage the tissue: comparative biology lessons from feather and hair models. *Exp Dermatol*. 2019;28(4):413-418. doi 10.1111/exd.13846

Gao Y., Shan Z., Jian C., Wang Y., Yao X., Li S., Ti X., Zhao G., Liu C., Zhang Q. HIB/SPOP inhibits Ci/Gli-mediated tumorigenesis by modulating the RNA polymerase II components stabilities. *iScience*. 2023;26(8):107334. doi 10.1016/j.isci.2023.107334

Ghafouri-Fard S., Khoshbakht T., Hussen B.M., Taheri M., Samsami M. Emerging role of non-coding RNAs in the regulation of Sonic Hedgehog signaling pathway. *Cancer Cell Int*. 2022;22(1):282. doi 10.1186/s12935-022-02702-y

Gorojankina T. Hedgehog signaling pathway: a novel model and molecular mechanisms of signal transduction. *Cell Mol Life Sci*. 2016; 73(7):1317-1332. doi 10.1007/s00018-015-2127-4

Harris L.G., Samant R.S., Shevde L.A. Hedgehog signaling: networking to nurture a promalignant tumor microenvironment. *Mol Cancer Res*. 2011;9(9):1165-1174. doi 10.1158/1541-7786.MCR-11-0175

Helwak A., Kudla G., Dudnakova T., Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. 2013;153(3):654-665. doi 10.1016/j.cell.2013.03.043

Huttlin E.L., Bruckner R.J., Paulo J.A., Cannon J.R., Ting L., Baltier K., Colby G., … Guruharsha K.G., Li K., Artavanis-Tsakonas S., Gygi S.P., Harper J.W. Architecture of the human interactome defines protein communities and disease networks. *Nature*. 2017; 545(7655):505-509. doi 10.1038/nature22366

Ingham P.W. Hedgehog signaling. *Curr Top Dev Biol*. 2022;149:1-58. doi 10.1016/bs.ctdb.2022.04.003

Ingham P.W., McMahon A.P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev*. 2001;15(23):3059-3087. doi 10.1101/gad.938601

Ingham P.W., Nakano Y., Seger C. Mechanisms and functions of Hedgehog signalling across the metazoa. *Nature Rev Genet*. 2011; 12(6):393-406. doi 10.1038/nrg2984

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9:S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Kolchanov N.A. ANDSystem: a cognitive system for the reconstruction and analysis of knowledge graphs (gene networks) based on the automated extraction of data from scientific publications, patents, and factual databases. *Nauka i Tekhnologii Sibiri*. 2022;4(7):122-125. Available at: https://scitech.sb-ras.ru/upload/iblock/010/5ttp14te9uu1r5g0suu7ka05ui8udynq/nit_2022_7.pdf (in Russian)

Ivanov R.A., Mukhin A.M., Kazantsev F.V., Mustafin Z.S., Afonnikov D.A., Matushkin Y.G., Lashin S.A. Orthoweb: a software package for evolutionary analysis of gene networks. *Vavilov J Genet Breed*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Jamieson C., Martinelli G., Papayannidis C., Cortes J.E. Hedgehog pathway inhibitors: a new therapeutic class for the treatment of acute myeloid leukemia. *Blood Cancer Discov*. 2020;1(2):134-145. doi 10.1158/2643-3230.BCD-20-0007

Jeffares D.C., Tomiczek B., Sojo V., dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol*. 2015;1201: 65-90. doi 10.1007/978-1-4939-1438-8_4

Jing J., Wu Z., Wang J., Luo G., Lin H., Fan Y., Zhou C. Hedgehog signaling in tissue homeostasis, cancers, and targeted therapies. *Signal Transduct Target Ther*. 2023;8(1):315. doi 10.1038/s41392-023-01559-5

Kenneth J.H. Big Data among Big Data: Genome Data. 2022. Available at: https://3billion.io/blog/big-data-among-big-data-genome-data/

Kim N.H., Lee A.Y. Oxidative stress induces skin pigmentation in melasma by inhibiting Hedgehog signaling. *Antioxidants* (*Basel*). 2023; 12(11):1969. doi 10.3390/antiox12111969

Kumar S., Balczarek K.A., Lai Z.C. Evolution of the *hedgehog* gene family. *Genetics*. 1996;142(3):965-972. doi 10.1093/genetics/142.3.965

Logan C.Y., Nusse R. The Wnt signaling pathway in development and disease. *Ann Rev Cell Dev Biol*. 2004;20:781-810. doi 10.1146/annurev.cellbio.20.010403.113126

Luo K. Signaling cross talk between TGF-β/Smad and other signaling pathways. *Cold Spring Harb Perspect Biol*. 2017;9(1):a022137. doi 10.1101/cshperspect.a022137

McIntyre G., Jackson Z., Colina J., Sekhar S., DiFeo A. *miR-181a*: regulatory roles, cancer-associated signaling pathway disruptions, and therapeutic potential. *Expert Opin Ther Targets*. 2024;28(12):1061-1091. doi 10.1080/14728222.2024.2433687

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nüsslein-Volhard C., Wieschaus E. Mutations affecting segment number and polarity in Drosophila. *Nature*. 1980;287(5785):795-801. doi 10.1038/287795a0

Oro A.E. The primary cilia, a 'Rab-id' transit system for Hedgehog signaling. *Curr Opin Cell Biol*. 2007;19(6):691-696. doi 10.1016/j.ceb.2007.10.008

Perrimon N., Pitsouli C., Shilo B.Z. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb Perspect Biol*. 2012;4(8):a005975. doi 10.1101/cshperspect.a005975

Regev A., Teichmann S.A., Lander E.S., Amit I., Benoist C., Birney E., Bodenmiller B., … Watt F., Weissman J., Wold B., Xavier R., Yosef N., Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*. 2017;6:e27041. doi 10.7554/eLife.27041

Rimkus T.K., Carpenter R.L., Qasem S., Chan M., Lo H.W. Targeting the sonic Hedgehog signaling pathway: review of Smoothened and GLI inhibitors. *Cancers* (*Basel*). 2016;8(2):22. doi 10.3390/cancers8020022

Roy S., Ingham P.W. Hedgehogs tryst with the cell cycle. *J Cell Sci*. 2002;115(Pt 23):4393-4397. doi 10.1242/jcs.00158

Rubin D.C. Intestinal morphogenesis. *Curr Opin Gastroenterol*. 2007;23(2):111-114. doi 10.1097/MOG.0b013e3280145082

Schermelleh L., Ferrand A., Huser T., Eggeling C., Sauer M., Biehlmaier O., Drummen G.P. Super-resolution microscopy demystified. *Nat Cell Biol*. 2019;21(1):72-84. doi 10.1038/s41556-018-0251-8

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Shimeld S.M., van den Heuvel M., Dawber R., Briscoe J. An amphioxus Gli gene reveals conservation of midline patterning and the evolution of hedgehog signalling diversity in chordates. *PLoS One*. 2007;2(9):e864. doi 10.1371/journal.pone.0000864

Skoda A.M., Simovic D., Karin V., Kardum V., Vranic S., Serman L. The role of the Hedgehog signaling pathway in cancer: a comprehensive review. *Bosn J Basic Med Sci*. 2018;18(1):8-20. doi 10.17305/bjbms.2018.2756

Song J.Y., Holtz A.M., Pinskey J.M., Allen B.L. Distinct structural requirements for CDON and BOC in the promotion of Hedgehog signaling. *Dev Biol*. 2015;402(2):239-252. doi 10.1016/j.ydbio.2015.03.015

Spielman S.J., Wilke C.O. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 2015;32(4):1097-1108. doi 10.1093/molbev/msv003

Spinella-Jaegle S., Rawadi G., Kawai S., Gallea S., Faucheu C., Mollat P., Courtois B., Bergaud B., Ramez V., Blanchet A.M., Adelmant G., Baron R., Roman-Roman S. Sonic hedgehog increases the commitment of pluripotent mesenchymal cells into the osteoblastic lineage and abolishes adipocytic differentiation. *J Cell Sci*. 2001;114(Pt. 11):2085-2094. doi 10.1242/jcs.114.11.2085

van der Weele C.M., Hospes K.C., Rowe K.E., Jeffery W.R. Hypoxia-sonic hedgehog axis as a driver of primitive hematopoiesis development and evolution in cavefish. *Dev Biol*. 2024;516:138-147. doi 10.1016/j.ydbio.2024.08.008

Varjosalo M., Taipale J. Hedgehog signaling. *J Cell Sci*. 2007;120(Pt. 1):3-6. doi 10.1242/jcs.03309

Vokes S.A., Ji H., McCuine S., Tenzen T., Giles S., Zhong S., Longabaugh W.J., Davidson E.H., Wong W.H., McMahon A.P. Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development*. 2007;134(10):1977-1989. doi 10.1242/dev.001966

Wang B., Fallon J.F., Beachy P.A. Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell*. 2000;100(4):423-434. doi 10.1016/s0092-8674(00)80678-9

Willis S., Day C.L., Hinds M.G., Huang D.C. The Bcl-2-regulated apoptotic pathway. *J Cell Sci*. 2003;116(Pt. 20):4053-4056. doi 10.1242/jcs.00754

Wilson C.W., Chuang P.T. Mechanism and evolution of cytosolic Hedgehog signal transduction. *Development*. 2010;137(13):2079-2094. doi 10.1242/dev.045021

Wu F., Zhang Y., Sun B., McMahon A.P., Wang Y. Hedgehog signaling: from basic biology to cancer therapy. *Cell Chem Biol*. 2017;24(3):252-280. doi 10.1016/j.chembiol.2017.02.010

Conflict of interest. The authors declare no conflict of interest.

Received July 22, 2025. Revised September 17, 2025. Accepted September 17, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):990-999

doi 10.18699/vjgb-25-104

Identification of proteins regulating phenotype-associated genes of M2 macrophages: a bioinformatic analysis

E.A. Antropova, I.V. Yatsyk, P.S. Demenkov, T.V. Ivanisenko, V.A. Ivanisenko

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
nzhenia@bionet.nsc.ru

Abstract. Macrophages are immune system cells that perform various, often opposing, functions in the organism depending on the incoming microenvironment signals. This is possible due to the plasticity of macrophages, which allows them to radically alter their phenotypic characteristics and gene expression profiles, as well as return to their original, non-activated state. Depending on the inductors acting on the cell, macrophages are activated into various functional states. There are five main phenotypes of activated macrophages: M1, M2a, M2b, M2c, and M2d. Although the amount of genome-wide transcriptomic and proteomic data showing differences between major macrophage phenotypes and non-activated macrophages (M0) is rapidly growing, questions regarding the mechanisms regulating gene and protein expression profiles in macrophages of different phenotypes still remain. We compiled lists of proteins associated with the macrophage phenotypes M1, M2a, M2b, M2c, and M2d (phenotype-associated proteins) and analyzed the data on potential mediators of macrophage polarization. Furthermore, using the computational system ANDSystem, we conducted a search and analysis of the relationships between potential regulatory proteins and the genes encoding the proteins associated with the M2 group phenotypes, obtaining estimates of the statistical significance of these relationships. The results indicate that the differences in the M2a, M2b, M2c, and M2d macrophage phenotypes may be attributed to the regulatory effects of the proteins JUN, IL8, NFAC2, CCND1, and YAP1. The expression levels of these proteins vary among the M2 group phenotypes, which in turn leads to different levels of gene expression associated with specific phenotypes.

Key words: macrophage phenotypes; expression regulation; proteomes; ANDSystem; automated text analysis

For citation: Antropova E.A., Yatsyk I.V., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Identification of proteins regulating phenotype-associated genes of M2 macrophages: a bioinformatic analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):990-999. doi 10.18699/vjgb-25-104

Funding. The work was supported by budget projects FWNR-2022-0020.

Выявление белков, регулирующих фенотип-ассоциированные гены макрофагов группы М2: биоинформатический анализ

Е.А. Антропова, И.В. Яцык, П.С. Деменков, Т.В. Иванисенко, В.А. Иванисенко

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
nzhenia@bionet.nsc.ru

Аннотация. Макрофаги – клетки иммунной системы, выполняющие в организме различные, часто противоположные функции в зависимости от поступающих сигналов микроокружения. Это возможно благодаря пластичности макрофагов, позволяющей кардинально менять фенотипические признаки и профили экспрессии генов, а также возвращаться в исходное, неактивированное состояние. В зависимости от действующих на клетку индукторов макрофаги поляризуются в различные функциональные состояния. Принято выделять пять основных фенотипов активированных макрофагов: M1, M2a, M2b, M2c и M2d. Хотя количество полногеномных транскриптомных и протеомных данных, показывающих различия между основными фенотипами макрофагов и неактивированными макрофагами (M0), растет стремительно, все еще остаются вопросы, касающиеся механизмов регуляции профилей экспрессии генов и белков у макрофагов разных фенотипов. Нами были составлены списки белков, ассоциированных с фенотипами макрофагов M1, M2a, M2b, M2c, M2d (фенотип-ассоциированные белки), проанализированы данные о возможных посредниках поляризации макрофагов. Далее с использованием компьютерной системы ANDSystem проведен поиск и анализ связей между потенциальными регуляторными белками и генами, кодирующими белки, ассоциированные с фенотипами группы М2, получены оценки стати-

стической значимости этих связей. Результаты указывают на то, что различия в фенотипах макрофагов M2a, M2b, M2c, M2d могут быть обусловлены регуляторными действиями белков JUN, IL8, NFAC2, CCND1 и YAP1. Уровень их экспрессии варьируется в зависимости от фенотипов группы М2, что в свою очередь приводит к различным уровням экспрессии генов, связанных с конкретными фенотипами.

Ключевые слова: фенотипы макрофагов; регуляция экспрессии; протеомы; система ANDSystem; автоматический анализ текстов

Introduction

Macrophages are immune system cells that play a key role in processes such as: maintaining body homeostasis (Mosser et al., 2021), defense against infections (Zhang M., Wang, 2014), proinflammatory and anti-inflammatory responses (Xu et al., 2013), tissue regeneration with concomitant stimulation of proliferation (Wynn, Vannella, 2016), and many others. The ability of macrophages to exhibit different functions through polarization (changing their functional state depending on signals from the microenvironment) is associated with their unique plasticity (Mills, 2012; Gurvich et al., 2020). Polarization leads to macrophages acquiring various phenotypes – functional states characterized by unique morphological, molecular and functional features, depending on the polarization inducers: proteins, peptides, polysaccharides, etc.

Each macrophage phenotype is characterized by a group of proteins (Martinez et al., 2008). These groups overlap, but different macrophage phenotypes can have radically different functions. For example, the M1 phenotype corresponds to proinflammatory macrophages, essential for the body's response to infections. M2a macrophages promote wound healing and clear the body of apoptotic cells (Murray et al., 2014). M2b macrophages are called regulatory for their ability to regulate T-helper cells, which leads to a switch in the immune response from proinflammatory to anti-inflammatory. M2c macrophages are necessary for tissue remodeling and the phagocytosis of apoptotic cells. M2d macrophages are called tumor-associated macrophages because they accompany tumor tissues (Zhang Q., Sioud, 2023).

In several studies, a link has been demonstrated between specific macrophage phenotypes and certain pathologies, as well as an association of disease outcomes with particular macrophage phenotypes. For example, patients with ovarian cancer exhibited a pronounced predominance of M1 phenotype macrophages over M2, which was associated with improved survival (Zhang M. et al., 2014). Additionally, the shift of macrophages from the M2 phenotype to M1 suppressed tumor metastasis (Yuan et al., 2017). Research on juvenile idiopathic arthritis in remission showed that the M2 macrophage group predominantly consisted of M2b and M2c, while the number of M2a macrophages was significantly reduced (Feng et al., 2021). In contrast, children with active juvenile idiopathic arthritis had a predominance of M2a and M2b macrophages, while the presence of M2c was decreased. The study of differences between macrophage phenotypes holds significant fundamental importance and also represents substantial practical interest for early disease diagnosis, prognosis, and management of disease progression (Zhang M. et al., 2014; Lampiasi, 2023).

It should be noted that there is conflicting information in the literature regarding the proteins and genes characterizing different macrophage phenotypes. For example, the fractalkine receptor (CX3CR1) is designated as a marker of the M2a phenotype in one publication (Joerink et al., 2011), while in another publication (Chhor et al., 2013), this protein is identified as a marker of the M1 phenotype. Metalloproteinase MMP12 is highlighted as a marker of the M1 phenotype (Hirani et al. 2021), but the article (Lee et al. 2014) shows that this protein is characteristic of the proteomes of the M2 phenotype and dendritic cells. The chemokine CXCL13 is described as an M1 marker in the study (Martinez et al. 2006), while in the work (van der Lans et al. 2015) it is noted as a marker of M2.

How do proteomes intersect in macrophages of different phenotypes to achieve significant functional differences? What molecular and genetic regulatory mechanisms underlie macrophage polarization? Despite the rapid accumulation of genome-wide transcriptomic and proteomic data characterizing the differences between the major macrophage phenotypes and their differences from non-activated macrophages (M0) (Gurvich et al., 2020; Oates et al., 2023), questions about how gene and protein expression profiles are regulated in macrophages of different phenotypes remain open.

The aim of this study was to identify mediator proteins that control the activity of phenotype-associated genes in different phenotypes of M2 macrophages. For this purpose, we used the ANDSystem information system, which is based on machine learning and artificial intelligence methods, including graph neural networks (Ivanisenko V.A. et al., 2015; Ivanisenko T.V. et al., 2024). ANDSystem provides automated analysis of scientific publication texts and factographic databases in the medical and biological domains. Currently, the ANDSystem knowledge base contains knowledge and facts extracted from more than 40 million scientific publications and patents, as well as factual databases, including information on molecular and genetic objects and processes that are important for the functioning of gene networks and their basic components: metabolic networks, signal transduction pathways, DNA-protein and protein-protein interaction networks. The effectiveness of ANDSystem has been demonstrated in a wide range of studies: reconstruction of molecular genetic mechanisms of asthma and hypertension comorbidity (Zolotareva et al., 2019), analysis of the plasma metabolome of patients with postoperative delirium (Ivanisenko V.A. et al., 2023), reconstruction of the hypermethylation regulatory network affecting the development of hepatocellular carcinoma in hepatitis C virus disease (Antropova et al., 2023).

In this work, the following tasks were addressed: 1) formation of phenotype-associated protein lists in macrophages of

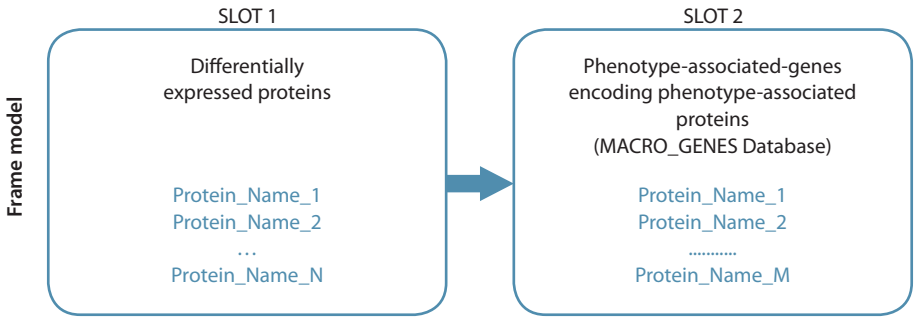


Fig. 1. Schematic diagram of a frame model for searching for regulatory links between differentially expressed proteins and phenotype-associated genes.

the main phenotypes (M1, M2a, M2b, M2c, M2d); 2) analysis of differential protein expression data in the M2 phenotype group as potential mediators of macrophage polarization; 3) analysis of regulatory relationships from mediator proteins to genes encoding phenotype-associated proteins using ANDSystem.

Materials and methods

Proteomic data on macrophages of different phenotypes. Two types of information about proteins in different macrophage phenotypes were used in the work:

1) Our curated database MACRO_GENES, containing lists of genes and proteins associated with macrophage phenotypes (Table S1)[1]. It was formed through manual analysis of scientific publications describing characteristic proteins that allow differentiation of macrophage phenotypes M1, M2a, M2b, M2c, M2d. Only those proteins, the presence of which in macrophages of certain phenotypes was confirmed by experimental data, were included in the MACRO_GENES database.
2) Proteomic data on differentially expressed proteins in M2a, M2b, M2c, and M2d macrophage phenotypes were obtained from the work by P. Li and colleagues (2022): approximately 200 proteins for each phenotype under consideration. Hereafter, such proteins will be referred to as regulatory proteins or differentially expressed proteins.

Search for potential regulators influencing the activity of phenotype-associated genes. The search for potential regulatory proteins influencing the activity of phenotype-associated genes was carried out using the knowledge base of the ANDSystem software and the ANDVisio software module included in this system (Demenkov et al., 2012; Ivanisenko V.A. et al., 2015; Ivanisenko T.V. et al., 2024). The ANDSystem knowledge base includes information on interactions between molecular biological objects (genes, proteins, metabolites, biological processes, etc.), obtained through automated analysis of over 40 million scientific publications and patents, as well as a large number of biomedical factual databases. The current version of this knowledge base contains information on over 36 million proteins from various organisms and approximately the same number of genes, 76 thousand metabolites, 100 million interactions, 21 thousand diseases, and more.

[1] Supplementary Tables S1–S3 and Figures S1–S6 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Antropova_Engl_29_7.pdf

To search for connections between regulatory proteins and phenotype-associated genes, the frame model software of the ANDSystem was used (Fig. 1). Step 1: The first slot of the frame was filled based on proteomic analysis data (Li et al., 2022) with a list of differentially expressed proteins for each phenotype (M2a, M2b, M2c, and M2d). Step 2: The second slot of the frame was filled with a list of phenotype-associated genes for the same phenotype from our curated MACRO_GENES database (Table S1). Step 3: Using the ANDVisio software module with the filled frame, regulatory connections described in the ANDSystem knowledge base were searched for the studied macrophage phenotype.

The analysis resulted in graphs of regulatory processes, in which nodes corresponded to differentially expressed proteins from the paper (Li et al., 2022) and phenotype-associated macrophage genes from the MACRO_GENES database. Edges connecting graph nodes corresponded to regulatory relationships between them.

Search for functionally significant regulatory proteins of phenotype-associated macrophage genes in regulatory process graphs. A key step in analyzing regulatory processes, associated with macrophage phenotype-associated genes and identified using frame models, is the search for functionally significant regulatory proteins (also called central nodes). Central nodes play a key role in signaling and coordinating regulatory processes. A wide range of methods have been developed to assess centrality (Ghasemi et al., 2014; Jalili et al., 2016; Ivanisenko V.A. et al., 2019). In our study, node centrality was assessed based on the number of interactions of the protein in question with phenotype-associated genes of the corresponding phenotype.

A high degree of centrality can be observed as a result of functional innovations between genes and proteins, as well as due to random factors. To distinguish between these situations, the statistical significance of the observed degree of centrality was assessed using the hypergeometric test. In this context, the hypergeometric test is used to measure the number of connections between a given protein and randomly determined phenotype-associated genes.

Here: M is the total number of genes represented in the ANDSystem knowledge base; N is the total number of genes with which a specific protein interacts in the ANDSystem knowledge base; n is the number of phenotype-associated

genes for a specific phenotype in the MACRO_GENES database; x is the observed number of interactions of the protein in question with phenotype-associated genes for a specific phenotype. Then, under the null hypothesis of a random distribution of interactions, the value of X obeys the hypergeometric distribution law:

$$X \sim Hypergeom(M, N, n),$$

and the p-value for the right-tailed (enrichment) test was calculated using the formula:

$$p = P(X \geq x) = \sum_{k=x}^{\min(N,n)} \frac{\binom{N}{k}\binom{M-N}{n-k}}{\binom{M}{n}}.$$

p-values were calculated using the SciPy Python library (scipy.stats.hypergeom). A Bonferroni correction was used to correct for multiple testing. At $p < 0.05$, the observed degree of centrality was considered statistically significant, and the corresponding protein was considered as a functionally significant regulatory protein controlling the expression of phenotype-associated genes.

Results and discussion

Our work aimed to identify regulatory proteins that influence genes, the expression of which differs between the M2a, M2b, M2c, and M2d macrophage phenotypes. Understanding the regulatory mechanisms that determine differences between macrophage phenotypes is not only of fundamental importance but also holds promise for applications in medicine and pharmacology, as the prevalence of a certain macrophage phenotype has been shown to be associated with the development and outcome of a number of pathologies (Zhang M. et al., 2014; Yuan et al., 2017; Feng et al., 2021).

General characteristics of phenotype-associated genes and proteins of macrophages M1, M2a, M2b, M2c, M2d

Table 1 presents a summary of our curated database, MACRO_GENES, of phenotype-associated genes encoding phenotype-associated proteins, i. e., proteins specific to macrophages of each of the phenotypes under consideration: M1, M2a, M2b, M2c, and M2d. The presence of proteins in specific phenotypes was confirmed by experimental data presented in the relevant publications. A detailed description of the gene information in MACRO_GENES is given in Table S1.

Some phenotype-associated proteins are used in experimental studies as specific markers for distinguishing macrophage phenotypes. In Table 1, the genes encoding such proteins are highlighted in green. If a protein is characteristic of a specific phenotype but is also considered a specific marker for another phenotype, the gene encoding it is highlighted in orange (Table 1). For example, the CCL2 protein is considered a marker for the M1 phenotype, but some publications indicate that it is also characteristic of the M2a and M2d phenotypes. Table 1 illustrates the complex pattern of marker intersections between different macrophage phenotypes.

Figure 2 shows a Venn diagram demonstrating the distribution of genes encoding phenotype-associated proteins across five macrophage types (M1, M2a, M2b, M2c, M2d). The diagram is constructed based on the information provided in Table 1. Note that the M1 and M2d phenotypes have the most matching proteins (17). The M2a/M2c and M1/M2c phenotype pairs have 15 and 13 common proteins, respectively. The M2b and M1 phenotypes have 11 matching proteins. A relatively small number of matching proteins (8) can be noted for the M2c and M2d phenotype pair. M2b has the fewest overlaps (6 proteins) with M2a.

Table 1. Lists of genes encoding phenotype-associated proteins of macrophages M1, M2a, M2b, M2c and M2d presented in the MACRO_GENES database

Macrophage phenotype	Genes encoding phenotype-associated proteins*
M1	*ARGI2, CAHM6, CCL2, CCL3, CCL4, CCL5, CCL7, CCL8, CCL15, CCL19, CCL20, CCR2, CCR7, CD38, CD80, CD86, CSF2, CXCL2, CXCL5, CXCL9, CXCL10, CXL11, FCG2B, FCG3A, FCGR1, GBP2, GBP5, GROA, HLA-DRA, HMGB1, IDO1, IFNA1, IFNB, IFNG, IL1A, IL1B, IL1R1, IL2RA, IL3, IL6, IL7RA, IL8, IL12B, IL15, IL15RA, IL17, IL18, IL23A, IRF1, IRF4, IRF7, ISG20, ITGAX, KCNN2, LAG3, MARCO, MET, MIF, MMP13, NAMPT, NFKB1, NOS2, PGH2, SOCS3, STAT1, TIMP1, TLR2, TLR4, TNFA, TNR5, TSP1, UBD, VEGFA*
M2a	*ALOX15, ARG1, CCL1, CCL2, CCL7, CCL8, CCL13, CCL14, CCL15, CCL17, CCL18, CCL22, CCL23, CCL24, CCL26, CCR2, CD200R1, CD209, CD274, CDH1, CDK11B, CLEC4A, CLEC7A, CLEC10A, CSF1R, CXCR1, CXCR2, EDN1, EGR2, FCER2, FGF2, FLT1, FN1, HAVCR2, HLA-DPA1, HLA-DPB1, HLA-DRA, HLA-DRB1, HLA-DRB3, HRH1, IGF1, IL1R1, IL1R2, IL4, IL10, IL27, IRF4, ITGAX, KLF2, LGALS3, MMP9, MMP14, MRC1, MRC2, MYC, PDCD1LG2, PGF, PPARG, PTGS1, RAMP1, SCN3A, SOCS1, TGFB1, TGM2, TREM2, VEGFA*
M2b	*CCL1, CCL4, CCL20, CD86, CD163, COX2, CXCL3, HLA-DRA, IFNA1, IL1B, IL1R1, IL6, IL10, MRC1, NOS2, PTPRC, SIGLEC10, SPHK1, TNFA, TNFSF14, VEGFA*
M2c	*ARG1, BCL3, C1QA, CCL8, CCL16, CCL18, CCL23, CCR2, CCR3, CCR10, CD14, CD163, CD300E, CDK11A, CDK11B, CSF1R, CX3CR1, CXCL12, CXCL13, CXCR4, EPAS1, F5, FCRLA, FPR1, GAS6, GXYLT2, HIF1A, IL1B, IL1R1, IL4R, IL10, IL21R, IRF3, IRF5, IRF8, ITGAX, JAK3, LIN7A, MAF, MARCO, MCTP2, MERTK, MMP2, MMP8, MMP14, MRC1, MRC2, MSR1, NOS2, PCOLCE2, PGF, PLOD2, SELENOP, SERPINA1, SH3PXD2B, SLAMF1, SOCS3, SPP1, SRPX2, STAT1, STAT3, STAT6, TGFB1, THBS1, TIMP1, TLR1, TLR2, TLR4, TLR5, TLR8, VCAN, VTCN1*
M2d	*ADORA2A, AIF1, C1QA, C1QC, CCL2, CCL3, CCL4, CCL5, CCL7, CD81, CD274, COX2, CSF3R, IL8, CXCL9, CXCL10, CXCL16, EGF, FCRL2, FGF2, FGFR1, GDF15, HLA-DMA, ID3, IDO1, IDO2, IL1A, IL1B, IL6, IL10, IRF7, LILRB4, MIF, MMP2, MMP9, MRC1, MSR1, NCAM1, NOS2, PDCD1LG2, PDGFB, TBX6, TGFBI, TNFA, VEGFA*

* Genes encoding proteins that are markers of various macrophages phenotypes are highlighted in green. Genes that are expressed in macrophages of a particular phenotype, according to some sources, but are markers of macrophages of a different phenotype, according to other sources, are highlighted in orange.

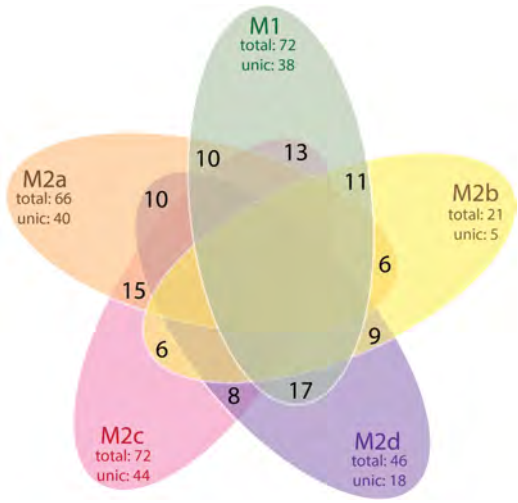


Fig. 2. Venn diagram for comparison of macrophage phenotypes M1, M2a, M2b, M2c, M2d according to the gene lists presented in the MACRO_GENES database.

General characteristics of differentially expressed proteins of the M2 macrophage group

To search for mediator proteins that transmit signals from macrophage polarization inducers to phenotype-associated genes, we used lists of differentially expressed proteins (compared to non-activated macrophages) from P. Li et al. (2022). The published data, summarized in Tables S2 and S3, indicate that the distribution of differentially expressed proteins across the four macrophage phenotypes (M2a, M2b, M2c, M2d) is characterized by significant overlap, i. e., there is no one-to-one correspondence between differentially expressed proteins and macrophage phenotypes. Therefore, to identify regulatory pathways that mediate macrophage polarization into different phenotypes, we required bioinformatic analysis of large volumes of molecular genetic data, conducted using the ANDSystem computer system.

Search for regulatory links from differentially expressed proteins to phenotype-associated genes of macrophages based on frame models

To analyze large volumes of published data on various macrophage phenotypes, we used the methods and information resources of computer-aided knowledge engineering implemented in the ANDSystem. Using the framework-based approach realized in this system, we searched for regulatory links between differentially expressed proteins and phenotype-associated genes in macrophages.

Regulatory process graphs were reconstructed, with nodes corresponding to differentially expressed proteins from the paper (Li et al., 2022) and phenotype-associated macrophage genes from the MACRO_GENES database. Edges connecting graph nodes corresponded to regulatory relationships between them. Figure 3 shows an example of a graph of potential regulatory relationships between differentially expressed proteins and phenotype-associated macrophage genes in the M2b phenotype.

Figure 3 shows that most phenotype-associated genes are regulated by more than one protein. Furthermore, most of the proteins shown in the figure are involved in the regulation of multiple genes. Similar regulatory relationship diagrams for M2a, M2c, and M2d macrophages are presented in the Supplementary Materials (Fig. S1–S3).

Identification of statistically significant regulators of phenotype-associated genes

Quantitative characteristics of regulatory links between differentially expressed proteins and phenotype-associated genes identified using frame models are shown in Section A of Table 2.

In the second stage of the analysis, centrality metrics characterizing the functional significance of differentially expressed proteins for the regulation of phenotype-associated genes were assessed. Centrality assessments allowed us to identify proteins regulating phenotype-associated genes with a Bonferroni-corrected statistical significance threshold of $p < 0.05$ (Table 2B). Accounting for the statistical significance of differentially expressed proteins based on centrality metrics led to a significant reduction in the number of nodes corresponding to phenotype-associated genes and the number of edges corresponding to regulatory events. For example, for the M2a phenotype, of the 40 differentially expressed proteins associated with phenotype-associated genes, 16 were statistically significantly associated with these genes (Table 2B). For the M2b phenotype, it was 6 out of 12 proteins. Accordingly, the number of regulatory events and phenotype-associated genes in the reconstructed graphs decreased.

Figure 4 shows the lists of differentially expressed proteins statistically significantly associated with the phenotype-associated genes of macrophages M2a, M2b, M2c and M2d. Proteins, the levels of which are elevated in specific macrophage phenotypes according to the study (Li et al., 2022), are highlighted in red. Proteins, the levels of which are decreased compared to non-activated macrophages are highlighted in blue. Green lines connect proteins with oppositely changing expression levels in macrophages of different phenotypes.

Figure 5 shows examples of schemes of statistically significant regulatory interactions between differentially expressed proteins of M2a macrophages and phenotype-associated genes reconstructed using the ANDSystem. The corresponding schemes for macrophages of the M2b, M2c, and M2d phenotypes are presented in Figures S4–S6. Figure 5 demonstrates regulatory connections using two alternative options for controlling molecular genetic processes in the same M2a macrophage phenotype: through an increase (Fig. 5*A*) and a decrease (Fig. 5*B*) in the levels of regulatory proteins. A description of the reconstructed connections obtained using frame models is given in Table 3.

As an example, let us consider the binding of the regulatory protein TGM2 (the lower protein in Figure 5*A*, marked with a blue asterisk). Proteomic data (Li et al., 2022) show that the level of this protein is elevated in the M2a phenotype compared to non-activated macrophages. According to information from the ANDSystem knowledge base obtained

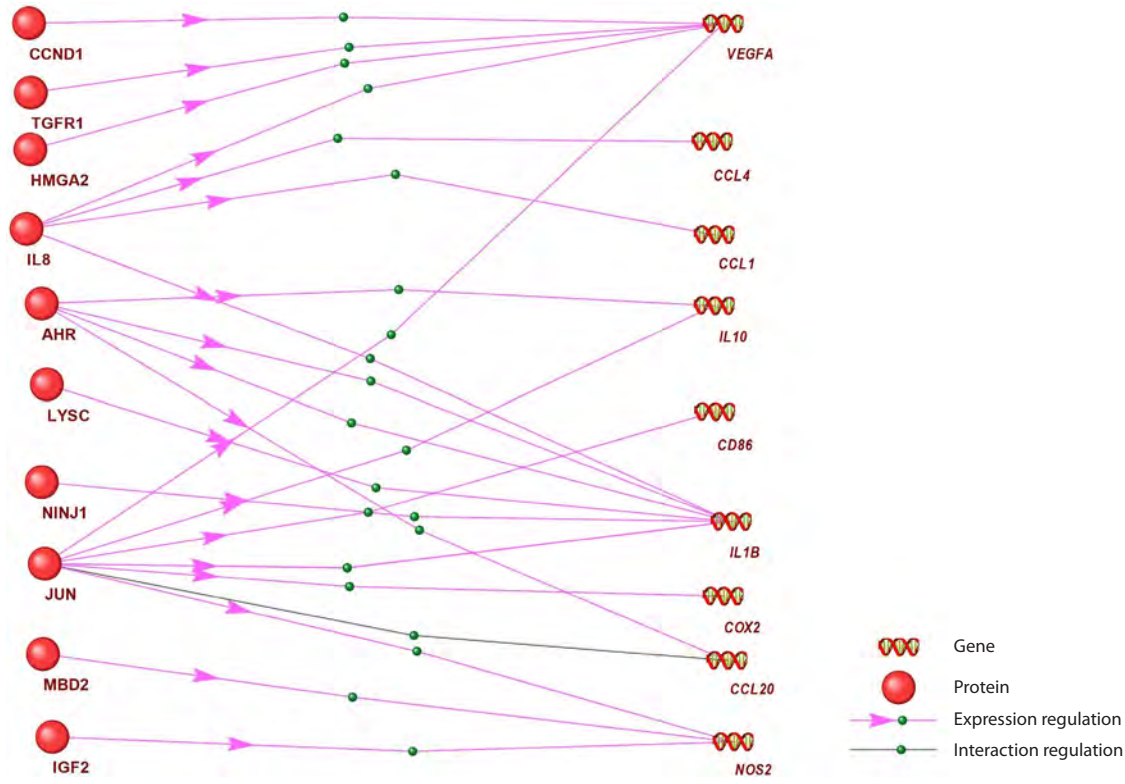


Fig. 3. A graph of potential regulatory links between differentially expressed proteins (left) and phenotype-associated genes (right) in M2b macrophages, presented in the ANDSystem interface. Green balls on the arrows in the interactive ANDSystem interface allow users to obtain additional information about specific regulatory links.

Table 2. Quantitative characteristics of potential regulatory links identified based on frame models

Components of frame models	M2a	M2b	M2c	M2d
A. Results of the first stage of analysis of the regulatory interactions reconstructed graph				
Differentially expressed proteins	40	12	41	43
Regulatory events	127	51	216	252
Phenotype-associated genes	26	12	41	31
B. Results of the second stage of the analysis (taking into account statistical estimates of the centrality of regulatory proteins, $p < 0.05$)				
Differentially expressed proteins	16	6	10	15
Regulatory events	85	19	89	133
Phenotype-associated genes	23	8	28	29

through its interface, in M2a macrophages, the TGM2 protein has an activating effect on the expression of the M2a phenotype-associated genes *CD274* and *FN1* (Liu et al., 2021; Sun et al., 2021), which is consistent with the data presented in Table 1 (the mentioned genes are marked with black asterisks in Figure 5*A*). TGM2 also has a suppressive effect on the *PPARG* gene (Maiuri et al., 2008), which is inconsistent with the data in Table 1 and indicates that the expression of this gene is also activated by some other factors, such as AHR, CDK4, CCL5, the level of which is increased in this phenotype (Fig. 5*A*).

Among the proteins with reduced levels (compared to non-activated macrophages) in M2a macrophages, as an example we consider the CBP protein, which regulates the phenotype-associated genes *CCL2*, *CD274*, and *CDH1* (Fig. 5*B*, marked with asterisks). According to information from the

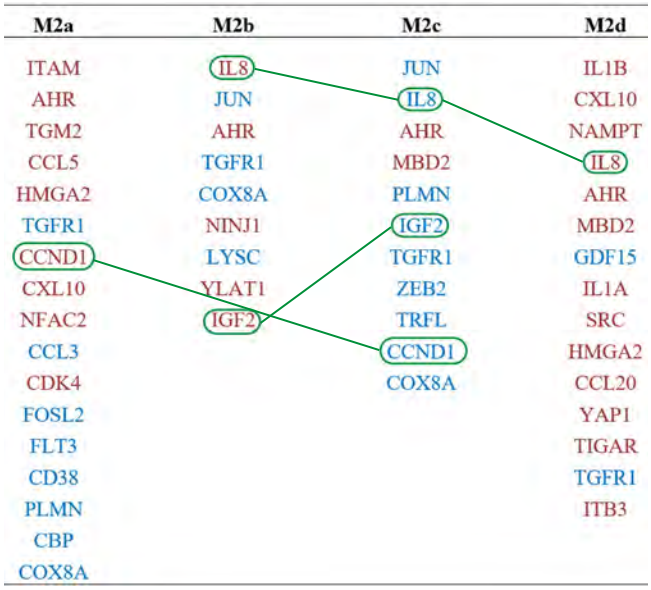


Fig. 4. Differentially expressed proteins statistically significantly ($p < 0.05$) associated with the phenotype-associated genes of macrophages M2a, M2b, M2c, and M2d. Proteins, the levels of which, according to (Li et al. 2022), are elevated in a particular phenotype are highlighted in red, while those, the levels of which are decreased compared to non-activated macrophages, are highlighted in blue. Green lines connect proteins with oppositely expressed changes in macrophages of different phenotypes.

ANDSystem knowledgebase, when CBP is suppressed in the M2a phenotype, *CCL2* expression increases (Huang et al., 2021), which is consistent with the data presented in Table 1 (MACRO_GENES database). At the same time, the CBP protein positively influences the expression of the phenotype-associated genes *CD274* and *CDH1* (Liu et al., 2020; Heng et al., 2021). It can be hypothesized that other regulators have a greater influence on the activity of these genes. Figure 5*A* shows that such regulators for the *CD274* gene may include the proteins AHR, CCL5, TGM2, and CDK4, the levels of which are elevated in the M2a phenotype (Fig. 5*A*, double green asterisks).

All statistically significant regulatory interactions identified in M2 macrophages between differentially expressed proteins and phenotype-associated genes are presented in Table 3. For M2a macrophages, these were interactions of nine regulatory proteins with increased levels (compared to non-activated macrophages), marked with arrows (↑), and four proteins with decreased levels (↓), regulating 23 phenotype-associated genes. For M2b, these were four upregulated and two downregulated proteins regulating eight phenotype-associated genes (see also Figure S4). For M2c, two upregulated and eight downregulated proteins regulating 28 genes were identified (see also Figure S5). For M2d, 13 upregulated and two downregulated proteins regulating 29 genes were found (see also Figure S6).

Thus, based on a computer analysis of differences in the proteomes of different macrophage phenotypes, as well as the use of large volumes of information accumulated in the ANDSystem knowledge base, some regulatory proteins were identified that mediate the action of macrophage polarization inducers on phenotype-associated macrophage genes. Future research is planned using frame models containing more slots reflecting the intermediate stages of action of macrophage polarization inducers on phenotype-associated macrophage genes. This will enable the identification of more subtle features of the regulatory pathways running from macrophage polarization inducers to phenotype-associated genes through the action of intermediary proteins.

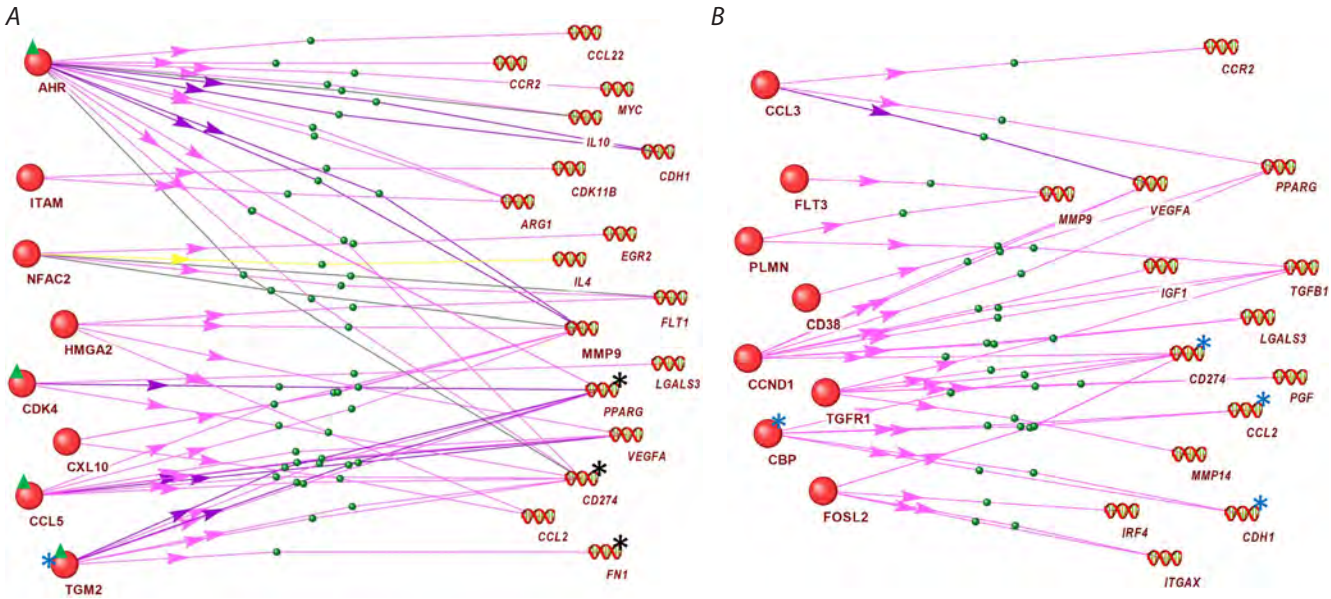


Fig. 5. Graph of the regulation of phenotype-associated gene expression (from the MACRO_GENES database) in M2a macrophages by differentially expressed regulatory proteins from the article (Li et al., 2022), statistically significantly associated with these genes: (*A*) through an increase and (*B*) through a decrease in the level of regulatory proteins in this phenotype. The blue asterisk in (*A*) indicates the TGM2 protein discussed in the text, black asterisks indicate its target genes; green triangles indicate the discussed proteins AHR, CDK4, CCL2, TGM2. Blue asterisks in (*B*) indicate the CBP protein discussed in the text and its target genes.

Table 3. Relationships between functionally significant differentially expressed regulatory proteins* and the phenotype-associated genes they regulate in M2 group macrophages, identified using frame models

M2a		M2b		M2c		M2d	
Protein	Target genes	Protein	Target genes	Protein	Target genes	Protein	Target genes
AHR (↑)**	*ARG1, MMP9, MYC, PPARG, CCL22, CCR2, CDH1, CD274, IL10*	AHR (↑)	*IL1B, CCL20, IL10*	AHR (↑)	*ARG1, CXCL12, STAT3, SOCS3, CCR2, HIF1A, TLR2, IL1B, IL10*	AHR (↑)	*IDO1, IDO2, IL1B, IL8,MMP9, NCAM1, CD274, IL1A, IL10*
TGM2 (↑)	*PPARG, CD274, FN1, MMP14*	IGF2 (↑)	*NOS2*	MBD2 (↑)	*CXCL12, NOS2, SOCS3*	IL1B (↑)	*FGF2, FGFR1, CXCL10, IL8, MMP9, VEGFA, TGFB1, CCL2, CCL3, CCL4, CCL5, NOS2, COX2, IL10, IL1A, TNF, PDCD1LG2, MIF*
CCL5 (↑)	*VEGFA, CD274, MMP9*	NINJ1 (↑)	*IL1B*				
NFAC2 (↑)	*EGR2, IL4, FLT1, MMP9*	IL8 (↑)	*CCL1, VEGFA*	TGFR1 (↓)	*MMP14, TGFB1, HIF1A, PGF*	MBD2 (↑)	*CXCL9, CXCL10, CD274, CCL5, MMP2, NOS2*
HMGA2 (↑)	*FLT1, MMP9, VEGFA, CCL2*	TGFR1 (↓)	*VEGFA*	ZEB2 (↓)	*MAF, MMP14, TGFB1,*	IL8 (↑)	*IL1B, FGF2, CXCL10, MMP9, CCL2, CD274, MMP2, VEGFA*
ITAM (↑)	*ARG1, CDK11B*	JUN (↓)	*CD86, COX2, NOS2, IL1B, IL10, CCL20, VEGFA*	CCND1 (↓)	*TGFB1, STAT3, HIF1A*	YAP1 (↑)	*IRF7, FGFR1, PDGFB, CD274*
CDK4 (↑)	*PPARG, LGALS3, CD274*			IL8 (↓)	*MMP14, MMP2, MMP8, TLR5, CXCL12, ARG1, CCR2, HIF1A, IL1B, CDK11B*	NAMPT (↑)	*IL1B, FGF2, IL8, MMP9, CCL2, CD274, MMP2*
CXL10 (↑)	*MMP9, CCL2*						
CCND1 (↓)	*VEGFA, TGFB1, CD274*			IGF2 (↓)	*CDK11B, MMP2, HIF1A, NOS2*	IL1A (↑)	*IL6, IL1B, VEGFA, CD274*
						CCL20 (↑)	*MMP9, CD274*
FLT3 (↓) ***	*MMP9*			PLMN (↓)	*TGFB1, CXCL12, CXCR4*	CXL10 (↑)	*IL1B, MMP9, CCL2, MMP2*
CCL3 (↓)	*CCR2, VEGFA, PPARG*			TRFL (↓)	*HIF1A, TLR4*	TIGAR (↑)	*MMP9, MMP2*
PLMN (↓)	*MMP9, TGFB1*			JUN (↓)	*TGFB1, CXCL12, MMP14, MMP2, ARG1, IL1B, NOS2, HIF1A, TLR2, TLR4, IL10, CX3CR1, THBS1, SERPINA1, BCL3, CCL18, ITGAX*	HMGA2 (↑)	*MMP9, CCL2, MMP2, VEFGA*
CD38 (↓)	*VEGFA*					SRC (↑)	*CXCL10, MMP2, VEFGA, NOS2*
TGFR1 (↓)	*TGFB1, CD274, MMP14, PGF*					ITB3 (↑)	*MMP2*
CBP (↓)	*CD274, CCL2, CDH1*					TGFR1 (↓)	*VEGFA, TGFB1, CD274*
FOSL2 (↓)	*CD274, IRF4, ITGAX*					GDF15 (↓)	*MMP9, CCL2, GDF15, CD274, MMP2*

* – Proteins selected based on the centrality criterion ($p < 0.05$); ** ↑ – proteins with increased expression levels; ***↓ – proteins with decreased expression levels.

Conclusion

A study of published data on phenotype-associated genes and proteomes of M2 macrophages, and a subsequent search for regulatory links between them using a frame-based approach implemented in the ANDSystem computer system, made it possible to identify potential regulatory proteins that mediate differences in gene expression in M2 macrophage phenotypes. The obtained results suggest that the differences between the M2a, M2b, M2c and M2d phenotypes may be associated, in particular, with the regulatory functions of the proteins JUN, IL8, NFAC2, CCND1 and YAP1, the level of which varies between phenotypes, leading to differences in the expression of phenotype-associated genes.

References

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Volianskaia A.R., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Kolchanov N.A., Lavrik I.N., Ivanisenko V.A. Reconstruction of the regulatory hypermethylation network controlling hepatocellular carcinoma development during hepatitis C viral infection. *J Integr Bioinform*. 2023;20(3): 20230013. doi 10.1515/jib-2023-0013

Chhor V., Le Charpentier T., Lebon S., Oré M.V., Celador I.L., Josserand J., Degos V., Jacotot E., Hagberg H., Sävman K., Mallard C., Gressens P., Fleiss B. Characterization of phenotype markers and neuronotoxic potential of polarised primary microglia *in vitro*. *Brain Behav Immun*. 2013;32:70-85. doi 10.1016/j.bbi.2013.02.005

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Feng D., Huang W.Y., Niu X.L., Hao S., Zhang L.N., Hu Y.J. Significance of macrophage subtypes in the peripheral blood of children with systemic juvenile idiopathic arthritis. *Rheumatol Ther*. 2021; 8(4):1859-1870. doi 10.1007/s40744-021-00385-x

Ghasemi M., Seidkhani H., Tamimi F., Rahgozar M., Masoudi-Nejad A. Centrality Measures in Biological Networks. *Curr Bioinform*. 2014;9:426-441. doi 10.2174/1574893611308666013

Gurvich O.L., Puttonen K.A., Bailey A., Kailaanmäki A., Skirdenko V., Sivonen M., Pietikäinen S., Parker N.R., Ylä-Herttuala S., Kekarainen T. Transcriptomics uncovers substantial variability associated with alterations in manufacturing processes of macrophage cell therapy products. *Sci Rep*. 2020;10(1):14049. doi 10.1038/s41598-020-70967-2

Heng W.S., Kruyt F.A.E., Cheah S.C. Understanding lung carcinogenesis from a morphostatic perspective: prevention and therapeutic potential of phytochemicals for targeting cancer stem cells. *Int J Mol Sci*. 2021;22(11):5697. doi 10.3390/ijms22115697

Hirani D., Alvira C.M., Danopoulos S., Milla C., Donato M., Tian L., Mohr J., … Seeger W., Khatri P., Al Alam D., Dötsch J., Alejandre Alcazar M.A. Macrophage-derived IL-6 trans-signaling as a novel target in the pathogenesis of bronchopulmonary dysplasia. *Eur Respir J*. 2021;59(2):2002248. doi 10.1183/13993003.02248-2020

Huang Y.H., Cai K., Xu P.P., Wang L., Huang C.X., Fang Y., Cheng S., Sun X.J., Liu F., Huang J.Y., Ji M.M., Zhao W.L. *CREBBP/EP300* mutations promoted tumor progression in diffuse large B-cell lymphoma through altering tumor-associated macrophage polarization via FBXW7-NOTCH-CCL2/CSF1 axis. *Signal Transduct Target Ther*. 2021;6(1):10. doi 10.1038/s41392-020-00437-8

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst Biol*. 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019; 20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Basov N.V., Makarova A.A., Venzel A.S., Rogachev A.D., Demenkov P.S., Ivanisenko T.V., Kleshchev M.A., Gaisler E.V., Moroz G.B., Plesko V.V., Sotnikova Y.S., Patrushev Y.V., Lomivorotov V.V., Kolchanov N.A., Pokrovsky A.G. Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(7):768-775. doi 10.18699/VJGB-23-89

Jalili M., Salehzadeh-Yazdi A., Gupta S., Wolkenhauer O., Yaghmaie M., Resendis-Antonio O., Alimoghaddam K. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Front Physiol*. 2016;7:375. doi 10.3389/fphys.2016.00375

Joerink M., Rindsjö E., van Riel B., Alm J., Papadogiannakis N. Placental macrophage (Hofbauer cell) polarization is independent of maternal allergen-sensitization and presence of chorioamnionitis. *Placenta*. 2011;32(5):380-385. doi 10.1016/j.placenta.2011.02.003

Lampiasi N. Macrophage polarization: learning to manage it 2.0. *Int J Mol Sci*. 2023;24(24):17409. doi 10.3390/ijms242417409

Lee J.T., Pamir N., Liu N.C., Kirk E.A., Averill M.M., Becker L., Larson I., Hagman D.K., Foster-Schubert K.E., van Yserloo B., Bornfeldt K.E., LeBoeuf R.C., Kratz M., Heinecke J.W. Macrophage metalloelastase (MMP12) regulates adipose tissue expansion, insulin sensitivity, and expression of inducible nitric oxide synthase. *Endocrinology*. 2014;155(9):3409-3420. doi 10.1210/en.2014-1037

Li P., Ma C., Li J., You S., Dang L., Wu J., Hao Z., Li J., Zhi Y., Chen L., Sun S. Proteomic characterization of four subtypes of M2 macrophages derived from human THP-1 cells. *J Zhejiang Univ Sci B*. 2022;23(5):407-422. doi 10.1631/jzus.B2100930

Liu J., He D., Cheng L., Huang C., Zhang Y., Rao X., Kong Y., … Jones K., Napier D., Lee E.Y., Wang C., Liu X. p300/CBP inhibition enhances the efficacy of programmed death-ligand 1 blockade treatment in prostate cancer. *Oncogene*. 2020;39(19):3939-3951. doi 10.1038/s41388-020-1270-z

Liu J., Liu Q., Zhang X., Cui M., Li T., Zhang Y., Liao Q. Immune subtyping for pancreatic cancer with implication in clinical outcomes and improving immunotherapy. *Cancer Cell Int*. 2021;21(1):137. doi 10.1186/s12935-021-01824-z

Maiuri L., Luciani A., Giardino I., Raia V., Villella V.R., D'Apolito M., Pettoello-Mantovani M., Guido S., Ciacci C., Cimmino M., Cexus O.N., Londei M., Quarantino S. Tissue transglutaminase activation modulates inflammation in cystic fibrosis via PPARgamma down-regulation. *J Immunol*. 2008;180(11):7697-7705. doi 10.4049/jimmunol.180.11.7697

Martinez F.O., Gordon S., Locati M., Mantovani A. Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. *J Immunol*. 2006;177(10):7303-7311. doi 10.4049/jimmunol.177.10.7303

Martinez F.O., Sica A., Mantovani A., Locati M. Macrophage activation and polarization. *Front Biosci*. 2008;13:453-461. doi 10.2741/2692

Mills C.D. M1 and M2 macrophages: oracles of health and disease. *Crit Rev Immunol*. 2012;32(6):463-488. doi 10.1615/critrevimmunol.v32.i6.10

Mosser D.M., Hamidzadeh K., Goncalves R. Macrophages and the maintenance of homeostasis. *Cell Mol Immunol*. 2021;18(3):579-587. doi 10.1038/s41423-020-00541-3

Murray P.J., Allen J.E., Biswas S.K., Fisher E.A., Gilroy D.W., Goerdt S., Gordon S., … Suttles J., Udalova I., van Ginderachter J.A., Vogel S.N., Wynn T.A. Macrophage activation and polarization: nomenclature and experimental guidelines. *Immunity*. 2014;41(1): 14-20. doi 10.1016/j.immuni.2014.06.008

Oates T.C., Moura P.L., Cross S., Roberts K., Baum H.E., Haydn-Smith K.L., Wilson M.C., Heesom K.J., Severn C.E., Toye A.M. Defining the proteomic landscape of cultured macrophages and their polarization continuum. *Immunol Cell Biol*. 2023;101(10):947-963. doi 10.1111/imcb.12687

Sun W., Qin Y., Wang Z., Dong W., He L., Zhang T., Zhang H. The *NEAT1_2*/miR-491 axis modulates papillary thyroid cancer invasion and metastasis through TGM2/NFκb/FN1 signaling. *Front Oncol*. 2021;11:610547. doi 10.3389/fonc.2021.610547

van der Lans A.A., Boon M.R., Haks M.C., Quinten E., Schaart G., Ottenhoff T.H., van Marken Lichtenbelt W.D. Cold acclimation affects immune composition in skeletal muscle of healthy lean subjects. *Physiol Rep*. 2015;3(7):e12394. doi 10.14814/phy2.12394

Wynn T.A., Vannella K.M. Macrophages in tissue repair, regeneration, and fibrosis. *Immunity*. 2016;44(3):450-462. doi 10.1016/j.immuni.2016.02.015

Xu W., Zhao X., Daha M.R., van Kooten C. Reversible differentiation of pro- and anti-inflammatory macrophages. *Mol Immunol*. 2013;53(3):179-86. doi 10.1016/j.molimm.2012.07.005

Yuan R., Li S., Geng H., Wang X., Guan Q., Li X., Ren C., Yuan X. Reversing the polarization of tumor-associated macrophages inhibits tumor metastasis. *Int Immunopharmacol*. 2017;49:30-37. doi 10.1016/j.intimp.2017.05.014

Zhang M., Wang C.C. Inflammatory response of macrophages in infection. *Hepatobiliary Pancreat Dis Int*. 2014;13(2):138-152. doi 10.1016/s1499-3872(14)60024-2

Zhang M., He Y., Sun X., Li Q., Wang W., Zhao A., Di W. A high M1/M2 ratio of tumor-associated macrophages is associated with extended survival in ovarian cancer patients. *J Ovarian Res*. 2014;7:19. doi 10.1186/1757-2215-7-19

Zhang Q., Sioud M. Tumor-associated macrophage subsets: shaping polarization and targeting. *Int J Mol Sci*. 2023;24:7493. doi 10.3390/ijms24087493

Zolotareva O., Saik O.V., Königs C., Bragina E.Y., Goncharova I.A., Freidin M.B., Dosenko V.E., Ivanisenko V.A., Hofestädt R. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Sci Rep*. 2019;9(1):16302. doi 10.1038/s41598-019-52762-w

Conflict of interest. The authors declare no conflict of interest.

Received August 09, 2025. Revised October 16, 2025. Accepted October 17, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1000-1008

doi 10.18699/vjgb-25-105

In silico reconstruction of the gene network for cytokine regulation of ASD-associated genes and proteins

N.M. Levanova[1], E.G. Vergunov[1, 2, 3], A.N. Savostyanov[1, 2, 3], I.V. Yatsyk[1], V.A. Ivanisenko[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
levanova@bionet.nsc.ru

Abstract. Accumulated evidence links dysregulated cytokine signaling to the pathogenesis of autism spectrum disorder (ASD), implicating genes, proteins, and their intermolecular networks. This paper systematizes these findings using bioinformatics analysis and machine learning methods. The primary tool employed in the study was the ANDSystem cognitive platform, developed at the Institute of Cytology and Genetics, which utilizes artificial intelligence techniques for automated knowledge extraction from biomedical databases and scientific publications. Using ANDSystem, we reconstructed a gene network of cytokine-mediated regulation of autism spectrum disorder (ASD)-associated genes and proteins. The analysis identified 110 cytokines that regulate the activity, degradation, and transport of 58 proteins involved in ASD pathogenesis, as well as the expression of 91 ASD-associated genes. Gene Ontology (GO) enrichment analysis revealed statistically significant associations of these genes with biological processes related to the development and function of the central nervous system. Furthermore, topological network analysis and functional significance assessment based on association with ASD-related GO biological processes allowed us to identify 21 cytokines exerting the strongest influence on the regulatory network. Among these, eight cytokines (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α) had the highest priority, ranking at the top across all employed metrics. Notably, eight of the 21 prioritized cytokines (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) are known targets of drugs currently used as immunosuppressants and antitumor agents. The pivotal role of these cytokines in ASD pathogenesis provides a rationale for potentially repurposing such inhibitory drugs for the treatment of autism spectrum disorders.

Key words: autism spectrum disorder (ASD); neurodevelopmental disorders; cytokines; automatic text analysis of scientific publications; ASD pathogenesis; ASD treatment; computer reconstruction of gene networks

For citation: Levanova N.M., Vergunov E.G., Savostyanov A.N., Yatsyk I.V., Ivanisenko V.A. *In silico* reconstruction of the gene network for cytokine regulation of ASD-associated genes and proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1000-1008. doi 10.18699/vjgb-25-105

Funding. The research was funded by the state budget project No. FWNR-2022-0020 and carried out at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (ICG SB RAS).

Компьютерная реконструкция генной сети цитокиновой регуляции генов и белков, ассоциированных с РАС

Н.М. Леванова[1], Е.Г. Вергунов[1, 2, 3], А.Н. Савостьянов[1, 2, 3], И.В. Яцык[1], В.А. Иванисенко[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
levanova@bionet.nsc.ru

Аннотация. Многочисленные исследования подтверждают связь нарушений цитокиновой регуляции с развитием расстройств аутистического спектра (РАС) на уровне генов, белков и их межмолекулярных взаимодействий. В работе эти данные были систематизированы с применением биоинформатического анализа и методов машинного обучения. Главным инструментом в исследовании являлась когнитивная система ANDSystem, разработанная в Институте цитологии и генетики СО РАН и задействующая методы искусственного интеллекта для автоматического извлечения информации из биомедицинских баз данных и текстов научных публикаций. С использованием ANDSystem была реконструирована ассоциативная генная сеть цитокиновой регуляции генов и белков, ассоциированных с РАС. В результате анализа удалось идентифицировать 110 цитокинов, которые, согласно воссозданной сети, регулируют активность, деградацию и транспорт 58 белков, вовлеченных в развитие РАС, а также экспрессию 91 гена, ассоциированного с этими расстройствами. Анализ перепредстав-

ленности биологических процессов Gene Ontology выявил статистически значимые ассоциации этих генов с процессами, связанными с развитием и работой центральной нервной системы. Анализ топологических характеристик сети и оценка функциональной значимости элементов сети через их ассоциацию с биологическими процессами Gene Ontology, связанными с РАС, позволили выделить 21 цитокин, оказывающий наибольшее влияние на элементы сети. Среди них наибольший приоритет имели восемь цитокинов (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α), которые занимали высокое положение по результатам всех использованных методик приоритизации. Кроме того, из 21 приоритетного цитокина выделяются восемь цитокинов (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17), которые являются мишенями препаратов, применяемых в качестве иммуносупрессантов и противоопухолевых средств. Установленная роль этих цитокинов в патогенезе РАС создает предпосылки для потенциального перепрофилирования препаратов, направленных на их ингибирование, для терапии расстройств аутистического спектра.

Ключевые слова: расстройства аутистического спектра (РАС); нарушения нейроразвития; цитокины; автоматический анализ текстов научных публикаций; патогенез РАС; терапия РАС; компьютерная реконструкция генных сетей

Introduction

DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth edition) classifies autism spectrum disorder (ASD) as a category of neurodevelopmental conditions exhibiting a substantial genetic component, with diagnosis predicated solely on behavioral criteria (American Psychiatric Association, 2013). The core diagnostic profile of ASD comprises persistent deficits in social communication and reciprocal social interaction, co-occurring with restricted, repetitive patterns of behavior, interests, or activities. Contemporary diagnostic frameworks mandate the manifestation of these symptoms during the early developmental period. While their severity can vary, certain individuals may develop compensatory strategies through learned behaviors, which can mask underlying deficits. A substantial heterogeneity is observed in the behavioral phenotypes associated with ASD (Van der Zee, Derksen, 2021). Furthermore, the neurophysiological features associated with autism were identified not only in diagnosed individuals but also in the general population (Harms et al., 2010; Tsai et al., 2013; Tseng et al., 2015).

ASD classification delineates idiopathic forms, lacking clear genetic correlates, from syndromic forms, which are defined by monogenic mutations and associated comorbid features (Ziats et al., 2021). A considerable subset of syndromic ASD cases is driven by mutations disrupting the mTOR signaling pathway, leading to its persistent hyperactivation (Ganesan et al., 2019). A prior bioinformatic analysis utilizing the SFARI Gene database (Abrahams et al., 2013) demonstrated that approximately 58 % of genes harboring ASD-associated mutations are directly linked to the mTOR signaling pathway (Trifonova et al., 2019). The mTOR protein (mechanistic target of rapamycin) is a serine/threonine kinase that serves as the central component of two protein complexes: mTORC1 and mTORC2. Rapamycin-sensitive mTORC1 responds to nutrient availability and growth factors, regulating cell growth and metabolism. mTORC2, in contrast, is largely rapamycin-insensitive and is activated in response to stress and growth factor signaling, regulating cell survival and proliferation processes (Ragupathi et al., 2024).

mTOR signaling pathway plays a critical regulatory role in diverse physiological processes, including cellular and tumor growth (Onore et al., 2017), immune function (Liu et al., 2015), as well as memory formation and neural circuit plasticity (Hoeffer, Klann, 2010). Furthermore, constitutive hyperactivation of this pathway has been shown to suppress autophagy (McMahon et al., 2012) and impair normal synaptic pruning mechanisms (Tang et al., 2014).

Synaptic pruning is a fundamental neurodevelopmental process involving the microglia-mediated elimination of superfluous synaptic connections persisting from development through adulthood. This refinement mechanism enhances the efficiency of neural transmission and facilitates the reallocation of metabolic and computational resources to behaviorally relevant circuits, thereby underlying effective learning and long-term memory formation (Navlakha et al., 2015). Impairments in this pruning cascade are implicated in the neuropathology of ASD, manifesting as an increase in dendritic spine and synaptic density across both supra- and infragranular layers of the frontal, temporal, and parietal cortices (Hutsler, Zhang, 2010).

Microglia, central to the process of synaptic pruning, are integral to the CNS immune environment, where their activity is modulated by cytokine signalling. Moreover, as a major source of pro-inflammatory cytokines in the brain, microglia function as critical orchestrators of neuroinflammatory processes and possess the capacity to induce or modulate diverse cellular responses (Smith et al., 2012). Postmortem analyses of individuals with ASD have revealed hallmarks of neuroinflammation associated with classical (M1) microglial activation, with documented elevations in interferon IFN-γ and cytokines IL-1β, IL-6, IL-12p40, TNF-α, and CCL2 in both brain tissue and cerebrospinal fluid (Vargas et al., 2005; Li et al., 2009; Morgan et al., 2010).

Cytokines provide regulatory signaling essential for normal early brain development, synaptic plasticity, and the preservation of brain homeostasis. Pronounced alterations in the cytokine milieu disrupt fundamental neurodevelopmental mechanisms such as neuronal migration and differentiation, ultimately leading to the emergence of behavioral deficits (Ashwood et al., 2011). Moreover, comparative analyses of plasma and serum cytokine levels further reveal statistically significant alterations in the immunological profile of individuals with ASD relative to neurotypical controls (Onore et al., 2017). Therefore, a systemic immune regulatory imbalance perpetuates a state of chronic neuroinflammation in ASD.

In this study, we employed artificial intelligence (AI)-based software tools to reconstruct associative gene networks, aiming to identify and systematize regulatory interactions

between cytokines and ASD-associated genes and proteins. The analysis was performed using the ANDSystem cognitive platform (Ivanisenko V.A. et al., 2015), a tool specifically designed for automated extraction and integration of data from scientific literature and biological databases.

The objective of this research was to reconstruct and analyze the gene network of cytokine-mediated regulation of ASD-associated genes and proteins, with the specific goal of identifying promising cytokine targets for ASD immunomodulation therapy.

Network analysis identified 110 cytokines regulating activity, degradation, and transport of 58 ASD-associated proteins, alongside influencing the expression of 91 ASD-related genes. Gene Ontology enrichment analysis revealed significant involvement of these genes in CNS development and function. Among the 21 cytokines exerting the greatest influence on the network, eight (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) are targeted by existing immunosuppressive and antitumor drugs. The identified role of these cytokines in ASD pathogenesis provides a strong foundation for exploring drug repurposing strategies targeting them.

Materials and methods

The study's first phase involved *in silico* reconstruction of a network mapping cytokine interactions with ASD-associated proteins and genes (consolidated gene network). To achieve the most comprehensive coverage of these regulatory interactions, five specialized gene subnetworks reflecting different pathways of cytokine influence were first reconstructed (Supplementary Table S1)[1]. These subnetworks were subsequently integrated into a consolidated gene network.

The second phase comprised a structural bioinformatic analysis of the integrated network and functional annotation of its components using Gene Ontology to identify ASD-relevant biological processes. This was followed by prioritization of cytokines according to their predicted regulatory impact on ASD-associated genes and proteins.

The final stage focused on identifying promising targets for immunomodulatory ASD therapy among the cytokines demonstrating the highest significance in the conducted analysis.

Stage 1. A set of ASD-associated genes (234 genes) was obtained from the SFARI Gene database (Abrahams et al., 2013) (https://gene.sfari.org). The sample included genes annotated in this database as having a high confidence of association with ASD (Category 1 according to the database's internal scoring system). Lists of cytokine genes (186 genes) and cytokine receptor genes (114 genes) were compiled using data extracted from the Human Protein Atlas (HPA) (https://www.proteinatlas.org/), a comprehensive knowledge base focused on the spatial localization and expression profiles of human proteins in tissues, cells, and organs (Uhlén et al., 2015).

Gene networks were reconstructed using the ANDVisio software (Demenkov et al., 2012), which utilizes data from the ANDSystem's knowledge base for network reconstruction and structural analysis. ANDSystem is designed for automated analysis of scientific publications and databases and employs ontological modeling, graph analysis, and natural language processing mechanisms (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2020, 2022, 2024).

A consolidated network was assembled from subnetworks reconstructed using ANDVisio's 'Pathway Wizard'. This tool enables the retrieval and visualization of gene networks from the ANDSystem knowledge base that match specified query templates. Five individual subnetworks were initially constructed using five distinct query templates (Table S1) and subsequently merged into a unified graph.

Stage 2. Gene Ontology (GO) term enrichment analysis for biological processes (GO_BP) (Ashburner et al., 2000) was performed on the consolidated gene network utilizing the DAVID bioinformatics platform (Huang et al., 2009; Sherman et al., 2022) (https://davidbioinformatics.nih.gov/). DAVID provides functional gene annotation and evaluates the statistical significance of GO term enrichment within gene sets against user-defined confidence thresholds.

Network topology analysis and cytokine ranking were performed using the statistical tools implemented in ANDVisio. Cytokines were evaluated based on two centrality metrics: betweenness centrality, defined as the fraction of the shortest paths traversing a node, and degree centrality, representing the number of its direct connections. Both parameters serve as measures of nodal influence within the network, where higher values correspond to greater functional significance. Furthermore, pathway-based prioritization of cytokines was conducted using a custom Python 3.10 script to assess their representation in ASD-associated biological pathways.

Stage 3. Cytokines identified through prior analysis were subsequently evaluated as potential targets for pharmacological intervention. This assessment incorporated data from the DrugBank (Knox et al., 2024) (https://go.drugbank.com/) and GETdb (Zhang et al., 2024) (https://togodb.org/db/getdb) databases.

Results of gene network reconstruction and analysis

Reconstruction of cytokine interactions with ASD-associated proteins and genes

During the initial research phase, five sub-networks were reconstructed using the Pathway Wizard software (Fig. 1). The subnetwork reconstruction utilized two datasets: ASD-associated gene set from the SFARI database (https://gene.sfari.org) and a list of cytokines and their receptors obtained from the Human Protein Atlas database (https://www.proteinatlas.org/).

Following automated reconstruction, all retrieved connections and network elements were manually reviewed against source publication texts to eliminate errors arising from inaccurate information extraction.

Integration of the reconstructed subnetworks produced a consolidated network representing cytokine interactions with ASD-associated proteins and genes (Fig. 2). This integrated network contained 1,112 nodes classified into two distinct types and 3,675 specific interactions between them, as detailed in Table 1.

Network analysis identified 110 regulatory cytokines (Fig. 2, I) targeting 58 ASD-associated proteins (Fig. 2, II) and 91 ASD-related genes (Fig. 2, III).

[1] Supplementary Table S1 is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Levanova_Engl_29_7.pdf

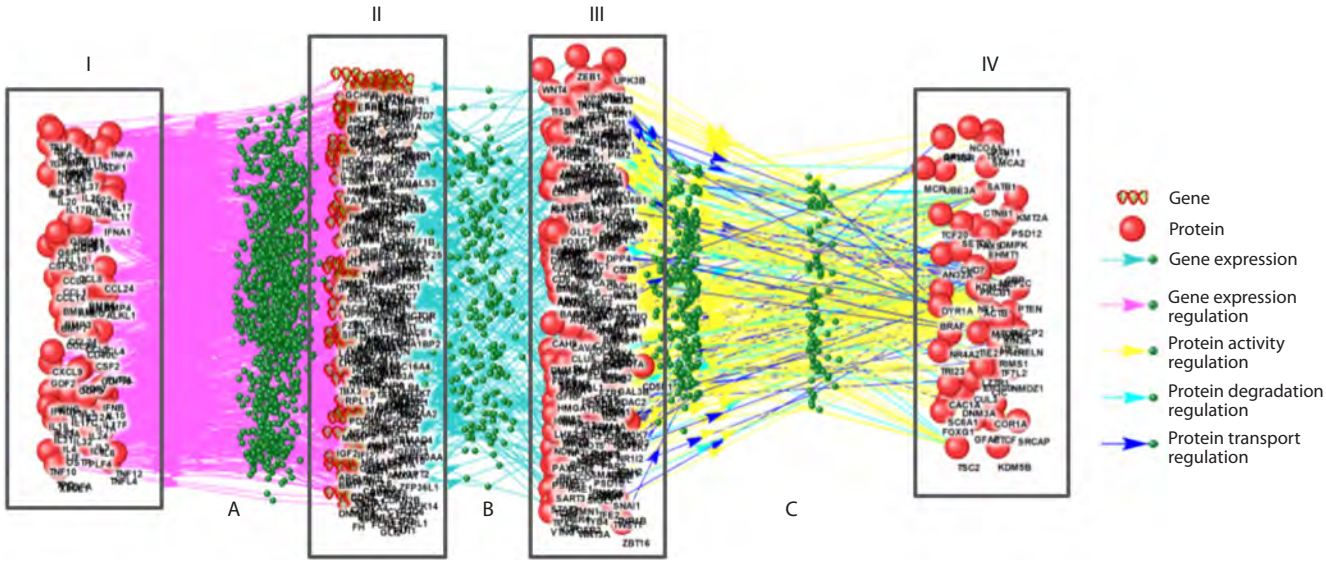


Fig. 1. Example of subnetwork reconstruction: modeling cytokine interactions with ASD-associated proteins via Pathway Wizard software using Template 1 from Supplementary Table S1.

Roman numerals indicate: I, cytokines regulating ASD-associated proteins and genes, II, mediator genes, III, mediator proteins, IV, ASD-associated proteins regulated by cytokines through signaling pathways. Letters denote: A, gene expression regulation, B, gene expression, C, regulation of protein activity, transport, and degradation.

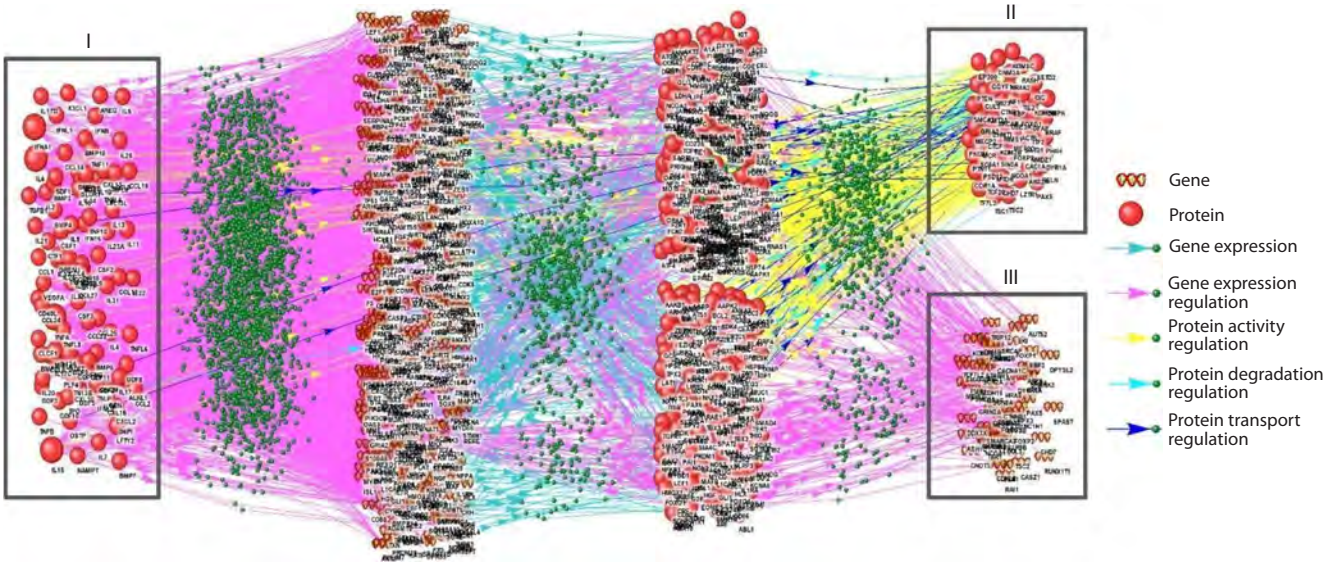


Fig. 2. Reconstructed consolidated network of cytokine interactions with ASD-related proteins and genes.

Roman numerals indicate: I, cytokines regulating ASD-associated proteins and genes, II, ASD-associated proteins regulated by cytokines, III, ASD-associated genes regulated by cytokines.

Table 1. Types and quantities of nodes and interactions in the consolidated gene network of cytokine interactions with ASD-associated proteins and genes

Interaction type	Count	Node type	Count
Activity regulation	369	Protein	621
Degradation regulation	64	Gene	491
Expression regulation	2,772		
Transport regulation	65		
Gene expression	409		

Table 2. Enrichment analysis of ASD-associated genes from the integrated network that are regulated by cytokines

No.	Biological process	Gene count	FDR
1	Excitatory postsynaptic potential	6	$9.7 \cdot 10^{-4}$
2	Regulation of dendritic spine development	4	$9.7 \cdot 10^{-4}$
3	Hippocampal development	6	$9.7 \cdot 10^{-4}$
4	Positive regulation of glutamatergic synaptic transmission	5	$8.5 \cdot 10^{-4}$
5	Neuron migration	7	$8.5 \cdot 10^{-4}$
6	Neurodevelopment	12	$7.6 \cdot 10^{-4}$
7	Negative regulation of neuronal apoptosis	8	$4.6 \cdot 10^{-4}$
8	Transmembrane calcium ion import	4	$4.6 \cdot 10^{-4}$

Note. FDR, false discovery rate.

Functional enrichment analysis of the cytokine-regulated gene set

Gene Ontology enrichment analysis was performed using the DAVID platform on the subset of ASD-associated genes identified as being under cytokine regulatory control in the reconstructed consolidated network. This analysis revealed significant enrichment (FDR < 0.05, false discovery rate) for 56 biological processes related to nervous system development and function. Specifically, these cytokine-regulated genes were overrepresented in processes including dendritic spine morphogenesis, hippocampal development, and neuronal migration (Table 2). Only the most statistically significant and biologically specific processes are presented in Table 2, while general cellular processes such as transcriptional regulation were excluded from the final selection.

Cytokine prioritization

To identify cytokines with the greatest impact on the regulatory network, we conducted multi-criteria prioritization based on three network topological and functional parameters: node degree, betweenness centrality, and enrichment in ASD-associated biological processes.

To evaluate the involvement of cytokines in ASD-associated biological processes, we developed a custom script that processes two primary inputs: cytokines identified through network reconstruction, and ASD-associated biological processes derived from Gene Ontology enrichment analysis of SFARI gene sets. The algorithm assessed each cytokine's involvement in the listed ASD-associated biological processes. This analysis identified 13 cytokines that participate in biological processes implicated in ASD (FDR < 0.05, Table 3).

To rank the cytokines by their influence within the network, two centrality metrics were employed: betweenness centrality and degree centrality. Betweenness centrality reflects the number of the shortest paths in a network that pass through a given node, while degree centrality is defined by the number of its direct connections to other nodes. These metrics quantify a node's influence on the network, as higher values indicate a more significant impact of the node. 15 most influential cytokines based on each metric are presented in Table 4.

Cytokines as potential targets for pharmacological intervention

Based on the data presented in Tables 3 and 4, a list of 21 potentially key regulators was compiled: BMP2, BMP4, BMP7, GDF2, GPI, IFN-γ, IFNL1, IL-10, IL-33, IL-15, IL-17, IL-1β, IL-2, IL-22, IL-4, IL-6, IL-8, OSTP, TGFB1, TNF-α, and VEGFA. Validation of this list against the GETdb database confirmed the status of these cytokines as promising pharmacological targets.

According to the DrugBank database records, 8 out of the 21 cytokines (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) are established targets for approved pharmaceuticals. Notably, four of these (IL-4, VEGFA, TNF-α, and IFN-γ) were also identified among the eight highest-priority candidates in our analysis, which were ranked based on a consensus across multiple prioritization metrics (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, and TNF-α).

In clinical practice, drugs targeting cytokines TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, and IL-17 are primarily used as immunosuppressants and antitumor agents. The therapeutic mechanisms of these agents principally involve either receptor blockade, utilizing cytokine antagonists, or direct cytokine neutralization through monoclonal antibodies.

Discussion

Analysis of Tables 3 and 4 identified 21 cytokines (BMP2, BMP4, BMP7, GDF2, GPI, IFN-γ, IFNL1, IL-10, IL-33, IL-15, IL-17, IL-1β, IL-2, IL-22, IL-4, IL-6, IL-8, OSTP, TGFB1, TNF-α, and VEGFA) as potential pharmacological targets, based on the GETdb database. Cross-referencing with the DrugBank database revealed that eight of them (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, and IL-17) are already targeted by approved therapeutics. A review of the existing literature confirms the critical role of specific pro-inflammatory cytokines (TNF-α, IL-6, IL-2, IL-1β, IFN-γ, VEGFA, IL-17A) in CNS development and function. These factors, secreted by classically activated microglia, are key drivers of neuroinflammation. Furthermore, dysregulation of specific cytokines, such as IL-6, IFN-γ, and IL-17A, during gestation, induced by maternal immune activation, may

Table 3. Prioritization of cytokines based on their representation in ASD-associated biological processes

No.	Cytokine	Number of ASD-associated biological processes	PadjValue	FDR
1	BMP2	6	$1.0 \cdot 10^{-6}$	$3.8 \cdot 10^{-5}$
2	IL-4	5	$5.0 \cdot 10^{-7}$	$2.7 \cdot 10^{-5}$
3	TGFB1	8	$1.9 \cdot 10^{-7}$	$2.1 \cdot 10^{-5}$
4	BMP4	6	$8.7 \cdot 10^{-6}$	$2.0 \cdot 10^{-4}$
5	IFN-γ	4	$7.2 \cdot 10^{-5}$	$1.5 \cdot 10^{-3}$
6	VEGFA	5	$8.8 \cdot 10^{-5}$	$1.6 \cdot 10^{-3}$
7	BMP7	4	$1.0 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$
8	TNF-α	5	$3.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-3}$
9	IL-33	3	$4.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-3}$
10	IL-10	3	$1.9 \cdot 10^{-3}$	$2.1 \cdot 10^{-2}$
11	IFNL1	2	$2.4 \cdot 10^{-3}$	$2.5 \cdot 10^{-2}$
12	GPI	2	$5.7 \cdot 10^{-3}$	$4.9 \cdot 10^{-2}$
13	GDF2	2	$5.7 \cdot 10^{-3}$	$4.9 \cdot 10^{-2}$

Note. FDR, false discovery rate, PadjValue, PValue with the Bonferroni correction.

Table 4. Prioritization of cytokines based on betweenness centrality and degree centrality

No.	Cytokine	Betweenness centrality	No.	Cytokine	Degree centrality
1	TNF-α	$19.3 \cdot 10^4$	1	TNF-α	100
2	IL-6	$10.4 \cdot 10^4$	2	IL-6	69
3	IL-4	$10.2 \cdot 10^4$	3	IL-4	65
4	TGFB1	$9.8 \cdot 10^4$	4	BMP4	52
5	BMP4	$9.0 \cdot 10^4$	5	TGFB1	51
6	VEGFA	$8.5 \cdot 10^4$	6	VEGFA	51
7	IL-2	$7.7 \cdot 10^4$	7	IL-2	47
8	IL-1β	$6.5 \cdot 10^4$	8	IL-1β	46
9	BMP2	$5.7 \cdot 10^4$	9	IL-10	46
10	OSTP	$5.5 \cdot 10^4$	10	BMP2	37
11	IL-10	$5.4 \cdot 10^4$	11	IFN-γ	34
12	IL-8	$4.0 \cdot 10^4$	12	IL-22	32
13	IFN-γ	$3.9 \cdot 10^4$	13	IL-17	32
14	IL-17	$3.1 \cdot 10^4$	14	OSTP	30
15	IL-15	$2.9 \cdot 10^4$	15	IL-8	29

Note. Betweenness centrality is defined as the number of the shortest paths in a network that pass through a particular node, while degree centrality represents the number of direct connections a node has with other elements in the network.

alter embryonic brain development and predispose to autism spectrum disorder (ASD) (Fujitani et al., 2022; Majerczyk et al., 2022).

Studies using maternal immune activation (MIA) mouse models demonstrate that CD4+ T-lymphocytes from affected offspring exhibit elevated IL-17A production (Morgan et al., 2010; Parkhurst et al., 2013). Furthermore, it was established that the activity of maternal RORγt-expressing pro-inflammatory T-cells (Th17), the primary source of IL-17A, is a prerequisite for the induction of ASD-like phenotypes in the offspring. It was further demonstrated that ASD-like phenotypes in the offspring require the activity of maternal RORγt-expressing Th17 cells, which are the primary source of IL-17A. Choi G.B. et al. (2016) demonstrated that both IL-17A neutralization and direct targeting of Th17 cells in pregnant mice prevent the development of MIA-induced behavioral abnormalities in their offspring. Conversely, the administration of IL-17A into the fetal brain was shown to cause disruptions

in cerebral hemisphere development and the manifestation of ASD-associated symptoms. These behavioral manifestations are linked to altered right-hemispheric activity, a region critical for adaptation mechanisms (Nikolaeva, Vergunov, 2020). This lateralized dysfunction is further supported by the significantly higher prevalence of left-handedness in children with ASD (Nikolaeva, Gaidamakina, 2018).

Paradoxically, despite the documented role of IL-17A in impairing CNS development, emerging evidence indicates its therapeutic potential for normalizing behavioral deficits in adult offspring of mothers with MIA. A study by M. Reed et al. (2020) demonstrated that lipopolysaccharide (LPS) therapy normalized behavior in adult offspring from mothers with immune activation (MIA); however, it was ineffective in monogenic models of autism spectrum disorder. This divergent outcome was attributed to variations in cytokine secretion, specifically a significantly lower production of IL-17A in response to LPS in monogenic models compared to MIA-induced counterparts.

In addition to pro-inflammatory cytokines, anti-inflammatory cytokine IL-4 is involved in ASD pathogenesis. This cytokine is critical for inducing the alternative activation pathway of microglia (M2 phenotype). Microglia in the M2 state exhibit anti-inflammatory and reparative functions, which include the secretion of numerous growth factors such as IGF-I, FGF, CSF1 and neurotrophic factors (Sica, Mantovani, 2012). Subsequently, these factors activate Trk receptors, a family of receptor tyrosine kinases involved in the regulation of synaptic plasticity.

Studies have identified a significant elevation of IL-4 levels in the amniotic fluid and maternal serum during pregnancy in women whose children were later diagnosed with ASD (Goines, Ashwood, 2013). The role of increased IL-4 concentration in ASD pathogenesis, however, remains unclear: it could either contribute to the development of pathology or represent a compensatory mechanism in response to inflammatory processes.

We hypothesize that repurposing established clinical cytokines offers a viable path for ASD therapy. To test this, we propose to initiate studies analogous to those by M. Reed et al. (2020), utilizing agents targeting the cytokines TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, and IL-17, with existing clinical applications. Planning of future research must account for the variable efficacy of cytokine interventions, which is influenced by disease etiology and developmental stage. A comprehensive approach should involve the use of rodent models that represent distinct methods of inducing ASD and its various forms, followed by a comparative analysis of the resulting data. This methodology will facilitate a more profound understanding of the effects of cytokines on the development and symptoms of ASD of diverse origins, as well as an assessment of the potential for repurposing the corresponding pharmaceutical agents for treating and alleviating ASD symptoms.

Conclusion

• Using the ANDSystem knowledge base and its components, we performed a computer-based reconstruction of five specialized gene subnetworks. These subnetworks represent distinct pathways through which cytokines influence proteins and genes associated with autism spectrum disorder (ASD), thereby providing a comprehensive mapping of cytokine interactions with ASD-associated biomolecules. Through the integration of these subnetworks into a unified model, a network for cytokine regulation of ASD-associated genes and proteins was reconstructed for the first time. The consolidated network comprises 1,112 nodes of two types (491 genes and 621 proteins) interconnected by 3,675 edges representing five distinct types of interactions.

• Analysis of the final gene network enabled the identification of 110 cytokines that regulate the activity, transport, and stability of network components implicated in ASD. Furthermore, 58 proteins and 91 genes involved in ASD pathogenesis, all of which are under cytokine regulation, were identified. Key characteristics of the network were defined, providing evidence for the significant role of cytokine-mediated regulation in ASD pathogenesis, and revealing specific cohorts of ASD-linked genes under cytokine control.

Subsequent Gene Ontology (GO) enrichment analysis for biological processes was performed on the subset of ASD-associated genes identified as being under cytokine regulatory control in the reconstructed interaction network. This analysis revealed 56 statistically significant biological processes related to neurodevelopment. Notable among these were dendritic spine morphogenesis, hippocampal development, neuronal migration, and the regulation of synaptic transmission.

• Cytokine prioritization was conducted to pinpoint key regulators, employing an analysis of network metrics (betweenness centrality and node degree) alongside an evaluation of functional relevance via linkage to ASD-associated GO biological processes. This approach yielded a set of 21 cytokines, with 8 (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α) ranking highest across all evaluated parameters.

Notably, 8 out of the 21 key cytokines (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) are targeted by existing, clinically approved drugs, highlighting an opportunity for repurposing immunomodulatory agents for ASD. The other 13 cytokines are potential targets for compounds in clinical development. Further *in vitro* and *in vivo* studies are required to delineate the precise mechanisms through which these cytokines influence neurodevelopment and to assess the therapeutic efficacy of their modulation.

References

Abrahams B.S., Arking D.E., Campbell D.B., Mefford H.C., Morrow E.M., Weiss L.A., Menashe I., Wadkins T., Banerjee-Basu S., Packer A. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*. 2013;4(1):36. doi 10.1186/2040-2392-4-36

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorder. American Psychiatric Publ., 2013. doi 10.1176/appi.books.9780890425596

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., … Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi 10.1038/75556

Ashwood P., Krakowiak P., Hertz-Picciotto I., Hansen R., Pessah I., Van de Water J. Elevated plasma cytokines in autism spectrum disorders provide evidence of immune dysfunction and are associated

with impaired behavioral outcome. *Brain Behav Immun*. 2011;25(1): 40-45. doi 10.1016/j.bbi.2010.08.003

Choi G.B., Yim Y.S., Wong H., Kim S., Kim H., Kim S.V., Hoeffer C.A., Littman D.R., Huh J.R. The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. *Science*. 2016;351(6276):933-939. doi 10.1126/science.aad0314

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Fujitani M., Miyajima H., Otani Y., Liu X. Maternal and adult interleukin-17A exposure and autism spectrum disorder. *Front Psychiatry*. 2022;13:836181. doi 10.3389/FPSYT.2022.836181/PDF

Ganesan H., Balasubramanian V., Iyer M., Venugopal A., Subramaniam M.D., Cho S.G., Vellingiri B. mTOR signalling pathway – a root cause for idiopathic autism? *BMB Rep*. 2019;52(7):424-433. doi 10.5483/BMBRep.2019.52.7.137

Goines P.E., Ashwood P. Cytokine dysregulation in autism spectrum disorders (ASD): possible role of the environment. *Neurotoxicol Teratol*. 2013;36:67-81. doi 10.1016/j.ntt.2012.07.006

Harms M.B., Martin A., Wallace G.L. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychol Rev*. 2010;20(3):290-322. doi 10.1007/S11065-010-9138-6

Hoeffer C.A., Klann E. mTOR signaling: at the crossroads of plasticity, memory and disease. *Trends Neurosci*. 2010;33(2):67-75. doi 10.1016/J.TINS.2009.11.003

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi 10.1038/nprot.2008.211

Hutsler J.J., Zhang H. Increased dendritic spine densities on cortical projection neurons in autism spectrum disorders. *Brain Res*. 2010; 1309:83-94. doi 10.1016/j.brainres.2009.09.120

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int J Mol Sci*. 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9(Suppl.2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. doi 10.1186/S12859-018-2567-6

Knox C., Wilson M., Klinger C.M., Franklin M., Oler E., Wilson A., Pon A., … Ackerman D., Jewison T., Sajed T., Gautam V., Wishart D.S. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. doi 10.1093/nar/gkad976

Li X., Chauhan A., Sheikh A.M., Patil S., Chauhan V., Li X.M., Ji L., Brown T., Malik M. Elevated immune response in the brain of autistic patients. *J Neuroimmunol*. 2009;207(1-2):111-116. doi 10.1016/j.jneuroim.2008.12.002

Liu Y., Zhang D.T., Liu X.G. mTOR signaling in T cell immunity and autoimmunity. *Int Rev Immunol*. 2015;34(1):50-66. doi 10.3109/08830185.2014.933957

Majerczyk D., Ayad E.G., Brewton K.L., Saing P., Hart P.C. Systemic maternal inflammation promotes ASD via IL-6 and IFN-γ. *Biosci Rep*. 2022;42(11):BSR20220713. doi 10.1042/BSR20220713

McMahon J., Huang X., Yang J., Komatsu M., Yue Z., Qian J., Zhu X., Huang Y. Impaired autophagy in neurons after disinhibition of mammalian target of rapamycin and its contribution to epileptogenesis. *J Neurosci*. 2012;32(45):15704-15714. doi 10.1523/JNEUROSCI.2392-12.2012

Morgan J.T., Chana G., Pardo C.A., Achim C., Semendeferi K., Buckwalter J., Courchesne E., Everall I.P. Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biol Psychiatry*. 2010;68(4):368-376. doi 10.1016/j.biopsych.2010.05.024

Navlakha S., Barth A.L., Bar-Joseph Z. Decreasing-rate pruning optimizes the construction of efficient and robust distributed networks. *PLoS Comput Biol*. 2015;11(7):e1004347. doi 10.1371/JOURNAL.PCBI.1004347

Nikolaeva E.I., Gaidamakina M.D. Specificity of lateral preference of mute preschoolers with autism spectrum disorders. *J Asymmetry*. 2018;12(4):367-371 (in Russian)

Nikolaeva E.I., Vergunov E.G. Functional Asymmetry of the Brain and Lateral Preferences: Reloaded. Evolutionary, Genetic, Psychological and Psychophysiological Approaches to Analysis. St. Petersburg, 2020 (in Russian)

Onore C., Yang H., Van de Water J., Ashwood P. Dynamic Akt/mTOR signaling in children with autism spectrum disorder. *Front Pediatr*. 2017;5:43. doi 10.3389/fped.2017.00043

Parkhurst C.N., Yang G., Ninan I., Savas J.N., Yates J.R., Lafaille J.J., Hempstead B.L., Littman D.R., Gan W.B. Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. *Cell*. 2013;155(7):1596-1609. doi 10.1016/j.cell.2013.11.030

Ragupathi A., Kim C., Jacinto E. The mTORC2 signaling network: targets and cross-talks. *Biochem J*. 2024;481(2):45-91. doi 10.1042/BCJ20220325

Reed M.D., Yim Y.S., Wimmer R.D., Kim H., Ryu C., Welch G.M., Andina M., King H.O., Waisman A., Halassa M.M., Huh J.R., Choi G.B. IL-17a promotes sociability in mouse models of neurodevelopmental disorders. *Nature*. 2020;577(7789):249-253. doi 10.1038/S41586-019-1843-6

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Sica A., Mantovani A. Macrophage plasticity and polarization: in vivo veritas. *J Clin Invest*. 2012;122(3):787-795. doi 10.1172/JCI59643

Smith J.A., Das A., Ray S.K., Banik N.L. Role of pro-inflammatory cytokines released from microglia in neurodegenerative diseases. *Brain Res Bull*. 2012;87(1):10-20. doi 10.1016/j.brainresbull.2011.10.004

Tang G., Gudsnuk K., Kuo S.H., Cotrina M.L., Rosoklija G., Sosunov A., Sonders M.S., … Peterson B.S., Champagne F., Dwork A.J., Goldman J., Sulzer D. Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron*. 2014;83(5): 1131-1143. doi 10.1016/j.neuron.2014.07.040

Trifonova E.A., Klimenko A.I., Mustafin Z.S., Lashin S.A., Kochetov A.V. The mTOR signaling pathway activity and vitamin D availability control the expression of most autism predisposition genes. *Int J Mol Sci*. 2019;20(24):6332. doi 10.3390/ijms20246332

Tsai A.C., Savostyanov A.N., Wu A., Evans J.P., Chien V.S.C., Yang H.H., Yang D.Y., Liou M. Recognizing syntactic errors in Chinese and English sentences: brain electrical activity in Asperger's syndrome. *Res Autism Spectr Disord*. 2013;7(7):889-905. doi 10.1016/j.rasd.2013.02.001

Tseng Y.L., Yang H.H., Savostyanov A.N., Chien V.S.C., Liou M. Voluntary attention in Asperger's syndrome: brain electrical oscillation and phase-synchronization during facial emotion recognition. *Res Autism Spectr Disord*. 2015;13-14:32-51. doi 10.1016/j.rasd.2015.01.003

Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., … Johansson F., Zwahlen M., Von Heijne G., Nielsen J., Pontén F. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. doi 10.1126/science.1260419

Van der Zee E., Derksen J.J.L. The power of systemizing in autism. *Child Psychiatry Hum Dev*. 2021;52(2):321-331. doi 10.1007/S10578-020-01014-4

Vargas D.L., Nascimbene C., Krishnan C., Zimmerman A.W., Pardo C.A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann Neurol*. 2005;57(1):67-81. doi 10.1002/ana.20315

Zhang Q., He Y., Lu Y.P., Wei Q.H., Zhang H.Y., Quan Y. GETdb: a comprehensive database for genetic and evolutionary features of drug targets. *Comput Struct Biotechnol J*. 2024;23:1429-1438. doi 10.1016/j.csbj.2024.04.006

Ziats C.A., Patterson W.G., Friez M. Syndromic autism revisited: review of the literature and lessons learned. *Pediatr Neurol*. 2021;114:21-25. doi 10.1016/J.PEDIATRNEUROL.2020.06.011

Conflict of interest. The authors declare no conflict of interest.

Received August 11, 2025. Revised September 12, 2025. Accepted September 15, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(6):1009-1019

doi 10.18699/vjgb-25-106

The gene network and knowledge base on human thermoregulation

E.V. Ignatieva[1], P.S. Demenkov[1], A.G. Bogomolov[1], R.A. Ivanov[1], S.A. Lashin[1, 2], A.D. Mikhailova[2], A.E. Alekseeva[2], N.S. Yudin[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
eignat@bionet.nsc.ru

Abstract. Reconstruction and analysis of gene networks regulating biological processes are among the modern methodological approaches for studying complex biological systems that ensure the vital activity of organisms. Thermoregulation is an important evolutionary acquisition of warm-blooded animals. Multiple physiological systems (nervous, cardiovascular, endocrine, respiratory, muscular, etc.) are involved in this process, maintaining stable body temperature despite changes in ambient temperature. This study aims to perform a computer reconstruction of the human thermoregulation gene network and present the results in the Termo_Reg_Human 1.0 knowledge base. The gene network was reconstructed using the ANDSystem software and information system, designed for the automated extraction of knowledge and facts from scientific publications and biomedical databases based on machine learning and artificial intelligence methods. The Termo_Reg_Human 1.0 knowledge base (https://www.sysbio.ru/ThermoReg_Human/) contains information about the human thermoregulation gene network, including a description of 469 genes, 473 proteins, and 265 microRNAs important for its functioning, interactions between these objects, and the evolutionary characteristics of the genes. Using the ANDVisio software tool (a module of ANDSystem), each gene, protein, and microRNA involved in the thermoregulation of the human body was prioritized according to its functional significance, i. e., the number of interactions with other objects in the reconstructed gene network. It was found that the key objects with the largest number of functional interactions in the human thermoregulation gene network included the *UCP1*, *VEGFA*, *PPARG* and *DDIT3* genes; STAT3, JUN, VEGFA, TLR4 and TNFA proteins; and the microRNAs hsa-mir-335 and hsa-mir-26b. We revealed that the set of 469 human genes from the network was enriched with genes whose ancestral forms originated at an early evolutionary stage (Unicellular organisms, the root of the phylostratigraphic tree) and at the stage of Vertebrata divergence.

Key words: heat; cold; gene network; database; microRNA; evolution; phylostratigraphy; gene age

For citation: Ignatieva E.V., Demenkov P.S., Bogomolov A.G., Ivanov R.A., Lashin S.A., Mikhailova A.D., Alekseeva A.E., Yudin N.S. The gene network and knowledge base on human thermoregulation. *Vavilovskii Zhurnal Genetiki i Selektsii* = *Vavilov J Genet Breed*. 2025;29(7):1009-1019. doi 10.18699/vjgb-25-106

Funding. The work was supported by the publicly funded project No. FWNR-2022-0020 of the Federal Research Center ICG SB RAS.

Acknowledgements. The authors would like to thank V.A. Ivanisenko for consulting us on the usage of ANDSystem.

Генная сеть и база знаний по терморегуляции организма человека

Е.В. Игнатьева[1], П.С. Деменков[1], А.Г. Богомолов[1], Р.А. Иванов[1], С.А. Лашин[1, 2], А.Д. Михайлова[2], А.Е. Алексеева[2], Н.С. Юдин[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
eignat@bionet.nsc.ru

Аннотация. Реконструкция и анализ генных сетей, регулирующих биологические процессы, – один из эффективных подходов к исследованию сложных систем обеспечения жизнедеятельности организмов. Терморегуляция – важное эволюционное приобретение человека и других теплокровных животных. Терморегуляция осуществляется при участии многих физиологических систем организма (нервной, сердечно-сосудистой, эндокринной, дыхательной, мышечной и т. д.), что способствует поддержанию относительно постоянной температуры тела в условиях колебания температуры окружающей среды. Цель работы – компьютерная реконструкция генной сети терморегуляции человека и представление полученных результатов в соответствующей базе знаний Termo_Reg_Human 1.0. Генная сеть реконструирована с использованием программно-информационной системы ANDSystem, предназначенной для

автоматизированного извлечения знаний и фактов из текстов научных публикаций и баз данных биомедицинской направленности, основанной на методах машинного обучения и искусственного интеллекта. База знаний Termo_Reg_Human 1.0 (https://www.sysbio.ru/ThermoReg_Human/) содержит информацию о генной сети терморегуляции человека, включая описание 469 генов, 473 белков и 265 микроРНК, значимых для ее функционирования; взаимодействиях между этими объектами, а также эволюционные характеристики генов. С использованием программного инструмента ANDVisio (модуля системы ANDSystem) проведена приоритизация каждого гена, белка и микроРНК, участвующих в терморегуляции организма человека по их функциональной нагруженности – количеству связей с другими объектами реконструированной генной сети. Установлено, что к числу ключевых объектов, имеющих наибольшее количество функциональных связей в генной сети терморегуляции человека, относятся гены *UCP1*, *VEGFA*, *PPARG*, *DDIT3*, белки STAT3, JUN, VEGFA, TLR4, TNFA и микроРНК hsa-mir-335 и hsa-mir-26b. Обнаружено обогащение генной сети терморегуляции генами, предковые варианты которых сформировались на эволюционных этапах появления одноклеточных организмов и дивергенции позвоночных.

Ключевые слова: тепло; холод; генная сеть; база данных; микроРНК; эволюция; филостратиграфия; возраст гена

Introduction

Humans and most other mammals are homoiothermic, capable of maintaining a relatively constant body temperature when the ambient temperature varies (Osvath et al., 2024). Human thermoregulation is carried out with the participation of: 1) thermoreceptors located on the body's surface and in the internal organs; 2) afferent neural signal transmission pathways; 3) thermoregulatory centers in the hypothalamus and other parts of the brain; 4) efferent neural pathways that control adaptive reactions (Nakamura, 2024). Such adaptive reactions include: a) shivering and nonshivering thermogenesis (chemical mechanisms of thermoregulation) (Ikeda, Yamada, 2020; Dumont et al., 2025); b) physical thermoregulation, including the regulation of heat transfer through evaporation and convection, as well as thermal insulation (Nakamura, 2011; Tattersall et al., 2012); c) behavioral reactions: avoidance of open areas of the Earth's surface characterized by extreme temperatures; crowding of individuals, etc. (Tattersall et al., 2012; Tansey, Johnson, 2015; McCafferty et al., 2017).

Chemical thermoregulation is carried out through heat production during skeletal muscle contractions (Blondin et al., 2019; Dumont et al., 2025), and nonshivering thermogenesis in brown adipose tissue (Tansey, Johnson, 2015; Ikeda, Yamada, 2020) and muscles (Blondin et al., 2019). Physical thermoregulation is carried out by changing the heat transfer from the body: conduction, radiation, perspiration, evaporation of water from the respiratory passages, thermal insulation due to the subcutaneous fat layer, piloerection (Nakamura, 2011; Tattersall et al., 2012). Both chemical and physical thermoregulatory processes are actively controlled by the neuroendocrine system (Charkoudian et al., 2017; Nakamura, 2024; Mittag, Kolms, 2025).

In addition, the thermoregulatory reactions are associated with changes in the cardiovascular system (Tansey, Johnson, 2015). Thus, thermoregulation is provided by a variety of biological processes occurring in the nervous, endocrine, cardiovascular, respiratory, muscular and other body systems. The genetic regulatory mechanisms controlling the above processes also play a significant role in thermoregulation (Festuccia et al., 2009; Rehman et al., 2013; Li et al., 2015; Horii et al., 2019; Xiao et al., 2019; Kudsi et al., 2022; Song et al., 2022; Valdivia et al., 2023).

Reconstruction and analysis of gene networks regulating biological processes are among the effective approaches to study complex biological systems that ensure vital activity of organisms (Ignatieva et al., 2017; Saik et al., 2018; Mustafin et al., 2019, 2021; Mikhailova et al., 2024). A large amount of experimental genetic data has been accumulated on the problem of thermoregulation, presented in tens of thousands of scientific publications and many specialized databases (e. g. KEGG Pathway, WikiPathways, MetaCyc, REACTOME, etc.). In this regard, in our work, we reconstructed the human thermoregulation gene network using the ANDSystem software and information system, designed for the automated extraction of knowledge and facts from the texts of scientific publications and biomedical databases using machine learning and artificial intelligence methods (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2024). The results obtained from the analysis of 30 million publications are accumulated in the specialized knowledge base of the ANDSystem in the form of a global knowledge graph (Ivanisenko T.V., 2024).

Information on the reconstructed human thermoregulatory gene network is presented in the Termo_Reg_Human 1.0. knowledge base (https://www.sysbio.ru/ThermoReg_Human/), including descriptions of 469 genes, 473 proteins and 265 microRNAs important for gene network functioning, as well as interactions between them.

Each gene, protein, and microRNA involved in human body thermoregulation was prioritized according to their functional load, i. e., the number of interactions with other objects of the reconstructed gene network, using the ANDVisio software tool (a module of the ANDSystem). The key objects with the largest number of functional interactions in the human thermoregulation gene network were found: the *UCP1*, *VEGFA*, *PPARG* and *DDIT3* genes, the STAT3, JUN, VEGFA, TLR4 and TNFA proteins, and microRNAs hsa-mir-335 and hsa-mir-26b.

The Termo_Reg_Human 1.0 knowledge base also presents the results of an evolutionary analysis of genes functioning in the thermoregulation gene network: this gene network was enriched with genes, the ancestral forms of which emerged at two important evolutionary stages corresponding to a) the appearance of unicellular organisms and b) the divergence of vertebrates.

Materials and methods

Lists of genes used for building a gene network. The list of human genes involved in thermoregulation was compiled based on the Gene Ontology, EntrezGene, and ANDSystem

databases (Ivanisenko V.A. et al., 2019) using the keywords shown in Supplementary Material S1[1].

Building of the gene network. The gene network of thermoregulation was built using the ANDSystem software and information system (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2024). ANDSystem, based on machine learning and artificial intelligence methods, is designed for the automated extraction of knowledge and facts about the structural and functional organization of gene networks from scientific publications and biomedical factographical databases. The information obtained in this way is accumulated in the specialized knowledge base of ANDSystem in the form of a global knowledge graph (Ivanisenko T.V. et al., 2024). Based on this information, a reconstruction of the graphs of target gene networks is carried out, the nodes of which correspond to molecular genetic objects (genes, RNA, proteins and metabolites), functioning as part of gene networks, and the edges connecting these nodes indicate the functional interactions between objects. Supplementary Material S2 provides a detailed description of the reconstruction process of the human thermoregulatory gene network.

Prioritization of genes, proteins, and microRNAs according to their functional significance in the human thermoregulation gene network. Prioritization of gene network nodes (genes, microRNAs and proteins) was performed using the ANDVisio software tool (a module of the ANDSystem). The number of interactions with other objects was calculated for a specific object in the human thermoregulation gene network graph. Next, the probability of obtaining the observed number of interactions for random reasons was estimated for each gene network object. Next, the probability of observing this number of interactions involving this specific object of the gene network by chance was estimated. The probability was calculated using a hypergeometric test:

$$p\text{-value} = \sum_{i=0}^{k} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}},$$

where: k – the number of interactions of this specific object (node) in the gene network; n – the number of objects (nodes) involved in the gene network under consideration; K – the number of interactions of this specific object (node) in the ANDSystem knowledge base global network graph; N – the total number of objects (nodes) in the ANDSystem knowledge base global graph (Ivanisenko V.A. et al., 2019).

When calculating the p-value, only objects of the same type (genes, proteins, microRNA) as the considered object of the human thermoregulation gene network were taken into account. Next, correction for multiple hypothesis testing was applied (Benjamini, Yekutieli, 2001), resulting in a P-adjusted value.

Analysis of the evolutionary characteristics of the genes. The analysis of the evolutionary characteristics of genes involved in the reconstructed gene network was carried out using the Orthoweb system (Ivanov et al., 2024), which calculates the phylostratigraphic index (PAI) of each gene, characterizing the evolutionary age of the gene. Details of the calculation procedure for the PAI index are described in Supplementary Material S2.

Functional annotation of genes. The identification of Gene Ontology terms associated with genes of a certain phylostratigraphic age was carried out using the DAVID web server and its GOTERM_BP_DIRECT dictionary (Sherman et al., 2022).

Implementation of the knowledge base on human thermoregulation. Data for the knowledge base information tables were extracted from text outputs of the ANDVisio program (a module of the ANDSystem) using original Python scripts. The online implementation of the knowledge base was performed using MySQL 5.1.73 and PHP 5.3.3. Apache HTTP Server 2.2.15 and Nginx 1.4.1 were used.

Results and discussion

Genes associated with thermoregulatory processes

The search through the Gene Ontology, EntrezGene, and ANDCell (the information component of ANDSystem) databases identified 467 protein-coding genes associated with thermoregulation, as well as two genes encoding microRNAs.

The gene network of human thermoregulation

Based on the list of human genes involved in thermoregulation mentioned above, the gene network of human thermoregulation was reconstructed using ANDSystem. The view of the entire reconstructed gene network is shown in Figure 1. The gene network includes 469 genes, 473 proteins, 265 microRNAs and 7,018 interactions between them. The number of proteins exceeds the number of genes because the gene network contains six genes that encode more than one protein due to alternative splicing or proteolytic cleavage of the precursor protein.

It should be noted that ANDSystem identifies two types of relationships between gene networks objects, based on the analysis of scientific literature and biomedical databases: direct molecular genetic interactions between gene network objects and indirect actions, i. e. relationships in which the effect of one gene network object on another is shown, but the molecular genetic mechanism of such effect remains unknown and/or may involve intermediate objects.

Figure 2 shows two fragments of the thermoregulatory gene network. Figure 2*a* illustrates molecular genetic interactions of the gene encoding the thermoreceptor TRPV1, which is activated when temperature increases. According to the ANDSystem knowledge base, *TRPV1* expression is regulated by interleukin 13 (IL13) and toll-like receptor 4 (TLR4). These regulatory relations are described in the articles (Rehman et al., 2015; Li et al., 2015) and can be categorized as “indirect”, since we are talking about the action of the cytokine IL13 (an extracellular signaling molecule) and the TLR4 receptor located on the cell membrane, which affect *TRPV1* expression through signal transduction pathways. In addition, *TRPV1* is coexpressed with other genes from the thermoregulation gene network, including thermoreceptor-encoding genes (*TRPM8*, *TRPA1*, *TRPV3*, *TRPV4*), as well as *NTRK1* encoding neurotrophic receptor tyrosine kinase 1. The experiments that

[1] Supplementary Materials S1–S7 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Ignatieva_Engl_29_7.pdf

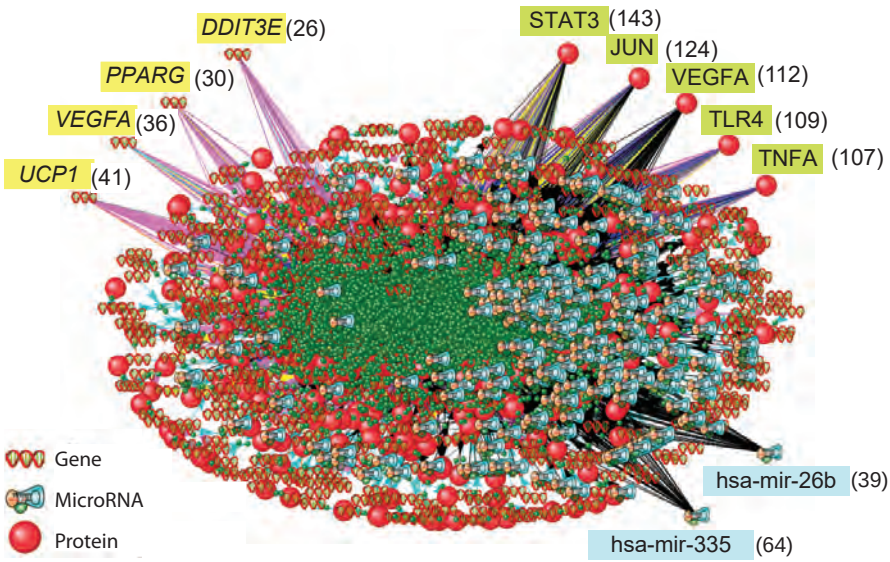


Fig. 1. The view of the entire gene network of human thermoregulation reconstructed using the ANDSystem tool.
The gene network includes 469 genes, 473 proteins, 265 microRNAs, and 7,018 interactions between these objects. Genes, proteins, and microRNAs with the highest number of interactions in the network are shown separately. Numbers in parentheses indicate the number of interactions in the network.

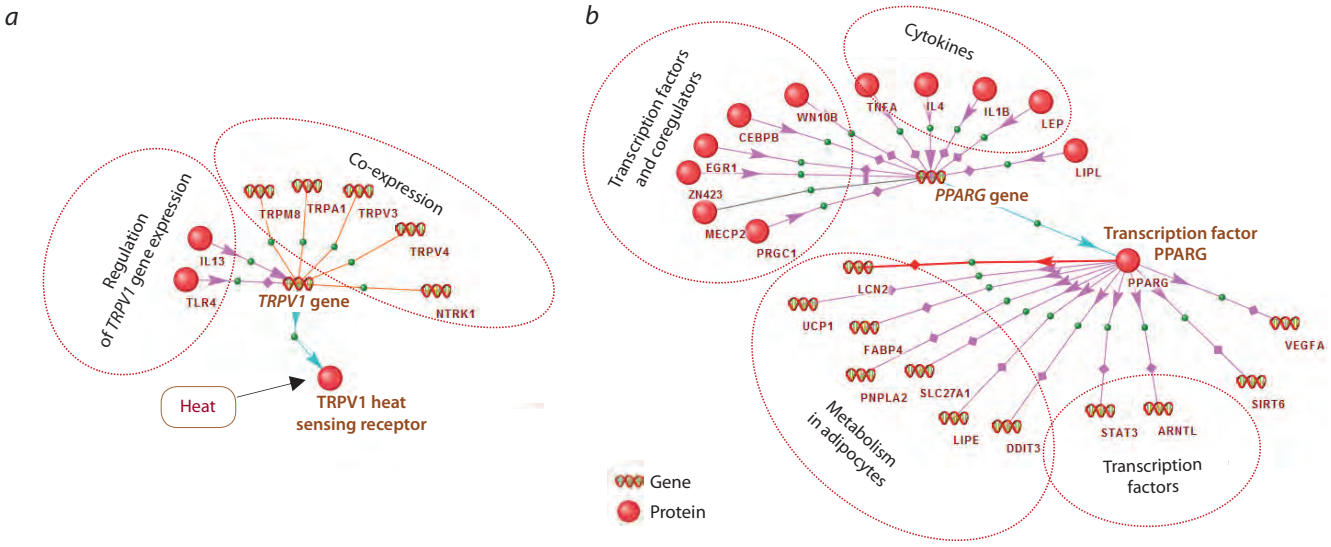


Fig. 2. The fragments of the thermoregulation gene network shown in Figure 1.
a – regulatory interactions involving the gene encoding the TRPV1 heat sensing receptor; *b* – regulatory interactions involving the PPARG gene and the encoded transcription factor PPARG.

revealed the coexpression of these genes are described in the research papers (Zhu, Oxford, 2007; Cao et al., 2009; Cheng et al., 2011; Gouin et al., 2012; Nguyen et al., 2017).

Figure 2*b* shows the regulatory relationships involving the *PPARG* gene and its encoded protein. *PPARG* expression is regulated by transcription factors ZN423, EGR1, CEBPB, which affect the level of transcription by interacting with DNA in the *PPARG* regulatory regions. *PPARG* expression is also regulated by transcription cofactors MECP2 and PRGC1/PGC-1-alpha and the WN10B protein, which activates the Wnt signaling cascade. In addition, cytokines TNF, IL4, IL1B, and LEP are involved in the regulation of *PPARG* expression. The transcription factor PPARG, encoded by the gene under consideration, controls the transcription of a) genes regulating metabolic processes in adipocytes: *LCN2*, *UCP1*, *FABP4*, *PNPLA2*, *SLC27A1*, *LIPE*, and *DDIT3;* b) genes encoding transcription factors *STAT3* and *ARNTL*; and c) the *SIRT6* gene encoding the NAD-dependent protein deacetylase. The references to scientific publications supporting these interactions are provided in Supplementary Material S3.

The Termo_Reg_Human knowledge base

At the next stage of the study, the Termo_Reg_Human 1.0. knowledge base (https://www.sysbio.ru/ThermoReg_Human/)

was developed. This knowledge base contains data on 469 genes, 473 proteins, and 265 microRNAs involved in human thermoregulation.

Termo_Reg_Human 1.0. contains four main tables: *Genes_evol*, *Proteins*, *MicroRNA* и *Genes_all* (the knowledge base scheme is shown in Figure 3).

The *Genes_evol* table contains a description of each of the 469 genes functioning as part of the human thermoregulation gene network, including: the EntrezGene GeneID, the number of interactions of the gene with other genes and proteins of the gene network, and the evidence type supporting the association of the gene with thermoregulation (Gene Ontology, ANDSystem, Entrez Gene). This table also presents such evolutionary characteristics for each protein-coding gene as the phylostratigraphic age index (PAI) and the divergence index (DI), calculated using the OrthoWeb software package (Ivanov et al., 2024).

The *Proteins* table contains data on proteins encoded by genes from the *Genes_evol* table. The description of each protein includes the UniProtKb Entry Name, the NCBI GeneID of the gene encoding the protein, the number of interactions the protein has in the gene network, and the names of the microRNAs that regulate protein expression.

The *MicroRNA* table contains information about microRNAs that regulate the expression of proteins involved in the network. These are two microRNAs encoded by genes from the list of 469 genes mentioned above, as well as additional microRNAs found using the ANDVisio program during the reconstruction of the network. The *MicroRNA* table shows for each microRNA: 1) microRNA name within the network; 2) official symbol of the gene encoding this microRNA; 3) the number of interactions involving this microRNA; 4) the names of proteins for which this microRNA acts as an expression regulator.

The fourth table, *Genes_all*, contains additional data on all 469 genes characterized in the *Genes_evol* table, as well as data on the genes encoding microRNAs included in the network using the ANDVisio program.

The web interface allows to view data on genes and proteins associated with thermoregulation, as well as to search for genes/proteins by identifiers or their names. In addition, a search for objects (genes, proteins, microRNAs) by the number of functional interactions in the network is available. The interface displays objects with a number of interactions exceeding the value specified by the user.

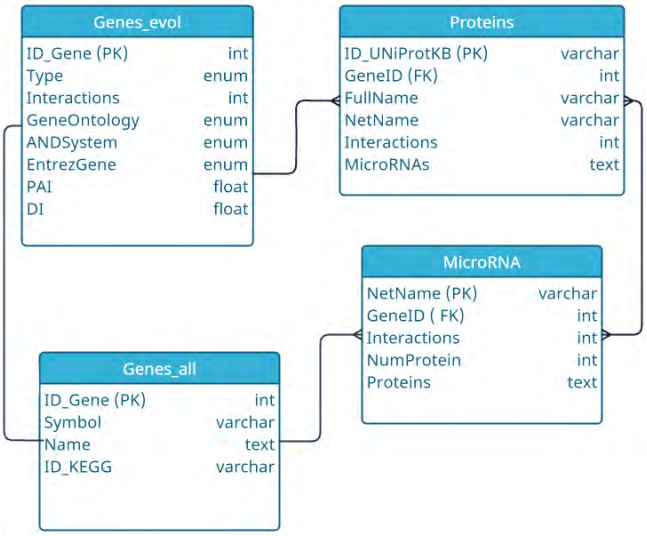


Fig. 3. Structure of the Termo_Reg_Human 1.0. knowledge base.

Using data from the Termo_Reg_Human 1.0 knowledge base in bioinformatics research

Prioritization of genes by the number of interactions in the gene network. Figure 4*a* shows the distribution of genes by the number of interactions with other objects of the human thermoregulation gene network (genes, proteins, and microRNAs). Most genes (373 out of 467) have a low number of interactions with other objects in the network (five or less). One fifth of all genes, that is, 90 genes, have from 6 to 25 interactions. Only four genes had more than 25 interactions: *UCP1* (41 interactions), *VEGFA* (36), *PPARG* (30), and *DDIT3* (26). A statistical analysis using the hypergeometric distribution confirmed that these four genes have significantly more interactions than would be expected by chance: the P-adjusted value varies from $2.44 \cdot 10^{-05}$ for the *DDIT3* gene to $1.20 \cdot 10^{-28}$ for the *UCP1* gene. Functional characteristics of these genes with the largest number of interactions are shown in Table 1.

The *UCP1* gene encodes the uncoupling protein 1 (called thermogenin), which is involved in one of the key processes of heat generation – nonshivering thermogenesis in brown adipose tissue (Wollenberg Valero et al., 2014). This protein, localized in the mitochondrial inner membrane, increases its permeability, dissipating the proton gradient generated in oxidative phosphorylation. As a result, the processes of oxidative phosphorylation and ATP synthesis are uncoupled, and heat is released (Ikeda, Yamada, 2020).

The *VEGFA* gene encodes vascular endothelial growth factor A (Naik et al., 2012). The resulting activation of the blood supply to tissues is important for thermoregulatory processes: heat exchange between the internal parts of the body and its surface, heat dissipation through evaporation and convection, etc. (Tansey, Johnson, 2015).

The *PPARG* gene encodes the transcription factor PPARG, which belongs to the nuclear receptor superfamily. PPARG controls the activity of genes governing the metabolism of fatty acids and glucose (Festuccia et al., 2009), and also activates the production of the UCP1 (uncoupling protein 1, thermogenin) in brown and beige adipocytes (Valdivia et al., 2023).

The *DDIT3* gene encodes CHOP (C/EBP homologous protein), a transcription factor from the C/EBP family regulating differentiation of adipocyte precursor cells into mature adipocytes, which play a crucial role in nonshivering thermogenesis (Okla et al., 2015).

Prioritization of proteins by the number of interactions in the gene network of thermoregulation. Analysis of the thermoregulation gene network revealed that proteins generally have more interactions than genes (Fig. 4*b*): the proportion of proteins that had no more than five interactions was less than half of their total number (144 out of 473). 55 % of the proteins (261 proteins) had from 6 to 30 interactions, 13 % of the proteins (63 proteins) had from 31 to 100 interactions. Five proteins (STAT3, JUN, VEGFA, TLR4, TNFA) had more

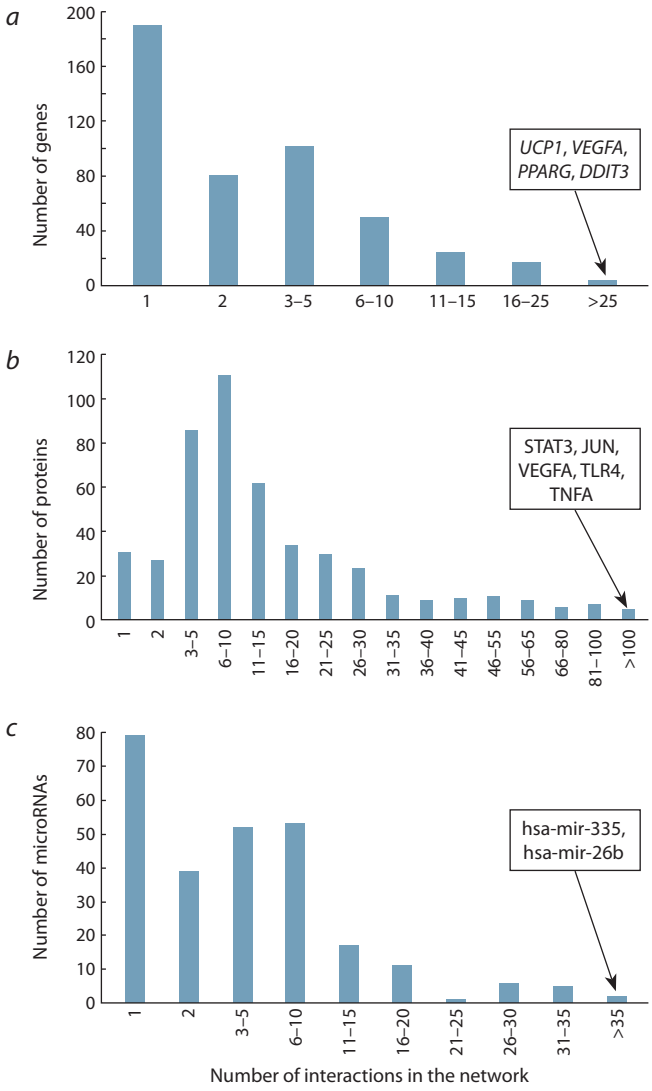


Fig. 4. Distribution of genes, proteins, and microRNAs involved in the thermoregulatory gene network according to the number of interactions in this network (based on information from the Termo_Reg_Human 1.0 knowledge base).

a – distribution of genes according to the number of interactions; *b* – distribution of proteins according to the number of interactions; *c* – distribution of microRNAs according to the number of interactions. The rectangular panels show the names of the genes, proteins, and microRNAs with the highest number of interactions.

than 100 interactions with other network objects. A statistical analysis using the hypergeometric distribution confirmed that these five proteins have a significantly greater number of interactions with the rest of the network objects than would be expected by chance: P-adjusted value ranged from $2.04 \cdot 10^{-18}$ for the TLR4 protein to $3.79 \cdot 10^{-43}$ for the STAT3 protein. The characteristics of these five proteins are given in Table 2.

STAT3 (143 interactions in the network) is a transcription factor acting at the final step of the JAK/STAT3 signal transduction pathway. STAT3 regulates adipocyte differentiation during the induction phase, and subsequent inactivation of the JAK/STAT3 pathway in these cells provides *UCP1* gene expression activation and the conversion of preadipocytes into mature brown fat cells (Song et al., 2022). In addition, STAT3 is involved in the signaling pathway activated by the heat sensing receptor TRPV1 in brain regions that control body temperature (Yoshida et al., 2016).

The JUN protein (124 interactions in the network) is a subunit of the transcription factor AP1 (the JUN/FOS heterodimer). JUN is involved in the regulation of cytokine expression, thereby controlling the inflammatory processes that are associated with elevated body temperature (Schonthaler et al., 2011; Johnson Rowsey, 2013). It has been shown that when the expression of the *JUN* gene in the liver is inactivated in liver-specific *c-Jun* knock-out mice, an increase in body temperature occurs due to the activation of the sympathetic nervous system and subsequent stimulation of UCP1 expression in brown fat (Xiao et al., 2019).

As mentioned above, the VEGFA protein, which has 112 interactions in the network, controls vascular endothelium growth (Naik et al., 2012), which is important for heat exchange between tissues and the external environment (Tansey, Johnson, 2015).

TLR4 (109 interactions in the network) is a transmembrane protein, toll-like receptor 4. It can be activated by lipopolysaccharides (LPS) found in bacterial cell walls, leading to an increase in body temperature in response to infection (Roth, Blatteis, 2014). Additionally, activation of the TLR4 receptor by lipopolysaccharides leads to oxidative stress, mitochondrial dysfunction, and inhibition of the brown adipocyte differentiation (Okla et al., 2018).

The TNFA protein, tumor necrosis factor, belongs to the cytokine family (107 interactions in the network). It activates, in particular, prostaglandin synthesis in endothelial cells. These prostaglandins act on neurons in the preoptic area of the hypothalamus, the brain's thermoregulatory center, leading to increased body temperature (Leon et al., 1998; Netea et al., 2000; Gil et al., 2007; Nakamura, 2024). TNFA has also been shown to have a direct effect on adipocytes *in vitro*, reducing the expression of thermogenin (UCP-1) (Valladares et al., 2001) and the enzyme triglyceride lipase ATGL/PNPLA2 (Kim et al., 2006). Thus, the cytokine TNFA plays an important role in thermoregulation, but its effect on body temperature depends on the type of cells affected by this cytokine.

Prioritization of microRNAs by the number of interactions in the gene network of thermoregulation. MicroRNAs regulate gene expression at the translational level. These RNAs bind to the mRNA targets within miRISC complex, inhibiting protein synthesis with or without transcript degradation (O'Brien et al., 2018). According to the Termo_Reg_Human 1.0 knowledge base, the thermoregulation gene network includes 265 microRNAs that are involved in regulating the expression of 297 genes. Data on these regulatory relationships was obtained from the miRTarBase, which contains experimentally confirmed information about interactions between microRNAs and their mRNA targets (Cui et al., 2025). The proportion of microRNAs having not more than five regulatory interactions in the network was 64 % (170 out of 265) (Fig. 4*c*). 35 % of the total set of microRNAs (93 out of 265) had from 6 to 30 interactions. Two microRNAs had the highest number of interactions (more than 35). These are hsa-mir-335 (64 interactions) and hsa-mir-26b (39 interactions). An assessment of the statistical significance of the number of interactions between these microRNAs and other

Table 1. Functional characteristics of genes with the highest number of interactions in the thermoregulatory network

Gene symbol	Number of interactions in the network	Role in thermoregulation	P-adjusted	PAI
UCP1	41	Encodes uncoupling protein 1, which is expressed in brown adipose tissue and enables heat generation through nonshivering thermogenesis (Wollenberg Valero et al., 2014; Ikeda, Yamada, 2020)	$1.2 \cdot 10^{-28}$	1
VEGFA	36	Encodes vascular endothelial growth factor A, which regulates tissue vascularization, facilitating heat exchange and heat transfer (Naik et al., 2012)	$1.8 \cdot 10^{-6}$	6
PPARG	30	Encodes a nuclear receptor that regulates adipocyte differentiation, fatty acid metabolism, and glucose uptake in fat cells (Festuccia et al., 2009)	$2.66 \cdot 10^{-7}$	6
DDIT3	26	Encodes the transcription factor CHOP, which plays a key role in adipogenesis (Okla et al., 2015)	$2.44 \cdot 10^{-6}$	7

Note. Genes are listed in descending order based on the number of interactions in the gene network.
Here and in Tables 2 and 3: P-adjusted indicates the probability of observing a given number of interactions in a network by chance, calculated using hypergeometric distribution with correction for multiple comparisons.

Table 2. Functional characteristics of proteins with the highest number of interactions in the network of thermoregulation

Protein	Number of interactions in the network	Role in thermoregulation	P-adjusted
STAT3	143	The transcription factor STAT3 regulates gene expression in brain regions that control thermoregulation (Yoshida et al., 2016), regulates the differentiation of adipocytes into brown fat cells, as well as *UCP1* gene expression (Song et al., 2022)	$3.79 \cdot 10^{-43}$
JUN	124	The transcription factor JUN regulates cytokine expression (Schonthaler et al., 2011; Johnson Rowsey, 2013) as well as *UCP1* gene expression in brown adipocytes (Xiao et al., 2019)	$3.78 \cdot 10^{-33}$
VEGFA	112	VEGFA (vascular endothelial growth factor A) was previously characterized in Table 1	$6.72 \cdot 10^{-28}$
TLR4	109	TLR4 is a cell surface receptor activated by lipopolysaccharides, which contributes to fever (Roth, Blatteis, 2014) and affects brown fat cell differentiation (Okla et al., 2018)	$2.04 \cdot 10^{-18}$
TNFA*	107	TNFA (tumor necrosis factor A) is a cytokine that can induce fever (Leon et al., 1998; Netea et al., 2000; Gil et al., 2007), and also affects gene expression in adipocytes (Valladares et al., 2001; Kim et al., 2006)	$1.78 \cdot 10^{-30}$

Note. Proteins are listed in descending order of the number of interactions in the gene network.
* TNFA is encoded by the *TNF* gene.

objects of the network using the ANDVisio program showed that microRNAs hsa-mir-335 and hsa-mir-26b regulate the expression of a significantly larger number of genes from the thermoregulatory network than would be expected by chance (P-adjusted < 0.01).

The two microRNAs mentioned above are important for thermoregulatory processes (Table 3). So, hsa-mir-335 regulates the expression of thermoreceptors TRPM8 and TRPV4, as well as the VEGFA protein, one of the key proteins for thermoregulation, which is involved in 112 interactions in the network. The hsa-mir-26b microRNA regulates the expression of JUN (Jun proto-oncogene, AP-1 transcription factor subunit), which is involved in 124 interactions in the network. As noted above, JUN affects the expression of thermogenin (uncoupling protein 1, UCP1) in brown fat cells (Xiao et al., 2019). This microRNA also regulates the expression of the EDN2 (endothelin-2) protein, which controls vasoconstriction, a process that mediates physical thermoregulation (Inoue et al., 1989).

The list of genes associated with thermoregulation we have created contains the *MIR21* and *MIRLET7c* genes. The microRNAs encoded by these genes, hsa-mir-21 and hsa-let-7c, regulate cellular processes in response to elevated temperature (Jiang et al., 2016; Permenter et al., 2019). The effect of the hsa-mir-21 and hsa-let-7c microRNAs on the expression of 15 and 5 proteins, respectively, was revealed in the reconstructed gene network (Table 3).

Among the proteins, the expression of which is regulated by hsa-mir-21, VEGFA (vascular endothelial growth factor A) was found to have 112 interactions in the network (Table 3). Multiple mentions of this protein in this report are an evidence of its important role in thermoregulation. Among the proteins, the expression of which is controlled by hsa-let-7c, the following were identified: a) COX2, a subunit of cytochrome c oxidase, involved in mitochondrial electron transport, encoded by the *MT-CO2* gene (Aich et al., 2018); b) DICER1, ribonuclease type III, involved in microRNA biogenesis (Wingo et al., 2015); c) CNOT3/NOT, CCR4-NOT transcription com-

Table 3. Characteristics of microRNAs with the highest functional significance within the network of human thermoregulation

MicroRNA	Gene encoding microRNA	Number of interactions in the network	P-adjusted	Regulated mRNAs*	Examples of functionally significant proteins encoded by mRNA targets of microRNA
MicroRNAs with the highest number of interactions in the network					
hsa-mir-335	*MIR3*	64	< 0.001	**TRPM8**, **TRPV4**, **VEGFA**, ANO1, ANO3, NPR3, AQP5, ARRDC3, ACVR2B, BAAT, CASQ1, CD14, CD36, CDKN1A, CRNN, DDIT3, DNAJC3, DBH, EIF2AK3, ELOVL6, FABP4, FOS, FOXO1, ABAT, GRB10, HDAC6, HMOX1, HSPA1A, HSPA1B, HSPB3, IGF2BP2, IGF1R, NFKBIA, IL1A, IL4, JAK2, KCNK4, KDM6B, LEPR, MOCOS, AVP, NOS3, NPY, NR1D1, NR2F6, NTSR1, PLA2G7, PTGS2, PPARGC1A, PTGES, RB1, SLC27A1, SCARA5, SCN9A, SQSTM1, STAT6, TCIM, TFE3, PTH2, TAC4, TMEM135, NGFR, TSHR, WNT10B	Thermoreceptors TRPM8 and TRPV4, and growth factor VEGFA, involved in 112 interactions in the network
hsa-mir-26b	*MIR26b*	39	< 0.01	**JUN**, **EDN2**, ACADM, ADRA2A, AGTR1, AKT1, BAG3, CASP8, CASQ1, CAV1, CD36, STUB1, CHORDC1, CRYAA, CXCR4, DNAJA2, DNAJA3, DNAJB4, EIF2AK3, EIF2B1, GRIK2, HADH, HMOX1, HSF1, IER5, NOX3, NRDC, NTSR1, PARK7, PDCL3, PTGS2, RBM3, RRAGC, SLC25A44, SMS, STAT6, VCP, TNFRSF11A, ZNF423	Transcription factor JUN, involved in 124 interactions in the network, EDN2 (endothelin-2), controlling vasoconstriction
microRNAs encoded by genes from the list of 469 genes associated with thermoregulation					
hsa-mir-21	*MIR21*	15	< 0.05	**VEGFA**, PRKAB2, ALMS1, APC, CPEB3, DAXX, DOCK7, EIF2S1, IL1B, PARP1, RB1, RDH11, RRAGC, SMARCA4, STAT3	VEGFA (vascular endothelial growth factor A), involved in 112 regulatory interactions in the network
hsa-let-7c	*MIRLET7c*	5	> 0.05	**MT-CO2/COX2**, **DICER1**, **CNOT3**, IP6K1, QKI	COX2, involved in mitochondrial electron transport (Aich et al., 2018), DICER1, involved in microRNA biogenesis (Wingo et al., 2015), CNOT3/NOT, participating in microRNA-mediated mRNA degradation (Wakiyama et al., 2022)

* mRNAs the translation of which is regulated by this microRNA (mRNAs encoding proteins described in the right column are underlined).

plex subunit 3, participating in microRNA-mediated mRNA degradation (Wakiyama, Takimoto, 2022).

Phylostratigraphic age of genes involved in the gene network of human thermoregulation (PAI-based analysis). The analysis of the evolutionary age of genes was carried out using the PAI (phylostratigraphic age index), the data on which were obtained from the *Genes_evol* information table from the Termo_Reg_Human 1.0 knowledge base. The phylostratigraphic age index was calculated using the Orthoweb system (Ivanov et al., 2024) as proposed in our previous studies (Mustafin et al., 2017). We constructed a distribution of PAI values for 467 protein-coding genes functioning in the thermoregulation gene network described in the Termo_Reg_Human 1.0 knowledge base (the *Thermoregulation_467* gene set, in Figure 5 this distribution is marked with orange bars). It turned out that this distribution has two maxima. The first of them is observed at PAI = 1 (176 genes, 38 % of their total list). The phylostratigraphic index PAI = 1 corresponds to the evolutionary stage of the emergence of unicellular organisms. The second peak is observed at PAI = 6 (100 genes associated with thermoregulation, 22 % of their total list). The phylostratigraphic index PAI = 6 corresponds to the evolutionary stage of the Vertebrata divergence.

To evaluate the statistical significance of the two peaks, a reference PAI index distribution was constructed for all human protein-coding genes (19,504 genes, the *all_CDS_19504* gene set, marked in blue in Figure 5), as it was done in our previous study (Mikhailova et al., 2024). This distribution also has two, but less noticeable, peaks. Using the chi-square method, the number of genes from the *Thermoregulation_467* gene set falling into peaks 1 and 6 was compared with the number of genes expected for random reasons in these peaks. In both cases, a difference was found between the observed and expected number with the level of significance $p < 0.05$

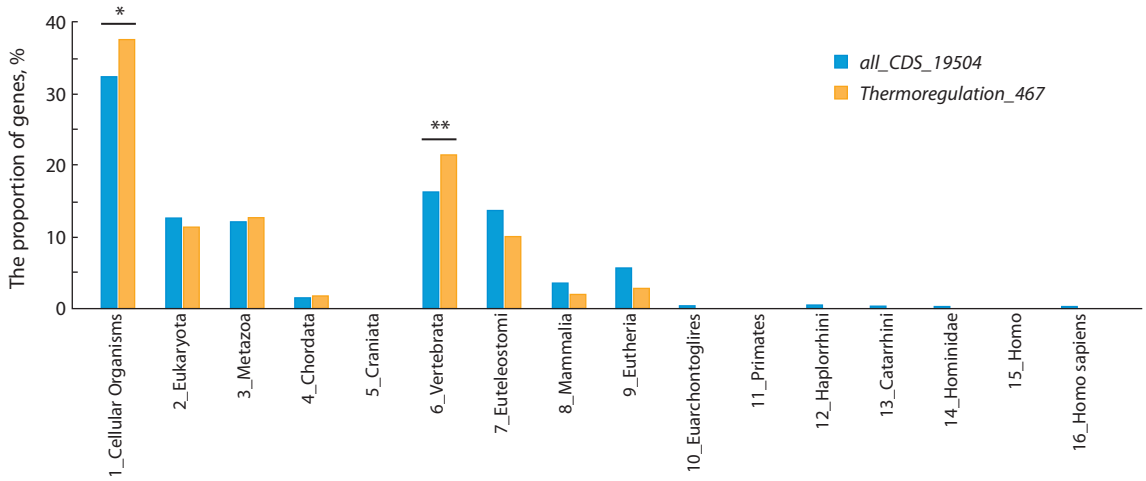


Fig. 5. Distribution of PAI values for protein-coding genes associated with thermoregulation (*Thermoregulation_467* set) and for all human protein-coding genes (*all_CDS_19504* set).

One asterisk (*) indicates a significant ($p < 0.05$) excess of the observed number of genes associated with thermoregulation corresponding to PAI = 1 (unicellular organisms, the root of the phylostratigraphic tree) over the expected number of genes with PAI = 1 calculated based on the distribution of PAI values for the complete set of protein-coding genes (*all_CDS_19504* set). Two asterisks (**) show a significant ($p < 0.01$) excess of the observed number of genes associated with thermoregulation corresponding to PAI = 6 (the stage of Vertebrata divergence) over their expected number.

and $p < 0.01$ (Supplementary Materials S4 and S5). Thus, it was shown that the gene network of thermoregulation was enriched with genes, the ancestral forms of which originated at the early evolutionary stage (emergence of unicellular organisms, the root of the phylostratigraphic tree) and at the stage of Vertebrata divergence.

Functional analysis of the genes from the *Thermoregulation_467* set performed using the DAVID tool showed that a group of genes with PAI = 1 is enriched with associations with the Gene Ontology terms related to transcription regulation (Supplementary Material S6), the most important mechanism for regulating gene expression in unicellular organisms. As for the group of genes with an index value of PAI = 6, it is enriched with genes involved in signal transduction (Supplementary Material S7), a vital process that ensures intercellular communications in a multicellular organism. This result is consistent with the idea that the interactions of a great number of physiological systems of the body (respiratory, circulatory, muscular, nervous, etc.) play a crucial role in the thermoregulation of the human body (Tansey, Johnson, 2015; Nakamura, 2024). In this case, the process of transcription provides genetic control over cell differentiation and formation of tissues involved in thermoregulation, and the coordination of the activity of physiological systems that ensure thermoregulation is carried out at the cellular level through signal transduction pathways.

Conclusion

In this study, a gene network comprising human genes, microRNAs, and proteins associated with thermoregulation was built. Additionally, the Termo_Reg_Human 1.0 knowledge base was developed to systematize current data on the molecular and genetic mechanisms underlying thermoregulatory processes. Based on data contained in the knowledge base, the prioritization of genes, proteins and microRNAs by the number of interactions in the network of thermoregulation was carried out, and the evolutionary characteristics of the genes were identified.

Enrichment of the thermoregulation gene network with genes, the ancestors of which were formed at the evolutionary stages of unicellular organisms and Vertebrata divergence, was revealed. The patterns in the evolution of the genes we discovered should be taken into account when developing new concepts for the emergence of endothermy across different animal taxa (Osvath et al., 2024).

References

Aich A., Wang C., Chowdhury A., Ronsör C., Pacheu-Grau D., Richter-Dennerlein R., Dennerlein S., Rehling P. COX16 promotes COX2 metallation and assembly during respiratory complex IV biogenesis. *eLife*. 2018;7:e32572. doi 10.7554/eLife.32572

Benjamini Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4)1165-1188. doi 10.1214/aos/1013699998

Blondin D.P., Haman F. Shivering and nonshivering thermogenesis in skeletal muscles. *Handb Clin Neurol*. 2018;156:153-173. doi 10.1016/B978-0-444-63912-7.00010-2

Cao D.S., Yu S.Q., Premkumar L.S. Modulation of transient receptor potential Vanilloid 4-mediated membrane currents and synaptic transmission by protein kinase C. *Mol Pain*. 2009;5:5. doi 10.1186/1744-8069-5-5

Charkoudian N., Hart E.C.J., Barnes J.N., Joyner M.J. Autonomic control of body temperature and blood pressure: influences of female sex hormones. *Clin Auton Res*. 2017;27(3):149-155. doi 10.1007/s10286-017-0420-z

Cheng W., Yang F., Liu S., Colton C.K., Wang C., Cui Y., Cao X., Zhu M.X., Sun C., Wang K., Zheng J. Heteromeric heat-sensitive transient receptor potential channels exhibit distinct temperature and chemical response. *J Biol Chem*. 2012;287(10):7279-7288. doi 10.1074/jbc.M111.305045

Cui S., Yu S., Huang H.Y., Lin Y.C.D., Huang Y., Zhang B., Xiao J., … Chen B., Zhang H., Fu J., Wang L., Huang H.-D. miRTarBase 2025: updates to the collection of experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 2025;53(D1):D147-D156. doi 10.1093/nar/gkae1072

Dumont L., Richard G., Espagnet R., Frisch F., Fortin M., Samson A., Bouchard J., … Dubreuil S., Guérin B., Turcotte É.E., Carpentier A.C., Blondin D.P. Shivering, but not adipose tissue thermogenesis, increases as a function of mean skin temperature in cold-exposed men and women. *Cell Metab*. 2025;37(9):1789-1805.e4. doi 10.1016/j.cmet.2025.06.010

Festuccia W.T., Blanchard P.-G., Turcotte V., Laplante M., Sariahmetoglu M., Brindley D.N., Deshaies Y. Depot-specific effects of the PPARγ agonist rosiglitazone on adipose tissue glucose uptake and metabolism. *J Lipid Res*. 2009;50(6):1185-1194. doi 10.1194/jlr.M800620-JLR200

Gil A., María Aguilera C., Gil-Campos M., Cañete R. Altered signalling and gene expression associated with the immune system and the inflammatory response in obesity. *Br J Nutr*. 2007;98(Suppl. 1):S121-S126. doi 10.1017/S0007114507838050

Gouin O., L'Herondelle K., Lebonvallet N., Le Gall-Ianotto C., Sakka M., Buhé V., Plée-Gautier E., Carré J.L., Lefeuvre L., Misery L., Le Garrec R. TRPV1 and TRPA1 in cutaneous neurogenic and chronic inflammation: pro-inflammatory response induced by their activation and their sensitization. *Protein Cell*. 2017;8(9):644-661. doi 10.1007/s13238-017-0395-5

Horii Y., Shiina T., Uehara S., Nomura K., Shimaoka H., Horii K., Shimizu Y. Hypothermia induces changes in the alternative splicing pattern of cold-inducible RNA-binding protein transcripts in a non-hibernator, the mouse. *Biomed Res*. 2019;40(4):153-161. doi 10.2220/biomedres.40.153

Ignatieva E.V., Igoshin A.V., Yudin N.S. A database of human genes and a gene network involved in response to tick-borne encephalitis virus infection. *BMC Evol Biol*. 2017;17(Suppl. 2):259. doi 10.1186/s12862-017-1107-8

Ikeda K., Yamada T. UCP1 dependent and independent thermogenesis in brown and beige adipocytes. *Front Endocrinol* (*Lausanne*). 2020;11:498. doi 10.3389/fendo.2020.00498

Inoue A., Yanagisawa M., Kimura S., Kasuya Y., Miyauchi T., Goto K., Masaki T. The human endothelin family: three structurally and pharmacologically distinct isopeptides predicted by three separate genes. *Proc Natl Acad Sci USA*. 1989;(8):2863-2867. doi 10.1073/pnas.86.8.2863

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanov R.A., Mukhin A.M., Kazantsev F.V., Mustafin Z.S., Afonnikov D.A., Matushkin Y.G., Lashin S.A. Orthoweb: a software package for evolutionary analysis of gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Jiang T., Wang X., Wu W., Zhang F., Wu S. Let-7c miRNA inhibits the proliferation and migration of heat-denatured dermal fibroblasts through down-regulating HSP70. *Mol Cells*. 2016;39(4):345-351. doi 10.14348/molcells.2016.2336

Johnson Rowsey P. Thermoregulation: cytokines involved in fever and exercise. *Annu Rev Nurs Res*. 2013;31:19-46. doi 10.1891/0739-6686.31.19

Kim J.Y., Tillison K., Lee J.H., Rearick D.A., Smas C.M. The adipose tissue triglyceride lipase ATGL/PNPLA2 is downregulated by insulin and TNF-α in 3T3-L1 adipocytes and is a target for transactivation by PPARγ. *Am J Physiol Endocrinol Metab*. 2006;291(1):E115-E127. doi 10.1152/ajpendo.00317.2005

Kudsi S.Q., Piccoli B.C., Ardisson-Araújo D., Trevisan G. Transcriptional landscape of *TRPV1*, *TRPA1*, *TRPV4*, and *TRPM8* channels throughout human tissues. *Life Sci*. 2022;308:120977. doi 10.1016/j.lfs.2022.120977

Leon L.R., White A.A., Kluger M.J. Role of IL-6 and TNF in thermoregulation and survival during sepsis in mice. *Am J Physiol*. 1998;275(1):R269-R277. doi 10.1152/ajpregu.1998.275.1.R269

Li Y., Adamek P., Zhang H., Tatsui C.E., Rhines L.D., Mrozkova P., Li Q., … Jawad A.B., Ghetti A., Yan J., Palecek J., Dougherty P.M. The cancer chemotherapeutic paclitaxel increases human and rodent sensory neuron responses to TRPV1 by activation of TLR4. *J Neurosci*. 2015;35(39):13487-13500. doi 10.1523/jneurosci.1956-15.2015

McCafferty D.J., Pandraud G., Gilles J., Fabra-Puchol M., Henry P.Y. Animal thermoregulation: a review of insulation, physiology and behaviour relevant to temperature control in buildings. *Bioinspir Biomim*. 2017;13(1):011001. doi 10.1088/1748-3190/aa9a12

Mikhailova A.D., Lashin S.A., Ivanisenko V.A., Demenkov P.S., Ignatieva E.V. Reconstruction and computer analysis of the structural and functional organization of the gene network regulating cholesterol biosynthesis in humans and the evolutionary characteristics of the genes involved in the network. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2024;28(8):864-873. doi 10.18699/vjgb-24-94

Mittag J., Kolms B. Hypothalamic control of heart rate and body temperature by thyroid hormones. *Rev Endocr Metab Disord*. 2025. doi 10.1007/s11154-025-09966-5

Mustafin Z.S., Lashin S.A., Matushkin Y.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics*. 2017;18(Suppl. 1):1427. doi 10.1186/s12859-016-1427-5

Mustafin Z.S., Zamyatin V.I., Konstantinov D.K., Doroshkov A.V., Lashin S.A., Afonnikov D.A. Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in *A. thaliana*. *Genes*. 2019;10(12):963. doi 10.3390/genes10120963

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Naik N.A., Bhat I.A., Afroze D., Rasool R., Mir H., Andrabi S.I., Shah S., Siddiqi M.A., Shah Z.A. Vascular endothelial growth factor A gene (VEGFA) polymorphisms and expression of VEGFA gene in lung cancer patients of Kashmir Valley (India). *Tumour Biol*. 2012;33(3):833-839. doi 10.1007/s13277-011-0306-y

Nakamura K. Central circuitries for body temperature regulation and fever. *Am J Physiol Regul Integr Comp Physiol*. 2011;301(5):R1207-R1228. doi 10.1152/ajpregu.00109.2011

Nakamura K. Central mechanisms of thermoregulation and fever in mammals. *Adv Exp Med Biol*. 2024;1461:141-159. doi 10.1007/978-981-97-4584-5_10

Netea M.G., Kullberg B.J., Van der Meer J.W. Circulating cytokines as mediators of fever. *Clin Infect Dis*. 2000;31(Suppl. 5):S178-S184. doi 10.1086/317513

Nguyen M.Q., Wu Y., Bonilla L.S., von Buchholtz L.J., Ryba N.J.P. Diversity amongst trigeminal neurons revealed by high throughput single cell sequencing. *PLoS One*. 2017;12(9):e0185543. doi 10.1371/journal.pone.0185543

O'Brien J., Hayder H., Zayed Y., Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front Endocrinol* (*Lausanne*). 2018;9:402. doi 10.3389/fendo.2018.00402

Okla M., Wang W., Kang I., Pashaj A., Carr T., Chung S. Activation of Toll-like receptor 4 (TLR4) attenuates adaptive thermogenesis via endoplasmic reticulum stress. *J Biol Chem*. 2015;290(44):26476-26490. doi 10.1074/jbc.M115.677724

Okla M., Zaher W., Alfayez M., Chung S. Inhibitory effects of Toll-like receptor 4, NLRP3 inflammasome, and interleukin-1β on white adipocyte browning. *Inflammation*. 2018;41(2):626-642. doi 10.1007/s10753-017-0718-y

Osvath M., Němec P., Brusatte S.L., Witmer L.M. Thought for food: the endothermic brain hypothesis. *Trends Cogn Sci*. 2024;28(11):998-1010. doi 10.1016/j.tics.2024.08.002

Permenter M.G., McDyre B.C., Ippolito D.L., Stallings J.D. Alterations in tissue microRNA after heat stress in the conscious rat: potential biomarkers of organ-specific injury. *BMC Genomics*. 2019;20(1):141. doi 10.1186/s12864-019-5515-6

Rehman R., Bhat Y.A., Panda L., Mabalirajan U. TRPV1 inhibition attenuates IL-13 mediated asthma features in mice by reducing airway epithelial injury. *Int Immunopharmacol*. 2013;15(3):597-605. doi 10.1016/j.intimp.2013.02.010

Roth J., Blatteis C.M. Mechanisms of fever production and lysis: lessons from experimental LPS fever. *Compr Physiol*. 2014;4(4):1563-1604. doi 10.1002/cphy.c130033

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Dosenko V.E., Zolotareva O.I., Choynzonov E.L., Hofestaedt R., Ivanisenko V.A. Search for new candidate genes involved in the comorbidity of asthma and hypertension based on automatic analysis of scientific literature. *J Integr Bioinform*. 2018;15(4):20180054. doi 10.1515/jib-2018-0054

Schonthaler H.B., Guinea-Viniegra J., Wagner E.F. Targeting inflammation by modulating the Jun/AP-1 pathway. *Ann Rheum Dis*. 2011;70(Suppl. 1):i109-i112. doi 10.1136/ard.2010.140533

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Song L., Cao X., Ji W., Zhao L., Yang W., Lu M., Yang J. Inhibition of STAT3 enhances UCP1 expression and mitochondrial function in brown adipocytes. *Eur J Pharmacol*. 2022;926:175040. doi 10.1016/j.ejphar.2022.175040

Tansey E.A., Johnson C.D. Recent advances in thermoregulation. *Adv Physiol Educ*. 2015;39(3):139-148. doi 10.1152/advan.00126.2014

Tattersall G.J., Sinclair B.J., Withers P.C., Fields P.A., Seebacher F., Cooper C.E., Maloney S.K. Coping with thermal challenges: physiological adaptations to environmental temperatures. *Compr Physiol*. 2012;2(3):2151-2202. doi 10.1002/cphy.c110055

Valdivia L.F.G., Castro É., Eichler R.A.D.S., Moreno M.F., de Sousa É., Jardim G.F.R., Peixoto Á.S., Moraes M.N., Castrucci A.M.L., Nedergaard J., Petrovic N., Festuccia W.T., Reckziegel P. Cold acclimation and pioglitazone combined increase thermogenic capacity of brown and white adipose tissues but this does not translate into higher energy expenditure in mice. *Am J Physiol Endocrinol Metab*. 2023;324(4):E358-E373. doi 10.1152/ajpendo.00217.2022

Valladares A., Roncero C., Benito M., Porras A. TNF-α inhibits UCP-1 expression in brown adipocytes via ERKs. Opposite effect of p38MAPK. *FEBS Lett*. 2001;493(1):6-11. doi 10.1016/s0014-5793(01)02264-5

Wakiyama M., Takimoto K. N-terminal Ago-binding domain of GW182 contains a tryptophan-rich region that confer binding to the CCR4-NOT complex. *Genes Cells*. 2022;27(9):579-585. doi 10.1111/gtc.12974

Wingo A.P., Almli L.M., Stevens J.S., Klengel T., Uddin M., Li Y., Bustamante A.C., … Bradley B., Binder E.B., Jin P., Gibson G., Ressler K.J. DICER1 and microRNA regulation in post-traumatic stress disorder with comorbid depression. *Nat Commun*. 2015;6:10106. doi 10.1038/ncomms10106

Wollenberg Valero K.C., Pathak R., Prajapati I., Bankston S., Thompson A., Usher J., Isokpehi R.D. A candidate multimodal functional genetic network for thermal adaptation. *PeerJ*. 2014;2:e578. doi 10.7717/peerj.578

Xiao F., Guo Y., Deng J., Yuan F., Xiao Y., Hui L., Li Y., … Chen Y., Ying H., Zhai Q., Chen S., Guo F. Hepatic c-Jun regulates glucose metabolism via FGF21 and modulates body temperature through the neural signals. *Mol Metab*. 2019;20:138-148. doi 10.1016/j.molmet.2018.12.003

Yoshida A., Furube E., Mannari T., Takayama Y., Kittaka H., Tominaga M., Miyata S. TRPV1 is crucial for proinflammatory STAT3 signaling and thermoregulation-associated pathways in the brain during inflammation. *Sci Rep*. 2016;6:26088. doi 10.1038/srep26088

Zhu W., Oxford G.S. Phosphoinositide-3-kinase and mitogen activated protein kinase signaling pathways mediate acute NGF sensitization of TRPV1. *Mol Cell Neurosci*. 2007;34(4):689-700. doi 10.1016/j.mcn.2007.01.005

Conflict of interest. The authors declare no conflict of interest.
Received September 18, 2025. Revised October 13, 2025. Accepted October 13, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1020-1030

doi 10.18699/vjgb-25-107

Searching for biological processes as targets for rheumatoid arthritis targeted therapy with ANDSystem, an integrated software and information platform

E.L. Mishchenko[1], I.V. Yatsyk[1] ✉, P.S. Demenkov[1], A.V. Adamovskaya[1], T.V. Ivanisenko[1], M.A. Kleshchev[1], V.A. Ivanisenko[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ woikin88@mail.ru

Abstract. Rheumatoid arthritis (RA) is a systemic autoimmune disease characterized primarily by joint involvement with progressive destruction of cartilage and bone tissue. To date, RA remains an incurable disease that leads to a significant deterioration in quality of life and patient disability. Despite a wide arsenal of disease-modifying antirheumatic drugs, approximately 40 % of patients show an insufficient response to standard treatment, highlighting the urgent need to identify new pharmacological targets. The aim of this study was to search for novel biological processes that could serve as promising targets for the targeted therapy of RA. To achieve this goal, we employed an approach based on the automated extraction of knowledge from scientific publications and biomedical databases using the ANDSystem software. This approach involved the reconstruction and subsequent analysis of two types of associative gene networks: a) gene networks describing genes and proteins associated with the development of RA, and b) gene networks describing genes and proteins involved in the functional responses to drugs used for the disease's therapy. The analysis of the reconstructed networks identified 11 biological processes that play a significant role in the pathogenesis of RA but are not yet direct targets of existing disease-modifying antirheumatic drugs. The most promising of these, described by Gene Ontology terms, include: a) the Toll-like receptor signaling pathway; b) neutrophil activation; c) regulation of osteoblast differentiation; d) regulation of osteoclast differentiation; e) the prostaglandin biosynthetic process, and f) the canonical Wnt signaling pathway. The identified biological processes and their key regulators represent promising targets for the development of new drugs capable of improving the efficacy of RA therapy, particularly in patients resistant to existing treatments. The developed approach can also be successfully applied to the search for new targeted therapy targets for other diseases.

Key words: rheumatoid arthritis; gene networks; targeted therapy; ANDSystem

For citation: Mishchenko E.L., Yatsyk I.V., Demenkov P.S., Adamovskaya A.V., Ivanisenko T.V., Kleshchev M.A., Ivanisenko V.A. Searching for biological processes as targets for rheumatoid arthritis targeted therapy with ANDSystem, an integrated software and information platform. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1020-1030. doi 10.18699/vjgb-25-107

Funding. The study was supported by budget projects FWNR-2022-0020.

Применение программно-информационной системы ANDSystem для поиска мишеней таргетной терапии ревматоидного артрита на основе анализа биологических процессов

Е.Л. Мищенко[1], И.В. Яцык[1] ✉, П.С. Деменков[1], А.В. Адамовская[1], Т.В. Иванисенко[1], М.А. Клещев[1], В.А. Иванисенко[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ woikin88@mail.ru

Аннотация. Ревматоидный артрит (РА) – системное аутоиммунное заболевание, сопровождающееся поражением преимущественно суставов с прогрессирующей деструкцией хрящевой и костной тканей. До настоящего времени РА остается неизлечимым заболеванием, приводящим к значительному ухудшению качества жизни и инвалидизации пациентов. Несмотря на наличие широкого арсенала базисных противовоспалительных препаратов, около 40 % пациентов демонстрируют недостаточный ответ на стандартное лечение, что подчеркивает

острую необходимость поиска новых фармакологических мишеней. Целью настоящей работы был поиск новых биологических процессов, которые могут служить перспективными мишенями для таргетной терапии РА. Для достижения поставленной цели был применен подход, основанный на автоматическом извлечении знаний из текстов научных публикаций и биомедицинских баз данных с помощью программно-информационной системы ANDSystem. Данный подход включал реконструкцию и последующий анализ ассоциативных генных сетей двух типов: а) генные сети, описывающие гены и белки, ассоциированные с развитием РА, и б) генные сети, описывающие гены и белки, вовлеченные в функциональные ответы на действие лекарств, применяемых для терапии заболевания. В результате анализа реконструированных сетей выявлено 11 биологических процессов, играющих значимую роль в патогенезе ревматоидного артрита, но до сих пор не являющихся прямыми мишенями существующих базисных противовоспалительных препаратов. К числу наиболее перспективных относятся следующие процессы, описываемые терминами онтологии генов: а) сигнальный путь Toll-подобных рецепторов; б) активация нейтрофилов; в) регуляция дифференцировки остеобластов; г) регуляция дифференцировки остеокластов; д) биосинтез простагландинов; е) канонический сигнальный путь Wnt. Выявленные биологические процессы и их ключевые регуляторы представляют собой перспективные мишени для разработки новых лекарственных средств, способных повысить эффективность терапии РА, в том числе у пациентов, резистентных к существующим методам лечения. Разработанный подход может быть успешно использован для поиска новых мишеней таргетной терапии и при других заболеваниях.

Ключевые слова: ревматоидный артрит; генные сети; таргетная терапия; ANDSystem

Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by systemic inflammation that primarily affects the joints and leads to progressive destruction of cartilage and bone tissue (Guo et al., 2018). According to the World Health Organization, RA affects approximately 0.5–0.6 % of the global population, occurring 2–3 times more frequently in women than in men, and is one of the leading causes of disability among working-age adults (Kvien et al., 2006; GBD 2023).

The pathogenesis of RA involves complex interactions between genetic factors, immune dysregulation, and environmental triggers, resulting in the activation of proinflammatory cytokines, infiltration of immune cells into the synovial membrane of the joints, and chronic inflammation (Firestein, McInnes, 2017). Despite significant progress in understanding the molecular mechanisms of RA, complete remission of the disease remains unattainable, and current therapeutic strategies are primarily aimed at preventing disease progression (Smolen et al., 2016).

Modern treatment strategies for rheumatoid arthritis are based on the use of several classes of drugs with anti-inflammatory effects (Ding et al., 2023; Smolen et al., 2023), including: a) conventional synthetic (cs) disease-modifying antirheumatic drugs (csDMARDs) such as methotrexate, leflunomide, sulfasalazine, and hydroxychloroquine; b) targeted synthetic (ts) DMARDs (tsDMARDs) such as tofacitinib and baricitinib; c) biologic DMARDs (bDMARDs), including inhibitors of tumor necrosis factor (infliximab, adalimumab), interleukin-6 (tocilizumab, sarilumab), interleukin-1 (anakinra), and anti-CD20 monoclonal antibodies (rituximab); d) nonsteroidal anti-inflammatory drugs (NSAIDs) for symptomatic treatment; and e) glucocorticoids (GCs) for rapid suppression of inflammation.

Particular attention in clinical practice is given to first-line drugs such as csDMARDs and tsDMARDs, which are capable of modulating immune responses at the level of intracellular signaling pathways and metabolism (van der Kooij et al., 2007). The action of tsDMARDs, in particular, targets specific genes encoding key components of the JAK/STAT signaling pathway. For instance, tofacitinib suppresses inflammation by specifically inhibiting Janus kinase 3 (JAK3), which plays a crucial role in cytokine signaling that regulates lymphocyte survival, proliferation, differentiation, and apoptosis (Adis Editorial, 2010). Although csDMARDs and tsDMARDs are effective in achieving remission in a substantial proportion of patients, their use is limited by side effects such as hepatotoxicity, immunosuppression, and the development of resistance (Olivera et al., 2020). Moreover, approximately 40 % of RA patients exhibit a poor response to therapy, and 5–20 % show no improvement at all with standard treatment (Smolen et al., 2016), highlighting the need to identify new molecular targets for the development of more effective therapeutic agents.

The development of rheumatoid arthritis involves a number of signaling pathways – including JAK/STAT, Notch, MAPK, Wnt, PI3K, SYK, and others – which regulate many biological processes implicated in the pathogenesis of the disease, such as the inflammatory response and remodeling of bone and cartilage tissue (Ding et al., 2023). These and other biological processes and signaling pathways can serve as potential targets for RA drug therapy. For example, experiments in laboratory mice have shown that treatment with CEP-33779 – a highly selective inhibitor of JAK2, a key component of the JAK/STAT signaling pathway – can reduce inflammatory manifestations in arthritis by suppressing cytokine production and the activation of T and B lymphocytes (Stump et al., 2011).

The aim of our study was to identify biological processes – new promising pharmacological targets for rheumatoid arthritis therapy – based on the reconstruction and analysis of a specific type of gene network known as an associative gene network (AGN).

A gene network is a group of coordinately functioning genes that control the phenotypic traits of an organism (Kolchanov et al., 2013). Interactions between genes within a gene network occur through their primary and secondary products – RNAs, proteins, and metabolites. An associative gene network represents an extension of the traditional gene network, integrating

genomic, molecular, phenotypic, and environmental entities and describing diverse types of interactions and associations among them (Demenkov et al., 2021).

To reconstruct AGNs, we used the ANDSystem software platform, which enables the automatic extraction of knowledge and facts from scientific publications and biomedical factual databases (Ivanisenko V.A. et al., 2019). To achieve this goal, the following tasks were addressed: a) reconstruction of an associative gene network for RA, including genes and proteins involved in the development of the disease; b) reconstruction of associative gene networks describing the mechanisms of action of drugs used in RA therapy, including genes and proteins participating in the functional response to these drugs; and c) identification, based on the reconstructed associative gene networks, of biological processes representing promising targets for RA therapy.

Based on the approach described above, 11 biological processes were identified that play a significant role in the development of rheumatoid arthritis but have not yet been recognized as direct targets of currently used disease-modifying antirheumatic drugs (DMARDs). These processes, described by Gene Ontology terms, include: a) the Toll-like receptor signaling pathway, b) neutrophil activation, c) regulation of osteoblast differentiation, d) regulation of osteoclast differentiation, e) prostaglandin biosynthetic process, and f) the canonical Wnt signaling pathway. The identified biological processes and their key regulators represent promising targets for the development of new therapeutic agents for rheumatoid arthritis. The approach implemented in this study can also be applied to the identification of novel targets for targeted therapy in other diseases.

Materials and methods

List of disease-modifying antirheumatic drugs (DMARDs). To compile a list of conventional synthetic DMARDs and targeted synthetic DMARDs used in the treatment of rheumatoid arthritis, we referred to the official document of the All-Russian Public Organization "Association of Rheumatologists of Russia" – "Clinical Guidelines: Rheumatoid Arthritis (ICD-10: M05, M06)" (Nasonov et al., 2024). This document provides a classification of drugs used for RA therapy, their pharmacotherapeutic characteristics, and Anatomical Therapeutic Chemical (ATC) classification codes. Based on these recommendations, the following list of drugs was compiled for further analysis: csDMARDs (methotrexate, leflunomide, sulfasalazine, hydroxychloroquine) and tsDMARDs (tofacitinib, baricitinib).

Reconstruction and analysis of associative gene networks. The reconstruction of associative gene networks was performed using the ANDSystem software and information platform (Ivanisenko V.A. et al., 2019, 2024; Ivanisenko T.V. et al., 2024). This system is based on methods of machine reading and artificial intelligence designed for the automatic extraction of knowledge and facts from large-scale genetic and biomedical data sources, such as scientific publications, patents, and factual databases.

Through the analysis of more than 40 million scientific articles and patents, as well as 150 factual databases, the ANDSystem knowledge base has accumulated biomedically significant information represented as semantic knowledge graphs, describing 12 types of biological entities (including genes, proteins, diseases, biological processes, drugs, etc.) and over 40 types of functional relationships among them. These relationships include gene expression regulation, protein degradation, modification, and transport, as well as physical interactions such as protein–protein and protein–ligand interactions.

In addition, the ANDSystem knowledge base contains descriptions of associative relationships linking genes, proteins, and metabolites with entities such as diseases, biological processes, and pharmaceutical compounds (Ivanisenko V.A. et al., 2019, 2024; Ivanisenko T.V. et al., 2024). The knowledge base also includes "marker" relationships, indicating that certain genes, proteins, biological processes, or phenotypic traits can serve as markers of specific diseases.

Identification of biological processes based on information from reconstructed associative gene networks. The analysis of overrepresented biological processes in the reconstructed associative gene networks was carried out using the DAVID web server, version 2021 (https://david.ncifcrf.gov/; Sherman et al., 2022), with default settings. DAVID evaluates the degree of overlap between the list of genes functioning within each reconstructed gene network and the lists of genes corresponding to biological processes described in the Gene Ontology (GO). Based on this comparison, the hypergeometric test was applied to calculate the probability that the observed overlap between gene lists could occur by chance. In our study, biological processes significantly associated with the reconstructed gene networks were identified using a p-value threshold of <0.05, corrected by the Bonferroni method. The biological processes that met this criterion were classified into two categories: a) biological processes significant for the rheumatoid arthritis gene network, and b) biological processes significant for the gene networks representing responses to csDMARD and tsDMARD therapies used in RA treatment.

Results

Reconstruction of the associative gene network of rheumatoid arthritis

Using the ANDSystem platform, we reconstructed the associative gene network of rheumatoid arthritis based on information contained in the ANDSystem knowledge base.

The graph of the reconstructed associative gene network had a star-shaped topology: the central node corresponding to the term "Rheumatoid arthritis" was connected by edges to other nodes of the network graph that represented proteins and genes associated with RA according to the ANDSystem knowledge base (Supplementary Fig. S1)[1]. In total, the graph contained 4,685 nodes, corresponding to 2,178 genes and 2,507 proteins (Table S1 in the Appendix), as well as 9,877 edges between the central node (rheumatoid arthritis) and the other nodes. Note that the number of edges exceeded the number of nodes. This is because the same node representing a gene or protein could be linked to the central node by multiple edges, each of which, according to the ANDSystem knowledge base, described a specific type of interaction between RA and a given gene or protein.

[1] Supplementary Figures S1 and S2 and Tables S1–S6 are available at: https://vavilovj-icg.ru/download/pict-2025-29/appx37.xlsx

Table 1. Characteristics of relationships between the central and peripheral nodes in the rheumatoid arthritis gene network

No.	Interaction type	Number of interactions	Proportion, %
	Regulatory interactions	4,381	44.4
1	Expression downregulation	93	0.9
2	Expression regulation	472	4.8
3	Expression upregulation	365	3.7
4	Activity downregulation	15	0.2
5	Activity regulation	26	0.3
6	Activity upregulation	10	0.1
7	Regulation	1,812	18.3
8	Upregulation	802	8.1
9	Downregulation	786	8.0
	Other Interactions	5,496	55.6
1	Association	4,449	45.0
2	Involvement	172	1.7
3	Marker	338	3.4
4	Risk factor	274	2.8
5	Treatment	263	2.7

* The percentage (%) indicates the proportion of a specific relationship type relative to the total number of relationships in the associative gene network of rheumatoid arthritis.

Table S1 lists the genes and proteins included in the reconstructed associative gene network of rheumatoid arthritis, which comprises, in particular, genes and proteins involved in the inflammatory process: interleukins (IL1, IL6, IL13, and others), members of the tumor necrosis factor (TNF) family, the key inflammatory regulator NF-κB, and genes and proteins functioning in the Wnt, JAK/STAT, Notch, MAPK, PI3K, and SYK signaling pathways, all of which are known to play a defining role in RA pathogenesis (Ding et al., 2024).

Table 1 presents a classification of 14 types of relationships between the central and peripheral nodes in the RA gene network. These relationships fall into two categories. The first category (regulatory relationships) comprises nine types, such as expression downregulation, expression upregulation, activity regulation, and others. For example, expression of interleukin-1 beta (IL1B) is increased in rheumatoid arthritis (Mohd et al., 2019), which is reflected in the ANDSystem knowledge base as an "expression upregulation" relationship between RA and the IL1B protein. Interleukin-6 (IL6) stimulates fibroblasts in the synovial membrane of the joints (Singh et al., 2021) and contributes to one of the symptoms of RA (bone loss), which is represented in ANDSystem as a "positive regulation" relationship between the disease "Rheumatoid arthritis" and the IL6 protein.

The second category (other relationships) includes five additional relationship types identified during the reconstruction of the RA gene network, describing situations in which a gene or protein is associated with RA in some way. For example, these may include structural or functional features of a gene if a mutation in that gene constitutes a risk factor for RA.

Based on the information contained in the associative gene network of rheumatoid arthritis and the ANDSystem knowledge base, it is possible to reconstruct the detailed mechanisms underlying the involvement of specific genes and proteins in the development of RA. Figure 1 illustrates, as an example, the regulatory interactions between genes and proteins functioning within the Wnt signaling pathway, which is regulated by proinflammatory cytokines such as interleukin-1 beta, tumor necrosis factor alpha (TNFA), and interleukin-6.

As shown in Figure 1, regulation of the Wnt signaling pathway in rheumatoid arthritis involves interleukin-1 beta, tumor necrosis factor alpha, and interleukin-6, which activate the expression of the *WNT5A* gene encoding the WNT5A protein – a ligand of FZD receptors participating in the non-canonical Wnt pathway (Miao et al., 2013). According to the ANDSystem data, WNT5A, in turn, activates the expression of the *IL1B* gene encoding interleukin-1 beta. Thus, *IL1B* and *WNT5A* mutually activate each other's expression, forming a positive feedback loop, indicated in Figure 1 by bold arrows.

Reconstruction of associative gene networks involved in functional responses to RA therapies

Figure 2 shows the AGN for responses to tsDMARDs (see also Table S2). It contains two nodes corresponding to the drug names (tofacitinib, baricitinib) and 157 edges linking these nodes to other nodes representing 22 proteins and 51 genes. As seen in Figure 2, according to the ANDSystem knowledge base, tofacitinib is characterized by a substantially larger number of interactions with proteins and genes (60) compared to baricitinib (26). In response to both drugs, genes involved in

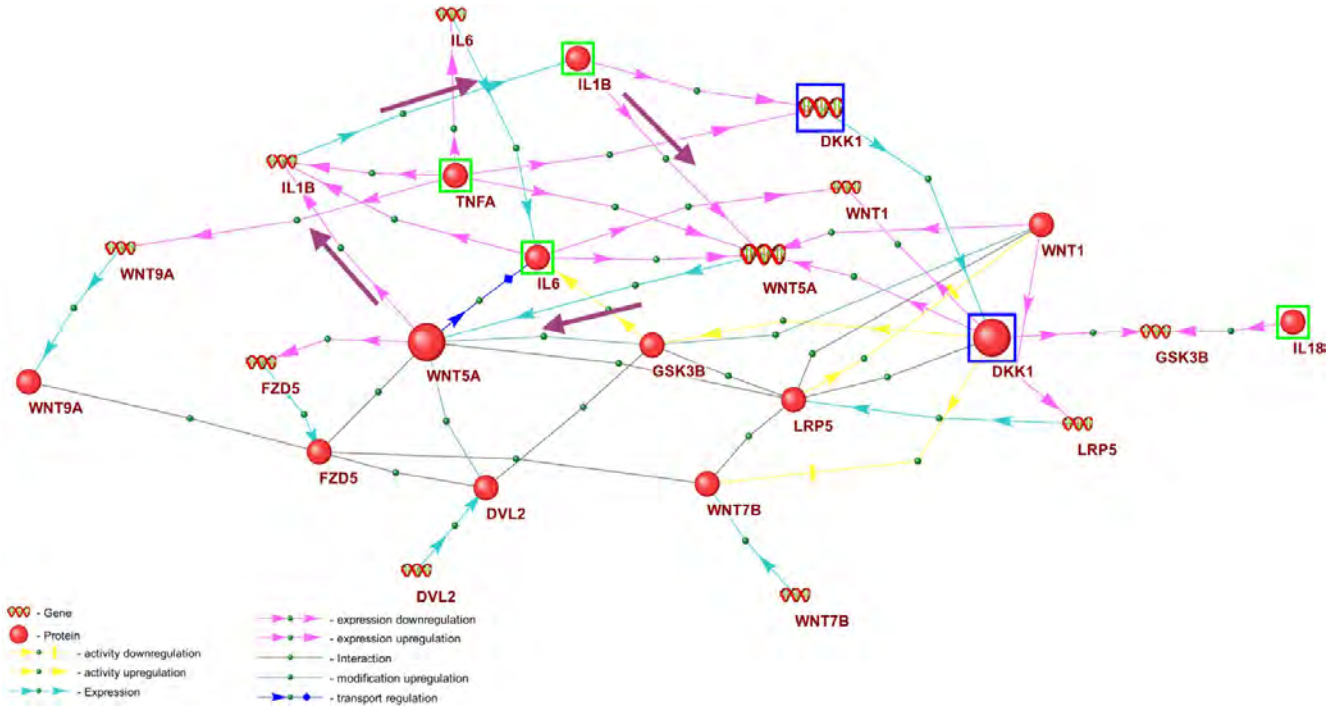


Fig. 1. Mechanism of regulation of key components of the Wnt signaling pathway by proinflammatory cytokines, reconstructed from the rheumatoid arthritis gene network in the ANDSystem knowledge base.
Proinflammatory cytokines are highlighted with green frames; components of the positive feedback regulatory loop are indicated with bold arrows; and the *DKK1* gene and its encoded protein Dickkopf-1 (DKK1) – an inhibitor of the canonical Wnt pathway – are shown in blue frames.

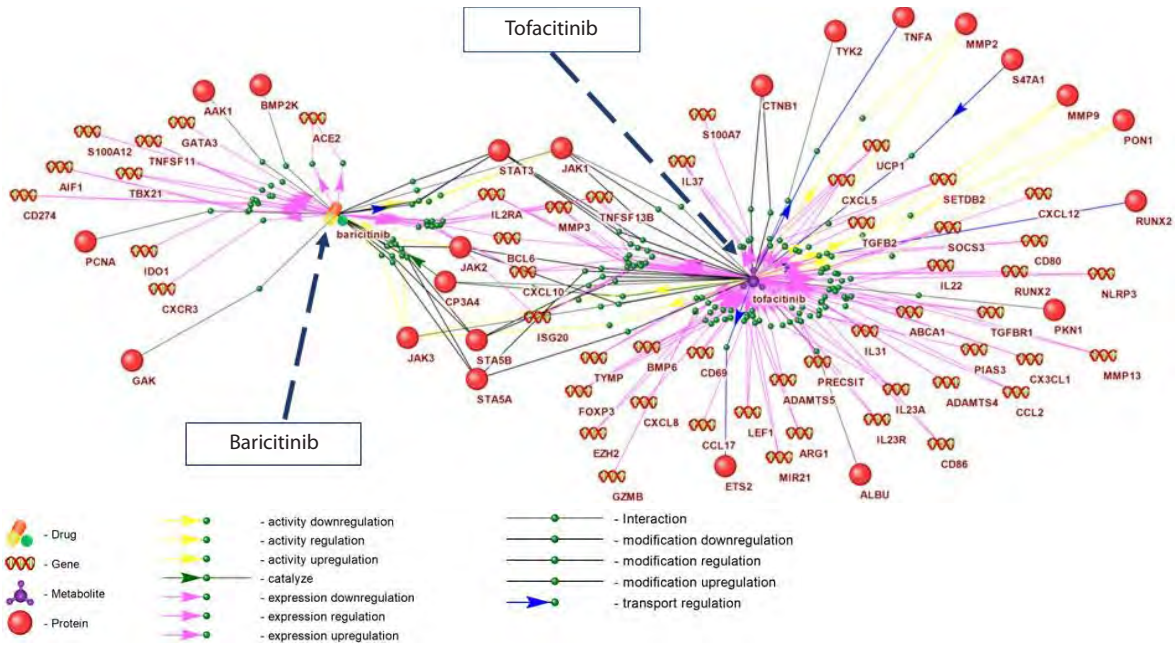


Fig. 2. Reconstructed associative gene network of the response to two targeted synthetic disease-modifying antirheumatic drugs – tofacitinib and baricitinib.

the inflammatory response – *MMP3*, *IL2RA*, *CXCL10* – and proteins (STAT3, STAT5A, JAK1, JAK2), members of the JAK/STAT pathway, were implicated.

Figure S2 presents the AGN for responses to csDMARDs (methotrexate, leflunomide, sulfasalazine, hydroxychloroquine). The graph contains 261 nodes, four of which correspond to the drug names (see also Table S2). The remaining nodes are connected to these four drug nodes by 485 edges and represent 106 proteins and 151 genes. The largest number of interactions in the csDMARD response AGN was observed for methotrexate (160). Proteins and genes associated with this drug include, in particular, IL1R1, TNFA, the inflammatory

Table 2. Distribution of interaction types in the reconstructed associative gene networks of the response to synthetic and targeted synthetic disease-modifying antirheumatic drugs

No.	Interaction type	csDMARD		tsDMARD	
		Interaction number	Interaction rate, %	Interaction number	Interaction rate, %
	Regulatory interactions	529	87.6	143	91.1
1	Expression downregulation	73	15.1	35	22.3
2	Expression regulation	158	32.6	57	36.3
3	Expression upregulation	64	13.2	12	7.6
4	Activity downregulation	18	3.7	6	3.8
5	Activity regulation	34	7.0	7	4.5
6	Activity upregulation	16	3.3	1	0.6
7	Modification downregulation	10	2.1	8	5.1
8	Modification regulation	8	1.6	9	5.7
9	Modification upregulation	4	0.8	3	1.9
10	Transport regulation	28	5.8	5	3.2
11	Degradation downregulation	5	1.0	No	No
12	Degradation regulation	6	1.2	No	No
13	Degradation upregulation	1	0.2	No	No
	Other interaction type	60	12.4	14	7.8
14	Catalyze	14	2.4	2	1.1
15	Physical interaction	46	7.8	12	6.7

transcription factor NFKB1, and caspases (CASP1, CASP3, CASP9). Hydroxychloroquine ranked second by number of interactions (73), being linked to proinflammatory cytokines such as IL1B and TNFA, as well as to catalase (CAT) and cytochromes involved in xenobiotic metabolism (CP2B6, CYP1B1). Sulfasalazine and leflunomide ranked third and fourth (26 and 17 interactions, respectively). Notably, some proteins in the csDMARD response AGN (e. g., IL1B, CCL2, TNFA, CASP3) are targets of multiple drugs.

The distribution of interaction types in the AGN of the response to csDMARDs and tsDMARDs is provided in Table 2. As can be seen from Table 2, regulatory interactions, particularly the regulation of gene expression, predominated among those in the AGN of the response to csDMARD and tsDMARD.

Identification of biological processes based on information from reconstructed associative gene networks

Using the DAVID web resource based on Gene Ontology, an overrepresentation analysis of biological processes in the reconstructed gene networks was performed for: a) the rheumatoid arthritis gene network and b) the gene networks of the response to two types of anti-inflammatory drugs (csDMARD and tsDMARD).

For the reconstructed associative gene networks of rheumatoid arthritis and the response to csDMARD and tsDMARD, 381, 64, and 44 overrepresented biological processes were identified, respectively. Most significant processes are characterized in Table 3 (for details, see Tables S4–S6). As seen in Table 3, the inflammatory response (GO identifier: GO:0006954) was statistically significantly overrepresented in both the RA gene network and the gene networks of the response to csDMARD and tsDMARD. It is interesting to note that the list of most significantly overrepresented processes for csDMARD response gene network included xenobiotic metabolic processes, which were not overrepresented in the tsDMARD gene network. For the tsDMARD response gene network, the JAK/STAT (GO identifier: GO:0007259, Table 3) and cytokine (GO identifier: GO:0019221, Table 3) signaling pathways were most significantly overrepresented.

For further analysis, from the 381 identified biological processes overrepresented in the RA AGN (Table 3), 71 processes were selected using the ANDSystem knowledge base, characterized by the interaction types “Regulation”, “Downregulation”, and “Upregulation” with the disease “Rheumatoid arthritis”. An intersection was performed between the list of 71 biological processes involved in the pathogenesis of RA and the lists of overrepresented biological processes for the AGN of the response to the csDMARD (64 processes) and tsDMARD (44 processes) drug groups. As a result, 59 biological processes were found that are involved in the pathogenesis of RA but are not included in the list of overrepresented pro-

Table 3. Results of the overrepresentation analysis of Gene Ontology (GO) biological processes for the associative gene networks of rheumatoid arthritis, as well as the gene networks of the response to synthetic disease-modifying antirheumatic drugs (csDMARD) and targeted synthetic disease-modifying antirheumatic drugs (tsDMARD)

Gene network	Overrepresented process number	The most statistically significant overrepresented biological processes		
		Identifier	Name	*p*-value*
Rheumatoid arthritis gene network	381	GO:0006954	Inflammatory response	$3.7 \cdot 10^{-123}$
		GO:0006955	Immune response	$8.2 \cdot 10^{-103}$
		GO:0007165	Signal transduction	$2.9 \cdot 10^{-60}$
csDMARD response gene network	64	GO:0006805	Xenobiotic metabolic process	$5.6 \cdot 10^{-21}$
		GO:0009410	Response to xenobiotic stimulus	$1.9 \cdot 10^{-19}$
		GO:0006954	Inflammatory response	$1.4 \cdot 10^{-14}$
tsDMARD response gene network	44	GO:0006954	Inflammatory response	$4.2 \cdot 10^{-16}$
		GO:0007259	Cell surface receptor signaling pathway via JAK/STAT	$2.2 \cdot 10^{-11}$
		GO:0019221	Cytokine-mediated signaling pathway	$3.0 \cdot 10^{-10}$

* $p < 0.05$.

Table 4. Biological processes for which no regulating drugs from the csDMARD and tsDMARD groups used in the therapy of rheumatoid arthritis have been identified

No.	The Gene Ontology identifier (GO)	The Gene Ontology biological process	The number of rheumatoid arthritis genes involved in the process	*p*-value*
1	GO:0034612	Response to tumor necrosis factor	58	$9.8 \cdot 10^{-23}$
2	GO:0031295	T cell costimulation	29	$3.3 \cdot 10^{-13}$
3	GO:0002224	Toll-like receptor signaling pathway	19	$8.3 \cdot 10^{-8}$
4	GO:0014823	Response to activity	26	$1.3 \cdot 10^{-7}$
5	GO:0034097	Response to cytokine	24	$8.64 \cdot 10^{-7}$
6	GO:0010468	Regulation of gene expression	53	$2.0 \cdot 10^{-3}$
7	GO:0045668	Negative regulation of osteoblast differentiation	27	$8.5 \cdot 10^{-5}$
8	GO:0042119	Neutrophil activation	12	$1.53 \cdot 10^{-3}$
9	GO:0045671	Negative regulation of osteoclast differentiation	15	$3.27 \cdot 10^{-2}$
10	GO:0001516	Prostaglandin biosynthetic process	12	$1.5 \cdot 10^{-2}$
11	GO:0060070	Canonical Wnt signaling pathway	30	$2.2 \cdot 10^{-2}$

* *p*-value – significance level of the overrepresentation of Gene Ontology terms for the set of genes associated with rheumatoid arthritis, with the Bonferroni correction.

cesses for the AGN of the response to the considered drugs. From these 59 processes, 48 were removed that, according to the ANDSystem knowledge base, are linked to the considered csDMARD (methotrexate, leflunomide, sulfasalazine, hydroxychloroquine) and tsDMARD (tofacitinib, baricitinib) drugs by interactions of the types "Regulation", "Downregulation", and "Upregulation".

This resulted in a list of 11 biological processes (Table 4). The identified processes are characterized by the following: firstly, these processes are involved in the pathogenesis of rheumatoid arthritis. Furthermore, no regulating csDMARDs and tsDMARDs have been identified for them. It is these processes that are of particular interest as targets for the development of drugs for rheumatoid arthritis therapy.

As seen from Table 4, the biological processes involved in the pathogenesis of rheumatoid arthritis but not regulated by disease-modifying antirheumatic drugs included: a) inflammatory responses (GO identifiers GO:0034097, GO:0034612, GO:0031295, GO:0002224); b) bone tissue morphogenesis (GO:0045668, GO:0045671); c) the canonical Wnt signal-

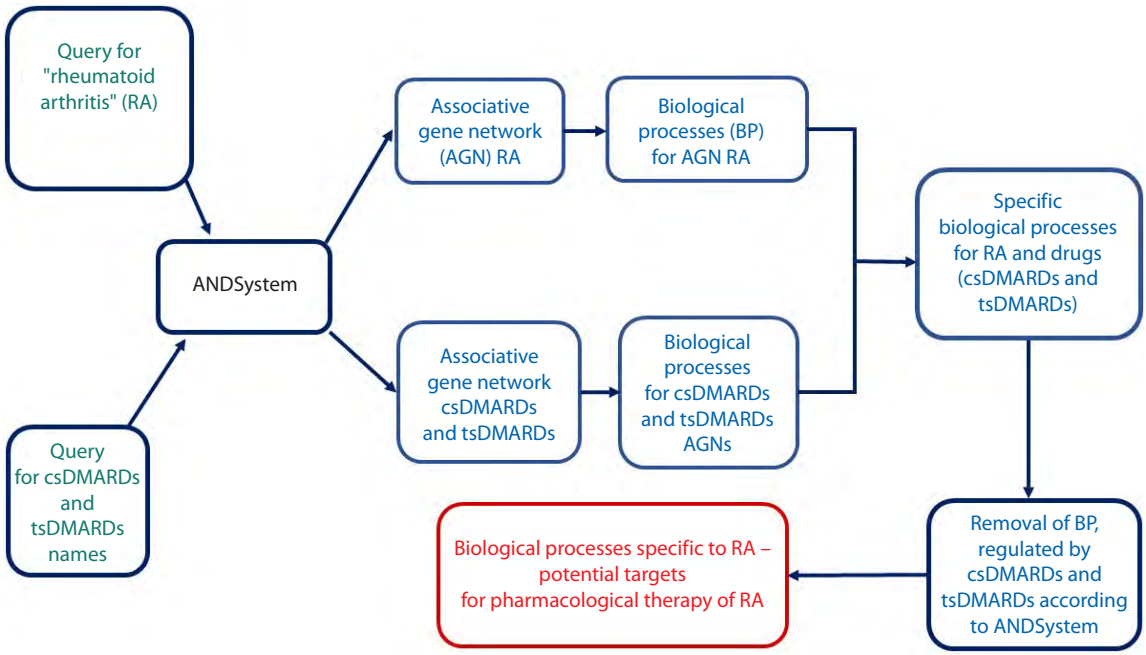


Fig. 3. Main stages for searching for biological processes promising as targets for the development of new antirheumatic drugs.

RA – rheumatoid arthritis; AGN – associative gene network; BP – biological process; csDMARD – conventional synthetic disease-modifying antirheumatic drugs (methotrexate, leflunomide, sulfasalazine, hydroxychloroquine); tsDMARD – targeted synthetic disease-modifying antirheumatic drugs (tofacitinib, baricitinib).

ing pathway (GO:0060070); d) prostaglandin biosynthesis (GO:0001516); e) response to activity (GO:0014823) and regulation of gene expression (GO:0010468).

Thus, we have conducted a search for biological processes – new promising pharmacological targets for RA therapy – based on the reconstruction and analysis of associative gene networks.

Figure 3 shows the schematic diagram, implemented in our work, for searching for biological processes that are new promising targets for the development of antirheumatic drugs.

Discussion

The search for new drug targets for the treatment of rheumatoid arthritis is important for modern medicine, given that up to 40 % of patients do not achieve a full response to existing therapy (Ding et al., 2023). In this regard, we have proposed a method for identifying biological processes as targets for new antirheumatic drugs, based on the reconstruction of associative gene networks and a comparative analysis of biological processes associated with rheumatoid arthritis and those regulated by the disease-modifying antirheumatic drugs currently used in clinical practice (Nasonov et al., 2024).

The ANDSystem knowledge base, which we used for reconstructing the gene networks, integrates accumulated information from scientific literature on the molecular mechanisms of drug action and disease pathogenesis, allowing for the discovery of new therapeutic targets at a systemic level, including biological processes, thereby increasing the efficacy of therapy and diagnostics. In our work, we reconstructed associative gene networks (AGNs) for rheumatoid arthritis, as well as AGNs describing the interactions of synthetic and targeted anti-inflammatory drugs with human genes and proteins. The analysis showed that the rheumatoid arthritis gene network is enriched with genes involved in the regulation of the inflammatory response, which corresponds to the well-known data on the leading role of systemic inflammation in the pathogenesis of this disease (Firestein, McInnes, 2017; Figus et al., 2021). It is therefore no coincidence that the reconstructed gene networks of proteins and genes targeted by csDMARDs (Fig. S2) and tsDMARDs (Fig. 2) primarily include genes and proteins involved in the functioning of the immune system.

According to the results of the functional annotation of genes, for conventional synthetic disease-modifying antirheumatic drugs, the list of statistically significantly overrepresented biological processes included processes related not only to inflammation but also to xenobiotic metabolism. This suggests that csDMARDs impose a significant load on the biochemical systems responsible for xenobiotic removal, potentially leading to serious adverse effects (Olivera et al., 2020).

On the other hand, for genes involved in the response to targeted synthetic disease-modifying antirheumatic drugs, xenobiotic metabolism processes were not significantly overrepresented. However, the list of overrepresented processes for tsDMARDs response gene network, along with inflammation, included processes related to the functioning of the JAK/STAT signaling pathway, which is crucial for pathogenesis of RA (Ding et al., 2023). This suggests a more targeted action of tsDMARD on the pathogenesis of RA and emphasizes the importance of developing targeted therapies to increase treatment efficacy and reduce side effects. However, the diversity and complexity of the interactions of biological processes leading to the development of RA, and the insufficient efficacy of therapy with existing disease-modifying antirheumatic drugs, necessitate the search for new targets for RA treatment (Smolen et al., 2016).

Our approach, based on the reconstruction of gene networks involved in the development of the disease and in the response to known drugs, as well as on a comparative analysis of the biological processes regulated by these gene networks, allowed us to identify 11 biological processes (Table 4). These processes are key to the pathogenesis of RA but are not targets of the anti-inflammatory drugs currently in use. It should be noted that the regulation of expression (GO:0010468) and the response to activity (GO:0014823) belong to a group of rather broad processes, covering many molecular mechanisms in the cell, which complicates the development of targeted drugs.

Literature analysis revealed that for processes such as the response to cytokines (GO:0034097), the response to tumor necrosis factor TNFA (GO:0034612), and T-cell co-stimulation (GO:0031295), there is evidence of their partial regulation by the currently used csDMARDs and tsDMARDs. For example, tsDMARDs like tofacitinib and baricitinib effectively block the JAK/STAT signaling pathways, which are downstream of cytokine and TNFA receptors, providing powerful suppression of inflammatory responses (Palmroth et al., 2021).

However, biological processes such as the Toll-like receptor signaling pathway, neutrophil activation, negative regulation of osteoblast differentiation, negative regulation of osteoclast differentiation, the canonical Wnt signaling pathway, and prostaglandin biosynthesis are not directly regulated by the disease-modifying antirheumatic drugs that are currently actively used by rheumatologists in accordance with clinical guidelines (Nasonov et al., 2024). Nevertheless, the biological processes and pathways listed above may be important for the pathogenesis of RA. For example, neutrophil activation plays an important role in inflammation in RA patients, and CXCR2 inhibitors, being investigated for other inflammatory conditions, could be adapted for RA (Alam et al., 2020).

It is known that the Wnt signaling pathway plays a significant role in fibroblast activation and synovial inflammation, as well as in bone resorption and joint destruction in the development of rheumatoid arthritis (Miao et al., 2013). The expression of genes encoding Wnt family proteins, which activate the Wnt signaling pathway, was increased in the synovium in rheumatoid arthritis, partly due to proinflammatory cytokines (Prajapati, Doshi et al., 2023). At the same time, the activation of the non-canonical Wnt signaling pathway, in turn, leads to an increased expression of inflammatory mediators, including the transcription factor NF-κB and cytokines (Miao et al., 2013), increasing inflammation.

According to the ANDSystem knowledge base (Fig. 1), interleukin-1 beta and the WNT5A protein mutually activate each other's expression, which may create a vicious cycle in the pathogenesis of rheumatoid arthritis. Therefore, modulating the Wnt signaling pathway may be a promising approach to reduce joint inflammation in RA. In particular, it has been shown that the NAV2 protein promotes the inflammatory response of fibrocyte-like synoviocytes by activating the Wnt signaling pathway in rheumatoid arthritis, and its inhibition can reduce joint inflammation in this disease (Wang R. et al., 2021).

On the other hand, proinflammatory cytokines – tumor necrosis factor-alpha and IL1B – according to ANDSystem (Fig. 2), can activate the expression of the *DKK1* gene, which encodes the Dickkopf-1 (DKK1) protein, an important inhibitor of the canonical Wnt signaling pathway (Rabelo et al., 2010). It has been shown that the serum level of DKK1 is elevated in patients with RA and correlates with the level of inflammation and the degree of bone destruction in the joints (Wang S.Y. et al., 2011). The activation of *DKK1* expression by proinflammatory cytokines in rheumatoid arthritis may lead to the suppression of the Wnt signaling pathway and, consequently, the activation of the RANK/RANKL signaling pathway in osteoclasts, increasing their activity and causing the bone loss characteristic of RA (Miao et al., 2013).

Thus, dysregulation of the Wnt signaling pathway may be the cause of changes in the biological processes of regulating osteoblast and osteoclast differentiation in RA, which, according to our study (Table 4), are potential targets for new antirheumatic drugs. Furthermore, DKK1 stimulates angiogenesis in the synovium and the formation of pannus – a pathologically altered synovial tissue that plays a crucial role in joint destruction in RA (Cici et al., 2019).

Thus, the Wnt signaling pathway is a promising target for the development of new antirheumatic drugs; however, its regulation in RA is very complex and depends on the type of tissues and cells, so further research is needed to reconstruct the gene network of this pathway in RA and analyze its structural and functional features in various cells and tissues.

Prostaglandins, particularly prostaglandin E2, are known to play an important role in the development of both acute inflammatory reactions and chronic inflammation (Kawahara et al., 2015), enhancing inflammatory processes by activating the expression of cytokine receptors and NFKB family proteins, which are key triggers of inflammation (Yao, Narumiya, 2019). Prostaglandin E2, an important mediator of inflammation in RA, is a target for a number of non-steroidal anti-inflammatory drugs (NSAIDs) for this disease (Park et al., 2006). The biosynthesis of prostaglandins (GO biological process identifier GO:0001516) is partially modulated by NSAIDs, such as celecoxib, but the development of more specific inhibitors could improve therapeutic outcomes (Gong et al., 2012).

It is known that toll-like receptors (TLRs) make an important contribution to the induction of inflammation, as their activation leads to increased activity of signaling pathways and a number of transcription factors such as nuclear factor κB (NF-κB), activator protein-1 (AP-1), and interferon regulatory factors (IRF), which induce the expression of proinflammatory cytokines – TNF, IL1β, IL6, and others (Kawasaki, Kawai, 2024). It has been shown that the expression of toll-like receptor genes is increased in the synovium of RA patients, and TLRs contribute significantly to the development of inflammation in RA, but therapeutic interventions targeting TLR signaling pathways have not yet been successfully introduced into clinical practice (Unterberger et al., 2021).

Thus, all the biological processes listed above play a major role in the development of RA, yet they are not regulated by the disease-modifying antirheumatic drugs currently used in clinical practice. Therefore, these biological processes and their key regulators can serve as targets for the development of new drugs for the treatment of rheumatoid arthritis.

It should be noted that rheumatoid arthritis is characterized by significant comorbidity with other diseases, including cardiovascular and respiratory diseases (Figus et al., 2021),

anxiety-depressive disorders (Hill et al., 2022), and osteoporosis (Llorente et al., 2020). In this regard, further work is planned to analyze the identified biological processes as a basis for the comorbidity of RA with other diseases.

Furthermore, this work did not identify targets at the gene level, which could be the subject of further research based on the analysis of the structural organization of gene networks.

Conclusion

In our work, we performed a computational reconstruction of associative gene networks for rheumatoid arthritis, as well as AGNs describing the interactions of synthetic and targeted anti-inflammatory drugs with human genes and proteins. Based on the analysis of these gene networks, a search for biological processes as new promising pharmacological targets for RA therapy was conducted. The proposed approach can also be used to search for new targets for therapy of other diseases where standard treatment methods show insufficient therapeutic effect.

References

Adis Editorial. Tofacitinib. *Drugs R D*. 2010;10(4):271-284. doi 10.2165/11588080-000000000-00000

Alam M.J., Xie L., Ang C., Fahimi F., Willingham S.B., Kueh A.J., Herold M.J., Mackay C.R., Robert R. Therapeutic blockade of CXCR2 rapidly clears inflammation in arthritis and atopic dermatitis models: demonstration with surrogate and humanized antibodies. *mAbs*. 2020;12(1):1856460. doi 10.1080/19420862.2020.1856460

Cici D., Corrado A., Rotondo C., Cantatore F.P. Wnt signaling and biological therapy in rheumatoid arthritis and spondyloarthritis. *Int J Mol Sci*. 2019;20(22):5552. doi 10.3390/ijms20225552

Demenkov P.S., Oshchepkova E.A., Ivanisenko T.V., Ivanisenko V.A. Prioritization of biological processes based on the reconstruction and analysis of associative gene networks describing the response of plants to adverse environmental factors. *Vavilov J Genet Breed*. 2021;25(5):580-592. doi 10.18699/VJ21.065

Ding Q., Hu W., Wang R., Yang Q., Zhu M., Li M., Cai J., Rose P., Mao J., Zhu Y.Z. Signaling pathways in rheumatoid arthritis: implications for targeted therapy. *Signal Transduct Target Ther*. 2023;8(1):68. doi 10.1038/s41392-023-01331-9

Figus F.A., Piga M., Azzolin I., McConnell R., Iagnocco A. Rheumatoid arthritis: extra-articular manifestations and comorbidities. *Autoimmun Rev*. 2021;20(4):102776. doi 10.1016/j.autrev.2021.102776

Firestein G.S., McInnes I.B. Immunopathogenesis of rheumatoid arthritis. *Immunity*. 2017;46(2):183-196. doi 10.1016/j.immuni.2017.02.006

GBD 2021 Rheumatoid Arthritis Collaborators. Global, regional, and national burden of rheumatoid arthritis, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol*. 2023;5(10):e594-e610. doi 10.1016/S2665-9913(23)00211-4

Gong L., Thorn C.F., Bertagnolli M.M., Grosser T., Altman R.B., Klein T.E. Celecoxib pathways: pharmacokinetics and pharmacodynamics. *Pharmacogenet Genomics*. 2012;22(4):310-318. doi 10.1097/FPC.0b013e32834f94cb

Guo Q., Wang Y., Xu D., Nossent J., Pavlos N.J., Xu J. Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Res*. 2018;6:15. doi 10.1038/s41413-018-0016-9

Hill J., Harrison J., Christian D., Reed J., Clegg A., Duffield S.J., Goodson N., Marson T. The prevalence of comorbidity in rheumatoid arthritis: a systematic review and meta-analysis. *Br J Community Nurs*. 2022;27(5):232-241. doi 10.12968/bjcn.2022.27.5.232

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation. *Sci Rep*. 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci*. 2024;25(21):11847. doi 10.3390/ijms252111847

Kawahara K., Hohjoh H., Inazumi T., Tsuchiya S., Sugimoto Y. Prostaglandin E_2-induced inflammation: relevance of prostaglandin E receptors. *Biochim Biophys Acta*. 2015;1851(4):414-421. doi 10.1016/j.bbalip.2014.07.008

Kawasaki T., Kawai T. Toll-like receptor signaling pathways. *Front Immunol*. 2014;5:461. doi 10.3389/fimmu.2014.00461

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2013;17(4/2):833-850 (in Russian)

Kvien T.K., Uhlig T., Ødegård S., Heiberg M.S. Epidemiological aspects of rheumatoid arthritis: the sex ratio. *Ann NY Acad Sci*. 2006;1069:212-222. doi 10.1196/annals.1351.019

Llorente I., García-Castañeda N., Valero C., González-Álvaro I., Castañeda S. Osteoporosis in rheumatoid arthritis: dangerous liaisons. *Front Med* (*Lausanne*). 2020;7:601618. doi 10.3389/fmed.2020.601618

Miao C.G., Yang Y.Y., He X., Li X.F., Huang C., Huang Y., Zhang L., Lv X.W., Jin Y., Li J. Wnt signaling pathway in rheumatoid arthritis. *Cell Signal*. 2013;25(10):2069-2078. doi 10.1016/j.cellsig.2013.04.002

Mohd Jaya F.N., Garcia S.G., Borràs F.E., Chan G.C.F., Franquesa M. Paradoxical role of Breg-inducing cytokines in autoimmune diseases. *J Transl Autoimmun*. 2019;2:100011. doi 10.1016/j.jtauto.2019.100011

Nasonov E.L., Lila A.M., Karateev D.E., Mazurov V.I. et al. Clinical Recommendations. Rheumatoid Arthritis. All-Russian Public Organization "Association of Rheumatologists of Russia", 2024. KR250 (in Russian)

Olivera P.A., Lasa J.S., Bonovas S., Danese S., Peyrin-Biroulet L. Safety of Janus kinase inhibitors in patients with inflammatory bowel diseases or other immune-mediated diseases: a systematic review and meta-analysis. *Gastroenterology*. 2020;158(6):1554-1573. doi 10.1053/j.gastro.2020.01.001

Palmroth M., Kuuliala K., Peltomaa R., Virtanen A., Kuuliala A., Kurttila A., Kinnunen A., Leirisalo-Repo M., Silvennoinen O., Isomäki P. Tofacitinib suppresses several JAK-STAT pathways in rheumatoid arthritis *in vivo* and baseline signaling profile associates with treatment response. *Front Immunol*. 2021;12:738481. doi 10.3389/fimmu.2021.738481

Park J.Y., Pillinger M.H., Abramson S.B. Prostaglandin E_2 synthesis and secretion: the role of PGE_2 synthases. *Clin Immunol*. 2006;119(3):229-240. doi 10.1016/j.clim.2006.01.016

Prajapati P., Doshi G. An update on the emerging role of Wnt/β-catenin, SYK, PI3K/AKT, and GM-CSF signaling pathways in rheumatoid arthritis. *Curr Drug Targets*. 2023;24(17):1298-1316. doi 10.2174/0113894501276093231206064243

Rabelo F.S., da Mota L.M., Lima R.A., Lima F.A., Barra G.B., de Carvalho J.F., Amato A.A. The Wnt signaling pathway and rheumatoid arthritis. *Autoimmun Rev*. 2010;9(4):207-210. doi 10.1016/j.autrev.2009.08.003

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Singh A.K., Haque M., Madarampalli B., Shi Y., Wildman B.J., Basit A., Khuder S.A., Prasad B., Hassan Q., Ouseph M.M., Ahmed S. Ets-2 propagates IL-6 trans-signaling mediated osteoclast-like changes in human rheumatoid arthritis synovial fibroblast. *Front Immunol*. 2021;12:746503. doi 10.3389/fimmu.2021.746503

Smolen J.S., Aletaha D., McInnes I.B. Rheumatoid arthritis. *Lancet*. 2016;388(10055):2023-2038. doi 10.1016/S0140-6736(16)30173-8

Smolen J.S., Landewé R.B.M., Bergstra S.A., Kerschbaumer A., Sepriano A., Aletaha D., Caporali R., ... van der Helm-van Mil A., van Duuren E., Vliet Vlieland T.P.M., Westhovens R., van der Heijde D. EULAR recommendations for the management of rheumatoid arthritis: 2022 update. *Ann Rheum Dis*. 2023;82(1):3-18. doi 10.1136/ard-2022-223356

Stump K.L., Lu L.D., Dobrzanski P., Serdikoff C., Gingrich D.E., Dugan B.J., Angeles T.S., Albom M.S., Ator M.A., Dorsey B.D., Ruggeri B.A., Seavey M.M. A highly selective, orally active inhibitor of Janus kinase 2, CEP-33779. *Arthritis Res Ther*. 2011;13(2):R68. doi 10.1186/ar3329

Unterberger S., Davies K.A., Rambhatla S.B., Sacre S. Contribution of toll-like receptors and the NLRP3 inflammasome in rheumatoid arthritis pathophysiology. *Immunotargets Ther*. 2021;10:285-298. doi 10.2147/ITT.S288547

van der Kooij S.M., de Vries-Bouwstra J.K., Goekoop-Ruiterman Y.P., van Zeben D., Kerstens P.J., Gerards A.H., van Groenendael J.H., Hazes J.M., Breedveld F.C., Allaart C.F., Dijkmans B.A. Limited efficacy of conventional DMARDs after initial methotrexate failure. *Ann Rheum Dis*. 2007;66(10):1356-1362. doi 10.1136/ard.2006.066662

Wang R., Li M., Wu W., Qiu Y., Hu W., Li Z., Wang Z., Yu Y., Liao J., Sun W., Mao J., Zhu Y.Z. NAV2 positively modulates inflammatory response through Wnt/β-catenin signaling in rheumatoid arthritis. *Clin Transl Med*. 2021;11(4):e376. doi 10.1002/ctm2.376

Wang S.Y., Liu Y.Y., Ye H., Guo J.P., Li R., Liu X., Li Z.G. Circulating Dickkopf-1 is correlated with bone erosion and inflammation in rheumatoid arthritis. *J Rheumatol*. 2011;38(5):821-827. doi 10.3899/jrheum.100089

Yao C., Narumiya S. Prostaglandin-cytokine crosstalk in chronic inflammation. *Br J Pharmacol*. 2019;176(3):337-354. doi 10.1111/bph.14530

Conflict of interest. The authors declare no conflict of interest.

Received August 3, 2025. Revised October 1, 2025. Accepted October 17, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1031-1040

doi 10.18699/vjgb-25-108

Mathematical models of iron metabolism: structure and functions

N.I. Melchenko [1], I.R. Akberdin [1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Research Center for Genetics and Life Sciences, Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia
akberdin@bionet.nsc.ru

Abstract. Mathematical models represent a powerful theoretical tool for studying complex biological systems. They provide an opportunity to track non-obvious interactions and conduct *in silico* experiments to address practical problems. Iron plays a key role in oxygen transport in the mammals. However, a high concentration of this microelement can damage cellular structures through the production of reactive oxygen species and can also lead to ferroptosis (programmed cell death associated with iron-dependent lipid peroxidation). The immune system contributes greatly to the regulation of iron metabolism – hypoferritinemia (decreased ferritin concentration in the blood) during infection – which is a result of the innate immune response. In the study of iron metabolism, many aspects of regulation remain insufficiently studied and require a deeper understanding of the structural-functional organization and dynamics of all components of this complex process in both normal and pathological conditions. Consequently, mathematical modeling becomes an important tool to identify key regulatory interactions and predict the behavior of the iron metabolism regulatory system in the human body under various conditions. This article presents a review of iron metabolism models applicable to humans presented in chronological order of their development to illustrate the evolution and priorities in modeling iron metabolism. We focused on the formulation of numerical problems in the analyzed models, their structure and reproducibility, thereby highlighting their advantages and drawbacks. Advanced models can numerically simulate various experimental scenarios: blood transfusion, signaling pathway disruption, mutation in the ferroportin gene, and chronic inflammation. However, existing mathematical models of iron metabolism are difficult to scale and do not account for the functioning of other organs and systems, which severely limits their applicability. Therefore, to enhance the utility of computational models in solving practical problems related to iron metabolism in the human body, it is necessary to develop a scalable and verifiable mathematical model of iron metabolism that considers interactions with other functional human systems (e. g., the immune system) and state-of-the-art standards for representing mathematical models of biological systems.

Key words: mathematical modeling; iron metabolism; ferritin; hepcidin; ordinary differential equations

For citation: Melchenko N.I., Akberdin I.R. Mathematical models of iron metabolism: structure and functions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1031-1040. doi 10.18699/vjgb-25-108

Funding. This study was conducted with the support of a state project FWNR-2022-0020 at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences.

Математические модели метаболизма железа: структура и функции

Н.И. Мельченко [1], И.Р. Акбердин [1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Научный центр генетики и наук о жизни, Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия
akberdin@bionet.nsc.ru

Аннотация. Математические модели представляют собой мощный теоретический инструмент для изучения сложных биологических систем. Они позволяют прослеживать неочевидные взаимодействия и проводить виртуальные эксперименты для решения практических задач. Железо играет ключевую роль в транспорте кислорода в организме млекопитающих. В то же время высокая концентрация этого микроэлемента может повреждать клеточные структуры за счет продукции активных форм кислорода, а также привести к ферроптозу (программируемая клеточная гибель в связи с железо-зависимым перекисным окислением липидов). Большой вклад в регуляцию метаболизма железа вносит иммунная система: гипоферритинемия (снижение концентрации ферритина в крови) на фоне инфекции является результатом врожденного ответа иммунной системы. В исследовании метаболизма железа многие аспекты регуляции остаются недостаточно изученными; требуется более глубокое понимание структурно-функциональной организации и динамики всех компонентов этого комплексного процесса

в норме и при патологии. Важным инструментом, позволяющим выявить наиболее существенные регуляторные взаимодействия и предсказать поведение метаболической системы регуляции железа в организме человека в разных условиях, становится математическое моделирование. Данная работа представляет обзор моделей метаболизма железа, применимых к человеку, в порядке их создания, что позволяет оценить историю развития и приоритеты в моделировании метаболизма железа. Мы акцентировали внимание на постановке численных задач в анализируемых моделях, их структуре и воспроизводимости, на основе чего выделили их недостатки и преимущества. Современные модели способны численно воспроизвести множество экспериментов: гемотрансфузию, нарушение сигнального пути; мутацию в гене ферропортина; хроническое воспаление. Однако существующие математические модели метаболизма железа сложно масштабировать, и они не учитывают работу других органов и систем, в связи с чем их применение остается крайне ограниченным. Для расширения применимости компьютерных моделей в решении практических задач, связанных с метаболизмом железа в организме человека, необходимо создать масштабируемую и верифицируемую математическую модель метаболизма железа с учетом взаимодействия с другими функциональными системами человека (например, иммунной) и современных стандартов представления математических моделей биологических систем.
Ключевые слова: математическое моделирование; метаболизм железа; ферритин; гепсидин; обыкновенные дифференциальные уравнения

Introduction

Iron plays a key role in oxygen transport in vertebrate organisms (Pantopoulos et al., 2012). In the human body, iron exists in multiple forms (Vogt et al., 2021). In blood plasma, iron is transported both in a free, transferrin-unbound form and in a transferrin-bound form, as part of hemoglobin. Iron is predominantly found in tissues either in a free form or bound to the iron storage protein ferritin. However, the majority of iron in the body is present in erythrocytes as hemoglobin.

Both iron excess and deficiency lead to adverse consequences. Iron deficiency results in iron-deficiency anemia, while iron overload causes toxic effects of free iron and triggers programmed cell death mediated by iron – ferroptosis (Xie et al., 2016). Therefore, vertebrates have a molecular genetic system orchestrating iron homeostasis. The main protein regulating iron metabolism is hepcidin. It binds to ferroportin (FPN), a protein that functions as the sole iron exporter in vertebrates. Hepcidin binding leads to ubiquitination, internalization, and degradation of FPN, thereby inhibiting iron export. Since FPN is highly expressed in duodenal enterocytes, iron-recycling macrophages, and hepatocytes, hepcidin-mediated inactivation and degradation of FPN reduce dietary iron absorption and limit the release of stored iron, thus lowering circulating iron levels (Xu et al., 2021). Hepcidin expression, in turn, is controlled by negative feedback from iron concentrations both in plasma and hepatocytes, as well as by the inflammatory response, predominantly mediated by IL-6 activity (Nemeth, Ganz, 2023).

Currently, many aspects of iron metabolism remain incompletely understood – for example, non-heme iron transport into enterocytes, allosteric regulation of hemoglobin, and hepcidin regulation (Ahmed et al., 2020; Nemeth, Ganz, 2023). Since experimental approaches cannot thoroughly uncover the complexity and hierarchical organization of the system of interacting components regulating iron metabolism in the human body, the reconstruction of a comprehensive model of iron metabolism that accounts for molecular interactions between various organs and systems will not only integrate these organizational levels of the molecular genetic iron metabolism system within a unified conceptual framework but also serve as a theoretical basis for *in silico* studies aimed at investigating the structural-functional organization and dynamics of interactions among system components. This, in turn, will provide a foundation for the development and evaluation of drug efficacy targeting various therapeutic sites within the iron metabolism system, considering functional interactions with the immune system.

Herein, we review existing models, assessing their advantages and disadvantages as well as their applicability in addressing fundamental and applied aspects of iron metabolism research.

Initial models of iron metabolism in the human body

Mathematical model of iron metabolism (Franzone et al., 1982)

The model developed by P.C. Franzone and colleagues was designed to numerically estimate the concentration of iron in various compartments of the body, as well as to study the effects of different treatment methods on patients with anemia of various origins. The metabolic processes in the model are distributed across the following compartments: intestinal mucosa, blood plasma, liver, reticuloendothelial cells, bone marrow, and erythrocytes. The model describes the intake of iron from food, its transport into plasma, storage in the liver, and participation in erythropoiesis. It takes into account the impact of erythropoietin on the proliferation and maturation of erythroid cells. The model also allows for the consideration of iron replenishment through donor blood and iron loss due to bleeding. To account for the process of iron return from erythrocytes to blood plasma, the model includes a component describing the destruction of erythrocytes by reticuloendothelial cells. Additionally, the model considers ineffective hematopoiesis, whereby some erythroid cells fail to complete differentiation (Fig. 1).

The model simulations were conducted on conditions such as blood donation in a healthy patient, blood transfusion after splenectomy in a patient with hemolytic anemia, as well as treatment of hypoplastic anemia using transfusions and androgens.

In the numerical experiment describing blood donation in healthy patients, the model shows complete recovery of hemoglobin levels in approximately 25–30 days. In turn, complete restoration of iron levels in the bone marrow takes 60 days, while recovery of iron levels in the storage pool requires more

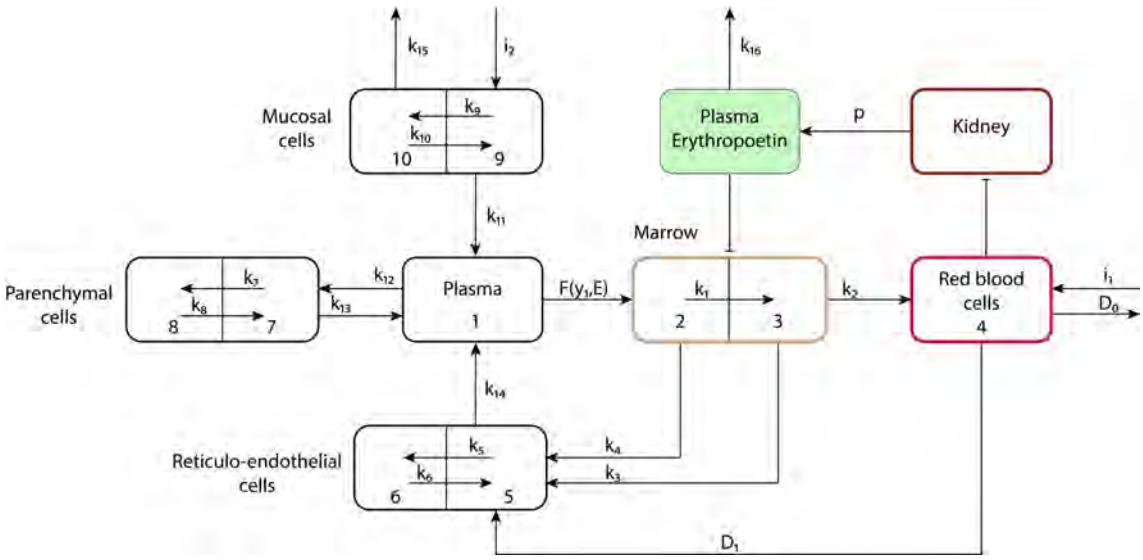


Fig. 1. Schematic representation (adapted from Franzone et al., 1982).

In the figure, the blocks represent the amount of iron in a specific organ or system, where 1 – blood plasma, 2 – maturing erythroid blood cells, 3 – mature erythroid blood cells, 4 – erythrocytes, 5 – macrophages, 6 – iron storage in macrophages, 7 – extravascular fluid, 8 – iron storage in hepatocytes, 9 – intestinal epithelial cells, 10 – iron storage in intestinal epithelial cells. The arrows indicate iron transport between organs and systems, where k_1, k_2, k_3… k_{16} are the rates of iron transport between the blocks, i_1 – iron influx due to blood donation, i_2 – iron influx from food, D_0 – iron loss due to bleeding, D_1 – transfer of iron to reticuloendothelial system cells as a result of phagocytosis, $F(y_1, E)$ – function describing the transfer of iron from plasma to erythroid cells, where y_1 is the amount of iron in the blood plasma, and E is the amount of erythropoietin, p – function of erythropoietin synthesis.

than two months, which corresponded to the literature data at the time of publication (Wadsworth, 1955; Liedén et al., 1975) and also aligns with data from recent studies (Kiss et al., 2015; Ziegler et al., 2015).

The model was also used to numerically investigate blood transfusion after splenectomy (removal of the spleen). The resulting model more accurately describes iron dynamics for patients after splenectomy. However, data from only one patient were used to validate this condition.

The proposed model was also used to study the effect of treating hypoplastic anemia with transfusions and androgens. However, these results have lost their relevance since such therapy is no longer used today (Killick et al., 2016). The authors of the developed model note that the system's equations can exhibit stiff behavior due to the large differences between the numerical values of transport rates when modeling anemic conditions. Considering the stiffness of the system, to achieve a compromise between accuracy and computational resources, the authors used the implicit trapezoidal method for the numerical solution of the system (Tavernini, 1973).

Given that Franzone and co-authors' model is one of the first models describing iron metabolism, it is significantly inferior to modern models. This model lacks descriptions of key participants in iron metabolism: hepcidin, ferritin, transferrin, and proteins regulating the expression of genes involved in iron metabolism (Iron Regulatory Proteins, IRP). The iron storage process is greatly simplified and represented by a linear coefficient. Despite this, the authors managed to simulate complex conditions such as blood transfusion after splenectomy in a patient with hemolytic anemia and treatment of hypoplastic anemia using transfusions and androgens. However, considering that data from only one patient was used to validate the numerical calculations of the model for each of these conditions, it is difficult to assess how applicable the numerical modeling results are to population data and how parameters might change when reproducing data on other patients.

Computational model of iron metabolism in the liver (Mitchell, Mendes, 2013)

The mathematical model proposed by S. Mitchell and P. Mendes in 2013 allows the numerical evaluation of processes related to iron transport into hepatocytes. The model enables quantitative prediction of the concentration of proteins synthesized in the liver that regulate iron metabolism. The model consists of 21 ordinary differential equations and includes two compartments: hepatocyte and plasma (Fig. 2).

Using the model built, the authors numerically analyzed the following physiological conditions: hereditary hemochromatosis types 1 and 3. To reproduce the state of type 1 hemochromatosis, a virtual knockdown of the human iron homeostasis regulator protein (HFE) was performed by reducing the synthesis constant 100-fold. The model could not quantitatively reproduce the result that mice with this pathology have liver iron levels three times higher than normal. This was due to the fixed concentration of intercellular transferrin-bound iron in the model, unlike that in mice, which show increased transferrin saturation as a result of increased intestinal iron absorption. Despite fixed extracellular conditions, the model predicts intracellular iron overload in hepatocytes. The hemochromatosis model also reproduced the dynamics observed in experiments with changes in dietary iron content. Increased dietary iron doubled ferroportin expression in the liver in both healthy mice and those with hemochromatosis. To reproduce the state of type 3 hemochromatosis, a virtual knockdown of TfR2 was

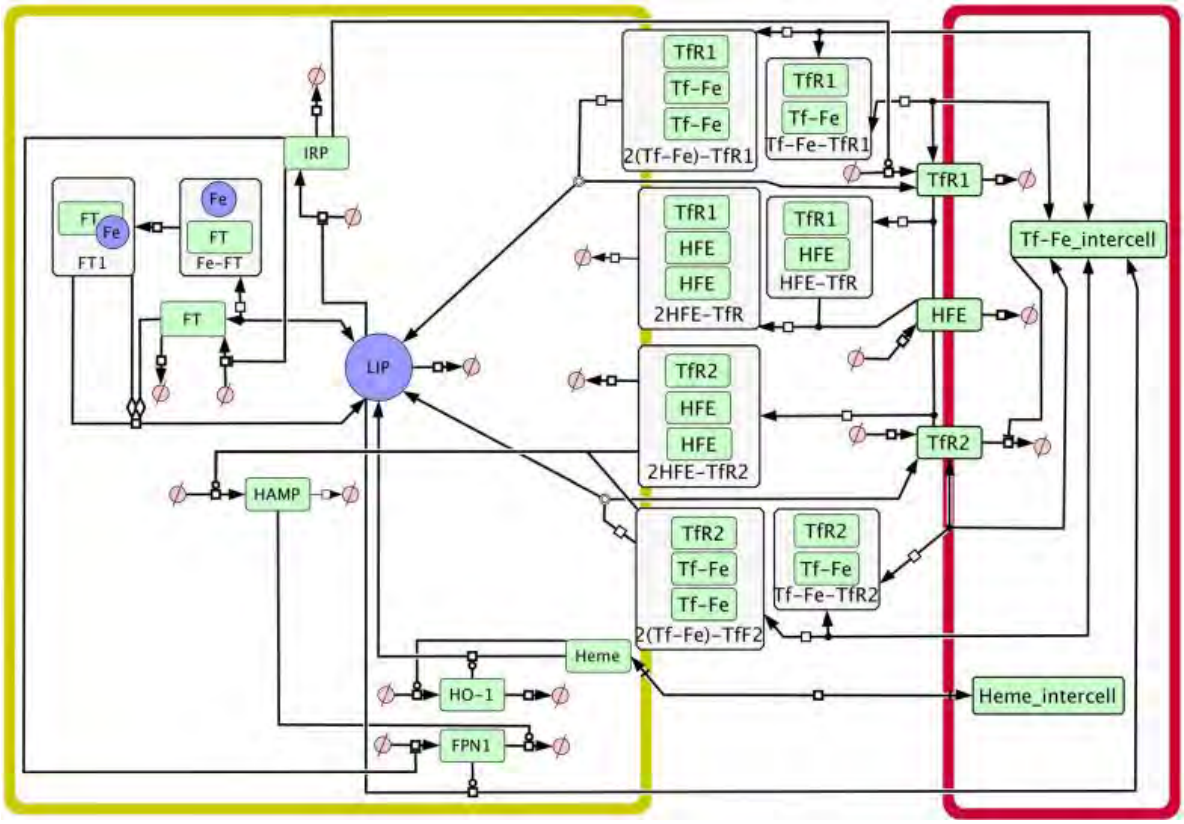


Fig. 2. Graphical representation of the model in the SBGN standard (Le Novère et al., 2009).

Arrows designate substance transport. Yellow compartment – hepatocyte, red compartment – plasma, LIP – labile iron pool, FT – ferritin, Fe – iron, HAMP – hepcidin, Heme – heme, HO-1 – heme oxygenase 1, IRP – iron regulatory proteins, FPN1 – ferroportin, TfR1 – transferrin receptor 1, TfR2 – transferrin receptor 2, Tf-Fe_intercell – plasma transferrin-bound iron (Mitchell, Mendes, 2013).

performed, also by reducing the synthesis constant 100-fold. Numerical analysis revealed an increase in hepcidin concentration and a decrease in ferroportin concentration, which was consistent with experimental data (Chua et al., 2010).

The model describes the iron transport into hepatocytes well, considering iron storage, export, and utilization for heme synthesis. We also comprehensively reproduced the authors' results both in the COPASI software (Hoops et al., 2006) and in the BioUML platform (Kolpakov et al., 2022). However, the model has some limitations: (1) the model lacks an important regulatory link in iron metabolism, namely the effect of hepcidin on iron absorption from the intestine; (2) fixed concentrations of heme and intercellular transferrin-bound iron are used; (3) due to limited availability of quantitative clinical data on human iron metabolism, various other data sources were integrated for parameterization, such as *in vitro* experiments and animal models; (4) the parameters reported in the study do not correspond to the model parameters in the supplementary material.

Modeling of the system iron regulation in various pathologies considering hepcidin-independent mechanisms (Enculescu et al., 2017)

The model by M. Enculescu and colleagues (2017) describes iron metabolism throughout the human body, taking into account intra- and extracellular regulatory mechanisms of iron metabolism. The authors focused primarily on the system regulation of iron metabolism via the hepcidin-ferroportin regulatory axis. The model describes iron content in seven compartments: serum, liver, spleen, bone marrow, erythrocytes, duodenum, and "other organs," representing iron distribution in the mouse body. Iron absorption and loss in the duodenum, as well as iron loss in the "other organs" compartment, are considered. The model explains inhibition of ferroportin transcription during inflammation and regulation of its translation by intracellular iron, as well as hepcidin-mediated post-translational destabilization of ferroportin. Iron export from peripheral organs is controlled by the iron exporter ferroportin (Fpn), which is predominantly localized on the plasma membrane of three cell types: duodenal enterocytes, macrophages, and hepatocytes. Fpn expression is described separately for each organ and regulated by three mechanisms: (1) inflammatory signals decrease Fpn mRNA transcription; (2) intracellular iron enhances Fpn mRNA translation; (3) Fpn protein turnover is increased by the soluble polypeptide hepcidin.

Hepcidin expression is activated by the iron-sensitive BMP6/SMAD pathway and an inflammatory signaling cascade involving cytokine production (primarily IL-6) and subsequent phosphorylation of the transcription factor STAT3 in hepatocytes (Fig. 3).

The authors' own data and previously published data were used for the model calibration. A total of 344 experimental measurements were obtained. The following assumptions were

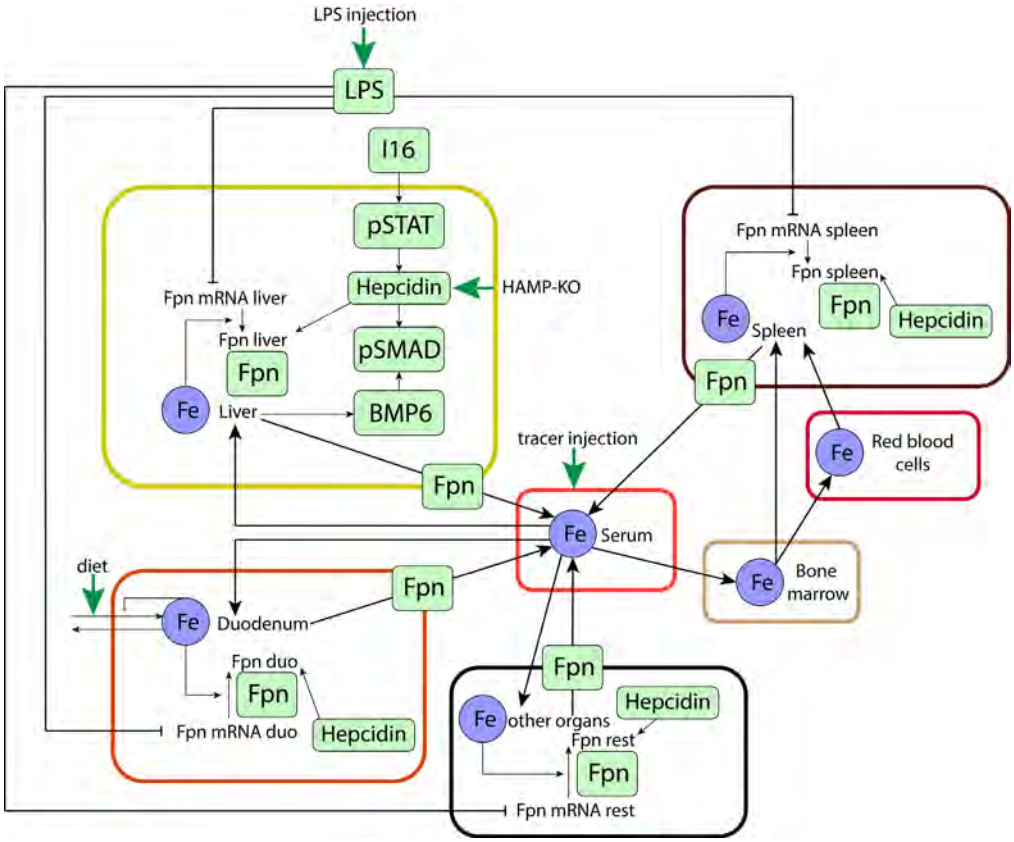


Fig. 3. Graphical representation of the model (Enculescu et al., 2017) in the SBGN standard (Le Novère et al., 2009).
LPS – lipopolysaccharides, Fpn – ferroportin, BMP6 – bone morphogenetic protein (regulatory protein), pSMAD, pSTAT – transcription factors. Black arrows indicate substance transport, green arrows designate substance input from outside the organism.

made for model parameterization: in some cases, homologous reactions in different compartments proceed with identical kinetic rate constants. Additionally, kinetic parameters of the hepcidin gene promoter model were fixed at values previously determined by the authors in the HuH7 cell culture system.

The following conditions were numerically investigated using the constructed model: administration of lipopolysaccharides (LPS) under iron overload; disruption of the BMP6 signaling pathway; mutation in the ferroportin gene leading to loss of ferroportin's ability to bind hepcidin; chronic inflammation.

The authors also used data from their own experiment to validate the model in the numerical analysis of LPS administration under iron overload. According to the experiment, male C57BL/6 mice were fed an iron-rich diet containing 100 times more iron than a normal diet for four weeks, followed by a single dose of LPS at 1 μg/kg. The experimental data corresponded to the model's predictions for most variables: iron in serum, liver, and duodenum; hepcidin content in the liver; BMP6 mRNA concentration; levels of pSTAT and pSMAD in the liver; mRNA and protein content of ferroportin in the liver. Deviations of the model approximation from experimental data were observed in the following indicators: iron content in the spleen and erythrocytes, ferroportin concentration in the spleen.

This study also provides a numerical analysis of the dynamic behavior of the iron regulation system when hepcidin feedback is blocked. Two situations were reproduced for this: (1) disruption of the BMP6 signaling pathway; (2) mutation in the ferroportin gene leading to the loss of ferroportin's ability to bind hepcidin.

To reproduce the first condition, SMAD expression was set to zero, whereas to reproduce the second condition, the parameter values describing hepcidin's effect on ferroportin degradation were also set to zero. Numerical simulations of the model in both cases showed an increase in iron concentration in the serum and liver and a decrease in iron concentration in the spleen, which was confirmed by experimental data. Moreover, as in the experiments, ferroportin resistance to hepcidin led to increased hepcidin expression, whereas the loss of SMAD signal transduction caused a significant decrease in hepcidin expression.

Then the authors hypothesized that hepcidin affects ferroportin in a tissue-specific manner. To model this situation, the authors sequentially set to zero the parameter values describing hepcidin's effect on ferroportin degradation in different tissues. The results of the numerical analysis demonstrated that only the elimination of hepcidin-mediated regulation of ferroportin in the duodenum has a system effect, leading to an increase in iron concentration in other organs. Meanwhile, modeling ferroportin resistant to hepcidin in the liver or spleen leads only to a local effect with a decrease in iron stores in the corresponding organ and minimal changes in other organs. Mouse models with tissue-specific resistance to hepcidin have

not yet been described. However, tissue-specific deletion of FPN in intestinal cells has been studied in mice. This study showed that deletion of FPN in intestinal cells leads to severe iron deficiency in blood, liver, and spleen.

The research team of the proposed model also applied it to conduct an *in silico* experiment studying chronic inflammation. Equations describing the kinetics of LPS and their effect on hepcidin were added to model the scenario. Numerical analysis of the model describing persistent inflammation showed an 85 % decrease in serum iron concentration; iron concentration in erythrocytes decreased over a longer period, stabilizing after two months at a value equal to 10 % of the normal level.

This investigation considers two mechanisms of ferroportin regulation: at the transcript level and regulation by hepcidin. To assess the contribution of each regulatory path, the authors modeled LPS responses when either the transcriptional or post-translational effect of LPS on ferroportin protein levels was eliminated. Numerical analysis indicated that the absence of hepcidin influence during inflammation resulted in a normal decrease in serum iron level (75 % of the original model version). In contrast, removal of transcriptional control of ferroportin during inflammation reduced hypoferriemia to 50 %. The authors concluded that removal of transcriptional control of ferroportin causes greater deviations in serum iron values from normal than removal of hepcidin control. This concludes that hypoferriemia arises as a result of a combination of hepcidin-dependent and independent mechanisms.

Among the limitations of the proposed model, the authors note varying degrees of parameter accuracy and the absence of description of iron binding to ferritin and its storage.

Erythropoiesis and iron metabolism model in humans (Schirm, Scholz, 2020)

A group of authors from the University of Leipzig developed a mathematical model (Schirm, Scholz, 2020) aimed at predicting the effects of treatments involving unproven therapeutic options, such as cytotoxic chemotherapy supported by iron and erythropoietin (EPO). The model is an extension of the authors' previous study on erythropoiesis modeling (Schirm et al., 2013), which was expanded by adding an extra module for iron metabolism. The original erythropoiesis module describes the dynamics of erythropoietic cell development, reflecting all the main stages of differentiation: stem cells, burst-forming units, colony-forming units, proliferating erythroblasts, maturing erythroblasts, and reticulocytes. This module also accounts for the effects of chemotherapy on erythropoiesis. The module describing iron metabolism includes the following compartments: hepcidin, non-transferrin-bound iron (NTBI) in plasma, the hemoglobin catabolic system, iron stores, transferrin bound to iron, and free transferrin (Fig. 4).

Within the framework of computational modeling, some simplifications of the complex physiological system were employed to reduce the number of unknown model parameters or due to the lack of quantitative data for humans. The model does not consider separate pools of Fe^{2+} and Fe^{3+} concentrations due to the absence of data, nor does it specify concentrations of transferrin saturated with one or two iron ions.

The following conditions were studied via the numerical analysis of the proposed model: (1) oral iron administration in healthy individuals; (2) intravenous injection of EPO with oral iron administration in healthy individuals; (3) iron deficiency; (4) intravenous iron administration in healthy individuals; (5) bleeding/phlebotomy; (6) chronic inflammation; (7) hemochromatosis.

To validated the model's numerical calculations, the authors harnessed the data from several clinical studies with different treatment modes (Rutherford et al., 1994; Souillard et al., 1996; Kiss et al., 2015). The authors numerically investigate the experimental scenario of Souillard and colleagues (1996), in which healthy athletes received 200 IU/kg of EPO on days 0, 2, 4, 7, and 10 without iron supplementation. The obtained *in silico* results for the quantity or concentration of reticulocytes, hemoglobin, erythrocytes, hematocrit, and ferritin generally differ from the clinical study data by no more than one standard deviation.

To validate the numerical results describing EPO administration with iron supplements, the authors used the data from by Rutherford and coauthors' study (1994). In this clinical trial, patients received EPO at a dose of 1,200 IU/kg per week with different dosing regimens and iron at a dose of 300 mg orally daily for 10 days. The modeling results for hematocrit, reticulocyte, ferritin concentrations, and transferrin saturation reflect the dynamics of these parameters in the clinical study very well. However, the numerical results for hemoglobin are underestimated.

S. Schirm and M. Scholz also conducted a numerical experiment on the donation of 500 mL of blood, both with and without iron supplementation. To validate the numerical results, the authors employed the clinical study by Kiss et al. (2015), which provided quantitative measurements of ferritin and hemoglobin dynamics. The numerical results for ferritin concentration calculated by the model differ from the clinical data by no more than one standard deviation in both scenarios, while the numerical results for hemoglobin dynamics in the iron supplementation scenario differ from the clinical data by more than one standard deviation over a large interval.

This study also included a virtual experiment aimed at a theoretical prediction for unused therapy. The Scholz group modeled the effect of CHOP-14 therapy supported by iron supplements and EPO on erythropoiesis and iron metabolism. CHOP-14 is a commonly accepted therapy for treating aggressive non-Hodgkin lymphomas, including drugs such as doxorubicin, cyclophosphamide, vincristine, and prednisolone. Currently, the therapy has been extended to R-CHOP, which also includes rituximab (Phan et al., 2010). This therapy is hematotoxic, so the authors considered the possibility of supplementing it with iron and EPO. To validate the numerical results in the *in silico* experiment of chemotherapy without iron and EPO supplementation, the data from a German research group on high-grade non-Hodgkin lymphoma (Pfreundschuh et al., 2004) were used. According to the numerical results of the *in silico* experiment, adding iron supplements together with EPO in patients undergoing CHOP-14 therapy slowed the decline in hemoglobin concentration. When iron supplements and EPO are administered on days 3, 7, and 21, the hemoglobin concentration on day 80 is approximately 11.2 g/dL, whereas without supportive therapy it is about 10.7 g/dL. With weekly administration of iron supplements

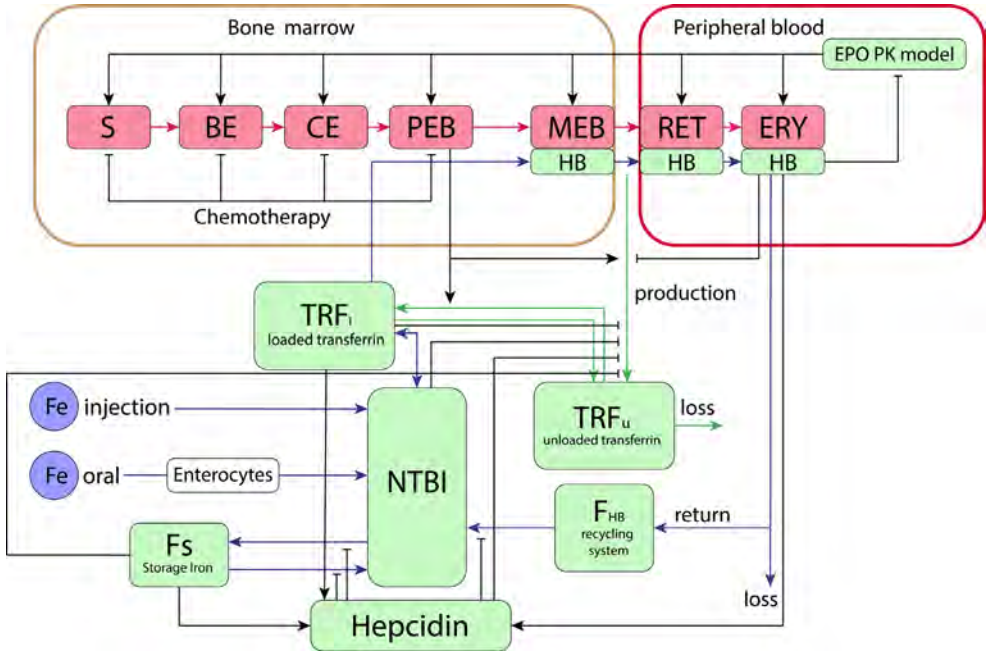


Fig. 4. Graphical representation of the model (Schirm, Scholz, 2020) in the SBGN standard (Le Novère et al., 2009).

S – stem cells, BE – burst-forming unit, CE – colony-forming unit, PEB – proliferating erythroblasts, MEB – maturing erythroblasts, RET – reticulocytes, ERY – erythrocytes, HB – hemoglobin, NTBI – non-transferrin-bound iron. Blue arrows indicate iron flow, green arrows represent transferrin flow, red arrows show the differentiation progression of erythroid lineage cells, and black arrows denote regulatory influences.

together with EPO starting from day 45, hemoglobin concentration recovers to 12.5 g/dL by day 80, while without supportive therapy hemoglobin concentration falls to 10.7 g/dL. It is important to note that EPO plays a significant role in hemoglobin recovery, as numerical results for supportive therapy with iron supplements alone practically did not differ from those without it.

The authors adhered to a modular approach and built the model upon their previous study by adding new components. A major advantage of this study is the validation using a large amount of data from various studies. The model demonstrated good agreement with clinical trials, as in most cases the differences between the model's numerical data and clinical results did not exceed one standard deviation. One drawback is the lower hemoglobin level predicted by the model compared to experimental measurements.

Model of iron sequestration by ferritin (Masison, Mendes, 2023)

P. Mendes and J. Masison developed a model describing the binding of iron ions by the protein ferritin. Ferritin consists of 24 subunits and is capable of binding about 4,300 iron atoms per ferritin molecule. Ferritin is an important participant in iron metabolism, so iron exchange models must include it. Such a model enables integrating the interaction of ferritin with iron ions into more complex models.

The model considered: (1) how iron bound to ferritin affects the dynamics of iron sequestration; (2) how the iron sequestration model with rate constants obtained experimentally *in vitro* can numerically reproduce experimental results obtained in cell lines; (3) the influence of ferritin subunit composition on the rate of iron sequestration; (4) the dependence of iron release dynamics from ferritin on the concentration of free iron and ferritin in the cell.

The model accounted for four chemical species: LIP – labile iron pool, soluble or readily soluble divalent iron in the cytoplasm; DFP – peroxo complex containing two iron atoms; core – iron incorporated into the mineralized ferrihydrite core; FT – 24 subunits of ferritin. The model included four reactions, three of which describe the process of iron sequestration by ferritin: oxidation converts two LIP into one DFP; nucleation converts two DFP into a new crystal core; mineralization adds one DFP to an existing core; and one reaction describes degradation of the intermediate product: reduction converts one DFP back into two LIP. The sequestration process is shown schematically in Fig. 5. The authors simplify and combine several of its components to construct a system of differential equations that reflects this biochemical process with sufficient accuracy. At the same time, they avoid excessive details and do not overload the model with variables.

The first reaction describes the oxidation of LIP to DFP and is represented by a Hill function:

$$\frac{kcat \times \dfrac{H + rO}{24 + rO} \times FT \times LIP^n}{Km^n + LIP^n},$$

kcat – catalytic turnover number, Km – Michaelis constant, *n* – Hill coefficient. The ferritin molecule consists of 24 subunits of two different types, H and L, and only the H subunits contain the active ferroxidase center. Therefore, molecules with different subunit compositions have different oxidation rates. To account for this, two additional parameters were

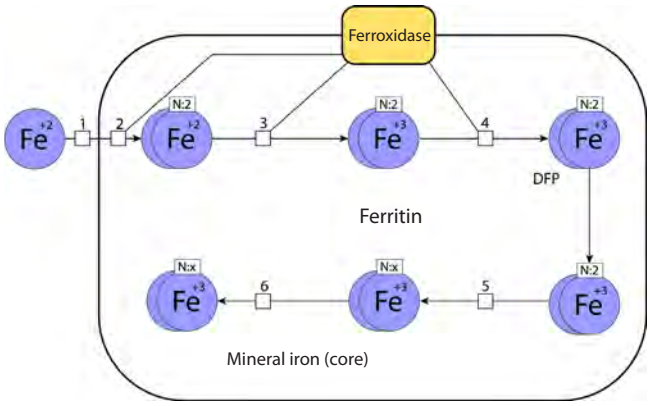


Fig. 5. Adapted graphical representation of the iron sequestration model (Masison, Mendes, 2023) in the SBGN standard (Le Novère et al., 2009).

The following reactions are shown: 1 – transport of Fe^{2+} into ferritin, 2 – binding of Fe^{2+} with ferroxidase, formation of DFP, 3, 4 – oxidation of Fe^{2+}, 5 – nucleation, 6 – mineralization.

used: H – the number of the H subunits (a value from 0 to 24); *rO* – a scaling factor representing the oxidation efficiency of the L homopolymer.

The parameter *rO* was included by the model authors because, despite the L subunits lacking a known ferroxidase, the L homopolymers still catalyze the formation of ferric iron (Fe^{3+}) within ferritin according to experimental data, although at a rate reduced by more than a quarter (Carmona et al., 2014). Since data on how oxidation occurs in the absence of the H subunit and the corresponding value of *rO* are limited, the value of *rO* was empirically set to two.

The second reaction, degradation of DFP, follows the law of mass action:

$$k_{deg} \times DFP.$$

The third reaction is nucleation:

$$kcat \times DFP^2 \times FT \times \frac{L + rN}{24 + rN} \times \frac{Ki^n}{Ki^n + core^n}.$$

It describes the process of forming a new crystal from two DFP molecules inside the ferritin molecule. New nuclei can also form within a ferritin molecule that already contains an existing core. The equation was empirically derived based on the law of mass action. The coefficients L and *rN* reflect how the ferritin subunit composition influences nucleation; due to limited information on this process, the coefficient *rN* was chosen to contribute significantly less to the nucleation rate range than the coefficient *rO* does to the oxidation rate range. The last factor included an inhibition constant and a Hill coefficient, allowing for the decrease in the probability of new crystal formation as the size of the existing core increases.

The fourth reaction is mineralization:

$$\frac{kcat \times DFP \times core}{Km + DFP} \times \frac{Ki^n}{Ki^n + core^n} \times \frac{4300^m - apc^m}{4300^m}.$$

The published data (Harrison et al., 1974) demonstrated that the rate of this reaction reaches a maximum at 1,500–2,000 iron atoms per core and decreases with further core growth. The second factor is needed to account for this process, while the third factor drives the rate to zero as the number of iron atoms per core (apc) approaches the maximum allowable value of 4,300.

To validate the model simulations, experimental data from different laboratories under various conditions were used. The model exhibited some differences compared to experimental data within the first 20 seconds: a stronger cooperative effect in the DFP mineralization rate and a faster attainment of steady-state concentration. Since the model's target context is cellular models, where the relevant time scale is minutes or longer, such differences from experimental data are not considered significant.

The authors conducted a virtual experiment investigating the influence of iron atoms in the core on the mineralization rate. The simulation revealed that the mineralization rate over time depends on the initial number of iron atoms per core (apc). Typically, the curves showing mineralization rate fall into three groups based on the initial apc. In the first group (<1,000 apc), the mineralization rate starts low, then increases as iron accumulates inside ferritin, and later decreases as the iron concentration in the solution drops. In the second group (1,000–3,000 apc), the mineralization rate starts high but rapidly declines due to decreasing iron concentration in the solution. Eventually, in the third group (>3,000 apc), the mineralization rate decreases throughout the simulation, as iron accumulation in the ferritin core slows down further mineralization.

Then the authors investigated the model behavior at ferritin and iron concentrations corresponding to those found in mammalian cells. The research team led by Mendes incorporated this model as a modular component into their previously developed model of iron metabolism in hepatocytes. The authors reported that the system's qualitative behavior remains similar to the original model before extension. However, the expanded model provided a deeper understanding and better assessment of iron storage mechanisms. Due to the increased detail of the new model, it becomes clear that the peak in ferritin-bound iron is driven by an increase in the concentration of DFP rather than the mineralized core – an important distinction since DFP is more readily released back into the cytoplasm. The numerical results of the models differed both over the time course and at equilibrium. The greatest differences appear after 1,000 seconds of simulation. In the original model, ferritin-bound iron content gradually increased, whereas in the new model, its concentration decreased. The authors of the original study hypothesized that this discrepancy may be related to new iron storage kinetics, which promotes a reduction in available iron through ferritin buffering, whereas in the original model, other mechanisms primarily influenced the kinetics of available iron.

Conclusion

The analysis of the presented mathematical models of iron metabolism reveals a tendency toward a progressive increase in their structural complexity over time (Supplementary Table S1)[1]. With the advancement of research, both the number of equations and the number of parameters in the models grow, indicating a pursuit of a more accurate and detailed description of biological processes. More recent models provide the simulation of a broader range of physiological and pathological states, expanding the possibilities for conduc-

[1] Supplementary Table S1 is available at:
https://vavilovj-icg.ru/download/pict-2025-29/appx38.pdf

ting *in silico* experiments. An exception is the latest model of iron sequestration by ferritin (Masison, Mendes, 2023), which is implemented according to a modular principle and was developed with the aim of integration into more complex systems. This approach ensures the flexibility and scalability of the model, which is important for further development and incorporation into multifactorial models of iron metabolism.

To deeper understand the iron metabolism, it is necessary to consider its interaction with the immune system, as it plays a key role in regulating iron homeostasis (Vogt et al., 2021). At the same time, the reduction of iron availability to pathogens and the production of reactive oxygen species can significantly affect the dynamics of infectious diseases (Weinberg, 2009). Inclusion of these factors in mathematical models will enable virtual experiments analyzing the impact of various infections on iron metabolism and assessing the long-term consequences of such interactions. This knowledge may be critically important for developing new approaches to treat diseases associated with iron metabolism disorders, as well as for understanding the pathogenesis of conditions such as anemia under chronic diseases, hemochromatosis, or post-viral syndromes, such as post-COVID syndrome.

Thus, integrating data on the interactions between the immune system and iron metabolism will not only deepen our understanding of these processes but may also pave the way for new opportunities for clinical research and therapeutic strategies. In this regard, the construction of a detailed model of iron metabolism that takes into account its interactions with the immune system represents a timely task, the solution of which will enable better understanding of the interplay between these two complex systems and allow the identification of key links in the pathology of iron metabolism in various diseases.

References

Ahmed M.H., Ghatge M.S., Safo M.K. Hemoglobin: structure, function and allostery. In: Hoeger U., Harris J. (Eds) Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins. Subcellular Biochemistry. Vol. 94. Springer, 2020;345-382. doi 10.1007/978-3-030-41769-7_14

Carmona U., Li L., Zhang L., Knez M. Ferritin light-chain subunits: key elements for the electron transfer across the protein cage. *Chem Commun*. 2014;50:15358-15361. doi 10.1039/C4CC07996E

Chua A.C.G., Delima R.D., Morgan E.H., Herbison C.E., Tirnitz-Parker J.E.E., Graham R.M., Fleming R.E., Britton R.S., Bacon B.R., Olynyk J.K., Trinder D. Iron uptake from plasma transferrin by a transferrin receptor 2 mutant mouse model of haemochromatosis. *J Hepatol*. 2010;52(3):425-431. doi 10.1016/j.jhep.2009.12.010

Enculescu M., Metzendorf C., Sparla R., Hahnel M., Bode J., Muckenthaler M.U., Legewie S. Modelling systemic iron regulation during dietary iron overload and acute inflammation: role of hepcidin-independent mechanisms. *PLoS Comput Biol*. 2017;13(1):e1005322. doi 10.1371/journal.pcbi.1005322

Franzone P.C., Paganuzzi A., Stefanelli M. A mathematical model of iron metabolism. *J Math Biol*. 1982;15(2):173-201. doi 10.1007/BF00275072

Harrison P.M., Hoy T.G., Macara I.G., Hoare R.J. Ferritin iron uptake and release. Structure–function relationships. *Biochem J*. 1974;143(2):445-451. doi 10.1042/bj1430445

Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N., Singhal M., Xu L., Mendes P., Kummer U. COPASI – a COmplex PAthway SImulator. *Bioinformatics*. 2006;22(24):3067-3074. doi 10.1093/bioinformatics/btl485

Killick S.B., Bown N., Cavenagh J., Dokal I., Foukaneli T., Hill A., Hillmen P., Ireland R., Kulasekararaj A., Mufti G., Snowden J.A., Samarasinghe S., Wood A., Marsh J.C.W. Guidelines for the diagnosis and management of adult aplastic anaemia. *Br J Haematol*. 2016;172(2):187-207. doi 10.1111/bjh.13853

Kiss J.E., Brambilla D., Glynn S.A., Mast A.E., Spencer B.R., Stone M., Kleinman S.H., Cable R.G.; National Heart, Lung, and Blood Institute (NHLBI) Recipient Epidemiology and Donor Evaluation Study-III (REDS-III). Oral iron supplementation after blood donation: a randomized clinical trial. *JAMA*. 2015;313(6):575-583. doi 10.1001/jama.2015.119

Kolpakov F., Akberdin I., Kiselev I., Kolmykov S., Kondrakhin Y., Kulyashov M., Kutumova E., Pintus S., Ryabova A., Sharipov R., Yevshin I., Zhatchenko S., Kel A. BioUML – towards a universal research platform. *Nucleic Acids Res*. 2022;50(W1):W124-W131. doi 10.1093/nar/gkac286

Le Novère N., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., … Sander C., Sauro H., Snoep J.L., Kohn K., Kitano H. The systems biology graphical notation. *Nat Biotechnol*. 2009;27(8):735-741. doi 10.1038/nbt.1558

Liedén G., Höglund S., Ehn L. Changes in certain iron metabolism variables after a single blood donation. *Acta Med Scand*. 1975;197(1-2):27-30. doi 10.1111/j.0954-6820.1975.tb04873.x

Masison J., Mendes P. Modeling the iron storage protein ferritin reveals how residual ferrihydrite iron determines initial ferritin iron sequestration kinetics. *PLoS One*. 2023;18(2):e0281401. doi 10.1371/journal.pone.0281401

Mitchell S., Mendes P. A computational model of liver iron metabolism. *PLoS Comput Biol*. 2013;9(11):e1003299. doi 10.1371/journal.pcbi.1003299

Nemeth E., Ganz T. Hepcidin and iron in health and disease. *Annu Rev Med*. 2023;74:261-277. doi 10.1146/annurev-med-043021-032816

Pantopoulos K., Porwal S.K., Tartakoff A., Devireddy L. Mechanisms of mammalian iron homeostasis. *Biochemistry*. 2012;51(29):5705-5724. doi 10.1021/bi300752r

Pfreundschuh M., Trümper L., Kloess M., Schmits R., Feller A.C., Rübe C., Rudolph C., Reiser M., Hossfeld D.K., Eimermacher H., Hasenclever D., Schmitz N., Loeffler M.; German High-Grade Non-Hodgkin's Lymphoma Study Group. Two-weekly or 3-weekly CHOP chemotherapy with or without etoposide for the treatment of elderly patients with aggressive lymphomas: results of the NHL-B2 trial of the DSHNHL. *Blood*. 2004;104(3):634-641. doi 10.1182/blood-2003-06-2095

Phan J., Mazloom A., Medeiros L.J., Zreik T.G., Wogan C., Shihadeh F., Rodriguez M.A., Fayad L., Fowler N., Reed V., Horace P., Dabaja B.S. Benefit of consolidative radiation therapy in patients with diffuse large B-cell lymphoma treated with R-CHOP chemotherapy. *J Clin Oncol*. 2010;28(27):4170-4176. doi 10.1200/JCO.2009.27.3441

Rutherford C.J., Schneider T.J., Dempsey H., Kirn D.H., Brugnara C., Goldberg M.A. Efficacy of different dosing regimens for recombinant human erythropoietin in a simulated perisurgical setting: the importance of iron availability in optimizing response. *Am J Med*. 1994;96(2):139-145. doi 10.1016/0002-9343(94)90134-1

Schirm S., Scholz M. A biomathematical model of human erythropoiesis and iron metabolism. *Sci Rep*. 2020;10(1):8602. doi 10.1038/s41598-020-65313-5

Schirm S., Engel C., Loeffler M., Scholz M. A biomathematical model of human erythropoiesis under erythropoietin and chemotherapy administration. *PLoS One*. 2013;8(6):e65630. doi 10.1371/journal.pone.0065630

Souillard A., Audran M., Bressolle F., Gareau R., Duvallet A., Chanal J.L. Pharmacokinetics and pharmacodynamics of recombinant human erythropoietin in athletes. Blood sampling and doping control. *Br J Clin Pharmacol*. 1996;42(3):355-364. doi 10.1046/j.1365-2125.1996.41911.x

Tavernini L. Linear multistep methods for the numerical solution of Volterra functional differential equations. *Appl Anal*. 1973;3(2):169-185. doi 10.1080/00036817308839063

Vogt A.-C.S., Arsiwala T., Mohsen M., Vogel M., Manolova V., Bachmann M.F. On iron metabolism and its regulation. *Int J Mol Sci*. 2021;22(9):4591. doi 10.3390/ijms22094591

Wadsworth G.R. Recovery from acute haemorrhage in normal men and women. *J Physiol*. 1955;129(3):583-593. doi 10.1113/jphysiol.1955.sp005380

Weinberg E.D. Iron availability and infection. *Biochim Biophys Acta*. 2009;1790(7):600-605. doi 10.1016/j.bbagen.2008.07.002

Xie Y., Hou W., Song X., Yu Y., Huang J., Sun X., Kang R., Tang D. Ferroptosis: process and function. *Cell Death Differ*. 2016;23(3):369-379. doi 10.1038/cdd.2015.158

Xu Y., Alfaro-Magallanes V.M., Babitt J.L. Physiological and pathophysiological mechanisms of hepcidin regulation: clinical implications for iron disorders. *Br J Haematol*. 2021;193(5):882-893. doi 10.1111/bjh.17252

Ziegler A.K., Grand J., Stangerup I., Nielsen H.J., Dela F., Magnussen K., Helge J.W. Time course for the recovery of physical performance, blood hemoglobin, and ferritin content after blood donation. *Transfusion*. 2015;55(4):898-905. doi 10.1111/trf.12926

Conflict of interest. The authors declare no conflict of interest.
Received May 26, 2025. Revised August 8, 2025. Accepted August 15, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1041-1050

doi 10.18699/vjgb-25-109

Identification and analysis of the connection network structure between the components of the immune system in children

D.S. Grebennikov [1, 2, 3], A.P. Toptygina[4], G.A. Bocharov [1, 2, 3]

[1] Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences (INM RAS), Moscow, Russia
[2] Moscow Center of Fundamental and Applied Mathematics at INM RAS, Moscow, Russia
[3] Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, Russia
[4] Gabrichevsky Research Institute for Epidemiology and Microbiology, Moscow, Russia

g.bocharov@inm.ras.ru

Abstract. Identification of the connections between the various functional components of the immune system is a crucial task in modern immunology. It is key to implementing the systems biology approach to understand the mechanisms of dynamic changes and outcomes of infectious and oncological diseases. The data characterizing an individual's immune status typically have a high-dimensional state space and a small sample size. To study the network topology of the immune system, we utilized previously published original data from Toptygina et al. (2023), which included measurements of the immune status in 19 healthy individuals (children, 9 boys and 10 girls, aged 1 to 2 years), i. e., the immune cells (42 subpopulations) obtained by flow cytometry; cytokine levels (13 types) obtained by multiplex analysis; and antibody levels (4 types) determined by using enzyme immunoassay. To correctly identify statistically significant correlations between the measured variables and construct the respective network graph, it is necessary to use an approach that takes into account the small size of the dataset. In this study, we implemented and analyzed an approach based on the regularized debiased sparse partial correlation (DSPC) algorithm to evaluate sparse partial correlations and identify the network structure of relationships in the immune system of healthy individuals (children) based on immune status data, which includes a set of indicators for subpopulations of immune cells, cytokine levels, and antibodies. For different levels of statistical significance, heatmaps of the partial correlations were constructed. The graph visualization of the DSPC networks was performed, and their topological characteristics were analyzed. It is found that with a limited measurements sample, the choice of a statistical significance threshold critically affects the structure of the partial correlations matrix. The final verification of the immunologically correct structure of the correlation-based network requires both an increase in the sample size and consideration of a priori mechanistic views and models of the functioning of the immune system components. The results of this analysis can be used to select the therapy targets and design combination therapies.

Key words: immune system; immune status; correlation analysis; partial correlations; network topology; graphs; DSPC algorithm

For citation: Grebennikov D.S., Toptygina A.P., Bocharov G.A. Identification and analysis of the connection network structure between the components of the immune system in children. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1041-1050. doi 10.18699/vjgb-25-109

Funding. The study was funded by the Russian Science Foundation (Grant Number 23-11-00116) (construction of correlation networks and analysis of the topology of connections graphs), and partially supported by the Moscow Center of Fundamental and Applied Mathematics at INM RAS (Agreement with the Ministry of Science and Higher Education of the Russian Federation No. 075-15-2025-347) (basic statistical data analysis in Section 2).

Идентификация и анализ сетевой структуры связей между компонентами иммунной системы у детей

Д.С. Гребенников [1, 2, 3], А.П. Топтыгина[4], Г.А. Бочаров [1, 2, 3]

[1] Институт вычислительной математики им. Г.И. Марчука Российской академии наук, Москва, Россия
[2] Отделение Московского центра фундаментальной и прикладной математики в ИВМ РАН, Москва, Россия
[3] Первый Московский государственный медицинский университет им. И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский университет), Москва, Россия
[4] Московский научно-исследовательский институт эпидемиологии и микробиологии им. Г.Н. Габричевского Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Москва, Россия

g.bocharov@inm.ras.ru

Аннотация. Идентификация связей между различными функциональными компонентами иммунной системы представляет собой чрезвычайно актуальную задачу современной иммунологии. Это необходимо для понимания механизмов динамики и исхода инфекционных и онкологических заболеваний при реализации

системно-биологического подхода. Параметры, характеризующие иммунный статус человека, отличаются большой размерностью пространства состояний при малой мощности выборки. Для изучения сетевой топологии иммунной системы нами использованы ранее опубликованные оригинальные данные (Toptygina et al, 2023) измерений показателей иммунного статуса у 19 здоровых индивидуумов – детей, 9 мальчиков и 10 девочек, в возрасте от одного до двух лет: популяций иммунных клеток (42 субпопуляции), полученных с помощью проточной цитометрии; уровней цитокинов (13 типов), полученных методами мультиплексного анализа; уровня антител (4 типа), определенных с помощью иммуноферментного анализа. Для корректного (статистически значимого) определения корреляционных связей между измеряемыми переменными и построения графа сетевой топологии может быть использован подход, который учитывает малый размер множества данных. В нашей работе был реализован и исследован подход, в основе которого лежит регуляризированный алгоритм скорректированных разреженных частных корреляций (DSPC) оценивания разреженных частных корреляций и идентификации сетевой структуры взаимосвязей в иммунной системе по данным иммунного статуса здоровых детей, включающего набор показателей субпопуляций клеток иммунной системы, уровня цитокинов и антител. Для разных уровней статистической значимости были построены тепловые карты частных корреляций, выполнена визуализация сетей частных корреляций в виде графов и проведен анализ их топологических характеристик. Получено, что при ограниченной выборке измерений выбор порога для уровня статистической значимости имеет принципиальное значение для формирования матрицы частных корреляций. Окончательная верификация иммунологически корректной структуры связей требует как увеличения размера выборки, так и сопряжения с априорными механизменными представлениями и моделями функционирования компонент иммунной системы. Результаты могут быть использованы для выбора мишеней терапии и формирования комбинированных воздействий.
Ключевые слова: иммунная система; иммунный статус; корреляционный анализ; частные корреляции; сетевая топология; графы; алгоритм DSPC

Introduction

The human immune system functions to maintain the antigenic homeostasis of the body's internal environment. It is a system with distributed parameters reflecting the spatial organization, phenotypic and clonal structure of its constituent cell populations. The cells of the immune system continuously interact with each other, and the balance of processes increasing or decreasing their activity underlies the development of productive or abortive reactions (Ng et al., 2013). Implementation of a systems biology approach to the investigation of the mechanisms determining the dynamics and outcome of infectious and oncological diseases requires identification of the structure of cellular interconnection networks in the immune system. An example of studying the connections network (network topology) between populations of cellular components of the immune system is provided in (Rieckmann et al., 2017), where the quantitative proteomics data were used for identification of the social architecture of immune cell interactions. The description of the network topology is associated with construction of a graph, with the vertices corresponding to specific cell populations of the immune system, and the edges representing connections of a diverse nature between the corresponding vertices.

To date, a large number (about 100 documented) of methods have been developed for analyzing the structural organization of intercellular interactions based on data of a diverse nature, including spatial and cellular transcriptomics, expression of ligand receptors, as well as intracellular signalling components (Armingol et al., 2024). They are used for the assessment of the connectivity indices or communication structures between cells, which provide the basis for building the graphs of connectivity networks. Both the biophysical and biochemical principles, and statistical data analysis methods in combination with machine learning, can be used to assess the strength of the intercellular connections.

The construction of a quantitative interactome of immune cells based on receptor proteins expressed on their surface is presented in (Shilts et al., 2022). It implements a number of graphs based on a set of physical connections between cells of the immune system in major human organs identified using multiplex immune and transcriptomic analysis technologies, genetic databases and biochemical methods for screening interactions between cells. Visualization of the transcriptome analysis data as a graph reflecting the genes co-expression is an integrative part of modern systemic vaccinology studies (Cortese et al., 2025).

The aim of our study was to implement a new approach to identifying the network structure of relationships in the immune system of a healthy individual based on the results of a correlation analysis of previously published data on the immune status of children aged one to two years. The data set includes the measurements of the immune status parameters, i. e. the subpopulations of immune cells, cytokine concentrations and antibody levels (Toptygina et al., 2023). The research objectives include the correlation analysis of children's immune status data to build heatmaps of partial correlations, visualization of the partial correlations networks as graphs, and analysis of the topological characteristics of the graph models.

The present work consists of four sections. The "Materials and methods" section describes the specific features of the source data, methods of correlation analysis, the correlation-based approach to identifying a network structure of relationships between the immune status parameters, and examines the topological properties of the corresponding graphs. Principal components analysis is performed. The "Results" section presents the results of network construction for various threshold levels of statistical significance of the correlations, an immunological interpretation of the corresponding network topologies, and a robustness analysis. The results of the work are discussed in the "Discussion" section.

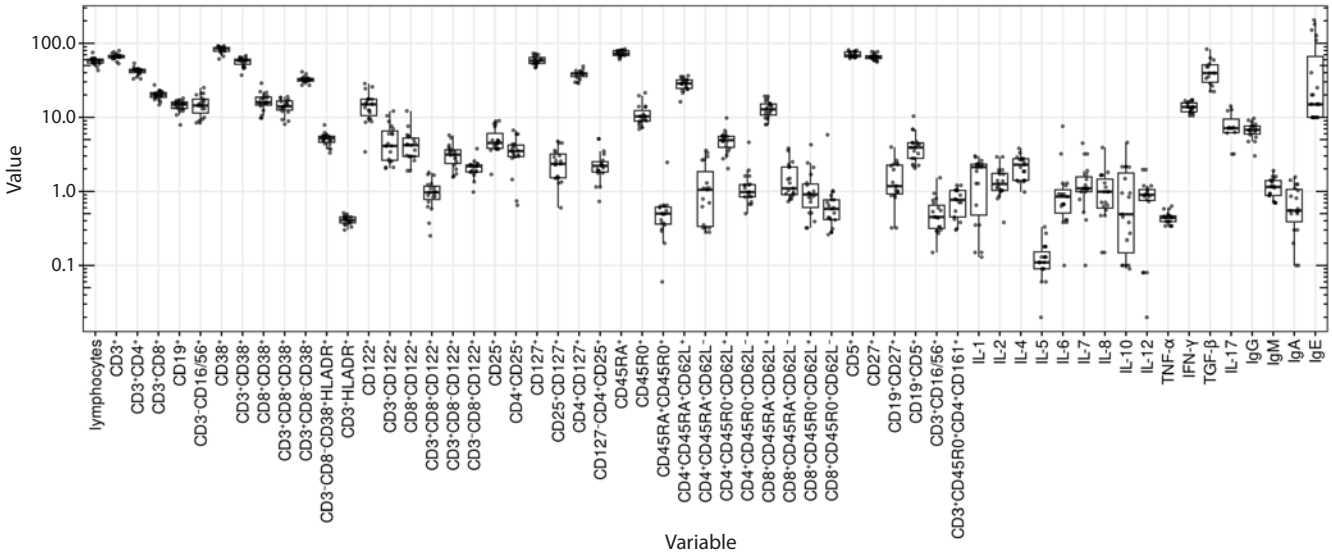


Fig. 1. Data on immune status in healthy individuals – children aged one to two years (adapted from Toptygina et al., 2023).

Individual measurements, median sample values, and 25–75 % quartiles are presented. The abscissa shows the names of the immune status indicators. The ordinate shows the percentage of cells (%), the levels of cytokines (pg/ml) and immunoglobulins A, M, G (g/l), IgE (IU/ml).

Materials and methods

Immune status data. To study the network topology of the immune system, we used previously published original data (Toptygina et al., 2023). The data are a set of measurements of immune status indicators in 19 healthy individuals, i. e., children aged one to two years: populations of immune cells (42 subpopulations) obtained by flow cytometry; cytokine levels (13 types) obtained by multiplex analysis; antibody levels (4 types) determined by enzyme immunoassay. The data samples are summarized in Figure 1 as individual measurements, median values, and 25 and 75 % quartiles. The distribution of the indicators does not follow either the normal or the log-normal behavior.

The data on the immune status of children are characterized by a large dimensionality of the state space (59) and a small sample size (19 patients), which is typical for systems biology studies (Basu et al., 2017). If the sample size is large enough, one can use the approach based on partial correlations in order to determine the relationships between the immune status parameters. Otherwise, an approach that takes into account the small size of the data set has to be implemented to correctly determine statistically significantly correlations between the measured variables and construct a network topology graph. It should be noted that all the children belonged to the same age group from one to two years old, which in medical practice is not customary to subdivide further. Due to the small size of the group (19 people), additional division by gender (10 girls and 9 boys) would have reduced the statistical power below the critical level required for the method used in our study.

Principal component analyses. The principal component analysis (PCA) was performed using the prcomp function in the R language, the factoextra R package (version 1.0.7) was used for visualization. To perform the PCA, the data were standardized, and the variables TGF-β, IL-17, and $CD3^+CD45R0^+CD4^+CD161^+$ were excluded from the analysis due to missing data. The analysis of the principal components (PCs) did not reveal the possibility of explaining the variance of the data by a small number of the components (Fig. 2*a*), and no correlation-based clusters of immune status variables exist in the first two PCs (Fig. 2*b*).

Methods of partial correlation analyses and reconstruction of the connection network. An alternative to the standard method of estimating partial correlations is an approach using regularization methods to estimate the matrix of partial correlations (Epskamp, Fried, 2018). The principle of regularization is based on the assumption that the number of connections in the constructed model network is significantly less than the number of observed variables, i. e. the real network is sparse. Accordingly, the LASSO method (Epskamp, Fried, 2018) is used as a regularizing correction that allows zeroing out insignificant correlations between variables (the number of edges in the graph). To analyze our data, we used this approach for the estimation of debiased sparse partial correlations matrix implemented in algorithm DSPC (Basu et al., 2017), which provides additional correction of estimates of the elements of the inverse covariance matrix, i. e. the elements of the partial correlations matrix. The estimates of the correlation matrix elements were represented as heatmaps and visualized as weighted networks, where the vertices (nodes) represent the immune status variables and the edges show correlations between them. The results of estimating the correlation-based relationships depend significantly on the algorithm parameters: 1) the value of the parameter λ for the regularization term in the form of ℓ_1 norm of the inverse covariance matrix; 2) the choice of the statistical significance level p for the predicted correlation relationship. Below, we study the effect of the p-value on the network topology of connections in the immune system.

To calculate the sparse partial correlations using the DSPC method, we used the Java application CorrelationCalculator

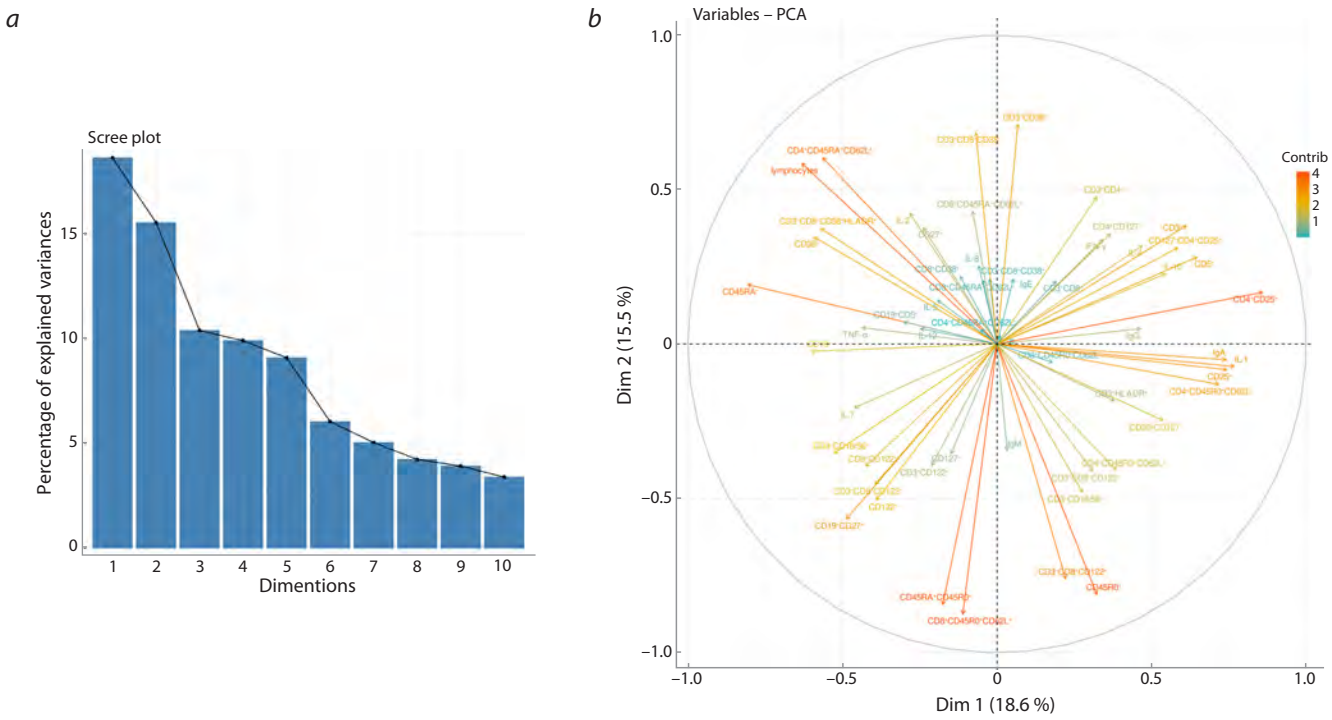


Fig. 2. Principal component analysis: *a* – fraction of explained variance; *b* – composition of the first two principal components.

(version 1.0.1) developed in (Basu et al., 2017). The original data were normalized, i. e. logarithmically transformed and standardized. A graphical representation of statistically significant correlations (for $p < 0.01$; 0.05; 0.1; 0.15) in the form of heatmaps and graphs of correlation networks was performed using the R packages igraph (version 1.6.0) and ggplot2 (version 3.5.2). The topological characteristics of the correlation networks graphs were calculated using the igraph package in R (version 1.6.0).

Results

In what follows, we study the effect of the *p*-value on the network topology of connections in the immune system. The conventionally considered statistical significance levels 0.01, 0.05, 0.1, 0.15 are analyzed.

Heatmap and connection graph for $p = 0.01$

The heatmap of partial correlations between immune status parameters for healthy children at a statistical significance threshold $p = 0.01$ is presented in Figure 3*a*. The corresponding graph of the network is shown in Figure 3*b*. This graph has 23 nodes and 12 edges (connections). In fact, connectivity in the network is missing. Figure 3*c* shows the distribution of immune response indicators with respect to the number of identified links between them. The node with the maximum number (2 in total) of correlations represents the CD4 T cell population ($CD3^+CD4^+$).

Heatmap and connection graph for $p = 0.05$

The heatmap of correlations between immune status parameters for healthy children at a statistical significance threshold $p = 0.05$ is presented in Figure 4*a*. The corresponding network graph is shown in Figure 4*b*. This graph has 53 nodes and 44 edges (connections). The cohesion of individual network components is strengthened, but overall, it is absent. Figure 4*c* shows the distribution of immune response indicators with respect to the number of identified links between them. The nodes with the maximum number of correlations (called hubs) represent the proinflammatory cytokines IL-8, IL-12, and central memory T cells ($CD4^+CD45RA^+CD62L^+$, $CD8^+CD45R0^+CD62L^+$), Th17 ($CD3^+CD45R0^+CD4^+CD161^+$) and activated NK cells ($CD3^-CD8^+CD122^+$). The maximum number of connections increases to three.

Heatmap and connection graph for $p = 0.1$

The heatmap of correlations between immune status parameters for healthy children at a statistical significance threshold $p = 0.1$ is presented in Figure 5*a*. The corresponding network graph is shown in Figure 5*b*. This graph has 59 nodes and 69 edges (connections). Figure 5*c* shows the distribution of immune response indicators with respect to the number of identified links between them. The nodes with the maximum number of correlations (four in this case) represent the cytokines IL-4, IL-12 inducing the cellular and humoral immunity, the terminally differentiated effector memory T cells ($CD4^+CD45RA^+CD62L^-$, $CD8^+CD45RA^+CD62L^-$), and Th17 cells ($CD3^+CD45R0^+CD4^+CD161^+$).

Heatmap and connection graph for $p = 0.15$

The heatmap of correlations between immune status parameters for healthy children at a statistical significance threshold $p = 0.15$ is presented in Figure 6*a*. The corresponding network graph is shown in Figure 6b. This graph has 59 nodes and

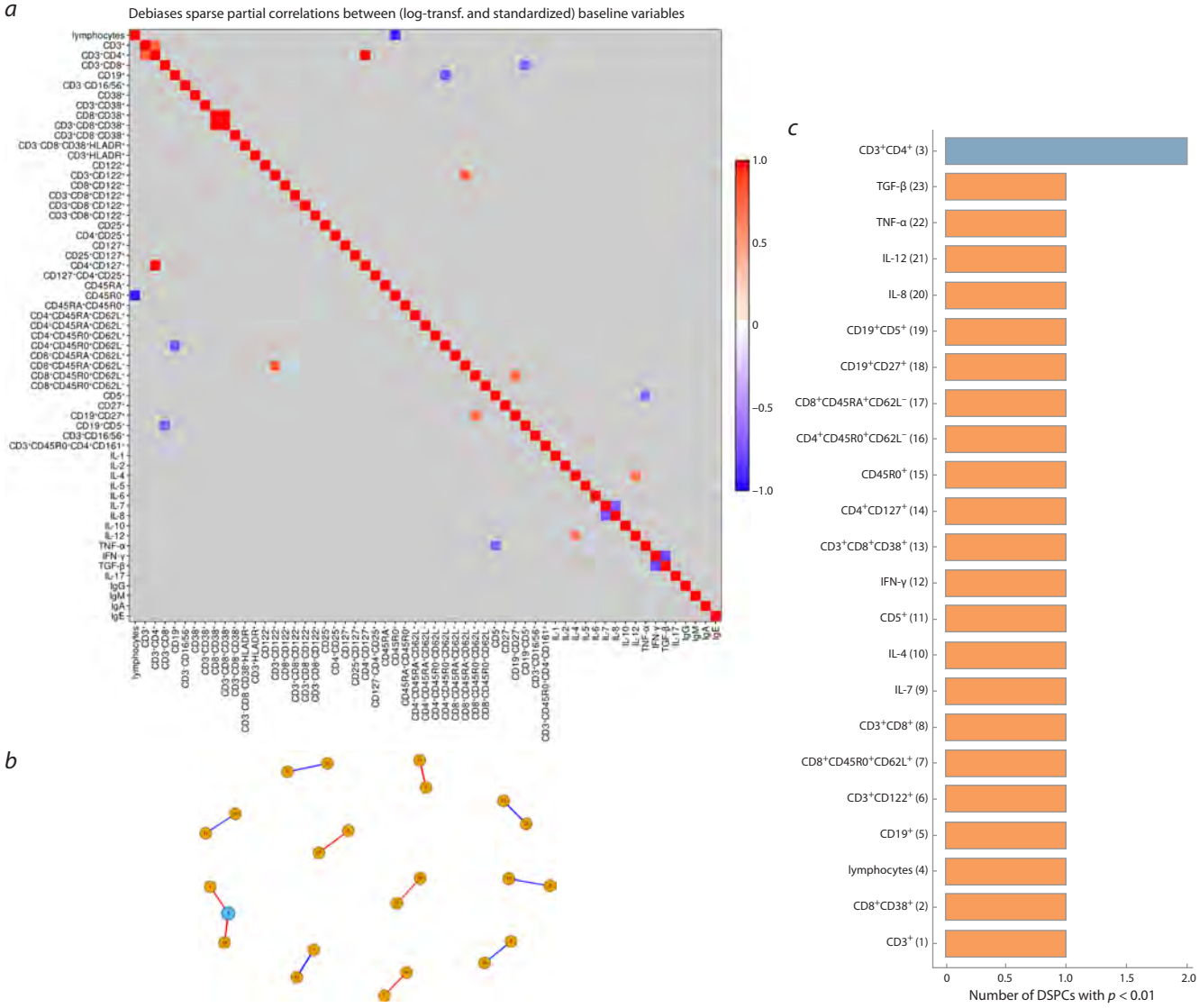


Fig. 3. Heatmap and network graph of immunological parameters in healthy children at a statistical significance level of $p = 0.01$:
a – heatmap of correlations between immune status indicators; *b* – graph of connections network at $p = 0.01$; *c* – characteristics of the complexity of the network of connections.

Here and in Figures 4–6: the node numbers correspond to the immune status parameters shown in *c*. The ordinate names the immune status indicators. The abscissa shows the degrees of the graph nodes. Positive correlations (red lines), negative correlations (blue lines), the thickness of the edges is proportional to the absolute values of the DSPC coefficients. The color of the nodes corresponds to the node index, i. e. the number of significant correlations.

106 edges (connections). Figure 6*c* shows the distribution of immune response indicators with respect to the number of identified links between them. The nodes with the maximum number of correlations (hubs) represent the immunoglobulins IgM, plasma cells ($CD3^-CD8^-CD38^+HLADR^+$), activated T cells ($CD3^+CD8^-CD38^+$, $CD8^+CD122^+$), and the double-positive activated cells ($CD45RA^+CD45R0^+$) reflecting the transition from naive to memory cells. The maximum number of connections increases to six.

Analysis of the robustness of correlation estimates

To assess the stability of the obtained DSPC correlation coefficients in relation to the sample size, a procedure was performed for generating ten different subsamples according to the vfold10 scheme. In most cases, it corresponds to the selection of 17 out of 19 measurements. The coefficient of variation (the ratio of the standard deviation to the mean value) of the DSPC coefficients estimated from the generated subsamples was chosen as a measure of stability (robustness). The estimated coefficients of variation are shown in Figure 7 for four levels of statistical significance in the form of heatmaps. Importantly, their absolute values do not exceed 0.1.

Comparative analysis of topological properties of graphs of correlations between indicators of immune status

The Table shows the results of calculating the topological characteristics of the constructed graphs of correlation networks

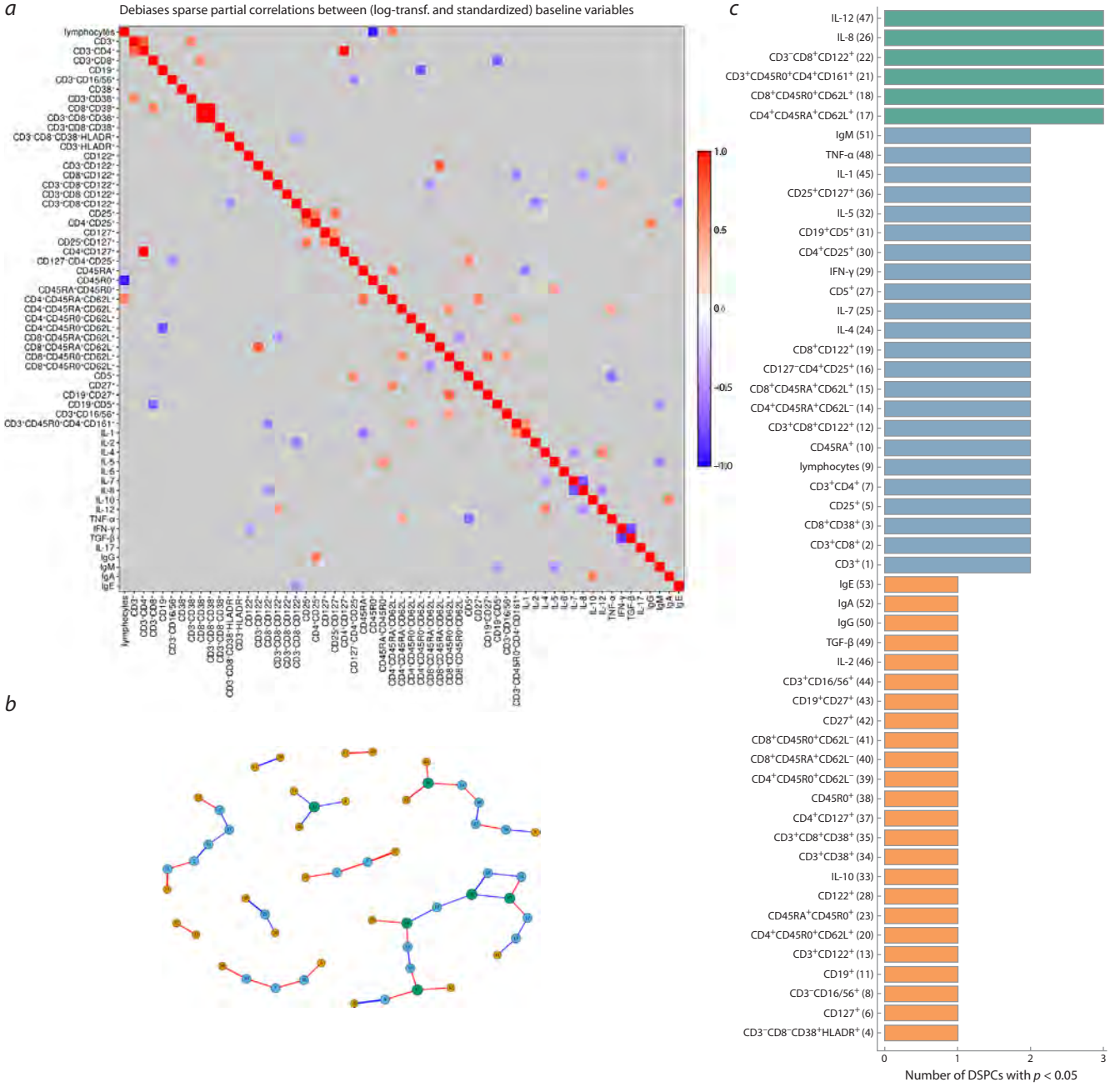


Fig. 4. Heatmap and network graph of immunological parameters in healthy children at a statistical significance level of $p = 0.05$:

a – heatmap of correlations between immune status indicators; *b* – graph of connections network at $p = 0.05$; *c* – characteristics of the complexity of the network of connections.

between immune status indicators for various thresholds of statistical significance. The following basic characteristics were considered: graph diameter, graph radius, girth of graph (the length of the smallest cycle contained in the graph), average path length, graph energy, spectral radius, edge density, clustering coefficient, average graph diversity (determined through entropy calculated by the weights of incident edges – the absolute values of the correlation coefficients DSPC), the number of separators, and the number of unconnected subgraphs.

The number of nodes, edges, and maximum node degrees grows with increasing statistical significance threshold. However, the graph diameter, radius, girth and average path length exhibit a non-monotonic dependence, initially increasing and then decreasing, which indicates a transformation of properties towards the "small world network" family. The graph energy and spectral radius increase monotonically with increasing threshold p. The clustering coefficient also increases, indicating that the graph nodes tend to cluster together. Interestingly, the number of cutting nodes and edges

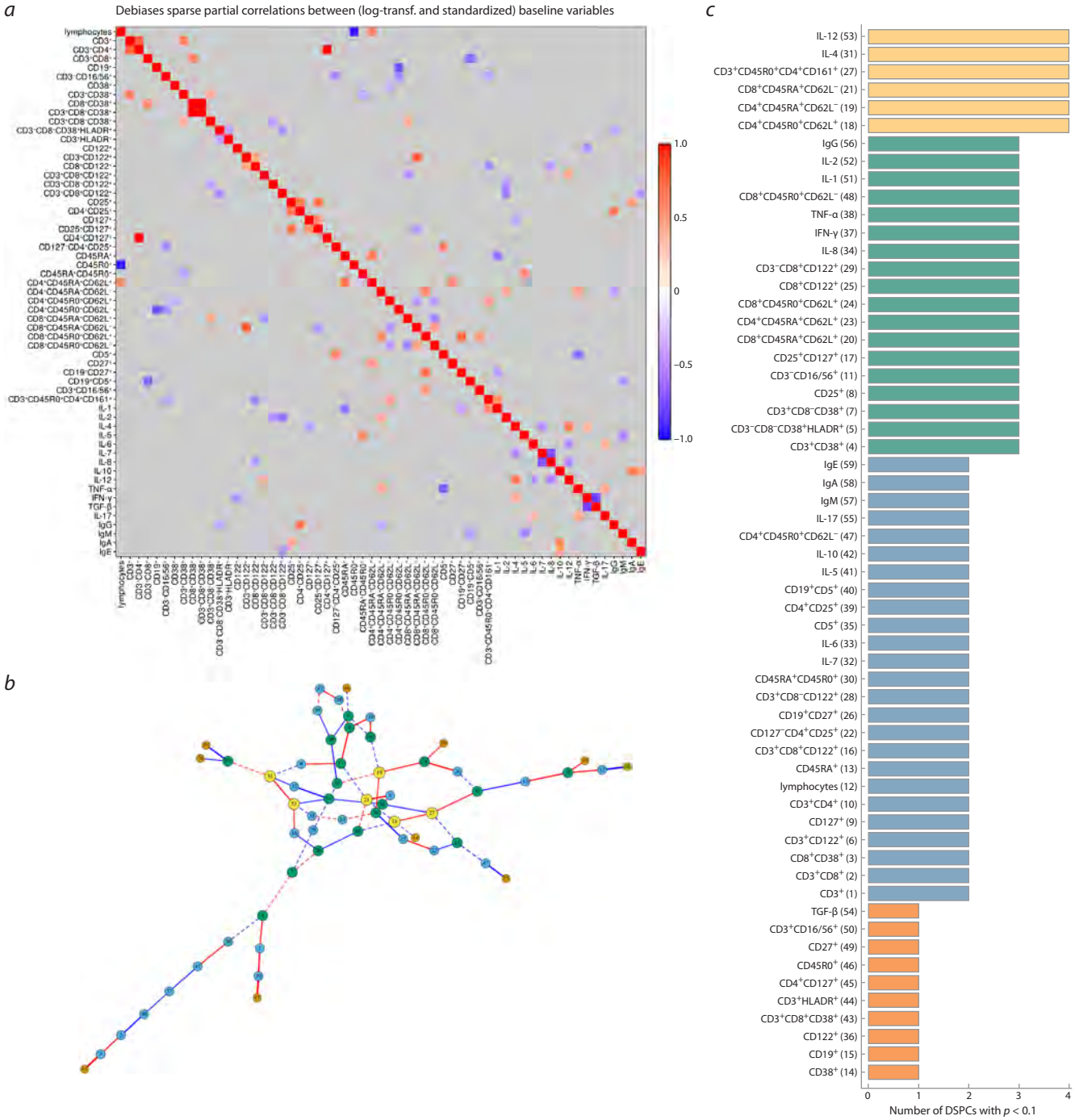


Fig. 5. Heatmap and network graph of immunological parameters in healthy children at a statistical significance level of $p = 0.1$:

a – heatmap of correlations between immune status indicators; *b* – graph of connections network at $p = 0.1$; *c* – characteristics of the complexity of the network of connections. Solid lines of the edges correspond to correlations with a significance level of $p < 0.05$, dashed lines, to $p < 0.1$.

decreases at $p = 0.15$, which may indicate an increase in the robustness of the connections graph. As expected, the number of disconnected subgraphs decreases.

Discussion

Identification of the connection structures between the various functional components of the immune system is an extremely urgent task of modern immunology. This is due to an extremely high number of measured characteristics, with a relatively small sample size, reflecting the situation in big data biomathematics, called the "curse of dimensionality". To analyze the relationships between immune status parameters, we implemented and analyzed an approach based on a regularized method for estimating sparse partial correlations implemented in the DSPC algorithm (Basu et al., 2017), which minimizes the number of false correlations. It is noted that the results of applying the algorithm may depend on the sample size, imputation of missing data, the nature of the true network

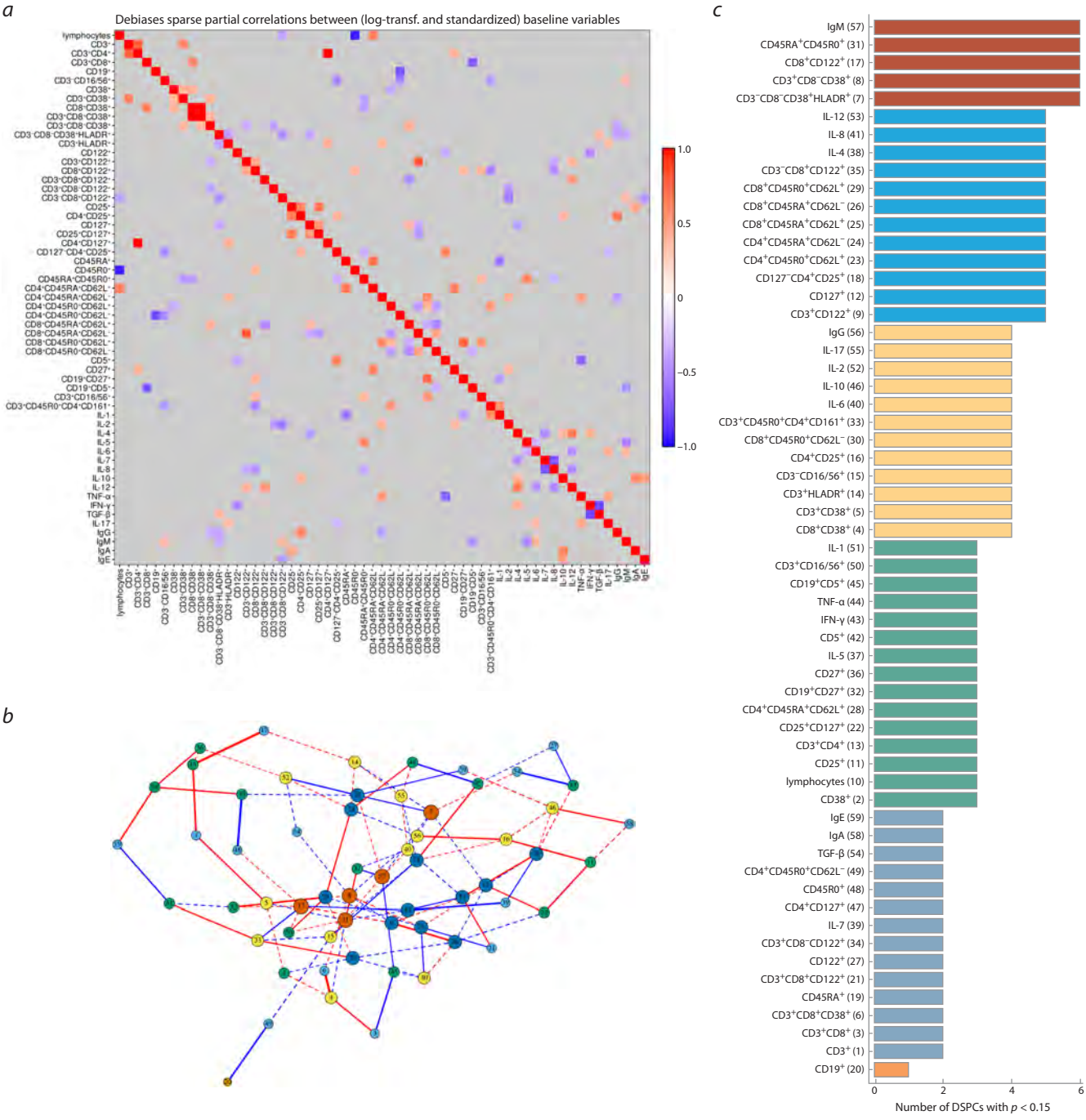


Fig. 6. Heatmap and network graph of immunological parameters in healthy children at a statistical significance level of $p = 0.15$:

a – heatmap of correlations between immune status indicators; *b* – graph of connections network at $p = 0.15$; *c* – characteristics of the complexity of the network of connections. Solid lines of the edges correspond to correlations with a significance level of $p < 0.05$, dashed lines, to $p < 0.15$.

structure and other aspects. Our work demonstrates that, given a limited sample size of measurements, an a priori assignment of the level of statistical significance is of fundamental importance for the formation of a matrix of partial correlations. Increasing the statistical significance threshold increases the complexity of the network topology generated by the DSPC-based approach. Final verification of the immunologically correct structure of connections requires both an increase in the sample size and conjugation with a priori mechanistic views and models of the functioning of the immune system components, i. e. the participation of clinical immunologists (Qiao et al., 2025). An important step in this direction was the development of the ImmunoGlobe tool for constructing and analyzing the network of interactions in the immune system (Atallah et al., 2020) using phenomenological information from the fundamental textbook "Janeway's Immunobiology" (Murphy, Weaver, 2017).

The aim of this work is to implement and introduce a new method for identifying relationships between cellular and humoral components of the immune systems. Identification

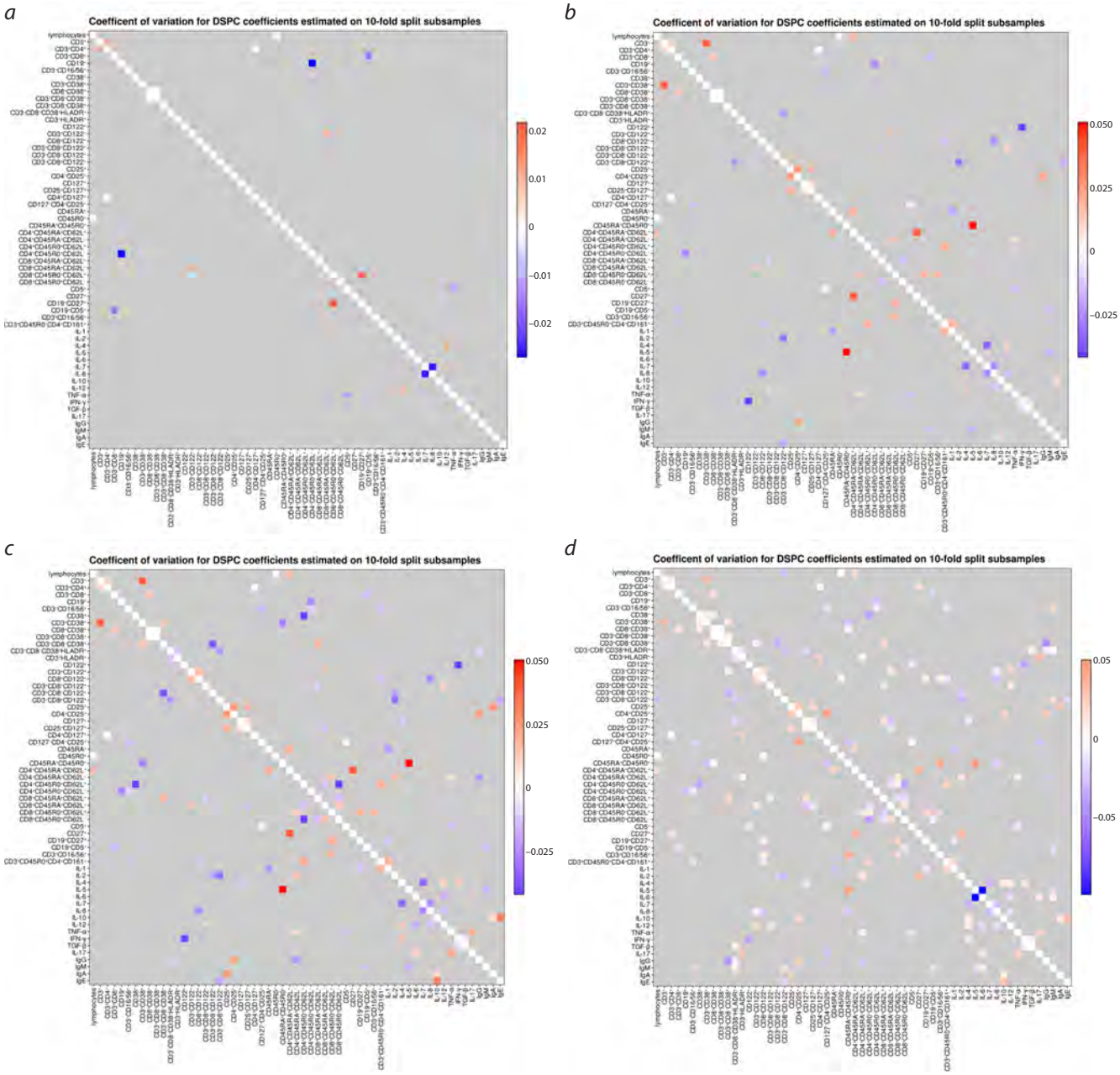


Fig. 7. Matrices of estimates of the variation coefficients for four significance levels: $p < 0.01$ (*a*); $p < 0.05$ (*b*); $p < 0.1$ (*c*); $p < 0.15$ (*d*).

of the network relationships between elements of immune status is central to the systems immunology approach, but the relevant analytical tools remain undeveloped. All currently existing verified concepts of immune networks are limited to schemes with no more than three or four components (antigen presentation, differentiation pathways, paracrine and autocrine interactions). For this reason, it is not possible to uniquely select and verify one of the presented networks. If we adhere to the generally accepted level of significance ($p = 0.05$), then we should give preference to the network constructed in the section "Heatmap and graph of connections for $p = 0.05$". Identifying the network structure of relationships between components of cellular and humoral immunity is a necessary element for the transition from a static description of immune status to a systems dynamics consideration of the maintenance of immune homeostasis.

Conclusion

The development of combination therapies for chronic diseases associated with induction of several components of the immune system requires understanding of the topology and strength of the structural connections in the system. Our study demonstrates for the first time that DSPC-based methods can be used to obtain consistent estimates of significant partial correlations for similar problems in a typical situation with a large number of measured immune status parameters and

Comparative analysis of topological properties of graphs of correlations between indicators of immune status for various significance thresholds

Topological characteristics	$p \le 0.01$	$p \le 0.05$	$p \le 0.1$	$p \le 0.15$
Number of nodes, n	23	53	59	59
Number of edges, m	12	44	69	106
Maximun digree, Δ_G	2	3	4	6
Diameter, D	2	11	17	7
Radius, r	1	1	9	4
Girth, g	0	4	3	3
Average path length, l_G	1.08	3.6	6.0	3.3
Energy, E_n	22.8	58.7	77.5	94.6
Spectral radius, ρ	1.4	2.3	3.0	4.3
Edge density, ρ_d	0.05	0.03	0.04	0.06
Clustering coefficient, C	0	0	0.026	0.055
Topological diversity of vertices, D_{DSPC}	0.04	0.54	0.82	0.96
Number of node separator, n_{cut}	1	27	21	2
Number of edge separatrors, m_{cut}	12	40	24	2
Number of unconnected subgraphs	11	10	1	1

a small number of patients. Translation of the results into biomedical practice to address the challenges of personalized treatment and prevention of immune-dependent pathological processes requires an active participation of clinicians in order to determine therapy targets and quantitatively predict its effectiveness.

References

Armingol E., Baghdassarian H.M., Lewis N.E. The diversification of methods for studying cell-cell interactions and communication. *Nat Rev Genet*. 2024;25(6):381-400. doi 10.1038/s41576-023-00685-8

Atallah M.B., Tandon V., Hiam K.J., Boyce H., Hori M., Atallah W., Spitzer M.H., Engleman E., Mallick P. ImmunoGlobe: enabling systems immunology with a manually curated intercellular immune interaction network. *BMC Bioinformatics*. 2020;21(1):346. doi 10.1186/s12859-020-03702-3

Basu S., Duren W., Evans C.R., Burant C.F., Michailidis G., Karnovsky A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics*. 2017;33(10):1545-1553. doi 10.1093/bioinformatics/btx012

Cortese M., Hagan T., Rouphael N., Wu S.Y., Xie X., Kazmin D., Wimmers F., … Subramaniam S., Mulligan M.J., Khurana S., Golding H., Pulendran B. System vaccinology analysis of predictors and mechanisms of antibody response durability to multiple vaccines in humans. *Nat Immunol*. 2025;26(1):116-130. doi 10.1038/s41590-024-02036-z

Epskamp S., Fried E.I. A tutorial on regularized partial correlation networks. *Psychol Methods*. 2018;23(4):617-634. doi 10.1037/met0000167

Murphy K., Weaver C. Janeway's Immunobiology. New York, NY: Garland Science/Taylor & Francis Group, 2017. ISBN 978-0-8153-4505-3 Available: https://inmunologos.wordpress.com/wp-content/uploads/2020/08/janeways-immunobiology-9th-ed_booksmedicos.org_.pdf

Ng C.T., Snell L.M., Brooks D.G., Oldstone M.B. Networking at the level of host immunity: immune cell interactions during persistent viral infections. *Cell Host Microbe*. 2013;13(6):652-664. doi 10.1016/j.chom.2013.05.014

Qiao L., Khalilimeybodi A., Linden-Santangeli N.J., Rangamani P. The evolution of systems biology and systems medicine: From mechanistic models to uncertainty quantification. *Annu Rev Biomed Eng*. 2025;27(1):425-447. doi 10.1146/annurev-bioeng-102723-065309

Rieckmann J.C., Geiger R., Hornburg D., Wolf T., Kveler K., Jarrossay D., Sallusto F., Shen-Orr S.S., Lanzavecchia A., Mann M., Meissner F. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol*. 2017;18(5):583-593. doi 10.1038/ni.3693

Shilts J., Severin Y., Galaway F., Müller-Sienerth N., Chong Z.S., Pritchard S., Teichmann S., Vento-Tormo R., Snijder B., Wright G.J. A physical wiring diagram for the human immune system. *Nature*. 2022;608(7922):397-404. doi 10.1038/s41586-022-05028-x. Erratum in: *Nature*. 2024;635(8037):E1. doi 10.1038/s41586-024-07928-6

Toptygina A., Grebennikov D., Bocharov G. Prediction of specific antibody- and cell-mediated responses using baseline immune status parameters of individuals received measles-mumps-rubella vaccine. *Viruses*. 2023;15(2):524. doi 10.3390/v15020524

Conflict of interest. The authors declare no conflict of interest.
Received July 30, 2025. Revised September 15, 2025. Accepted September 19, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1051-1061

doi 10.18699/vjgb-25-110

Self-learning virtual organisms in a physics simulator: on the optimal resolution of their visual system, the architecture of the nervous system and the computational complexity of the problem

M.S. Zenin[1], A.P. Devyaterikov[2], A.Yu. Palyanov [1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
palyanov@iis.nsk.su

Abstract. Vision plays a key role in the lives of various organisms, enabling spatial orientation, foraging, predator avoidance and social interaction. In species with relatively simple visual systems, such as insects, effective behavioral strategies are achieved through high neural specialization, adaptation to specific environmental conditions, and the use of additional sensory systems such as olfaction or hearing. Animals with more complex vision and nervous systems, such as mammals, have greater cognitive abilities and flexibility, but this comes with increased demands on the brain's energy costs and computational resources. Modeling the features of such systems in a virtual environment could allow researchers to explore the fundamental principles of sensorimotor integration and the limits of cognitive complexity, as well as test hypotheses about the interaction between perception, memory and decision-making mechanisms. In this work, we implement and investigate a model of virtual organisms with a visual system operating in a three-dimensional physical environment using the Unity ML-Agents software – one of the most high-performance simulation platforms currently available. We propose a hierarchical control architecture that separates locomotion and navigation tasks between two modules: (1) visual perception and decision-making, and (2) coordinated control of limb movement for locomotion in the physical environment. A series of numerical experiments was conducted to examine the influence of visual system parameters (e.g, resolution of the "first-person" view), environmental configuration and agent architectural features on the efficiency and outcomes of reinforcement learning (using the PPO algorithm). The results demonstrate the existence of an optimal range of resolutions that provide a trade-off between computational complexity and success in accomplishing the task, while excessive dimensionality of sensory inputs or action space leads to slower learning. We performed system performance profiling and identified key bottlenecks in large-scale simulations. The discussion considers biological parallels, highlighting cases of high behavioral efficiency in insects with relatively low-resolution visual systems, and the potential of neuroevolutionary approaches for adapting agent architectures. The proposed approach and the results obtained are of potential interest to researchers working on biologically inspired artificial agents, evolutionary modeling, and the study of cognitive processes in artificial systems.

Key words: virtual organism; computational modeling; computational complexity; vision system; neural network; simulator; PPO; reinforcement learning; Unity ML-Agents

For citation: Zenin M.S., Devyaterikov A.P., Palyanov A.Yu. Self-learning virtual organisms in a physics simulator: on the optimal resolution of their visual system, the architecture of the nervous system and the computational complexity of the problem. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1051-1061. doi 10.18699/vjgb-25-110

Самообучающиеся виртуальные организмы в физическом симуляторе: об оптимальном разрешении их зрительной системы, архитектуре нервной системы и вычислительной сложности задачи

М.С. Зенин[1], А.П. Девятериков[2], А.Ю. Пальянов [1, 2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
palyanov@iis.nsk.su

Аннотация. Зрение играет ключевую роль в жизни множества различных организмов, обеспечивая ориентацию в пространстве, поиск пищи, избегание хищников и социальное взаимодействие. У видов с относительно простой зрительной системой, таких как насекомые, эффективная поведенческая стратегия достигается за счет высокой специализации нейронов, адаптации к конкретным условиям среды, а также благодаря дополнительным сенсорным системам – обонянию или слуху. У животных с более сложным зрением и нервной системой, таких как млекопитающие, когнитивные возможности и способности к адаптации значительно выше, однако выше и энергозатраты на работу мозга. Моделирование особенностей таких систем в виртуальной среде позволило бы исследовать фундаментальные принципы функционирования и обучения когнитивных систем, включая механизмы восприятия, памяти, принятия решений и их взаимодействие. В данной работе объектом исследования являются виртуальные организмы, обладающие зрительной системой и функционирующие в трехмерной физической среде на базе Unity ML-Agents – одного из наиболее высокопроизводительных современных симуляторов. Предложенная иерархическая архитектура управления, разделяющая когнитивные задачи между двумя модулями – зрительного восприятия/принятия решений и управления координированным движением конечностей для перемещения в физической среде – показала существенно большую скорость и эффективность обучения по сравнению с единой системой. Проведена серия численных экспериментов, направленных на выявление влияния параметров зрительной системы, конфигурации среды и архитектурных особенностей агентов на успешность их обучения с подкреплением (алгоритм PPO). Показано, что существует диапазон разрешений, обеспечивающий компромисс между вычислительной сложностью и успешностью выполнения задачи, а избыточная размерность сенсорных входных данных или пространства действий приводит к замедлению обучения. Должное внимание уделено также оценке вычислительной сложности системы и профилированию производительности ее основных компонентов. Полученные результаты представляют потенциальный интерес в контексте исследований искусственных агентов, вдохновленных биологическими системами, эволюционного моделирования, включая нейроэволюционные подходы для создания более адаптивных и умных агентов, и изучения когнитивных процессов в них.
Ключевые слова: виртуальный организм; компьютерное моделирование; вычислительная сложность; зрительная система; нейронная сеть; симулятор; PPO; обучение с подкреплением; Unity ML-Agents

Introduction

Modeling cognitive activity, behavior, and evolutionary mechanisms in virtual environments constitutes an important step toward the development of artificial intelligence systems capable of learning, adaptation, and interaction with complex environments (Bongard, 2013; Stanley et al., 2019). The advancement of such systems has been facilitated by modern agent-based learning platforms, in particular Unity ML-Agents (Juliani et al., 2018), which allow the creation of fully featured three-dimensional simulations incorporating physics, vision, and multiple trainable agents.

Despite the relatively small number of neurons due to their small body size (compared, for instance, to mammals), the nervous systems of many invertebrates, including insects, exhibit remarkably complex, diverse, and adaptive behavior. For example, ants possess approximately 250,000 neurons, which is several orders of magnitude less than mammals (a mouse has about $7.1 \cdot 10^7$), but these insects are capable of solving complex tasks of navigation, social interaction, coordination of collective actions, and route memory (Chittka, Niven, 2009). Moreover, according to a number of studies, certain species of ants are capable of passing the mirror test, a behavioral indicator of self-awareness (Cammaerts M.-C.T., Cammaerts R., 2015). This makes them unique among insects and highlights the potential of minimal but efficiently organized nervous systems, which are of considerable interest to modern science.

Insect visual systems also serve as a source of inspiration for the design of artificial agents. In particular, compound eyes provide a wide field of view and high refresh rates, enabling efficient responses to rapidly changing stimuli (Land, Nilsson, 2012). However, their angular resolution is significantly inferior to that of humans, but this limitation is compensated by high sensitivity to movement and the capacity for learning at the level of entire behavioral sequences.

These considerations give rise to several fundamental research questions: what are the minimal requirements for an agent's visual system that enable successful adaptation to its environment? What control architecture ensures cognitive modularity under constrained computational resources? In other words, how to construct an "artificial organism" – an agent with simple but functional elements of perception and decision-making. The present study addresses these questions by investigating virtual organisms endowed with vision and operating in a 3D environment, with a focus on their ultimate cognitive efficiency, scalability, and capacity for learning in tasks of search and navigation.

The interest in structures that enable movement with minimal design complexity is also evident in engineering systems. For example, a recent study (Song et al., 2022) examines the control of hybrid soft limbs, reflecting the pursuit of structurally simple but functionally efficient solutions for motion control. The body model of the virtual organism used in the present study, in terms of degrees of freedom and segment composition, is comparable to those employed in such constructions. This makes it possible to regard it as comparable in complexity to its physical counterparts.

In our previously published work (Devyaterikov, Palyanov, 2022), we presented a simulator of the evolution of virtual organisms in a 3D environment, where each agent was equipped with a visual system and a neural network for processing sensory input. The system was based on a combination of neuroevolution and agent–environment interaction, enabling agents to perform elementary cognitive tasks that required the use of vision (such as searching for "food" necessary for "survival") and allowing the assessment of agent survivability

within a population. The present work provides estimates of the computational complexity of calculations related to physics (agent bodies, the environment, and their interactions), first-person 3D rendering for each agent, and the operation of their neural networks. In addition, it introduces a new hierarchical agent model and presents the results of a quantitative analysis of training time, speed, and efficiency as a function of visual system resolution. The (Aksoy, Camlitepe, 2018) study provides data on the number of ommatidia (photosensitive sensors) for various ant species (from 100 to 3,000). Roughly approximating such vision with a square pixel matrix, this corresponds to a visual resolution from 10×10 to 55×55.

The present work combines reinforcement learning methods (PPO (Schulman et al., 2017)), convolutional neural networks (O'Shea, Nash, 2015), approaches to hierarchical agent training (Vezhnevets et al., 2017), and practical analysis of resource-saving simulation schemes (Peng et al., 2018). We demonstrate that a hierarchical agent approach (e. g., a "Walker/Searcher" pair) enables more stable and interpretable behavior while reducing training time at a comparable level of task complexity.

Particular attention is given to investigating the impact of visual system resolution on agent learning rate, with an assessment of the minimal input image size at which the ability to perform visual search and navigation tasks is preserved. Such investigations are relevant both for biologically inspired modeling and for the development of compact and efficient AI agent architectures capable of functioning under limited computational resources (Hassabis, Humaran, 2017; Zador, 2019).

In addition, this study examines the effect of task decomposition strategies (navigation and locomotion) on training efficiency. This approach provides deeper insights into the principles underlying cognitive modularity and distributed control in complex agent systems (Botvinick et al., 2020; Tschantz et al., 2020). The introduced Searcher agent, relying exclusively on visual perception, interacts with the Walker agent, responsible for physical movement. Such a scheme enhances the adaptability of the model and improves the interpretability of agent behavior.

Thus, the aim of the present work is to conduct a systematic investigation of the limits of cognitive complexity in agents equipped with visual systems, to develop optimal control architectures and perceptual parameters, and to evaluate the performance and scalability of the proposed system implemented on the Unity ML-Agents platform.

Materials and methods

Problem statement. The problem under consideration is formulated in terms of a Markov decision process, where the agent interacts with a three-dimensional physical environment and learns to maximize cumulative reward. The task performed by the agent is described below:

Environment E: a square arena bounded by walls. Targets with radius r appear randomly within the arena and must be collected. Once a target is reached, a new one is generated.

Agent state s_t: consists of an RGB image from the first-person camera of size $h \times w \times 3$, long with a vector of control parameters (joint angles of the limbs and the corresponding torques).

Agent action a_t: a single scalar value representing a normalized rotation angle in the interval $[-1, 1]$. This parameter determines the direction of the agent's body movement. The actual rotation angle is defined as $\theta = a_t * \theta_{\max}$, where $\theta_{\max}$ is the maximum allowable rotation angle specified in the experimental parameters. In different experimental series, various values of this parameter were used, which allowed us to investigate its impact on policy efficiency (results are reported in Section "Results with varying rotation angles"). The restriction to a single control variable is due to the fact that low-level locomotion tasks (coordination of limbs and balance maintenance) are delegated to a separate Walker module, enabling the focus to remain on the cognitive aspects of the task, i. e., perception and decision-making.

Reward function $R(s_t, a_t)$: an agent receives a positive reward for successfully reaching the target.

Objective: to maximize the cumulative reward over an episode of time T, i. e., to develop a policy that enables efficient navigation in the environment and target collection based on visual information.

One of the goals of our study is to identify the minimal input image resolution at which the agent can still successfully learn within a reasonable amount of time. The formal problem formulation is as follows:

Training success is defined as achieving an average reward of at least $R_{goal} = 5$ per episode (where the reward is granted for target collection by the agent). The value of R_{goal} was determined experimentally. As shown in the training results (see Section "Dependence of learnability on image resolution"), an untrained agent, due to random wandering, attains on average no more than 2.

Training time of the agent until reaching the threshold value: $T(N) \in \mathbb{R}_+$.

Average reward $R(Res, T)$ achieved by the agent after training with input resolution $Res = h \times w \times 3$ over time T.

Admissible set of resolutions $Res \in \mathbb{N}$, from $20 \times 20 \times 3$ to $100 \times 100 \times 3$ with a step of 20 and with an additional case of $84 \times 84 \times 3$, used as the default resolution in Unity ML-Agents.

It is required to find $\min_{r \in Res} T(N)$, where $R(Res, T) \geq R_{goal}$, that is, the minimal training time over admissible resolution for which the achieved reward meets or exceed the threshold R_{goal}.

Simulator architecture. The proposed system employs a hierarchical control architecture for the agent, separating perception and motion functions across two levels. The lower-level agent (Walker) is responsible for physical locomotion in the environment, relying on local sensors and a pre-trained locomotion model. The higher-level agent (Searcher) receives visual input from the camera and decides on the movement direction, transmitting a control signal to the Walker agent in the form of a normalized rotation angle. This approach makes it possible to isolate the complex problem of sensorimotor transformation (from image to action) from the tasks of motion stabilization and limb coordination. As a result, training of the Searcher becomes faster and more stable, since it controls only a single variable. The internal communication between agents is implemented within the Unity environment through the transmission of the direction parameter to the Walker controller. In the training mode, the Searcher agent processes

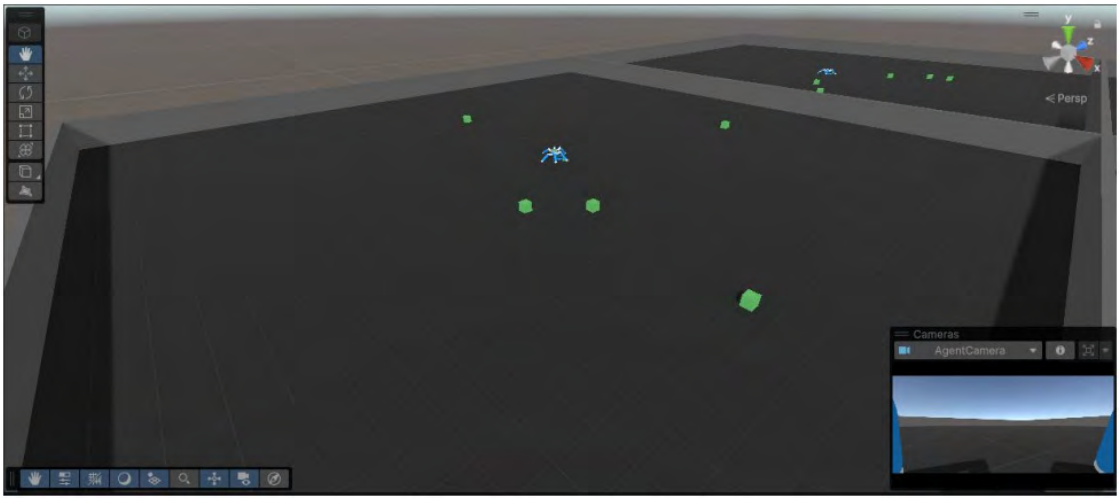


Fig. 1. Unity model of the Walker agent, with the first-person camera view shown in the bottom right corner. Two environments, the agents, and a number of targets are also presented.

visual data and generates a rotation angle, which is used as the control parameter for selecting the body orientation at the next step. The Walker, in turn, executes the specified direction, ensuring movement in the intended direction.

During simulation, the environment is dynamically updated: after a target (a unit of "food" required for survival) is collected by the Searcher agent, a new one is generated at a random position (to maintain the number of available "food" units at a constant level). When the agent falls or the maximum number of steps is reached, the episode is reset. The architecture supports parallel execution of multiple environments, each containing one Searcher and one Walker, which enables training to be scaled within the Unity ML-Agents framework.

Simulation environment. For the experiments, we selected the modern Unity ML-Agents platform, which demonstrates high performance and provides convenient tools for building complex three-dimensional simulations with reinforcement learning integration. Unity also offers built-in support for parallel environments, visual sensors, and integration with the PyTorch library.

Each environment represents a bounded square arena (DynamicPlatform) with walls, a floor, and randomly placed targets that the agent must collect. The platform size is fixed, and the target spawn coordinates are uniformly sampled across the available area. When the agent collides with a target, it disappears and is immediately replaced by a new one. The walls are impenetrable and serve as physical boundaries of the environment.

Simulation parameters are specified via the CrawlerSettings component and include the simulation tick rate of the physical world, gravity, episode duration (max_step – the number of simulation steps at which the agent receives observations and performs actions), and the number of parallel environments. If the agent falls (detected by body contact with the floor), the environment is automatically reset. Each parallel environment contains one Searcher agent, embedding a nested Walker, equipped with an individual camera mounted at the front of the head, which supplies the agent's neural network with a stream of first-person visual information.

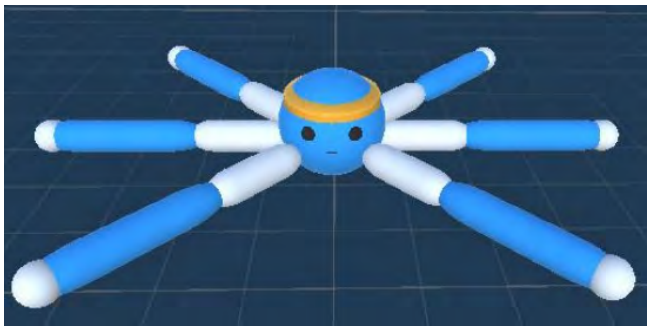


Fig. 2. Walker agent model in Unity.

The number of simultaneously running environments (num_envs) depended on the agent type: for the Walker agent, which does not use visual input, 10 environments were employed, while for the Searcher agent, four environments were used. This configuration enabled efficient utilization of GPU resources and accelerated data collection through parallel interaction with the environment. For each environment, actions data, observations, and rewards were collected independently and synchronized with the training strategy in Python via the Unity ML-Agents gRPC interface. Figure 1 presents a view of the simulation from the observer's perspective, showing two environments, the agents, and a number of targets.

Walker agent model. The lower-level agent (Walker) is a complex articulated model with six limbs, implemented in the Unity engine using the Rigidbody and ConfigurableJoint components. Each limb consists of two segments: upper and lower – with three degrees of freedom (resulting in a total of 18 degrees of freedom for all legs). This design enables the agent to perform realistic locomotion and maintain stability during movement. The agent model in the Unity environment is shown in Figure 2.

The control system is implemented through the JointDrive Controller module, which converts control signals into desired joint angles and forces. The control parameters are represented as a vector of dimension 30: 18 values control joint angles,

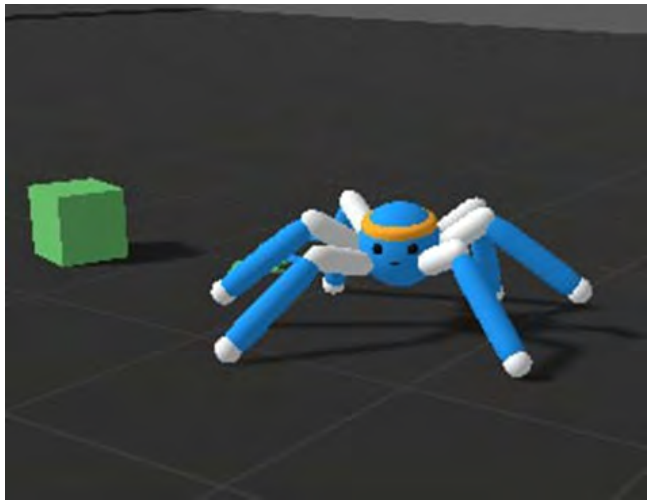


Fig. 3. Walker agent model in motion.

and 12 correspond to the torques applied to them. Specifically, for each of the six legs, the upper segment is controlled by two angles (rotation about the *X* and *Y* axes), and the lower segment by one angle (rotation about the *X* axis), yielding 18 control parameters in total. In addition, for each of these 12 segments, a control force is specified, determining the intensity of movement, which yields another 12 parameters. At each step, the agent receives observations that include information on current joint angles, velocities, surface contacts, target direction vector, body orientation, and ground raycast data. The Walker agent model in motion is shown in Figure 3.

The neural network architecture of the Walker consists of three fully connected layers with LeakyReLU activation functions and two outputs: an actor (30 action parameters) and a critic estimating the value function (Fig. 4*a*). The input layer has a dimensionality of 223 (vector features and joint parameters), while the hidden layers each contain 512 neurons. The total size of the model is 655,903 parameters and 1,567 neurons, making it lightweight enough for real-time training.

The reward function for the Walker agent is defined based on the deviation of the agent's current body velocity from the target velocity and the alignment of its movement direction with the specified vector. This enables the agent to learn purposeful locomotion in the desired direction while maintaining physical stability. After training, the Walker agent is used in inference mode as part of the Searcher agent, providing stable execution of movement.

During training, the critic block receives the same input as the actor – the state feature vector. Based on these data, it learns to approximate the expected cumulative reward the agent will obtain in the future if it continues to act according to the current policy. At the early stages of training, this estimate is inaccurate, but it is gradually refined through backpropagation of the error, grounded in the actual rewards received by the agent. Thus, the critic does not initially "know" what is good or bad – it learns to distinguish this by comparing predicted rewards with the real rewards accumulated during simulations.

After training, the Walker agent is used in inference mode as part of the Searcher agent, ensuring stable motion execution based on the deviation of the current body velocity from the target and the alignment of the movement direction with the specified vector. This allows the agent to learn purposeful locomotion in the desired direction while maintaining physical stability.

Searcher agent model. The higher-level agent (Searcher) is responsible for perceiving the environment and selecting the direction of body movement. Unlike the Walker agent, it does not interact directly with the physical components of the simulation but instead controls the Walker by transmitting a normalized rotation angle in the interval [−1, 1]. Thus, the Searcher serves as a cognitive module that implements a target-search strategy based on visual information. The primary input source for the Searcher agent is the image obtained

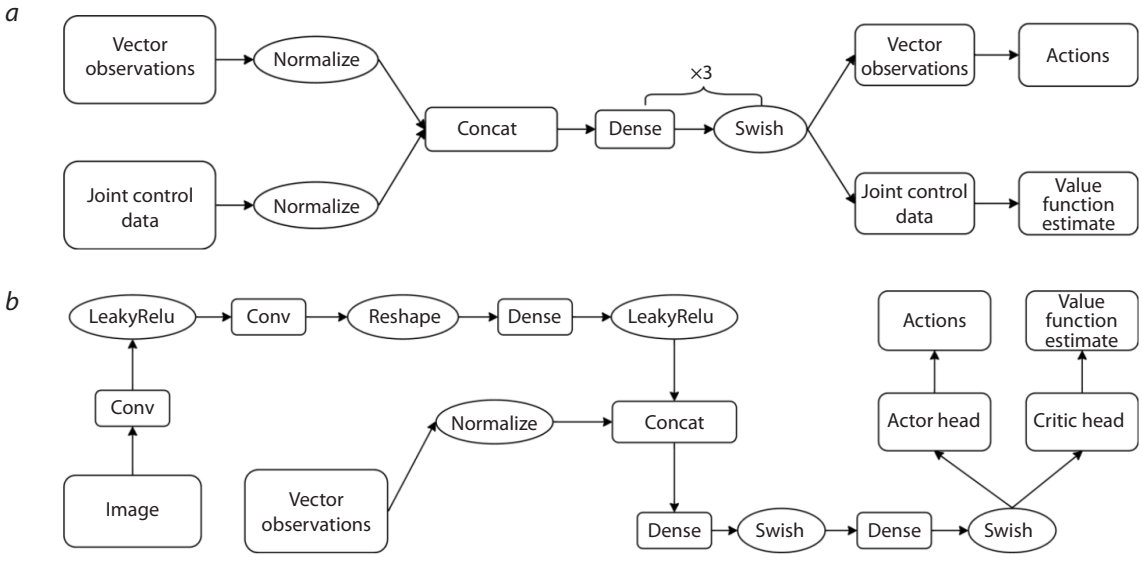


Fig. 4. Schematic representation of the Walker (*a*) and Searcher (*b*) agent's neural network architecture.

from a camera mounted on the agent's body (at the front of the head). The camera is oriented forward and positioned at a height corresponding to the head of the virtual organism. The image resolution varies across experiments from 20 × 20 to 100 × 100 pixels, with increments of 20 in each dimension (three-channel RGB), allowing for analysis of the impact of visual load and frame resolution on the model's learning performance.

For image processing, a convolutional neural network is employed, consisting of two convolutional layers (Conv2D), a flattening layer (Flatten), and subsequent fully connected layers. The output of the visual input processing is concatenated with vector observations and fed into two output layers: the actor (a single value representing the rotation angle) and the critic (value function estimate). The activation functions used are LeakyReLU and Swish. A schematic representation of the Searcher agent's neural network architecture is presented in Figure 4*b*.

The Searcher agent is trained using the Proximal Policy Optimization (PPO) algorithm with a continuous action space. The objective function is to maximize the cumulative reward for collecting targets in the arena. Upon colliding with a target, the agent receives a positive reward; upon colliding with a wall or remaining inactive, it is penalized. When max_step is exceeded or the body falls, the simulation episode terminates and a new one begins.

Unlike the Walker agent, which is pre-trained once and then used only to execute the learned behavior (inference mode), the Searcher agent is trained from scratch, and its neural network includes image processing, which increases computational costs but enables the realization of biologically plausible behavior based solely on visual perception. This makes it possible to model cognitive constraints and analyze the impact of visual resolution on the speed and stability of learning.

Training algorithms and hyperparameters. The PPO algorithm is a gradient-based policy optimization method that belongs to the family of actor-critic approaches. Such methods combine the training of a policy and a value function. By avoiding abrupt policy updates, in contrast to classical methods of this type, PPO is designed to improve the stability and reliability of training. The Actor, the component responsible for selecting an action in each state, implements the agent's policy. The Critic, in turn, evaluates how good the chosen action was by using the value function. This approach combines the advantages of stochastic action selection (important for exploration of the environment) with the evaluation of these actions based on accumulated experience.

The PPO algorithm operates within the framework of a Markov decision process (S, A, P, R, γ), where S – the set of states, A – the set of actions, $P(s'|s, a)$ – the state transition probability, $R(s, a)$ – the reward function, $\gamma \in [0, 1]$ – the discount factor.

The parameterized policy $\pi_\theta(a|s)$ defines the probability of selecting action a in state s, where θ signifies the parameters of the actor neural network. The critic $V_\phi(s)$ is an approximation of the value function $V^\pi(S) = E[R_t|s_t = s]$, with parameters ϕ, where $R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$ is the discounted sum of future rewards. In PPO, instead of direct gradient updates, the so-called clipped objective function is used:

$$L^{CLIP}(\theta) = \hat{E}_t[(r_t(\theta) \cdot \hat{A}_t, clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) \cdot \hat{A}_t)],$$

where: $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{Old}}(a_t|s_t)}$ – the probability ratio between the new and the old policy, $\varepsilon \in (0, 1)$ the clipping parameter, typically $\varepsilon = 0.1$ or 0.2, $\hat{A}_t$ – the advantage estimate.

If the new action deviates too strongly from the old one (i. e., r_t falls outside the interval $[1 - \varepsilon, 1 + \varepsilon]$), the gradient is suppressed. This prevents abrupt changes in the policy.

To estimate $\hat{A}_t$, the generalized advantage estimation (GAE) is used:

$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t),$$

where $\lambda \in [0, 1]$ – is the smoothing parameter. This method improves training stability by reducing variance.

The loss function in PPO consists of:

- the policy loss L^{CLIP},
- the value critic loss (MSE between the predicted $V(s_t)$ and the target value),
- an entropy bonus to encourage action diversity:

$$L_t^{CLIP+VF+S} = E_t[L^{CLIP}(\theta) - c_1 \cdot (V(s_t) - V_t^{target})^2 + c_2 \cdot H[\pi_\theta](s_t)],$$

where $H[\pi]$ is the policy entropy and c_1, c_2 are the corresponding coefficients.

A schematic representation of the proximal policy optimization algorithm is shown below:

Algorithm: PPO

for iteration = 1, 2, … **do**
  for actor = 1, 2, …, N **do**
    run policy $\pi_{\theta_{Old}}$ in environment for T timesteps
    compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  optimize surrogate L w.r.t. θ, with K epochs and minibatch size $M \leq NT$
  $L_t = E_t[L^{CLIP}(\theta) - c_1 \cdot (V(s_t) - V_t^{target})^2 + c_2 \cdot H[\pi_\theta](s_t)]$
  $\theta_{Old} \leftarrow \theta$
end for

where N is the number of parallel actors collecting data over T time steps, and K is the number of epochs. Neural networks are used to approximate the target policy and the value function.

The choice of PPO in this work is motivated by several factors: the algorithm supports continuous action spaces, which is critical for the locomotion of virtual organisms with multi-joint limbs. The update constraint allows the agent's policy to evolve incrementally without disrupting previously learned behaviors. PPO can also be effectively applied in architectures incorporating convolutional neural networks (CNNs) that process images from the agents' cameras. In addition, the Unity ML-Agents environment provides a built-in PPO implementation, which simplifies configuration and accelerates the cycle of computational experiments.

The actor network receives state features (velocities, joint positions, surface contacts, etc.) together with visual data processed through convolutional layers. The agent's objective is to maximize the reward associated with locomotion and stability while moving in the chosen direction. PPO enables

smooth adaptation of the policy to complex dynamics and noisy feedback from the environment.

For the Walker agent, the action space is represented by a vector of 30 continuous values (18 for joint angles and 12 for actuation forces/torques controlling joint movements), whereas the Searcher agent controls only a single parameter – the movement direction (a normalized rotation angle in the range [−1, 1]). Both models are trained asynchronously using multiple parallel environment simulations (from 4 to 10), which enables efficient data collection and accelerates the optimization process.

The main training parameters are (detailed in the documentation (Juliani et al., 2018)):

- algorithm: PPO (proximal policy optimization);
- framework: Unity ML-Agents + PyTorch backend;
- learning_rate: 3×10^{-4}. A coefficient that determines the step size when updating neural network parameters;
- batch_size – the size of the data batch used for one training step: Searcher: 1,024, Walker: 2,048;
- buffer_size: 10,240. The number of environment interactions used for one training cycle. Configured as a multiple of batch_size × num_envs;
- num_epochs: 3. The number of optimizer passes (epochs) over one data buffer before it is updated;
- gamma (discount factor): Searcher: 0.99, Walker: 0.995;
- lambda (GAE): 0.95;
- clip_range: 0.2.

The Walker agent was trained separately in an isolated environment until stable and straight locomotion was achieved. The average number of steps to convergence was approximately 2–3 million. After this stage, the model weights were fixed, and the agent was used only in inference mode.

The Searcher agent was trained independently of the Walker. The average number of steps per experiment ranged from 5 to 10 million, depending on the environment configuration (camera resolution, max_step, number of target objects in the environment, etc.).

Simulation parameters were specified through YAML configurations of ML-Agents. To ensure stable and reproducible results, a fixed parameter was used to set the initial value for the random number generator applied in both the environment and training (random_seed), along with consistent settings: when the number of environments (num_envs) was changed, buffer_size was necessarily adjusted proportionally, as required by the ML-Agents framework.

All experiments were conducted on a computer equipped with a CUDA-compatible GPU (see Section “System performance and profiling”). The software versions used were: Unity 2022.3, ML-Agents 21.0, PyTorch 2.0.1, and Python 3.10.

Experiments. The experimental part of the study (numerical experiments) was aimed at investigating the influence of visual system parameters, environment configuration, and architectural constraints on the training efficiency of agents. All experiments were carried out in isolated environments using a fixed Walker agent model and a trainable Searcher agent. The main directions of investigation were as follows:

1. Impact of camera image resolution on learnability. A range of resolutions was considered: 20×20, 40×40, 60×60, 80×80, 84×84 (the default resolution for Unity ML-Agents), and 100×100 pixels. For each of these, a separate training of the Searcher was conducted under otherwise identical parameters. The objective was to determine the minimal resolution at which the agent consistently achieves the target behavior (Reward ≥5).
2. Impact of speed control capability. In one of the experiments, the Searcher agent was additionally given the ability to control the target movement speed (a second continuous output parameter). The objective was to determine whether this would lead to more flexible behavior or instead complicate the learning task.
3. Variation of maximum rotation angle. The Searcher agent transmits a body rotation command. In different experiments, the maximum allowable rotation angles were tested: 90, 120, 180, and 270°. The hypothesis examined was that larger angles may simplify navigation but make the behavior less precise and stable.
4. Impact of episode length (max_step parameter). In the experiments, two values of the max_step parameter were considered: 5,000 and 20,000. The value max_step = 5,000 was used as the baseline, as it allowed the agent to receive rewards quickly enough and provided timely feedback to the learning algorithm. The value 20,000 was considered as an alternative, applicable to tasks with longer action sequences and delayed rewards.
5. Verification with manual control. To validate the behavior of the trained Walker model, manual control of the agent was implemented (via the A/D keys, left/right). This made it possible, on the one hand, to confirm that the observed effects (e. g., halting of movement) were caused by body dynamics rather than the Searcher agent’s policy, and on the other hand, to test whether a human, using the same type of control, could successfully perform the target-search task (an assessment of controllability and environment perception).

All experiments were recorded using the Unity ML-Agents logging system and analyzed in TensorBoard, a visualization tool for monitoring the training process that allows real-time plotting of reward dynamics, loss functions, simulation speed, and other metrics. The success criteria are described in Section “Problem statement”.

Results

Dependence of learnability on image resolution

The results of the series of experiments with different input image resolutions showed that the minimal resolution at which the agent consistently achieved the target behavior (average reward ≥5) was 84×84 pixels. At resolutions of 20×20, 40×40, and 60×60, training required substantially more time, although the trend toward improvement was preserved. The resolution of 100×100 also allowed the target reward to be reached, but training at 84×84 was slightly faster due to lower computational load. The results of this experiment are presented as TensorBoard plots in Figure 5.

Impact of speed control on training

The addition of a second control parameter (movement speed) increased the dimensionality of the action space and significantly complicated training. The agent required more time to

converge (approximately 33 % longer under otherwise identical conditions), and the resulting behavior was less stable – for the given task, speed control is largely a redundant parameter. This supports the simple hypothesis that increasing the number of degrees of freedom requires a more complex policy and hinders model training. The results of this experiment are shown as a TensorBoard plot in Figure 6.

Results with varying rotation angles

The best results were obtained with a maximum rotation angle of 90°. Increasing the angle to 120° led to a slight decrease in stability, while at 180 and 270°, the agent did not reach the target reward level, requiring longer and less efficient training. This indicates that an excessively wide action space hinders the development of a stable navigation policy.

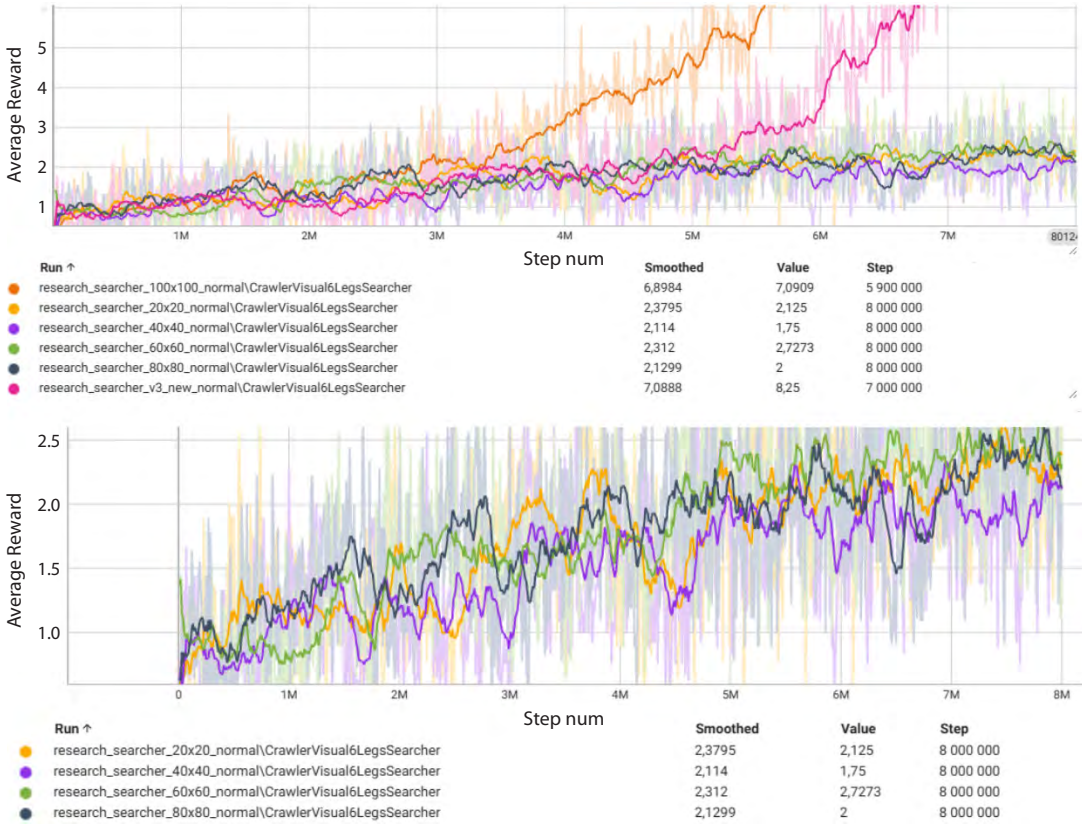


Fig. 5. Training results of the Searcher agent at different input image resolutions.

The upper panel shows the average reward curves for all investigated resolutions; the magenta curve corresponds to 84×84, and the orange curve to 100 × 100. The lower panel presents the same data with the dominant curves removed, allowing a more detailed view of the remaining variants (20×20, 40×40 и 60×60, 80×80).

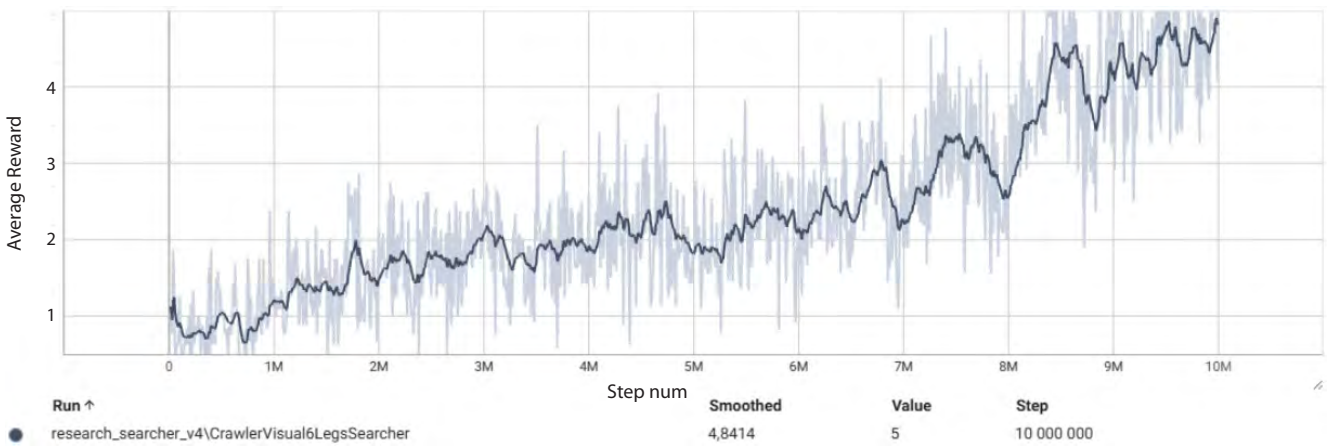


Fig. 6. Training results of the Searcher agent with input image resolution 84×84 and the addition of a second control parameter (movement speed) alongside the primary one (rotation angle).

Impact of episode length (max_step)

With max_step = 5,000, the agent demonstrated stable training, receiving timely feedback on goal achievement. Increasing the episode length to 20,000 did not improve training quality, while simulation time and resource load increased. Therefore, max_step = 5,000 was used as the primary setting, as it provided a balance between training efficiency and computational cost.

System performance and profiling

To evaluate the scalability and computational efficiency of the simulator, profiling of key system components was conducted under varying visual sensor resolutions and numbers of parallel environments. All measurements were performed on a machine equipped with an NVIDIA GeForce RTX 3070 GPU and an AMD Ryzen 5 7500F CPU (6 cores, 12 threads, 3.7 GHz base clock, 5.0 GHz in turbo mode).

The contribution of main simulation components to computational costs:

- Physics Engine – less than 1 ms per step, virtually independent of resolution;
- graphics and sensors (Camera.Render, PostProcess) – from 3.2 to 9.5 ms depending on resolution (almost linear dependence);
- neural network (PyTorch Inference) – approximately 35 ms per step when using convolutional architecture for Searcher;
- Unity–Python communication (gRPC, serialization) – from 45 to 60 ms. With an increasing number of agents, this component becomes one of the main system bottlenecks, since communication costs (serialization/deserialization, data exchange) grow proportionally to the number of agents;
- other (UI, garbage collection, VSync) – up to 20 % of runtime, may increase during active debugging.

At a resolution of 84 × 84 with four parallel agents, the average simulation step time was approximately 3.6 ms, corresponding to about 278 steps per second. At a resolution of 100 × 100, the step time increased to 3.8 ms, reducing performance to roughly 263 steps per second. All measurements were conducted without scene visualization. In all experiments with the Searcher agent, the number of simultaneously running environments was set to 4.

Thus, the main limiting factor in scaling is not physics or rendering, but data exchange between Unity and Python. This should be considered when planning large-scale experiments or transitioning to population-level modeling. A working prototype for reproducing the results is available in the repository at: https://github.com/DerpyFox/organism_simulator.

Discussion

Results interpretation

The obtained results demonstrate that the success of training agents with visual perception directly depends on the resolution of the input image. Too low a resolution (up to 60 × 60) leads to a loss of spatial structure of the scene and the agent's inability to develop a stable strategy. On the other hand, resolutions above 84 × 84 do not provide a noticeable gain in efficiency but increase the computational load. This confirms the existence of an optimal range of visual perception, comparable to that evolutionarily formed in insects: their vision developed to be sufficiently detailed for performing behavioral tasks (Chittka, Niven, 2009).

Despite the observed dependence between visual system resolution and the success of agent training, it should be noted that in nature there are organisms capable of effective behavior even with extremely low visual resolution. For example, in some ant species, as mentioned in the introduction, the visual system is comparable in scale to a resolution of about 10 × 10, yet this does not prevent them from confidently navigating, locating food, interacting with their environment, and even passing the mirror test (Cammaerts M.-C.T., Cammaerts R., 2015). Such efficiency is determined not only by vision but also by the developed olfactory system, which plays a key role in perceiving the surrounding world. In addition, the neural systems of real insects may possess a range of properties that enhance their effectiveness. These systems were shaped through long evolutionary processes and are adapted to specific living conditions and the typical tasks of a living organism – for example, navigating in complex environments, searching for food, and interacting with conspecifics. They exhibit a high degree of neuronal specialization and mechanisms of adaptation to changing stimuli. Such "tuning" to real-world conditions makes it possible to efficiently process even limited or fragmentary sensory signals, including visual, olfactory, and mechanosensory inputs.

The addition of speed control and the increase in rotation angle showed that even a slight expansion of the action space leads to slower learning. Thus, it is important to maintain a balance between the expressiveness of the model and its learnability. The division of perception and body control tasks between the Searcher and Walker agents proved to be critical for achieving stable behavior.

Biological parallels and cognitive efficiency

The results resonate with principles observed in insects: minimal but functionally redundant visual systems enable successful navigation and real-time decision-making. Similarly, the proposed architecture allows the agent to achieve target strategies with limited resolution and a relatively small neural network.

When the obtained results are considered in the context of real biological systems, a parallel can be drawn with the evolutionary trade-offs that arise between sensory accuracy, computational cost, and behavioral adaptability. For example, the visual systems of insects such as fruit flies (~150,000 neurons) or honeybees (~960,000 neurons) provide basic object recognition and spatial orientation with a minimal number of neurons and extremely limited bandwidth (Menzel, 2012). These organisms do not possess high-resolution visual systems, but they achieve high efficiency through a combination of rapid response, sensorimotor architecture, and decision-making strategies (Chittka, Niven, 2009). Such considerations are well illustrated by insects with a high level of social organization. In ants, division of labor and communication are shaped not only as innate behavioral patterns but also as the result of flexible adaptation at the level of individual workers. The distribution of roles within a colony may vary depending on age, physiological state, and the current situation, while

information transfer between ants is achieved through a wide range of signals (Chittka, Muller, 2009). Thus, even simple agents with limited cognitive capacities can achieve high efficiency through the organization of interactions and simple behavioral rules.

Agents in our simulator demonstrate the ability for adaptive behavior even at relatively low visual resolutions (e. g., 84 × 84 pixels), which allows further analogies to be drawn with minimal cognitive systems in nature. Such models can be employed as artificial systems that reproduce key behavioral aspects of simple organisms and serve as a basis for generating hypotheses about the neurophysiological mechanisms of perception and behavior in invertebrates.

System limitations

The main limitation of the system lies in the communication overhead between the Unity environment and the PyTorch training framework. Even with high computational performance of the processing units, serialization and data transfer via gRPC become the bottleneck. In addition, at this stage, the environment remains limited in complexity: it lacks obstacles, dynamic topography, and inter-agent interactions. Finally, the agent architectures are fixed and do not undergo evolution or temporal adaptation (only parameter weights change, while network topology remains unchanged).

Future directions

Further development is possible in several directions. The introduction of neuroevolutionary mechanisms (e. g., the NEAT approach – NeuroEvolution of Augmenting Topologies) would make it possible to investigate not only changes in neural network weights but also the evolutionary optimization of network structure. This is particularly relevant in the context of energy costs: with excessive brain complexity, resource consumption increases, whereas in simpler environments it may be advantageous to reduce the number of neurons. In this way, agents could autonomously adapt the size and potentially the architecture of their neural networks, reducing redundancy under conditions of low cognitive load. In biological systems, even a slight increase in nervous system complexity can lead to a noticeable rise in energy expenditure. For example, in the fly *Calliphora vicina*, the retina alone consumes about 8 % of the organism's resting metabolic rate (Niven, Laughlin, 2008). In humans, by contrast, the brain accounts for only about 2 % of body mass yet consumes up to 20 % of the body's energy (Attwell, Laughlin, 2001). These data indicate that the benefit of reducing the number of neurons or decreasing the complexity of the sensory system can be substantial.

Introducing environmental elements involving resource competition (multiple agents, a limited number of targets, and the ability of more advanced agents to select and solve more complex cognitive tasks from those available in the system, thereby gaining additional advantages) would make it possible to analyze behavioral strategies at the population level.

A promising direction is the addition of an olfactory model – a sensory channel based on short-term "traces" in the environment, analogous to pheromone markings in ants. Such traces may decay over time, differ in content (e. g., distinguishing between a satiated and a hungry ant), and influence an agent's trajectories, thereby reinforcing elements of indirect communication and collective behavior. It would also be reasonable to incorporate memory and recurrent modules into the Searcher model to study navigation under partial observability.

Conclusion

This study was aimed at the quantitative and qualitative evaluation of architectural and sensory parameters in the task of training visually guided agents in a three-dimensional simulation. We proposed and implemented a hierarchical control model in which the locomotion agent (Walker) functions as a low-level executor of movements, while the perception and navigation agent (Searcher) makes strategic decisions based on visual information.

A systematic analysis demonstrated that even under limited sensory input (due to low resolution), agents are capable of developing stable behavioral strategies, provided that the model and environmental conditions are designed with cognitive load in mind. It was established that a resolution of 84 × 84 pixels offers a compromise between computational efficiency and minimal cognitive adequacy, whereas increasing the dimensionality of the action space without a corresponding increase in training resources leads to degraded performance.

Our results support the hypothesis that minimally complex neural network agents can realize sophisticated behavioral patterns under conditions of limited sensory perception, where the agent receives only partial information about the environment. These findings are consistent with observed examples of cognitive efficiency in invertebrates, such as ants and bees, and open up prospects for the use of such models in biological modeling, robotics, and research in the field of neuroevolution.

In the future, the system may be extended toward population-level simulations incorporating competition, inter-agent communication, and strategy adaptation in a changing environment. The architecture can be further enhanced with memory modules, recurrent connections, or neuroevolutionary mechanisms, enabling the study of more complex cognitive phenomena in virtual populations.

It was also shown that the use of visual information, despite its expressiveness, requires substantial computational resources and, in some cases, may be less efficient than simpler sensory models. These observations highlight the importance of sensory architecture choice when designing minimally sufficient cognitive agents.

Another key finding was the recognition of the critical role of environment design and training structure in the success of modeling. Initial attempts to train behavior through a single neural network model that combined locomotion and strategy did not lead to the emergence of the ability to detect and collect targets ("food" units), due to difficulties in balancing rewards and formulating the task. The introduction of a functionally separated approach (search and locomotion) made it possible to achieve a substantial improvement in learnability and behavioral stability.

Thus, the obtained results demonstrate the potential of neuro-agent systems in biologically inspired modeling tasks and provide a foundation for further experiments aimed at exploring the limits of cognitive complexity under constrained perceptual and control resources.

References

Aksoy V., Camlitepe Y. Spectral sensitivities of ants – a review. *Anim Biol*. 2018;68(1):55-73. doi 10.1163/15707563-17000119

Attwell D., Laughlin S.B. An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab*. 2001;21(10):1133-1145. doi 10.1097/00004647-200110000-0000

Bongard J.C. Evolutionary robotics. *Commun ACM*. 2013;56(8):74-83. doi 10.1145/2493883

Botvinick M., Wang J.X., Dabney W., Miller K.J., Kurth-Nelson Z. Deep reinforcement learning and its neuroscientific implications. *Neuron*. 2020;107(4):603-616. doi 10.1016/j.neuron.2020.06.014

Cammaerts M.C. The visual perception of the ant *Myrmica ruginodis* (Hymenoptera: Formicidae). *Biologia*. 2012;67(6):1165-1174. doi 10.2478/s11756-012-0112-z

Cammaerts M.-C.T., Cammaerts R. Are ants (Hymenoptera, Formicidae) capable of self recognition? *J Sci*. 2015;5(7):521-532

Chittka L., Muller H. Learning, specialization, efficiency and task allocation in social insects. *Commun Integr Biol*. 2009;2(2):151-154. doi 10.4161/cib.7600

Chittka L., Niven J. Are bigger brains better? *Curr Biol*. 2009;19(21):R995-R1008. doi 10.1016/j.cub.2009.08.023

Devyaterikov A.P., Palyanov A.Y. A software system for modeling evolution in a population of organisms with vision, interacting with each other in 3D simulator. *Vavilov J Genet Breed*. 2022;26(8):780-786. doi 10.18699/VJGB-22-94

Hassabis D., Humaran D., Summerfield C., Botvinick M. Neuroscience-inspired artificial intelligence. *Neuron*. 2017;95(2):245-258. doi 10.1016/j.neuron.2017.06.011

Juliani A., Berges V.-P., Teng E., Cohen A., Harper J., Elion C., Goy C., Gao Y., Henry H., Matter M., Lange D. Unity: A general platform for intelligent agents. *arXiv*. 2018. doi 10.48550/arXiv.1809.02627

Land M.F., Nilsson D.-E. Animal Eyes. Oxford University Press, 2012. Available: https://www.softouch.on.ca/kb/data/Animal%20Eyes.pdf

Menzel R. The honeybee as a model for understanding the basis of cognition. *Nat Rev Neurosci*. 2012;13(11):758-768. doi 10.1038/nrn3357

Niven J.E., Laughlin S.B. Energy limitation as a selective pressure on the evolution of sensory systems. *J Exp Biol*. 2008;211:1792-1804. doi 10.1242/jeb.017574

O'shea K., Nash R. An introduction to convolutional neural networks. *arXiv*. 2015. doi 10.48550/arXiv.1511.08458

Peng P., Wen Y., Yang Y., Yuan Q., Tang Z., Long H., Wang J. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play StarCraft combat games. *arXiv*. 2017. doi 10.48550/arXiv.1703.10069

Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms. *arXiv*. 2017. doi 10.48550/arXiv.1707.06347

Song K.-Y., Behzadfar M., Zhang W.-J. A dynamic pole motion approach for control of nonlinear hybrid soft legs: A preliminary study. *Machines*. 2022;10(10):875. doi 10.3390/machines10100875

Stanley K.O., Clune J., Lehman J., Miikkulainen R. Designing neural networks through neuroevolution. *Nat Mach Intell*. 2019;1(1):24-35. doi 10.1038/s42256-018-0006-z

Tschantz A., Anil K.S., Christopher L.B. Learning action-oriented models through active inference. *PLoS Comput Biol*. 2020;16(4):e1007805. doi 10.1371/journal.pcbi.1007805

Vezhnevets A.S., Osindero S., Schaul T., Heess N., Jaderberg M., Silver D., Kavukcuoglu K. FeUdal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70 (ICML'17). JMLR.org, 2017;3540-3549

Zador A.M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun*. 2019;10(1):3770. doi 10.1038/s41467-019-11786-6

Conflict of interest. The authors declare no conflict of interest.

Received July 30, 2025. Revised September 23, 2025. Accepted September 23, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1062-1072

doi 10.18699/vjgb-25-111

Molecular dynamic analysis of the functional role of amino acid residues V99, F124 and S125 of human DNA dioxygenase ABH2

M. Zhao [1], T.E. Tyugashev [2], A.T. Davletgildeeva [2], N.A. Kuznetsov [1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
kuznetsov@1bio.ru

Abstract. The ABH2 enzyme belongs to the AlkB-like family of Fe(II)/α-ketoglutarate-dependent dioxygenases. Various non-heme dioxygenases act on a wide range of substrates and have a complex catalytic mechanism involving α-ketoglutarate and an Fe(II) ion as a cofactor. Representatives of the AlkB family catalyze the direct oxidation of alkyl substituents in the nitrogenous bases of DNA and RNA, providing protection against the mutagenic effects of endogenous and exogenous alkylating agents, and also participate in the regulation of the methylation level of some RNAs. DNA dioxygenase ABH2, localized predominantly in the cell nucleus, is specific for double-stranded DNA substrates and, unlike most other human AlkB-like enzymes, has a fairly broad spectrum of substrate specificity, oxidizing alkyl groups of such modified nitrogenous bases as, for example, N^1-methyladenosine, N^3-methylcytidine, 1,N^6-ethenoadenosine and 3,N^4-ethenocytidine. To analyze the mechanism underlying the enzyme's substrate specificity and to clarify the functional role of key active-site amino acid residues, we performed molecular dynamics simulations of complexes of the wild-type ABH2 enzyme and its mutant forms containing amino acid substitutions V99A, F124A and S125A with two types of DNA substrates carrying methylated bases N^1-methyladenine and N^3-methylcytosine, respectively. It was found that the V99A substitution leads to an increase in the mobility of protein loops L1 and L2 involved in binding the DNA substrate and changes the distribution of π-π contacts between the side chain of residue F102 and nitrogenous bases located near the damaged nucleotide. The F124A substitution leads to the loss of π-π stacking with the damaged base, which in turn destabilizes the architecture of the active site, disrupts the interaction with the iron ion and prevents optimal catalytic positioning of α-ketoglutarate in the active site. The S125A substitution leads to the loss of direct interaction of the L2 loop with the 5′-phosphate group of the damaged nucleotide, weakening the binding of the enzyme to the DNA substrate. Thus, the obtained data revealed the functional role of three amino acid residues of the active site and contributed to the understanding of the structural-functional relationships in the recognition of a damaged nucleotide and the formation of a catalytic complex by the human ABH2 enzyme.

Key words: DNA repair; base methylation; human DNA dioxygenase ABH2; MD modeling; functional role of amino acid residues

For citation: Zhao M., Tyugashev T.E., Davletgildeeva A.T., Kuznetsov N.A. Molecular dynamic analysis of the functional role of amino acid residues V99, F124 and S125 of human DNA dioxygenase ABH2. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1062-1072. doi 10.18699/vjgb-25-111

Funding. The work was carried out within the framework of state assignment No. 121031300041-4.

Acknowledgments. The work was performed using the resources of the NSU Computing Center.

Молекулярно-динамический анализ функциональной роли аминокислотных остатков V99, F124 и S125 ДНК-диоксигеназы человека ABH2

М. Чжао [1], Т.Е. Тюгашев [2], А.Т. Давлетгильдеева [2], Н.А. Кузнецов [1, 2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
kuznetsov@1bio.ru

Аннотация. ДНК-диоксигеназа человека ABH2 относится к семейству AlkB-подобных негемовых диоксигеназ, которые действуют на широкий спектр субстратов и обладают сложным каталитическим механизмом с участием α-кетоглутарата и иона Fe(II) в качестве кофактора. Представители семейства AlkB катализируют прямое

окисление алкильных заместителей в азотистых основаниях ДНК и РНК, обеспечивая защиту от мутагенного воздействия эндогенных и экзогенных алкилирующих агентов, а также участвуя в регуляции уровня метилирования некоторых РНК. Фермент ABH2, локализованный преимущественно в ядре клетки, проявляет специфичность к двуцепочечным ДНК-субстратам и, в отличие от большинства других AlkB-подобных ферментов человека, обладает довольно широким спектром субстратной специфичности, окисляя алкильные группы таких модифицированных азотистых оснований, как, например, N^1-метиладенозин, N^3-метилцитидин, 1,N^6-этеноаденозин и 3,N^4-этеноцитидин. В данной работе с целью анализа механизма, обеспечивающего субстратную специфичность фермента, и выяснения функциональной роли аминокислотных остатков в составе активного центра нами выполнено молекулярно-динамическое моделирование комплексов фермента ABH2 дикого типа и его мутантных форм, содержащих аминокислотные замены V99A, F124A или S125A, с двумя типами ДНК-субстратов, несущих метилированные основания N^1-метиладенин или N^3-метилцитозин. Установлено, что замена V99A приводит к увеличению подвижности белковых петель L1 и L2, участвующих в связывании ДНК-субстрата, и изменяет распределение π-π-контактов боковой цепи остатка F102 с азотистыми основаниями, расположенными рядом с поврежденным нуклеотидом. Замена F124A приводит к потере π-π-стэкинга с поврежденным основанием, что, в свою очередь, дестабилизирует архитектуру активного центра, вызывает нарушение взаимодействия с ионом железа и препятствует оптимальному каталитическому позиционированию α-кетоглутарата в активном центре. Замена S125A приводит к потере прямого взаимодействия петли L2 с 5′-фосфатной группой поврежденного нуклеотида, ослабляя связывание фермента с ДНК-субстратом. Таким образом, полученные данные позволили установить функциональную роль трех аминокислотных остатков активного центра и расширить понимание структурно-функциональных связей в процессах узнавания поврежденного нуклеотида и формирования каталитического комплекса ферментом ABH2 человека.
Ключевые слова: репарация ДНК; метилирование оснований; ДНК-диоксигеназа человека ABH2; МД-моделирование; функциональная роль аминокислотных остатков

Introduction

The stability of genetic information encoded in the form of nucleotide sequences in DNA is extremely important for normal functioning and survival of individual cells, organisms, and species as a whole (Travers, Muskhelishvili, 2015). At the same time, cellular DNA of all living organisms is regularly subjected to damaging effects of various endogenous and exogenous factors, such as chemically reactive reagents and metabolites, ionizing and UV radiation, and others (Ougland et al., 2015). Living organisms evolved multiple different repair pathways for damage occurring in genomic DNA, some of which are represented by a single enzyme, while others involve sequential and coordinated work of entire enzymatic cascades (Yi et al., 2009; Li et al., 2013; Müller, Hausinger, 2015; Ougland et al., 2015).

So, among enzymes participating in recognition and removal of non-bulky individual damage to DNA nitrogenous bases, the following can be distinguished: 1) DNA glycosylases that remove damaged nitrogenous bases with the formation of apurinic/apyrimidinic sites in DNA, which are then processed with restoration of the original DNA structure by other enzymes of the base excision repair (BER) pathway (Ringvoll et al., 2006; Chen et al., 2010; Li et al., 2013); 2) O^6-alkylguanine-DNA-alkyltransferases (AGT) that transfer the alkyl adduct to their own cysteine residue (Ringvoll et al., 2006); 3) photolyases responsible for the removal of UV-induced photodamage such as cyclobutane pyrimidine dimers and pyrimidine-pyrimidine photoproducts (Yi, He, 2013); 4) dioxygenases of the AlkB family, belonging to the superfamily of Fe(II)/α-ketoglutarate(αKG)-dependent dioxygenases that use non-heme iron as a cofactor and αKG as a cosubstrate for direct oxidation of alkyl groups in damaged DNA bases (Yang et al., 2009; Yi et al., 2009; Kuznetsov et al., 2021). It should be noted that the diversity of repair pathways for non-bulky DNA lesions is related to the great diversity of possible chemical modifications of nitrogenous bases.

Representatives of the Fe(II)/αKG- dependent dioxygenase AlkB family found in humans have attracted particular interest in recent years due to their participation in the repair of alkylated DNA bases. It is believed that enzymes of this family may play an important role in the progression of some oncological diseases since they are often overexpressed in tumor cells and neutralize the effect of alkylating drugs used in chemotherapy. ABH2 is one of the first identified human representatives of the AlkB-like dioxygenase family (Duncan et al., 2002; Aas et al., 2003). It is known that changes in ABH2 expression levels affect the efficiency of removal of certain toxic DNA damages in tumor cells, making this enzyme a potential marker for cancer diagnostics and a possible therapeutic target (Wilson et al., 2018).

To date, it is known that ABH2 exhibits activity against at least 8 different alkylated DNA bases, namely N^1-methyladenosine (m^1A), N^3-methylcytidine (m^3C), N^3-methylthymidine (m^3T), N^3-ethylthymidine (N^3-EtT), 1,N^6-ethenoadenosine (εA), 3,N^4-ethenocytidine (εC), 1,N^2-ethenoguanosine (1,N^2-εG), and 5-methylcytidine (m^5C) (Fig. 1) (Falnes, 2004; Ringvoll et al., 2006, 2008; Bian et al., 2019).

Methylation is the most common type of DNA base damage caused by exposure to alkylating agents (Sall et al., 2022), and m^1A and m^3C are substrates most effectively removed by ABH2 from double-stranded DNA (dsDNA) (Duncan et al., 2002; Aas et al., 2003; Xu et al., 2021). D.H. Lee et al. showed that ABH2 oxidizes m^1A and m^3C in the context of dsDNA at least twice as efficiently compared to single-stranded DNA (ssDNA) (Lee et al., 2005).

Currently known structural data allow suggestion of specific features of ABH2 enzyme functioning and the mechanism providing its substrate specificity. ABH2 contains a highly conserved catalytic domain – a double-stranded β-helical domain (DSBH) of the Fe(II)/αKG-dependent dioxygenase superfamily. The unstructured N-terminal fragment of ABH2 also includes a proliferating cell nuclear antigen (PCNA)

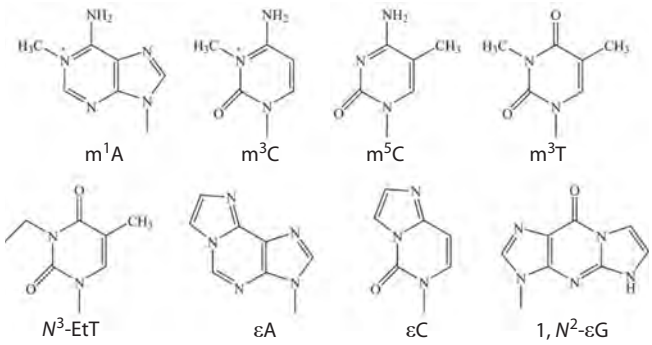


Fig. 1. Alkylated nitrogenous bases that are substrates for human DNA dioxygenase ABH2.

binding motif (Xu et al., 2021). A triplet consisting of two histidine amino acid residues and one aspartate (H171, H236, and D173) coordinates the Fe(II) cofactor in the enzyme's active site (Giri et al., 2011; Xu et al., 2021). D173 amino acid residue, through interaction with R254 residue, also participates in formation of a hydrogen bond network including N159, Y161, R248, T252, and R254 amino acid residues, that coordinate the αKG cosubstrate in the enzyme's active site (Waheed et al., 2020).

The ABH2 active site is surrounded by four functional loops, L1-L4 (Fig. 2). These loops play a key role in stabilizing the position of the DNA substrate in the enzyme's active site (Xu et al., 2021). Loop L1, including amino acid residues 98–107, contains a hydrophobic hairpin V101-F102-G103, through which "testing" of base pair stability in the substrate occurs. If a damaged base forms an unstable pair with its partner from the complementary strand, V101 and F102 residues induce flipping of the damaged nucleotide into the active site. Herewith the vacated space in the DNA duplex is filled by F102 residue, stabilizing the flipped-out position of the nucleotide through π-π interaction with the surrounding bases (Chen et al., 2010, 2014; Yi et al., 2012; Xu et al., 2021).

Loop L2, including amino acid residues 122–129, together with loop L1 forms the so-called "nucleotide recognition lid" (NRL). Y122 amino acid residue participates in a hydrogen bond network forming the catalytically competent state of the enzyme's active site (Davletgildeeva et al., 2023), S125 residue forms a hydrogen bond with the 5′-phosphate of the flipped damaged nucleotide; F124 and H171 amino acid residues form π-π stacking with the flipped nitrogenous base (Chen et al., 2010, 2014; Lenz et al., 2020). S125 amino acid residue also participates in forming the wall of the damage-binding pocket alongside V99, R110, and I168 residues (Davletgildeeva et al., 2023).

It should be noted that V99 holds an important position in the network of hydrophobic residues formed by V101, V108, F124, and L127 (Monsen et al., 2010). Loop L3, including amino acid residues 198–213, and loop L4, including amino acid residues 237–247, play an important role in binding to the dsDNA substrate. R198, R203, and K205 amino acid residues in loop L3 and the RKK sequence (R241-K242-K243) in loop L4 form contacts with the DNA strand complementary to the

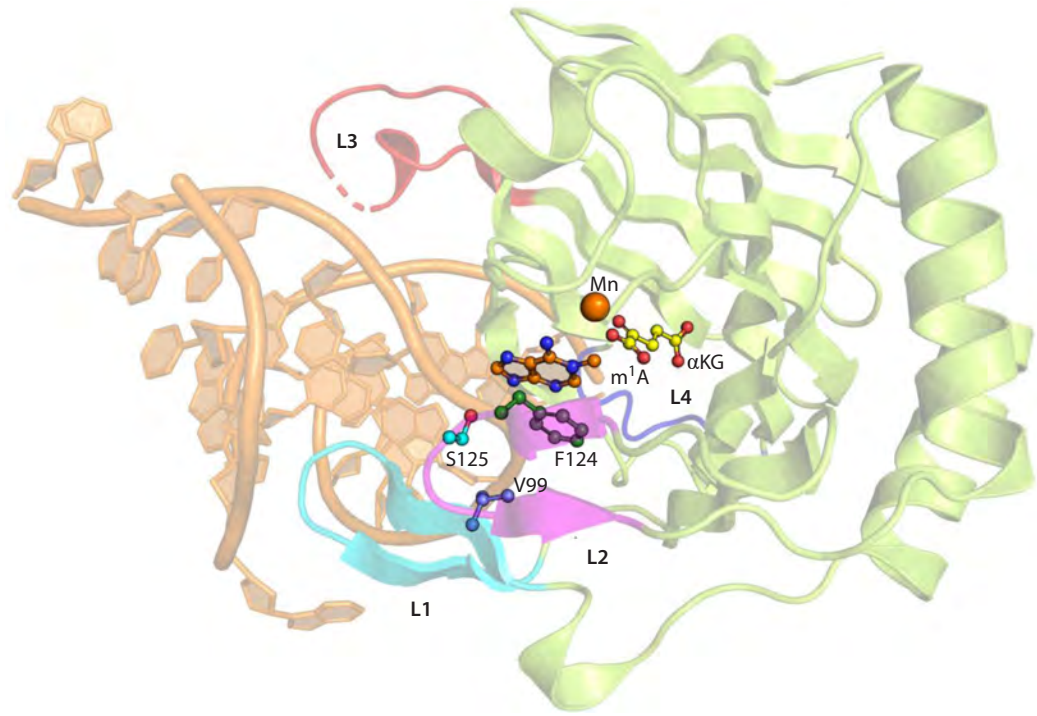


Fig. 2. Crystal structure of the ABH2 complex with dsDNA containing m^1A (PDB ID 3BUC).

Loops L1–L4 are marked, damaged nitrogenous base m^1A, αKG and Mn^{2+} ion, as well as the amino acid residues considered in this work (V99, F124, and S125) are shown.

damaged strand, thereby ensuring effective binding of the dsDNA substrate by the ABH2 enzyme (Yang et al., 2008; Yi et al., 2009; Waheed et al., 2020).

Molecular dynamic analysis of structural data and experimental verification of activity of recombinant preparations of wild-type ABH2 and several of its mutant forms, conducted by our group previously, allowed establishment of the role of Y122, I168, and D173 amino acid residues, which form direct contacts with bases m^1A, m^3C, as well as m^5C, in the active site pocket (Davletgildeeva et al., 2023). Comparative analysis of enzymes revealed the influence of substitutions of these amino acid residues on the enzyme's catalytic activity, and only a slight decrease in DNA binding efficiency. The obtained data suggested that these residues are responsible for precision positioning of the flipped damaged nucleotide in the active site, which ensures effective catalytic reaction (Davletgildeeva et al., 2023).

It should be noted that the broad spectrum of substrate specificity of the ABH2 enzyme and the complex catalytic mechanism of action, including cofactor and cosubstrate for reaction implementation, complicate detailed studies of the molecular mechanism of damaged DNA recognition and catalytically competent complex formation as well as local conformational changes affecting catalytic reaction efficiency. Due to the above, in the present study, with the aim of further elucidating the mechanism of substrate specificity of human DNA dioxygenase ABH2 using molecular dynamics methods, analysis of the functional role of three amino acid residues, V99, F124, and S125, participating in the formation of the pocket where the flipped-out damaged nucleotide is located, was conducted.

Materials and methods

Complex models were built based on crystallographic structures of the ABH2-dsDNA complexes with metal ion (Mn^{2+}) and αKG: 3BUC (for m^1A), and 3RZJ (for m^3C) (Yang et al., 2008; Yi et al., 2012). DNA sequence changes, correction of unresolved amino acid residues and enzyme modifications were performed using Chimera and Modeller (Šali, Blundell, 1993), protonation optimization of ionizable groups was done using the H++ server (Anandakrishnan et al., 2012). MD modeling was performed in GROMACS (Abraham et al., 2015). The complex was placed in a dodecahedral cell with TIP3P water and 50 mM KCl (Jorgensen et al., 1983; Joung, Cheatham, 2008), the AMBER14SB/OL15 force field was used to describe the complex (Cornell et al., 1996; Zgarbová et al., 2011, 2015; Maier et al., 2015).

Parameterization for m^1A, m^3C and αKG was performed using the Antechamber module (AMBER package), RESP charges were calculated on the REDD server, topologies of modified residues were converted to GROMACS format using ACPYPE (Bayly et al., 1993; Wang et al., 2004, 2006; Vanquelef et al., 2011; Sousa da Silva, Vranken, 2012).

In order to preserve octahedral coordination geometry of Fe^{2+} ion under possible active site perturbations introduced by amino acid residue substitutions, a distributed charge model was used to describe the ion (Jiang et al., 2016). The following parameters were used for MD calculations: system energy minimization by the steepest descent method, van der Waals interaction cutoff value set to 10 Å, long-range Coulomb interactions accounted for by the PME (Particle Mesh Ewald) method (Essmann et al., 1995), hydrogen atom covalent bond vibration restriction by the LINCS method (Hess et al., 1997).

After minimization, the system was heated to 310 K in NVT ensemble for 500 ps using a V-rescale thermostat (Bussi et al., 2007). Then equilibration in NPT ensemble was performed for 1 ns, pressure was maintained at 1 bar using a Parrinello–Rahman barostat (Parrinello, Rahman, 1981).

Classical molecular dynamics calculations were performed for 250 ns duration at least three times. Trajectory analysis was performed using built-in GROMACS tools and the MDTraj library (McGibbon et al., 2015). Distribution changes between stable states of wild-type ABH2 enzyme complexes and its mutant forms with DNA substrates are shown in distance distribution graphs between key atoms during modeling. Interatomic distance distributions in MD trajectory are presented as histograms with 0.1 Å step and step height equal to the percentage of trajectory frames in which the distance falls within the corresponding range of values. For each trajectory, the sum of fractions across the entire distance range equals 100 %.

Results and discussion

Model of the ABH2 V99A enzyme-substrate complex with damaged DNA

When modeling enzyme-substrate complexes both with the m^1A-containing dsDNA substrate (hereafter m^1A-DNA, Fig. 3*a*, *b*), and with the m^3C-containing dsDNA substrate (hereafter m^3C-DNA, Fig. 3*c*, *d*), the V99A substitution led to changes both in the region of loops L1 and L2 interacting with the nucleotide flipped into the enzyme's active site and the adjacent dsDNA region, and in the cosubstrate binding region. Thus, in the model complex with m^1A-DNA, the side chain of F124 amino acid residue lost π-π stacking interactions with the base of the nucleotide flipped into the active site (Fig. 3*a*, *b*). This reduced the lifetime of the hydrogen bond between the hydroxyl group of Y122 and the exocyclic amino group of the damaged base (Fig. 4*a*). In the model complex with the m^3C-containing dsDNA substrate, partial loss of contact between the hydroxyl group of the Y122 side chain and the carboxyl group of the E175 side chain also occurred (Fig. 4*b*), which also disrupted the contact network stabilizing the flipped-out base.

The V99A substitution induced a change in the position of F102 residue, which intercalates into DNA and is part of loop L1. Herewith, in the complex with m^1A-DNA, redistribution of π-π contacts formed by F102 occurred from the nitrogenous base of the complementary strand in the wild-type enzyme (dG in Fig. 3*a*) to the nitrogenous base of the damaged strand in case of ABH2 V99A (dA in Fig. 3*b*).

The values of the dihedral angle C-Cα-Cβ-Cγ at F102 residue were 148.1 ± 55.3° for wild-type enzyme and 127.2 ± 47.7° for the V99A mutant form, indicating stability of these positions during molecular dynamics. Meanwhile, in the complex with m^3C-DNA, the V99A substitution induced a significant increase in the mobility of its side chain (dihedral angle C-Cα-Cβ-Cγ equals 135.6±58.6° and 100.2±100.3° for

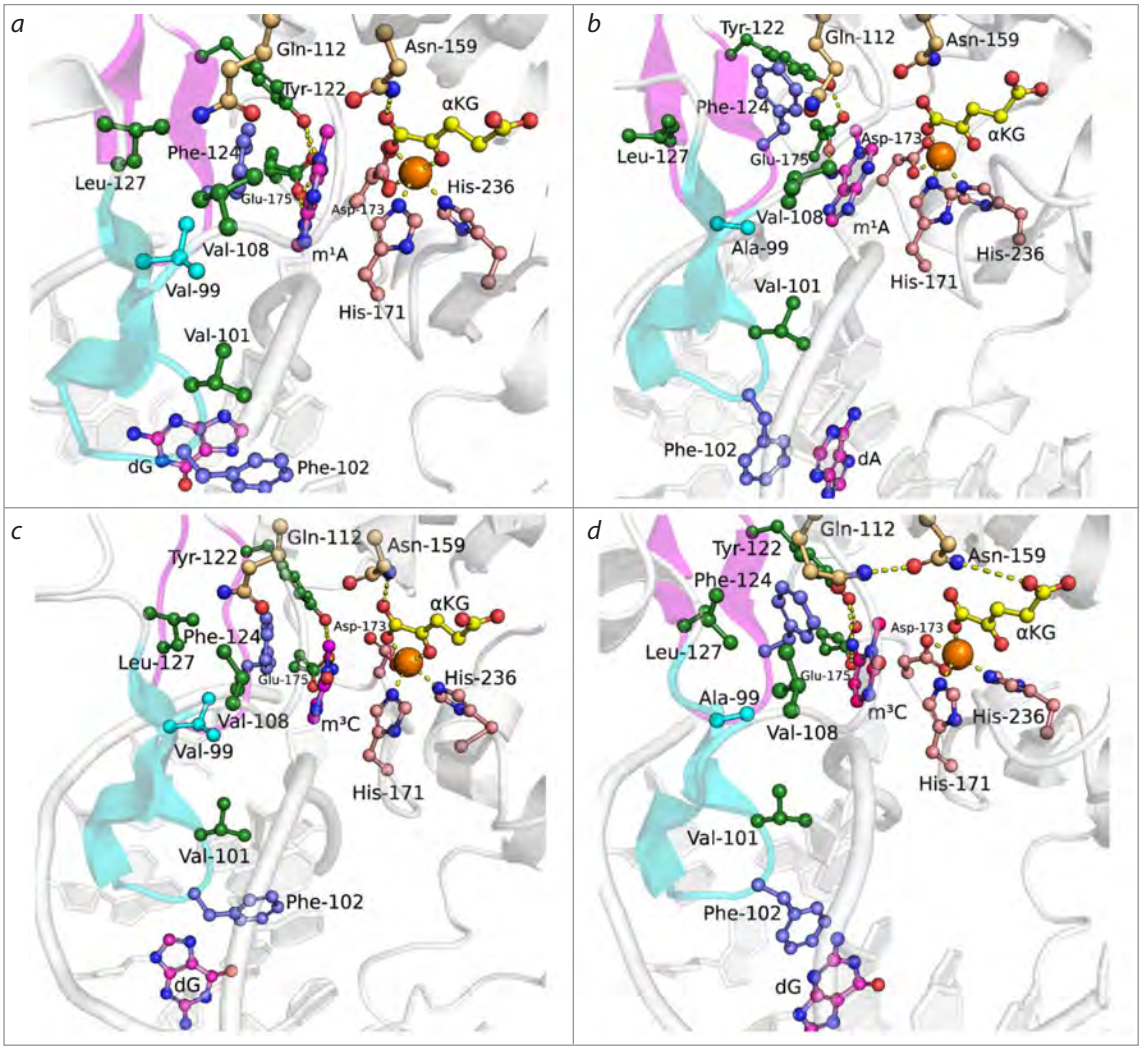


Fig. 3. Representative MD structures of ABH2 WT in complex with m^1A-DNA (*a*) and m^3C-DNA (*c*), and ABH2 V99A in complex with m^1A-DNA (*b*) and m^3C-DNA (*d*).

Key amino acid residues of the active site, damaged nitrogenous base, αKG and Mn^{2+} ion are shown. Loops L1 (blue) and L2 (pink) are highlighted with corresponding colors.

the wild-type enzyme and V99A, respectively). Increased mobility of F102 residue led to guanine complementary to m^3C (dG in Fig. 3*c*, *d*) acquiring the opportunity to return inside the DNA structure in the mutant enzyme complex, entering into π-π contact with the side chain of F102, while this guanine was completely flipped out from the DNA double strand in the wild-type enzyme complex.

The V99A substitution also induced changes in interaction with the cosubstrate, which led to αKG adopting a catalytically unfavorable conformation for half of the total modeling trajectory time. Changes in position of hydrophobic residues V108, F124, L127, and L129 in loops L1 and L2 lead to reorientation of amino acid residues Q112 and N159. In turn, in the wild-type enzyme, the side chain of N159 is one of the elements of the contact network supporting catalytically competent orientation of the cosubstrate, forming a hydrogen bond with the α-carboxyl group of αKG. Convergence of side chains of Q112 and N159 residues in the ABH2 V99A mutant form (Fig. 4*c*) leads to transfer of the hydrogen bond of the amide group of N159 from the α-carboxyl group of αKG (Fig. 4*d*) to the ω-carboxyl group of αKG (Fig. 4*e*, *f*), provoking its displacement from the optimal position for catalysis.

Thus, modeling results allow the suggestion that the V99A substitution, leading to disruptions in the binding of both substrate and cosubstrate in the enzyme's active site, should cause significant activity reduction. These data are in a good agreement with experimental results obtained previously for the V99A mutant form, revealing significant reduction (Monsen et al., 2010) or complete loss (Davletgildeeva et al., 2025) of ABH2 V99A catalytic activity toward dsDNA substrates containing m^1A or m^3C as damage.

Model of the ABH2 F124A enzyme-substrate complex with damaged DNA

To determine the functional role of F124 residue, modeling of complexes of the ABH2 F124A mutant form with m^1A- and m^3C-containing dsDNA was performed (Fig. 5). Detailed analysis of distribution changes of distances between key

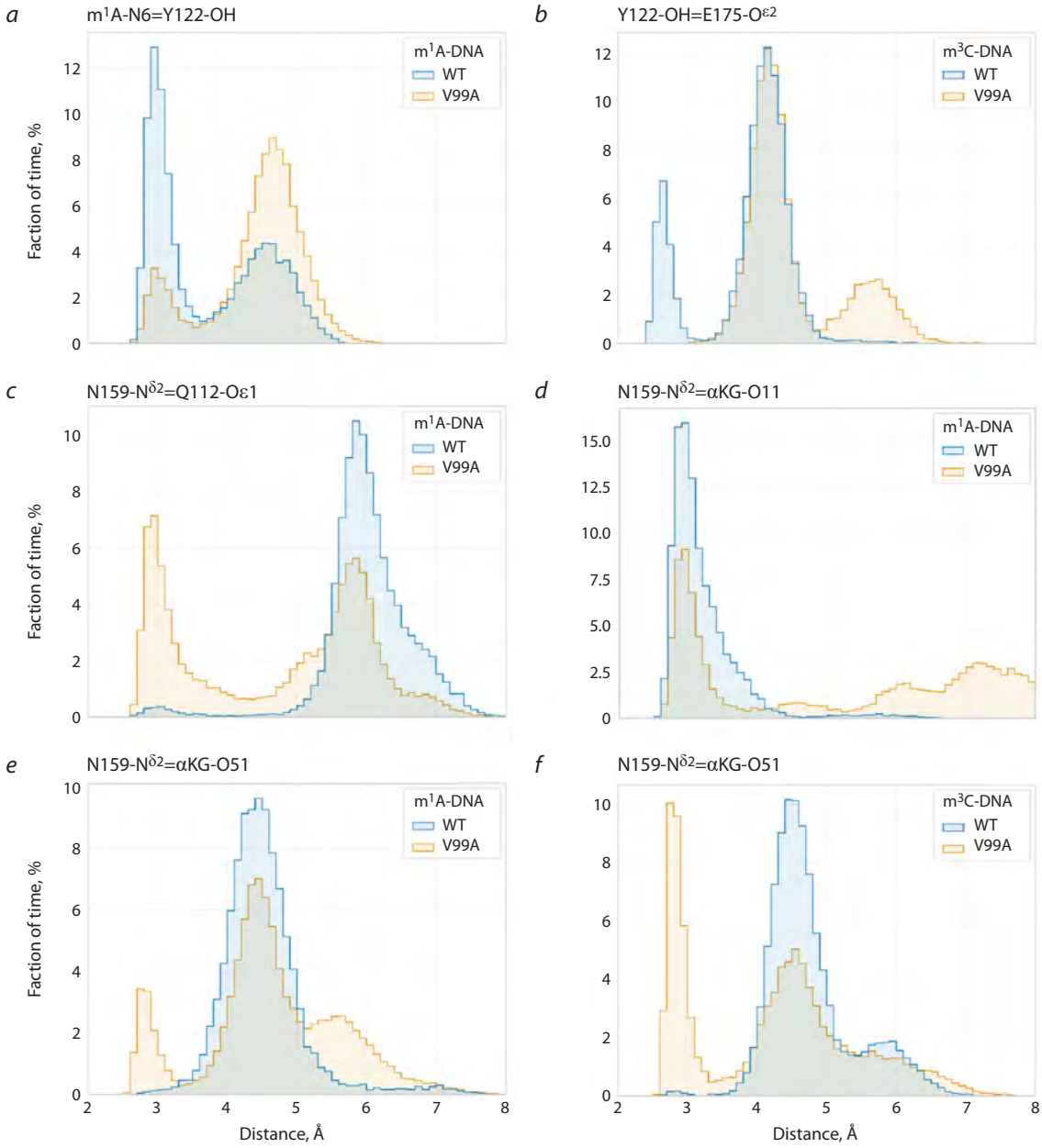


Fig. 4. Distance distributions between key atoms when modeling complexes of the wild-type ABH2 enzyme and its V99A mutant form with DNA substrates.

atoms of the active site in case of F124A substitution revealed destabilization of both the flipped methylated nitrogenous base and αKG in the enzyme's active site.

Thus, the F124A substitution, directly leading to loss of π-π stacking between the F124 side chain and the nitrogenous base, induces rotation and displacement of the flipped base from the enzyme's active site, with concomitant loss of hydrogen bonds with side chains of Y122, D173, E175 residues (Fig. 6*a*, *b*). The hydrogen bond between the hydroxyl group of S125 residue and the corresponding phosphate group of the nucleotide backbone is also lost, reflecting deterioration of contact between loop L2 and DNA (Fig. 6*c*).

The cosubstrate also loses catalytically competent position as a result of restructuring of the hydrogen bond network involving amino acid residues coordinating it. The amide group of N159 maintains a hydrogen bond predominantly with the ω-carboxyl group of αKG instead of the α-carboxyl group (Fig. 6*d*). Destabilization of the cosubstrate position is reflected in changes in the nature of contacts between side chains of Y161 and R248 residues and the ω-carboxyl group of αKG. If in the wild-type enzyme complex, stable hydrogen bonds are maintained between the guanidinium group of R248 and O2 atom of the ω-carboxyl group of αKG, and between the hydroxyl group of Y161 and O1 atom of the ω-carboxyl group, then in the ABH2 F124A mutant form complex, expansion of these distance distributions occurs, indicating contact destabilization (Fig. 6*d*, *e*).

The results of modeling indicate that amino acid residue F124 plays an important role in the structure of the ABH2 enzyme active site. This conclusion agrees with data (Chen et

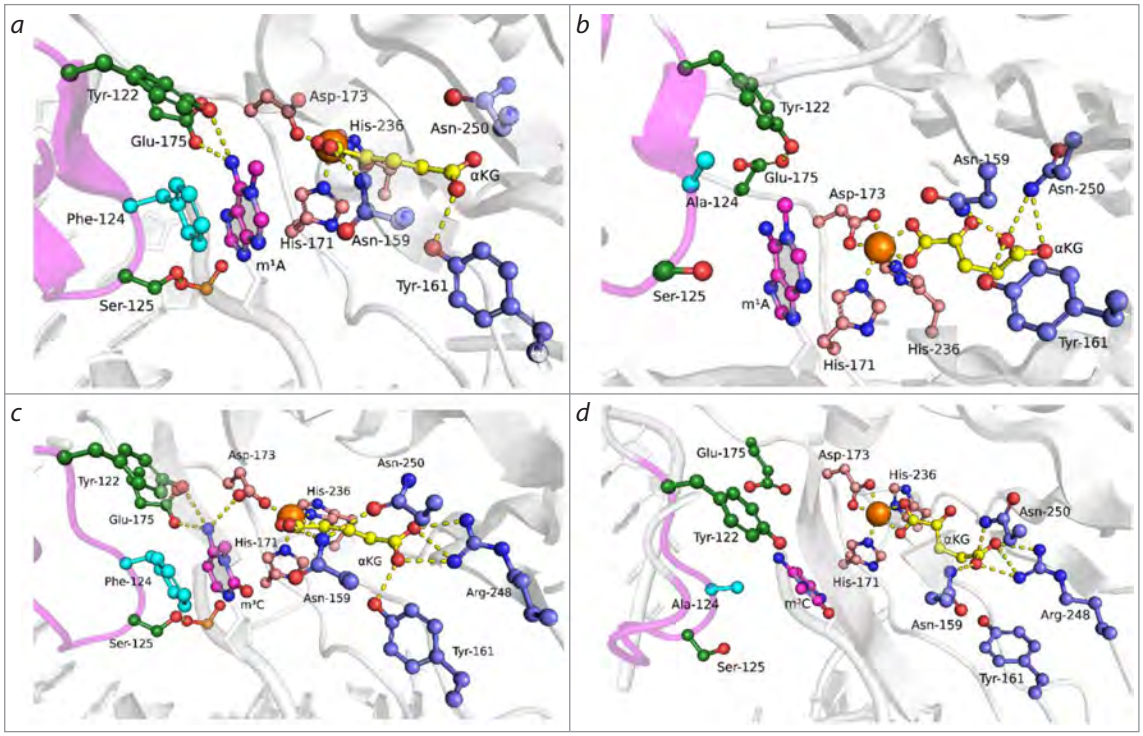


Fig. 5. Representative MD structures of complexes ABH2 WT with m^{1}A-DNA (*a*) and m^{3}C-DNA (*c*), and ABH2 F124A with m^{1}A-DNA (*b*) and with m^{3}C-DNA (*d*).

Key amino acid residues of the active site, damaged nitrogenous base, αKG and Mn2+ ion are shown. Loop L2 is highlighted with color (pink).

al., 2010; Monsen et al., 2010), as well as with data obtained previously in our laboratory (Davletgildeeva et al., 2025), according to which the ABH2 F124A mutant form completely lost catalytic activity toward m^{1}A- and m^{3}C-containing DNA substrates.

Model of the ABH2 S125A enzyme-substrate complex with damaged DNA

The S125A substitution in the ABH2 enzyme causes loss of the hydrogen bond between the hydroxyl group of the amino acid residue and the 5′-phosphate group of the damaged nucleotide, leading to loss of direct interaction of loop L2 with m^{1}A- (Fig. 7*a*, *b*) and m^{3}C-DNA (Fig. 7*c*, *d*). Analysis of distance changes between key residues of the active site showed that in the enzyme complex with m^{1}A-DNA, loss of loop L2 interaction with DNA causes loss of the hydrogen bond between the hydroxyl group of Y122 residue from L2 and the exocyclic amino group of m^{1}A (Fig. 8*a*). At the same time, convergence of guanidinium groups of R110 and R172 residues with the O3′ atom of the nucleotide of the flipped nitrogenous base and the O5′ atom of the nucleotide located 5′ to the flipped nitrogenous base, respectively, occurs (Fig. 8*b*, *c*). Thus, in case of DNA substrate containing m^{1}A, the S125A substitution leads to R110 and R172 amino acid residues binding more strongly to the DNA sugar-phosphate backbone.

Unlike the ABH2 S125A enzyme complex with m^{1}A-DNA, in the model complex with m^{3}C-DNA, convergence of guanidinium groups of R110 and R172 residues with the sugar-phosphate backbone does not occur (Fig. 7*c*, *d*). Meanwhile, compared to the WT enzyme, in case of S125A substitution, stability of the hydrogen bond between the side chain of E175 residue and the exocyclic amino group of m^{3}C decreases (Fig. 8*d*).

Deterioration of direct contact with the flipped base and possible compensatory restructuring in case of S125A substitution in the ABH2 active site agrees with the results obtained by B. Chen et al., since their work showed that the ABH2 S125A mutant form retains catalytic activity toward dsDNA containing m^{1}A as damage (Chen et al., 2010). However, in a later work (Davletgildeeva et al., 2025), it was shown that this substitution leads to loss of ABH2 catalytic activity toward both m^{3}C- and m^{1}A-containing DNA under the used reaction conditions. This suggests that compensatory restructuring that occurs according to modeling data in the ABH2 structure upon S125A substitution cannot fully preserve the enzyme's catalytic activity on all types of DNA substrates.

Conclusion

Introduction of the V99A substitution into the ABH2 enzyme affected other amino acid residues forming the hydrophobic network of which the substituted residue is a part. This led to negative influence on functional loops L1 and L2, causing destabilization of their position, which, in turn, led to reorientation or displacement of key amino acid residues, Y122, E175, and F102, comprised in these loops. Additionally, the V99A substitution led to a catalytically unfavorable conformation of αKG in the enzyme's active site. The obtained data confirm the role of V99 amino acid residue as an important participant in intraprotein coordination

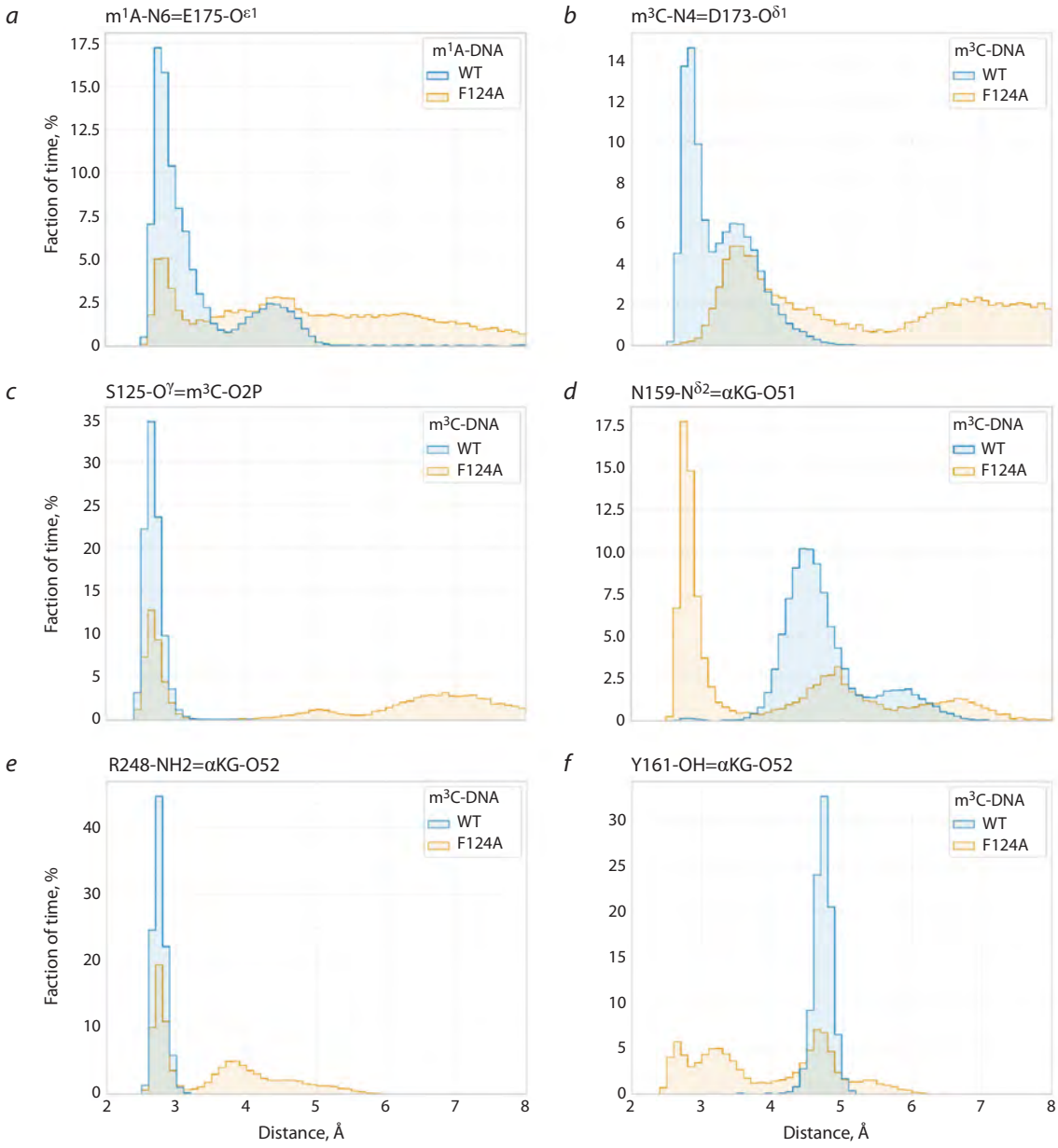


Fig. 6. Distance distributions between key atoms when modeling complexes of the wild-type ABH2 enzyme and its F124A mutant form with DNA substrates.

necessary for effective oxidation of methyl groups in damaged DNA bases by the ABH2 enzyme.

Substitution of amino acid residue F124, localized in NRL, led to significant displacement of both L1 and L2 loops and the damaged base itself relative to each other due to loss of π-π stacking with the damaged nitrogenous base. This substitution also led to changes in Fe^{2+} ion coordination, both through changes in coordination type by the αKG molecule and through additional coordination by D173 amino acid residue. The obtained data suggest extreme importance of F124 amino acid residue in the catalytic process carried out by ABH2 DNA dioxygenase.

The S125A substitution led to loss of direct interaction of loop L2 with the 5′-phosphate group of the damaged nucleotide; however, according to MD modeling data, this contact can be partially compensated by formation of bonds between R110 and R172 amino acid residues and the DNA sugar-phosphate backbone. It should be noted that such contact compensation was found only in case of the ABH2 S125A complex with m^1A-containing DNA substrate, but not in case of m^3C, which indirectly indicates a more complex mechanism responsible for recognition of different damages in the enzyme's active site.

Thus, the MD modeling data obtained in the present work for complexes of human ABH2 DNA dioxygenase mutant forms containing V99A, F124A, or S125A amino acid substitutions with m^1A- and m^3C-containing DNA substrates indicate the important role of all three amino acid residues in ensuring formation of a catalytically competent state of the active site when interacting with damaged DNA.

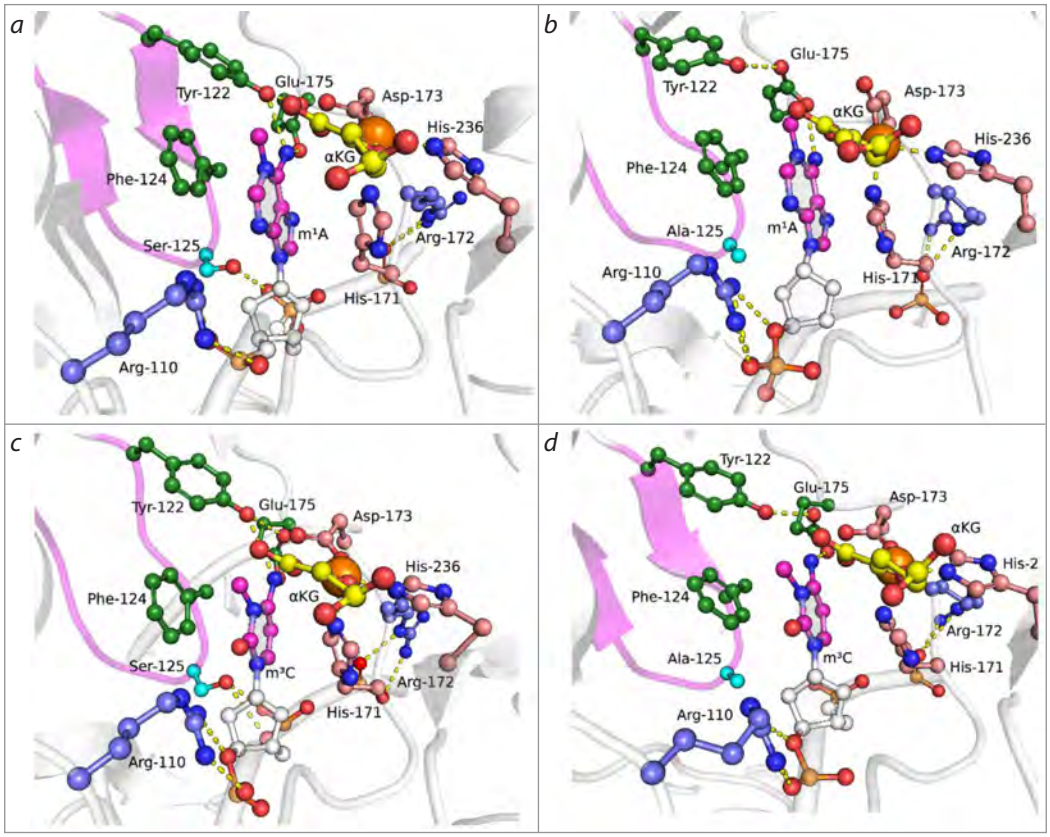


Fig. 7. Representative MD structures of complexes ABH2 WT with m^1A-DNA (*a*) and m^3C-DNA (*c*), and ABH2 S125A with m^1A-DNA (*b*) and with m^3C-DNA (*d*).

Key amino acid residues of the active site, damaged nitrogenous base, αKG and Mn^{2+} ion are shown. Loop L2 is highlighted with color (pink).

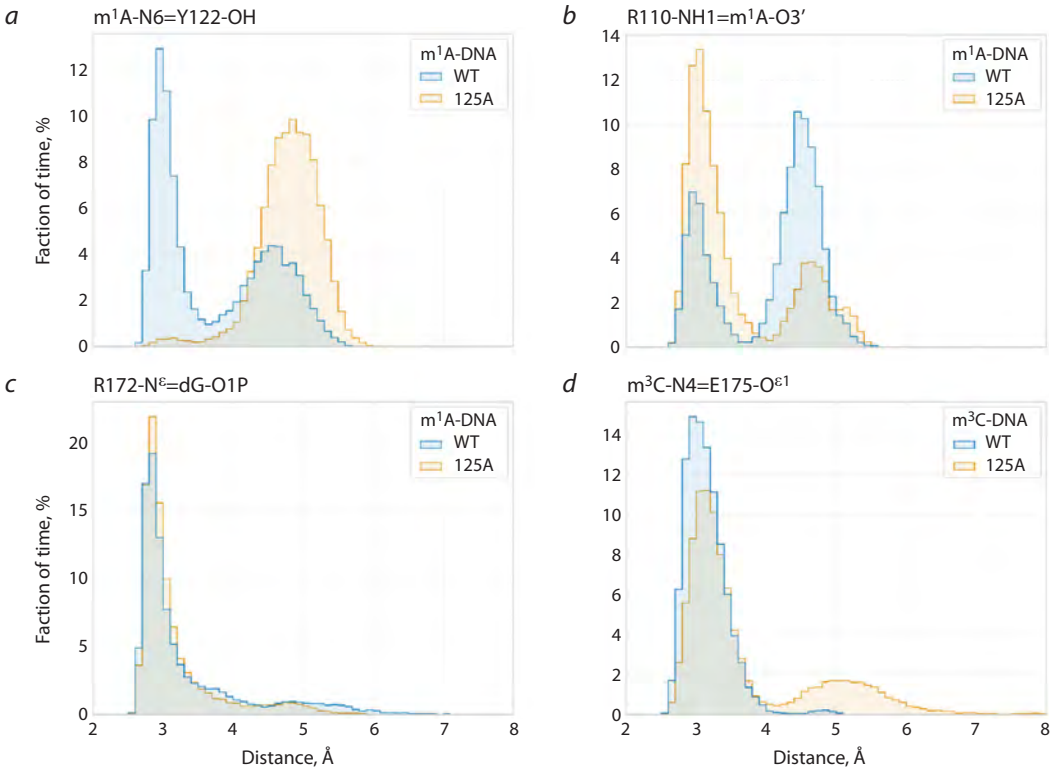


Fig. 8. Distance distributions between key atoms when modeling complexes of the wild-type ABH2 enzyme and its S125A mutant form with DNA substrates.

References

Aas P.A., Otterlei M., Falnes P., Vågbø C.B., Skorpen F., Akbari M., Sundheim O., Bjørås M., Slupphaug G., Seeberg E., Krokan H.E. Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature*. 2003;421:859-863. doi 10.1038/nature01363

Abraham M.J., Murtola T., Schulz R., Páll S., Smith J.C., Hess B., Lindahl E. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1-2:19-25. doi 10.1016/j.softx.2015.06.001

Anandakrishnan R., Aguilar B., Onufriev A.V. *H*++ 3.0: automating p*K* prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res*. 2012;40:W537-W541. doi 10.1093/nar/gks375

Bayly C.I., Cieplak P., Cornell W., Kollman P.A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem*. 1993;97:10269-10280. doi 10.1021/j100142a004

Bian K., Lenz S.A.P., Tang Q., Chen F., Qi R., Jost M., Drennan C.L., Essigmann J.M., Wetmore S.D., Li D. DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine *in vitro*. *Nucleic Acids Res*. 2019;47(11):5522-5529. doi 10.1093/nar/gkz395

Bussi G., Donadio D., Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys*. 2007;126(1):014101. doi 10.1063/1.2408420

Chen B., Liu H., Sun X., Yang C.-G. Mechanistic insight into the recognition of single-stranded and double-stranded DNA substrates by ABH2 and ABH3. *Mol Biosyst*. 2010;6(11):2143-2149. doi 10.1039/c005148a

Chen B., Gan J., Yang C. The complex structures of ALKBH2 mutants cross-linked to dsDNA reveal the conformational swing of β-hairpin. *Sci China Chem*. 2014;57:307-313. doi 10.1007/s11426-013-5029-z

Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W., Kollman P.A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*. 1995;117(19):5179-5197. doi 10.1021/ja00124a002

Davletgildeeva A.T., Tyugashev T.E., Zhao M., Kuznetsov N.A., Ishchenko A.A., Saparbaev M., Kuznetsova A.A. Individual contributions of amido acid residues Tyr122, Ile168, and Asp173 to the activity and substrate specificity of human DNA dioxygenase ABH2. *Cells*. 2023;12(14):1839. doi 10.3390/cells12141839

Davletgildeeva A.T., Tyugashev T.E., Zhao M., Ishchenko A.A., Saparbaev M., Kuznetsov N.A. Role of individual amino acid residues directly involved in damage recognition in active demethylation by ABH2 dioxygenase. *Int J Mol Sci*. 2025;26:6912. doi 10.3390/ijms26146912

Duncan T., Trewick S.C., Koivisto P., Bates P.A., Lindahl T., Sedgwick B. Reversal of DNA alkylation damage by two human dioxygenases. *Proc Natl Acad Sci USA*. 2002;99(26):16660-16665. doi 10.1073/pnas.262589799

Essmann U., Perera L., Berkowitz M.L., Darden T., Lee H., Pedersen L.G. A smooth particle mesh Ewald method. *J Chem Phys*. 1995;103:8577-8593. doi 10.1063/1.470117

Falnes P. Repair of 3-methylthymine and 1-methylguanine lesions by bacterial and human AlkB proteins. *Nucleic Acids Res*. 2004;32:6260-6267. doi 10.1093/nar/gkh964

Giri N.C., Sun H., Chen H., Costa M., Maroney M.J. X-ray absorption spectroscopy structural investigation of early intermediates in the mechanism of DNA repair by human ABH2. *Biochemistry*. 2011;50(22):5067-5076. doi 10.1021/bi101668x

Hess B., Bekker H., Berendsen H.J.C., Fraaije J.G.E.M. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*. 1997;18(12):1463-1472. doi 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Jiang Y., Zhang H., Tan T. Rational design of methodology-independent metal parameters using a nonbonded dummy model. *J Chem Theory Comput*. 2016;12(7):3250-3260. doi 10.1021/acs.jctc.6b00223

Jorgensen W.L., Chandrasekhar J., Madura J.D., Impey R.W., Klein M.L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79(2):926-935. doi 10.1063/1.445869

Joung I.S., Cheatham T.E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B*. 2008;112:9020-9041. doi 10.1021/jp8001614

Kuznetsov N.A., Kanazhevskaya L.Y., Fedorova O.S. DNA demethylation in the processes of repair and epigenetic regulation performed by 2-ketoglutarate-dependent DNA dioxygenases. *Int J Mol Sci*. 2021;22:10540. doi 10.3390/ijms221910540

Lee D.H., Jin S.G., Cai S., Chen Y., Pfeifer G.P., O'Connor T.R. Repair of methylation damage in DNA and RNA by mammalian AlkB homologues. *J Biol Chem*. 2005;280(47):39448-39459. doi 10.1074/jbc.M509881200

Lenz S.A.P., Li D., Wetmore S.D. Insights into the direct oxidative repair of etheno lesions: MD and QM/MM study on the substrate scope of ALKBH2 and AlkB. *DNA Repair* (*Amst*). 2020;96:102944. doi 10.1016/j.dnarep.2020.102944

Li P., Gao S., Wang L., Yu F., Li J., Wang C., Li J., Wong J. ABH2 couples regulation of ribosomal DNA transcription with DNA alkylation repair. *Cell Rep*. 2013;4:817-829. doi 10.1016/j.celrep.2013.07.027

Maier J.A., Martinez C., Kasavajhala K., Wickstrom L., Hauser K.E., Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*. 2015;11:3696-3713. doi 10.1021/acs.jctc.5b00255

McGibbon R.T., Beauchamp K.A., Harrigan M.P., Klein C., Swails J.M., Hernández C.X., Schwantes C.R., Wang L.-P., Lane T.J., Pande V.S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys J*. 2015;109:1528-1532. doi 10.1016/j.bpj.2015.08.015

Monsen V.T., Sundheim O., Aas P.A., Westbye M.P., Sousa M.M.L., Slupphaug G., Krokan H.E. Divergent β-hairpins determine double-strand versus single-strand substrate recognition of human AlkB-homologues 2 and 3. *Nucleic Acids Res*. 2010;38:6447-6455. doi 10.1093/nar/gkq518

Müller T.A., Hausinger R.P. AlkB and its homologues. DNA repair and beyond. In: Schofield C., Hausinger R. (Eds) 2-Oxoglutarate-Dependent Oxygenases. Royal Society Chemistry. 2015;246-262. doi 10.1039/9781782621959-00246

Ougland R., Rognes T., Klungland A., Larsen E. Non-homologous functions of the AlkB homologs. *J Mol Cell Biol*. 2015;7(6):494-504. doi 10.1093/jmcb/mjv029

Parrinello M., Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys*. 1981;52(12):7182-7190. doi 10.1063/1.328693

Ringvoll J., Nordstrand L.M., Vagbo C.B., Talstad V., Reite K., Aas P.A., Lauritzen K.H., Liabakk N.B., Bjork A., Doughty R.W., Falnes P.O., Krokan H.E., Klungland A. Repair deficient mice reveal mABH2 as the primary oxidative demethylase for repairing 1meA and 3meC lesions in DNA. *Embo J*. 2006;25:2189-2198. doi 10.1038/sj.emboj.7601109

Ringvoll J., Moen M.N., Nordstrand L.M., Meira L.B., Pang B., Bekkelund A., Dedon P.C., Bjelland S., Samson L.D., Falnes P.Ø., Klungland A. AlkB homologue 2 – mediated repair of ethenoadenine lesions in mammalian DNA. *Cancer Res*. 2008;68(11):4142-4149. doi 10.1158/0008-5472.CAN-08-0796

Šali A., Blundell T.L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234(3):779-815. doi 10.1006/jmbi.1993.1626

Sall S.O., Berens J.T.P., Molinier J. DNA damage and DNA methylation. In: Jasiulionis M.G. (Ed.) Epigenetics and DNA Damage. Academic Press, 2022;3-16. doi 10.1016/B978-0-323-91081-1.00005-4

Sousa da Silva A.W., Vranken W.F. ACPYPE – AnteChamber PYthon Parser interfacE. *BMC Res Notes*. 2012;5:367. doi 10.1186/1756-0500-5-367

Travers A., Muskhelishvili G. DNA structure and function. *FEBS J*. 2015;282(12):2279-2295. doi 10.1111/febs.13307

Vanquelef E., Simon S., Marquant G., Garcia E., Klimerak G., Delepine J.C., Cieplak P., Dupradeau F.-Y. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res*. 2011;39:W511-W517. doi 10.1093/nar/gkr288

Waheed S.O., Ramanan R., Chaturvedi S.S., Lehnert N., Schofield C.J., Christov C.Z., Karabencheva-Christova T.G. Role of structural dynamics in selectivity and mechanism of non-heme Fe(II) and 2-oxoglutarate-dependent oxygenases involved in DNA repair. *ACS Cent Sci*. 2020;6(5):795-814. doi 10.1021/acscentsci.0c00312

Wang J., Wolf R.M., Caldwell J.W., Kollman P.A., Case D.A. Development and testing of a general amber force field. *J Comput Chem*. 2004;25:1157-1174. doi 10.1002/jcc.20035

Wang J., Wang W., Kollman P.A., Case D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*. 2006;25(2):247-260. doi 10.1016/j.jmgm.2005.12.005

Wilson D.L., Beharry A.A., Srivastava A., O'Connor T.R., Kool E.T. Fluorescence probes for ALKBH2 allow the measurement of DNA alkylation repair and drug resistance responses. *Angew Chem Int Ed Engl*. 2018;57(39):12896-12900. doi 10.1002/anie.201807593

Xu B., Liu D., Wang Z., Tian R., Zuo Y. Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. *Cell Mol Life Sci*. 2021;78:129-141. doi 10.1007/s00018-020-03594-9

Yang C.G., Yi C., Duguid E.M., Sullivan C.T., Jian X., Rice P.A., He C. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature*. 2008;452:961-965. doi 10.1038/nature06889

Yang C.G., Garcia K., He C. Damage detection and base flipping in direct DNA alkylation repair. *Chembiochem*. 2009;10(3):417-423. doi 10.1002/cbic.200800580

Yi C., He C. DNA repair by reversal of DNA damage. *Cold Spring Harb Perspect Biol*. 2013;5:a012575. doi 10.1101/cshperspect.a012575

Yi C., Yang C.G., He C. A non-heme iron-mediated chemical demethylation in DNA and RNA. *Acc Chem Res*. 2009;42(4):519-529. doi 10.1021/ar800178j

Yi C., Chen B., Qi B., Zhang W., Jia G., Zhang L., Li C.J., Dinner A.R., Yang C.-G., He C. Duplex interrogation by a direct DNA repair protein in search of base damage. *Nat Struct Mol Biol*. 2012;19:671-676. doi 10.1038/nsmb.2320

Zgarbová M., Otyepka M., Sponer J., Mládek A., Banáš P., Cheatham T.E., Jurečka P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theory Comput*. 2011;7(9):2886-2902. doi 10.1021/ct200162x

Zgarbová M., Šponer J., Otyepka M., Cheatham T.E., Galindo-Murillo R., Jurečka P. Refinement of the sugar-phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J Chem Theory Comput*. 2015;11(12):5723-5736. doi 10.1021/acs.jctc.5b00716

Conflict of interest. The authors declare no conflict of interest.

Received June 6, 2025. Revised September 12, 2025. Accepted September 12, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1073-1083

doi 10.18699/vjgb-25-112

Structural basis of the phosphoramidate *N*-benzimidazole group's influence on modified primer extension efficiency by Taq DNA polymerase

A.A. Berdugin [1, 2], V.M. Golyshev [1, 2], A.A. Lomzov [1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
lomzov@1bio.ru

Abstract. We recently proposed a novel class of nucleic acid derivatives – phosphoramidate benzoazole oligonucleotides (PABAOs). In these compounds, one of the non-bridging oxygen atoms is replaced by a phosphoramidate *N*-benzoazole group, such as benzimidazole, dimethylbenzimidazole, benzoxazole, or benzothiazole. Studies of the properties of these derivatives have shown that their use in PCR enhances the specificity and selectivity of the analysis. The study investigates the effect of phosphoramide *N*-benzimidazole modification of DNA primers on their elongation by Taq DNA polymerase using molecular dynamics simulations. We examined perfectly matched primer-template complexes with modifications at positions one through six from the 3'-end of the primer. Prior experimental work demonstrated that the degree of elongation suppression depends on the modification position: the closer to the 3'-end, the stronger the inhibition, with maximal suppression observed for the first position, especially in mismatched complexes. Furthermore, incomplete elongation products were experimentally observed for primers modified at the fourth position. Our molecular dynamics simulations and subsequent analysis revealed the molecular mechanisms underlying the interaction of modified primers with the enzyme. These include steric hindrance that impedes polymerase progression along the modified strand and local distortions in the DNA structure, which explain the experimentally observed trends. We established that both different stereoisomers of the phosphoramidate groups and conformers of the phosphoramidate *N*-benzimidazole moiety differentially affect the structure of the enzyme-substrate complex and the efficiency of Taq DNA polymerase interaction with the modified DNA complex. Modification of the first and second internucleoside phosphate from the 3'-end of the primer causes the most significant perturbation to the structure of the protein-nucleic acid complex. When the modification is located at the fourth phosphate group, the *N*-benzimidazole moiety occupies a specific pocket of the enzyme. These findings provide a foundation for the rational design of specific DNA primers bearing modified *N*-benzimidazole moieties with tailored properties for use in PCR diagnostics.

Key words: *N*-benzimidazole oligonucleotides; PABAO; molecular dynamics; structure; Taq DNA polymerase; molecular diagnostics

For citation: Berdugin A.A., Golyshev V.M., Lomzov A.A. Structural basis of the phosphoramidate *N*-benzimidazole group's influence on modified primer extension efficiency by Taq DNA polymerase. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1073-1083. doi 10.18699/vjgb-25-112

Funding. This work was financially supported by the Russian Science Foundation (project No. 23-74-01116, https://rscf.ru/project/23-74-01116/) for the construction and initial analysis of model systems, and by the Russian state-funded project for ICBFM SB RAS (grant number 123021600208-7) for molecular dynamics simulations and analysis of the resulting data.

Структурные основы влияния фосфорамидной *N*-бензимидазольной группы на эффективность удлинения модифицированного праймера Taq ДНК-полимеразой

А.А. Бердюгин [1, 2], В.М. Голышев [1, 2], А.А. Ломзов [1, 2]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
lomzov@1bio.ru

Аннотация. Недавно нами был предложен новый класс производных нуклеиновых кислот – фосфорамидные бензоазольные олигонуклеотиды. В них один из немостиковых атомов кислорода замещен на фосфорамидную *N*-бензоазольную группу: бензимидазольную, диметилбензимидазольную, бензоксазольную или бензотиазольную. Изучение свойств таких производных показало, что их применение в ПЦР увеличивает специфичность и селективность анализа. Данное исследование посвящено изучению влияния фосфорамидной *N*-бензимидазольной модификации ДНК-праймеров на эффективность их удлинения Taq ДНК-полимеразой при

помощи метода молекулярной динамики. Мы рассматривали совершенные комплексы нуклеиновых кислот с модификациями в положениях с первого по шестое считая от 3′-конца праймера. Ранее было показано, что степень подавления элонгации зависит от положения модификации: чем ближе к 3′-концу, тем сильнее ингибирование, а максимальное подавление наблюдается при модификации в первом положении, особенно в несовершенных комплексах. Кроме того, в экспериментах наблюдались продукты неполного удлинения праймеров с модификацией в четвертом положении. Проведенные компьютерное моделирование и анализ позволили выявить молекулярные механизмы взаимодействия модифицированных праймеров с ферментом, включая стерические препятствия для продвижения полимеразы по модифицированной цепи и локальные нарушения структуры ДНК, которые объясняют наблюдаемые экспериментально закономерности. Установлено, что как различные стереоизомеры фосфорамидных групп, так и конформеры фосфорамидной *N*-бензимидазольной группы по-разному влияют на структуру фермент-субстратного комплекса и эффективность взаимодействия Taq ДНК-полимеразы с модифицированным ДНК комплексом. Модификация первого и второго межнуклеозидного фосфатного остатка с 3′-конца праймера в наибольшей степени возмущает структуру белково-нуклеинового комплекса, а при расположении модификации в четвертом фосфатном остатке *N*-бензимидазольная модификация располагается в кармане фермента. Полученные результаты открывают перспективы для рационального конструирования специфичных, обладающими заранее заданными свойствами ДНК праймеров с модифицированными *N*-бензимидазольными межнуклеотидными звеньями для использования в ПЦР диагностике.
Ключевые слова: *N*-бензимидазольные олигонуклеотиды; ФАО; молекулярная динамика; структура; Taq ДНК-полимераза; молекулярная диагностика

Introduction

DNA-dependent DNA polymerase I from the bacterium *Thermus aquaticus* (Taq DNA polymerase) is a widely used enzyme for nucleic acid amplification by the polymerase chain reaction (PCR) in various applications. It possesses DNA polymerase and 5′→3′ exonuclease activities but lacks proofreading 3′→5′ exonuclease activity (Terpe, 2013). This enzyme is widely used for the detection of nucleic acids (NA) and single-nucleotide variants (point mutations) in diagnostic applications for various diseases, using diverse PCR-based methods such as real-time PCR, allele-specific PCR, and digital PCR (Kalendar et al., 2022; Starza et al., 2022). Allele-specific PCR is based on the inhibition of primer elongation when primers form duplexes with the template strand containing one or more mismatches at or near the 3′-end of the primer (Rejali et al., 2018). Often, a single nucleotide substitution that disrupts full complementarity between the primer and the DNA template does not provide sufficient specificity for polymorphism detection. To enhance specificity, additional single-nucleotide mismatches and/or structural modifications are introduced into the primer. These modifications can be incorporated either into the nucleobase or into the ribose-phosphate backbone and are typically positioned near the 3′-end of the primer (Kutyavin, 2011; Ishige et al., 2018; Chubarov et al., 2023). In particular, substitution of the non-bridging oxygen atom in the phosphodiester backbone affects both the thermodynamic stability of the primer–template duplex and the coordination of the terminal 3′-OH group within the enzyme's active site. For example, incorporation of a phosphorothioate modification at the terminal or penultimate internucleotide phosphate linkage from the 3′-end of the primer results in only a modest reduction in elongation efficiency (5–15 %) while simultaneously enhancing amplification specificity (Di Giusto, King, 2003). Introduction of phosphoryl guanidine modifications into primer structures likewise alters the efficiency and selectivity of target nucleic acid sequence detection (Chubarov et al., 2020).

Recently, a novel class of nucleic acid derivatives, phosphoramidate benzazole oligonucleotides (PABAOs), was developed at the Institute of Chemical Biology and Fundamental Medicine SB RAS (Vasilyeva et al., 2023). In PABAOs, the non-bridging oxygen atom of the phosphate moiety is substituted by an *N*-benzazole group (*N*-benzimidazole, *N*-benzoxazole, or *N*-benzothiazole) (Fig. 1). PABAOs can be synthesized using standard automated solid-phase phosphoramidite chemistry.

To date, the physicochemical properties of several *N*-benzazole derivatives of NA have been investigated (Golyshev et al., 2024; Yushin et al., 2024; Novgorodtseva et al., 2025) and their potential use as primers in PCR, including allele-specific PCR, has been shown (Chubarov et al., 2024). We have examined the elongation efficiency of 13-mer primers containing an *N*-benzimidazole modification on a 22-mer DNA template using Taq DNA polymerase (Golyshev et al., 2025). When the modification is introduced at the first or second internucleotide phosphate from the 3′-end of the primer in perfectly matched duplexes, full-length extension

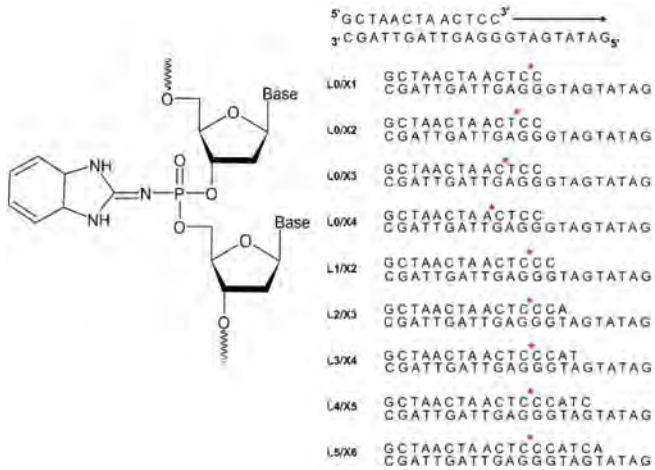


Fig. 1. Structure of a dinucleotide step of phosphoramidate benzazole oligonucleotides containing an *N*-benzimidazole group and the model systems used in this study.

The position of the phosphoramidate *N*-benzimidazole group is indicated by a red asterisk.

occurs with an efficiency of approximately 50 %. In contrast, for duplexes containing a single-nucleotide mismatch at the penultimate base pair from the 3′-end of the primer, the yield of full-length product is markedly reduced. Incorporation of the modification at the third position typically results in the smallest decrease in full-length product yield among the studied positions. Furthermore, for all perfectly matched duplexes bearing the modification, a distinct aborted elongation product was consistently observed, corresponding to a partially elongated primer in which the modification was at the fourth position from the 3′-end.

In this work, we used molecular dynamics (MD) simulations to elucidate the experimental patterns of PABAO primer elongation by Taq DNA polymerase. Our study focused on how the phosphoramidate *N*-benzimidazole group, positioned at various sites along the primer, affects the structure and dynamics of the enzyme–substrate complex. To this end, we constructed molecular models and carried out MD simulations of both the native (unmodified) and a series of modified nucleic acid substrates containing the *N*-benzimidazole modification at the 1st through 6th internucleotide phosphate positions from the 3′-end of the primer, as well as their complexes with Taq DNA polymerase. The simulation results correlate well with experimental data and provide a mechanistic explanation for the effects observed *in vitro*.

Methods

Model building. The structure of the Taq polymerase–DNA complex was constructed based on the experimentally determined crystal structure with PDB ID: 1QTM as follows. The protein coordinates, including the bound nucleoside triphosphate (dNTP) and magnesium ions, were retained from this structure. The DNA complex of the template strand with the primer was modeled by building a protein–nucleic acid complex using AlphaFold3 software (Abramson et al., 2024). As input for these calculations, we provided the amino acid sequence of *Thermus aquaticus* DNA polymerase I (UniProt ID: P19821), along with the nucleotide sequences of the DNA template and either the unextended or partially extended primers, an incoming deoxyribonucleoside triphosphate (dNTP), and two Mg^{2+} ions in catalytic site. The resulting AlphaFold3-predicted structure was then superimposed onto the experimentally determined structure 1QTM by aligning the protein backbone based on Cα atoms of equivalent residues. Subsequently, the native nucleic acid components in the 1QTM structure were replaced with the DNA duplexes generated by AlphaFold3. For each constructed model, the original dNTP was substituted with the nucleotide triphosphate complementary to the base in template at the active site, ensuring correct base pairing for the elongation step under investigation.

Since the *N*-benzimidazole modification generally requires additional space for proper geometric accommodation within the DNA/Taq polymerase complex, we employed amino acid side-chain rotamer libraries (Shapovalov, Dunbrack, 2011) implemented in UCSF Chimera (Pettersen et al., 2004) to minimize van der Waals clashes between protein atoms and bulky modification.

Partial atomic charges for amino acid residues in each complex were assigned using the pdb2pqr software (version 3.7.1) (Unni et al., 2011). The pH was set to 8.3 to match the experimental primer extension conditions (Golyshev et al., 2025). As a result, certain complexes exhibited differences in the protonation states of specific charged residues. Out of the 36 modeled complexes, seven displayed distinct protonation patterns. In the complexes L0/X2/R1, L0/X2/R2, L0/X3/R1, L0/X3/R2, and L1/X2/R2 (notation defined below), the residues LYS540, ASP610, LYS663, and ASP785 were found in their protonated forms. In the complexes L0/X4/R1 and L0/X4/R2, the residues LYS663, LYS762, and GLU786 were also protonated.

The primer/template complexes were obtained from the protein–nucleic acid complex by removing all residues except those belonging to the DNA strands.

Molecular dynamics simulation. Structural investigations of complexes formed between native or modified DNA and Taq DNA polymerase were carried out using molecular dynamics (MD) simulations and subsequent analysis with the AMBER20 software package (Case et al., 2020). Simulations were performed using parallel computing on both central processing units (CPUs) and graphics processing units (GPUs) with CUDA architecture. All MD calculations employed the ff19SB force field (Tian et al., 2020) for Taq polymerase, the OL21 force field (Zgarbová et al., 2021) for native DNA, and gaff2 parameters for the *N*-benzimidazole-modified phosphate residues. Parameters for magnesium and sodium ions were taken from (Li Z. et al., 2020). These force fields represent the most up-to-date and rigorously validated options currently recommended by the AMBER developers for reliable biomolecular simulations. Parameters for the deoxyribonucleoside triphosphates (dNTPs) were adopted from (Meagher et al., 2003), which remain the only published and widely accepted dNTP parameters compatible with the AMBER force field family.

MD simulation protocol. Initial models were first relaxed in implicit solvent (saltcon = 0.10 M, igb = 1, T = 1 K) using the conjugate gradient method for 2,500 steps. The systems were then solvated in an octahedral box of OPC water molecules (Izadi et al., 2014), with a minimum distance of 14 Å between any solute atom and the box boundary. Sodium ions (Na^+) were added to neutralize the total charge of the periodic cell. Subsequently, the solvated systems underwent restrained energy minimization for 10,000 steps (with the first 200 steps performed using the steepest descent algorithm), applying positional restraints of 1.0 kcal/(mol·Å²) on all complex' heavy atoms to prevent structural distortion during initial solvent relaxation. Following minimization, the systems were gradually heated from 0 to 300 K over 2 ns under constant volume (NVT ensemble), using Langevin dynamics for temperature control (ntt = 3, gamma_ln = 1.0). Pressure was then equilibrated to 1 atm over an additional 1 ns using a Monte Carlo barostat (NPT ensemble). A final unrestrained energy minimization was performed for 10,000 steps (first 200 steps: steepest descent) to remove any residual clashes after equilibration. A time step of 2 fs was used throughout, with bonds involving hydrogen atoms constrained via the SHAKE algorithm. And at the final stage, MD simulation was carried out for 100 ns with parameters similar to the heating stage, but without imposing positional restrictions on the atoms of the model system.

The MD simulation trajectories were analyzed using the cpptraj module from the AMBER20 package (Roe, Cheatham, 2013). For each trajectory, the 10 most representative structures were identified through hierarchical clustering analysis, using the average-linkage algorithm and root-mean-square deviation (RMSD) of backbone atoms as the distance metric.

Molecular graphics were prepared using UCSF Chimera version 1.15 (Pettersen et al., 2004).

Results

Selection and construction of molecular models

The structural and dynamic properties of PABAO complexes with Taq DNA polymerase were investigated using a comprehensive set of model systems. We employed the DNA complex formed by the primer 5′-GCTAACTAACTCC-3′ and the template strand 5′-GATATGATGGGAGTTAGTTAGC-3′, which was previously characterized in our experimental study of modified primer elongation efficiency (Golyshev et al., 2025). It has been shown that the introduction of benzoazole modifications at various positions of the primer affects the efficiency and specificity of its extension. As part of this work, MD modeling of a set of protein-nucleic acid complexes, as well as individual DNA complexes, was carried out. Both native DNA complexes and complexes containing *N*-benzimidazole modifications at the internucleotide phosphate groups from the 1st to the 6th position from the 3′-end of the primer were considered. To evaluate the effect of primer elongation and to obtain more reliable insights, we analyzed oligonucleotide complexes containing unextended primers with *N*-benzimidazole modifications positioned at 1 through 4 internucleotide phosphate from the 3′-end of the primer. In addition, we examined systems in which the primer initially bearing the *N*-benzimidazole modification at the first position was extended by 1 to 5 nucleotides. Following such elongation, the modification was at positions 2 through 6 relative to the new 3′-end of the primer. The sequences of the model oligonucleotide complexes and their corresponding nomenclature are provided in Figure 1.

Model construction was carried out based on the crystal structure with Protein Data Bank identifier (PDB ID) 1QTM, as described in the Methods section. This structure represents a fragment of *Thermus aquaticus* DNA polymerase I in its closed conformation, bound to a dideoxyribonucleoside triphosphate (ddNTP) and Mg^{2+}, and lacking exonuclease domain. The modification was introduced into the primer by replacing the native phosphate group with a phosphoramidate bearing an *N*-benzimidazole moiety (Fig. 1). Both stereoisomers of the phosphoramidate linkage (Sp and Rp) were considered in our study.

Analysis of the constructed molecular models of modified DNA in complex with Taq polymerase revealed that, for each phosphoramidate stereoisomer (Sp and Rp), the *N*-benzimidazole group can adopt two distinct orientations. These orientations correspond to the dihedral angle OP–P–N–C (where OP is the bridging oxygen, P is the phosphorus atom, N is the benzimidazole nitrogen, and C is the adjacent carbon in the heterocycle) of approximately –100 or +100°. Preliminary molecular dynamics simulations of the protein–nucleic acid complexes indicated that no transitions occurred between these two orientations of the *N*-benzimidazole group during the simulation timescale. Therefore, we explicitly considered both conformers (rotamers). For the model DNA complexes, we adopted the following nomenclature: L*i*/X*j*/R*k* and L*i*/X*j*/S*k*, where i = 0–5 denotes the number of nucleotides by which the primer has been elongated, j = 1–6 indicates the position of the internucleotide phosphate (counting from the 3′-end of the primer) at which the *N*-benzimidazole modification is introduced, k = 1, 2 specifies the rotameric conformation of the benzimidazole group for each phosphoramidate stereoisomer. For the rotamers R1 and S2, the dihedral angle defined by the atoms OP2–P–N–C (for the *Rp* isomer) or OP1–P–N–C (for the *Sp* isomer) was approximately –100°. In contrast, for rotamers R2 and S1, the corresponding dihedral angle adopted a value of approximately +100°. In these configurations, the spatial orientation of the benzoazole ring in the R1 and S1 rotamers directs the modified group away from the major groove of the DNA duplex, whereas in the R2 and S2 rotamers, the benzoazole ring is oriented toward the minor groove (Fig. 2). For modeling, 36 complexes were built with modified DNA and three with native DNA – non-extended and two extended by 3 and 5 nt (L0, L3 and L5). Simulations were also carried out for all DNA from these models.

During the construction of the protein–DNA complexes L3/X4/S2, L0/X4/S2, L0/X2/S2, and L4/X5/S2, significant steric clashes were observed between the *N*-benzimidazole–modified DNA residue and the surrounding protein residues. In these cases, either the initial models were too distorted to proceed with stable MD simulations, or during the early stages of simulation (within the first few nanoseconds), the S2 rotamer spontaneously converted to the S1 conformation to relieve the clashes. To enable simulations with the S2 rotamer, we started from the relaxed structure of the corresponding S1 complex and performed 25 ns of restrained MD simulation in which a flat-bottom harmonic potential was applied to the dihedral angle OP1–P–N–C to gradually drive the system toward the S2 conformation (during the first 0.2 ns, the force constant of the restraint was linearly increased from 0 to 1, while the flat-bottom potential was defined with "walls" at –130.0 to –125.0° and –115.0 to –110.0°, the force constant for the restraining potential was set to 200.0 kcal/mol/rad). Following this restrained relaxation, the rotamer of the modified residue adopted the desired S2 conformation within the protein–DNA complex. Subsequently, a 100-ns unrestrained production MD trajectory was generated from this stabilized structure. This trajectory was analyzed using the same protocols applied to all other simulated systems.

Conformational flexibility analysis

Stability of the protein–nucleic acid complex

During MD simulations, the protein structure in certain models underwent noticeable conformational rearrangements, as evidenced by a pronounced increase in root-mean-square deviation (RMSD) values for the protein backbone (Fig. S1)[1]. In these trajectories, the RMSD exhibited considerable fluctuations during the first 50 ns, indicating incomplete equilibration. To ensure robust and reliable analysis, we extended the

[1] Supplementary Figures S1–S10 and Tables S1–S6 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Berdyugin_Engl_29_7.pdf

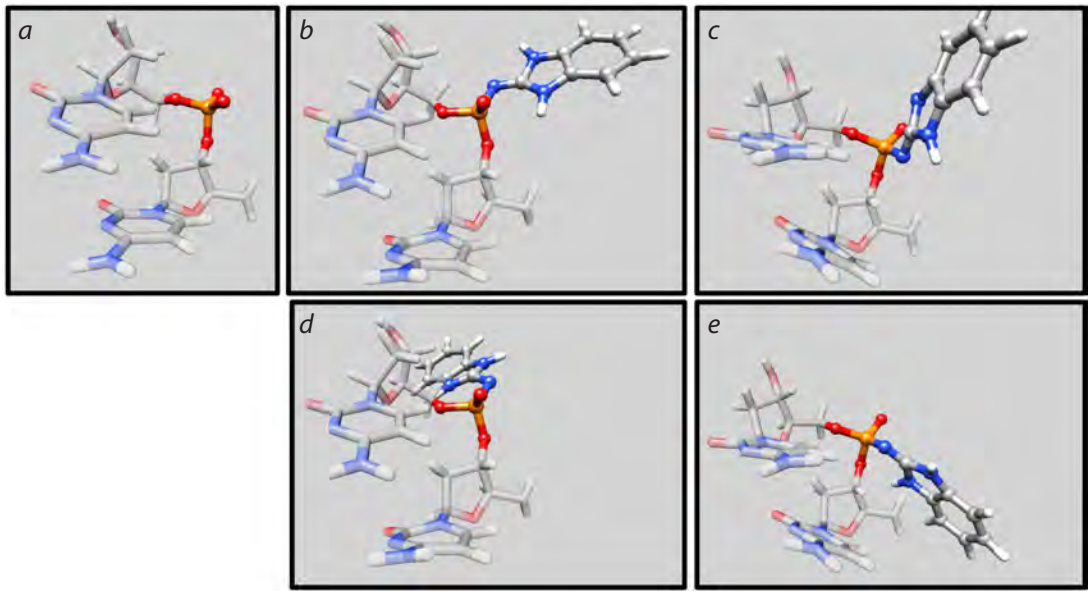


Fig. 2. Spatial structure of DNA dinucleotide steps: native (*a*) and modified for the studied stereoisomers and conformers (*b*) R1, (*c*) S1, (*d*) R2 and (*e*) S2.

simulations of these specific complexes by an additional 50 ns beyond the initial 100-ns run, allowing the systems to reach an equilibrium. The RMSD profiles for the full 150-ns trajectories are shown in Figure S1. For all subsequent structural and dynamic analyses, we used only the final 50 ns.

Analysis of the MD trajectories revealed that the single-stranded region of the template strand exhibited high conformational flexibility and, as expected, did not adopt any stable or preferred conformation during the simulations. Due to its intrinsic disorder and lack of defined structural features, this single-stranded segment was excluded from further structural analysis. Figure S2 shows the RMSD profiles along the trajectories for all studied complexes. It is evident that, over the 50-ns analysis segment, all structures remain stable, as indicated by the plateauing of RMSD values after an initial brief increase during the first 1–5 ns. The average RMSD value across all analyzed complexes is 2.63 ± 0.29 Å, with a mean standard deviation along the trajectory of 0.39 ± 0.11 Å.

Protein structural stability

To assess structural changes in the protein during MD simulations, RMSD time profiles were calculated for the protein Cα atoms over the last 50 ns of each trajectory, using the first frame of the respective analysis segment as the reference structure (Fig. S3). The presented data clearly indicate that, following initial relaxation during the first 50 ns, the protein structure remains highly stable in all modeled complexes.

The analysis of RMSD distributions across the trajectories, presented in Figure S4, shows that RMSD values remain within a narrow range, below 3.5 Å, and the distributions themselves are relatively sharp, confirming the high conformational stability of the protein throughout the simulations. The presence of multiple peaks in some RMSD distributions indicates that the system samples several distinct yet closely related conformational substates during the simulation. This observation is corroborated by the subsequent hierarchical cluster analysis (see below), which identifies multiple populated clusters corresponding to these substates. Importantly, the structural differences between these clusters are minor.

Stability of the DNA structure within the complex

To assess DNA structural changes during MD simulations, we calculated the RMSD over the last 50 ns of each trajectory, using the first frame of this segment as the reference structure (Fig. S5). For this analysis, we considered two distinct representations of the nucleic acid component: the duplex region only and the full DNA construct, including the single-stranded 5′-overhang of the template strand. This is attributed to the high conformational flexibility of the single-stranded overhang. As shown in the data, the duplex region of the DNA remains highly stable in all trajectories after the initial 50 ns. The RMSD analysis along the trajectories for DNA in complex with the protein performed both including and excluding the single-stranded template overhang revealed a significant difference in the average RMSD values and their standard deviations (averaged across all models). When the single-stranded overhang was included, the mean RMSD was 3.46 ± 0.97 Å, with a trajectory-wise standard deviation of 0.84 ± 0.31 Å. In contrast, when only the duplex region (primer–template hybrid) was considered, the mean RMSD dropped significantly to 1.97 ± 0.77 Å, with a much lower standard deviation of 0.39 ± 0.12 Å. Thus, to ensure a reliable and meaningful structural analysis, we excluded the single-stranded DNA segment from our evaluations, as it adopted highly variable conformations along the MD trajectories and did not exhibit a stable or functionally relevant orientation within the complex.

Stability of the structure for simulated free DNA

RMSD analysis of DNA trajectories in the absence of protein revealed significantly higher conformational mobility compared to the DNA within the Taq polymerase complex

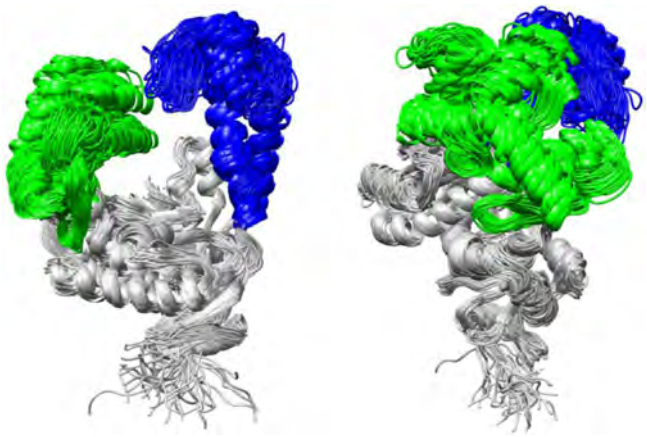


Fig. 3. Superposition of the most representative protein structures from the MD trajectories of all studied complexes, obtained by hierarchical clustering.

The palm domain is shown in gray, the thumb domain in blue, and the fingers domain in green. Protein structures were aligned based on the palm domain to highlight conformational differences in the mobile domains. The panel on the right shows the same superposition rotated by 90° around the vertical axis relative to the left panel, providing a side view of domain arrangements.

(Fig. S6). For the full DNA construct (including the single-stranded overhang), the average RMSD and its standard deviation (averaged across all models) were 5.11 ± 1.72 and 1.29 ± 0.61 Å, respectively. When the single-stranded region was excluded, these values decreased to 2.45 ± 0.41 and 0.50 ± 0.12 Å. These results clearly demonstrate that Taq polymerase substantially restricts the conformational flexibility of both the duplex and single-stranded regions of DNA upon complex formation. Moreover, the greater spread in RMSD values (evidenced by higher standard deviations) for free DNA indicates a broader ensemble of sampled conformations, whereas the protein-bound DNA adopts a more constrained and homogeneous structural state.

Analysis of protein, DNA, and protein–nucleic acid complex structures

To evaluate the impact of the *N*-benzimidazole modification on protein conformation, we calculated pairwise RMSD values between Cα atoms of the most representative structures (i. e., cluster centroids) extracted from the last 50 ns of each MD trajectory via hierarchical clustering. These RMSD values were used to construct a two-dimensional heatmap (Fig. S7), which visualizes structural similarities and differences across all simulated complexes. The analysis revealed that the average RMSD between native and modified complexes is very similar, with a mean value of ~2.60 Å, indicating that the overall protein fold is largely preserved regardless of the presence, position, or stereochemistry of the modification. However, when comparing individual modified systems, spanning different modification positions (X1–X6), stereoisomers (Rp/Sp), and rotamers (R1/R2, S1/S2), the pairwise RMSD values exhibit a broader range, from 1.31 to 4.37 Å. Notably, the average RMSD of each structure relative to all others falls within a relatively narrow interval of 2.33–3.26 Å (Table S1), confirming that all modeled complexes adopt globally similar conformations. The average RMSD values for each modification position, averaged over both stereoisomers and rotamers follow the trend: X1 < X2 < X6 < X4 < X3 < X5. This ordering indicates that modifications at positions X3 and X5 induce the largest structural perturbations in Taq polymerase, whereas modifications near the 3′-terminus (X1, X2) are best accommodated with minimal impact on the protein conformation. Furthermore, when RMSD values are averaged across all modification positions for each rotamer/stereoisomer type, the following trend emerges: S1 > R1 > R2 > S2. This sequence correlates directly with the spatial orientation of the *N*-benzimidazole group relative to the DNA duplex, the benzimidazole moiety toward the major groove leading to greater steric interference with polymerase residues.

Comparison of the most representative structures from the MD trajectories across all model complexes reveals that structural differences are primarily localized to the fingers and thumb domains, while the palm domain remains remarkably stable in all systems (Fig. 3). Additionally, the N-terminal region of the protein exhibits high conformational flexibility. Such variations are associated both with the conformational mobility of the thumb and fingers domains and with the effect of modification on their arrangement.

Structure of DNA

It is well established that nucleic acid (NA) substrates undergo significant conformational rearrangements upon binding to DNA polymerases compared to their solution-state structures (Vinogradova, Pyshnyi, 2010). Key structural changes commonly observed in experimentally determined polymerase–DNA complexes include: sugar pucker conformational shifts, narrowing of the minor groove, and induction of a pronounced bend in the DNA duplex at the active site. To characterize these effects in our systems, we compared the structures of the DNA substrate in the free state (i. e., without protein) and in complex with Taq DNA polymerase, using the most representative conformations identified by hierarchical clustering of the MD trajectories. RMSDs between the duplex regions of the free and protein-bound DNA structures were calculated for all combinations of stereoisomers (Rp and Sp), rotamers (R1/R2 and S1/S2), and extension states (elongated and nonelongated primers). These RMSD values are summarized in Table S2.

The average RMSD between the duplex regions of DNA in the free state and in complex with Taq polymerase across all modeled systems is approximately 2.4 Å. The largest structural deviation was observed for the L0/X4/S1 complex, with an RMSD of 3.3 Å. This pronounced difference is attributed to a marked widening of the minor groove in the protein-bound state. In this orientation, the modification effectively shields the nucleobases from solvent exposure and induces local stretching of the sugar–phosphate backbone. In contrast, the smallest RMSD values (i. e., the highest structural similarity between free and bound DNA) were found for modifications at positions X5 and X6 (Table S2). Furthermore, the RMSD between unmodified and modified DNA substrates – both in complex with Taq polymerase – averages ~1.75 Å. Notably, this deviation is smaller for modifications oriented toward the major groove, as these conformers minimize direct contacts with the protein.

The average RMSD which computed across all rotamers and stereoisomers for the DNA duplex in complex with Taq polymerase is approximately 2.0 Å. Lower RMSD values are observed for systems in which the *N*-benzimidazole modification adopts a consistent spatial orientation. Structural analysis further reveals that, even in cases of pronounced interactions between the modification and protein residues, the overall architecture of the duplex region remains largely unperturbed. In general, the structure of a substrate with a modification largely depends on which regions of the protein it interacts with, which is determined by both the isomer and the conformer of the *N*-benzimidazole residue.

The structural parameters of the investigated nucleic acid substrates are predominantly characteristic of B-form DNA. However, localized deviations from ideal B-form DNA are observed in the vicinity of the 3′-end of the primer and at the site of *N*-benzimidazole modification. In particular, for nonelongated model systems (L0), a pronounced increase in the Roll and Buckle parameters was detected for AT base pairs adjacent to the catalytic center. For both extended and unextended complexes, the propeller twist angle of these AT base pairs was consistently negative, a feature more typical for A-tract DNA than canonical B-DNA (Strahs, Schlick, 2000). The Inclination of base pairs relative to the helical axis increased the closer the *N*-benzimidazole modification was positioned to the catalytic center. In contrast, this deviation markedly decreased in complexes with an elongated duplex region (L1–L5). Notably, the average Twist value across all systems remained approximately 34°, independent of duplex length or the presence and position of the modification. This constancy in Twist suggests that the helical packing density of the DNA duplex is largely preserved.

In all studied complexes, a significant widening of the DNA minor groove (defined as the distance between phosphorus atoms on opposite strands) was observed in the region adjacent to the catalytic center, reaching 15–18 Å. In modified complexes, this widening increased further with the length of the duplex region (i. e., in L1–L5 systems), which corresponds to the progressive displacement of the modification away from the 3′-end of the primer. In contrast, native (unmodified) complexes exhibited a much smaller degree of minor groove width increase. No clear correlation was found between the structural parameters of the nucleic acid substrate and the specific spatial orientation of the modification. This suggests that the position of the modification relative to the 3′-primer terminus dominates its impact on global DNA conformation within the polymerase complex.

Analysis of sugar pucker conformations in the DNA duplex reveals that, in most cases, deoxyribose adopts the C2′-endo conformation which is characteristic of canonical B-form DNA. However, near the 3′-end of the primer, specific nucleotides, particularly those adjacent to the catalytic site, exhibit C1′-exo or O4′-endo sugar puckers. These non-canonical sugar conformations are indicative of local structural strain and are commonly associated with the catalytically active state of DNA polymerases.

The presence of the modification in the DNA strand within the Taq polymerase complex caused significant deviation from canonical planar base pairing only in the case of terminal and penultimate base pairs when the modification was located at the first or second position of the primer. Structural analysis shows that the modification does not affect the nature of base pairing: Watson–Crick pairs with standard hydrogen bond lengths are formed, except for the terminal base pairs – a finding previously observed both experimentally and in MD simulations (Nonin et al., 1995; Zgarbová et al., 2014). Thus, the modification at the first internucleotide phosphate residue exerted the greatest influence on the local DNA structure within the polymerase complex. Overall, the presence of the modification does not significantly alter the DNA structure, either in free duplexes or in the enzyme–substrate complex.

An analysis of the *N*-benzimidazole group orientation within the DNA duplex was performed for both the free state and the protein-bound complex. This was done by examining the dihedral angle around the P–N bond, defined by the non-bridging phosphate oxygen (OP1 for the Rp isomer and OP2 for the Sp isomer), the phosphorus atom, the nitrogen atom, and the carbon atom of the benzoazole ring. The analysis revealed considerable flexibility of the modified residue and the possibility of interconversion between rotameric states (Fig. S8).

Population analysis of the dihedral angles along the MD trajectories shows that, for both elongated and nonelongated systems, free DNA exhibits generally similar conformational preferences (Fig. S8, S9). The data indicate that the Rp isomer of the modified residue is predominantly oriented toward the minor groove, whereas the Sp isomer preferentially points toward the major groove, corresponding to a dihedral angle of approximately +100°. In some cases, the modification flips away from the duplex, corresponding to an angle of about –100° (rotamers R1 and S2). The lower population of this outward orientation is attributed to the hydrophobic nature of the benzimidazole group, which tends to minimize solvent exposure by interacting with the DNA strands. In most cases, the distributions for the two stereoisomers are qualitatively similar: when two peaks are present for one isomer, they are typically also observed for the other. Differences in peak amplitudes suggest that the conformational space for the modification is not fully sampled within the 50-ns trajectory of each individual model. However, when the angular probability distributions are aggregated across all modification positions for each stereoisomer, the average dihedral angles for rotamers 1 and 2 of each isomer nearly coincide (Fig. 4), indicating consistent conformational preferences irrespective of modification position.

In the protein-bound complex, the orientation of the modification undergoes significant changes compared to free DNA (Fig. 4). The plots of dihedral angle values and their probability distributions (Fig. 4, S9, and S10) show that, along the MD trajectories, angles are observed not only between the two main peaks characteristic of free DNA (+100° and –100°), but also shifted beyond these values to larger absolute magnitudes. This indicates substantial interactions between the modified residue and the protein, which constrain and redirect the conformational preferences of the *N*-benzimidazole group relative to its behavior in the unbound state.

Comparison of the average probability distributions for different stereoisomers in the complexes shows that they differ significantly both from each other across modification positions and from the distributions observed for free DNA

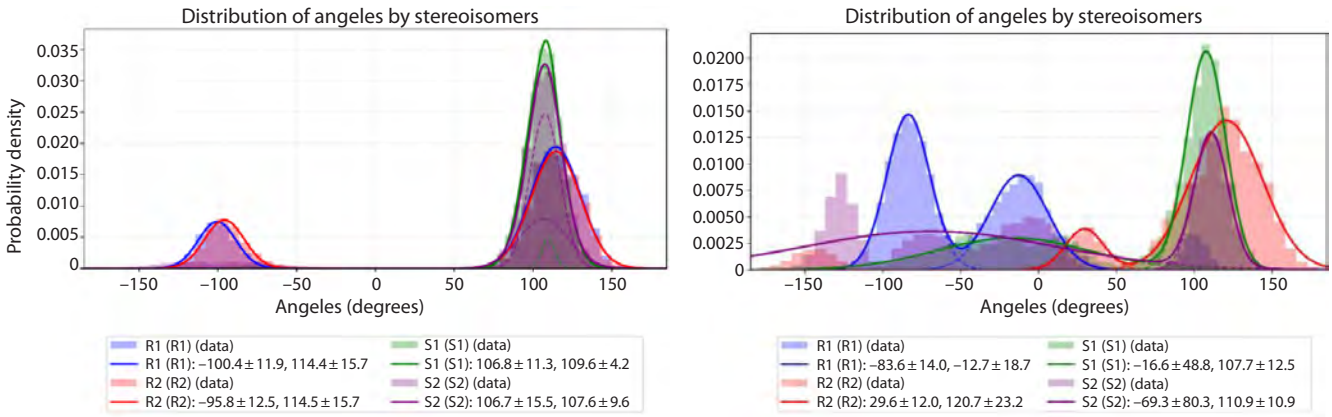


Fig. 4. Dihedral angle values of the P–N bond in the phosphoramidate linkage for rotamers 1 and 2 along MD trajectories of free DNA (left) and DNA in complex with the protein (right), aggregated across all studied models.

(Fig. 4, S8, and S10). Notably, the probability distributions for rotamers R1 and R2 are markedly distinct. The main peak for R1 is located around –80°, corresponding to an orientation of the modification toward the major groove (i. e., away from the DNA helix). This is attributed to the fact that, in the polymerase complex, the native phosphate backbone is tightly coordinated by specific amino acid residues; consequently, the bulkier phosphoramidate modification is sterically expelled from the minor groove. In contrast, the primary peak for R2 appears near +100°, indicating that the modification is directed into the minor groove. For the S1 rotamer, the dominant angle is +100°, but the modification is oriented toward the major groove – a consequence of the opposite stereochemistry at the phosphorus center compared to the Rp series. The S2 rotamer exhibits a markedly different behavior: its probability distribution shows multiple peaks of comparable amplitude spread across nearly the entire angular range, indicating that the modification can adopt diverse spatial orientations depending on its position in the primer chain (Xj). This conformational heterogeneity is driven by specific, position-dependent interactions with the protein environment.

It should be noted that, for all examined stereoisomers, a distinct peak appears around 0° (Fig. 4), corresponding to an orientation in which the modification points away from the DNA helix. In this conformation, one of the amino groups of the five-membered ring of the *N*-benzimidazole moiety forms a hydrogen bond with the non-bridging oxygen atom of the adjacent phosphate group. The absence of such orientations in free DNA indicates that this conformation is specifically stabilized by additional interactions with the protein, highlighting the role of the polymerase in shaping the conformational landscape of the modified backbone.

Analysis of interactions of modification with Taq polymerase

A hierarchical cluster analysis of the last 50 ns of each MD trajectory was performed to identify the most representative structures. The spatial arrangement of the *N*-benzoazole groups relative to the polymerase active site was examined, and the number of protein atoms in contact with the modification was quantified. Contact maps between the modification and Taq polymerase were also generated. All amino acid residues with at least one atom located within 3 Å of the modified phosphate group were considered to be in direct interaction with the modification (Tables S3 and S4). The DNA duplex region that engages with Taq polymerase spans 5–8 base pairs, and approximately 40 amino acid residues participate in this interaction. These residues are involved in nucleic acid recognition, substrate stabilization, and catalysis (Eom et al., 1996; Li Y. et al., 1998).

Analysis of contacts between the phosphoramidate *N*-benzimidazole moiety and Taq polymerase revealed several key patterns. First, in the complexes L0/X1/R1, L1/X2/R2, L0/X3/R2, L2/X3/R1, and L4/X5/R2, the *N*-benzimidazole group was accommodated within protein pockets. Moreover, for the fourth modification position (X4) with the R stereoisomer, both rotamers (X4/R1 and X4/R2) occupied a pocket, forming stable interactions between the modification's electronegative atoms and the protein's positively charged arginine residues (Fig. 5).

Overall, modifications at positions 1–5 form an extensive network of hydrogen bonds and van der Waals contacts with the protein, whereas interactions for the 6th position are considerably weaker. Stereochemistry also strongly influences the binding mode: Sp stereoisomers preferentially interact

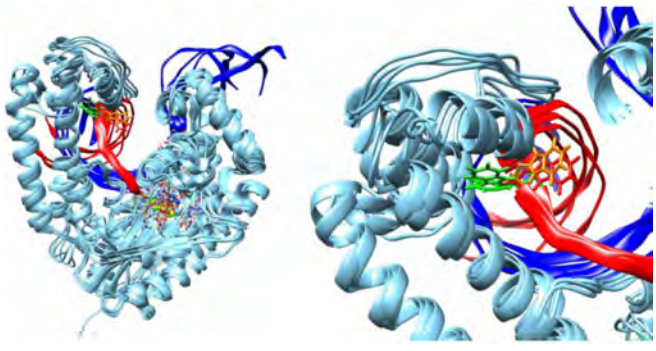


Fig. 5. Structural comparison of the L0/X4 complexes: overall view (left) and close-up of the modification interaction region with the thumb domain of the enzyme (right).

Taq DNA polymerase is shown in blue, the DNA template strand in blue, and the primer in red. The modified *N*-benzoazole groups are displayed as atomic models, with Sp isomers colored red and orange, and Rp isomers in light and dark green.

with positively charged residues, while Rp stereoisomers more frequently engage in contacts with hydrophobic amino acids. Sp isomers are often oriented toward the major groove, effectively shielding the heterocyclic bases of the duplex from solvent exposure. In contrast, Rp isomers are predominantly directed away from the DNA and toward the protein surface. The presence of the modification frequently disrupts the regular nucleic acid structure due to interactions of the *N*-benzimidazole group with protein pockets, which induce strain in the sugar–phosphate backbone. Introduction of the modification at the first or second position of the primer leads to significant distortion of the terminal and penultimate base pairs. Moreover, in the complexes L1/X2/S2, L0/X3/R1, and L0/X2/R1, disruption of Watson–Crick base pairing near the modification site is observed.

According to the literature data, residue Arg660 from the fingers domain coordinates the phosphate group at the first position of the primer from the 3′-end, Arg587 from the palm domain coordinates the second internucleotide phosphate, and Arg536 from the thumb domain interacts with the fourth phosphate (Vinogradova, Pyshnyi, 2010). The presence of the *N*-benzimidazole modification is expected to neutralize the negative charge of the phosphate group and introduce steric hindrance that impedes coordination of the phosphate by arginine residues, which should reduce the catalytic rate. However, structural analysis shows that, in the case of Sp isomers at positions 2 and 4, the non-bridging oxygen atom of the phosphate moiety is still coordinated by Arg536 for both rotamers. Similarly, for Rp isomers with the modification at positions 1 or 2, at least one rotamer retains coordination of the phosphate oxygen by the corresponding arginine residue.

Both rotamers of the Rp isomer at the fourth position are accommodated within a hydrophobic pocket of the thumb domain, whereas the Sp isomer shows minimal interaction with the protein. As a result, the Rp-modified phosphate group impedes translocation of the polymerase to the next position along the DNA strand, which is required for incorporation of the subsequent nucleotide onto the primer. This steric and dynamic blockage most likely explains the accumulation of incomplete elongation products observed experimentally when the modification is located at the fourth position.

We have previously shown (Golyshev et al., 2025) that in primer elongation experiments with Taq DNA polymerase using primers bearing the *N*-benzimidazole modification, incorporation of the modification at the second position results in the smallest reduction in elongation efficiency for perfectly matched complexes. This correlates with the lowest number of contacts observed between the modification and the protein among the first three internucleotide phosphate positions. Furthermore, in all perfectly matched modified complexes, a distinct band corresponding to a partially extended primer, with the modification located at the 4th position from the 3′-end, was clearly observed. This effect is most pronounced for primers carrying modifications at the 1st and 3rd positions. These experimental observations correlate well with structural data showing that both rotamers of the R stereoisomer at position 4 (X4/R1 and X4/R2) are accommodated within a protein pocket and form stable interactions with the enzyme (Fig. 5).

Thus, steric interactions of Rp isomers with protein pockets can slow down – or, as in the case of the fourth modification position, block – the translocation of Taq DNA polymerase along the substrate. This is experimentally confirmed by the reduced polymerization rate and the appearance of abortive elongation products of the modified primer containing the phosphoramidate *N*-benzimidazole group.

Substrate–polymerase interaction energy

The interactions described in the previous section are reflected in the binding energetics between the enzyme and its substrate. Therefore, we calculated the interaction energy between the nucleic acid substrate and Taq polymerase using the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) calculation method, based solely on the MD trajectory of the protein–DNA complex. To minimize fluctuations in the computed free energy arising from the high flexibility of the single-stranded template overhang, only the duplex region of the nucleic acid substrate was included in the energy calculations. The energies of the DNA, protein, their complex, and the resulting binding (complexation) energies are reported in Tables S5 and S6. Analysis of the interaction energies between the modified nucleic acid substrates and Taq polymerase revealed the following trends: 1) for native (unmodified) complexes, the binding energy (in absolute value) increased with duplex length, reflecting stronger stabilization of longer primer–template hybrids within the polymerase active site; 2) in contrast, no clear correlation was observed between binding energy and duplex length for modified complexes; 3) notably, modifications at the 5th and 6th internucleotide phosphate positions exhibited weaker binding compared to all other model systems, which correlates with the reduced number of contacts between DNA and the protein observed in these cases.

In the case of nonelongated model systems (L0), which have the shortest duplex region, the complexation energy was, on average, significantly lower (~ –200 kcal/mol) than that of extended complexes (~ –180 kcal/mol). For the majority of complexes, S stereoisomers exhibited more favorable (i. e., more negative) binding energies compared to their Rp counterparts. This is likely due to the greater accessibility of the non-bridging oxygen atom of the modified phosphate group in the Sp configuration, facilitating its coordination by protein residues. Among the two rotamers, the S1 conformation – in which the *N*-benzimidazole group is oriented toward the major groove – consistently displayed the most favorable binding energy, as this orientation leaves the non-bridging phosphate oxygen exposed for interaction with amino acid side chains. It should be noted, no direct correlation was found between the number of protein atoms in proximity to the modification (Table S3) and the computed binding energy. However, the strongest enzyme–substrate binding was observed for the complexes L0/X4/R1 and L0/X4/R2, in which the modification is buried within a protein pocket and engages with the largest number of amino acid residues (Table S4).

Conclusion

In this work, we employed molecular simulation and analysis to investigate the structure, dynamics, and interaction energetics of DNA substrates containing a phosphoramidate *N*-benzimidazole group at various positions within the primer strand in complex with Taq DNA polymerase. We found that

both the position of the modification near the 3′-end of the primer and its stereochemistry significantly influence interactions with the enzyme. Within the enzyme–substrate complex, two stable rotamers were identified for each phosphoramidate stereoisomer (Rp and Sp). Analysis of the stereochemical effects revealed that Rp isomers generally exhibit stronger interactions with the polymerase, with the most pronounced binding observed when the modification is located at the fourth internucleotide phosphate from the 3′-end of the primer. Structural analysis of both DNA and protein showed no major global rearrangements in either biopolymer upon modification. Structural perturbations induced by the *N*-benzimidazole group were either minor or strictly localized. The greatest impact on local DNA conformation within the polymerase complex was observed for modifications at the first internucleotide phosphate position.

These computational findings correlate well with experimental data on the processing of PABAO primers by Taq DNA polymerase. In particular, they explain: 1) the reduced rate of full-length product formation for modified primers, 2) the accumulation of incomplete elongation products when the modification is located at the fourth position from the 3′-end of the primer, and 3) the significant decrease in primer elongation efficiency upon modification at the first position (Chubarov et al., 2024; Golyshev et al., 2025).

The results of this study provide a molecular basis for understanding how the phosphoramidate *N*-benzimidazole group affects the elongation of PABAO primers. These insights will be instrumental in the rational design of PABAO structures for applications in molecular diagnostics using PCR-based methods. Furthermore, the pronounced differences in polymerase interaction efficiency between Rp and Sp isomers of PABAOs highlight the need to develop stereoselective synthesis methods for these oligonucleotides. Such approaches would enable precise control over the stereochemistry of the phosphoramidate linkage, thereby allowing fine-tuning of the biochemical and biophysical properties of phosphoramidate benzazole oligonucleotides for optimized performance in diagnostic assays.

References

Abramson J., Adler J., Dunger J., Evans R., Green T., Pritzel A., Ronneberger O., … Bapst V., Kohli P., Jaderberg M., Hassabis D., Jumper J.M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500. doi 10.1038/s41586-024-07487-w

Case D.A., Belfon K., Ben-Shalom I.Y., Brozell S.R., Cerutti D.S., Cheatham T.E. III, Cruzeiro V.W.D., … Wu X., Xiong Y., Xue Y., York D.M., Kollman P.A. Amber 20. San Francisco, Univ. of California, 2020. Available at: https://ambermd.org/doc12/Amber20.pdf

Chubarov A.S., Oscorbin I.P., Filipenko M.L., Lomzov A.A., Pyshnyi D.V. Allele-specific PCR for *KRAS* mutation detection using phosphoryl guanidine modified primers. *Diagnostics*. 2020;10(11):872. doi 10.3390/diagnostics10110872

Chubarov A.S., Oscorbin I.P., Novikova L.M., Filipenko M.L., Lomzov A.A., Pyshnyi D.V. Allele-specific PCR for PIK3CA mutation detection using phosphoryl guanidine modified primers. *Diagnostics*. 2023;13(2):250. doi 10.3390/diagnostics13020250/S1

Chubarov A.S., Baranovskaya E.E., Oscorbin I.P., Yushin I.I., Filipenko M.L., Pyshnyi D.V., Vasilyeva S.V., Lomzov A.A. Phosphoramidate azole oligonucleotides for single nucleotide polymorphism detection by PCR. *Int J Mol Sci*. 2024;25(1):617. doi 10.3390/ijms25010617

Di Giusto D., King G.C. Single base extension (SBE) with proofreading polymerases and phosphorothioate primers: improved fidelity in single-substrate assays. *Nucleic Acids Res.* 2003;31(3):e7. doi 10.1093/nar/gng007

Eom S.H., Wang J., Steitz T.A. Structure of Taq polymerase with DNA at the polymerase active site. *Nature*. 1996;382(6588):278-281. doi 10.1038/382278A0

Golyshev V.M., Yushin I.I., Gulyaeva O.A., Baranovskaya E.E., Lomzov A.A. Properties of phosphoramide benzoazole oligonucleotides (PABAOs). I. Structure and hybridization efficiency of *N*-benzimidazole derivatives. *Biochem Biophys Res Commun*. 2024;693:149390. doi 10.1016/j.bbrc.2023.149390

Golyshev V.M., Morozova F.V., Berdugin A.A., Kozyreva E.A., Baranovskaya E.E., Yushin I.I., Lomzov A.A. Structural and thermodynamic insights for enhanced SNP detection using *N*-benzimidazole oligonucleotides. *J Phys Chem B*. 2025;129(44):11409-11420. doi 10.1021/acs.jpcb.5c04047

Ishige T., Itoga S., Matsushita K. Locked nucleic acid technology for highly sensitive detection of somatic mutations in cancer. *Adv Clin Chem*. 2018;83:53-72. doi 10.1016/bs.acc.2017.10.002

Izadi S., Anandakrishnan R., Onufriev A.V. Building water models: a different approach. *J Phys Chem Lett*. 2014;5(21):3863-3871. doi 10.1021/jz501780a

Kalendar R., Baidyussen A., Serikbay D., Zotova L., Khassanova G., Kuzbakova M., Jatayev S., Hu Y.G., Schramm C., Anderson P.A., Jenkins C.L.D., Soole K.L., Shavrukov Y. Modified "Allele-specific qPCR" method for SNP genotyping based on FRET. *Front Plant Sci*. 2022;12:747886. doi 10.3389/fpls.2021.747886

Kutyavin I.V. Use of base modifications in primers and amplicons to improve nucleic acids detection in the real-time snake polymerase chain reaction. *Assay Drug Dev Technol*. 2011;9(1):58-68. doi 10.1089/adt.2010.0303

Li Y., Korolev S., Waksman G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J*. 1998;17(24):7514-7525. doi 10.1093/emboj/17.24.7514

Li Z., Song L.F., Li P., Merz K.M. Systematic parametrization of divalent metal ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB water models. *J Chem Theory Comput*. 2020;16(7):4429-4442. doi 10.1021/acs.jctc.0c00194

Meagher K.L., Redman L.T., Carlson H.A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem*. 2003;24(9):1016-1025. doi 10.1002/jcc.10262

Nonin S., Leroy J.L., Guéron M. Terminal base pairs of oligodeoxynucleotides: imino proton exchange and fraying. *Biochemistry*. 1995;34(33):10652-10659. doi 10.1021/bi00033a041

Novgorodtseva A.I., Vorob'ev A.Y., Lomzov A.A., Vasilyeva S.V. Synthesis and physicochemical properties of new phosphoramide oligodeoxyribonucleotides. I. N-caffeine derivatives. *Bioorg Chem*. 2025;157:108313. doi 10.1016/j.bioorg.2025.108313

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi 10.1002/jcc.20084

Rejali N.A., Moric E., Wittwer C.T. The effect of single mismatches on primer extension. *Clin Chem*. 2018;64(5):801-809. doi 10.1373/clinchem.2017.282285

Roe D.R., Cheatham T.E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*. 2013;9(7):3084-3095. doi 10.1021/ct400341p

Shapovalov M.V., Dunbrack R.L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*. 2011;19(6):844-858. doi 10.1016/j.str.2011.03.019

Starza I.D., Eckert C., Drandi D., Cazzaniga G.; EuroMRD Consortium. Minimal residual disease analysis by monitoring immunoglobulin and T-cell receptor gene rearrangements by quantitative PCR and droplet digital PCR. *Methods Mol Biol.* 2022;2453:79-89. doi 10.1007/978-1-0716-2115-8_5

Strahs D., Schlick T. A-tract bending: insights into experimental structures by computational models. *J Mol Biol.* 2000;301(3):643-663. doi 10.1006/jmbi.2000.3863

Terpe K. Overview of thermostable DNA polymerases for classical PCR applications: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol.* 2013;97(24):10243-10254. doi 10.1007/s00253-013-5290-2

Tian C., Kasavajhala K., Belfon K.A.A., Raguette L., Huang H., Migues A.N., Bickel J., Wang Y., Pincay J., Wu Q., Simmerling C. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J Chem Theory Comput.* 2020;16(1):528-552. doi 10.1021/acs.jctc.9b00591

Unni S., Huang Y., Hanson R.M., Tobias M., Krishnan S., Li W.W., Nielsen J.E., Baker N.A. Web servers and services for electrostatics calculations with APBS and PDB2PQR. *J Comput Chem.* 2011; 32(7):1488-1491. doi 10.1002/jcc.21720

Vasilyeva S.V., Baranovskaya E.E., Dyudeeva E.S., Lomzov A.A., Pyshnyi D.V. Synthesis of oligonucleotides carrying inter-nucleotide *N*-(benzoazole)-phosphoramide moieties. *ACS Omega*. 2023; 8(1):1556-1566. doi 10.1021/acsomega.2c07083

Vinogradova O.A., Pyshnyi D.V. Selectivity of enzymatic conversion of oligonucleotide probes during nucleotide polymorphism analysis of DNA. *Acta Naturae*. 2010;2(1):40-58. doi 10.32607/ 20758251-2010-2-1-36-52

Yushin I.I., Golyshev V.M., Novgorodtseva A.I., Lomzov A.A. Properties of phosphoramide benzoazole oligonucleotides (PABAOs). II. Structure and hybridization efficiency of *N*-benzoxazole derivatives. *Biochem Biophys Res Commun.* 2024;740:150997. doi 10.1016/j.bbrc.2024.150997

Zgarbová M., Otyepka M., Šponer J., Lankaš F., Jurečka P. Base pair fraying in molecular dynamics simulations of DNA and RNA. *J Chem Theory Comput.* 2014;10(8):3177-3189. doi 10.1021/ct500120v

Zgarbová M., Šponer J., Jurečka P. Z-DNA as a touchstone for additive empirical force fields and a refinement of the Alpha/Gamma DNA torsions for AMBER. *J Chem Theory Comput.* 2021;17(10):6292-6301. doi 10.1021/acs.jctc.1C00697

Conflict of interest. The authors declare no conflict of interest.
Received July 31, 2025. Revised September 9, 2025. Accepted September 9, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1084-1096

doi 10.18699/vjgb-25-113

Prediction of interactions between the SARS-CoV-2 ORF3a protein and small-molecule ligands using the ANDSystem cognitive platform, graph neural networks, and molecular modeling

T.V. Ivanisenko, P.S. Demenkov, M.A. Kleshchev, V.A. Ivanisenko

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
itv@bionet.nsc.ru

Abstract. In recent years, artificial intelligence methods based on the analysis of heterogeneous graphs of biomedical networks have become widely used for predicting molecular interactions. In particular, graph neural networks (GNNs) effectively identify missing edges in gene networks – such as protein–protein interaction, gene–disease, drug–target, and other networks – thereby enabling the prediction of new biological relationships. To reconstruct gene networks, cognitive systems for automatic text mining of scientific publications and databases are often employed. One such AI-driven platform, ANDSystem, is designed for automatic knowledge extraction of molecular interactions and, on this basis, the reconstruction of associative gene networks. The ANDSystem knowledge base contains information on more than 100 million interactions among diverse molecular genetic entities (genes, proteins, metabolites, drugs, etc.). The interactions span a wide range of types: regulatory relationships, physical interactions (protein–protein, protein–ligand), catalytic and chemical reactions, and associations among genes, phenotypes, diseases, and more. In the present study, we applied attention-based graph neural networks trained on the ANDSystem knowledge graph to predict new edges between proteins and ligands and to identify potential ligands for the SARS-CoV-2 ORF3a protein. The accessory protein ORF3a plays an important role in viral pathogenesis through ion-channel activity, induction of apoptosis, and the ability to modulate endolysosomal processes and the host innate immune response. Despite this broad functional spectrum, ORF3a has been explored far less as a pharmacological target than other viral proteins. Using a graph neural network, we predicted five small molecules of different origins (metabolites and a drug) that potentially interact with ORF3a: N-acetyl-D-glucosamine, 4-(benzoylamino)benzoic acid, austocystin D, bictegravirum, and L-threonine. Molecular docking and MM/GBSA affinity estimation indicate the potential ability of these compounds to form complexes with ORF3a. Localization analysis showed that the binding sites of bictegravir and 4-(benzoylamino)benzoic acid lie in a cytosolic surface pocket of the protein that is solvent-exposed; L-threonine binds within the intersubunit cleft of the dimer; and austocystin D and N-acetyl-D-glucosamine are positioned at the boundary between the cytosolic surface and the transmembrane region. The accessibility of these binding sites may be reduced by the influence of the lipid bilayer. The binding energetics for bictegravirum were more favorable than for 4-(benzoylamino)benzoic acid (docking score −7.37 kcal/mol; MM/GBSA ΔG −14.71 ± 3.12 kcal/mol), making bictegravirum a promising candidate for repurposing as an ORF3a inhibitor.

Key words: ANDSystem; SARS-CoV-2; ORF3a; gene networks; graph neural networks; protein–ligand interaction prediction; bictegravirum; 4-(benzoylamino)benzoic acid; molecular docking; potential therapeutic agents

For citation: Ivanisenko T.V., Demenkov P.S., Kleshchev M.A., Ivanisenko V.A. Prediction of interactions between the SARS-CoV-2 ORF3a protein and small-molecule ligands using the ANDSystem cognitive platform, graph neural networks, and molecular modeling. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1084-1096. doi 10.18699/vjgb-25-113

Funding. This study was funded by the budgetary project of the Federal Research Center Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), "Systems biology and bioinformatics: reconstruction, analysis, and modeling of the structural-functional organization and evolution of gene networks in humans, animals, plants, and microorganisms" No. FWNR-2022-0020.

Предсказание взаимодействий белка ORF3a SARS-CoV-2 с низкомолекулярными лигандами с использованием когнитивной платформы ANDSystem, графовых нейронных сетей и молекулярного моделирования

Т.В. Иванисенко, П.С. Деменков, М.А. Клещев, В.А. Иванисенко

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
itv@bionet.nsc.ru

Аннотация. В последние годы методы искусственного интеллекта, основанные на анализе гетерогенных графов биомедицинских сетей, получили широкое распространение для предсказания молекулярных взаимодействий. В частности, графовые нейронные сети (graph neural networks, GNN) позволяют эффективно выявлять отсутствующие ребра в генных сетях, таких как сети белок-белковых взаимодействий, ген–заболевание, лекарство–мишень и др., и тем самым предсказывать новые биологические связи. Для реконструкции генных сетей часто применяют когнитивные системы автоматического анализа текстов научных публикаций и баз данных. Одна из таких платформ, базирующаяся на методах искусственного интеллекта,– ANDSystem, предназначенная для автоматического извлечения знаний о молекулярных взаимодействиях и на этой основе – реконструкции ассоциативных генных сетей. База знаний ANDSystem содержит сведения о более чем 100 млн взаимодействий между различными молекулярно-генетическими объектами (гены, белки, метаболиты, лекарства и др.). Взаимодействия представлены широким спектром типов: регуляторные связи, физические взаимодействия (белок–белок, белок–лиганд), каталитические и химические реакции, ассоциации между генами, фенотипами, заболеваниями и др. В настоящем исследовании мы применили графовые нейронные сети с механизмом внимания, обученные на графе знаний ANDSystem, для предсказания новых ребер между белками и лигандами и поиска потенциальных лигандов для белка ORF3a SARS-CoV-2. Вспомогательный белок ORF3a SARS-CoV-2 играет важную роль в патогенезе вируса за счет ион-канальной активности, индукции апоптоза и способности модулировать эндолизосомальные процессы и врожденный иммунитет хозяина. Несмотря на широкий спектр функций, ORF3a как фармакологическая мишень изучен значительно меньше, чем другие вирусные белки. Применение графовой нейронной сети позволило нам предсказать пять малых молекул разного происхождения (метаболиты и лекарство), потенциально взаимодействующих с ORF3a: N-ацетил-D-глюкозамин, 4-(бензоиламино)бензойная кислота, аустоцистин D, биктегравир и L-треонин. Молекулярный докинг и оценка аффинности методом MM/GBSA подтвердили потенциальную способность этих соединений образовывать комплексы с ORF3a. Анализ локализации показал, что сайты связывания биктегравира и 4-(бензоиламино)бензойной кислоты расположены в цитозольной поверхностной области белка, доступной растворителю; L-треонин связывается в межсубъединичной щели димера, а аустоцистин D и N-ацетил-D-глюкозамин – на границе между цитозольной поверхностью и трансмембранной областью. Доступность этих сайтов связывания может быть снижена из-за влияния липидного бислоя. Энергетические характеристики связывания у биктегравира по сравнению с 4-(бензоиламино)бензойной кислотой оказались более высокими (–7.37 ккал/моль в докинге; –14.71 ± 3.12 ккал/моль по MM/GBSA), что делает его перспективным кандидатом для репозиционирования как ингибитора ORF3a. Взаимодействие биктегравира с ORF3a может нарушать связывание ORF3a с белком хозяина VPS39 – субъединицей комплекса HOPS, участвующего в слиянии аутофагосом и поздних эндосом с лизосомами. Это, в свою очередь, может снимать индуцируемую ORF3a блокаду данного процесса и тем самым способствовать восстановлению аутофагического потока и лизосомной деградации вирусных компонентов.

Ключевые слова: ANDSystem; SARS-CoV-2; ORF3a; генные сети; графовые нейронные сети; предсказание белок–лиганд взаимодействий; биктегравир; 4-(бензоиламино)бензойная кислота; молекулярный докинг; потенциальные лекарства

Introduction

The development of antiviral drugs is a priority due to the risk of global pandemics and the emergence of new variants of pathogenic viruses during such events, as demonstrated by the COVID-19 pandemic caused by SARS-CoV-2 (Ng et al., 2022). SARS-CoV-2 is an enveloped betacoronavirus with a positive-sense single-stranded RNA genome of approximately 29.9 kb; the genome encodes structural (S, E, M, N) as well as several nonstructural proteins that ensure replication and virion assembly (Naqvi et al., 2020). Because these proteins determine key stages of the viral life cycle, drug development efforts have focused primarily on three main targets: the main protease (3CLpro/Mpro), the RNA-dependent RNA polymerase (RdRp), and the S glycoprotein (Spike protein) (Boby et al., 2023).

A combination of experimental and computational approaches has been used to discover and optimize inhibitors of these targets: *de novo* design, high-throughput screening, and repurposing of known drugs (von Delft et al., 2023). This approach has yielded compounds with confirmed antiviral activity *in vitro* and *in vivo* and has enabled clinical strategies for treating COVID-19, including protease and polymerase inhibition. In particular, the antiviral nirmatrelvir/ritonavir (Paxlovid), which targets the main protease Nsp5 (nonstructural protein 5) of SARS-CoV-2, received full FDA approval on May 25, 2023, for the treatment of adults with COVID-19 (FDA, 2023). The drug remdesivir (Veklury), which targets the viral RNA-dependent RNA polymerase (RdRp, nsp12), was approved by the FDA in October 2020 (FDA, 2020). In parallel, alternative approaches are being developed to block fusion of the viral and cellular membranes during SARS-CoV-2 entry. In particular, peptide inhibitors complementary to the HR1/HR2 domains of the S2 subunit of the Spike protein prevent formation of the six-helix bundle (6-HB) – a key structure that mediates membrane fusion – and thereby block viral entry (Dong et al., 2024).

Among the promising classes of pharmacological targets are accessory viral proteins that modulate the interactions of SARS-CoV-2 with host cellular systems. One such protein is ORF3a. It is predominantly localized to late endosomes and

lysosomes, where it co-localizes with the human lysosomal proteins LAMP1 and cathepsin D (Zhang J. et al., 2021; Hinkle et al., 2025). ORF3a forms ion channels (viroporin activity) (Zhang J. et al., 2022), induces apoptosis through oxidative stress and caspase activation (Zhang Y. et al., 2021), activates the NLRP3 inflammasome (the ORF3a–NLRP3–ASC cascade) (Zhang J. et al., 2022), and suppresses interferon signaling pathways, thereby enhancing viral pathogenicity (Zhang J. et al., 2022).

ORF3a is a dimeric membrane protein with three transmembrane helices and a large cytosolic C-terminal domain, as shown by cryo-EM (Kern et al., 2021). It interacts with the human protein VPS39 – a component of the HOPS complex – and this interaction blocks fusion of autophagosomes with lysosomes. A short tyrosine-based sorting signal motif, YXXΦ (Y, tyrosine; X, any amino acid; Φ, a hydrophobic residue), present in ORF3a as the sequence YNSV (residues 160–163), plays a key role in binding ORF3a to VPS39 (Stephens et al., 2025). The point mutation Y160A, which disrupts this motif, abolishes co-immunoprecipitation with VPS39 and lifts the block on autophagosome-lysosome fusion (Zhang Y. et al., 2021).

In recent years, artificial intelligence methods capable of uncovering hidden patterns in large biomedical datasets have seen increasingly widespread use in pharmacology and related fields. Graph neural networks (GNNs) are regarded as a particularly promising direction, as they enable the integration of heterogeneous biological information and the prediction of novel interactions in complex networks that have not previously been reported in the literature. An early study that played a notable role in shaping this approach was conducted by M. Zitnik et al. (2018), which showed that graph convolutional neural networks can model drug–disease interactions and predict drug side effects.

This approach has since advanced rapidly: studies have integrated diverse data sources (external databases, abstracts and full texts of scientific publications, patents, electronic medical records, etc.), predicted protein-ligand and protein–protein interactions, and identified targets for drug repurposing using GNNs (Stokes et al., 2020; Gaudelet et al., 2021). In particular, the compound halicin was identified as a candidate with antibacterial activity against resistant strains; using a graph neural network, this molecule was shown to have bactericidal effects against *Mycobacterium tuberculosis*, carbapenem-resistant Enterobacteriaceae, as well as multidrug-resistant strains of *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Clostridioides difficile* (Stokes et al., 2020).

Methods for reconstructing and analyzing gene and associative networks are increasingly used to identify pharmacological targets at the human genome scale (Ali, Alrashid, 2025). Against this backdrop, cognitive systems and knowledge-engineering methods that automate the extraction of facts from the literature and specialized databases – and construct biomedical knowledge graphs – are being actively developed. In such graphs, nodes represent genes, proteins, metabolites, diseases, drugs, and other biomedical entities, while edges represent their interactions (regulatory relationships, protein–protein interactions, disease associations, etc.). Notable resources implementing this approach include STRING (Nicholson, Greene, 2020; Szklarczyk et al., 2023), QIAGEN Ingenuity Pathway Analysis (Krämer et al., 2014), GeneGo/MetaCore (Clarivate), and others.

We previously developed the cognitive platform ANDSystem, designed for the reconstruction of associative gene networks. It brings together two strands: 1) automatic knowledge extraction from scientific publications and biological databases using semantic-linguistic templates and rules (Ivanisenko V.A. et al., 2015, 2019), and 2) integration of statistical and machine-learning methods, including graph neural networks, to predict and add new protein–protein interactions to the network (Ivanisenko N.V. et al., 2024).

The ANDSystem knowledge base (KB) contains information on more than 100 million interactions among various types of molecular genetic entities (genes, RNAs, proteins, metabolites, drugs), as well as cellular- and organism-level entities such as cells, biological processes, diseases, and phenotypic traits. Interactions are classified into 49 types, including regulatory relationships (regulation of expression, activity, stability, transport, etc.), physical interactions (protein–protein, protein–ligand), chemical interactions (catalytic reactions, post-translational modifications, etc.), and associative links (gene–disease, gene–phenotype, biological process–disease, etc.). Of particular note are "marker" relationships, which indicate that a gene, biological process, or phenotypic trait serves as an indicator of an associated disease or phenotype. In addition, the KB includes "risk factor" interactions, in which a gene, process, disease, phenotypic trait, or other entity is considered a risk factor for the associated disease (Ivanisenko V.A. et al., 2019).

A distinctive feature of ANDSystem is its web-based module ANDDigest, designed for searching and analyzing PubMed publications using ontological dictionaries (Ivanisenko T.V. et al., 2020, 2022). The module supports complex queries that simultaneously take into account multiple types of entities from the ANDSystem dictionaries, as well as user-specified refining keywords. Search results are presented in graphical form with in-text annotation of the detected entities, options for sorting and filtering (by date, source citation counts, and other parameters), visualization of the year-by-year dynamics of mentions of the annotated entities, and links to external databases.

ANDSystem has been used to address a wide range of tasks based on the reconstruction and analysis of gene networks: reconstruction of the hepatitis C virus interactome (Saik et al., 2016); prioritization of genes associated with susceptibility to tuberculosis (Bragina et al., 2016); systems studies of preeclampsia (Glotov et al., 2015); analysis of the comorbidity of asthma and tuberculosis (Bragina et al., 2014); investigation of endothelial apoptosis in lymphedema (Saik et al., 2019); analysis of gene expression and the proteomic profile of clinical *Helicobacter pylori* strains associated with early stages of gastric cancer (Momynaliev et al., 2010); proteome stability in the Mars-500 project (Larina et al., 2015); interpretation of metabolomic data in studies of postoperative delirium (Ivanisenko V.A. et al., 2024); and the melanoma response to THz radiation (Butikova et al., 2025). Applying ANDSystem to the analysis of plasma metabolomic data from patients with COVID-19 made it possible to reconstruct gene

networks describing the molecular genetic pathways through which SARS-CoV-2 proteins influence metabolic disturbances during infection (Ivanisenko V.A. et al., 2022). It was shown that nonstructural coronavirus proteins play a particularly important role in such networks.

In the present study, we used graph neural networks with an attention mechanism (Veličković et al., 2017) to predict new ligands of the ORF3a protein among metabolites and drugs represented in the ANDSystem knowledge base. Using a model we trained on the ANDSystem knowledge graph, five small molecules of endogenous and exogenous origin were predicted to potentially interact with ORF3a:

1. N-acetyl-D-glucosamine – a monomer of the natural polysaccharide chitin. According to molecular modeling data, it can form stable complexes with four SARS-CoV-2 proteins: the Spike protein (PDB ID: 6M0J), the nucleocapsid phosphoprotein N (PDB ID: 6WKP), the S protein (PDB ID: 6X79), and the 3CLpro protease (PDB ID: 7JVZ), and may potentially elicit an immune response against the virus (Baysal et al., 2021; Tekin, 2023).
2. 4-(benzoylamino)benzoic acid – an amide derivative of benzoic acid. This compound exhibits antiviral activity against and Rift Valley fever virus (Islam et al., 2018).
3. Austocystin D – a polyketide metabolite of fungi of the genus *Aspergillus* with cytotoxic and antineoplastic activity (Marks et al., 2011).
4. Bictegravir – a small-molecule integrase inhibitor used to treat HIV infection (Sax et al., 2023). Studies have shown its high binding affinity to the Spike protein (Ahsan, Sajib, 2021; Sun et al., 2021) and to the main protease of SARS-CoV-2 (Mpro, PDB ID: 6LU7) (Oner et al., 2023).
5. L-threonine – an essential amino acid involved in protein synthesis, glycosylation, and regulation of the immune response. Evidence indicates that L-threonine levels change in various viral infections, including COVID-19, reflecting metabolic reprogramming in response to infection (Barberis et al., 2020). Several studies have shown that amino acid profiles, including threonine, can serve as biomarkers of COVID-19 severity and are involved in regulating inflammatory responses and mucosal barrier functions (Páez-Franco et al., 2021).

Molecular docking and binding free energy calculations indicated that bictegravir and 4-(benzoylamino)benzoic acid are the most promising candidates for experimental validation. For bictegravir, binding energies of −7.37 kcal/mol (AutoDock Vina) and −14.71 ± 3.12 kcal/mol (MM/GBSA) were obtained, indicating higher affinity compared with 4-(benzoylamino) benzoic acid (−5.68 kcal/mol and −11.01 ± 3.58 kcal/mol, respectively). Bictegravir is therefore of particular interest as a candidate for drug repurposing studies.

Materials and methods

The ANDSystem cognitive system. ANDSystem is a cognitive platform for the automated extraction of facts and knowledge from scientific publication texts and factual databases, their integration into a unified ontological model (a knowledge graph), and the reconstruction of associative gene networks (Ivanisenko V.A. et al., 2015, 2019). In the knowledge graph, vertices correspond to molecular genetic entities (genes, RNA transcripts, proteins, metabolites, drugs) as well as cellular- and organism-level objects (cell types, biological processes, diseases, phenotypic traits). Edges represent relationships between entities, including regulatory relationships (effects on expression, activity, stability, transport, etc.), physical contacts (protein–protein, protein–ligand interactions), chemical relationships (catalytic reactions, post-translational modifications, etc.), and associative links (gene–disease, gene–phenotype, process–disease, etc.). In its current version, the ANDSystem knowledge graph contains more than 1.5 million nodes and over 100 million edges.

For recognition of biomedical entity names and extraction of context-dependent relationships, ANDSystem uses more than 20,000 semantic linguistic templates and rules; in addition, large language models are employed, which improves the recall and precision of automated analysis of textual sources. To predict new interactions – particularly protein–protein interactions – graph neural networks (GNNs) trained on the ANDSystem knowledge graph, which is built from the scientific literature and specialized databases, are used (Ivanisenko T.V. et al., 2024).

ANDSystem includes the ANDDigest module – a specialized web-based system for searching and analyzing PubMed publications grounded in the ANDSystem ontological model and using dictionaries covering 13 types of biomedical entities (Ivanisenko T.V. et al., 2020, 2022). The ANDDigest database contains indexed and annotated PubMed texts, as well as computed characteristics and statistical co-occurrence measures for biomedical entities, which are used in subsequent stages of analysis and knowledge extraction.

Obtaining vector representations of nodes in the ANDSystem knowledge graph. To compute vector representations of nodes in the ANDSystem knowledge graph, we used a graph neural network with an attention mechanism (GAT) based on TransformerConv (the PyTorch Geometric package, version 2.5.3) (Fey, Lenssen, 2019). The network architecture comprised four hidden layers with 256 neurons each. Every node in the ANDSystem knowledge graph was described by a 13-dimensional binary vector in which a value of "1" indicated the object's membership in one of the 13 dictionary types defined by the ANDSystem ontology. Each edge was encoded by a 50-dimensional vector: the first 49 components corresponded to different interaction types and took values of 0 or 1 depending on whether the given type of relationship was present between the node pair in the knowledge graph, and the last component contained a numerical estimate of their co-occurrence (the *p*-value). This measure reflects the statistical significance of the joint mention of the object pair in PubMed abstracts and was computed using the ANDDigest module. The final node vector representations produced by the neural network had a dimensionality of 256.

The attention mechanism in each hidden layer comprised four independent heads that computed the contribution of neighboring nodes, that is, nodes connected to the node under consideration by edges in the ANDSystem graph. In doing so, it took into account both the features of the neighboring nodes themselves and the features of the edges linking them (relationship types and the *p*-value). The loss function was the logistic loss (Mao et al., 2023) with a temperature parameter

$\tau = 0.2$. Parameters were optimized using AdamW (Zhou et al., 2024).

Given the large size of the ANDSystem knowledge graph, to speed up training, the model was not trained on the entire graph at once but on subgraphs automatically generated from it. For each target node, a subgraph was constructed that included the node itself and its neighbors within at most three hops. At each "neighborhood level" (i. e., at distances of 1, 2, or 3 hops), the number of neighboring nodes considered was limited: up to 15 at the first level, 10 at the second, and 5 at the third. These neighbors were selected at random.

The computations were performed on a workstation with six NVIDIA GeForce RTX 4090 GPUs (24 GB of memory each); all programs were written in Python version 3.12.11.

Fully connected neural network. To predict new interactions (edges) between proteins and metabolites in the ANDSystem knowledge graph, a fully connected neural network (multilayer perceptron) was used. The size of the input layer matched the dimensionality of the vector representation of a pair of nodes (512). The model architecture included three consecutive hidden layers with 512, 256, and 128 neurons. Each hidden layer used the Rectified Linear Unit activation function (ReLU) (Glorot et al., 2011):

$$f(x) = \max(0, x).$$

The output layer contained a single neuron, the value of which reflected the probability of an edge existing between two nodes. For each protein–metabolite node pair, the neural network returned a value from 0 to 1, interpreted as the probability of an interaction between that pair. A standard threshold of 0.5 was used for classification: values above this threshold were interpreted as the presence of an interaction, and values below, as its absence (Harris, 2021).

From the ANDSystem knowledge graph, 250,000 object pairs were randomly selected, each consisting of one entity of type "protein" and the other of type "metabolite"; these pairs were treated as positive examples. As negative examples, an equal number of protein-metabolite pairs were randomly assembled from the set of all proteins and metabolites under the condition that the corresponding edge was absent from the original knowledge graph.

For each pair (u, v), we constructed a composite feature vector of length 512 (with node embedding dimensionality $d = 256$), comprising four blocks: 1) vector representation of the protein e_u; 2) vector representation of the metabolite e_v; 3) element-wise absolute difference $|e_u - e_v|$; 4) element-wise product (Hadamard product) $e_u \times e_v$.

The resulting array of vectors was split in an 80, 10, 10 % ratio into training, validation, and test subsets, respectively. The training subset was used to fit the model parameters during training; the test subset served for interim performance assessment and selection of the model's optimal hyperparameters; and the validation subset was used only to evaluate the accuracy of the final model after training. In each subset, the ratio of positive to negative examples was 1:1.

The model's performance after each training epoch (i. e., after the model had processed the entire training set) was evaluated on the test dataset using the Matthews correlation coefficient (MCC) (Chicco, Jurman, 2020), given by the formula:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}},$$

where *TP* (true positives) – the number of object pairs correctly classified by the model as interacting; *TN* (true negatives) – the number of object pairs correctly classified by the model as non-interacting; *FP* (false positive) – the number of object pairs incorrectly classified by the model as interacting; *FN* (false negative) – the number of object pairs incorrectly classified by the model as non-interacting.

Training was conducted over 83 epochs; the achieved MCC was 0.9542, indicating high model accuracy. The neural network was implemented using PyTorch version 2.4.1.

Molecular docking was used for an initial assessment of affinity via the docking score (Vina score) and for building protein–ligand complex models. The Vina score used at this stage is an empirical estimate of the binding energy (kcal/mol); more negative values correspond to higher predicted affinity. Calculations were performed with AutoDock Vina 1.2.0 (Python API) (Trott, Olson, 2010; Eberhardt et al., 2021). Docking was carried out in a blind-docking mode, defining a search region that encompassed the entire surface of the ORF3a protein.

The most energetically favorable protein–ligand conformations (minimum Vina scores) were used as the starting structures for estimating the binding free energy (ΔG) by the MM/GBSA method.

MM/GBSA evaluation. ΔG was calculated using the AmberTools package (Case et al., 2023). The method accounts for molecular mechanics energies and solvation contributions (the generalized Born model) with a nonpolar component proportional to the solvent-accessible surface area, and provides an approximate thermodynamic descriptor of complex stability. The three-dimensional structure of the SARS-CoV-2 ORF3a protein was obtained from the Protein Data Bank (PDB ID: 6XDC).

Results

Prediction of new protein–ligand interactions using graph neural networks

The analysis workflow employed in ANDSystem to predict new interactions with graph neural networks is shown in Figure 1.

An associative human gene network at the whole-genome scale was exported from the ANDSystem knowledge base. The network included all 13 object types (including genes, proteins, metabolites, diseases, and others) and 49 interaction types (regulatory relationships: regulation of expression, activity, stability, transport, etc.; physical interactions: protein–protein, protein–ligand, etc.). In total, the graph contained about 310,000 nodes connected by 48 million edges. To obtain vector representations of nodes in the knowledge graph, a graph neural network with an attention mechanism was trained; an F1 score of 0.8003 was reached by epoch 230.

Based on the obtained vector representations of proteins and metabolites in the ANDSystem knowledge graph, a multilayer perceptron was trained as a binary classifier to predict

edges missing from the graph. Training lasted 83 epochs; the achieved MCC was 0.9542. The trained model was then used to predict protein–metabolite edges for the ORF3a protein. In total, 38,172 potential links of this protein with small molecules of endogenous and exogenous origin were analyzed – including human metabolites and those of other organisms, as well as drugs, inorganic molecules, and ions – and five novel interactions not present in the ANDSystem knowledge base were identified.

In Figure 2, the ORF3a interaction network is shown: edges initially present in the ANDSystem knowledge base are depicted in black, and new links predicted by the graph neural network and the binary classification model are shown in red. The knowledge base contained 19 interactions extracted from scientific publications, including both direct physical contacts and associative links between ORF3a and small molecules. For example, physical interactions experimentally confirmed by fluorescence and UV-visible spectroscopy were reported for chlorin and cationic porphyrins; in the same study, molecular docking indicated complex formation for related porphyrins (bacteriochlorin, tetraphenylporphyrin, TPP) (Lebedeva et al., 2021). As an example of an associative link, one can cite the ORF3a–bradykinin association discussed in the context of an intensified "bradykinin storm" via ORF3a/NS7b interaction in COVID-19 (Messina et al., 2021).

The group of predicted interactions comprised five candidates: N-acetyl-D-glucosamine (a chitin monomer and a precursor for glycosylation); 4-(benzoylamino)benzoic acid (a derivative of benzoic acid); austocystin D (a polyketide metabolite of Aspergillus fungi); bictegravir (an HIV integrase inhibitor; a medicinal drug); and L-threonine (an essential amino acid).

Molecular docking and binding energy evaluation

To assess the ability of the five predicted small molecules to physically interact with ORF3a, we performed molecular docking using AutoDock Vina and, for the resulting 3D complex models, recalculated the binding free energy (ΔG) by the MM/GBSA method (Table 1). The docking score (Vina score), which provides an empirical estimate of affinity, was used for the relative ranking of ligands, whereas the MM/GBSA ΔG values were considered an approximate thermodynamic descriptor of complex stability.

According to AutoDock Vina, the highest predicted affinity was shown by austocystin D (−8.296 kcal/mol) and bictegravir (−7.368 kcal/mol); intermediate affinities, by N-acetyl-D-glucosamine (−6.242 kcal/mol) and 4-(benzoylamino)benzoic acid (−5.682 kcal/mol); and the lowest affinity, by L-threonine (−4.89 kcal/mol).

According to MM/GBSA, the most negative (i. e., lowest) ΔG was obtained for austocystin D (−21.67 ± 2.30 kcal/mol), followed by L-threonine (−19.04 ± 2.15) and N-acetyl-D-glucosamine (−16.76 ± 2.58), whereas bictegravir (−14.71 ± 3.12) and 4-(benzoylamino)benzoic acid (−11.01 ± 3.58) had ΔG values of smaller magnitude.

Taken together, the docking scores (Vina score) and the ΔG estimates from the MM/GBSA method indicate the potential formation of ORF3a complexes with the analyzed small molecules, serving as complementary criteria for the computational assessment of affinity.

The 3D models of ORF3a complexes with the ligands under study, constructed based on the results of molecular docking, are shown in Figure 3. According to cryo-EM data, ORF3a forms a dimer; each subunit contains three transmembrane helices and a large cytosolic C-terminal domain (Kern et al., 2021). ORF3a is predominantly localized to the membranes of the Golgi apparatus, endosomes, and lysosomes, participating in the regulation of vesicular transport and lysosomal exocytosis; it is also detected at the plasma membrane (Hinkle et al., 2025).

It is known that ORF3a interacts with VPS39 (the HOPS complex) and blocks the fusion of autophagosomes with

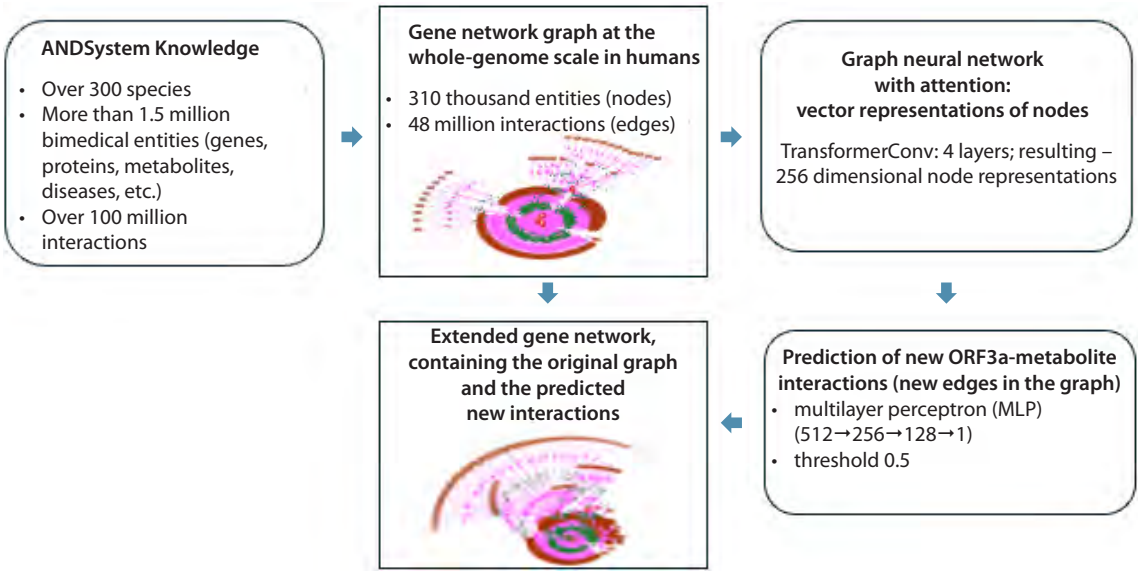


Fig. 1. Schematic representation of the computational pipeline for predicting new interactions between human proteins and metabolites based on analysis of the ANDSystem knowledge graph.

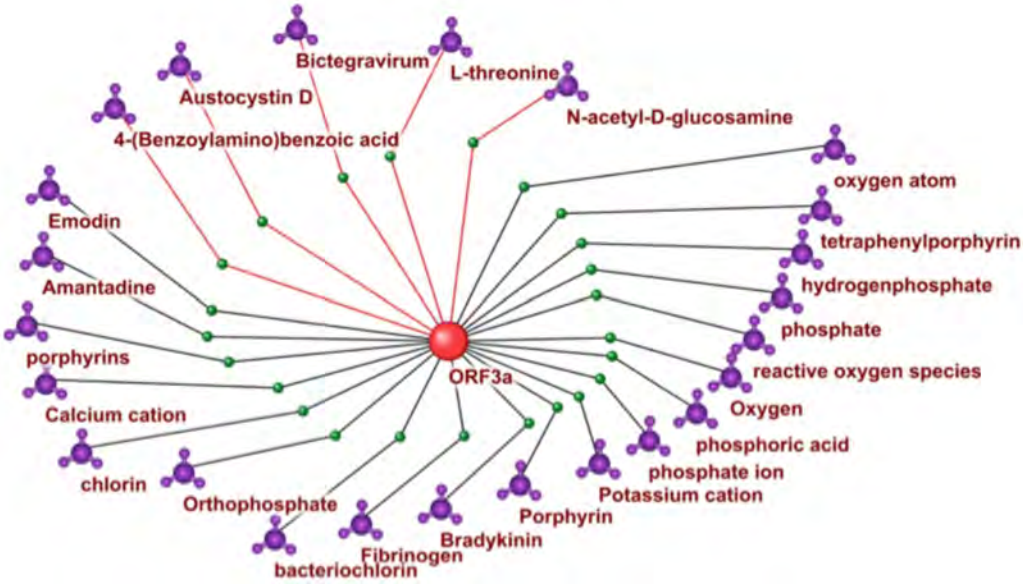


Fig. 2. Interaction network of ORF3a with small molecules reconstructed using ANDSystem.

Dark lines indicate interactions supported by scientific publications; red lines indicate interactions predicted by the graph neural network: N-acetyl-D-glucosamine, 4-(benzoylamino)benzoic acid, austocystin D, bictegravir, and L-threonine.

Table 1. Calculated ORF3a–ligand binding metrics from AutoDock Vina and MM/GBSA

No.	Ligand	AutoDock Vina (kcal/mol)*	MM/GBSA (kcal/mol)**
1	Austocystin D	–8.296	–21.67 ± 2.3
2	Bictegravirum	–7.368	–14.71 ± 3.12
3	N-acetyl-D-glucosamine	–6.242	–16.76 ± 2.58
4	4-(Benzoylamino)benzoic acid	–5.682	–11.01 ± 3.58
5	L-threonin	–4.89	–19.04 ± 2.15

* AutoDock Vina docking score (kcal/mol); ** binding free energy ΔG (kcal/mol) estimated by the MM/GBSA method.

lysosomes, leading to the accumulation of unfused autophagosomes and facilitating viral evasion of degradation (Zhang J. et al., 2021; Miller et al., 2023). For clarity, the corresponding region of the protein involved in the interaction with VPS39 is highlighted with a box in the Figure 3.

According to the docking results, the binding sites of L-threonine, bictegravir, and 4-(benzoylamino)benzoic acid are located on the cytosolic surface of the dimer and partially overlap with the ORF3a–VPS39 binding region (Fig. 3*a*). L-threonine binds at the intersubunit interface (inter-subunit cleft) of ORF3a, is deeply buried there, and is essentially solvent-inaccessible. Bictegravir and 4-(benzoylamino) benzoic acid occupy solvent-exposed surface regions of the protein (Fig. 4). Austocystin D and N-acetyl-D-glucosamine bind at the boundary between the cytosolic surface and the transmembrane domain (Fig. 3*b*).

Details of hydrogen (H-) and hydrophobic contacts between the ligands and ORF3a amino acid residues are given in Table 2 and illustrated in Figure 5. N-acetyl-D-glucosamine forms multiple H-bonds with residues Lys61, Ile63, Thr64, Arg126, and others. 4-(Benzoylamino)benzoic acid forms H-bonds with Ser165 and Asp226, as well as hydrophobic contacts with Val225 and Val228. Austocystin D forms H-bonds with Ser165, Glu226, His227, and Asn234 and hydrophobic contacts with His227. Bictegravir forms three H-bonds (Ser165, Glu226, Asn234). L-threonine, located deep in the intersubunit cleft at the dimer interface, forms multiple H-bonds (with six residues) and hydrophobic contacts with Ile186.

Discussion

Building on our previous work with GraphSAGE for predicting protein–protein interactions (Ivanisenko T.V. et al., 2024), in this study, we applied a graph neural network with an attention mechanism to predict interactions of the SARS-CoV-2 ORF3a protein with small molecules on the ANDSystem knowledge graph and identified five candidate ligands: N-acetyl-D-glucosamine, 4-(benzoylamino)benzoic acid, austocystin D, bictegravir, and L-threonine.

Unlike the GraphSAGE architecture, attention-based models update node representations by explicitly weighting the

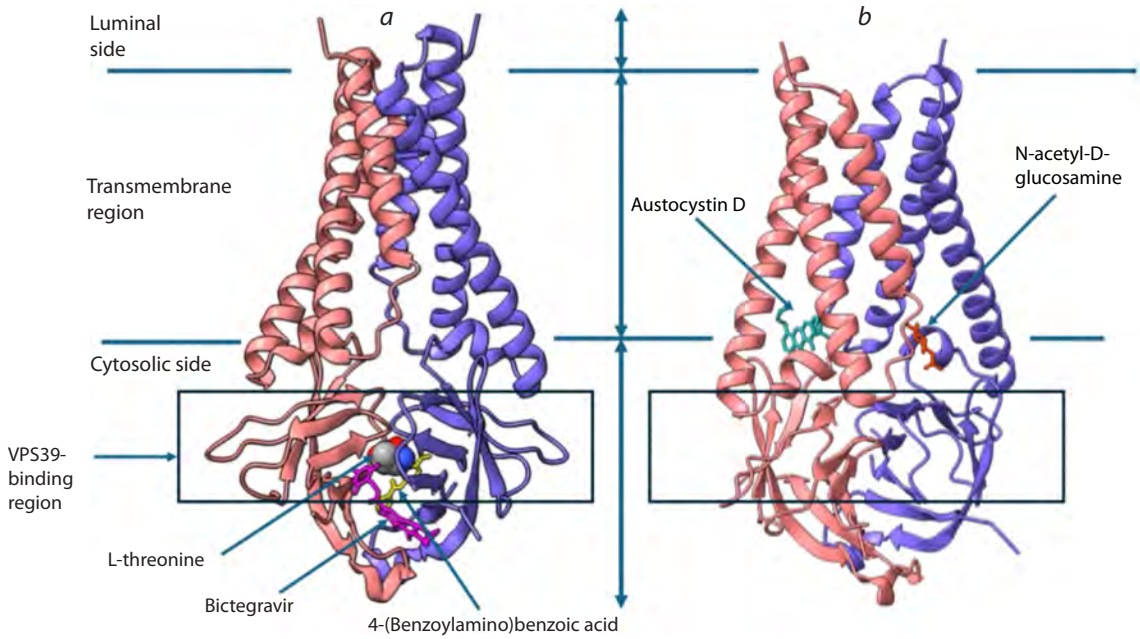


Fig. 3. Spatial structures of ORF3a complexes with the analyzed ligands.

a – ORF3a complex with L-threonine, bictegravir, and 4-(benzoylamino)benzoic acid; *b* – ORF3a complex with austocystin D and N-acetyl-D-glucosamine. The protein is shown in a ribbon representation; the two subunits of the dimer are colored differently. In panel (*b*), the protein structure is rotated to better display the ligands. Ligands are shown in a stick representation; their positions are indicated by arrows. L-threonine is shown in a space-filling (spheres) representation for clarity. Lines mark the regions of the protein corresponding to its position within the membrane (Kern et al., 2021): cytosolic side, transmembrane region, and luminal side (the lumen of the Golgi apparatus and endo-/lysosomes). The boxed area denotes the region involved in interaction with the VPS39 protein. Images were generated in ChimeraX.

contributions of their neighbors: more informative relations receive higher weights, and less informative ones, lower weights. Multiple attention heads operate in parallel, and their outputs are then aggregated into the final node vector, enabling a more precise accounting of the local graph context (Wu et al., 2021).

To validate these predictions, we performed molecular docking and estimated the binding free energy (ΔG) of the protein–ligand complexes using the MM/GBSA method. The calculations showed that the predicted binding sites of austocystin D and N-acetyl-D-glucosamine are located at the boundary between the cytosolic surface and the transmembrane domain of ORF3a, whereas L-threonine, bictegravir, and 4-(benzoylamino)benzoic acid bind on the cytosolic side of the dimer; moreover, the binding regions of bictegravir and 4-(benzoylamino)benzoic acid partially overlap with the ORF3a–VPS39 interaction region.

The interaction of ORF3a with the host protein VPS39, a subunit of the homotypic fusion and protein sorting (HOPS) complex that regulates the late stages of endosome–lysosome compartment fusion, is well characterized (Zhang J. et al., 2021; Miller et al., 2023). It hinders the fusion of autophagosomes and late endosomes with lysosomes, thereby suppressing autophagic flux – a key pathway for the degradation of viral components.

The functional significance of the interaction interface between ORF3a and VPS39 is supported by the presence of an YXXΦ motif in the cytosolic domain of ORF3a (Y, tyrosine; X, any amino acid; Φ, a hydrophobic residue).

In ORF3a, this motif is present as the sequence YNSV (residues 160–163). Studies (Zhang J. et al., 2021; Miller et al., 2023) have shown that the point mutation Y160A disrupts co-immunoprecipitation of ORF3a with VPS39 and lifts the blockade of HOPS-dependent fusion, partially restoring autophagic flux.

It can be hypothesized that the predicted locations of the binding sites for bictegravir and 4-(benzoylamino)benzoic acid could influence the formation and/or stability of the ORF3a–VPS39 complex, making them promising candidates for functional intervention at the HOPS-dependent stage of autophagosome–lysosome fusion.

Taken together across metrics (Vina score and MM/GBSA ΔG), bictegravir shows more negative values – indicating higher predicted affinity – than 4-(benzoylamino) benzoic acid (Vina score −7.37 kcal/mol and MM/GBSA ΔG −14.71 ± 3.12 kcal/mol vs. −5.68 kcal/mol and −11.01 ± 3.58 kcal/mol, respectively). In addition, bictegravir is a licensed HIV integrase inhibitor (the drug Biktarvy) (Gallant et al., 2017), making it a promising repurposing candidate. A potential mechanism of action for bictegravir as a therapeutic for COVID-19 could be inhibition of the ORF3a interaction with the host protein VPS39, which in turn would neutralize ORF3a's ability to block fusion of endosome–lysosome compartments and promote degradation of viral components in lysosomes. In turn, 4-(benzoylamino)benzoic acid may be of interest as an aromatic carboxamide fragment for targeting protein–protein interaction interfaces within the ORF3a structure (Marks et al., 2011).

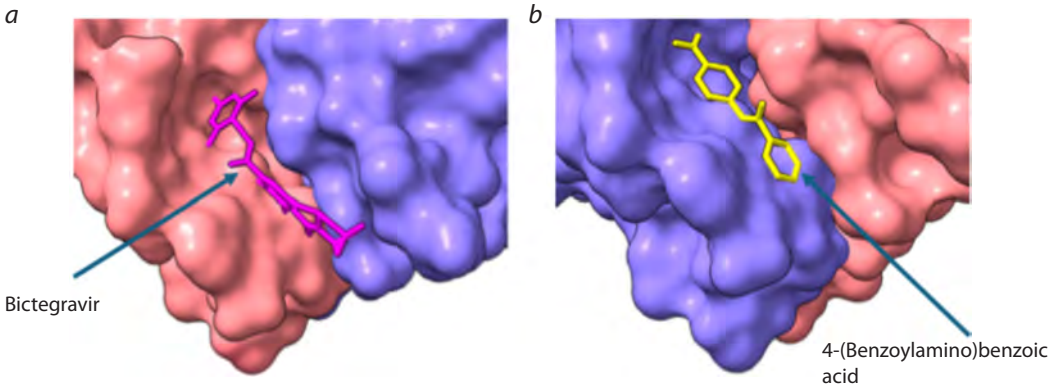


Fig. 4. Surface of ORF3a bound to bictegravir (*a*) and 4-(benzoylamino)benzoic acid (*b*).
Images were generated in ChimeraX.

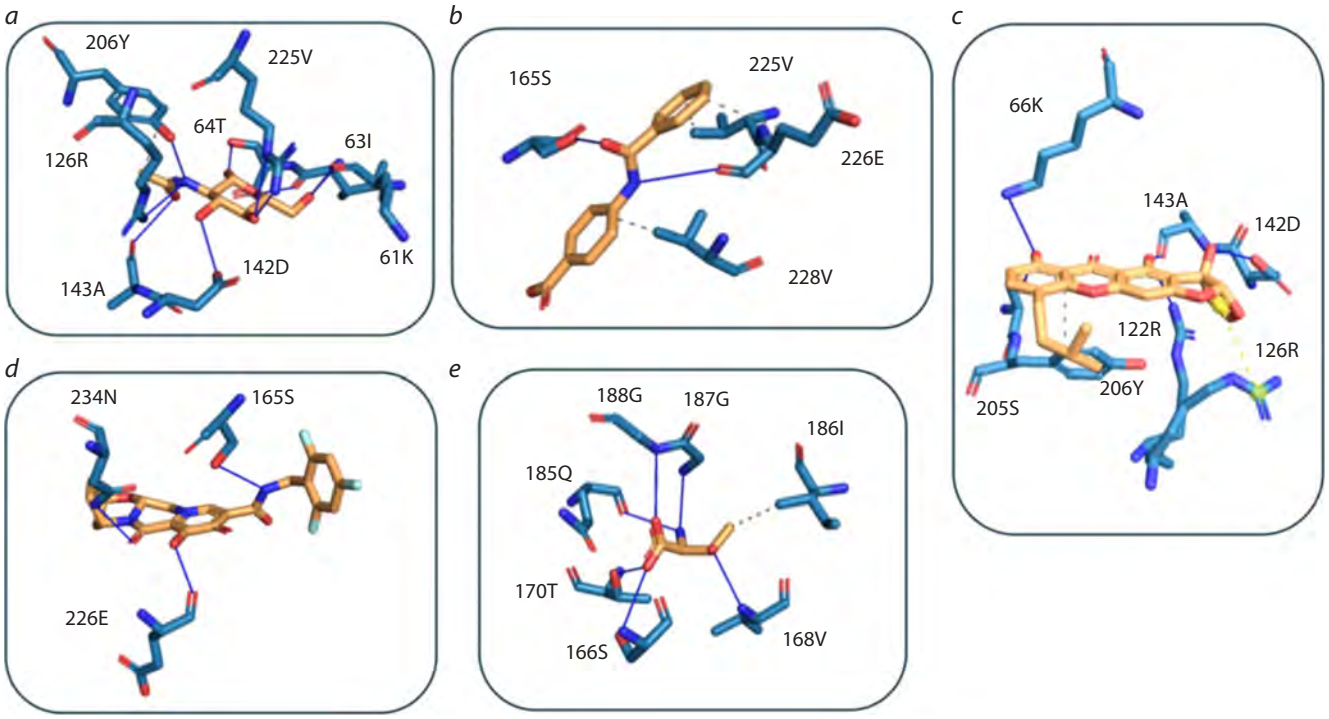


Fig.5. Detailed representation of the interactions of the analyzed ligands with ORF3a amino acid residues.
a – N-acetyl-D-glucosamine; *b* – 4-(benzoylamino)benzoic acid; *c* –austocystin D; *d* – ictegravir; L-threonine. The ligand is shown in yellow and amino acid residues in blue. Hydrogen bonds are shown as solid lines; hydrophobic interactions are shown as dashed lines. Images were generated in PyMOL.

Conclusion

Our approach – predicting new protein–ligand interactions on the ANDSystem knowledge graph followed by molecular docking and estimation of binding ΔG via the MM/GBSA method – enabled us to identify promising small-molecule ligand candidates for the SARS-CoV-2 ORF3a protein. Among the selected compounds, bictegravir and 4-(benzoylamino)benzoic acid are of greatest interest: their predicted sites lie on the cytosolic surface of ORF3a and partially overlap with the ORF3a–VPS39 interaction region. Based on energetic estimates, bictegravir shows more negative Vina score and ΔG values: AutoDock Vina, −7.37 kcal/mol; MM/GBSA, −14.71 ± 3.12 kcal/mol. For 4-(benzoylamino) benzoic acid, comparable but smaller-magnitude values were obtained: −5.68 kcal/mol and −11.01 ± 3.58 kcal/mol, respectively.

A limitation of this study is the lack of explicit consideration of the lipid bilayer: the calculations were performed without embedding the protein in a membrane, which may affect the conformation of ORF3a and the energetic contributions associated with ligand penetration into the hydrophobic environment. As a next step, molecular dynamics in a membrane model with recalculation of binding energies could be performed, followed by experimental validation of the results.

Table 2. Molecular interactions of the ORF3a protein with ligands, obtained from analysis of the reconstructed ORF3a–ligand complexes using the PLIP (Protein-Ligand Interaction Profiler) web server

Ligand	Amino acid residue numbe*	Amino acid residue**	Distance, Å	Interaction type
Austocystin D	165B	SER	2.45	H-bond
	226A	GLU	3.70	
	227A	HIS	2.83	
	234B	ASN	2.85	
	227A	HIS	3.59	Hydrophobic
Bictegravir	165A	SER	2.44	H-bond
	226B	GLU	2.20	
	234A	ASN	2.35	
N-acetyl-D-glucosamine	61B	LYS	2.07	H-bond
	63B	ILE	2.17	
	63B	ILE	2.69	
	64B	THR	3.15	
	122A	ARG	2.07	
	122A	ARG	2.52	
	126A	ARG	2.36	
	142A	ASP	2.45	
	143A	ALA	3.03	
	206A	TYR	2.87	
	206A	TYR	3.69	Hydrophobic
4-(Benzoylamino) benzoic acid	165B	SER	2.53	H-bond
	226A	GLU	3.36	
	225A	VAL	3.93	Hydrophobic
	226A	GLU	3.69	
	228A	VAL	3.55	
L-threonine	166B	SER	3.28	H-bond
	168B	VAL	2.25	
	170A	THR	2.08	
	185A	GLN	2.34	
	187A	GLY	2.36	
	188A	GLY	2.85	
	186B	ILE	3.69	Hydrophobic

* Amino acid residue numbering follows the ORF3a sequence; the chain identifier is given according to the PDB structure 6XDC.
** The amino acid involved in the interaction is indicated.

Taken together, the *in silico* results identify bictegravir as a priority candidate for experimental studies of its interaction with ORF3a – including within a drug-repurposing framework – and provide a foundation for further optimization of small molecules targeting this protein.

References

Ahsan T., Sajib A.A. Repurposing of approved drugs with potential to interact with SARS-CoV-2 receptor. *Biochem Biophys Rep.* 2021;26:100982. doi 10.1016/j.bbrep.2021.100982

Ali S.I.M., Alrashid S.Z. A review of methods for gene regulatory networks reconstruction and analysis. *Artif Intell Rev.* 2025;58:256. doi 10.1007/s10462-025-11257-z

Barberis E., Timo S., Amede E., Vanella V.V., Puricelli C., Cappellano G., Raineri D., … Rolla R., Chiocchetti A., Baldanzi G., Marengo E., Manfredi M. Large-scale plasma analysis revealed new mechanisms and molecules associated with the host response to SARS-CoV-2. *Int J Mol Sci.* 2020;21(22):8623. doi 10.3390/ijms21228623

Baysal Ö., Abdul Ghafoor N., Silme R.S., Ignatov A.N., Kniazeva V. Molecular dynamics analysis of N-acetyl-D-glucosamine against specific SARS-CoV-2's pathogenicity factors. *PLoS One.* 2021;16(5):e0252571. doi 10.1371/journal.pone.0252571

Boby M.L., Fearon D., Ferla M., Filep M., Koekemoer L., Robinson M.C., COVID Moonshot Consortium, … Zaidmann D., Zhang I., Zidane H., Zitzmann N., Zvornicanin S.N. Open science discovery of potent non-covalent SARS-CoV-2 main protease inhibitors. *Science.* 2023;380(6640):eabo7201. doi 10.1126/science.abo7201

Bragina E.Y., Tiys E.S., Freidin M.B., Koneva L.A., Demenkov P.S., Ivanisenko V.A., Kolchanov N.A., Puzyrev V.P. Insights into pathophysiology of dystropy through the analysis of gene networks: an example of bronchial asthma and tuberculosis. *Immunogenetics.* 2014;66(7-8):457-465. doi 10.1007/s00251-014-0786-1

Bragina E.Y., Tiys E.S., Rudko A.A., Ivanisenko V.A., Freidin M.B. Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks. *Infect Genet Evol.* 2016;46:118-123. doi 10.1016/j.meegid.2016.10.030

Butikova E.A., Basov N.V., Rogachev A.D., Gaisler E.V., Ivanisenko V.A., Demenkov P.S., Makarova A.-L.A., … Pokrovsky A.G., Vinokurov N.A., Kanygin V.V., Popik V.M., Shevchenko O.A. Metabolomic and gene networks approaches reveal the role of mitochondrial membrane proteins in response of human melanoma cells to THz radiation. *Biochim Biophys Acta Mol Cell Biol Lipids.* 2025;1870(2):159595. doi 10.1016/j.bbalip.2025.159595

Case D.A., Aktulga H.M., Belfon K., Cerutti D.S., Andrés Cisneros G., Cruzeiro V.W.D., Forouzesh N., … Roitberg A., Simmerling C.S., York D.M., Nagan M.C., Merz K.M. Jr. AmberTools. *J Chem Inf Model.* 2023;63(20):6183-6191. doi 10.1021/acs.jcim.3c01153

Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21:6. doi 10.1186/s12864-019-6413-7

Clarivate. MetaBase & MetaCore: Early Research Intelligence Solutions. Available at: http://clarivate.com/life-sciences-healthcare/research-development/discovery-development/early-research-intelligence-solutions/

Dong M., Galvan Achi J.M., Du R., Rong L., Cui Q. Development of SARS-CoV-2 entry antivirals. *Cell Insight.* 2024;3(1):100144. doi 10.1016/j.cellin.2023.100144

Eberhardt J., Santos-Martins D., Tillack A.F., Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J Chem Inf Model.* 2021;61(8):3891-3898. doi 10.1021/acs.jcim.1c00203

Fey M., Lenssen J.E. Fast graph representation learning with PyTorch Geometric. *arXiv.* 2019. doi 10.48550/arXiv.1903.02428

Gallant J.E., Thompson M., DeJesus E., Voskuhl G.W., Wei X., Zhang H., Martin H. Antiviral activity, safety, and pharmacokinetics of bictegravir as 10-day monotherapy in HIV-1-infected adults. *J Acquir Immune Defic Syndr.* 2017;75(1):61-66. doi 10.1097/QAI.0000000000001306

Gaudelet T., Day B., Jamasb A.R., Soman J., Regep C., Liu G., Hayter J.B.R., Vickers R., Roberts C., Tang J., Roblin D., Blundell T.L., Bronstein M.M., Taylor-King J.P. Utilizing graph machine learning within drug discovery and development. *Brief Bioinform.* 2021;22(6):bbab159. doi 10.1093/bib/bbab159

Glorot X., Bordes A., Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS). 2011;315-323. Available at: https://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf

Glotov A.S., Tiys E.S., Vashukova E.S., Pakin V.S., Demenkov P.S., Saik O.V., Ivanisenko T.V., Arzhanova O.N., Mozgovaya E.V., Zainulina M.S., Kolchanov N.A., Baranov V.S., Ivanisenko V.A. Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). *BMC Syst Biol.* 2015;9(Suppl. 2):S4. doi 10.1186/1752-0509-9-S2-S4

Gwon Y.-D., Strand M., Lindqvist R., Nilsson E., Saleeb M., Elofsson M., Överby A.K., Evander M. Antiviral activity of benzavir-2 against emerging flaviviruses. *Viruses.* 2020;12(3):351. doi 10.3390/v12030351

Harris J.K. Primer on binary logistic regression. *Fam Med Community Health.* 2021;9(Suppl. 1):e001290. doi 10.1136/fmch-2021-001290

Hinkle J.J., Trychta K.A., Wires E.S., Osborn R.M., Leach J.R., Faraz Z.F., Svarcbahs R., Richie C.T., Dewhurst S., Harvey B.K. Subcellular localization of SARS-CoV-2 E and 3a proteins along the secretory pathway. *J Mol Histol.* 2025;56(2):98. doi 10.1007/s10735-025-10375-w

Islam M., Strand M., Saleeb M., Svensson R., Baranczewski P., Artursson P., Wadell G., Ahlm C., Elofsson M., Evander M. Anti-Rift Valley fever virus activity *in vitro*, pre-clinical pharmacokinetics and oral bioavailability of benzavir-2, a broad-acting antiviral compound. *Sci Rep.* 2018;8:1925. doi 10.1038/s41598-018-20362-9

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics.* 2020;21(Suppl 11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int J Mol Sci.* 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci.* 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an associative network discovery system for automated literature mining in the field of biology. *BMC Syst Biol.* 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics.* 2019;20(1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., … Karpenko T.N., Ve-

lichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci Rep.* 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci.* 2024;25(21):11847. doi 10.3390/ijms252111847

Kern D.M., Sorum B., Mali S.S., Hoel C.M., Sridharan S., Remis J.P., Toso D.B., Kotecha A., Bautista D.M., Brohawn S.G. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat Struct Mol Biol.* 2021;28(7):573-582. doi 10.1038/s41594-021-00619-0

Krämer A., Green J., Pollard J. Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics.* 2014;30(4):523-530. doi 10.1093/bioinformatics/btt703

Larina I.M., Pastushkova L.Kh., Tiys E.S., Kireev K.S., Kononikhin A.S., Starodubtseva N.L., Popov I.A., Custaud M.-A., Dobrokhotov I.V., Nikolaev E.N., Kolchanov N.A., Ivanisenko V.A. Permanent proteins in the urine of healthy humans during the Mars-500 experiment. *J Bioinform Comput Biol.* 2015;13(1):1540001. doi 10.1142/S0219720015400016

Lebedeva N.S., Gubarev Y.A., Mamardashvili G.M., Zaitceva S.V., Zdanovich S.A., Malyasova A.S., Romanenko J.V., Koifman M.O., Koifman O.I. Theoretical and experimental study of interaction of macroheterocyclic compounds with ORF3a of SARS-CoV-2. *Sci Rep.* 2021;11:19481. doi 10.1038/s41598-021-99072-8

Mao A., Mohri M., Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. In: International Conference on Machine Learning (ICML). 2023;23803-23828. Available at: https://proceedings.mlr.press/v202/mao23b/mao23b.pdf

Marks K.M., Park E.S., Arefolov A., Russo K., Ishihara K., Ring J.E., Clardy J., Clarke A.S., Pelish E.P. The selectivity of austocystin D arises from cell-line-specific drug activation by cytochrome P450 enzymes. *J Nat Prod.* 2011;74(4):567-573. doi 10.1021/np100429s

Messina F., Giombini E., Montaldo C., Sharma A.A., Zoccoli A., Sekaly R.P., Locatelli F., Zumla A., Maeurer M., Capobianchi M.R., Lauria F.N., Ippolito G. Looking for pathways related to COVID-19: confirmation of pathogenic mechanisms by SARS-CoV-2-host interactome. *Cell Death Dis.* 2021;12(8):788. doi 10.1038/s41419-021-03881-8

Miller A.N., Houlihan P.R., Matamala E., Cabezas-Bratesco D., Lee G.Y., Cristofori-Armstrong B., Dilan T.L., Sanchez-Martinez S., Matthies D., Yan R., Yu Z., Ren D., Brauchi S.E., Clapham D.E. The SARS-CoV-2 accessory protein Orf3a is not an ion channel. *eLife.* 2023;12:e84477. doi 10.7554/eLife.84477

Momynaliev K.T., Kashin S.V., Chelysheva V.V., Selezneva O.V., Demina I.A., Serebryakova M.V., Alexeev D., Ivanisenko V.A., Aman E., Govorun V.M. Functional divergence of *Helicobacter pylori* related to early gastric cancer. *J Proteome Res.* 2010;9(1):254-267. doi 10.1021/pr900586w

Naqvi A.A.T., Fatima K., Muhammad T., Fatima U., Singh I.K., Singh A., Atif S.M., Hariprasad G., Hasan G.M., Hassan M.I. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies. *Int J Biol Sci.* 2020;16(10):1708-1724. doi 10.7150/ijbs.45127

Ng T.I., Correia I., Seagal J., DeGoey D.A., Schrimpf M.R., Hardee D.J., Noey E.L., Kati W.M. Antiviral drug discovery for the treatment of COVID-19 infections. *Viruses.* 2022;14(5):961. doi 10.3390/v14050961

Nicholson D.N., Greene C.S. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J.* 2020;18:1414-1421. doi 10.1016/j.csbj.2020.05.017

Oner E., Demirhan I., Miraloglu M., Yalin S., Kurutas E.B. Investigation of antiviral substances in COVID-19 by molecular docking: in silico study. *Afr Health Sci.* 2023;23(1):23-36. doi 10.4314/ahs.v23i1.4

Páez-Franco J.C., Torres-Ruiz J., Sosa-Hernández V.A., Cervantes-Díaz R., Romero-Ramírez S., Pérez-Fragoso A., Meza-Sánchez D.E., Germán-Acacio J.M., Maravillas-Montero J.L., Mejía-Domínguez N.R., Ponce-de-León A., Ulloa-Aguirre A., Gómez-Martín D., Llorente L. Metabolomics analysis reveals a modified amino acid metabolism that correlates with altered oxygen homeostasis in COVID-19 patients. *Sci Rep.* 2021;11(1):6350. doi 10.1038/s41598-021-85788-0

Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.* 2016;218:40-48. doi 10.1016/j.virusres.2015.12.003

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics.* 2019;12(Suppl. 2):117. doi 10.1186/s12920-019-0492-9

Sax P.E., Arribas J.R., Orkin C., Lazzarin A., Pozniak A., DeJesus E., Maggiolo F., … Hindman J.T., Martin H., Baeten J.M., Wohl D.; GS-US-380-1489 and GS-US-380-1490 study investigators. bictegravir/emtricitabine/tenofovir alafenamide as initial treatment for HIV-1: five-year follow-up from two randomized trials. *EClinicalMedicine.* 2023;59:101991. doi 10.1016/j.eclinm.2023.101991

Stephens E.B., Kunec D., Henke W., Vidal R.M., Greishaber B., Saud R., Kalamvoki M., Singh G., Kafle S., Trujillo J.D., Ferreyra F.M., Morozov I., Richt J.A. The role of the tyrosine-based sorting signals of the ORF3a protein of SARS-CoV-2 in intracellular trafficking and pathogenesis. *Viruses*. 2025;17(4):522. doi 10.3390/v17040522

Stokes J.M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., Donghia N.M., MacNair C.R., … Church G.M., Brown E.D., Jaakkola T.S., Barzilay R., Collins J.J. A deep learning approach to antibiotic discovery. *Cell.* 2020;180(4):688-702. doi 10.1016/j.cell.2020.01.021

Sun C., Zhang J., Wei J., Zheng X., Zhao X., Fang Z., Xu D., Yuan H., Liu Y. Screening, simulation, and optimization design of small-molecule inhibitors of the SARS-CoV-2 spike glycoprotein. *PLoS One.* 2021;16(1):e0245975. doi 10.1371/journal.pone.0245975

Szklarczyk D., Kirsch R., Koutrouli M., Nastou K., Mehryary F., Hachilif R., Gable A.L., Fang T., Doncheva N.T., Pyysalo S., Bork P., Jensen L.J., von Mering C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023; 51(D1):D638-D646. doi 10.1093/nar/gkac1000

Tekin E.D. Investigation of the effects of N-acetylglucosamine on the stability of the spike protein in SARS-CoV-2 by molecular dynamics simulations. *Comput Theor Chem.* 2023;1222:114049. doi 10.1016/j.comptc.2023.114049

Trott O., Olson A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010;31(2):455-461. doi 10.1002/jcc.21334

U.S. Food and Drug Administration (FDA). FDA approves first treatment for COVID-19 (Veklury/remdesivir): press release. 22 Oct 2020. Available at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-covid-19

U.S. Food and Drug Administration. FDA approves first oral antiviral for treatment of COVID-19 in adults: press announcement. 2023-05-25. Available at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-oral-antiviral-treatment-covid-19-adults

Veličković P., Cucurull G., Casanova A., Romero A., Liò P., Bengio Y. Graph attention networks. *arXiv.* 2017. doi 10.48550/arXiv.1710.10903

von Delft A., Hall M.D., Kwong A.D., Purcell L.A., Saikatendu K.S., Schmitz U., Tallarico J.A., Lee A.A. Accelerating antiviral drug dis-

covery: lessons from COVID-19. *Nat Rev Drug Discov.* 2023;22(7): 585-603. doi 10.1038/s41573-023-00692-8

Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P.S. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst.* 2021;32(1):4-24. doi 10.1109/TNNLS.2020.2978386

Zhang J., Cruz-Cosme R., Zhuang M.W., Liu D., Liu Y., Teng S., Wang P.-H., Tang Q. A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. *Signal Transduct Target Ther.* 2021;6(1):192. doi 10.1038/s41392-021-00564-w

Zhang J., Ejikemeuwa A., Gerzanich V., Nasr M., Tang Q., Simard J.M., Zhao R.Y. Understanding the role of SARS-CoV-2 ORF3a in viral pathogenesis and COVID-19. *Front Microbiol.* 2022;13:854567. doi 10.3389/fmicb.2022.854567

Zhang Y., Sun H., Pei R., Mao B., Zhao Z., Li H., Lin Y., Lu K. The SARS-CoV-2 protein ORF3a inhibits fusion of autophagosomes with lysosomes. *Cell Discov.* 2021;7:31. doi 10.1038/s41421-021-00268-z

Zhou P., Xie X., Lin Z., Yan S. Towards understanding convergence and generalization of AdamW. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(9):6486-6493. doi 10.1109/TPAMI.2024.3382294

Zitnik M., Agrawal M., Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics.* 2018; 34(13):i457-i466. doi 10.1093/bioinformatics/bty294

Conflict of interest. The authors declare no conflict of interest.

Received September 25, 2025. Revised October 24, 2025. Accepted October 27, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1097-1108

doi 10.18699/vjgb-25-114

The effect of dimeric bisbenzimidazoles on the activity of DNA repair enzymes TDP1, TDP2, PARP1 and PARP2

N.S. Dyrkheeva [1], I.A. Chernyshova [1], A.F. Arutyunyan [2], A.L. Zakharenko [1], M.M. Kutuzov [1], K.N. Naumenko [1], A.S. Venzel [3], V.A. Ivanisenko [3], S.M. Deyev [4], A.L. Zhuze [2], O.I. Lavrik [1]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] The Engelhardt Institute of Molecular Biology of the Russian Academy of Sciences, Moscow, Russia
[3] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, Russia

lavrik@1bio.ru

Abstract. Oncological diseases remain a leading cause of pathological mortality worldwide, making the development of anticancer drugs a critical focus in medicinal chemistry. A promising strategy to enhance therapeutic efficacy and reduce chemotherapy-induced toxicity involves the combined inhibition of DNA repair enzymes and topoisomerases. Of particular interest are minor-groove DNA ligands, which exhibit potent inhibition of DNA-dependent enzymes while having low toxicity and mutagenicity. A number of research groups, including ours, are developing inhibitors of DNA repair enzymes that act simultaneously on several targets: tyrosyl-DNA phosphodiesterase 1/2 (TDP1/TDP2), poly(ADP-ribose) polymerase 1 (PARP1)/TDP1, topoisomerase 1 (TOP1)/TDP1. Such bifunctional inhibitors are designed to resolve the problem of tumor cell resistance to known chemotherapy drugs and increase the effectiveness of the latter. In this study, we evaluated the inhibitory activity of 22 minor-groove DNA ligands – bis- and trisbenzimidazoles against four key repair enzymes: TDP1, TDP2, PARP1, and PARP2. Four series of dimeric compounds and their monomeric units were studied. The difference in inhibitory activity of dimeric bisbenzimidazoles depending on the structure of the compound and the enzyme is shown. Our findings reveal distinct structure-activity relationships, with monomeric and dimeric ligands exhibiting potent TDP1 inhibition at micromolar to submicromolar IC_{50} values (half-maximal inhibitory concentration). Notably, dimeric compounds from the $DB_2Py(n)$ and $DB_3P(n)$ series demonstrated superior TDP1 inhibition compared to their monomers. In contrast, all tested compounds showed negligible activity against the other three repair enzymes; so, the compounds demonstrate specificity to TDP1. It should be noted that in this work, in the experiments with TDP1 and TDP2, the effect of the tested compounds as narrow-groove ligands binding to DNA was excluded, and their direct effect on the enzyme was investigated. The results of molecular docking suggest the possibility of direct interaction of active compounds with the active center of TDP1. According to the results of modeling, the inhibitors are located in the binding region of the 3′-end of DNA in the active site of TDP1 and could form stable bonds with the catalytically significant TDP1 residues His263 and His493. These interactions probably provide the high inhibitory activity of the compounds observed in biochemical experiments.

Key words: tyrosyl-DNA phosphodiesterase 1 (TDP1); TDP1 inhibitor; inhibitory activity; TDP2; PARP1; PARP2; DNA-ligands; bisbenzimidazole derivatives

For citation: Dyrkheeva N.S., Chernyshova I.A., Arutyunyan A.F., Zakharenko A.L., Kutuzov M.M., Naumenko K.N., Venzel A.S., Ivanisenko V.A., Deyev S.M., Zhuze A.L., Lavrik O.I. The effect of dimeric bisbenzimidazoles on the activity of DNA repair enzymes TDP1, TDP2, PARP1 and PARP2. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1097-1108. doi 10.18699/vjgb-25-114

Funding. The work was supported by the Russian Science Foundation, grant 25-74-30006 (enzymes purification and activity) and state-funded project for ICBFM SB RAS, grant number 125012300658-9 (oligonucleotide synthesis and infrastructure).

Acknowledgements. The authors express their gratitude to the Center for Collective Use (CCU) "Bioinformatics" for the computational resources and their software. The authors thank Rashid O. Anarbaev (Institute of Chemical Biology and Fundamental Medicine SB RAS) for providing the TDP2 enzyme preparation.

Влияние димерных бисбензимидазолов на активность ферментов репарации ДНК тирозил-ДНК-фосфодиэстераз 1 и 2 и поли(АДФ-рибоза)полимераз 1 и 2

Н.С. Дырхеева [1], И.А. Чернышова [1], А.Ф. Арутюнян [2], А.Л. Захаренко [1], М.М. Кутузов [1], К.Н. Науменко [1], А.С. Вензель [3], В.А. Иванисенко [3], С.М. Деев [4], А.Л. Жузе [2], О.И. Лаврик [1]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук, Москва, Россия
[3] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук, Москва, Россия
lavrik@1bio.ru

Аннотация. Онкологические заболевания остаются одной из главных причин патологической смертности в мире, что определяет дизайн противораковых препаратов как ключевое направление медицинской химии. Комбинация ингибиторов ферментов репарации ДНК с ингибиторами топоизомераз – перспективный подход для усиления противоракового действия и снижения токсичности химиотерапии. Особый интерес представляют узкобороздочные ДНК-лиганды, способные эффективно ингибировать ДНК-зависимые ферменты, обладая при этом низкой токсичностью и мутагенностью. Ряд исследовательских групп, включая нашу, разрабатывает ингибиторы ферментов репарации ДНК, действующие одновременно на несколько взаимосвязанных мишеней {тирозил-ДНК-фосфодиэстеразы 1/2 (TDP1/TDP2), поли(АДФ-рибоза)полимераза 1 (PARP1)/TDP1, топоизомераза 1 (TOP1)/TDP1}. Такие бифункциональные ингибиторы призваны решить проблему резистентности опухолевых клеток к известным химиопрепаратам и повысить эффективность последних. В настоящем исследовании представлены данные скрининга ингибирующей активности 22 узкобороздочных лигандов, взаимодействующих с ДНК, – бис- и трисбензимидазолов – в отношении четырех ферментов репарации: TDP1, TDP2, PARP1 и PARP2. Изучены четыре серии димерных соединений и их мономерных единиц. Показана разница в ингибирующей активности димерных бисбензимидазолов в зависимости от структуры соединения и фермента. Мономерные и димерные бисбензимидазолы эффективно ингибируют активность TDP1 в микромолярном и субмикромолярном диапазоне IC_{50} (концентрация полумаксимального ингибирования). Димерные соединения групп $DB_2Py(n)$ и $DB_3P(n)$ проявили более значительную ингибирующую активность в отношении ферментативной реакции с участием TDP1 по сравнению с мономерами, входящими в их состав. Для всех исследованных соединений была показана низкая ингибирующая способность в отношении остальных трех ферментов репарации ДНК, т. е. наблюдается их специфическое воздействие именно на TDP1. Следует отметить, что в данной работе в экспериментах с TDP1 и TDP2 было исключено действие исследуемых соединений как узкобороздочных лигандов, связывающихся с ДНК, и исследовано их непосредственное воздействие на фермент. По результатам молекулярного докинга можно предположить возможность прямого взаимодействия изучаемых соединений с активным центром TDP1. Согласно результатам моделирования, ингибиторы располагаются в области связывания 3′-конца ДНК с активным центром TDP1 и могут образовывать устойчивые связи с каталитически значимыми остатками активного центра His263 и His493. Эти взаимодействия, вероятно, обеспечивают высокую ингибирующую активность соединений, наблюдаемую в биохимических экспериментах.

Ключевые слова: тирозил-ДНК фосфодиэстераза 1 (TDP1); ингибитор TDP1; ингибирующая активность; TDP2; PARP1; PARP2; ДНК-лиганды; производные бисбензимидазола

Introduction

Nowadays, DNA repair enzymes are actively studied by various researchers to understand the mechanisms of maintaining genetic stability and preventing the development of various diseases. Disruptions in DNA repair systems lead to the accumulation of modified bases, DNA breaks, and other damages, which increase the risk of developing oncological and other diseases. The study of DNA repair system functioning helps to identify the causes of hereditary diseases, neurodegenerative dysfunctions associated with repair defects, and develop new methods for the therapy and prevention of oncological diseases.

In recent years, considerable attention has been paid to DNA repair enzymes as targets for drug development. Researchers are actively searching for new compounds that suppress the activity of DNA repair enzymes to enhance the efficacy of anticancer therapy. Inhibition of enzymes involved in repair increases the effectiveness of antitumor therapy, as this leads to cancer cell death due to the accumulation of DNA damage caused by chemotherapy or radiation therapy. Currently, such repair enzymes as tyrosyl-DNA phosphodiesterases 1 and 2 (TDP1 and TDP2) and poly(ADP-ribose) polymerases 1 and 2 (PARP1 and PARP2) are considered promising targets for drug development (Pommier et al., 2014; Curtin, Szabo, 2020; Zakharenko et al., 2023).

TDP1 is a DNA repair enzyme that participates in the removal of covalent adducts of topoisomerase 1 (TOP1) from DNA, catalyzing the hydrolysis of the phosphodiester bond between the Tyr723 residue of TOP1 and the 3′-phosphate group in the single-strand DNA break generated by TOP1. TDP1 is also capable of removing other DNA-protein adducts located at the 3′-end of DNA and various other damage at the 3′-end of DNA (Comeaux, van Waardenburg, 2014; Kawale, Povirk, 2018). TDP2 catalyzes the hydrolysis of covalent adducts between DNA and the Tyr804 residue of the active center of topoisomerase 2 (TOP2) (Pommier et al., 2010). TDP2 removes covalent adducts from DNA located at the 5′-end of DNA through hydrolysis of the 5′-phosphodiester bond, resulting in the formation of DNA with a free 5′-phosphate (Pommier et al., 2014). TDP1 and TDP2 are capable of taking

over each other's functions to some extent, since TDP1 has low activity in the cleavage of 5′-phosphotyrosyl bonds, while TDP2 has low activity in the cleavage of 3′-phosphotyrosyl bonds (Zeng et al., 2012; Pommier et al., 2014).

Today, topoisomerase inhibitors are widely used in clinical practice as anticancer drugs. The most widely used topoisomerase inhibitors are topotecan and irinotecan, which suppress the activity of topoisomerase 1, as well as etoposide, targeting topoisomerase 2 (Pommier et al., 2010). Their mechanism of action consists in the formation of covalent adducts of topoisomerases with DNA, replication arrest, which ultimately leads to the suppression of cell proliferation. Various researchers have expressed the opinion (Pommier et al., 2014; Zakharenko et al., 2015) that the use of TDP1 and TDP2 inhibitors, which enhance the efficacy of topoisomerase inhibitors, may allow reduction of the dose of these rather toxic drugs and, consequently, the toxicity of therapy. Today, the search for TDP1 inhibitors is actively underway (Zakharenko et al., 2023; Zhang M. et al., 2025). As TDP1 inhibitors, derivatives of natural compounds such as usnic acid, berberines, coumarins, nucleosides, and steroids are particularly notable (Zakharenko et al., 2023), which are effective inhibitors of the purified TDP1 enzyme and topotecan sensitizers in experiments conducted on cellular and mouse cancer models (Zakharenko et al., 2023; Kornienko et al., 2024). Among TDP2 inhibitors, deazaflavins are worth noting, being among the most active inhibitors found to date for this enzyme (Marchand et al., 2016).

The enzymes PARP1 and PARP2 are key regulators of DNA repair and other cellular processes. These enzymes catalyze the DNA-dependent synthesis of the branched polymer poly(ADP-ribose) (PAR) and subsequent ADP-ribosylation of proteins. ADP-ribosylation of proteins is a post-translational modification that is induced in response to DNA damage. PARP1 participates in various DNA repair pathways (Ray Chaudhuri, Nussenzweig, 2017; Lavrik, 2020). PARP2 is also a DNA-dependent PARylation agent and can partially replace PARP1 (Lavrik, 2020; Szanto et al., 2024); therefore, the search for PARP1 and PARP2 inhibitors is an urgent task of modern medicinal chemistry. In clinical practice, such PARP1 and PARP2 inhibitors as olaparib, rucaparib, niraparib, veliparib, and talazoparib are currently approved for use in the treatment of ovarian, fallopian tube, breast, and peritoneal cancer (Kim D.-S. et al., 2021). The inhibitors used today work on the principle of synthetic lethality to destroy cancer cells with defects in the homologous recombination system (for example, with BRCA1/2 mutations), converting single-strand DNA breaks into double-strand breaks that cannot be effectively repaired, leading to cancer cell death. The active sites of PARP1 and PARP2 are very similar (Schreiber et al., 2006; Hoch, Polo, 2019); therefore, the currently known inhibitors most often act on both enzymes, as well as on other enzymes of the PARP family, due to the similarity of their active center that binds nicotinamide adenine dinucleotide (NAD^+) and initiates the synthesis of poly(ADP-ribose), therefore the search for selective inhibitors of each of these enzymes is actively conducted (Johannes et al., 2024). PARP inhibitors approved for clinical use are quite toxic and cause severe side effects, so the search for new inhibitors actively continues (Murai et al., 2014; Kim D.-S. et al., 2021; Johannes et al., 2024).

Small-molecule DNA-binding agents are an extremely promising class of compounds for the search of new inhibitors of repair enzymes. Of particular interest are minor-groove DNA ligands capable of inhibiting DNA-dependent enzymes, while not possessing high toxicity and mutagenicity, and being well soluble in water. Such DNA ligands have a low level of DNA geometry alteration and absence of covalent crosslink formation when forming a complex with DNA (Arutyunyan et al., 2023a).

Our research group has significant experience both in experimental investigation of potential inhibitors at the level of individual protein targets, cells, and animal models (Zakharenko et al., 2023), and in the application of molecular docking and modeling methods to study the mechanisms of interaction of small molecules with target proteins. Effective TDP1 inhibitors have been found that inhibit the recombinant TDP1 enzyme in the submicromolar concentration range. The lead compounds were topotecan sensitizers in experiments conducted on cell cultures and mouse tumor models (Zakharenko et al., 2023; Kornienko et al., 2024). We have developed and investigated inhibitors of PARP1, PARP2, and PARP3 based on conjugates of ADP and morpholino nucleosides using structural modeling of the active sites of these enzymes (Sherstyuk et al., 2019; Chernyshova et al., 2024).

This work presents screening data of twenty-two minor-groove ligands as inhibitors of TDP1, TDP2, PARP1, and PARP2. The studied compounds are bis- and trisbenzimidazole derivatives. Four monomeric compounds – MB_2, MB_2(Ac), MB_2Py(Ac), MB_3 – as well as four series of dimeric derivatives were investigated. The dimeric derivatives were obtained by condensation of monomeric subunits with dicarboxylic acids DB_2P(n), DB_2Py(n), and DB_3P(n), where (n) is the number of methylene units in the linker (Fig. 1).

It was shown that the activity of the compounds varies depending on their structure and the type of enzymatic target. The studied compounds exhibited pronounced inhibitory activity against TDP1, and the observed correlation indicates an increase in inhibitor activity upon introduction of additional binding blocks into its structure, such as a pyrrolecarboxamide fragment for the DB_2Py(n) series, or when using a combination of three benzimidazole blocks in the monomeric subunit. Despite the fact that extremely high IC_{50} values were observed for the DB_3(n) series, this phenomenon can be explained by the high propensity of members of this series of compounds to aggregation, since the introduction of a piperazine fragment into the linker in the DB_3P(n) series led to the obtaining of inhibitors with the lowest IC_{50} values, which indirectly confirms our assumption. In order to elucidate the possible mechanism of their inhibitory action for this enzyme, molecular docking was performed, the results of which suggest the presence of direct interaction between the active compounds and the TDP1 enzyme. According to the constructed binding model, the inhibitors are located in the region of the DNA-binding pocket of TDP1 and are capable of forming stable contacts with the catalytically important amino acid residues His263 and His493. The efficacy of these compounds as TDP1 inhibitors was confirmed by experimental data. The results of the work can be used for the rational design of new, even more effective TDP1 inhibitors.

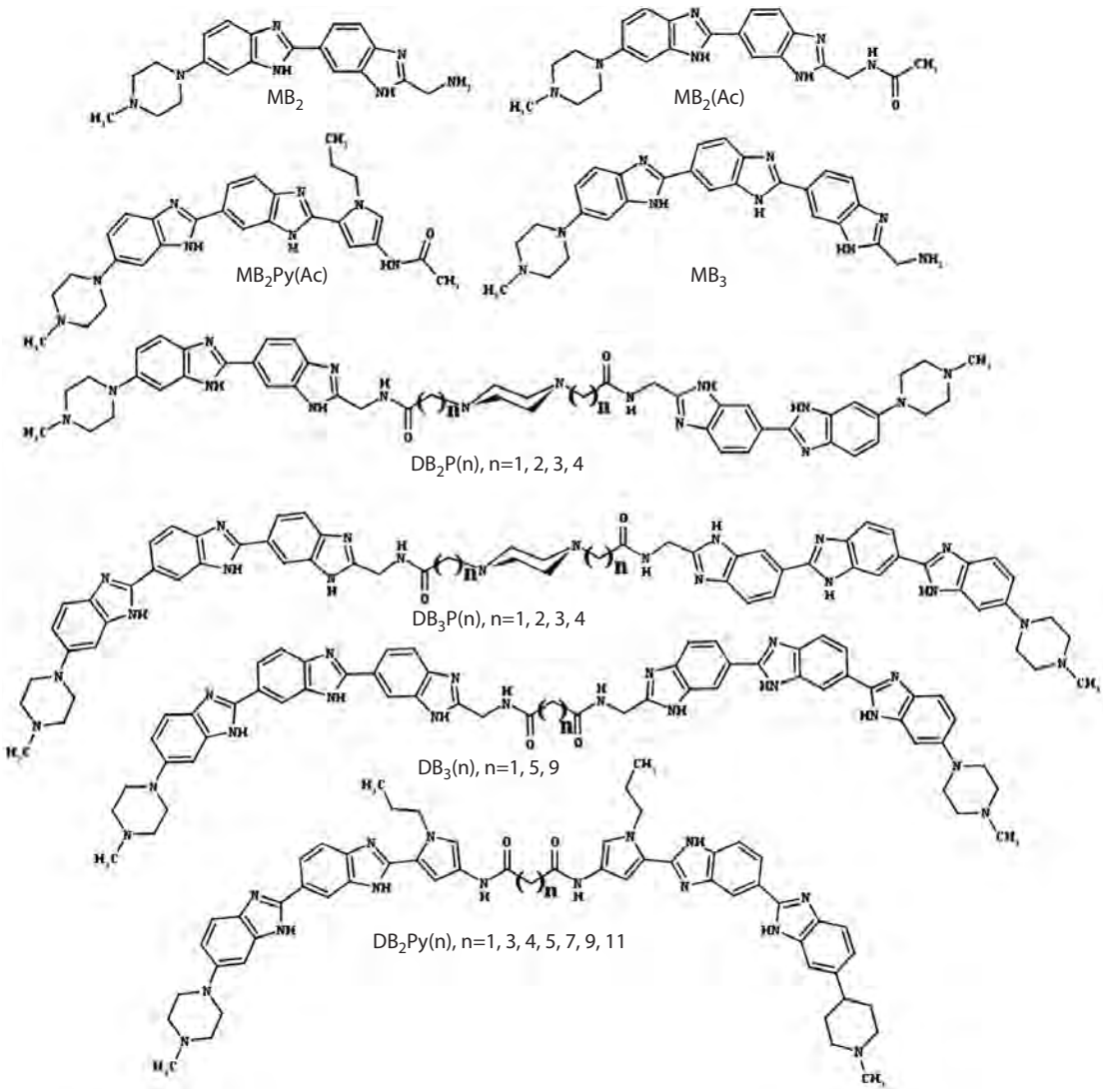


Fig. 1. Structures of bisbenzimidazole derivatives studied in this work.

Materials and methods

Materials and reagents. The studied compounds were synthesized at the Engelhardt Institute of Molecular Biology in the Laboratory of DNA-Protein Interactions according to previously developed methods (Ivanov et al., 2015; Arutyunyan et al., 2023a, b; Susova et al., 2024). The list of IUPAC names of the compounds is provided in the Supplementary Materials[1].

Recombinant human proteins tyrosyl-DNA phosphodiesterase 1 (TDP1) and tyrosyl-DNA phosphodiesterase 2 (TDP2) were expressed in the *E. coli* system, poly(ADP-ribose) polymerase 1 (PARP1) and poly(ADP-ribose) polymerase 2 (PARP2) were expressed in insect cells using a baculovirus expression system and purified as described in (Sukhanova et al., 2004; Sherstyuk et al., 2019; Dyrkheeva et al., 2020, 2021).

The oligonucleotide 5′-FAM-AAC GTC AGG GTC TTC C-BHQ1-3′ was synthesized at the Laboratory of Nucleic Acid Chemistry, Institute of Chemical Biology and Fundamental Medicine (Novosibirsk, Russia), according to (Zakharenko et al., 2015).

[1] Supplementary Tables S1, S2 and Figs S1–S4 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Dyrkheeva_Engl_29_7.pdf

Determination of TDP1 activity. The reaction mixture (200 μl) for real-time fluorescent detection of TDP1 enzyme activity (Zakharenko et al., 2015) contained TDP1 reaction buffer (50 mM Tris-HCl, pH 8.0, 50 mM NaCl, and 7 mM β-mercaptoethanol), 50 nM oligonucleotide 5′-FAM-AAC GTC AGG GTC TTC C-BHQ1-3′, the test compound at various concentrations, and TDP1 at a final concentration of 1.5 nM. The reaction mixtures were incubated at a constant temperature of 26 °C in a POLARstar OPTIMA microplate fluorometer (BMG LABTECH, GmbH, Ortenberg, Germany). Fluorescence intensity (Ex485/Em520 nm) was measured every minute for 10 min. Mean values of half-maximal inhibitory concentration (IC_{50} – the concentration of the compound that inhibited 50 % of enzyme activity compared to the untreated control well containing only enzyme and substrate) were determined using a dose-response curve of the fluorescence signal level versus inhibitor concentration and calculated using MARS Data Analysis 2.0 (BMG LABTECH). Kinetic curves were obtained in at least three independent experiments and statistically processed in OriginPro 8.6.0 (OriginLab, Northampton, Massachusetts, USA).

Determination of TDP2 activity. For determination of TDP2 enzyme activity, an oligonucleotide 5′-tyrosine-AAC GTC AGG GTC TTC C-FAM-3′ containing a 6-FAM label at the 3′-end and an L-tyrosine residue attached via the phenolic OH group to the 5′-terminal phosphate was used as substrate, synthesized at the Russian-French-Japanese Laboratory of Bionanotechnology of Novosibirsk State University as described in (Dyrkheeva et al., 2021). The substrate at a concentration of 100 nM was incubated with TDP2 at a concentration of 200 nM in the absence or presence of inhibitor (500 μM) for 10 min at 37 °C in buffer containing 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 7 mM β-mercaptoethanol (Dyrkheeva et al., 2021). The reaction was stopped by addition of PAGE loading buffer (TBE, 10 % formamide, 7 M urea, 20 mM EDTA, 0.1 % xylene cyanol, and 0.1 % bromophenol blue). The samples were then heated at 90 °C for 5 min. The enzymatic reaction products were separated by electrophoresis in 20 % denaturing PAGE with 7 M urea at an acrylamide to bisacrylamide ratio of 19:1. A high-resolution Typhoon FLA 9500 laser scanner (GE Healthcare, Chicago, Illinois, USA) was used for gel scanning and visualization, and the data were analyzed using QuantityOne 4.6.7 software (Bio-Rad Laboratories, Inc., Hercules, California, USA). At least three independent experiments were performed, and statistical processing was carried out using OriginPro 8.6.0 (OriginLab, Northampton, Massachusetts, USA).

Determination of PARP1 and PARP2 activity. For determination of PARP1 and PARP2 enzyme activity in the presence and absence of test compounds, radiolabeled [^{32}P]-NAD$^+$ was synthesized from α-[^{32}P]-ATP according to the protocol (Sherstyuk et al., 2019). The auto-poly(ADP-ribosyl)ation reaction was performed in buffer for PARP1: 50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 150 mM NaCl, and 7 mM β-mercaptoethanol, as well as 2 A_{260} units/ml activated DNA, 0.3 mM [^{32}P]-NAD$^+$ at 37 °C. The reaction was initiated by addition of PARP1 to 200 nM and carried out for 2 min.

The buffer for PARP2 contained: 50 mM Tris-HCl, pH 8.0, 3 mM spermine, 150 mM NaCl, and 7 mM β-mercaptoethanol, 2 A_{260} units/ml activated DNA, 0.6 mM [^{32}P]-NAD$^+$ at 37 °C. The reaction was initiated by addition of PARP2 to 600 nM, and the reaction mixtures were incubated for 5 min. The reaction was stopped by placing 5 μl aliquots on Whatman 1 paper filters impregnated with 5 % trichloroacetic acid (TCA). The filters were washed with 5 % TCA four times and air-dried after removal of TCA with 90 % ethanol. The incorporation of the radioactive label into the reaction product was calculated using a Typhoon FLA 9500 scanner (GE Healthcare, Chicago, Illinois, USA). At least three independent experiments were performed.

Molecular modeling. To evaluate the interaction of the studied compounds with the TDP1 enzyme, we performed molecular docking followed by analysis of the resulting complexes. The study included preparation of protein and ligand structures, molecular docking, energy minimization of compounds in the binding site, and assessment of inhibitor affinity using the Vinardo, X-Score, and REF2015 scoring functions.

The crystal structure of TDP1 (PDB ID: 8V0B) was used as the target protein structure. Missing loops in the model were reconstructed based on AlphaFold2 prediction (Jumper et al., 2021) performed in ColabFold (Mirdita et al., 2022) without using multiple sequence alignment (MSA) and using 8V0B as a template.

Hydrogen atoms were then added to the resulting model and charges were calculated using the DockPrep utility in UCSF Chimera (Pettersen et al., 2004). The inhibitor structures were prepared in OpenBabel (O'Boyle et al., 2011): hydrogens were added, partial charges were calculated, and geometry minimization was performed.

Molecular docking was performed using the UCSF DOCK 6.11 software package (Allen et al., 2015). Full-atom flexible docking over the entire protein surface was used. At the first stage of docking, the core fragments of the inhibitors (MB_2(Ac), MB_2Py(Ac)) were positioned, after which full-length molecules were docked with subsequent minimization of their energy in the binding site. Up to nine best conformations by GridScore were requested for each compound. From the nine conformations obtained for each ligand, the structure with the minimum RMSD relative to the optimal conformation of the core fragment was selected. In cases where DOCK6 returned fewer than nine unique conformations (due to clustering, energy filtering, or failure to generate additional conformers), selection was performed from all available conformations (Table S1).

Final assessment of the inhibitors' binding ability to the protein was performed using several independent scoring functions: ContinuousScore from DOCK 6, Vinardo (Quiroga, Villarreal, 2016), X-Score (Wang R. et al., 2002), and REF2015 in the PyRosetta4 environment (Chaudhury et al., 2010; Alford et al., 2017) according to the protocol of Moretti et al. (2016). ContinuousScore is a scoring function in DOCK 6 that accounts for van der Waals interactions, electrostatic interactions, internal ligand energy, and penalties for steric clashes through direct calculation of interatomic distances. Vinardo is a scoring function for docking that accounts for the contribution of hydrogen bonds, hydrophobic and van der Waals interactions, as well as corrections for non-optimal ligand positioning. The X-Score scoring function consists of three components: HPScore, HMScore, and HSScore, based on different empirical principles for assessing ligand-protein affinity. In this study, the averaged X-Score was used, reflecting the influence of hydrophobic, polar, and electrostatic contacts. The full-atom REF2015 scoring function implemented in PyRosetta includes contributions from van der Waals, electrostatic, hydrogen bonding, solvation, and additional atom pair interactions and allows correct ranking of inhibitor positions close in energy.

To validate the molecular docking results and assess the stability of the predicted complex over time, molecular dynamics simulation of the TDP1 complex with the lead compound DB_2Py(1), which had shown the best inhibitory activity, was performed. The simulation was carried out using the OpenMM 8 package (Eastman et al., 2017). A detailed protocol of the molecular dynamics simulation is presented in the Supplementary Materials.

Results

In this work, the ability of four series of small-molecule dimeric DNA ligands DB_2P(n), DB_2Py(n), DB_3(n), DB_3P(n) as well as their monomeric units MB_2, MB_2(Ac), MB_2Py(Ac),

and MB_3 (Fig. 1) to inhibit the activity of recombinant DNA repair enzymes TDP1 and TDP2, PARP1 and PARP2 was studied for the first time (see the Table).

The first group of studied compounds represents dimeric derivatives of the monomeric bisbenzimidazole ligand MB_2, a derivative of the widely studied minor-groove DNA ligand Hoechst 33258, in which the hydroxyphenyl group is replaced by a more hydrophilic aminomethylene fragment – $DB_2P(n)$. As a linker for compounds of this group, 1,4-piperazine-dialkyldicarboxylic acids containing a methylene, ethylene, propylene, or butylene spacer were used (Fig. 1). This series was also supplemented with the monomeric derivative $MB_2(Ac)$, acylated at the aminomethylene fragment, which structurally brings this compound, compared to MB_2, closer to half of the dimeric compound $DB_2P(n)$ and makes it a more appropriate reference for comparison. The $DB_2P(n)$ series differs from other ligand series by the presence of a positively charged 1,4-piperazine introduced into the linker, which improves ligand solubility and may increase ligand affinity for the enzyme.

The next group of compounds are derivatives of the monomeric trisbenzimidazole compound MB_3, which can be considered as a derivative of MB_2 containing one additional benzimidazole fragment, which increases the number of potentially possible hydrogen bonds in the inhibitor-TDP1 complex. Dimeric derivatives of MB_3 are represented by two series of compounds – $DB_3P(n)$, also dimerized with 1,4-piperazinedialkyldicarboxylic acids, and $DB_3(n)$, where n-alkyldicarboxylic acids are used as linkers. The $DB_3(n)$ and $DB_3P(n)$ series are characterized by the presence of trisbenzimidazoles in the structure, and $DB_3P(n)$, also by the presence of 1,4-piperazine in the linker.

The third group of compounds includes derivatives of the monomeric compound $MB_2Py(Ac)$, which is an isosteric analog of MB_3, due to the fact that the pyrrolecarboxamide fragment contained in its structure can act as a hydrogen

Inhibitory activity of test compounds against TDP1, TDP2, PARP1, and PARP2

No.	Compounds	IC_{50} TDP1, μM	TDP2	PARP1	PARP2
			% of residual activity (500 μM)		
1	MB_2	2 ± 1	~100	~100	~100
2	$MB_2(Ac)$	1.5 ± 0.5	~100	~100	~100
3	$DB_2P(1)$	6 ± 4	66 ± 7	57 ± 16	~100
4	$DB_2P(2)$	9 ± 3	44 ± 11	51 ± 15	80 ± 20
5	$DB_2P(3)$	4.1 ± 0.6	36 ± 7	37 ± 10	64 ± 16
6	$DB_2P(4)$	2.3 ± 0.3	44 ± 11	33 ± 13	85 ± 13
7	$MB_2Py(Ac)$	5 ± 2	~100	~100	~100
8	$DB_2Py(1)$	0.25 ± 0.05	55 ± 3	~100	~100
9	$DB_2Py(3)$	0.41 ± 0.09	70 ± 11	~100	~100
10	$DB_2Py(4)$	0.4 ± 0.15	~100	~100	~100
11	$DB_2Py(5)$	0.35 ± 0.13	~100	~100	~100
12	$DB_2Py(7)$	0.28 ± 0.01	~100	~100	~100
13	$DB_2Py(9)$	0.30 ± 0.08	~100	~100	~100
14	$DB_2Py(11)$	0.9 ± 0.1	~100	~100	~100
15	MB_3	0.70 ± 0.05	~100	65 ± 15	~100
16	$DB_3(1)$	>50	70 ± 6	55 ± 13	~100
17	$DB_3(5)$	>50	65 ± 10	62 ± 16	~100
18	$DB_3(9)$	>50	~100	~100	~100
19	$DB_3P(1)$	0.10 ± 0.05	~100	70 ± 12	~100
20	$DB_3P(2)$	0.11 ± 0.01	~100	40 ± 5	~100
21	$DB_3P(3)$	0.20 ± 0.05	~100	47 ± 14	~100
22	$DB_3P(4)$	0.15 ± 0.03	~100	48 ± 15	~100

Note. For IC_{50} values and percentage of residual enzyme activity in the presence of inhibitor, the Table shows mean values ± standard deviation (at least three replicates).

atom donor at the carboxamide nitrogen for hydrogen bond formation, in a position analogous to benzimidazole. Dimeric derivatives are represented by the $DB_2Py(n)$ series containing n-alkyldicarboxylic acids as a linker. This series is represented by a set of compounds containing 1, 3, 4, 5, 7, 9, and 11 methylene units, which allowed for a more accurate assessment of the dependence of the inhibitory activity of compounds on spacer length. The $DB_2Py(n)$ series differs from the $DB_3(n)$ series by the presence, in addition to the bisbenzimidazole structure, of a pyrrolecarboxamide structure, which is a fragment of the AT-specific antibiotic netropsin.

Using the real-time fluorescence analysis method, half-maximal inhibitory concentration (IC_{50}) values of the studied compounds (see the Table) were obtained in the reaction of BHQ1 cleavage from the 3′-end of the oligonucleotide by TDP1, which led to an increase in FAM fluorescence at the 5′-end of the chain (Zakharenko et al., 2015). It should also be noted that a single-stranded oligonucleotide was used as substrate to exclude the binding of dimeric bisbenzimidazoles as minor-groove ligands to the DNA substrate and direct their action toward the enzymatic target.

From the data obtained for the monomeric compounds MB_2 and $MB_2(Ac)$ and their dimeric derivatives $DB_2P(n)$, at n = 1, 2, 3, 4, the IC_{50} values were in the micromolar range, and dimerization did not lead to an increase in the inhibitory activity of the studied compounds. At the same time, for dimers of the monomeric $MB_2Py(Ac)$, which has an IC_{50} value of 5 ± 2 μM, the half-inhibitory concentration parameter value decreased significantly, ranging from 0.25 to 0.90 μM. Similarly, the transition from monomeric MB_3 to the dimeric $DB_3P(n)$ series led to an increase in the inhibitory activity of the compounds, although not as pronounced; however, dimeric derivatives of MB_3 that do not contain a piperazine fragment in the linker – $DB_3(n)$ compounds – showed the lowest level of activity among all the inhibitors tested in this work. The fact that the IC_{50} values for these compounds (see the Table) deviate so strongly from the overall data set is most likely due to the fact that $DB_3(n)$ compounds possess an extended and planar geometry, as well as a rigid linker, which prevents optimal positioning of compounds of this type in the enzyme active site (Fig. 1).

Thus, according to the experimental data, all compounds studied in this work, except for the $DB_3(n)$ group, effectively inhibit TDP1 activity at micromolar and submicromolar concentrations. A structure-activity correlation is observed, consisting of a decrease in concentration to achieve the half-maximal inhibition effect with an increase in the number of blocks containing hydrogen bond donors in the compound. In particular, dimerization is one of the simple approaches to increasing such structures in one molecule, which leads to a nonlinear increase in the binding constant (Neudachina, Lakiza, 2014). A decrease in IC_{50} is also observed upon introduction of a piperazine fragment into the linker structure, which may be due to an increase in the hydrophilicity of the molecules. The results obtained allowed us to establish a structure-activity correlation, as well as to assess the contribution of dimerization to the increase of the inhibitory capacity of the studied compounds.

To study the effect of the studied compounds on TDP2 activity, we tested the ability of this enzyme to remove the tyrosine residue from the 5′-end of the oligonucleotide substrate in the absence and presence of inhibitors, as described in (Dyrkheeva et al., 2021). All compounds of the $DB_2P(n)$ group, as well as $DB_2Py(n)$, at n = 1, 3 and $DB_3(n)$, at n = 1, 5 at a concentration of 500 μM inhibited enzyme activity by approximately 50 %, while all other compounds showed no inhibitory activity (see the Table). Thus, all tested compounds showed a significantly lower propensity to inhibit TDP2 compared to TDP1. Interestingly, the $DB_2P(n)$ group inhibited TDP1 less effectively and TDP2 more effectively than compounds of other groups.

The next step of our work was to test the ability of the studied compounds to inhibit PARP1 and PARP2, that is, their enzymatic activity in the poly(ADP-ribose) (PAR) synthesis reaction, at a rather high concentration range of compounds. All studied compounds showed low efficiency in inhibiting these two enzymes. The most active compounds were those of the $DB_2P(n)$ group, representatives of which with n = 2, 3, 4 reduced the activity of PARP1 and PARP2 at a concentration of 500 μM. Inhibitory action was also observed for compounds of the $DB_3(n)$ and $DB_3P(n)$ series at a concentration of 500 μM, while these compounds exhibited inhibitory activity only in the PAR synthesis reaction catalyzed by PARP1, but not PARP2 (see the Table).

Since, according to the experimental data, all studied compounds, with the exception of the $DB_3(n)$ group, effectively inhibit TDP1 activity, we further performed an *in silico* evaluation of the ability of compounds of the $DB_2P(n)$ and $DB_2Py(n)$ groups to bind to the TDP1 enzyme in order to elucidate the possible molecular mechanism of their inhibitory action. For this purpose, full-atom flexible molecular docking over the entire surface of the TDP1 protein (PDB ID: 8V0B) was performed for $DB_2P(n)$ and $DB_2Py(n)$ compounds.

According to the docking results obtained, it can be assumed that conformations with minimum calculated energy for each inhibitor form interactions in the TDP1 active site, near His263 and His493 residues (Fig. 2*a*), similarly to compound $MB_2(Ac)$ (Fig. S1). An additional analysis of the binding ability of dimeric compounds to TDP1 was performed using the Vinardo, X-Score, and REF2015 scoring functions in the PyRosetta environment (Table S2). The obtained scoring function values suggest high affinity of the studied inhibitors of the $DB_2P(n)$ and $DB_2Py(n)$ groups for TDP1. It should be noted that complete correlation of the parameters obtained by docking (Table S2) with the IC_{50} values found experimentally (see the Table) is not observed, which can be explained by the contribution of hydrophobic linkers, which are difficult to account for in energy calculations.

According to molecular modeling data, compound $MB_2(Ac)$ (Fig. 2*b*), which is the monomeric unit for dimeric derivatives $DB_2P(n)$, may form a hydrogen bond with His263 and a π-cation interaction with His493, which could potentially lead to blocking of the TDP1 catalytic act. In addition to interactions with catalytically active residues, $MB_2(Ac)$ may form hydrophobic contacts with Tyr204 and Ala520, as well as a hydrogen bond with Phe259, which could enhance the inhibitory action of this compound. In contrast to $MB_2(Ac)$, compound $MB_2Py(Ac)$ (Fig. 2*c*) appears to interact with only one catalytic residue – His493 – through hydrogen bond formation. Such a difference in interactions could be the reason for the higher inhibitory activity of $MB_2(Ac)$ compared to

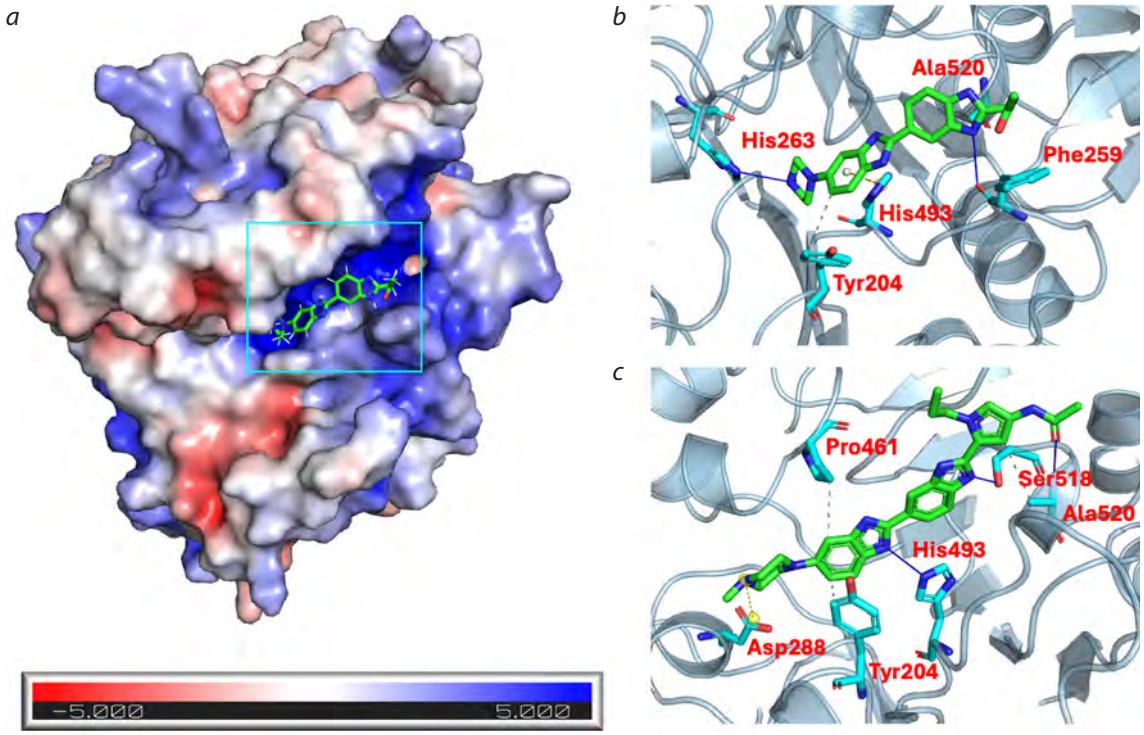


Fig. 2. *a*, Structure of TDP1 (PDB ID: 8V0B) with inhibitor MB_2(Ac) located in the positively charged region of the TDP1 active site. The protein surface is colored according to the electrostatic potential distribution calculated using APBS (Jurrus et al., 2018). The DNA-binding region of TDP1 is highlighted by a rectangular frame. Below is a scale of TDP1 surface electrostatic potential values (in units of kT/e, where kT/e ≈ 25.7 mV at 298 K). Color scale: red indicates negative potential (–5 kT/e), white indicates neutral (0 kT/e), blue indicates positive potential (+5 kT/e). *b*, *c*, Predicted conformations of inhibitors MB_2(Ac) and MB_2Py(Ac) (green) in complex with TDP1 with contacting residues (cyan).

MB_2Py(Ac), which is consistent with experimental data (see the Table).

Analysis of interactions using PLIP (Protein–Ligand Interaction Profiler) (Salentin et al., 2015) for predicted TDP1 complexes with dimeric compounds of the DB_2Py(n) group (Fig. S2) showed that these compounds form a greater number of protein-ligand contacts (hydrogen bonds and hydrophobic interactions) compared to the MB_2Py(Ac) monomer. In particular, compound DB_2Py(1) forms hydrogen bonds with Ser400 and Ser403, as well as hydrophobic interactions with Pro463 – the residues of these amino acids are located in the ligand binding site with the TDP1 active center, which likely contributes to stabilization of the interacting dimer fragment in the enzyme active site. The data obtained from docking analysis, characterizing the larger contact surface area of dimeric DB_2Py(n) compounds with TDP1 compared to the MB_2Py(Ac) monomer, correlate with the decrease in IC_{50} values for dimers, which indicates an increase in the affinity of these compounds for the enzyme active site (see the Table). According to the data obtained, hydrophobic interactions with Pro461 and/or Tyr204 residues localized in the TDP1 active site may also contribute to increasing the inhibitory activity of DB_2Py(n) group compounds.

Analysis of interactions of compounds from the DB_2P(n) group with TDP1 showed that analogous amino acid residues participate in complex formation, with the exception of Tyr204, with which DB_2P(n) compounds, unlike DB_2Py(n), apparently do not interact (Fig. S3). In addition, possible differences in the nature of interactions with the same amino acids were noted. For example, for the Lys519 residue in the case of DB_2P(n) compounds, formation of hydrogen bonds with nitrogen atoms of the piperazine fragment through the N1 atom of the side chain can be assumed. At the same time, two types of interactions with Lys519 are predicted in DB_2Py(n) compounds: a hydrogen bond between the backbone nitrogen atom of Lys519 and the oxygen atom in the pyrrolecarboxamide group (in DB_2Py(1), DB_2Py(4), DB_2Py(7), DB_2Py(9)), as well as a π-cation interaction between pyrrole and the Lys519 side chain (in DB_2Py(3) and DB_2Py(5)) (Fig. S2).

For compound DB_2Py(1), which demonstrated the highest inhibitory activity (lowest IC_{50} value) among the studied derivatives, additional molecular dynamics modeling in the predicted complex with TDP1 was performed. Analysis of the MD trajectory showed that the TDP1-DB_2Py(1) complex maintains stability throughout the simulation time. RMSD values of the ligand were in the range of 1.5–3.0 Å (Fig. S4), which indicates stable binding of DB_2Py(1) in the protein active site without signs of dissociation or significant conformational rearrangements. The data obtained confirm the strength of the formed complex and are consistent with the high biological activity of this compound.

It should be noted that our analysis of molecular contacts, as well as the scoring function values obtained according to molecular docking results, indicate the ability of compounds of both analyzed groups – DB_2P(n) with an aliphatic linker and DB_2Py(n) with a piperazine fragment in the linker – to

form a stable complex with TDP1. Nevertheless, experimental data show differences in their inhibitory activity: compounds with an aliphatic linker demonstrate higher inhibition efficiency compared to compounds containing a piperazine ring. This difference cannot be fully explained based on contact analysis, which suggests a possible difference in the conformational mobility of these groups of compounds. In particular, the inclusion of a piperazine fragment in the central part of the linker apparently restricts its flexibility, which affects the dynamics of inhibitor interaction with the active site, prevents optimal positioning of the inhibitor in the enzyme active site and, consequently, reduces its inhibitory activity.

Discussion

TDP1 plays a key role in eliminating DNA damage located at the 3′-end of DNA, stabilized by anticancer drugs used in clinical practice, such as topotecan and irinotecan, which are derivatives of the natural compound camptothecin (Comeaux, van Waardenburg, 2014; Kawale, Povirk, 2018). Consequently, TDP1 activity may be a possible cause of tumor resistance to TOP1 inhibitors used in the clinic. Currently, searches for combined TOP1 and TDP1 inhibitors are actively underway (Conda-Sheridan et al., 2013; Nguyen et al., 2015; Zhang X.-R. et al., 2018; Hu et al., 2021;Yang et al., 2023).

Furthermore, since it is known that the activities of TDP1 and TDP2 overlap, albeit to a minor extent (Pommier et al., 2014), the ability of these enzymes to perform each other's functions makes the combined use of inhibitors of these two enzymes or the creation of agents capable of simultaneously inhibiting both TDP1 and TDP2 quite promising. Simultaneous suppression of the activity of these two enzymes can be used to enhance the efficacy of a large set of clinically important anticancer drugs, TOP1 and TOP2 inhibitors. Triple TOP1/TDP1/TDP2 inhibitors have also been discovered, which exhibit moderate activity against TDP1 and weak activity against TDP2 (Wang P. et al., 2017). The most effective TDP2 inhibitors to date are deazaflavins, which exhibit synergy with etoposide *in vitro* at non-toxic concentrations (Marchand et al., 2016), and some effective TDP2 inhibitors from other compound classes have also been found (Yang et al., 2021; Zhang Y. et al., 2021).

It is known that the N-terminal domain of TDP1 directly binds to the C-terminal domain of PARP1, and TDP1 undergoes PARylation by PARP1 in order to be recruited to the TOP1-DNA adduct (Das et al., 2014; Lebedeva et al., 2015). PARylation of TDP1 stimulates its recruitment to sites with damaged DNA without affecting the catalytic activity of this enzyme (Chowdhuri, Das, 2021). It has also been shown that PARP1 can interact with TDP1, forming protein-protein contacts (Moor et al., 2015). It was established that the combination of TDP1 knockdown and inhibition of PARP1 activity with rucaparib reduces cell proliferation more significantly than these methods of enzyme function suppression separately (Fam et al., 2013). Therefore, there is a suggestion in the literature that the anticancer effect of TOP1 inhibitors can be significantly enhanced by simultaneous inhibition of PARP1 and TDP1 (Smith et al., 2005; Alagoz et al., 2014; Das et al., 2014; Murai et al., 2014; Elsayed et al., 2016; Matsuno et al., 2018; Jing et al., 2020; Kim J.W. et al., 2020; Chowdhuri, Das, 2021; Flörkemeier et al., 2022). The interaction between PARP1 and TDP1 enzymes has been demonstrated in a number of publications (Das et al., 2014; Moor et al., 2015), which makes the search for dual TDP1 and PARP1 inhibitors relevant.

Previously, we discovered dual TDP1 and TDP2 inhibitors, as well as triple TDP1, TDP2, and PARP1 inhibitors (Dyrkheeva et al., 2021) – usnic acid thioethers that weakly inhibit TDP2 and PARP1; therefore, the search for new compounds capable of acting on two or three functionally interacting targets simultaneously is relevant. In this work, the ability of a series of minor-groove DNA ligands to inhibit TDP1, TDP2, PARP1, and PARP2 enzymes was tested. Effective inhibitors acting on all four enzymes simultaneously were not found, but it was shown that these compounds inhibit TDP1. The DNA ligands studied in this work are capable of inhibiting DNA-dependent enzymes through binding to double-stranded DNA. However, in the present work we showed that they are capable of selectively inhibiting TDP1, since the experiments were conducted in the absence of double-stranded DNA as an alternative target.

The results of molecular docking and analysis of intermolecular interactions suggest that most of the studied compounds of the $DB_2P(n)$ and $DB_2Py(n)$ groups may possess high affinity for the TDP1 enzyme and form stable complexes with its catalytic center. Interactions with key catalytic residues of the TDP1 protein active site were predicted for all compounds.

Conclusion

In this work, a study of the effect of dimeric bis- & tris-benzimidazoles on the activity of DNA repair enzymes – TDP1, TDP2, PARP1, and PARP2 – was conducted. The main results showed that all studied inhibitors, except compounds of the $DB_3(n)$ series, effectively inhibit TDP1. The most active were compounds $DB_2Py(n)$ and $DB_3P(n)$, capable of inhibiting TDP1 in the submicromolar concentration range. The studied compounds demonstrate high selectivity, with minimal effect on the activity of other tested enzymes.

Based on the results of molecular docking, it is proposed that the studied active inhibitors are localized in the region of the DNA-binding pocket of TDP1 and may form stable interactions with the catalytically important residues His263 and His493. These interactions likely underlie the observed high inhibitory activity.

An important result is also the establishment of the structure-activity relationship. Dimerization had a mixed effect on the inhibitory effect: compounds of the $DB_2Py(n)$ and $DB_3P(n)$ series were significantly (by an order of magnitude) more active than the corresponding monomers; in the $DB_2P(n)$ series, the inhibitory activity was influenced not only by dimerization, but also by linker length and the introduction of 1,4-piperazine bearing two positive charges into the linker. The $DB_3(n)$ series was inactive, unlike the monomer. Introduction of the piperazine fragment into the linker in the $DB_3P(n)$ series led to pronounced inhibitory activity compared to $DB_3(n)$ without such a fragment. We propose that the enhancement of the inhibitory effect is related to the introduction of two positive charges into the linker and to the increase in the number of possible contacts of ligands with the enzyme active site.

Overall, based on the results of this work, new strategies for the development of cancer therapy may be proposed. The

obtained data also highlight the potential of dimeric bis- & tris-benzimidazoles as safe and effective tools for targeted regulation of DNA repair enzymes.

References

Alagoz M., Wells O.S., El-Khamisy S.F. TDP1 deficiency sensitizes human cells to base damage via distinct topoisomerase I and PARP mechanisms with potential applications for cancer therapy. *Nucleic Acids Res*. 2014;42(5):3089-3103. doi 10.1093/nar/gkt1260

Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., DiMaio F.P., Park H., Shapovalov M.V., … Das R., Baker D., Kuhlman B., Kortemme T., Gray J.J. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*. 2017; 13(6):3031-3048. doi 10.1021/acs.jctc.7b00125

Allen W.J., Balius T.E., Mukherjee S., Brozell S.R., Moustakas D.T., Lang P.T., Case D.A., Kuntz I.D., Rizzo R.C. DOCK 6: impact of new features and current docking performance. *J Comput Chem*. 2015;36(15):1132-1156. doi 10.1002/jcc.23905

Arutyunyan A.F., Kostyukov A.A., Korolev S.P., Gottikh M.B., Kaluzhny D.N., Susova O.Yu., Zhuze A.L. DNA sequence-specific ligands. 19. Synthesis, spectral properties, virological and biochemical studies of $DB_3(n)$ fluorescent dimeric trisbenzimidazoles. *Mol Biol*. 2023a;57(3):512-521. doi 10.1134/s0026893323030020

Arutyunyan A.F., Kostyukov A.A., Lushpa V.A., Mineev K.S., Korolev S.P., Gottikh M.B., Klimova R.R., Kushch A.A., Kalabina K.V., Susova O.Yu., Zhuze A.L. DNA sequence-specific ligands. XX. Synthesis, spectral properties, virological and biochemical studies of fluorescent dimeric trisbenzimidazoles $DB_3P(n)$. *Med Chem Res*. 2023b;32(3):587-599. doi 10.1007/s00044-023-03017-x

Chaudhury S., Lyskov S., Gray J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*. 2010;26(5):689-691. doi 10.1093/bioinformatics/btq007

Chernyshova I., Vasil'eva I., Moor N., Ivanisenko N., Kutuzov M., Abramova T., Zakharenko A., Lavrik O. Aminomethylmorpholino nucleosides as novel inhibitors of PARP1 and PARP2: experimental and molecular modeling analyses of their selectivity and mechanism of action. *Int J Mol Sci*. 2024;25(23):12526. doi 10.3390/ijms252312526

Chowdhuri S.P., Das B.B. Top1-PARP1 association and beyond: from DNA topology to break repair. *NAR Cancer*. 2021;3(1):zcab003. doi 10.1093/narcan/zcab003

Comeaux E.Q., van Waardenburg R.C. Tyrosyl-DNA phosphodiesterase I resolves both naturally and chemically induced DNA adducts and its potential as a therapeutic target. *Drug Metab Rev*. 2014;46(4):494-507. doi 10.3109/03602532.2014.971957

Conda-Sheridan M., Reddy P.V.N., Morrell A., Cobb B.T., Marchand C., Agama K., Chergui A., Renaud A., Stephen A.G., Bindu L.K., Pommier Y., Cushman M. Synthesis and biological evaluation of indenoisoquinolines that inhibit both tyrosyl-DNA phosphodiesterase I (Tdp1) and topoisomerase I (Top1). *J Med Chem*. 2013;56(1):182-200. doi 10.1021/jm3014458

Curtin N.J., Szabo C. Poly(ADP-ribose) polymerase inhibition: past, present and future. *Nat Rev Drug Discov*. 2020;19(10):711-736. doi 10.1038/s41573-020-0076-6

Das B.B., Huang S.N., Murai J., Rehman I., Amé J.-C., Sengupta S., Das S.K., Majumdar P., Zhang H., Biard D., Majumder H.K., Schreiber V., Pommier Y. PARP1-TDP1 coupling for the repair of topoisomerase I-induced DNA damage. *Nucleic Acids Res*. 2014;42(7): 4435-4449. doi 10.1093/nar/gku088

Dyrkheeva N., Anarbaev R., Lebedeva N., Kuprushkin M., Kuznetsova A., Kuznetsov N., Rechkunova N., Lavrik O. Human tyrosyl-DNA phosphodiesterase 1 possesses transphosphooligonucleotidation activity with primary alcohols. *Front Cell Dev Biol*. 2020;8: 604732. doi 10.3389/fcell.2020.604732

Dyrkheeva N.S., Filimonov A.S., Luzina O.A., Orlova K.A., Chernyshova I.A., Kornienko T.E., Malakhova A.A., … Burakova E.A., Stetsenko D.A., Zakian S.M., Salakhutdinov N.F., Lavrik O.I. New hybrid compounds combining fragments of usnic acid and thioether are inhibitors of human enzymes TDP1, TDP2 and PARP1. *Int J Mol Sci*. 2021;22(21):11336. doi 10.3390/ijms222111336

Eastman P., Galvelis R., Peláez R.P., Abreu C.R.A., Farr S.E., Gallicchio E., Gorenko A., … Wang Y., Zhang I., Chodera J.D., De Fabritiis G., Markland T.E. OpenMM 8: molecular dynamics simulation with machine learning potentials. *J Phys Chem B*. 2024;128(1):109-116. doi 10.1021/acs.jpcb.3c06662

Elsayed W., El-Shafie L., Hassan M.K., Farag M.A., El-Khamisy S.F. Isoeugenol is a selective potentiator of camptothecin cytotoxicity in vertebrate cells lacking TDP1. *Sci Rep*. 2016;6(1):26626. doi 10.1038/srep26626

Fam H.K., Walton C., Mitra S.A., Chowdhury M., Osborne N., Choi K., Sun G., … Aparicio S., Triche T.J., Bond M., Pallen C.J., Boerkoel C.F. TDP1 and PARP1 deficiency are cytotoxic to rhabdomyosarcoma cells. *Mol Cancer Res*. 2013;11(10):1179-1192. doi 10.1158/1541-7786.mcr-12-0575

Flörkemeier I., Hillmann J.S., Weimer J.P., Hildebrandt J., Hedemann N., Rogmans C., Dempfle A., Arnold N., Clement B., Bauerschlag D.O. Combined PARP and dual topoisomerase inhibition potentiates genome instability and cell death in ovarian cancer. *Int J Mol Sci*. 2022;23(18):10503. doi 10.3390/ijms231810503

Hoch N.C., Polo L.M. ADP-ribosylation: from molecular mechanisms to human disease. *Genet Mol Biol*. 2019;43(Suppl.1):e20190075. doi 10.1590/1678-4685-GMB-2019-0075

Hu D.-X., Tang W.-L., Zhang Y., Yang H., Wang W., Agama K., Pommier Y., An L.-K. Synthesis of methoxy-, methylenedioxy-, hydroxy-, and halo-substituted benzophenanthridinone derivatives as DNA topoisomerase IB (TOP1) and tyrosyl-DNA phosphodiesterase 1 (TDP1) inhibitors and their biological activity for drug-resistant cancer. *J Med Chem*. 2021;64(11):7617-7629. doi 10.1021/acs.jmedchem.1c00318

Ivanov A.A., Koval V.S., Susova O.Yu., Salyanov V.I., Oleinikov V.A., Stomakhin A.A., Shalginskikh N.A., Kvasha M.A., Kirsanova O.V., Gromova E.S., Zhuze A.L. DNA specific fluorescent symmetric dimeric bisbenzimidazoles DBP(n): the synthesis, spectral properties, and biological activity. *Bioorg Med Chem Lett*. 2015;25(13):2634-2638. doi 10.1016/j.bmcl.2015.04.087

Jing C.-B., Fu C., Prutsch N., Wang M., He S., Look A.T. Synthetic lethal targeting of TET2-mutant hematopoietic stem and progenitor cells (HSPCs) with TOP1-targeted drugs and PARP1 inhibitors. *Leukemia*. 2020;34(11):2992-3006. doi 10.1038/s41375-020-0927-5

Johannes J.W., Balazs A.Y.S., Barratt D., Bista M., Chuba M.D., Cosulich S., Critchlow S.E., … Xue L., Yao T., Zhang K., Zhang A.X., Zheng X. Discovery of 6-Fluoro-5-{4-[(5-fluoro-2-methyl-3-oxo-3,4-dihydroquinoxalin-6-yl)methyl]piperazin-1-yl}-*N*-methylpyridine-2-carboxamide (AZD9574): a CNS-penetrant, PARP1-selective inhibitor. *J Med Chem*. 2024;67(24):21717-21728. doi 10.1021/acs.jmedchem.4c01725

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., … Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi 10.1038/s41586-021-03819-2

Jurrus E., Engel D., Star K., Monson K., Brandi J., Felberg L.E., Brookes D.H., … Krasny R., Wei G., Holst M.J., McCammon J.A., Baker N.A. Improvements to the APBS biomolecular solvation software suite. *Protein Sci*. 2018;27(1):112-128. doi 10.1002/pro.3280

Kawale A.S., Povirk L.F. Tyrosyl-DNA phosphodiesterases: rescuing the genome from the risks of relaxation. *Nucleic Acids Res*. 2018; 46(2):520-537. doi 10.1093/nar/gkx1219

Kim D.-S., Camacho C.V., Kraus W.L. Alternate therapeutic pathways for PARP inhibitors and potential mechanisms of resistance. *Exp Mol Med*. 2021;53(1):42-51. doi 10.1038/s12276-021-00557-3

Kim J.W., Min A., Im S.-A., Jang H., Kim Y.J., Kim H.-J., Lee K.-H., Kim T.-Y., Lee K.W., Oh D.-Y., Kim J.-H., Bang Y.-J. TDP1 and TOP1 modulation in olaparib-resistant cancer determines the efficacy of subsequent chemotherapy. *Cancers*. 2020;12(2):334. doi 10.3390/cancers12020334

Kornienko T.E., Chepanova A.A., Zakharenko A.L., Filimonov A.S., Luzina O.A., Dyrkheeva N.S., Nikolin V.P., Popova N.A., Salakhutdinov N.F., Lavrik O.I. Enhancement of the antitumor and antimetastatic effect of topotecan and normalization of blood counts in mice with Lewis carcinoma by Tdp1 inhibitors – new usnic acid derivatives. *Int J Mol Sci*. 2024;25(2):1210. doi 10.3390/ijms25021210

Lavrik O.I. PARPs' impact on base excision DNA repair. *DNA Repair*. 2020;93:102911. doi 10.1016/j.dnarep.2020.102911

Lebedeva N.A., Anarbaev R.O., Sukhanova M., Vasil'eva I.A., Rechkunova N.I., Lavrik O.I. Poly(ADP-ribose)polymerase 1 stimulates the AP-site cleavage activity of tyrosyl-DNA phosphodiesterase 1. *Biosci Rep*. 2015;35(4):e00230. doi 10.1042/BSR20140192

Marchand C., Abdelmalak M., Kankanala J., Huang S.-Y., Kiselev E., Fesen K., Kurahashi K., Sasanuma H., Takeda S., Aihara H., Wang Z., Pommier Y. Deazaflavin inhibitors of tyrosyl-DNA phosphodiesterase 2 (TDP2) specific for the human enzyme and active against cellular TDP2. *ACS Chem Biol*. 2016;11(7):1925-1933. doi 10.1021/acschembio.5b01047

Matsuno Y., Hyodo M., Fujimori H., Shimizu A., Yoshioka K. Sensitization of cancer cells to radiation and topoisomerase I inhibitor camptothecin using inhibitors of PARP and other signaling molecules. *Cancers*. 2018;10(10):364. doi 10.3390/cancers10100364

Mirdita M., Schütze K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi 10.1038/s41592-022-01488-1

Moor N.A., Vasil'eva I.A., Anarbaev R.O., Antson A.A., Lavrik O.I. Quantitative characterization of protein-protein complexes involved in base excision DNA repair. *Nucleic Acids Res*. 2015;43(12):6009-6022. doi 10.1093/nar/gkv569

Moretti R., Bender B.J., Allison B., Meiler J. Rosetta and the design of ligand binding sites. In: Stoddard B. (Ed.) Computational Design of Ligand Binding Proteins. Methods in Molecular Biology. Vol. 1414. New York: Humana Press, 2016;47-62. doi 10.1007/978-1-4939-3569-7_4

Murai J., Marchand C., Shahane S.A., Sun H., Huang R., Zhang Y., Chergui A., Ji J., Doroshow J.H., Jadhav A., Takeda S., Xia M., Pommier Y. Identification of novel PARP inhibitors using a cell-based TDP1 inhibitory assay in a quantitative high-throughput screening platform. *DNA Repair*. 2014;21:177-182. doi 10.1016/j.dnarep.2014.03.006

Neudachina L., Lakiza N. Physico-Chemical Principles of the Use of Coordination Compounds. Ekaterinburg, 2014 (in Russian)

Nguyen T.X., Abdelmalak M., Marchand C., Agama K., Pommier Y., Cushman M. Synthesis and biological evaluation of nitrated 7-, 8-, 9-, and 10-hydroxyindenoisoquinolines as potential dual topoisomerase I (Top1)–tyrosyl-DNA phosphodiesterase I (TDP1) inhibitors. *J Med Chem*. 2015;58(7):3188-3208. doi 10.1021/acs.jmedchem.5b00136

O'Boyle N.M., Banck M., James C.A., Morley C., Vandermeersch T., Hutchison G.R. Open Babel: an open chemical toolbox. *J Cheminform*. 2011;3(1):33. doi 10.1186/1758-2946-3-33

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi 10.1002/jcc.20084

Pommier Y., Leo E., Zhang H., Marchand C. DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem Biol*. 2010;17(5):421-433. doi 10.1016/j.chembiol.2010.04.012

Pommier Y., Huang S.N., Gao R., Das B.B., Murai J., Marchand C. Tyrosyl-DNA-phosphodiesterases (TDP1 and TDP2). *DNA Repair*. 2014;19:114-129. doi 10.1016/j.dnarep.2014.03.020

Quiroga R., Villarreal M.A. Vinardo: a scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS One*. 2016;11(5):e0155183. doi 10.1371/journal.pone.0155183

Ray Chaudhuri A., Nussenzweig A. The multifaceted roles of PARP1 in DNA repair and chromatin remodelling. *Nat Rev Mol Cell Biol*. 2017;18(10):610-621. doi 10.1038/nrm.2017.53

Salentin S., Schreiber S., Haupt V.J., Adasme M.F., Schroeder M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res*. 2015;43(W1):W443-W447. doi 10.1093/nar/gkv315

Schreiber V., Dantzer F., Ame J.-C., de Murcia G. Poly(ADP-ribose): novel functions for an old molecule. *Nat Rev Mol Cell Biol*. 2006;7(7):517-528. doi 10.1038/nrm1963

Sherstyuk Y.V., Ivanisenko N.V., Zakharenko A.L., Sukhanova M.V., Peshkov R.Y., Eltsov I.V., Kutuzov M.M., Kurgina T.A., Belousova E.A., Ivanisenko V.A., Lavrik O.I., Silnikov V.N., Abramova T.V. Design, synthesis and molecular modeling study of conjugates of ADP and morpholino nucleosides as a novel class of inhibitors of PARP-1, PARP-2 and PARP-3. *Int J Mol Sci*. 2019;21(1):214. doi 10.3390/ijms21010214

Smith L.M., Willmore E., Austin C.A., Curtin N.J. The novel poly(ADP-ribose) polymerase inhibitor, AG14361, sensitizes cells to topoisomerase I poisons by increasing the persistence of DNA strand breaks. *Clin Cancer Res*. 2005;11(23):8449-8457. doi 10.1158/1078-0432.ccr-05-1224

Sukhanova M.V., Khodyreva S.N., Lavrik O.I. Poly(ADP-ribose) polymerase-1 inhibits strand-displacement synthesis of DNA catalyzed by DNA polymerase β. *Biochemistry* (*Moscow*). 2004;69(5):558-568. doi 10.1023/b:biry.0000029855.68502.fa

Susova O.Y., Karshieva S.S., Kostyukov A.A., Moiseeva N.I., Zaytseva E.A., Kalabina K.V., Zusinaite E., Gildemann K., Smirnov N.M., Arutyunyan A.F., Zhuze A.L. Dimeric bis-benzimidazole-pyrroles $DB_2Py(n)$ – AT-site-specific ligands: synthesis, physicochemical analysis, and biological activity. *Acta Naturae*. 2024;16(1):86-100. doi 10.32607/actanaturae.27327

Szanto M., Yelamos J., Bai P. Specific and shared biological functions of PARP2 – is PARP2 really a lil' brother of PARP1? *Expert Rev Mol Med*. 2024;26:e13. doi 10.1017/erm.2024.14

Wang P., Elsayed M.S.A., Plescia C.B., Ravji A., Redon C.E., Kiselev E., Marchand C., Zeleznik O., Agama K., Pommier Y., Cushman M. Synthesis and biological evaluation of the first triple inhibitors of human topoisomerase 1, tyrosyl-DNA phosphodiesterase 1 (Tdp1), and tyrosyl-DNA phosphodiesterase 2 (Tdp2). *J Med Chem*. 2017;60(8):3275-3288. doi 10.1021/acs.jmedchem.6b01565

Wang R., Lai L., Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*. 2002;16(1):11-26. doi 10.1023/a:1016357811882

Yang H., Zhu X.-Q., Wang W., Chen Y., Hu Z., Zhang Y., Hu D.-X., Yu L.-M., Agama K., Pommier Y., An L.-K. The synthesis of furoquinolinedione and isoxazoloquinolinedione derivatives as selective Tyrosyl-DNA phosphodiesterase 2 (TDP2) inhibitors. *Bioorg Chem*. 2021;111:104881. doi 10.1016/j.bioorg.2021.104881

Yang H., Qin C., Wu M., Wang F., Wang W., Agama K., Pommier Y., Hu D., An L. Synthesis and biological activities of 11- and 12-substituted benzophenanthridinone derivatives as DNA topoisomerase IB and tyrosyl-DNA phosphodiesterase 1 inhibitors. *ChemMedChem*. 2023;18(10):e202200593. doi 10.1002/cmdc.202200593

Zakharenko A., Khomenko T., Zhukova S., Koval O., Zakharova O., Anarbaev R., Lebedeva N., Korchagina D., Komarova N., Vasiliev V., Reynisson J., Volcho K., Salakhutdinov N., Lavrik O. Synthesis and biological evaluation of novel tyrosyl-DNA phosphodiesterase 1 inhibitors with a benzopentathiepine moiety. *Bioorg Med Chem*. 2015;23(9):2044-2052. doi 10.1016/j.bmc.2015.03.020

Zakharenko A.L., Luzina O.A., Chepanova A.A., Dyrkheeva N.S., Salakhutdinov N.F., Lavrik O.I. Natural products and their derivatives as inhibitors of the DNA repair enzyme tyrosyl-DNA phos-

phodiesterase 1. *Int J Mol Sci*. 2023;24(6):5781. doi 10.3390/ijms24065781

Zeng Z., Sharma A., Ju L., Murai J., Umans L., Vermeire L., Pommier Y., Takeda S., Huylebroeck D., Caldecott K.W., El-Khamisy S.F. TDP2 promotes repair of topoisomerase I-mediated DNA damage in the absence of TDP1. *Nucleic Acids Res*. 2012;40(17):8371-8380. doi 10.1093/nar/gks622

Zhang M., Wang Z., Su Y., Yan W., Ouyang Y., Fan Y., Huang Y., Yang H. TDP1 represents a promising therapeutic target for overcoming tumor resistance to chemotherapeutic agents: progress and potential. *Bioorg Chem*. 2025;154:108072. doi 10.1016/j.bioorg.2024.108072

Zhang X.-R., Wang H.-W., Tang W.-L., Zhang Y., Yang H., Hu D.-X., Ravji A., Marchand C., Kiselev E., Ofori-Atta K., Agama K., Pommier Y., An L.-K. Discovery, synthesis, and evaluation of oxynitidine derivatives as dual inhibitors of DNA topoisomerase IB (TOP1) and tyrosyl-DNA phosphodiesterase 1 (TDP1), and potential antitumor agents. *J Med Chem*. 2018;61(22):9908-9930. doi 10.1021/acs.jmedchem.8b00639

Zhang Y., He X., Yang H., Liu H., An L. Robustadial A and B from *Eucalyptus globulus* Labill. and their anticancer activity as selective tyrosyl-DNA phosphodiesterase 2 inhibitors. *Phytotherapy Res*. 2021;35(9):5282-5289. doi 10.1002/ptr.7207

Conflict of interest. The authors declare no conflict of interest.

Received August 7, 2025. Revised September 30, 2025. Accepted September 30, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1109-1121

doi 10.18699/vjgb-25-115

Computer modeling of spatial dynamics and primary genetic divergence for a population system in a ring areal

M.P. Kulakov [1], O.L. Zhdanova [2], E.Ya. Frisman [1]

[1] Institute for Complex Analysis of Regional Problems of the Far Eastern Branch of the Russian Academy of Sciences, Birobidzhan, Russia
[2] Institute of Automation and Control Processes of the Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia
k_matvey@mail.ru

Abstract. One of the main goals of modern evolutionary biology is to understand the mechanisms that lead to the initial differentiation (primary divergence) of populations into groups with genetic traits. This divergence requires reproductive isolation, which prevents or hinders contact and the exchange of genetic material between populations. This study explores the potential for isolation based not on obvious geographical barriers, population distance, or ecological specialization, but rather on hereditary mechanisms, such as gene drift and flow and selection against heterozygous individuals. To this end, we propose and investigate a dynamic discrete-time model that describes the dynamics of frequencies and numbers in a system of limited populations coupled by migrations. We consider a panmictic population with Mendelian inheritance rules, one-locus selection, and density-dependent factors limiting population growth. Individuals freely mate and randomly move around a one-dimensional ring-shaped habitat. The model was verified using data from an experiment on the box population system of *Drosophila melanogaster* performed by Yu.P. Altukhov et al. With rather simple assumptions, the model explains some mechanisms for the emergence and preservation of significant genetic differences between subpopulations (primary genetic divergence), accompanied by heterogeneity in allele frequencies and abundances within a homogeneous area. In this scenario, several large groups of genetically homogeneous subpopulations form and independently develop. Hybridization occurs at contact sites, and polymorphism is maintained through migration from genetically homogeneous nearby sites. It was found that only disruptive selection, directed against heterozygous individuals, can sustainably maintain such a spatial distribution. Under directional selection, divergence may occur for a short time as part of the transitional evolutionary process towards the best-adapted genotype. Because of the reduced adaptability of heterozygous (hybrid) individuals and low growth rates in these sites (hybrid zones), gene flow between adjacent sites with opposite genotypes (phenotypes) is significantly impeded. As a result, the hybrid zones can become effective geographical barriers that prevent the genetic flow between coupled subpopulations.

Key words: metapopulation; migration; spatiotemporal dynamics; mathematical modeling; genetic divergence; gene flow; hybrid zones; isolation

For citation: Kulakov M.P., Zhdanova O.L., Frisman E.Ya. Computer modeling of spatial dynamics and primary genetic divergence for a population system in a ring areal. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1109-1121. doi 10.18699/vjgb-25-115

Funding. This work was carried out within the framework of the state task of the Institute for Complex Analysis of Regional Problems of the Far Eastern Branch of the Russian Academy of Sciences.

Компьютерное моделирование пространственной динамики и первичной генетической дивергенции в системе популяций на кольцевом ареале

М.П. Кулаков [1], О.Л. Жданова [2], Е.Я. Фрисман [1]

[1] Институт комплексного анализа региональных проблем Дальневосточного отделения Российской академии наук, Биробиджан, Россия
[2] Институт автоматики и процессов управления Дальневосточного отделения Российской академии наук, Владивосток, Россия
k_matvey@mail.ru

Аннотация. Одна из ключевых задач современной эволюционной биологии – изучение процессов, приводящих к первичному разделению (дивергенции) популяций на различающиеся генотипами группы особей. Для дивергенции очевидно необходима репродуктивная изоляция, которая делает невозможным контакт особей или существенно затрудняет обмен генетической информацией между популяциями. Настоящее исследование изучает возможность изоляции, в основе которой лежат не очевидные географические барьеры, удаленность

популяций или экологическая специализация, а лишь наследственные механизмы, дрейф и поток генов, а также отбор против гетерозигот. Для этого предложена и исследована динамическая модель с дискретным временем, которая описывает динамику частот и численностей в системе миграционно связанных лимитированных популяций. Рассматривается панмиктичная популяция с менделевскими правилами наследования, монолокусным отбором, действием плотностно-зависимых факторов, лимитирующих рост численности. Особи свободно скрещиваются и перемещаются вдоль одномерного кольцевого ареала. Модель верифицирована с использованием данных эксперимента над ящичной системой популяций дрозофил *Drosophila melanogaster*, проведенного под руководством Ю.П. Алтухова. При достаточно простых предположениях модель описывает некоторые механизмы возникновения и сохранения на однородном ареале существенных генетических различий (первичной генетической дивергенции), сопровождаемых неоднородностью в частотах аллелей и численностях. В этом случае формируется несколько больших групп генетически однородных субпопуляций, развивающихся независимо. В местах их контакта активно идет гибридизация, а полиморфизм сохраняется за счет миграции с сопредельных однородных участков. Обнаружено, что устойчиво поддерживать такое пространственное распределение может только дизруптивный (разрывающий) отбор, направленный против гетерозигот. При движущем отборе дивергенция существует непродолжительное время, как часть переходного процесса. За счет пониженной приспособленности гетерозигот (гибридов) и низкой скорости роста на этих участках (зонах гибридизации) существенно затрудняется обмен генами между смежными участками с противоположными гомозиготными генотипами (фенотипами). В результате зоны гибридизации выполняют функцию географического барьера, который фактически останавливает обмен генов между разными группами в случае смежной симпатрии.

Ключевые слова: метапопуляция; миграция; пространственно-временная динамика; математическое моделирование; генетическая дивергенция; поток генов; гибридные зоны; изоляция

Introduction

Genetic divergence cannot occur without effective mechanisms of reproductive isolation and stopping the gene flow between populations. This can be caused by large distances between populations (allopatry), which cannot be overcome during the lifetime of individuals, or by geographical barriers that prevent the transfer of genes. However, even if populations of the same species live in the same or adjacent areas (sympatry or parapatry) they can differ significantly in their traits. Although individuals from these populations can interact and produce viable, fertile hybrids, there is no blurring of parental traits. Several mechanisms support the reproductive isolation and the divergence between different forms, including selection against hybrids, which often have lower fitness than parental populations.

There are sufficient examples of reproductive isolation, where different subpopulations have accumulated sufficient differences even when they live sympatrically and have developed effective measures to prevent hybridization. For instance, recognition signals related to phonetic features and used in mating behavior contribute to the stabilization of extreme forms of a characteristic. Thus, the mating calls of certain frog species (such as *Microhyla carolinensis* and *M. olivacea*, *Litoria verreauxii* and *L. v. alpina*) differ greatly in the contact zone where their ranges overlap, but do not differ significantly in areas where they do not occur together (Blair, 1955a; Littlejohn, 1965; Smith et al., 2003). In addition, the body sizes of different frog forms differ greatly in the contact zone, which complicates the mating process (Blair, 1955b).

Prezygotic isolation of sympatric forms of the same species or subspecies is often followed by ecological specialization, which does not prevent copulatory behavior between individuals with different traits and their hybridization, but only makes it unlikely. For example, the periods of sexual activity for two species of *Rhagoletis pomonella* are determined by the time of fruiting of the trees they were born on and lay their eggs on – hawthorn and apple (Filchak et al., 2000). These two races of flies of *R. pomonella* differ in their sensory processing of key fruit odors: while some individuals are attracted to apple and avoid hawthorns, others choose hawthorn and avoid apples, which significantly hinders their contact (Tait et al., 2021). The mating preferences of hybrids are not entirely clear. However, when two races of *R. pomonella* are interbred in the laboratory, a lower conception rate is recorded (Yee, Goughnour, 2011), which signals some selection against hybrids and persistent divergence in nature caused by specialization of flies.

There are a few examples of hybridization where it does not have obvious negative effects, such as reduced fitness or a catastrophic decline in the reproductive success of hybrids (heterozygotes). For example, intraspecific variability in some birds is often expressed as differences in plumage coloration. At the same time, there is a clear divergence in traits between different parts of a large range, and stable hybrid zones exist over long periods of time in areas where the ranges overlap. The populations of the carrion crow and hooded crow (*Corvus corone* and *C. cornix*) are well known in Siberian (between the Ob and Yenisei rivers) and European hybrid zones (Haring et al., 2012; Poelstra et al., 2014; Kryukov, 2019; Blinov, Zheleznova, 2020), or northern flicker hybrid zone (*Colaptes auratus cafer* and *C. a. auratus*) in USA (Aguillon, Rohwer, 2022). Another example is the hybridization of the great tit (*Parus major*) and Japanese tit (*P. minor*) in the Amur region (Kapitonova et al., 2012).

A genetic mechanism supporting isolation based on innate mating preferences has been identified in crows: they prefer to choose partners who are similar to themselves rather than exotic individuals. The process of forming phenotypes in carrion and hooded crows is linked to chromosomal inversion, which affects both feather coloration and the visual perception

of feather colors, as well as certain aspects of reproductive behavior (Poelstra et al., 2014). However, in areas where hybridization occurs, which apparently arises simultaneously with different colorations, mating preferences turn out to be more diverse and complete isolation does not occur. This is because the inverted chromosome region of the hooded crow is inherited in its entirety and does not recombine with the homologous regions of the carrion crow.

One simple model for studying genetic divergence is a linear chain or ring of partially isolated subpopulations that exchange genes. The studies on such models show that gene flow between subpopulations coupled by migration can lead to stable geographic variability of a trait and the maintenance of hybrid zones only with disruptive selection. With directional selection, stable divergence is impossible and can only occur as part of a transition process under special initial conditions (Bazykin, 1972; Frisman, 1986; Yeaman, Otto, 2011; Láruson, Reed, 2016). For chains of connected populations with different topologies, it has been found that divergence occurs more often in linear chains and rings, and less often in fully connected networks (with global connectivity) (Láruson, Reed, 2016; Sundqvist et al., 2016).

At the same time, for many natural populations with significant divergence in characteristics and sometimes with known isolating mechanisms, it can be difficult to identify a specific adaptive trait that disruptive selection acts upon. This may be due to hidden traits, such as innate immune factors or the major histocompatibility complex, which are not directly related to an external trait that we currently observe in individuals, such as feather coloration in birds, skin or coat patterns, beak shape and size, or behavioral characteristics. The observed spatial distribution of a trait does not directly indicate the causes or type of selection that led to this divergence in the past. However, it can be successfully linked to the observed trait and serve as an indicator or marker of fitness, particularly for species with wide ranges, heterogeneous environmental conditions, significant divergence, and a high degree of polymorphism (Orsini et al., 2008; Murphy et al., 2010).

This work is part of a series of studies investigating the basic mechanisms of primary genetic divergence in systems of panmictic populations of diploid organisms coupled by migration and selection directed against heterozygotes (Zhdanova, Frisman, 2023; Kulakov, Frisman, 2025). We propose a dynamic discrete-time model that takes into account the action of density-dependent factors limiting population growth, genetic drift (through certain perturbations of initial conditions), natural selection, and migration of individuals between adjacent sites. The model is verified based on data from laboratory experiments with box populations of Drosophila (*Drosophila melanogaster*) conducted under the supervision of Yu.P. Altukhov, which showed significant divergence in allele structure at the α-glycerophosphate dehydrogenase (α-Gdph) locus between groups of adjacent boxes (Altukhov et al., 1979; Altukhov, Bernashevskaya, 1981; Altukhov, 2003).

In this article, we analyze the processes of selection and migration (gene flow) that form and maintain the heterogeneous spatial distribution of allele frequencies, based on multiple computer simulations of a model. We investigate the role of hybrid zones with high proportions of heterozygous individuals in the *α-Gdph* gene and demonstrate that these zones separate monomorphic groups of boxes apart and do not allow the most adapted genotype to spread throughout the entire ring area.

Material and methods

The study is based on an original mathematical model – a system of coupled nonlinear maps (discrete-time equations) that describes the dynamics of genotype frequencies and subpopulation abundances. The migration of individuals and gene flow between subpopulations are described using a migration matrix with random coefficients. We use the MT19937 random number generator (Matsumoto et al., 1998), available in the GSL numerical computation library. This generator has an extremely long period ($\sim 10^{6,000}$) and low correlation, passing most statistical tests for randomness in its pseudo-random number sequences.

To validate the model, we use data from an experiment on the *D. melanogaster* ring system, conducted by a team led by Yu.P. Altukhov. The data consist of allele frequencies at the locus encoding the α-Gdph enzyme, as well as the numbers of flies in each box at different stages of the experiment (Altukhov, 2003). We estimate model parameters using the least squares method.

Numerical experiments are conducted with the author's software package, including the computer implementation of a mathematical model, visualization of the results, and analysis of dynamic regimes.

Model of local population

We consider a population of diploid organisms where between two adjacent generations, the following sequence of elementary population processes occurs: zygote formation from gametes, natural selection on zygotes (individuals), migration (dispersal) between adjacent subpopulations, and production of new gametes. We focus on populations in which the adaptive diversity is determined by a single locus with two alleles (*A* and *a*), which are inherited co-dominantly. The phenotype of individuals is strictly determined by their genotype. The population is panmictic, and Mendelian inheritance rules apply. This means that the population contains individuals with genotypes *AA*, *Aa*, and *aa*. At time *t*, these genotypes have abundances $N_1(t)$, $N_2(t)$, and $N_3(t)$, respectively, and frequencies $q_1(t) = N_1(t)/N(t)$, $q_2(t) = N_2(t)/N(t)$, and $q_3(t) = N_3(t)/N(t)$ (where $N(t) = N_1(t) + N_2(t) + N_3(t)$ is the total population size).

Let us assume that the genotypes differ in their reproductive abilities, which is expressed by differences in gamete production rates or individual survival rates. Denote the intensity of gamete production for individuals with genotypes *AA*, *Aa*, and *aa* as g_{AA}, g_{Aa} and g_{aa}, respectively, taking into account the death of some gametes before they combine into zygotes in the next generation. Additionally, let W_{AA}, W_{Aa} and W_{aa} represent the proportion of zygotes (or individuals) with the corresponding genotype that survive the natural selection and have the ability to migrate (disperse).

In cases where gamete production intensity does not depend on parental genotypes, i. e., $g_{AA} = g_{Aa} = g_{aa} = g$, the equations for genotype frequencies in a local panmictic population can be expressed as:

$$\begin{cases} q_1(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_1) q_1^*(t), \\ q_2(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_2) q_2^*(t), \\ q_3(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_3) q_3^*(t), \end{cases} \quad (1)$$

where $q_1^*(t) = (q_1(t) + q_2(t)/2)^2$,

$q_2^*(t) = 2(q_1(t) + q_2(t)/2)(q_3(t) + q_2(t)/2)$,

$q_3^*(t) = (q_3(t) + q_2(t)/2)^2$ are the genotype frequencies immediately after gametes combine into zygotes, but before selection and migration of individuals (Zhdanova, Frisman, 2023; Kulakov, Frisman, 2025). The parameter s_k is the selection coefficient for zygotes with the corresponding genotype, which links the fitness W_k of each genotype and the gamete production rate g_k as follows: $1 + s_k = gW_k$ (k =*AA*, *Aa*, *aa*). In system (1), the normalization factor

$$\overline{W}(t) = 1 + s_1 q_1^*(t) + s_2 q_2^*(t) + s_3 q_3^*(t) \quad (2)$$

is equal to the average (generalized) fitness, and its value determines the population growth rate. If there are no factors limiting the growth, the population size changes according to the following equation:

$$N(t+1) = \overline{W}(t) N(t). \quad (3)$$

The number of individuals with each genotype is determined by ratios: $N_k(t+1) = q_k(t+1)N(t+1) = (1+s_k)q_k^*(t)N(t+1)$ (k=*AA*, *Aa*, *aa*).

Of all the types of genetic selection determined by values s_1, s_2, and s_3, disruptive selection is the most interesting ($s_2 < s_1$ and $s_2 < s_3$), as system (1) demonstrates bistability. Early studies show that this type of selection is responsible for the emergence and fixation of genetic differences in different parts of a homogeneous area, even when environmental and other factors are not considered.

At the same time, on a large temporal scale, the growth of actual evolving populations is limited by environmental factors. This growth limitation can be described by a nonlinear dependence of selection and gamete production parameters on the abundance of genotypes or the total population density in model (1)–(3). It is easy to show that if the rates of gamete production are equal for all genotypes, then there is no difference between the limiting gamete production rate (g) and the intensity of selection (W_{ij}) in case of competition for a common resource. Therefore, without loss of generality, we can assume that:

$$W_{ij} = w_{ij} F(N), \quad (4)$$

where w_{ij} is the maximum proportion of individuals with genotype ij (*AA*, *Aa*, or *aa*) that survive after natural selection under minimal competition (at low density), F is the function that describes the effect of density-dependent growth limitation, and N is the total population size. Considering (4), the frequency dynamics equations (1) will not change their form, except for replacing W_{ij} with w_{ij} and gW_{ij} with $1+s_k$, while the population equations (3) will have a nonlinear dependency on density:

$$N(t+1) = \overline{W}(t) N(t) F(N(t)). \quad (5)$$

In populations of diploid organisms, exchange of gametes often requires contact between individuals. The probability of this decreases significantly at low densities, i. e., there is a direct correlation between the average individual fitness and the population density – the Allee effect (Allee, 1958). As a result, when the population size falls below a certain critical value N_0, population growth becomes impossible and effective natural selection ceases to operate. Instead, only genetic drift determines the evolutionary trajectory of the population. Therefore, to describe these density-dependent limiting factors, we can use a function of the following form:

$$F(N) = a\varphi(N)\exp(-N/K), \quad (6)$$

where $\varphi(N)$ is a sigmoid function equal to:

$$\varphi(N) = \frac{1}{1 + e^{-h(N-N_0)}}, \quad (7)$$

with parameter $h \geq 2$, which defines the slope angle of the sigmoid at point N_0. The value of N_0 determines the minimum population size required for simple reproduction (1:1). The parameter K defines the ecological capacity of the habitat, and a defines the average number of offspring per individual with an average fitness of 1. These two parameters determine the steady-state (equilibrium) population size $\overline{N} \approx K \ln(a\overline{W})$. Using (7), we can rewrite the equation (5) for population dynamics as follows:

$$N(t+1) = rN(t)\varphi(N(t))\exp(-N(t)/K), \quad (8)$$

where $r = a\overline{W}(t)$ is the total reproductive capacity of all genotypes.

When $r > 1$, equation (8) has three fixed points [$N(t+1) = N(t)$]: 0, N_0 and $\overline{N} \approx K \ln(a\overline{W})$. If $N < N_0$, the number of surviving offspring $N(t+1)$ is less than the number of their ancestors $N(t)$, and the population inevitably declines, which corresponds to a strong Allee effect. If $N_0 < N < \overline{N}$ and $r > 1$, there are enough breeders and the population size increases. With $N > \overline{N}$, the population size exceeds the carrying capacity of the habitat, and the population abundance falls to a steady-state of $\overline{N}$.

Let us now consider populations that are coupled by migration and evolve in the way described above.

Dynamic model with gene flow

One method for studying the dynamics and evolution of dispersed population systems (metapopulations) is to conduct laboratory experiments using populations in boxes that are connected by narrow corridors. In these experiments, environmental conditions, growth parameters, selection, and migration can be carefully controlled. Typically, the connected boxes (chambers) form closed chains of subpopulations that exchange a small number of individuals (Fig. 1*a*). These population systems are often constructed in laboratory settings, for example, for *D. melanogaster* (Altukhov et al., 1979; Altukhov, Bernashevskaya, 1981; Dey, Joshi, 2006), or *Escherichia coli* (Keymer et al., 2006).

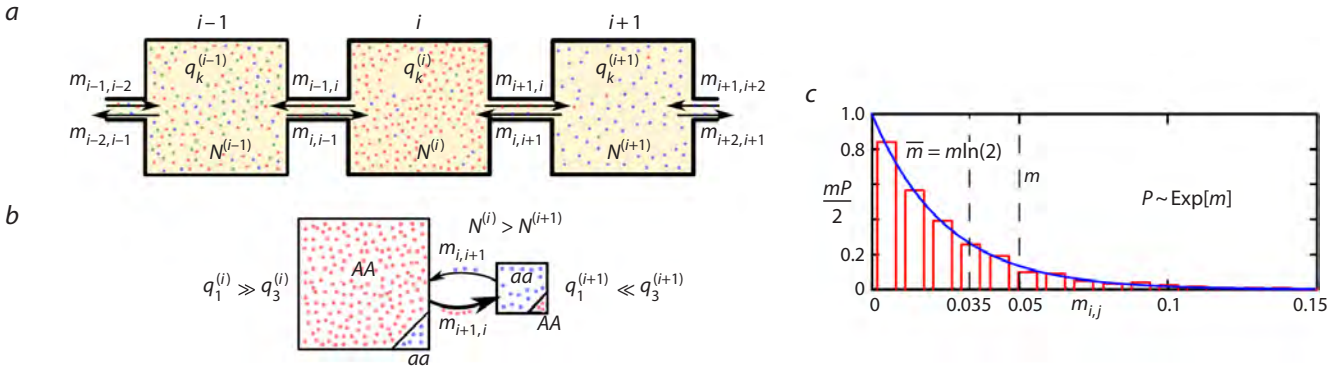


Fig. 1. *a*, Scheme of the population system – boxes coupled by narrow migration corridors. *b*, Illustration showing that gene flow between populations of different sizes can significantly change the genotype in a small population, but has no effect on a large population. *c*, The probability density of an exponentially distributed random value of the migration coefficient $m_{i,j}$.

Consider a system of n boxes, or subpopulations, and each box is numbered from 1 to n (Fig. 1*a*). Let $0 \le m_{i,j} < 1$ denote the proportion of individuals from the total population size that move from box j to box i ($m_{i,j}$ is the migration coefficient). The emigrants consist of individuals with three studied genotypes, so it is true that $m_{i,j}N^{(j)} = m_{i,j}q_{AA}^{(j)}N^{(j)} + m_{i,j}q_{Aa}^{(j)}N^{(j)} + m_{i,j}q_{aa}^{(j)}N^{(j)}$.

Then, for a system of subpopulations coupled by migration, the equations for frequency dynamics (1) and abundance dynamics (8) take the following forms:

$$\begin{cases} q_k^{(i)}(t+1) = \dfrac{1+s_k}{G^{(i)}(t)}\Big(q_k^{(i)^*}(t)N^{(i)}(t)\big(1-m_{i-1,i}-m_{i+1,i}\big)+ \\ \qquad + q_k^{(i-1)^*}(t)N^{(i-1)}(t)m_{i,i-1} + q_k^{(i+1)^*}(t)N^{(i+1)}(t)m_{i,i+1}\Big), \\ N^{(i)}(t+1) = N^{(i)^*}(t+1)\big(1-m_{i-1,i}-m_{i+1,i}\big)+ \\ \qquad + N^{(i-1)^*}(t+1)m_{i,i-1} + N^{(i+1)^*}(t+1)m_{i,i+1}, \end{cases} \quad (9)$$

where k = 1, 2, 3 are the numbers of the groups of individuals with the genotypes *AA*, *Aa*, and *aa*, respectively, $q_k^{(i)^*}$ are the frequencies before migration, and $N^{(i)^*}(t+1) = a\overline{W}^{(i)}(t)N^{(i)}(t)F\big(N^{(i)}(t)\big)$ is the abundance of the ith subpopulation after selection but before migration. The normalization coefficient G is equal to:

$$G^{(i)}(t) = \overline{W}^{(i)}(t)\big(1-m_{i-1,i}-m_{i+1,i}\big)N^{(i)}(t) + \\ + \overline{W}^{(i-1)}(t)m_{i,i-1}N^{(i-1)}(t) + \overline{W}^{(i+1)}(t)m_{i,i+1}N^{(i+1)}(t), \quad (10)$$

where $\overline{W}^{(i)}(t) = 1 + s_1q_1^{(i)^*}(t) + s_2q_2^{(i)^*}(t) + s_3q_3^{(i)^*}(t)$. To close the chain of subpopulations into a ring, we assume that the 1st box is connected to the 2nd and nth, the nth to the (n–1)th and 1st, i. e., the following mapping applies to the site number: $i \to i \bmod n$. In system (9), the factor $(1 - m_{i-1,i} - m_{i+1,i})$ is the proportion of individuals that stayed in the ith box after migrating to the two neighboring boxes; $m_{i,i-1}$ and $m_{i,i+1}$ are the proportions of individuals from (i–1) and (i+1)-subpopulations that migrated to the ith box.

Equations (9) demonstrate that the intensity of gene flow from each subpopulation is not only dependent on the frequencies of genotypes within the native site, as was the case for the local population, but also on the absolute number of individuals. This is clearly evident from the assumption that migrants consist of individuals with all three possible genotypes. Therefore, the flow of migrants from a small population consisting, for example, solely of *aa* homozygotes, has a minimal impact on a larger population consisting mainly of *AA* homozygotes (Fig. 1*b*). Conversely, the flow from a larger population can quickly change the frequencies even at a low migration rate. Note that, in some cases, this mechanism clearly violates the assumption of panmixia at the scale of the entire metapopulation, as changes in the frequency of non-comparable subpopulations are determined more by the genetic structure of immigrants than by random mating, genetic drift, or natural selection.

The flow of genes and individuals between subpopulations can be either completely deterministic or random. In the first case, the number and genetic structure of migrants depend on factors such as population density at the source and sink sites, or external environmental factors like food (taxis) and energy flows (phototaxis). In the second case, both the direction and proportion of migrants vary randomly from generation to generation, without any clear pattern.

Below, we will only consider random migration. To describe this, we do the following. For each season number t, we randomly select two migration coefficients $m_{i-1,i}$ and $m_{i+1,i}$, which are equal to the proportions of individuals that leave the ith site and migrate to adjacent sites. We ignore the possibility of more distant dispersal. Each pair of values $m_{i-1,i}$ and $m_{i+1,i}$ will be generated independently using an exponentially distributed random variable generator with an expected value of $m/2$ and a median of $m\ln(2)$.

Figure 1*c* shows a histogram of the distribution of 200 replicates, each consisting of 30 pairs of independent random values for migration coefficients (n = 30 and m = 0.05), along with the graph of the theoretical probability density function. Both curves are scaled to the same distribution parameter $\lambda = 2m^{-1}$. This value corresponds to a situation where approximately half of all migration coefficients are less than or equal to $m\ln(2) \approx 0.035$, and their average is $\overline{m} = m/2 = 0.025$.

Next, we consider the dynamic regimes in the system (9)–(10) with random migration, using parameter values obtained from experimental data.

Model verification

There are two ways to verify the model and search for conditions of primary genetic divergence. First, we can perform a series of simulations to ensure that the system (9) generates regimes corresponding to genetic divergence with only reduced heterozygote fitness. Secondly, we need to compare the results of simulations with the empirical data. However, this can be challenging, as despite all the available research and data, most natural populations with clear divergence in traits across space are initially highly heterogeneous.

The ideal solution may involve using data from a carefully designed animal experiment. In the mentioned experiment, conducted under the supervision of Yu.P. Altukhov, evolutionary processes were studied in a system consisting of 30 boxes connected by narrow tubes and inhabited by *D. melanogaster* flies (Altukhov et al., 1979; Altukhov, Bernashevskaya, 1981). The randomness of migration was provided by uniform environmental conditions (lighting and food) and random rotation of the ring system of connected boxes. During the experiment, the spatial distribution and abundance dynamics, as well as the frequency of alleles at the autosomal esterase-6 (*Est-6*) and α-glycerophosphate dehydrogenase (*α-Gdph*) loci, were analyzed. By the 60th generation, a clear and stable differentiation of allele distribution at the *α-Gdph* locus formed between groups of adjacent boxes.

Some parameters are immediately known from the description of the original experiment, such as the migration coefficient ($m \approx 0.03$) and the number of boxes ($n = 30$). Initially, a few heterozygous individuals for the considered loci (150 pairs, from 1 to 37 in each box) were placed in the boxes, i.e. $q_2^{(i)}(0) = 1$. At the same time, a large panmictic population was established, which was similar in size and initial frequency to the system of connected boxes. Based on the frequency dynamics of the *A* allele at the *α-Gdph* locus in a large population, we can easily estimate the selection parameters s_k (see the Table). As a basis for our study, we used the values of s_k derived from earlier work (Zhdanova, Frisman, 2023), where they were obtained using a one-dimensional equation for the frequency of allele *A* of the *α-Gdph* locus. The pattern of change in the frequency of allele *A* in the experiment closely matches the typical solution of model (1), with disruptive selection ($s_2 < s_1$ and $s_2 < s_3$) rather than directional selection ($s_1 > s_2 > s_3$ or $s_3 > s_2 > s_1$).

Based on the initial conditions ($N^{(i)}(0) = 1 \ldots 37$, $\sum N^{(i)}(0) = 300$), the population growth pattern, and the limiting number of individuals in each box ($\overline{N}^{(i)} \approx 135$), as well as in the local panmictic population, we can easily calculate the parameters for population growth, including values of a, h, N_0 and K, which are shown in the Table.

The average migration coefficient $\overline{m} = 0.025$ in the Table and the median value of $m\ln(2) \approx 0.035$ indicate that in most cases, the number of migrants does not exceed 4–5 individuals, which is similar to the results of the original experiment.

The greatest difficulty in verifying the model (9) involves selecting initial distributions of allele frequencies and abundances that yield final distributions similar to those presented in Chapter 4 of the book (Altukhov, 2003). In order to select initial conditions, we generate a set of initial frequencies and abundances using a feature of the experiment: individuals of the same sex are randomly included in some boxes and do not produce offspring. To describe this, let us create a vector of random numbers as follows: $N^{(i)}(0) \sim U[0, 37]$, so that $\sum N^{(i)}(0) \approx 300$, and let some boxes be initially empty ($N^{(i)}(0) = 0$). As a result, since $0 \le N^{(i)}(0) < N_0$ (lower than the effective number of breeders), in subsequent generations, the boxes will still remain empty and will be recolonized by migrants from neighboring boxes, the genetic structure of which may already differ significantly from the original one due to random genetic drift and selection. However, there may not be enough migrants to effectively sustain the subpopulation, and the box may remain empty for several generations.

Because the initial numbers in all boxes are below the effective population size (N_e), the natural selection is not effective, and we cannot ignore the effect of random genetic drift. The authors of the outlined experiment assumed $N_e \approx 50$. This means that after the 2nd or 3rd generation, the effect of deterministic selection processes begins to dominate over random processes that change allele frequencies. It would be difficult to directly describe genetic drift in the model (9) without significant modification or transitioning to a simulation model. Instead, we "simulate" the result of genetic drift by using the most likely initial frequency distribution, which is typically formed in model (1). With disruptive selection (s_k values from the Table), system (1) predicts that the frequencies of offspring genotypes in the 2nd and 3rd generations from completely heterozygous ancestors (with $q_2(0) = 1$) will be approximately $q_1 \approx 0.27$, $q_2 \approx 0.46$ and $q_3 \approx 0.27$. We can assume that, for the first few generations, genetic drift will randomly shift the frequencies away from their initial values while the population sizes remain below the effective population size N_e. As a result, the observed genetic divergence in the system of coupled populations can be equally explained by the initial differences in both population sizes and frequencies, caused by the initial genetic drift prior to reaching the effective size in each subpopulation.

To fit the initial frequencies, we generate two independent vectors of random numbers: $q_1^{(i)}(0) \sim U[0,1]$ and $q_2^{(i)}(0) \sim U[0,1]$ $\left(q_3^{(i)}(0) = 1 - \left(q_1^{(i)}(0) + q_2^{(i)}(0)\right)\right)$, and estimate how much the "true" initial frequencies may vary from the theoretical values of 0.27, 0.46, and 0.27 due to drift, so that after 50–60 generations, model (9) approximately describes

Values of parameters for model (9)

n	$\overline{m}$	s_1	s_2	s_3	a	h	N_0	K
30	0.025	0.244	0.069	0.227	3.6	5	5	90

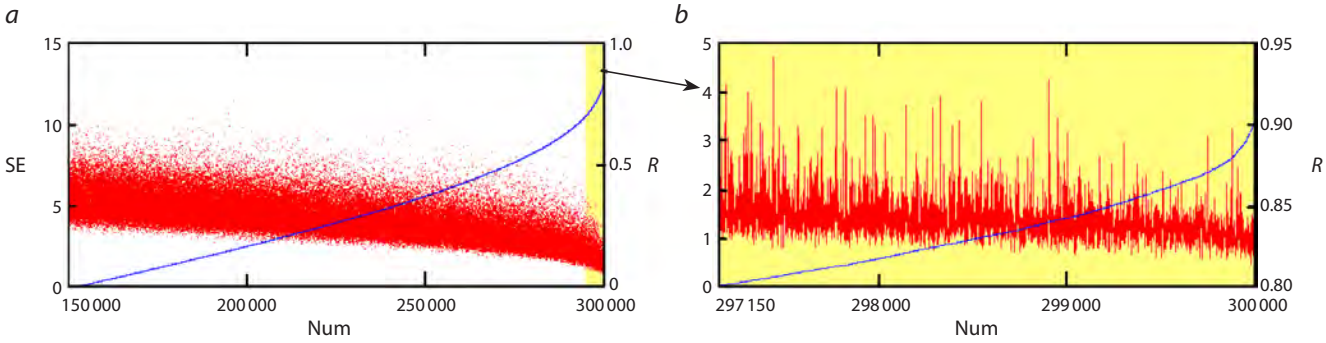


Fig. 2. Squared errors SE and correlation coefficients *R* for 300,000 initial conditions are ranked in order of increasing *R*. The Num is the "number" of initial conditions.

the real distribution of allele *A* frequencies at the *α-Gdph* locus. After examining 300,000 randomly selected initial frequencies and abundances, we found that only about 100 replicas most accurately describe the actual distribution, with the following distribution of initial frequencies:

$$q_1^{(i)}(0) = 0.25 \pm 0.1,$$
$$q_2^{(i)}(0) = 0.41 \pm 0.1, \qquad (11)$$
$$q_3^{(i)}(0) = 0.33 \pm 0.16.$$

This shows that we obtain a slightly lower frequency of heterozygotes and a shift towards homozygosity with the *aa* genotype than those predicted by model (1). Note that the experimental data also showed a slight shift in the average frequency of allele *A* towards allele *a* in the 5th generation, despite the lower fitness of s_3. Therefore, it would be reasonable to choose initial frequencies within these ranges. From a new set of 300,000 initial conditions of type (11), about 3,000 describe the actual frequency distribution quite well (Fig. 2). To assess the quality of the approximation, we used the correlation coefficient *R* between the actual and model frequency distributions of allele *A* at the *α-Gdph* locus in generation *t*, as well as the squared error SE:

$$\mathrm{SE}(t) = \sum_{i=1}^{n} \left(Q^{(i)}(t) - (q_1^{(i)}(t) + 0.5q_2^{(i)}(t))\right)^2.$$

Simulation results

We now consider the verification of equations (9) and analyze the mechanisms leading to stable genetic divergence.

Figure 3*a* shows two diagrams of the spatiotemporal dynamics in system (9) for the parameter values from the Table, using the most favorable initial conditions (Fig. 2*b*).

In the first diagram, the pixel color encodes the predominant genotype at site *i* and time *t*; in the second diagram, it encodes the population size. Figure 3*a* shows that at the initial stages, all subpopulations are polymorphic and contain all three genotypes (shown in green). Over time, driven by selection and the dispersal of individuals within the distributed system, an equilibrium state is established. This state corresponds to a stable genetic divergence that persists for a long time (including for $t \gg 200$). In one part of the boxes, only individuals with the *AA* genotype (red) are present; in another, only those with the *aa* genotype (blue) are found; polymorphic subpopulations with a high frequency of heterozygotes (green) are located between them. In the diagram, the subpopulation numbered $i = 16$, along with its neighbors, maintains polymorphism for $t \gg 200$. The second diagram shows changes in population size, where pink corresponds to the maximum values (~135) and black to the minimum ones. This diagram reveals several boxes that were initially empty, demonstrating that their location does not correlate with the final distribution of genotypes.

As can be seen from Figure 3*b*, model (9) describes the observed frequency distribution quite well. However, in all simulation runs (i. e., replicas with varied migration coefficients, $m_{i,j}$), the distribution similar to that observed in the *Drosophila* experiments emerges slightly earlier – around the 50th generation rather than the 60th. This discrepancy could be attributed to inaccurately estimated growth parameters since the equations (9) seem to describe a slightly faster population growth and evolutionary rate than is observed in reality. Alternatively, genetic drift processes, which were simulated using random initial frequencies, may have prevailed over selection for a longer period in the real experiment than we assumed (e. g., for 2–3 generations until the population size reached an effective $N_e \approx 50$). However, there is another probable explanation. In the experiments with *D. melanogaster*, the sex and age composition of all subpopulations was artificially maintained to prevent generation overlap. Specifically, all adult individuals were removed from the boxes after the females laid eggs. However, the sex ratio varied considerably between boxes throughout the experiment. Some boxes exhibited a significant deficit of females, while others had a pronounced shortage of males. Consequently, not all females were able to produce offspring before the removal time, and some males fertilized multiple females. This violation of panmixia likely skewed the data, as each complete removal event set back the evolutionary process slightly. These complex processes are not fully captured by the relatively simple model (9), which is why it predicts a slightly faster rate of evolution.

In Figure 3*c*, the final 100 distributions (for $t = 100 \ldots 200$) of the total population size for each genotype are superimposed. The figure shows that, due to fluctuations in the number of

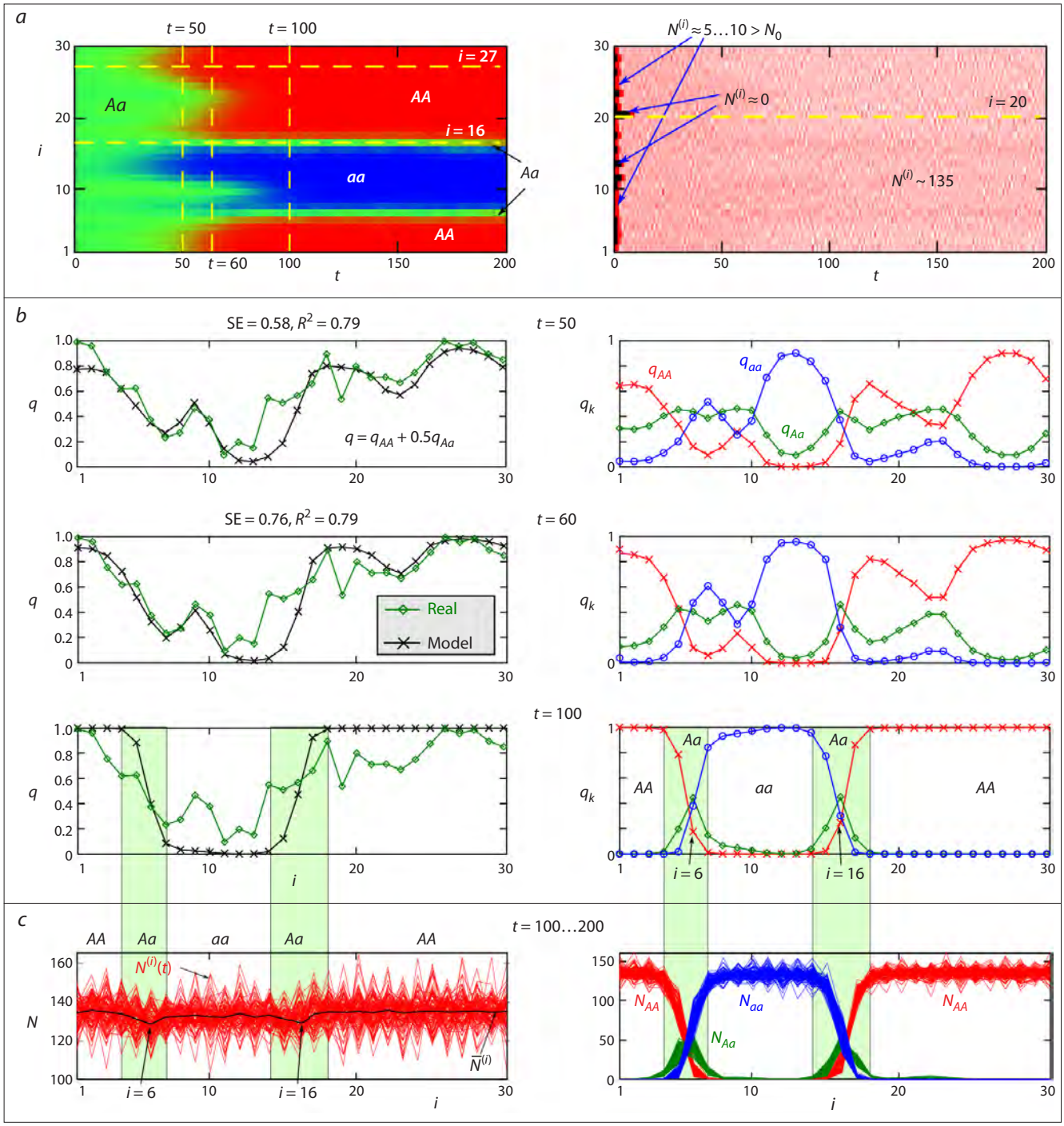


Fig. 3. *a*, Spatiotemporal dynamics of genotype frequencies and population sizes in the system of migration-coupled populations described by model (9). *b*, Modeled and observed frequency (q) distributions of the allele *A* at the α-*Gdph* locus and the frequency (q_k) of zygotes at the 50th, 60th, and 100th generations. *c*, The distribution of the total population size across the area (left), along with its components represented by the numbers of individuals with genotypes *AA*, *Aa* и *aa*.

migrants, the population size in different boxes undergoes irregular, non-synchronous oscillations. Furthermore, it is evident that the polymorphic subpopulations (i = 6 and 16) have a lower average abundance ($\overline{N}^{(i)}$) than the surrounding monomorphic subpopulations, which is consistent with the significant frequency of heterozygotes in these populations.

As shown in the first diagram of Figure 3*a*, the subpopulations evolve at different rates. This rate is determined by how close the initial population size of a subpopulation is to the effective size (N_e) and how close its initial allele frequency is to its final state (q = 1 or 0). For instance, the diagram highlights box i = 27, where the frequency of allele *A* was among the first to reach fixation (q = 1). Notably, this subpopulation

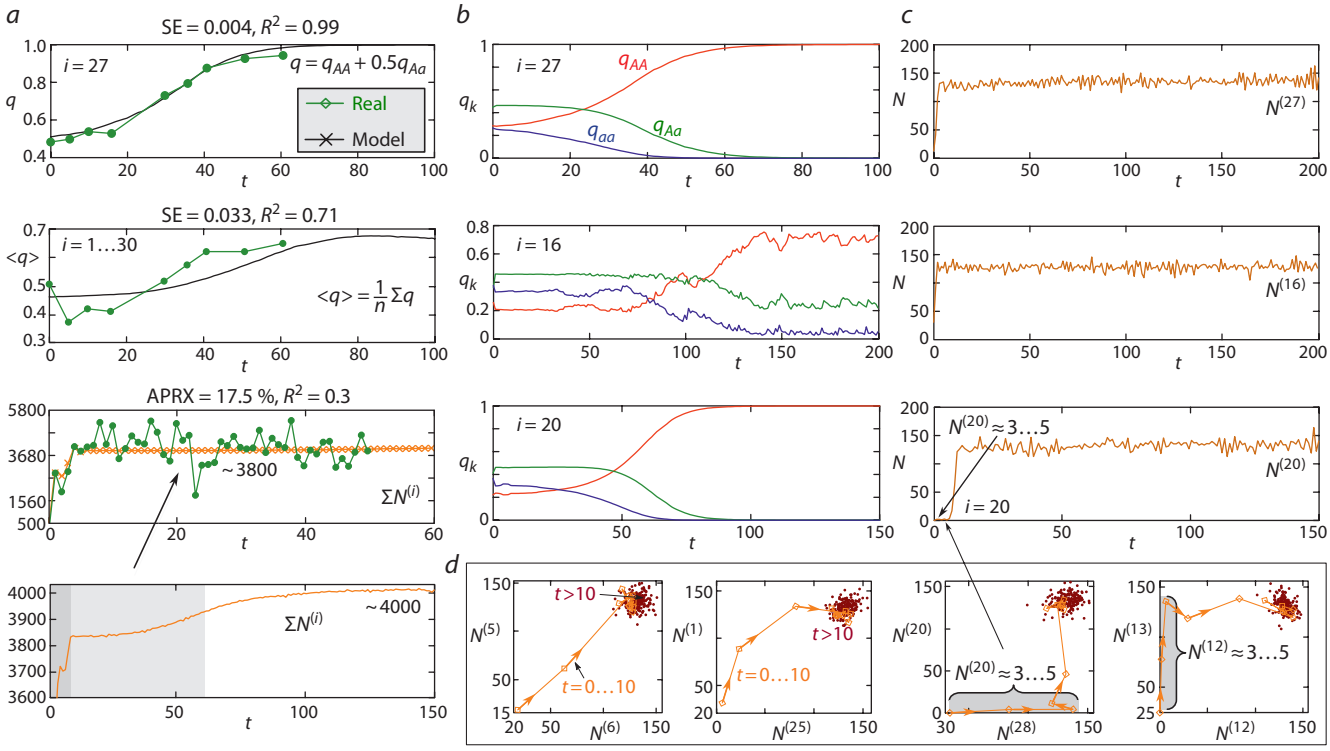


Fig. 4. *a*, Modeled and observed dynamics of the frequency *q* of allele *A* at the α-*Gdph* locus and the total number of populations of *D. melanogaster* in the box system. SE is the squared error, R^2 is the coefficient of determination, and APRX is the approximation error. Model dynamics of genotype frequencies (*b*) and population sizes (*c*) of the subpopulations highlighted in Fig. 3*a*. *d*, Phase portraits illustrating the group dynamics of the two subpopulations; the light brown color denotes the stage of rapid box colonization, and brown indicates the transition to the maximum population size.

evolves similarly to a large panmictic population (the first graph in Fig. 4*a*). Other subpopulations, as a rule, evolve more slowly.

Figure 4 demonstrates the correlation between the dynamics of allele frequencies and population sizes predicted by model (9) and the actual experimental data. Figure 4*a* shows that the modeled and experimentally observed average frequency of allele *A* across all 30 boxes follow a similar trend, stabilizing at a value of $q \approx 0.65$. The discrepancy between the modeled and observed average frequency at time point $t = 5$ can be explained by the fact that model (9) does not directly account for genetic drift, which occurred in the experimental population; instead, its effect is simulated solely through random perturbations of the frequency in the polymorphic population.

The third graph, Figure 4*a*, shows the observed and modeled total population sizes for the system of 30 subpopulations. The fourth graph (Fig. 4*a*) shows that the transition to the maximum population size proceeds through three stages: explosive growth over 2–3 generations from a small number of founders; reaching a quasi-stationary level with a total size of approximately $\sum N^{(i)} \sim 3800$ individuals, at which point there is already a distinct differentiation of genotypes by box groups, but the system still remains sufficiently polymorphic (Fig. 3*b* at $t = 50$); and a transition to the final distribution (Fig. 3*b* at $t = 100$) and the maximum total population size of approximately 4,000 individuals. As can be seen, model (9) describes only the general trends of population growth, which is explained by the fact that its behavior is, in principle, the only possible type of dynamics at $r = a\overline{W} < e^2 \approx 7.38$. Furthermore, equation (8), which describes the dynamics of a local population, does not account for sex and age structure or many other factors that undoubtedly caused irregular fluctuations in the experimental populations. More importantly, model (9) describes only the reproductive core of the population system – females and an equal number of males – and does not consider the fact that some males could have remained single and constituted the majority of migrants. As a result, the modeled population size is lower than the actual observed size.

At the same time, the modeled dynamics of the total population size, $\sum N^{(i)}$, result from non-synchronous fluctuations of each subpopulation around a stationary value of approximately 135 individuals per box (Fig. 4*c*, *d*). Summing these values smooths out all differences in the sizes of the subpopulations. Despite heterogeneities in the initial distributions of individuals, population growth in the first 5 generations – driven by increased fitness – occurs synchronously in almost all boxes (the first and second panels in Fig. 4*d*). The exception are boxes that were initially empty or had an insufficient number of breeders (the third and fourth in Fig. 4*d*). For these boxes, a non-zero population size of approximately 3–5 individuals is maintained solely by migrants. In all other boxes, the numbers slowly reach their maximum values and fluctuate around them (dark dots in Fig. 4*d*).

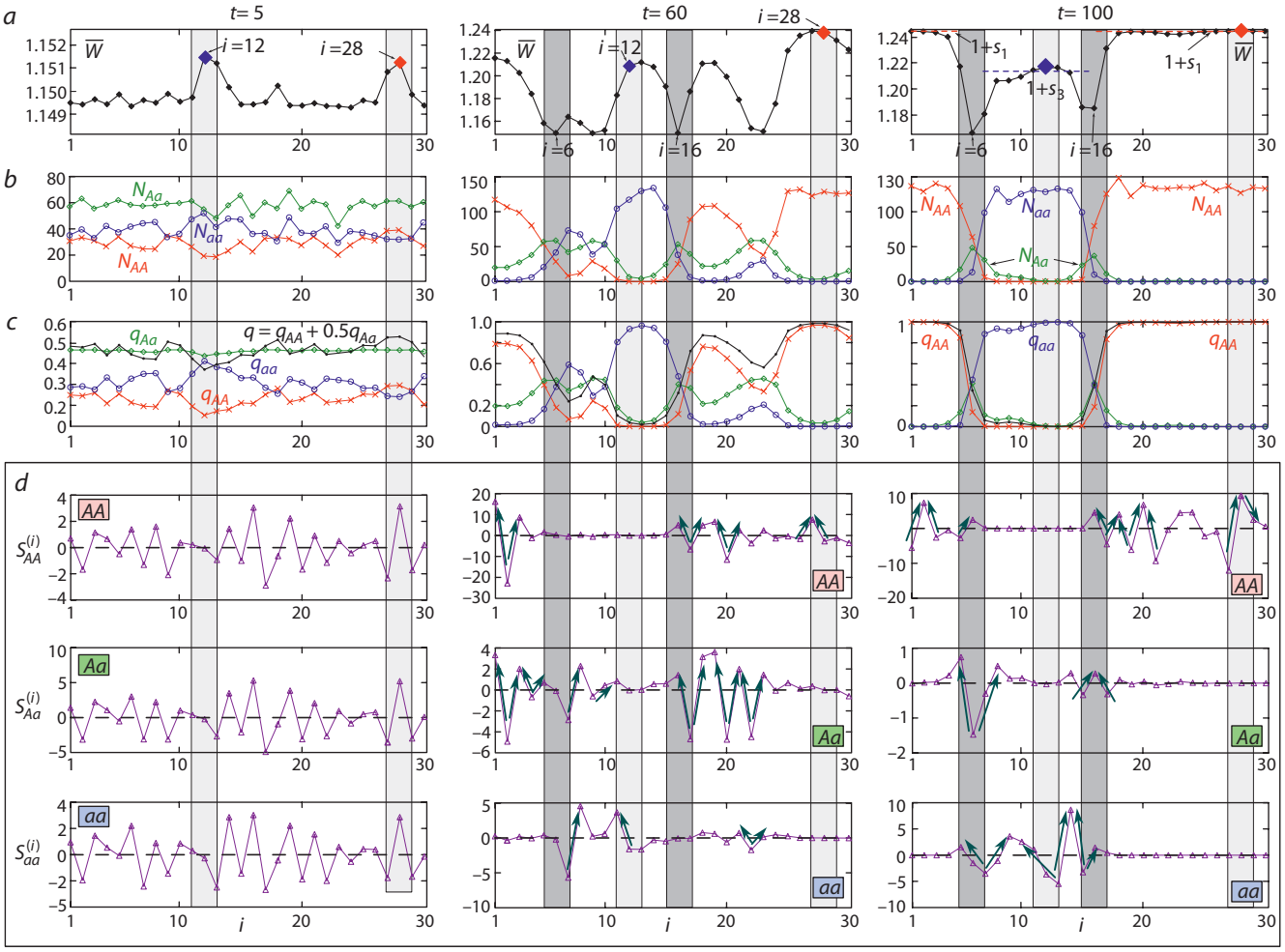


Fig. 5. Distribution of the average fitness values for each subpopulation before migration (*a*), population sizes (*b*), and frequencies (*c*) of the *AA*, *Aa*, and *aa* genotypes. *d*, Distribution of the migration balance values for each site.

The graphs highlight the areas where groups with the *AA* ($i = 28$) and *aa* ($i = 12$) genotypes are formed, as well as areas with active hybridization of individuals ($i = 6$ and $i = 16$). The arrows on the balance charts indicate the flow directions of individuals with the corresponding genotypes.

We now consider the mechanisms that could generate and maintain the observed spatial divergence in allelic composition within this experimental population system.

Analysis of migration flows

One of the reasons for the observed differentiation between the subpopulations is revealed by the small declines in population size in boxes $i = 6$ and $i = 16$, where polymorphism was maintained (boxes designated as *Aa* in Fig. 3). These declines become apparent only in the final distribution, as these boxes are surrounded by subpopulations with opposite genotypes and have a large population number. However, the presence of such subpopulations indicates only the possible mechanisms for maintaining divergence, rather than the reasons of its initial occurrence. These boxes can be considered as the hybrid zones, the allelic composition of which is maintained solely through migration and gene flow from sites inhabited by individuals with fixed opposite genotypes.

To study the mechanisms of the formation and maintenance of divergence, we will consider changes in the average fitness in each box $\overline{W}^{(i)}$ (Fig. 5*a*), the numbers of individuals of each genotype $N_k^{(i)}$ (Fig. 5*b*), and allele frequencies $q_k^{(i)}$ (Fig. 5*c*) over time. We will also assess the contribution of migration to the process of natural selection and the transition to the final frequency distribution. The migration balance of individuals with genotype k ($k = AA$, Aa, or aa) in the subpopulation i will be calculated using the following formula:

$$S_k^{(i)} = m_{i,i+1} q_k^{(i+1)} N^{(i+1)^*} + m_{i,i-1} q_k^{(i-1)} N^{(i-1)^*} - \\ - (m_{i-1,i} + m_{i+1,i})\, q_k^{(i)} N^{(i)^*}, \qquad (12)$$

where $q_k^{(i)} N^{(i)^*}$ represents the number of individuals with genotype k after selection, but before migration. This value is equal to the difference between the number of arrivals (the first two terms) at the site with index i and the number of departures (the third term) of individuals. The value of S indicates whether the size of the subpopulation with index i has increased ($S > 0$) or decreased ($S < 0$) due to migration (Fig. 5*d*). By comparing these three values, we can easily determine the directions of migration (arrows in Fig. 5*d*).

When selecting the initial conditions, it was found that the experimentally observed frequency distribution in model (9) occurs when the initial frequencies are shifted toward the

prevalence of homozygotes with the *aa* genotype. Note that the *AA* and *aa* genotypes differ in fitness by approximately 11 %. This means that for the most adapted *AA* genotype to become fixed, it must overcome this fitness threshold for a small proportion of subpopulations. However, a rarer set of circumstances is required for the less adapted *aa* genotype to avoid complete displacement, allowing both traits to be maintained.

Figure 5*a* shows that after a period of rapid growth until the 5th generation, two sites are distinguished, numbered $i = 12$ and $i = 28$, in which the frequency distribution yields the highest values of both average fitness $\overline{W}^{(i)}$ and total reproductive potential $a\overline{W}^{(i)}$ among all others. Although this difference is small (1 % for *aa* and 0.7 % for *AA*), it proves sufficient to initiate the separation of individuals of the same genotype near these boxes. This likely required a frequency shift in more than one site. Figures 5*b* and 5*c* show the distributions of population sizes and genotype frequencies, respectively. It can be observed that near site $i = 12$ at $t = 5$, there are at least six boxes with an increased number of *aa* homozygotes (and $q < 0.5$) relative to their surroundings. This implies that the flow of migrants from this region for any random $m_{i,j}$ is primarily represented by this genotype, which promotes its fixation. Site $i = 28$ has only one neighboring box with a high number of *AA* homozygotes (and $q > 0.5$), but this proves sufficient to fix the best-adapted genotype. Until approximately generation 50, sites $i = 12$ and $i = 28$ maintain the highest rates of fitness increase, exhibit frequencies closer to their final values ($q = 1$ or $q = 0$), and clearly support larger numbers of the corresponding genotype compared to their surroundings. As a result, migrants from these boxes are more genetically homogeneous than those from other boxes, and even the stochastic migration does not alter the overall evolutionary trend – homozygotes displace the less adapted heterozygotes.

On the migration balance $S_k^{(i)}$ graphs (Fig. 5*d*), it can be observed that at the initial stages ($t = 5$), the distribution of both the direction and intensity of individual flows between sites appeared largely random and comparable across different genotypes. As spatial differentiation progresses and better-adapted individuals displace less adapted ones, homogeneous areas with the largest population sizes ($i = 12$ and $i = 28$) begin to contribute more significantly to migration than highly polymorphic areas. By the 60th generation, two monomorphic groups with opposite traits, *AA* and *aa*, reach their largest sizes (*AA* – 17 boxes, *aa* – 8 boxes) and come into contact. However, since they have by then accumulated a sufficient number of individuals and their population sizes prove to be comparable, the resulting migrant flows also become comparable, despite the 11 % difference in fitness. As a result, in the hybrid zones near sites numbered $i = 6$ and $i = 16$, two equally large streams of individuals with opposite genotypes converge, ensuring a non-zero number of heterozygotes in these boxes. The outflow from these boxes is much weaker and is barely sufficient to maintain a low level of polymorphism in their vicinity. However, it is these hybrid zones that slow down the flows of homozygous individuals of different forms, preventing the better-adapted *AA* genotype from achieving complete fixation throughout its range.

Discussion

The verification of model (9) against the experimental data from Yu.P. Altukhov's study on box populations of *D. melanogaster*, along with the analysis of scenarios underlying the formation of heterogeneous distributions of allele frequencies and population sizes, requires further clarification.

First, it is necessary to discuss the reason for the pronounced differences in fitness observed among genotypes with different allele combinations of the α-Gdph enzyme, as revealed by estimates of the selection coefficients s_k. It is quite plausible that the *α-Gdph* locus serves as a marker of disruptive selection operating within the system, acting not directly on the *α-Gdph* gene itself, but on closely linked adaptive genes. This may explain certain discrepancies between the observed and modeled distributions and frequency dynamics, since the overall adaptive effect and direction of selection – even for genes strongly linked to *α-Gdph* – are not simply additive. Instead, they result from more complex interactions, such as polygenic or complementary gene effects, epistasis, or multigene interaction.

Note that a significant difference in fitness is not a necessary condition for genetic divergence in model (1). It has been previously demonstrated that spatial differentiation can occur even with small differences in fitness. The degree of difference between genotypes, as well as the migration coefficient, determines the rate at which stable divergence is achieved, and the size of the resulting monomorphic subpopulations and hybrid zones (Kulakov, Frisman, 2025).

Despite the limitations noted above, the proposed model allows to analyze the processes that led to the primary genetic divergence observed in the experiment. It was found that the combined effect of genetic drift, density-dependent limitation, and gene flow – before the effective population size N_e and the minimum number of breeders N_0 were reached – resulted in some boxes accidentally containing a higher number of less adapted *aa* individuals than the more adapted *AA* ones. As a result, subpopulations with even a slight deviation in allele frequencies from the theoretically expected values (typical for a local panmictic population) reached the highest average fitness and population growth rate earlier than others. As emigrants carry the allelic composition of their source subpopulation, clusters of boxes with either *AA* or *aa* genotypes form around these rapidly growing groups. Gradually, these genotypes displace the less-adapted heterozygous *Aa* individuals and occupy the largest number of sites. The interaction between the two migrant streams, carrying *AA* and *aa* genotypes, maintains a non-zero number of heterozygous individuals in certain boxes, creating hybrid zones. On the one hand, their presence preserves the genetic diversity of the entire metapopulation. On the other hand, these zones prevent the fittest individuals from occupying the entire range.

This evolutionary scenario can be considered universal for several reasons. The divergence of natural populations is always preceded by the emergence of mutants with a new trait in certain areas. For such a trait to become fixed, especially if it confers no significant immediate advantage, strong reproductive isolation from the parental population is required. This may be a case of disruptive selection, which is manifested

not only in the reduced fitness of heterozygotes (hybrids) but also in positive assortative mating, which further diminishes the reproductive success of small hybrid populations. For instance, in the case of the hooded and carrion crow mentioned in the Introduction, the primary isolating mechanism appears to be based on mating preferences. For crows, plumage color is significantly associated with innate perception of potential partners, which substantially reduces the likelihood of mating between dissimilar morphs but allows for crossbreeding between already hybrid individuals or between hybrid and "pure" forms (Poelstra et al., 2014; Kryukov, 2019).

Unlike seasonal migration, the dispersal of individuals and colonization of new sites is a slow process that unfolds over multiple generations. Consequently, the remote parts of a new area will be inhabited only by the descendants of the original migrants. During this gradual expansion, individuals will inevitably interbreed with local populations. The model proposed in this paper demonstrates that such dispersal will inevitably cease if the recipient site is inhabited by individuals possessing a different trait than the migrants, due to potential selection against hybrids. In the case of crows, assortative mating will restrict interbreeding between the different morphs in newly colonized areas, thereby significantly reducing the likelihood of further expansion. In the ring populations' system of *Drosophila*, the reduced fitness of heterozygotes decreases hybrid fertility and prevents their descendants from dispersing further. Consequently, for species where dispersal is a multigenerational process, hybrid zones act as significant barriers. They effectively impede the movement of individuals possessing one trait into areas occupied by individuals with another trait, without the need for those areas to be permanently settled, and with a high probability of producing hybrid offspring. If a more rapid dispersal mechanism is possible, this dynamic can change dramatically.

Conclusion

The dynamic model proposed in this paper enables a detailed investigation of the mechanisms underlying primary genetic divergence. These mechanisms are attributed to differences in genotype fitness, settlement patterns, migration, and the formation of stable hybrid zones. The model demonstrates the possibility of reproductive isolation between different forms of diploid organisms, which arises not only from geographical isolation, habitat remoteness, or ecological specialization but also from hereditary mechanisms, genetic drift, gene flow, and selection against heterozygotes. This type of selection results in stable spatial genotype differentiation, maintained by hybrid zones that act as effective barriers to the introgression of divergent traits.

Thus, disruptive selection is demonstrated to play a crucial role – an effect that can be detected through certain marker genes but is not always apparent from external morphology. Consequently, it may be far more widespread in nature than previously believed.

References

Aguillon S.M., Rohwer V.G. Revisiting a classic hybrid zone: movement of the northern flicker hybrid zone in contemporary times. *Evolution*. 2022;76(5):1082-1090. doi 10.1111/evo.14474

Allee W.C. The Social Life of Animals. Beacon Press, 1958

Altukhov Yu.P. Genetic Processes in Populations. Moscow: Akademkniga Publ., 2003 (in Russian)

Altukhov Yu.P., Bernashevskaya A.G. Experimental modeling of genetic processes in a population system of *Drosophila melanogaster* corresponding to a circular stepping-stone model: 2. Stability of allelic composition and periodic relationship of allele frequency with distance. *Soviet Genetics*. 1981;17(6):1052-1059 (in Russian)

Altukhov Yu.P., Bernashevskaya A.G., Milishnikov A.N. Experimental modeling of genetic processes in the population system of *Drosophila melanogaster* corresponding to the ring step model. *Soviet Genetics*. 1979;15(4):646-655 (in Russian)

Bazykin A.D. Reduced fitness of heterozygotes in a system of adjacent populations. *Soviet Genetics*. 1972;8(11):155-161 (in Russian)

Blair W.F. Mating call and stage of speciation in the *Microhyla olivacea–M. carolinensis* complex. *Evolution*. 1955a;9(4):469-480. doi 10.1111/j.1558-5646.1955.tb01556

Blair W.F. Size difference as a possible isolation mechanism in Microhyla. *Am Nat*. 1955b;89(848):297-301. doi 10.1086/281894

Blinov V.N., Zheleznova T.K. Black *Corvus corone* and grey *C. cornix* crows: controversial issues about status (races, semispecies or species?), origin (allo- or sympatric?) and the phenomenon of stable hybrid zones. *Russkiy Ornitologicheskiy Zhurnal = Russian Ornithological Journal*. 2020;29(1958):3596-3601 (in Russian)

Dey S., Joshi A. Stability via asynchrony in *Drosophila* metapopulations with low migration rates. *Science*. 2006;312(5772):434-436. doi 10.1126/science.1125317

Filchak K., Roethele J., Feder J. Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. *Nature*. 2000; 407(6805):739-742. doi 10.1038/35037578

Frisman E.Y. Primary Genetic Divergence (Theoretical analysis and modeling). Vladivostok, 1986 (in Russian)

Haring E., Däubl B., Pinsker W., Kryukov A., Gamauf A. Genetic divergences and intraspecific variation in corvids of the genus *Corvus* (Aves: Passeriformes: Corvidae) – a first survey based on museum specimens. *J Zool Syst Evol Res*. 2012;50(3):230-246. doi 10.1111/j.1439-0469.2012.00664.x

Kapitonova L.V., Formozov N.A., Fedorov V.V., Kerimov A.B., Selivanova D.S. Peculiarities of behavior and ecology of the Great tit *Parus major* Linneus, 1758 and Japanese tit *P. minor* Temmink et Schlegel, 1848 as possible factors of maintaining the stability of speciesspecific phenotypes in the area of sympatry and local hybridization in the Amur Region. *Dal'nevostochnyy Ornitologicheskiy Zhurnal = Far Eastern Journal of Ornithology*. 2012;3:37-46 (in Russian)

Keymer J.E., Galajda P., Muldoon C., Park S., Austin R.H. Bacterial metapopulations in nanofabricated landscapes. *Proc Natl Acad Sci USA*. 2006;103(46):17290-17295. doi 10.1073/pnas.0607971103

Kryukov A.P. Phylogeography and hybridization of corvid birds in the Palearctic Region. *Vavilov J Genet Breed*. 2019;23(2):232-238. doi 10.18699/VJ19.487

Kulakov M., Frisman E.Ya. Primary genetic divergence in a system of limited population coupled by migration in a ring habitat. *Mathematical Biology and Bioinformatics*. 2025;20(1):1-30. doi 10.17537/2025.20.1 (in Russian)

Láruson Á.J., Reed F.A. Stability of underdominant genetic polymorphisms in population networks. *J Theor Biol*. 2016;390:156-163. doi 10.1016/j.jtbi.2015.11.023

Littlejohn M.J. Premating isolation in the *Hyla ewingi* complex (Anura: Hylidae). *Evolution*. 1965;19(2):234-243. doi 10.2307/2406376

Matsumoto M., Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans Model Comput Simul*. 1998;8(1):3-30. doi 10.1145/272991.272995

Murphy M.A., Dezzani R., Pilliod D.S., Storfer A. Landscape genetics of high mountain frog metapopulations. *Mol Ecol*. 2010;19(17): 3634-3649. doi 10.1111/j.1365-294X.2010.04723.x

Orsini L., Corander J., Alasentie A., Hanski I. Genetic spatial structure in a butterfly metapopulation correlates better with past than present demographic structure. *Mol Ecol*. 2008;17(11):2629-2642. doi 10.1111/j.1365-294X.2008.03782.x

Poelstra J.W., Vijay N., Bossu C.M., Lantz H., Ryll B., Müller I., Baglione V., Unneberg P., Wikelski M., Grabherr M.G., Wolf J.B.W. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*. 2014;344(6190):1410-1414. doi 10.1126/science.1253226

Smith M.J., Osborne W., Hunter D. Geographic variation in the advertisement call structure of *Litoria verreauxii* (Anura: Hylidae). *Copeia*. 2003;4:750-758. doi 10.1643/HA02-133.1

Sundqvist L., Keenan K., Zackrisson M., Prodöhl P., Kleinhans D. Directional genetic differentiation and relative migration. *Ecol Evol*. 2016;6(11):3461-3475. doi 10.1002/ece3.2096

Tait C., Kharva H., Schubert M., Kritsch D., Sombke A., Rybak J., Feder J.L., Olsson S.B. A reversal in sensory processing accompanies ongoing ecological divergence and speciation in *Rhagoletis pomonella*. *Proc Biol Sci*. 2021;288(1947):20210192. doi 10.1098/rspb.2021.0192

Yeaman S., Otto S.P. Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*. 2011;65(7):2123-2129. doi 10.1111/j.1558-5646.2011.01277.x

Yee W.L., Goughnour R.B. Mating frequencies and production of hybrids by *Rhagoletis pomonella* and *Rhagoletis zephyria* (Diptera: Tephritidae) in the laboratory. *Can Entomol*. 2011;143(1):82-90. doi 10.4039/n10-047

Zhdanova O.L., Frisman E.Y. On the genetic divergence of migration-coupled populations: modern modeling based on the experimental results of Yu.P. Altukhov et al. *Russ J Genet*. 2023;59:614-622. doi 10.1134/S1022795423060133

Conflict of interest. The authors declare no conflict of interest.
Received July 29, 2025. Revised September 2, 2025. Accepted September 5, 2025.

EVOLUTIONARY BIOINFORMATICS
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1122-1128

doi 10.18699/vjgb-25-116

Asymmetry of nucleotide substitutions in tRNAs indicates common descent of modern organisms from a thermophilic ancestor

I.I. Titov [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
titov@bionet.nsc.ru

Abstract. The nature of the last universal common ancestor (LUCA) of all living organisms remains a controversial issue in biology. There is evidence of both thermophilic and mesophilic LUCA origin. The increasing complexity of the cellular apparatus during the evolution from early life forms to modern organisms could have manifested itself in long-term evolutionary changes in the nucleotide composition of genetic sequences. This work is devoted to the identification of such trends in tRNA sequences. The results of an evolutionary analysis of single-nucleotide substitutions in tRNAs of 123 species from three domains – Bacteria, Archaea and Eukaryota – are presented. A universal vector of directed evolutionary change in tRNA sequences has been discovered, in which substitutions of guanine (G) to adenine (A) and cytosine (C) to uracil (U) occur more frequently than the reverse. The most striking asymmetry in the number of substitutions is observed in the following transitions: a) purine-to-purine, where G→A outnumbers A→G, b) pyrimidine-to-pyrimidine, where C→U outnumbers U→C, and c) purine-to-pyrimidine and vice versa, where G→U outnumbers U→G. As a result, tRNAs could lose "strong" three-hydrogen-bond complementary pairs formed by guanine and cytosine and fix "weak" two-hydrogen-bond complementary pairs formed by adenine and uracil. 16 out of 20 tRNA families are susceptible to the detected change in sequence composition, which corresponds to the significance level $p = 0.006$ according to the one-sided binomial test. The identified pattern indicates a high GC content in the common ancestor of modern tRNAs, supporting the hypothesis that the last universal common ancestor (LUCA) lived in a hotter environment than do most contemporary organisms.

Key words: evolution; thermophile; mutations; tRNA; transition matrix; last universal common ancestor

For citation: Titov I.I. Asymmetry of nucleotide substitutions in tRNAs indicates common descent of modern organisms from a thermophilic ancestor. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1122-1128. doi 10.18699/vjgb-25-116

Funding. The work was supported by budget project No. FWNR-2022-0020.

Асимметрия нуклеотидных замен в тРНК свидетельствует об общем происхождении современных организмов от термофильного предка

И.И. Титов [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
titov@bionet.nsc.ru

Аннотация. Природа последнего универсального общего предка (last universal common ancestor, LUCA) всех ныне живущих организмов до сих пор остается актуальной проблемой биологии. Существуют свидетельства в пользу того, что LUCA был как термофилом, так и мезофилом. Усложнение клеточного аппарата в ходе эволюции от ранних форм жизни к современным организмам могло проявиться в долговременных эволюционных изменениях нуклеотидного состава генетических последовательностей. Выявлению подобных тенденций в последовательностях тРНК посвящена эта работа. Представлены результаты эволюционного анализа точечных нуклеотидных замен в тРНК 123 видов трех доменов: Bacteria, Archaea и Eukaryota. Обнаружен универсальный вектор направленного эволюционного изменения последовательностей тРНК, при котором замены гуанина (G) и цитозина (C) на аденин (A) и урацил (U) суммарно происходят чаще обратных. Наиболее ярко асимметрия числа замен наблюдается в следующих переходах: а) между пуринами в преобладании числа замен G→A над

числом замен A→G; б) между пиримидинами в преобладании C→U над U→C, а также в) при переходе из пурина в пиримидин и наоборот – в преобладании G→U над U→G. В результате эволюционного процесса тРНК могли терять «сильные» комплементарные пары с тремя водородными связями, формируемые гуанином и цитозином, и фиксировать «слабые» комплементарные пары с двумя водородными связями, образуемые аденином и урацилом. Обнаруженному изменению состава последовательностей были подвержены 16 из 20 семейств тРНК, что соответствует уровню статистической значимости $p = 0.006$ согласно одностороннему биномиальному тесту. Выявленная закономерность свидетельствует о высоком GC-содержании в последовательности общего предка современных тРНК и, следовательно, подтверждает предположение о том, что самая молодая из гипотетических общих предковых клеток, от которой произошли все ныне живущие организмы (последний универсальный общий предок, LUCA), обитала в более горячей среде, нежели ныне живущие организмы.

Ключевые слова: эволюция; термофил; мутации; тРНК; матрица перехода; последний универсальный общий предок

Introduction

Despite extensive research, the nature of the last universal common ancestor (LUCA) of all living organisms remains a pressing problem in biology. According to recent studies (Moody et al., 2024), LUCA arose approximately 4.2 billion years ago and possessed the basic elements of the cellular apparatus of modern prokaryotes (genes and molecular genetic systems for transcription and translation, including tRNAs). There is a debate about whether LUCA was a thermophile (Di Giulio, 2000; Weiss et al., 2016; Moody et al., 2024) or a mesophile (Galtier et al., 1999; Cantine, Fournier, 2017).

The increase in cellular complexity during the evolution from early life forms to modern organisms could have manifested itself in long-term evolutionary changes in the nucleotide composition of genetic sequences. Thus, in the work (Jordan et al., 2005), using the method of unrooted parsimony (Rickert et al., 2025), patterns of systematic unidirectional changes in the amino acid composition of proteins during their evolution from ancestral forms were identified: an increase in the content of the amino acids Cys, Met, His, Ser and Phe due to a decrease in the content of the amino acids Pro, Ala, Glu and Gly. In the work (Galtier et al., 1999), a comparison of LUCA ribosomal RNAs and those of modern species based on GC content was conducted, the results of which were subsequently criticized (Di Giulio, 2000). Of interest is the work (Men et al., 2022), in which fragments of LUCA ribosomal RNAs (16S, 5S, and 23S rRNA) that are evolutionarily conserved in modern sequences and correspond to sites of rRNA interaction with ribosome proteins were reconstructed. However, this study examined rRNA nucleotide sequences in the binary purine-pyrimidine code and, therefore, did not assess the G/C content of the RNA. Therefore, evolutionary changes in the RNA nucleotide composition from LUCA to modern species have not been definitively established.

In this regard, it seemed interesting to study long-term trends in changes in the nucleotide composition of RNA sequences, namely tRNA molecules, which are the most important element of translation systems in all organisms.

In our study, we examined the molecular evolution of 20 isoacceptor tRNA families, each of which mediates the transfer of a specific amino acid during translation. These tRNA families were analyzed for 123 organisms from three domains: Bacteria, Archaea and Eukaryota.

Phylogenetic analysis was performed using the unrooted parsimony method (Jordan et al., 2005). Single nucleotide substitutions were identified that became fixed in tRNAs during their evolution from ancestral sequences to modern ones, and it was shown that substitutions of guanine (G) or cytosine (C) for adenine (A) or uracil (U) are fixed more often than substitutions of A or U for G or C. This shapes a view of predominantly unidirectional evolutionary change of tRNA sequences, during which they lost “strong” complementary pairs with three hydrogen bonds formed by guanine and cytosine, and fixed “weak” complementary pairs with two hydrogen bonds formed by adenine and uracil. This feature was characteristic of 16 of the 20 tRNA families, with a significance level of $p < 0.006$ according to the one-sided binomial test.

The obtained results indicate a high content of G/C in the nucleotide sequences of tRNAs of the common ancestor of modern Bacteria, Archaea and Eukaryota and, therefore, support the assumption that the last universal common ancestor, LUCA, lived in a hotter environment than living organisms, i. e., was a thermophile or heat-loving mesophile (moderate thermophile). This conclusion is based on the fact that the content of G and C nucleotides in nucleotide sequences is associated with the optimal temperature of the organisms’ habitat, in connection with which genetic macromolecules (DNA, RNA) can be considered as a kind of molecular thermometers, and the content of G/C in them as an indicator of the temperature of the habitat.

Materials and methods

The tRNA nucleotide sequences of three domains (Bacteria, Archaea and Eukaryota) were taken from a curated database presented in the paper (Sprinzl et al., 1998, Supplementary Material S1)[1]. The database contained an alignment of tRNA sequences “most compatible with the tRNA phylogeny and known three-dimensional structures of tRNA” (Sprinzl et al., 1998). Each tRNA was assigned to its amino acid by the database authors.

The procedure for generating a sample of nucleotide sequences for evolutionary analysis was as follows. 1) For each of the 123 organisms, 20 tRNA groups were considered. Each group included a tRNA interacting with one of the 20 amino acids. Possible horizontal transfer (Soucy et al., 2015), as well as transitions between groups as a result of remodeling (a change in the isoacceptor group as a result of an anticodon change, for which only about 20 cases are currently known

[1] Supplementary Materials S1 and S2 are available at: https://vavilovj-icg.ru/download/pict-2025-29/appx41.zip

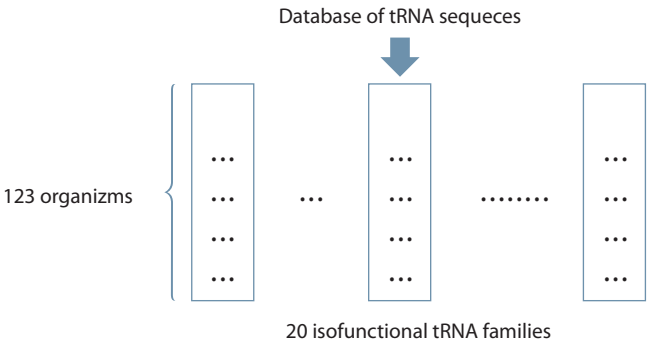


Fig. 1. Scheme of building the sample from the tRNA sequence database.

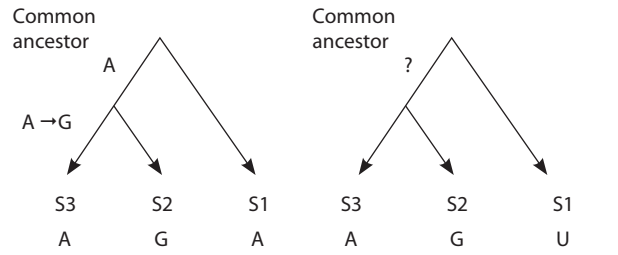


Fig. 2. Search for nucleotide substitutions using the unrooted maximum parsimony method on the simplest trees of three closest tRNAs.

The identified A→G substitution in the group of two closest relatives, S2 and S3, is shown on the left, and the uninformative substitution is shown on the right.

(Bermudez-Santana et al., 2010; Velandia-Huerto et al., 2016; Romanova et al., 2020)) were not considered. 2) For each position of the nucleotide sequences of this group corresponding to a specific organism and amino acid, the frequencies of four nucleotides were calculated, and the nucleotide with the highest frequency was assigned to the position in question; considering all positions of the sequences of the group, a consensus sequence of the tRNA group was constructed. 3) For a consensus sequence corresponding to a particular group of tRNAs, its similarity to each of the nucleotide sequences of the multiple alignment included in the group under consideration was assessed, and the sequence closest to the consensus was selected from this group.

Thus, a sample of tRNA nucleotide sequences for evolutionary analysis was formed, containing 20 × 123 = 2,460 typical tRNA sequences (Fig. 1). Each sequence in this sample was most typical for one of the isofunctional tRNA families of a given organism (out of 123).

Following (Jordan et al., 2005), identification of nucleotide substitutions recorded during the evolution of the nucleotide sequences of each isofunctional tRNA family was carried out based on the unrooted maximum parsimony method on phylogenetic trees with three vertices (Fig. 2) using the Dnapars program (Phylip package, Phylip, https://phylip web.github.io/phylip).

When analyzing a specific family of isoacceptor tRNAs, the following procedure was performed. For each S1 nucleotide sequence of 123 tRNA sequences in the family, the closest (in terms of similarity) S2 nucleotide sequence was identified, followed by the closest S3 sequence to S2 (Fig. 2), so that S2 and S3 formed a pair of closest relatives. This resulted in the formation of a phylogenetic triad in which S1 was the "outgroup" relative to the pair S2 and S3.

The unrooted maximum parsimony method assumes that if a nucleotide is found at a certain position in the sequence that is identical in S1, S2 and S3, then this nucleotide was present at the same position in the tRNA in the common ancestor of S1, S2 and S3. If, however, a different nucleotide is observed in S3, then a single nucleotide substitution occurred along the branch leading to S3. If all three nucleotides were different, then, following (Jordan et al., 2005), this position was considered uninformative and excluded from consideration. This method does not require stationarity and reversibility of the evolutionary process (Klopfstein et al., 2015).

Results

Following the approach of (Jordan et al., 2005) and considering nucleotide changes between the sequences of the closest ancestors and descendants, we constructed a mutational transition matrix for each of the 20 aligned tRNA families. Table 1 shows an example of such a matrix for the $tRNA^{Cys}$ family. Off-diagonal elements $M_{i,k}$ (i, k = 1,…,4) characterize the total number of single substitutions in the $tRNA^{Cys}$ sequences of nucleotide i to nucleotide k. Diagonal elements $M_{i,k}$ correspond to conserved positions. Rows and columns with gaps in the alignments (–) mainly corresponded to the variable loop region and were omitted for quantitative assessments.

Table 1 shows that among the nucleotide substitutions identified for the $tRNA^{Cys}$ family, the most frequently observed were transitions, i. e. substitutions between purines ($N_{G\to A}$ = 139 and $N_{A\to G}$ = 113) and between pyrimidines ($N_{C\to U}$ = 177 and $N_{U\to C}$ = 138).

It is noteworthy that the number of substitutions of "strong" nucleotides with "weak" ones (G→A, G→U, C→A, C→U), which is 417, exceeds the number of substitutions of "weak" nucleotides with "strong" ones (A→G, A→C, U→C, U→G), which is 340. This indicates an evolutionary trend toward a decrease in the G/C content of tRNAs in favor of an increase in the A/U content. The effect we identified, described above, was termed nucleotide substitution asymmetry.

We arrive at qualitatively similar conclusions by examining mutational transitions in the $tRNA^{Glu}$ family (Table 2). In this family, the number of substitutions of "strong" nucleotides with "weak" ones is 454, and the number of substitutions of "weak" nucleotides with "strong" ones is 302.

A similar analysis was performed for all 20 isoacceptor tRNA families (Supplementary Material S2). Next, we estimated the asymmetry effect for all isoacceptor tRNA families. For this purpose, we calculated a general substitution matrix by summing the corresponding elements of all 20 isoacceptor tRNA family matrices (Supplementary Material S2). For all tRNAs, the number of identified single substitutions was 24,653, and the number of uninformative substitutions was 2,083.

The diagonal elements of the resulting matrix (Table 3) characterize the average nucleotide composition of tRNAs from the studied species: 32.9 % (G), 27.8 % (C), 21.0 % (U), 18.3 % (A), as well as the content of "strong" G + C nucleotides (60.7 %) and "weak" ones (39.3 %). Transitions are represented by four out of the twelve off-diagonal ele-

Table 1. Matrix of the number of single-nucleotide substitutions in $tRNA^{Cys}$ sequences

From\to	A	C	G	U	–
A	1,526	38	113	44	27
C	43	2,292	74	177	25
G	139	91	2,469	58	6
U	45	138	51	1,492	23
–	20	27	0	39	3,131

Note. Here and in Tables 2 and 3: green indicates the number of substitutions of "strong" nucleotides (G and C, which form complementary pairs with three hydrogen bonds) with "weak" nucleotides (A and U, which form complementary pairs with two hydrogen bonds). Yellow indicates the number of substitutions of "weak" nucleotides A and U with "strong" nucleotides G and C. The column marked with a "–" sign indicates the number of substitutions at alignment positions corresponding to deletions.

Table 2. Matrix of the number of single-nucleotide substitutions in $tRNA^{Glu}$ sequences

From\to	A	C	G	U	–
A	1,353	40	101	57	37
C	52	2,526	105	184	35
G	167	105	2,389	51	9
U	58	124	37	1,608	27
–	30	35	0	23	2,956

ments. The proportion of transitions in the total number of substitutions was 56 %.

As in most partial matrices for individual families of isoacceptor tRNAs (see, for example, Tables 1 and 2), in Table 3, the number of substitutions of "strong" nucleotides with "weak" ones (shown in green) exceeds the number of substitutions of "weak" nucleotides with "strong" ones (marked in yellow): cf. $N_{G\to A}$ = 3451 and $N_{A\to G}$ = 2949, $N_{C\to U}$ = 3963 and $N_{U\to C}$ = 3468, $N_{G\to U}$ = 1421 and $N_{U\to G}$ = 1261, $N_{C\to A}$ = 963 and $N_{A\to C}$ = 952.

To quantitatively assess the asymmetry of substitutions $A_{F\to Z}$, the relative difference was calculated, defined as the doubled difference of two values divided by their sum – the number of substitutions between nucleotides F and Z, where F, Z∈(A, U, G, C):

$$A_{F\to Z} = \frac{2(N_{F\to Z} - N_{Z\to F})}{N_{F\to Z} + N_{Z\to F}}. \quad (1)$$

Table 4 presents the results of $A_{F\to Z}$ calculations based on (1) and Table 3. The asymmetry in the number of substitutions was: 0.16 for G→A and A→G; 0.14 for C→U and U→C; 0.12 for G→U and U→G. The remaining transitions were slightly asymmetric: from 0.008 to 0.028 (Table 4).

Based on Table 3, we can also calculate the balance of losses and gains of B_F for the F-type nucleotide:

$$B_F = \sum_Z (N_{Z\to F} - A_{F\to Z}). \quad (2)$$

Table 3. Matrix of the number of nucleotide substitutions identified by the unrooted parsimony method for tRNAs, summarized for all isoacceptor families

From\to	A	C	G	U	–
A	28,841	952	2,949	1,273	853
C	963	43,778	1,829	3,963	951
G	3,451	1881	51,756	1,421	330
U	1,272	3,438	1,261	32,994	715
–	666	862	210	867	53,981

Table 4. Asymmetry of nucleotide substitutions in tRNAs

$A_{G\to A}$	$A_{C\to U}$	$A_{G\to U}$	$A_{G\to C}$	$A_{A\to U}$	$A_{C\to A}$
0.16	0.14	0.12	0.028	0.008	0.011

Table 5 shows the total decrease in the number of "strong" G/C nucleotides in the studied nucleotide sequences of all analyzed tRNA families by 1,198 (714 G + 484 C) due to the evolutionary gain of the same number of weak A/G nucleotides (512 A + 686 U). Considering the total number of G, C, A, and U nucleotides in the studied tRNA sequences, the changes in the number of these nucleotides during the evolution of tRNA families, normalized by their number, were –0.014, –0.011, +0.018, and +0.021 for G, C, A, and U, respectively (Table 5).

The nucleotide substitution matrices for all 20 isoacceptor tRNA families are given in Supplementary Material S2. Table 6, obtained from these 20 matrices, shows the arithmetic differences $N_{F\to Z} - N_{Z\to F}$ (F, Z ∈(A, U, G, C)) between the numbers of all possible types of nucleotide substitutions fixed in the evolution of 20 isoacceptor families of tRNAs. Each variant of the arithmetic difference in the number of F→Z and Z→F substitutions corresponds to a specific column in Table 6. Each row in this table corresponds to a specific isoacceptor family of tRNAs. The last column shows the relative difference in the number of substitutions, $A_{S\to W}$, of "strong" nucleotides, S∈(G, C) with "weak" nucleotides, W∈(A, U), determined by equation (1).

Table 6 shows that 16 tRNA families are characterized by a positive value of the relative difference in the number of substitutions, $A_{S\to W} > 0$. At the same time, four families of tRNAs (bottom lines) are characterized by a negative difference, < 0. Of these four families of tRNAs, for three tRNAs ($tRNA^{Gly}$, $tRNA^{Thr}$ and $tRNA^{Val}$), the observed negative trend, i. e. the predominance of W→S substitutions over S→W, is insignificant ($-0.06 \le A_{S\to W} \le -0.03$), and only for $tRNA^{Lys}$, the predominance of W→S substitutions over S→W is pronounced ($A_{S\to W} = -0.34$).

A one-sided binomial test was used to assess the significance of the predominance of positive values $A_{S\to W}$ characterizing the relative difference between a) the number of substitutions of "strong" nucleotides with "weak" nucleotides (S→W) and b) the number of substitutions of "weak" nucleotides with

Table 5. Characteristics of the composition and evolutionary dynamics of the studied nucleotide sequences of all analyzed tRNA families

Characteristics of the composition and evolutionary dynamics	G	C	A	U
Total number of conserved nucleotides of four types in trees of unrooted parsimony for the studied tRNA sequences	51,756	43,778	28,841	32,994
Average content of nucleotides of four types in the studied tRNA sequences	32,9	27.8	18.3	21.0
Changes in the number of nucleotides of four types during the evolution of tRNA families	–714	–484	+512	+686
Changes in the number of nucleotides of four types during the evolution of tRNA families, normalized by their number	–0,014	–0,011	+0,018	+0,021

Table 6. Arithmetic differences $N_{F\to Z} - N_{Z\to F}$ (F, Z∈(A, U, G, C)) between the numbers of nucleotide substitutions of all possible types fixed in the process of evolution of 20 isoacceptor families of tRNAs

tRNA	$N_{G\to A} - N_{A\to G}$	$N_{C\to U} - N_{U\to C}$	$N_{G\to U} - N_{U\to G}$	$N_{G\to C} - N_{C\to G}$	$N_{A\to U} - N_{U\to A}$	$N_{C\to A} - N_{A\to C}$	$A_{S\to W}$*
Ala	–5	36	21	0	–20	–4	0.13
Arg	20	41	4	6	21	21	0.14
Asn	45	30	10	4	–10	–11	0.19
Asp	32	4	20	2	–2	13	0.21
Cys	26	39	7	17	–1	5	0.20
Gln	–4	–2	21	–3	10	31	0.11
Glu	66	60	14	0	–1	12	0.40
His	52	2	–18	10	–4	–17	0.04
Ile	25	–2	5	16	13	8	0.12
Leu	62	89	25	6	–13	25	0.14
Met	34	45	–11	14	7	–9	0.12
Phe	20	44	7	4	19	8	0.24
Pro	29	21	24	–9	2	14	0.20
Ser	50	105	61	–5	–12	–32	0.19
Trp	44	13	6	0	3	–4	0.16
Tyr	44	48	7	–3	–23	5	0.24
Gly	–11	4	–7	9	5	6	–0.04
Thr	–21	–26	–17	–14	–5	0	–0.06
Val	12	12	–23	–5	7	–19	–0.03
Lys	–18	–58	–31	3	5	–41	–0.34

* The last column shows the value of the relative difference in the number of substitutions between "strong" and "weak" nucleotides, $A_{S\to W} = 2(N_{S\to W} - N_{W\to S})/(N_{S\to W} + N_{W\to S})$, where S∈(G, C), W∈(A, U).

"strong" nucleotides (W→S) fixed during the evolution of 20 tRNA families (Lehmann, 2012). In our case, the level of significance was calculated as the probability p of random observation of 16 matrices out of 20 with substitutions in favor of a decrease in the number of "strong" G/C nucleotides: see expression (3). At the same time, it was assumed that the number of recorded substitutions of types S→W and W→S was the same on average.

$$p = \sum_{l=16}^{l=20} C_l^{20} 0.5^{20} = 0.0059. \quad (3)$$

Using (3), the statistical hypothesis of the asymmetry of evolutionary substitution matrices in the direction of G and C nucleotide loss and A and U nucleotide gain was accepted with a significance level of $p < 0.006$.

Discussion

Our analysis of the evolution of 20 isoacceptor tRNA families of 123 species of the three domains (Bacteria, Archaea and Eukaryota) from their ancestral forms revealed a tendency to decrease the G/C composition of tRNAs in favor of an increase in the A/U composition. This effect was called the asymmetry of nucleotide substitutions. It consisted in the evolutionary loss of "strong" nucleotides G and C, capable of forming energy-advantageous complementary pairs with three hydrogen bonds, and the gain of "weak" nucleotides A and U, which form less stable complementary pairs with two hydrogen bonds. 16 out of the 20 tRNA families were affected by the detected change in sequence composition, which corresponds to the significance level of $p < 0.006$ according to the one-sided binomial test.

The results suggest that the last universal common ancestor, LUCA, lived in a hotter environment than currently living organisms; i. e. it was a thermophile or a thermophilic mesophile (moderate thermophile). This conclusion is substantiated by the fact that the content of nucleotides G and C in nucleotide sequences is associated with the optimal temperature of organisms (Dutta, Chaudhuri, 2010), in connection with which genetic macromolecules (DNA, RNA) can be considered as a kind of molecular thermometers, and their G/C content is an indicator of the temperature of the environment.

Early Earth conditions must have determined the energetic, metabolic, biochemical, and environmental features of LUCA. According to (Di Giulio, 2000; Weiss et al., 2016), LUCA lived in hot springs, the high temperature of which facilitates the course of biochemical reactions and molecular genetic processes, but requires thermodynamic and kinetic stability of biomolecular structures, the thermodynamic fluctuations of which are more pronounced the higher the temperature of the environment. Modern thermophiles are adapted to high temperatures due to the high content of nucleotides G and C in the genome (Dutta, Chaudhuri, 2010), which form stronger complementary bonds with each other. And this is especially important for the thermal stability of structural RNAs, including tRNAs.

It should be noted that four out of the 20 families of tRNAs studied in our work do not follow the general trend of losing "strong" nucleotides. The reasons that determined the peculiarities of the evolution of these tRNAs could vary. For example, two families, $tRNA^{Gly}$ and $tRNA^{Val}$, correspond to chemically simple, so-called "Miller" amino acids. Presumably, these amino acids were part of the most ancient proteins and the nucleotide composition of their tRNAs could have had time to reach their individual evolutionary equilibrium, albeit different from the average for all tRNAs. However, overall, comparing the G/C composition of tRNAs in organisms living at different temperatures, our results suggest that modern organisms, on average, live in colder environments than LUCA.

Conclusion

A universal vector of directed evolutionary change in tRNA sequences has been discovered, in which the substitution of guanine (G) and cytosine (C) with adenine (A) and uracil (U) in total occurs more often than the reverse. As a result of the evolutionary process, tRNAs could lose "strong" complementary pairs with three hydrogen bonds, formed by guanine and cytosine, and fix "weak" complementary pairs with two hydrogen bonds, formed by adenine and uracil. 16 out of the 20 tRNA families were affected by the detected change in sequence composition, which corresponds to the level of statistical significance $p = 0.006$ according to the one-sided binomial test. This pattern suggests high G/C content in the sequence of the common ancestor of modern tRNAs and, therefore, supports the assumption that the youngest of the hypothetical common ancestral cells, from which all currently living organisms descended (the last universal common ancestor, LUCA), lived in a hotter environment than currently living organisms.

References

Bermudez-Santana C., Attolini C.S.-O., Kirsten T., Engelhardt J., Prohaska S.J., Steigele S., Stadler P.F. Genomic organization of eukaryotic tRNAs. *BMC Genomics*. 2010;11(1):270. doi 10.1186/1471-2164-11-270

Cantine M.D., Fournier G.P. Environmental adaptation from the origin of life to the last universal common ancestor. *Orig Life Evol Biosph*. 2017;48(1):35-54. doi 10.1007/s11084-017-9542-5

Di Giulio M. The universal ancestor lived in a thermophilic or hyperthermophilic environment. *J Theor Biol*. 2000;203(3):203-213. doi 10.1006/jtbi.2000.1086

Dutta A., Chaudhuri K. Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: indications for thermal adaptation. *FEMS Microbiol Lett*. 2010;305(2):100-108. doi 10.1111/j.1574-6968.2010.01922.x

Galtier N., Tourasse N., Gouy M. A non hyperthermophilic common ancestor to extant life forms. *Science*. 1999;283(5399):220-221. doi 10.1126/science.283.5399.220

Jordan I.K., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., Sunyaev S. A universal trend of amino acid gain and loss in protein evolution. *Nature*. 2005;433(7026):633-638. doi 10.1038/nature03306

Klopfstein S., Vilhelmsen L., Ronquist F. A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Syst Biol*. 2015;64(6):1089-1103. doi 10.1093/sysbio/syv052

Lehmann E.L. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? In: Rojo J. (Ed.) Selected Works of E.L. Lehmann. Selected Works in Probability and Statistics. Boston, MA: Springer, 2012;201-208. doi 10.1007/978-1-4614-1412-4_19

Men Y., Lu G., Wang Y., Lin J., Xie Q. Reconstruction of the rRNA sequences of LUCA, with bioinformatic implication of the local similarities shared by them. *Biology*. 2022;11(6):837. doi 10.3390/biology11060837

Moody E.R.R., Álvarez-Carretero S., Mahendrarajah T.A., Clark J.W., Betts H.C., Dombrowski N., Szánthó L.L., … Spang A., Pisani D., Williams T.A., Lenton T.M., Donoghue P.C.J. The nature of the last universal common ancestor and its impact on the early Earth system. *Nat Ecol Evol*. 2024;8(9):1654-1666. doi 10.1038/s41559-024-02461-1

Rickert D.A., Fan L.W.-T., Hahn M.W. Inconsistency of parsimony under the multispecies coalescent. *Theor Popul Biol*. 2025;166:56-69. doi 10.1016/j.tpb.2025.09.004

Romanova E.V., Bukin Y.S., Mikhailov K.V., Logacheva M.D., Aleoshin V.V., Sherbakov D.Yu. Hidden cases of tRNA gene duplication and remolding in mitochondrial genomes of amphipods. *Mol Phylogenet Evol*. 2020;144:106710. doi 10.1016/j.ympev.2019.106710

Soucy S.M., Huang J., Gogarten J.P. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 2015;16(8):472-482. doi 10.1038/nrg3962

Sprinzl M., Horn C., Brown M., Ioudovitch A., Steinberg S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*. 1998;26(1):148-153. doi 10.1093/nar/26.1.148

Velandia-Huerto C.A., Berkemer S.J., Hoffmann A., Retzlaff N., Romero Marroquín L.C., Hernández-Rosales M., Stadler P.F., Bermúdez-Santana C.I. Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genomics*. 2016;17(1):617. doi 10.1186/s12864-016-2927-4

Weiss M.C., Sousa F.L., Mrnjavac N., Neukirchen S., Roettger M., Nelson-Sathi S., Martin W.F. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*. 2016;1(9):16116. doi 10.1038/nmicrobiol.2016.116

Conflict of interest. The author declares no conflict of interest.

Received September 13, 2025. Revised October 7, 2025. Accepted October 7, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1129-1136

doi 10.18699/vjgb-25-117

Assessing the dependence of brain activity on individual single-nucleotide variability of genetic markers of major depressive disorder using principal component analysis

K.A. Zorina [1], A.A. Kriveckiy [4], V.S. Karmanov [4], A.N. Savostyanov [1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[4] Novosibirsk State Technical University, Novosibirsk, Russia

a.savostianov@g.nsu.ru

Abstract. Major depressive disorder (MDD) is one of the most widespread mental illnesses, which necessitates the search for factors of increased predisposition to this disorder. Single nucleotide polymorphisms in genes of the brain's neurotransmitter systems are often considered as molecular genetic markers of MDD. Indicators of individual single nucleotide variability in neurotransmitter genes are used to assess the risk of MDD before its symptomatology at the behavioral level. However, the predictive capabilities of analyzing genomic variations to assess the risk of depression are not yet sufficiently reliable and are complemented by behavioral and neurophysiological information about patients. Neurophysiological markers of MDD provide the most reliable estimates of the severity of pathological symptoms, but they reflect a person's state at the time of examination, and not a predisposition to the occurrence of this pathological state and do not allow assessing the risk of its appearance in the future. Major depressive disorder is often accompanied by abnormalities in a person's ability to control motor responses, including the ability to voluntary suppress inappropriate behavior. The "stop-signal paradigm" (SSP) is an experimental method for assessing the functional balance between the inhibitory and activation systems of the brain during targeted movements. Combined with EEG recording, this experimental method allows for the consideration of not only participants' behavioral characteristics, such as speed or accuracy of responses, but also the brain's neurophysiological features associated with behavior control. The objective of this study was to evaluate the relationship between EEG responses in the stop-signal paradigm and individual single nucleotide variability in candidate genes for MDD detection. Dimensionality in the original genetic and neurophysiological experimental data was reduced by principal component analysis (PCA) to subsequently detect an association between EEG response components recorded during the control of random motor responses and single nucleotide variations in genes, the variability of which is associated with MDD risk. Variability in these genes has been shown to be associated with the amplitude of brain responses under the conditions of test subjects using the PCA method. The results obtained can be used to develop systems for the early diagnosis of depression, identify individual patterns of impairment in the brain, select methods for correcting the disease and control the effectiveness of therapy.

Key words: stop-signal paradigm; EEG; event related potentials; single nucleotide polymorphisms (SNPs); major depressive disorder; principal component analysis; regression analysis

For citation: Zorina K.A., Kriveckiy A.A., Karmanov V.S., Savostyanov A.N. Assessing the dependence of brain activity on individual single-nucleotide variability of genetic markers of major depressive disorder using principal component analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1129-1136. doi 10.18699/vjgb-25-117

Funding. This research was funded by Budget Project No. FWNR-2022-0020.

Оценка зависимости показателей мозговой активности от индивидуальной однонуклеотидной вариабельности генетических маркеров большого депрессивного расстройства с использованием анализа главных компонент

К.А. Зорина [1], А.А. Кривецкий [4], В.С. Карманов [4], А.Н. Савостьянов [1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[4] Новосибирский государственный технический университет, Новосибирск, Россия

a.savostianov@g.nsu.ru

Аннотация. Большое депрессивное расстройство (БДР) относится к наиболее широко распространенным психическим заболеваниям, что обусловливает необходимость поиска факторов повышенной предрасположенности к этому нарушению. В качестве молекулярно-генетических маркеров БДР часто рассматривают однонуклеотидные полиморфизмы генов нейромедиаторных систем мозга. Показатели индивидуальной однонуклеотидной вариабельности в генах нейромедиаторов применяются для оценки риска появления БДР до проявления его симптоматики на поведенческом уровне. Однако прогностические возможности анализа геномных вариаций для оценки риска депрессии до настоящего времени недостаточно надежны и дополняются поведенческой и нейрофизиологической информацией о пациентах. Нейрофизиологические маркеры БДР дают наиболее надежные оценки выраженности патологической симптоматики, но они отражают состояние человека в момент обследования, а не предрасположенность к возникновению этого патологического состояния и не позволяют выполнить оценку риска его появления в будущем. Большое депрессивное расстройство часто сопровождается отклонениями в способности человека контролировать двигательные реакции, включая возможность произвольно подавлять неадекватное поведение. «Стоп-сигнал парадигма» (ССП) – экспериментальный метод для оценки функционального баланса между тормозными и активационными системами головного мозга в условиях выполнения целенаправленных движений. Объединенный с регистрацией ЭЭГ, этот экспериментальный метод позволяет учитывать не только поведенческие характеристики участников, такие как скорость или точность ответов, но и нейрофизиологические особенности головного мозга, ассоциированные с контролем над поведением. Цель настоящего исследования заключалась в оценке зависимости между особенностями ЭЭГ реакций в условиях парадигмы стоп-сигнал и индивидуальной однонуклеотидной вариабельностью в генах-кандидатах для выявления БДР. Размерность в исходных генетических и нейрофизиологических экспериментальных данных была снижена при помощи анализа главных компонент (PCA) для последующего выявления ассоциации между компонентами ЭЭГ реакций, регистрируемыми в условиях контроля произвольных двигательных реакций, и однонуклеотидными вариациями в генах, изменчивость которых ассоциирована с риском БДР. Установлено, что изменчивость в этих генах ассоциирована с амплитудными показателями мозговых ответов в условиях тестирования испытуемых методом ССП. Полученные результаты могут быть использованы для разработки систем ранней диагностики депрессии, выявления индивидуальных паттернов нарушения в работе головного мозга, подбора методов коррекции заболевания и контроля над эффективностью терапии.
Ключевые слова: стоп-сигнал парадигма; ЭЭГ; вызванные потенциалы; однонуклеотидные полиморфизмы; большое депрессивное расстройство; метод главных компонент; регрессионный анализ

Introduction

Major depressive disorder (MDD), also known as clinical depression, is a psychiatric disorder characterized by symptoms including depressed mood, loss of interest or pleasure in previously enjoyable activities, fatigue or loss of energy, alterations in sleep and appetite, difficulties with concentration and memory, as well as feelings of guilt and low self-esteem (DSM-5, 2013). MDD ranks among the most prevalent psychiatric disorders (Wong, Licinio, 2001). Susceptibility to various forms of depressive disorders is known to depend on both genetic factors and individual life experiences, particularly during the period preceding the onset of MDD symptoms (Cross-Disorder Group, 2013; Northoff, 2013; Haase, Brown, 2015; Ivanov et al., 2019; Whitney et al., 2019). For many years, the monoamine theory of depression was considered the most plausible, and allelic polymorphisms in genes encoding components of the brain's monoaminergic neurotransmitter systems have frequently been investigated as molecular markers of depression susceptibility (Willner et al., 2013). However, attempts to predict depression risk based solely on genetic data have generally proven unsatisfactory (Duncan et al., 2014; Halldorsdottir, Binder, 2017), as depression is a multifactorial disorder arising from the interplay of multiple genetic and environmental factors (Ivanov et al., 2019; Wang et al., 2025). Consequently, the identification of reliable biomarkers for depression necessitates the concurrent use of not only genetic but also neurophysiological indicators reflecting the functional state of the human brain.

Neurophysiological markers of depression may include the amplitude and latency of event-related potentials (ERPs) derived from electroencephalography (EEG) (Stone et al., 2025). It is well established that depression is frequently associated with impairments in inhibitory control, manifesting at both behavioral and neurophysiological levels (Shetty et al., 2025). An example of a method used to assess individual capacity for behavioral self-control is the stop-signal paradigm (SSP) (Band et al., 2003). This experimental paradigm provides an objective measure of the functional balance between brain activation systems that govern goal-directed actions and inhibitory systems responsible for suppressing inappropriate behavior.

A major challenge in the comprehensive investigation of depression lies in the need to account for a large number of variables, the interrelationships of which are not initially evident to the researcher. This challenge can be addressed through the application of dimensionality reduction techniques designed to uncover latent dependencies among factors. In particular, principal component analysis (PCA) is widely employed to reduce the dimensionality of original datasets and to identify the most informative features (Gewers et al., 2021). PCA transforms the original variables into a lower-dimensional space, thereby reducing the number of parameters under analysis and mitigating redundancy inherent in high-dimensional data (Subasi, Gursoy, 2010).

The aim of the present study was to investigate the association between neurophysiological measures recorded during the stop-signal paradigm and individual single-nucleotide variability in genes linked to an elevated risk of depression.

In this work, we analyzed genetic and neurophysiological data obtained from the publicly available ICBrainDB, developed by researchers at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), and the Institute of Neuroscience and Medicine, and hosted on the ICG SB RAS website (Ivanov et al., 2022). Candidate genes for MDD had been previously selected

through a bioinformatic analysis of scientific publications retrieved from open-access databases containing information on depressive spectrum disorders diagnosed in the studied individuals (Ivanov et al., 2019).

Materials and methods

Participant sample. The sample comprised 212 individuals for whom both genomic and EEG data were analyzed. Among them, 47 participants residing in Novosibirsk had a clinically diagnosed major depressive disorder, while 165 participants had no diagnosed depression; of these, 67 resided in Novosibirsk, 50 in Yakutsk, and 48 in Khandyga, Sakha Republic.

Experimental design. Participants performed a series of tasks in a stop-signal paradigm modified by A.N. Savostyanov and colleagues (2009). During the task, one of two visual stimuli was presented on the screen; upon the appearance of the target stimulus, participants were required to press a button on the keyboard. On a subset of trials, a stop-signal appeared shortly after the target stimulus, instructing the participant to abort the already initiated motor response. Across the experiment, each participant completed 135 trials, 35 of which included a stop-signal. EEG was recorded using a 128-channel NVX-132 amplifier. Electrodes were positioned according to the international 10-5 system, with AFz serving as the ground electrode and Cz as the reference. The signal bandwidth was set between 0.3 and 100 Hz, and the sampling rate was 1,000 Hz.

EEG signal processing. Raw EEG recordings contained non-neural noise, including ocular movement artifacts, facial muscle activity, cardiac electrical activity, and vascular artifacts. All non-neural artifacts were removed using independent component analysis (ICA), implemented in the EEGLAB toolbox (Delorme, Makeig, 2004). ICA is a computational algorithm that decomposes multichannel data into statistically independent components. In contrast, PCA identifies components characterized by high mutual dependence.

From the preprocessed EEG data, two types of epochs were extracted: go-epochs (intervals of brain activity time-locked to the participant's button press following the target visual stimulus) and stop-epochs (intervals corresponding to successful inhibition of the motor response after stop-signal presentation). Epoching for go-trials was performed relative to the onset of the target stimulus, whereas for stop-trials it was aligned to the onset of the stop-signal. Within go-epochs, two distinct EEG peaks were identified: a premotor peak (400–600 ms post-stimulus) and a postmotor peak (700–800 ms post-stimulus). The premotor peak reflects brain activity associated with motor preparation, whereas the postmotor peak corresponds to neural processes occurring during movement execution.

In stop-epochs, two additional peaks were identified, either preceding or following the suppression of the motor response. These peaks and their corresponding time windows were determined based on visual inspection of event-related potential (ERP) waveforms recorded at electrode C3, which overlies the motor cortex of the left hemisphere. Using the ERPLAB toolbox (Lopez-Calderon, Luck, 2014), for each of these peaks, the following quantitative measures were computed separately for each participant and each EEG channel: peak maximum amplitude, mean amplitude within the peak window, and peak latency. Since each participant completed 100 go-trials and 35 stop-trials during the experiment, brain responses were averaged across trials for each participant. EEG channels were grouped into 12 spatially defined regions: left frontal, medial frontal, right frontal, left temporal, left central, medial central, right central, right frontal, left parietal, medial parietal, right parietal, and a combined occipital group. Consequently, the initial EEG dataset comprised 144 parameters: 12 (electrode groups) × 2 (ERP peaks) × 3 (quantitative measures: maximum amplitude, mean amplitude, latency) × 2 (experimental conditions: go or stop).

Genetic data. Genetic material, collected as either whole blood or buccal epithelial cells, was obtained from all participants. Targeted sequencing of 164 genes was performed using this material. These genes were selected based on prior reconstruction and analysis of a gene network associated with susceptibility to MDD (Ivanov et al., 2019). Targeted sequencing libraries were prepared for these 164 genes, and high-coverage next-generation sequencing (NGS) was conducted for all participants. For each allele of every gene in the list, a binary variability index was assigned for each participant relative to the reference genome (Ivanov et al., 2022). If a participant's allele sequence matched the reference genome exactly, the variability index was set to 0; if one or more nucleotide substitutions were present, the index was set to 1 (regardless of the number of substitutions within the allele). Across all participants, 799 single-nucleotide polymorphisms were identified in 121 of the 164 sequenced genes. No nucleotide substitutions were detected in any participant for the remaining 43 genes. Thus, the total number of input genetic parameters was 242 (121 genes × 2 alleles per gene).

Results

As previously stated, the aim of this study was to assess the association between EEG responses recorded during the stop-signal paradigm and individual single-nucleotide variability in candidate genes linked to MDD risk. To achieve this objective, a multi-stage analysis of the experimental data was conducted, and the results are presented below.

Task 1. Identification of MDD candidate genes exhibiting significant associations between single-nucleotide variability and EEG measures

To address Task 1, a series of linear models was constructed, wherein each EEG parameter served as a dependent variable and the binary indicator of the presence or absence of single-nucleotide variants (SNVs) in a specific gene served as the independent variable. The term "linear model series" refers to separate linear regression analyses performed for each unique pair of "EEG parameter – single-nucleotide variability" (Table 1). Given 144 EEG parameters and 242 genetic parameters, the initial number of parameter pairs subjected to linear regression totaled 34,848. An individual linear regression model was formulated as follows:

$$EEG\_parameter = B_0 + B_1 + e.$$

Here, B_1 represents the binary predictor coded as 0 (no nucleotide substitution in the allele) or 1 (at least one substitution present).

The dependent variable was a quantitative EEG measure, while the predictor was the binary indicator of nucleotide substitution presence in a given gene allele. If at least one

Table 1. Example of a parameter pair used in linear regression analysis.
The first parameter is individual variability in the *ADRA2B* gene; the second is the amplitude of the premotor ERP peak in the right parietal cortex

Participant ID	*ADRA2*B gene (0 – no variability; 1 – variability present)	Amplitude of the postmotor EEG peak in the "go" condition in the right parietal cortex, uV
D_Nov_001	0	1.68
D_Yak_2016_001	1	8.05

substitution was present in one allele, the binary indicator was assigned a value of 1. The two alleles of the same gene were treated as two distinct binary predictors. This approach enabled testing whether single-nucleotide variability in each candidate gene was associated with alterations in a given EEG parameter.

In addressing Task 1, multiple comparisons were corrected using the Benjamini–Hochberg procedure (False discovery rate, FDR) to control the expected proportion of false rejections of the null hypothesis (Benjamini, Hochberg, 1995). The FDR method is more statistically powerful than the Bonferroni correction and is particularly advantageous when the number of tested hypotheses is large or when minimizing false positives is prioritized over strict per-hypothesis control of Type I error.

Associations were tested between all 144 EEG measures and variability in each of the 121 genes in which at least one SNV was detected in at least one participant. This analysis revealed statistically significant associations (FDR-corrected significance threshold $q < 0.05$) for only five genes – *ADRA2B*, *TF*, *HCRTR2*, *WFS1*, and *PENK* – and four EEG measures recorded during go-epochs in the medial frontal, right parietal, left parietal, and combined occipital cortical regions (Table 2). Notably, significant associations for three genes (*ADRA2B*, *TF*, *HCRTR2*) were observed across three cortical regions (right parietal, left parietal, and occipital), whereas for the remaining two genes (*WFS1* and *PENK*), significant associations were confined to the medial frontal cortex. These five genes were subsequently included in further analyses.

Table 2 summarizes the linear regression results linking EEG measures to polymorphisms in MDD candidate genes. It lists 11 most significant "gene–EEG measure" pairs with the lowest FDR-corrected *p*-values (*q*-values), along with their uncorrected *p*-values. All reported associations are significant at FDR < 0.05.

The average frequency of single-nucleotide variants for each of the five selected genes across the entire participant sample is presented in Table 3. The prevalence of variant carriers for these genes ranged from approximately one-third to two-thirds of participants, ensuring sufficient variability for robust statistical analysis.

Task 2. Dimensionality reduction of neurophysiological data using principal component analysis

In addressing Task 2, PCA with prior feature standardization was applied to reduce the dimensionality of the EEG dataset (Rokhlin et al., 2010). From the original set of 144 EEG variables, 15 principal components were extracted. The Figure demonstrates that these 15 components collectively account for approximately 80 % of the total variance in the original EEG parameters, thereby capturing the majority of inter-individual variability.

Task 3. Assessment of the influence of variability in MDD candidate genes on integrated measures of brain activity derived from PCA

In Task 3, for each of the five selected genes showing statistically significant associations with specific EEG measures (Table 2), a regression analysis was performed between the principal components (PCs) and the binary indicators of polymorphism presence. Unlike in Task 1, where regression was conducted on individual EEG parameters, here the analysis was performed on integrated composite measures (the principal components) that collectively explain 80 % of the total inter-individual variance in the EEG data (see the Figure).

Among the 15 PCA-derived components of brain activity, only the third principal component (PC3) exhibited a statistically significant association with genetic variability in the MDD candidate genes. This finding is summarized in Table 4, which presents the results of statistical significance testing for the effects of genetic variability in the five candidate genes on the three most informative PCA components.

To provide a neurophysiological interpretation of the observed associations, factor loadings for the third principal component (PC3) were computed for each of the original EEG measures. In the context of PCA, a factor loading represents the correlation coefficient between an original variable and a principal component, indicating the strength and direction of their association. The factor loadings of the original EEG measures for PC3 are presented in Table 5. As evident from these results, PC3 is most strongly associated with brain activity in occipito-parietal cortical regions and, to a somewhat lesser extent, with frontal cortical activity. This cortical topography is characteristic of functional processes involved in attentional control during visual stimulus recognition. Furthermore, it is apparent that both premotor and postmotor ERP peaks-across both go- and stop-episodes contributed most substantially to this component.

Task 4. Prediction of candidate gene variability based on composite EEG measures (solving the inverse problem)

To address this task, logistic regression with L1 regularization (Flach, 2016) was employed to predict the presence or absence of single-nucleotide variants in MDD candidate genes using the first 15 EEG-derived principal components (PC1–PC15). Unlike linear regression, which models continuous dependent variables, logistic regression is designed for binary outcomes.

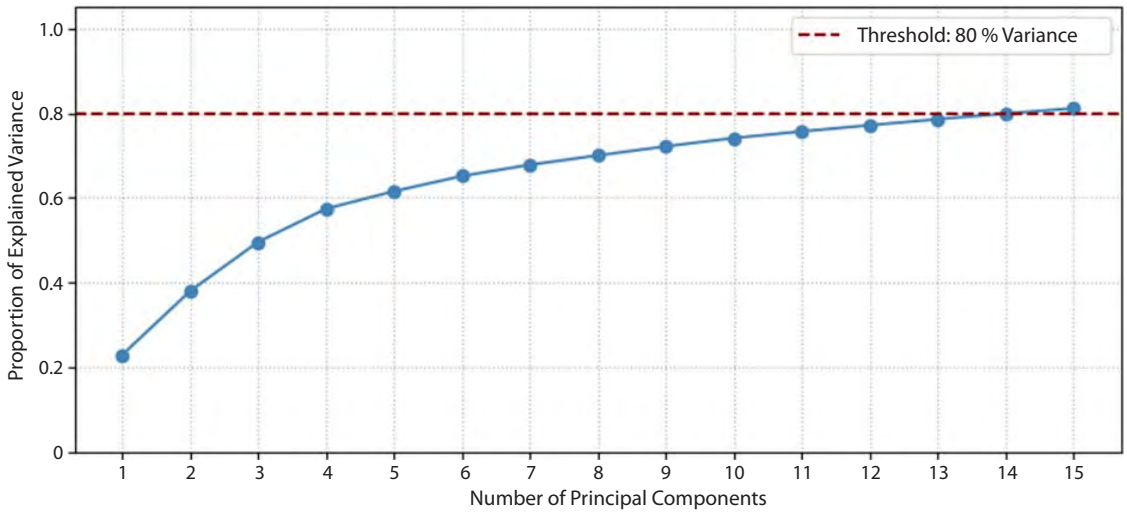
Table 2. Results of the association analysis between the amplitude of the postmotor ERP peak in go-episodes of the stop-signal paradigm and binary variability in MDD candidate genes

Gene with identified variability	Cortical region in which EEG responses depended on gene variability	Significance level (*p*-value)	FDR-corrected significance level (*q*-value)
ADRA2B	right parietal	7.35E-34	1.15E-29
	left parietal	9.88E-26	1.03E-22
	occipital	2.65E-28	5.91E-25
TF	right parietal	1.48E-32	7.70E-29
	left parietal	1.66E-26	1.85E-23
	occipital	1.34E-29	4.17E-26
HCRTR2	right parietal	1.48E-32	7.70E-29
	left parietal	1.66E-26	1.85E-23
	occipital	1.34E-29	4.17E-26
WFS1	medial frontal	4.93E-27	8.53E-24
PENK	medial frontal	5.44E-28	1.06E-24

Table 3. Mean number of single-nucleotide variants in selected MDD* candidate genes

MDD candidate gene	Mean variability**	Standard deviation (Std)	Percentage of individuals with no substitutions in this gene
ADRA2B	0.29	0.45	70.54
TF	0.34	0.47	65.75
HCRTR2	0.34	0.47	65.75
WFS1	0.65	0.48	34.93
PENK	0.63	0.48	36.98

* Data are shown only for genes exhibiting significant associations between genetic variability and EEG measures. ** In this context, mean values represent the proportion of participants in the sample who carried at least one nucleotide substitution in the respective gene.



Cumulative variance explained by principal component analysis of EEG data.
The red dashed line indicates the 80 % variance threshold.

In our case, the logistic model aimed to estimate the probability of genetic variability in MDD candidate genes based solely on EEG-derived features, thus constituting the inverse problem. The input features consisted of the first 15 principal components extracted from the original EEG parameter space, while the target variables were binary indicators of deviation from the human reference genome in the five genes previously shown to exhibit significant associations with EEG components: *ADRA2B*, *WFS1*, *PENK*, *TF*, and *HCRTR2*. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), computed via 5-fold stratified cross-validation. The accuracy estimates

(AUC values) and their standard deviations across the five cross-validation folds are presented in Table 6. As shown in Table 6, prediction accuracy for binary genetic variability in three of the five candidate genes ranged from 0.73 to 0.78, with standard deviations between 0.13 and 0.18. These results indicate that the presence of binary variability in MDD candidate genes can be predicted from EEG data recorded during the stop-signal paradigm with 70–80 % accuracy, thereby providing convergent evidence for a robust link between genetic susceptibility and neurophysiological phenotypes.

Thus, the sequential accomplishment of the four objectives of our study enabled us to: (1) identify a list of candidate genes for MDD, the variability of which is associated with measures of brain activity during behavioral control; (2) determine composite characteristics of brain activity accounting for 80 % of the variance in EEG data; (3) identify an integrated measure of brain activity most robustly associated with single-nucleotide variability in MDD candidate genes; and (4) solve the inverse problem by predicting variability in MDD candidate genes based on EEG-derived measures.

Discussion

A fundamental challenge in identifying candidate genes for most psychiatric disorders is that the behavioral effects of single-nucleotide variations in any individual selected gene are relatively weak (Duncan et al., 2014). Depression exemplifies a disorder for which no direct and unambiguous associations with specific g-enes have been established (Halldorsdottir, Binder, 2017). This suggests that the genetic underpinnings of depression are highly heterogeneous across individuals and cannot be reduced to a small set of genes and their mutations.

This has motivated a shift in focus from analyzing the contribution of individual genes or mutations toward investigating interconnected complexes of genes, their protein products, and metabolites. Such gene complexes are referred to as "gene networks" (Kolchanov et al., 2013). A gene network may encompass dozens to hundreds of genes, along with the multitude of proteins and metabolites they encode. Previously, using bioinformatic approaches, fragments of a gene network implicated in susceptibility to major depressive disorder (MDD) were reconstructed (Ivanov et al., 2019). In the same study, a comprehensive dataset was assembled, integrating psychometric, neurophysiological, and genetic data reflecting the analysis of SNPs across 164 genetic loci incorporated into the depression-related gene network (Ivanov et al., 2022). The aim of the present study was to identify genes associated not only with psychometric traits but also with neurophysiological characteristics of brain activity, which may likewise be considered as manifestations of depression.

Table 4. Results of linear regression between the first three EEG principal components (PC1–PC3) and variability in the five MDD candidate genes*

Gene, the variability of which influenced brain activity	Significance level (*p*-value)
PC3	
WFS1	**0.0055**
TF	**0.0065**
HCRTR2	**0.0065**
PENK	**0.0065**
ADRA2B	**0.0258**
PC1	
ADRA2B	0.3297
TF	0.2844
HCRTR2	0.2844
WFS1	0.2876
PENK	0.2844
PC2	
TF	0.3109
HCRTR2	0.3109
WFS1	0.3028
PENK	0.3109
ADRA2B	0.3933

* Results are ordered by the significance level of the linear regression.

Table 5. Factor loadings of original brain activity measures for PC3

EEG parameter	Factor loading* for PC3 ($p < 0.05$)
Occipital cortex, postmotor peak, stop-episodes	0.24
Right parietal cortex, premotor peak, stop-episodes	0.22
Left parietal cortex, premotor peak, stop-episodes	0.21
Right frontal cortex, postmotor peak, stop-episodes	0.19
Occipital cortex, postmotor peak, go-episodes	0.19
Right parietal cortex, postmotor peak, go-episodes	0.18
Medial parietal cortex, postmotor peak, go-episodes	0.17
Medial parietal cortex, premotor peak, stop-episodes	0.17
Left frontal cortex, postmotor peak, stop-episodes	0.17
Left parietal cortex, postmotor peak, go-episodes	0.16

* Factor loading denotes the correlation coefficient between an EEG measure and the integrated score of PC3.

Table 6. Results of logistic regression for predicting the presence of mutations using the 15 EEG principal components

Gene	AUC – measure of the accuracy in predicting variability	Standard deviation
WFS1	0.78	0.15
PENK	0.76	0.18
TF	0.75	0.13
HCRTR2	0.75	0.13
ADRA2B	0.73	0.15

Behavioral control is one of the core cognitive functions in humans, and its impairment constitutes a symptom of numerous neuropsychiatric disorders. In the present study, we analyzed the relationship between parameters of human ERPs and the presence of single-nucleotide variations in candidate genes for MDD within a combined sample comprising both healthy individuals and those diagnosed with depressive disorder. Our results demonstrate that the amplitude of the postmotor positivity in go-trials of the stop-signal paradigm is associated with binary variability in five MDD candidate genes: *ADRA2B*, *TF*, *HCRTR2*, *WFS1*, and *PENK* (Table 2).

Associations with genetic variability were observed not only for several localized EEG measures reflecting cortical activity in specific brain regions during brief phases of task performance but also for an integrated measure of global brain activity derived via PCA, which captures more general features of the nervous system's functional state (Table 4). This integrated brain activity measure significantly influenced by genetic variability reflects the engagement of cortical regions involved in visual signal perception and voluntary attentional control (Table 5). Furthermore, we demonstrated that these integrated EEG measures can serve as predictors of single-nucleotide variability in MDD candidate genes with 70–80 % accuracy when applying logistic regression (Table 6), thereby indicating the feasibility of solving the inverse problem: predicting genetic variability from neurophysiological data.

Additional findings from our prior work indicate that ERP amplitudes during performance of the stop-signal paradigm are positively correlated with the severity of depressive symptoms (Zorina et al., 2025). Thus, a coherent link emerges between specific genes, the variability of which is associated both with depression at the behavioral level and with a neurophysiological marker of elevated depressive symptomatology. Information from Ivanov et al. (2019) further clarifies the biological roles of these genes: (a) *ADRA2B* encodes the alpha-2B adrenergic receptor, a member of the G protein-coupled receptor family; (b) *TF* encodes transferrin; (c) *HCRTR2* encodes hypocretin (orexin) receptor type 2; (d) *WFS1* encodes wolframin; and (e) *PENK* encodes the proenkephalin precursor protein. Our new findings indicate that variability in these MDD candidate genes is associated with brain activity parameters reflecting an individual's capacity for behavioral self-control – a function impaired in MDD – thereby supporting the existence of a composite genetic-neurophysiological marker linked to depression risk.

Conclusion

The present analysis revealed statistically significant associations between polymorphisms in the *ADRA2B*, *TF*, *HCRTR2*, *WFS1*, and *PENK* genes and EEG signal characteristics recorded during performance of the stop-signal paradigm. Principal component analysis effectively reduced data dimensionality and enabled the identification of the most informative indices of integrated brain activity. Logistic regression models demonstrated that EEG-derived parameters can predict, with moderate accuracy, the presence of single-nucleotide substitutions in MDD candidate genes. These results may facilitate the assessment of complex interdependencies between genetic and neurophysiological markers associated with depression.

Limitations. This study did not specifically evaluate differences between clinically diagnosed patients with depression and healthy participants. A more detailed comparison of the identified associations between neurophysiological and molecular biological markers of depression remains an objective for future, more granular analyses currently planned in our ongoing research.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA: American Psychiatric Publishing, 2013;87-122. doi 10.1176/appi.books.9780890425596

Band G.P.H., van der Molen M.W., Logan G.D. Horse-race model simulations of the stop-signal procedure. *Acta Psychol.* 2003; 112(2):105-142. doi 10.1016/s0001-6918(02)00079-3

Benjamini Y., Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B.* 1995; 57(1):289-300. doi 10.1111/j.2517-6161.1995.tb02031.x

Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a Genome-wide analysis. *Lancet.* 2013;381(9875):1371-1379. doi 10.1016/s0140-6736(12)62129-1

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neuroscience Methods.* 2004;134(1):9-21. doi 10.1016/j.jneumeth.2003.10.009

Duncan L.E., Pouastri A.R., Smoller J.W. Mind the gap: Why many geneticists and psychological scientists have discrepant views about gene–environment interaction (G×E) research. *Am Psychol.* 2014; 69(3):249-268. doi 10.1037/a0036320

Flach P.A. ROC Analysis. In: Sammut C., Webb G. (Eds). Encyclopedia of Machine Learning and Data Mining. Springer, 2016. doi 10.1007/978-1-4899-7502-7_739-1

Gewers F.L., Ferreira G.R., de Arruda H.F., Silva F.N., Comin C.H., Amancio D.R., da Costa L.F. Principal component analysis: A natural approach to data exploration. *ACM Comput Surv.* 2021;54(4):70. doi 10.1145/3447755

Haase J., Brown E. Integrating the monoamine, neurotrophin and cytokine hypotheses of depression: A central role for the serotonin transporter? *Pharmacol Ther.* 2015;147:1-11. doi 10.1016/j.pharmthera.2014.10.002

Halldorsdottir T., Binder E.B. Gene × environment interactions: From molecular mechanisms to behavior. *Annu Rev Psychol.* 2017;68: 215-241. doi 10.1146/annurev-psych-010416-044053

Ivanov R., Zamyatin V., Klimenko A., Matushkin Y., Savostyanov A., Lashin S. Reconstruction and analysis of gene networks of human neurotransmitter systems reveal genes with contentious manifestation for anxiety, depression, and intellectual disabilities. *Genes.* 2019;10(9):699. doi 10.3390/genes10090699

Ivanov R., Kazantsev F., Zavarzin E., Klimenko A., Milakhina N., Matushkin Y.G., Savostyanov A., Lashin S. ICBrainDB: An integrated

database for finding associations between genetic factors and EEG markers of depressive disorders. *J Pers Med*. 2022;12(1):53. doi 10.3390/jpm12010053

Kolchanov N.A., Ignatyeva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2013;17(4/2):833-850 (in Russian)

Lopez-Calderon J., Luck S.J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci*. 2014;8:213. doi 10.3389/fnhum.2014.00213

Northoff G. Gene, brains, and environment – genetic neuroimaging of depression. *Curr Opin Neurobiol*. 2013;23(1):133-142. doi 10.1016/j.conb.2012.08.004

Rokhlin V., Szlam A., Tygert M. A randomized algorithm for principal component analysis. *SIAM J Matrix Anal Appl*. 2010;31(3):1100-1124. doi 10.1137/080736417

Savostyanov A.N., Tsai A.C., Liou M., Levin E.A., Lee J.D., Yurganov A.V., Knyazev G.G. EEG-correlates of trait anxiety in the stop-signal paradigm. *Neurosci Lett*. 2009;449(2):112-116. doi 10.1016/j.neulet.2008.10.084

Shetty T., Kashyap H., Mehta U.M., Binu V.S. Executive function and emotion regulation in depressive and anxiety disorders: A cross-sectional study. *Indian J Psychol Med*. 2025. doi 10.1177/02537176251340586

Stone B., Desrochers P.C., Nateghi M., Chitadze L., Yang Y., Cestero G.I., Bouzid Z., … Bremner J.D., Inan O.T., Sameni R., Lynn S.K., Bracken B.K. Decoding depression: Event related potential dynamics and predictive neural signatures of depression severity *J Affect Disord*. 2025;391:119893. doi 10.1016/j.jad.2025.119893

Subasi A., Gursoy M.I. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Syst Appl*. 2010;37(12):8659-8666. doi 10.1016/j.eswa.2010.06.065

Wang Z., Zellers S., Piirtola M., Aaltonen S., Salvatore J., Dick D., Kuhn S., Kaprio J. Gene-environment interplay in the relationship between the visibility of the environment and self-reported depression in early midlife: a Finnish twin cohort study. *medRxiv*. 2025. doi 10.1101/2025.05.23.25328215

Whitney D.G., Shapiro D.N., Peterson M.D., Warschausky S.A. Factors associated with depression and anxiety in children with intellectual disabilities. *J Intellect Disabil Res*. 2019;63(5):408-417. doi 10.1111/jir.12583

Willner P., Scheel-Krüger J., Belzung C. The neurobiology of depression and antidepressant action. *Neurosci Biobehav Rev*. 2013;37(10 Pt. 1):2331-2371. doi 10.1016/j.neubiorev.2012.12.007

Wong M.L., Licinio J. Research and treatment approaches to depression. *Nat Rev Neurosci*. 2001;2(5):343-351. doi 10.1038/35072566

Zorina K.A., Kriveckiy A.A., Klemeshova D.I., Bocharov A.V., Karmanov V.S. Using machine learning methods to search for EEG and genetic markers of depressive disorder. In: IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation. 2025;1790-1793. doi 10.1109/EDM65517.2025.11096763

Conflict of interest. The authors declare no conflict of interest.

Received June 25, 2025. Revised September 10, 2025. Accepted September 11, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1137-1144

doi 10.18699/vjgb-25-118

Senescent cell accumulation is associated with T-cell imbalance in the skin

K.S. Matveeva, S.K. Kolmykov, T.S. Sokolova, D.R. Salimov, D.V. Shevyrev

Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia
dr.daniil25@mail.ru

Abstract. Organismal aging is accompanied by the accumulation of senescent cells – damaged, non-functional cells that exhibit cell cycle arrest, resistance to apoptosis, metabolic dysfunction, and production of a wide range of pro-inflammatory substances. The age-related accumulation of these cells is associated with impaired tissue function, contributes to chronic inflammation (inflammaging), and promotes the development of various age-associated diseases. Conversely, the elimination of senescent cells restores tissue functions and positively affects overall metabolism. Under normal conditions, senescent cells are removed by the innate immune system; however, the efficiency of this process declines with age. The involvement of adaptive immunity and the role of T cells in the clearance of senescent cells remain poorly understood. The aim of this study was to identify alterations in local T cell immunity associated with the accumulation of senescent cells in human skin. The analysis was performed on publicly available single-cell RNA-sequencing data from skin biopsies, and the senescent status was assessed using the SenePy algorithm with Gaussian mixture models. It was found that the emergence of senescent cells occurs heterogeneously across cell types within the tissue. The accumulation of these cells is associated with alterations in the $CD4^+$ to $CD8^+$ T cell ratio, as well as with an increased abundance of regulatory T cells. Functional analysis revealed that these quantitative age-related shifts were accompanied by more pronounced activation of regulatory T cells together with features of anergy and exhaustion in $CD8^+$ T cells, whereas functional changes in $CD4^+$ T cells were heterogeneous. These findings underscore the importance of adaptive immunity in maintaining tissue homeostasis and suggest potential age-related dysfunction of tissue-resident T cells. Understanding the mechanisms underlying the interaction between adaptive immunity and senescent cells is crucial for the development of senolytic vaccines and other immunological approaches aimed at enhancing endogenous elimination of senescent cells.

Key words: senescence; adaptive immunity; regulatory T cells; single-cell transcriptome; aging; genetic signatures; tissue-resident T cells; senescent cell elimination; skin

For citation: Matveeva K.S., Kolmykov S.K., Sokolova T.S., Salimov D.R., Shevyrev D.V. Senescent cell accumulation is associated with T-cell imbalance in the skin. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7): 1137-1144. doi 10.18699/vjgb-25-118

Funding. This work was supported by Russian Scientific Foundation, project No. 24-15-20003, https://rscf.ru/project/24-15-20003/

Старение кожи связано с локальным дисбалансом в Т-клеточном иммунитете

К.С. Матвеева, С.К. Колмыков, Т.С. Соколова, Д.Р. Салимов, Д.В. Шевырев

Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия
dr.daniil25@mail.ru

Аннотация. Старение организма сопровождается накоплением поврежденных нефункциональных клеток, которые называют сенесцентными. Эти клетки находятся в состоянии ареста клеточного цикла, устойчивы к апоптозу, имеют нарушенный метаболизм, а также продуцируют широкий спектр провоспалительных факторов – цитокинов, хемокинов, протеаз, молекул адгезии и продуктов арахидонового каскада. Накопление таких клеток с возрастом связано с нарушением функций тканей, способствует хроническому воспалению (inflammaging) и развитию различных возраст-ассоциированных заболеваний. В свою очередь, элиминация сенесцентных клеток восстанавливает тканевые функции и позитивно сказывается на общем метаболизме. В норме сенесцентные клетки удаляются системой врожденного иммунитета, однако с возрастом эффективность этого процесса падает. При этом участие адаптивного иммунитета и роль Т-лимфоцитов в удалении сенесцентных клеток остаются неизученными. Целью исследования был поиск изменений в локальном Т-клеточном иммунитете, которые связаны с накоплением сенесцентных клеток в

коже человека. Анализ проводился на открытых данных РНК секвенирования единичных клеток биоптатов кожи. Сенесцентный статус клеток оценивали при помощи алгоритма SenePy с применением смешанных гауссовских моделей. Было выявлено, что появление клеток с выраженными признаками сенесцентности в пределах ткани происходит неравномерно среди клеточных типов. Накопление этих клеток ассоциировано с изменением соотношения популяций $CD4^+$ и $CD8^+$ лимфоцитов, а также сопряжено с увеличением содержания регуляторных Т-лимфоцитов. В ходе функционального анализа обнаружено, что данные количественные изменения с возрастом сопровождаются более выраженной активацией регуляторных Т-лимфоцитов совместно с анергией и истощением $CD8^+$ лимфоцитов, тогда как функциональные изменения $CD4^+$ лимфоцитов имеют гетерогенный характер. Полученные результаты подчеркивают значение адаптивного иммунитета в поддержании тканевого гомеостаза и указывают на потенциальную дисфункцию эффекторных тканевых Т-лимфоцитов, которая возникает с возрастом. Понимание механизмов взаимодействия адаптивного иммунитета с сенесцентными клетками важно в контексте разработки сенолитических вакцин и других иммунологических подходов, направленных на усиление эндогенной элиминации сенесцентных клеток.

Ключевые слова: сенесцентность; адаптивный иммунитет; регуляторные Т-лимфоциты; транскриптом единичных клеток; старение; генетические сигнатуры; тканерезидентные Т-лимфоциты; элиминация сенесцентных клеток; кожа

Introduction

Cellular senescence is a state of irreversible cell cycle arrest triggered by diverse stressors, including replicative exhaustion, DNA damage, telomere shortening, oxidative stress, and oncogene activation (Regulski, 2017; Di Micco et al., 2021). Senescent cells exhibit resistance to apoptosis, diminished cellular function, metabolic dysregulation, and multiple aberrations in protein quality control machinery. A hallmark feature of these cells is their sustained secretion of a broad array of pro-inflammatory mediators, collectively termed the senescence-associated secretory phenotype (SASP). The SASP is widely regarded as a primary driver of chronic, low-grade inflammation associated with aging, commonly referred to as inflammaging. Although senescence serves as an important tumor-suppressive mechanism, the prolonged persistence and accumulation of senescent cells in tissues disrupt tissue homeostasis, impair organ function, and contribute to the pathogenesis of age-related and degenerative diseases (Di Micco et al., 2021; Liao et al., 2021; Witham et al., 2023).

Preclinical studies in animal models have demonstrated that targeted elimination of senescent cells improves tissue function and metabolism, extends healthspan and lifespan, and attenuates the progression of age-associated pathologies (Yousefzadeh et al., 2019; Yang et al., 2023). Under physiological conditions, senescent cells are efficiently cleared by the immune system, with innate immune mechanisms being the most extensively characterized in this context. Natural killer (NK) cells recognize senescent cells primarily via the activating receptor NKG2D and eliminate them through perforin–granzyme-mediated cytotoxicity and interferon-gamma (IFN-γ) secretion (Antonangeli et al., 2019). Invariant natural killer T (iNKT) cells can also target senescent cells upon activation by glycolipid antigens (Arora et al., 2021). Furthermore, SASP-derived factors recruit macrophages, which contribute to the clearance of senescent cells during tissue remodeling (Song P. et al., 2020). However, with advancing age, the immune system's capacity to eliminate senescent cells declines – likely due to immunosenescence – resulting in increased senescent cell burden, chronic inflammation, tissue dysfunction, and heightened susceptibility to age-related diseases (Song S. et al., 2020; Hense et al., 2024).

Despite extensive research into the physiological clearance of senescent cells, the role of adaptive immunity in their elimination remains poorly understood (Matveeva et al., 2024). Conventional experimental approaches often inadequately reproduce the complex three-dimensional tissue architecture essential for critical interactions between adaptive immune system and senescent cells. A substantial proportion of T lymphocytes resides in peripheral tissues, does not recirculate, and exhibits functional properties distinct from those of circulating peripheral T cells (Li et al., 2025). Conversely, senescent cells are predominantly localized within the parenchyma and stroma of organs, where they can shape a unique microenvironment that modulates the efficacy of immune surveillance (Zhang W. et al., 2024). In this context, single-cell RNA sequencing (scRNA-seq) data derived directly from tissues hold particular significance. Such data enable the identification of senescent cells across diverse cell types and facilitate the assessment of key features of adaptive immunity, including the composition of specific T-cell subsets and their functional competence. By preserving the native tissue context, scRNA-seq datasets from multiple organs allow for the correlation of senescent cell burden with both quantitative and qualitative alterations in T-lymphocyte populations – the principal effectors of adaptive immunity (Kim S., Kim C., 2021).

In this study, we utilized publicly available scRNA-seq data to evaluate whether age-related accumulation of senescent cells in tissues is associated with alterations in the tissue-resident T-cell pool. It is currently accepted that cellular senescence manifests differently across distinct cell types (Cohn et al., 2023). Moreover, robust and universal molecular markers of senescence applicable to all senescent cell types remain elusive. Consequently, we employed the SenePy algorithm to infer cellular senescence status. Unlike conventional differential expression analyses, SenePy identifies co-expression gene network clusters associated with aging (Sanborn et al., 2025). Skin aging is a multifaceted process driven by cumulative exposure to diverse damaging factors throughout life. Key hallmarks of skin aging include the accumulation of senescent cells, disruption of dermal extracellular matrix architecture, degradation of elastic fibers, and impairment of barrier function (Shin et al., 2025). In the present study, the identification of senescent cells within each human skin cell

type, combined with quantification of various T-lymphocyte subpopulations, revealed significant age-related alterations in tissue-resident T cells that were associated with the accumulation of senescent cells.

Materials and methods

For this analysis, we used publicly available single-cell RNA sequencing (scRNA-seq) datasets deposited in the NCBI Gene Expression Omnibus (GEO) and the Genome Sequence Archive for Human (GSA-Human). Skin biopsy samples from healthy donors (n = 32; age range: 18–76 years) were automatically retrieved from these repositories (see Supplementary Materials, Table S1)[1].

Unique Molecular Identifier (UMI) count matrices were generated from raw sequencing reads using the 10x Genomics Cell Ranger pipeline (v9.0.1). Subsequent processing of count matrices and associated metadata was primarily performed using the Scanpy toolkit (Wolf et al., 2018). Prior to downstream analysis, low-quality cells were filtered out based on the following criteria: (i) total UMI counts <500 or >5 median absolute deviations (MAD); (ii) number of detected genes >5 MAD; and (iii) mitochondrial gene expression >15 % or >4 MAD from the median. Doublets were identified and removed using the Scrublet package (Wolock et al., 2019).

Following quality control, samples were integrated into a unified dataset and prepared for clustering. This preprocessing pipeline included: (i) library-size normalization to a target sum of 10,000 UMIs per cell (scanpy.pp.normalize_total(target_sum=1e4)); (ii) log-transformation; (iii) scaling; (iv) dimensionality reduction via principal component analysis (PCA); and (v) batch-effect correction using the Harmony algorithm (Korsunsky et al., 2019). Cell-type annotation was performed on log-normalized data using CellTypist (Domínguez et al., 2022), which employs pre-trained logistic regression models. Specifically, we applied the "Adult_Human_Skin" model (Reynolds et al., 2021), which encompasses annotations for diverse dermal, epidermal, and immune cell populations in human skin. To validate and refine automated annotations, cells were further clustered using the Leiden algorithm. Cluster identities were cross-referenced with CellTypist predictions, and manual curation of annotations was performed where necessary. The full data processing workflow is illustrated in Figure 1. Particular attention was devoted to the accurate annotation of T-lymphocyte subpopulations. To this end, the T-cell cluster was isolated from the integrated dataset and reprocessed starting from the original UMI count matrix to ensure a more precise representation of T-cell heterogeneity in reduced-dimensional space. Annotations were refined as needed based on this focused re-analysis. Samples exhibiting insufficient representation of specific cell types were excluded from relevant downstream analyses at corresponding stages of the study.

Canonical markers of cellular senescence are highly cell type-specific and poorly reflect the true senescent state *in vivo*. Therefore, cellular senescence status was assessed using the SenePy algorithm, published in 2025 (Sanborn et al., 2025), which enables discrimination between bona fide senescence-associated markers and genes, the expression of which is

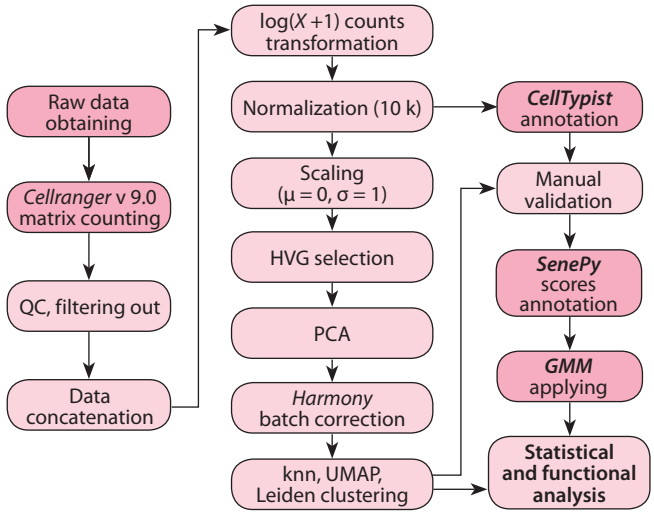


Fig. 1. Schematic representation of the data processing workflow.

elevated for reasons unrelated to senescence. Within this algorithm, the identification of genes potentially associated with age-related accumulation of senescent cells is performed under the following criteria: the gene must be expressed in fewer than 5 % of cells from young donors, and in more than 1 % but fewer than 20 % of cells from older donors. Additionally, either the proportion of cells expressing the gene in aged individuals must be at least 2.5-fold higher than in young individuals, or the absolute increase in the proportion of expressing cells (i. e., the difference between old and young donors) must exceed 5 %. This strategy enables the identification of cell type-specific genetic signatures of senescence within a given tissue, thereby allowing more accurate detection of senescent cells in *ex vivo* samples compared to conventional approaches. Each cell is assigned a continuous numerical metric – the "SenePy score" – reflecting the degree to which its gene expression profile aligns with the corresponding cell type-specific senescence signature.

Following SenePy scoring, Gaussian Mixture Models (GMMs) were fitted to the distribution of SenePy scores within each annotated cell type. Depending on the shape of the score distribution, models comprised either two or three components. The threshold for classifying a cell as senescent was defined as the value lying between the two rightmost GMM components. This approach enabled a quantitative estimation of the fraction of cells exhibiting robust senescence features within each cell population.

Correlation analyses were performed using the spearmanr() function from the scipy.stats module to compute Spearman's rank correlation coefficient and associated p-values. To account for multiple comparisons, Bonferroni correction was applied.

Differentially expressed genes (DEGs) in T-lymphocyte populations from young and old donors were identified using the rank_genes_groups() function from the Scanpy package, employing the Mann–Whitney U test. Genes were considered differentially expressed if they met the following criteria: false discovery rate (FDR) < 0.01, presence in more than 10 % of cells within the target group, and detection in fewer than

[1] Supplementary Tables S1–S4 and Fig. S1 are available at: https://vavilovj-icg.ru/download/pict-2025-29/appx42.zip

Fig. 2. Cell type annotation of human skin using the CellTypist tool.

DC – dendritic cells; KC – keratinocytes; LE – lymphoid epithelial cells; Tc – cytotoxic T lymphocytes (classical phenotype: $CD3^+CD8^+$); Th – T helper cells (classical phenotype: $CD3^+CD4^+$); Treg – regulatory T cells (classical phenotype: $CD3^+CD4^+FoxP3^+$); VE – vascular endothelial cells.

50 % of cells in the comparison group. Functional enrichment analysis of the identified DEGs was performed in the R programming language using the enricher() function from the clusterProfiler package (Yu et al., 2021). Gene sets from the C5 (ontology gene sets) and C7 (immunologic signature gene sets) collections of the Molecular Signatures Database (MSigDB; Subramanian et al., 2005) were used as reference annotations. Significantly enriched gene sets were manually grouped into functional categories.

Results

To identify senescent cells in human skin tissues, we adapted and applied the recently published SenePy algorithm (Sanborn et al., 2025), followed by Gaussian Mixture Modeling (GMM). The analysis was performed on the major skin cell populations previously annotated (Fig. 2).

As a result, we observed a significant age-associated increase in the proportion of senescent cells across multiple cell types in human skin samples (Fig. 3). Specifically, the fraction of senescent cells rose with age in tissue-resident dendritic cells, macrophages, T lymphocytes, keratinocytes, melanocytes, fibroblasts, pericytes, and endothelial cells. Notably, the rate of accumulation varied between cell types, reflecting the heterogeneity of aging processes among distinct cellular populations within the same tissue.

Our analysis revealed a significant age-related accumulation of cells exhibiting senescence features in the skin, consistent with prior evidence implicating cellular senescence as a key hallmark of tissue aging (Childs et al., 2015). The overall proportion of senescent cells across all cell types also showed

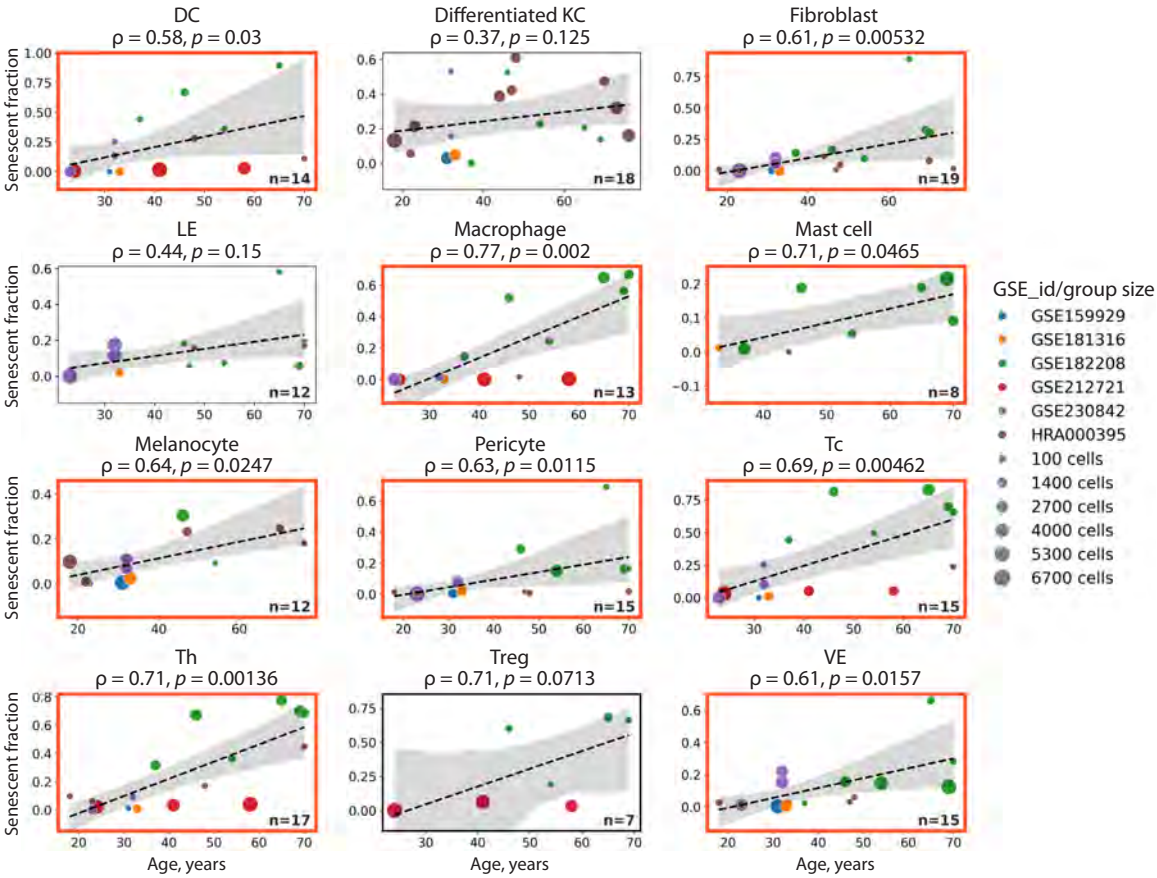


Fig. 3. Correlations between the accumulation of senescent cells in distinct human skin cell types and donor age.

For each cell type, samples with cell counts below 2SD (standard deviations) from the mean across all donors were excluded from the analysis. Statistically significant correlations are highlighted with red boxes. DC – dendritic cells; KC – keratinocytes; LE – lymphoid epithelial cells; Tc – cytotoxic T lymphocytes; Th – T helper cells; Treg – regulatory T cells; VE – vascular endothelial cells.

a positive correlation with donor age (Fig. 4), indicating a progressive disruption of tissue homeostasis. Given that senescent cells are characterized by a stable cell cycle arrest and thus lack proliferative capacity, their age-dependent accumulation is likely attributable to a decline in the efficiency of mechanisms responsible for their clearance.

Therefore, in the next step, we sought to investigate how the proportions of major T-lymphocyte subpopulations in the skin change with age. Correlation analysis did not reveal statistically significant age-related changes in the proportions of the three T-lymphocyte subpopulations examined, nor in key immunological indices (Fig. 5). Given the absence of detectable age-associated alterations among tissue-resident T lymphocytes, we next sought to explore potential associations between T-lymphocyte populations and the accumulation of senescent cells independent of chronological age.

Different cell types may exhibit varying rates of aging or differing immunogenicity of their senescent counterparts, which could account for the observed heterogeneity in age-related accumulation of senescent cells. Therefore, we first sought to determine whether any alterations in skin T-lymphocyte populations were associated with the burden of senescent cells. Specifically, we assessed the relationship between the accumulation of senescent cells within each cell type and the relative abundance of T-lymphocyte subpopulations (Fig. S1). We found a significant increase in total T-lymphocyte frequency associated with the accumulation of senescent pericytes, as well as modest trends ($p < 0.07$) toward elevated regulatory T-cell (Treg) proportions correlating with senescent cell burden in certain cell types.

In the next step, we examined how the proportions of different T-lymphocyte populations vary with the total burden of senescent cells across all cell types. We observed a significant increase in the relative abundance of both T helper (Th) cells and regulatory T (Treg) cells as the cumulative number of senescent cells rose (Fig. 6). Moreover, we noted a statistically significant elevation in the "tissue immunoregulatory

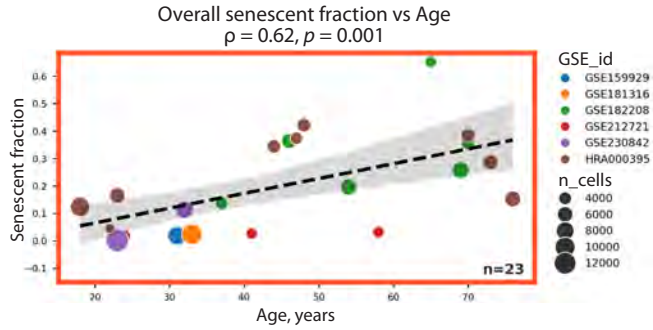


Fig. 4. Proportion of senescent cells across all cell types as a function of donor age.

index" – defined as the Th/Tc ratio – which reflects a shift toward T helper dominance over cytotoxic T lymphocytes.

Thus, we identified a significant association between the accumulation of senescent cells in human skin and an imbalance in T-cell immunity. This imbalance was characterized by an increased proportion of regulatory T cells and T helper cells, accompanied by a relative decrease in cytotoxic T lymphocytes. Notably, these alterations were not directly correlated with chronological age, underscoring the specific role of interactions between T-cell immunity and senescent cells, independent of aging per se.

The age-independent shifts in the tissue-resident T-lymphocyte pool observed in earlier analyses highlight the involvement of adaptive immunity in tissue aging processes. However, these findings do not provide insight into the functional states of Treg cells, Th, or cytotoxic T lymphocytes. To further characterize the functional implications of these changes, we performed differential gene expression analysis followed by functional enrichment profiling of T-lymphocyte populations (see Materials and methods), comparing cells from older versus younger donors (Fig. 7).

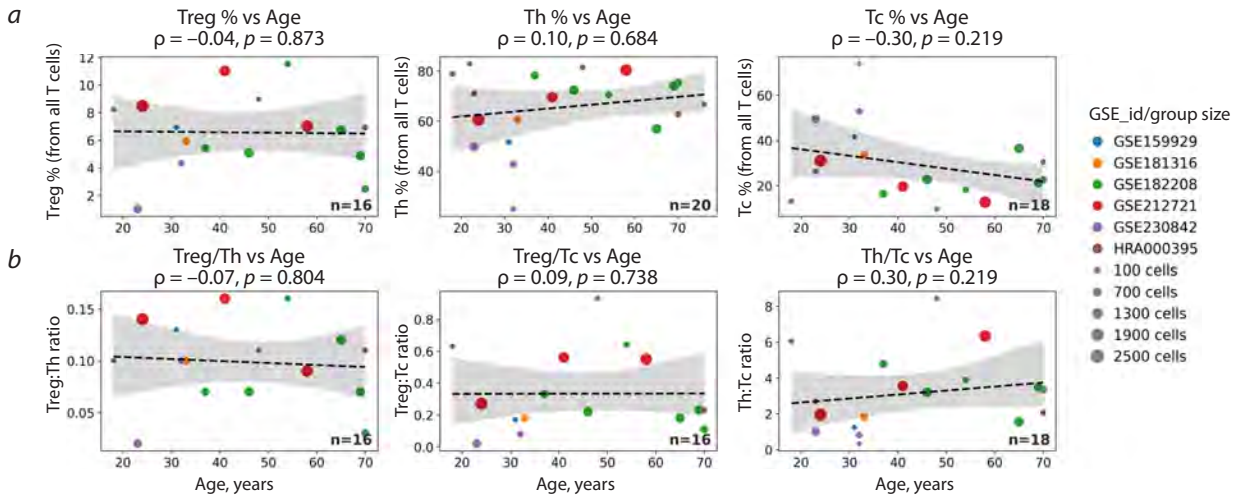


Fig. 5. Age-related changes in the proportions of major T-lymphocyte populations (*a*) and their ratios (*b*).

The immunological indices shown – Th/Tc, Treg/Tc, and Treg/Th ratios – are widely used to assess immune status with greater precision and sensitivity in various pathological or compromised conditions. In this figure, the proportion of each T-lymphocyte subset is expressed relative to the total number of T lymphocytes, thereby reflecting the balance among subpopulations within the entire pool of skin-resident T cells. Treg – regulatory T cells; Th – T helper cells; Tc – cytotoxic T lymphocytes.

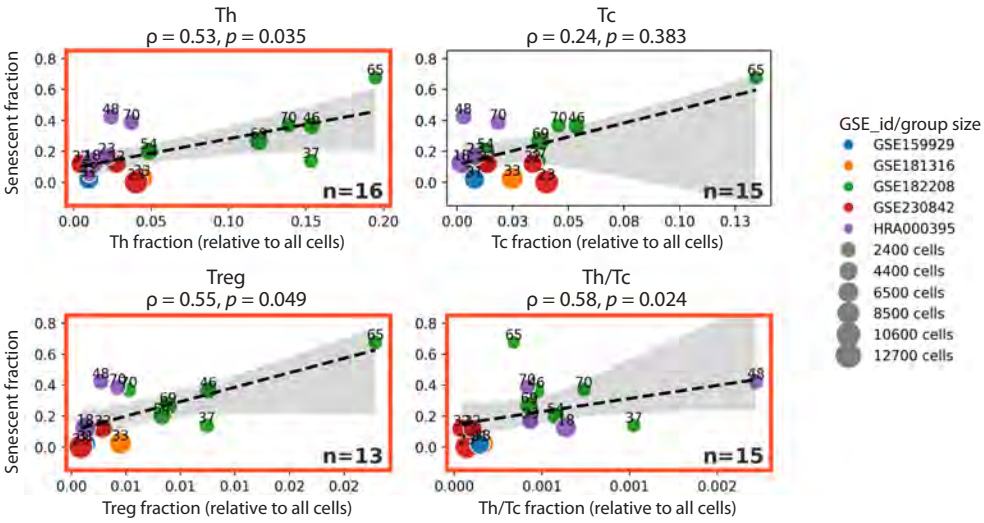


Fig. 6. Proportions of major T-lymphocyte populations relative to the total number of senescent cells.

In this figure, the abundance of each T-lymphocyte subset is expressed as a fraction of the total cell count across all cell types, rather than as a proportion of the total T-cell pool. This approach captures age-independent shifts in T-lymphocyte representation within the entire skin cellular landscape and more accurately reflects biologically relevant changes associated with the accumulation of senescent cells. Th – T helper cells; Tc – cytotoxic T lymphocytes; Treg – regulatory T cells.

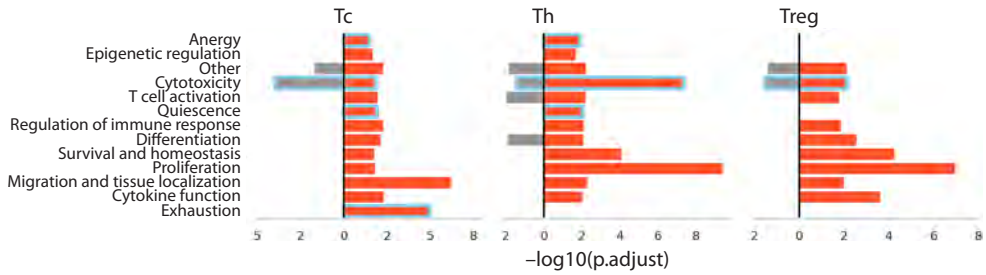


Fig. 7. Results of functional enrichment analysis of differentially expressed genes (DEGs) in tissue-resident T-lymphocyte populations from older versus younger donors.

Red bars represent enrichment of functional pathways by upregulated genes, while gray bars indicate enrichment by downregulated genes. The X-axis shows the $-\log_{10}$-transformed FDR-corrected *p*-value, such that higher values correspond to stronger enrichment. Tc – cytotoxic T lymphocytes; Th – T helper cells; Treg – regulatory T cells.

Functional enrichment analysis revealed statistically significant overrepresentation of biological pathways associated with enhanced functional activity of T helper (Th) cells, including tissue adaptation, differentiation, and response to cytokines involved in their homeostasis. Additionally, enrichment of pathways characteristic of quiescent and anergic states was observed in this population (highlighted with blue boxes). Notably, however, these Th cells did not exhibit clear molecular signatures of exhaustion. In contrast, age-related alterations in cytotoxic T lymphocytes were associated with enrichment of pathways typical of quiescence, anergy, and exhaustion. Intriguingly, this Tc population also displayed significant downregulation of pathways directly linked to their effector function – particularly cytotoxicity. Conversely, regulatory T cells showed no evidence of quiescence, anergy, or exhaustion. Instead, similar to Th cells, Treg cells exhibited heightened functional and proliferative activity. Moreover, this population demonstrated significant enrichment of genes involved in differentiation and response to homeostatic cytokines – specifically IL-2, IL-7, and IL-15 – which are essential for the maintenance and survival of tissue-resident regulatory T cells (Table S2).

Thus, functional enrichment analysis of differentially expressed genes (DEGs) identified from scRNA-seq data revealed distinct functional states across T-lymphocyte subsets. Cytotoxic T lymphocytes exhibited clear signatures of exhaustion and reduced functional activity. In contrast, regulatory T cells displayed heightened functional activity and showed no evidence of exhaustion or anergy. Changes in the Th population were more heterogeneous: alongside increased functional activity, these cells also exhibited features characteristic of anergy and quiescence.

Discussion

The accumulation of senescent cells is a hallmark of tissue aging and is closely linked to the development of chronic, low-grade systemic inflammation – termed "inflammaging" – which constitutes a major risk factor for age-related

diseases (Franceschi et al., 2018). Using a modern algorithm for identifying senescence-associated gene signatures, we demonstrated that the proportion of cells exhibiting senescence features increases with age in human skin. Importantly, this accumulation is not uniform across all cell types, underscoring the heterogeneity of aging trajectories among distinct cellular populations and highlighting the multifaceted nature of tissue aging (Ge et al., 2022).

The immune system plays a central role in the surveillance and clearance of senescent cells. The pro-inflammatory secretome of senescent cells – commonly referred to as the senescence-associated secretory phenotype (SASP) – recruits innate immune effectors such as macrophages, neutrophils, natural killer (NK) cells, and NKT cells, which contribute to the recognition and elimination of senescent cells (Song P. et al., 2020). Although emerging evidence implicates T lymphocytes in these processes, the role of adaptive immunity in senescent cell clearance remains incompletely understood (Matveeva et al., 2024). Our findings reveal that the burden of senescent cells in human skin is associated with a local imbalance in T-cell immunity, suggesting that T lymphocytes actively participate in regulating senescent cell homeostasis. Notably, higher senescent cell loads correlated with an increased proportion of regulatory T cells and an elevated Th/Tc ratio. This shift points toward the establishment of an immunosuppressive microenvironment that may facilitate immune evasion by senescent cells (Zhang W. et al., 2024). This interpretation is further supported by functional profiling of T-cell populations in older donors. Cytotoxic T lymphocytes exhibited molecular signatures of exhaustion and diminished effector potential, whereas both Treg and Th cells displayed heightened functional activity and signs of tissue adaptation. Collectively, these quantitative and qualitative alterations in the skin-resident T-cell compartment in aged individuals may promote peripheral tolerance to senescence-associated antigens. This aligns with the hypothesis that aging impairs the immune system's capacity to recognize and efficiently eliminate senescent cells, thereby contributing to their progressive accumulation (Song P. et al., 2020).

It is well established that senescent cells not only generate a pro-inflammatory milieu but also can actively suppress effector T-cell functions and evade immune surveillance (Lorenzo et al., 2022). For instance, certain SASP-derived chemokines selectively recruit Treg-cells, while senescence-driven polarization of monocytes toward an M2-like macrophage phenotype suppresses cytotoxic T-cell activation (Zhang X. et al., 2024). Moreover, aging-associated activation of endogenous retroelements – particularly LINE-1 – triggers an IFN-γ-mediated response (Zhang X. et al., 2020). This antiviral-like response may fuel chronic inflammation and drive T-cell exhaustion, a phenotype strikingly reminiscent of the cytotoxic T-cell dysfunction observed in our cohort of older donors.

In summary, our data indicate that the skin T-cell compartment undergoes substantial functional remodeling with age. The decline in cytotoxic activity coupled with enhanced regulatory T-cell function may foster immunological tolerance, thereby enabling the persistence and accumulation of senescent cells and contributing to inflammaging. We propose that this represents an active process of peripheral tolerance to senescence-associated antigens, wherein the aging immune system progressively loses its ability to detect and eliminate senescent cells. The identified imbalance in tissue-resident T-lymphocyte populations thus constitutes a promising therapeutic target for interventions aimed at restoring immune surveillance and promoting the clearance of senescent cells.

Conclusion

In this study, we employed bioinformatic analyses of publicly available scRNA-seq data derived from skin biopsies of healthy donors to identify aging-associated alterations in tissue-resident adaptive immunity. We demonstrated that skin aging – manifested as the accumulation of senescent cells across multiple cell types – is associated with a shift in the balance between Th and cytotoxic T lymphocytes, as well as an increased proportion of Treg cells. Functional enrichment analysis further revealed a general decline in cytotoxic potential among tissue T cells, concurrent with enhanced regulatory activity. These changes likely reflect compensatory adaptations within the tissue T-cell compartment in response to the persistent accumulation of senescent cells and the resulting chronic inflammatory microenvironment. In this context, the observed T-cell remodeling appears to promote an immunosuppressive milieu, potentially contributing to the age-related decline in the efficiency of senescent cell clearance.

scRNA-seq data provide a powerful tool for investigating immune-senescence interactions at the tissue level. Preservation of the tissue cellular context enables the identification of physiologically relevant aging signatures and facilitates the analysis of gene programs associated with activation or suppression of specific immune components. Nevertheless, this approach has inherent limitations. The loss of spatial tissue architecture precludes direct assessment of cell-to-cell interactions, while technical artifacts introduced during sample preparation and data integration from multiple sources necessitate rigorous preprocessing, batch-effect correction, and normalization – steps that may introduce substantial uncertainty into the results. Therefore, to gain a deeper understanding of the role of adaptive immunity in the surveillance and elimination of senescent cells, future studies should integrate scRNA-seq with spatial transcriptomics, histological validation, and methods capable of defining the antigen specificity of T and B cells. Additionally, longitudinal analyses of T- and B-cell receptor repertoires will be essential to elucidate dynamic changes in antigen recognition during aging and their functional consequences for immune-mediated clearance of senescent cells.

References

Antonangeli F., Zingoni A., Santoni A., Soriani A. Senescent cells: living or dying is a matter of NK cells. *J Leukoc Biol*. 2019;105(6):1275-1283. doi 10.1002/jlb.mr0718-299r

Arora S., Thompson P.J., Wang Y., Bhattacharyya A., Apostolopoulou H., Hatano R., Naikawadi R.P., Shah A., Wolters P.J., Koliwad S., Bhattacharya M., Bhushan A. Invariant natural killer T cells coordinate removal of senescent cells. *Med*. 2021;2(8):938-950. doi 10.1016/j.medj.2021.04.014

Childs B.G., Durik M., Baker D.J., van Deursen J.M. Cellular senescence in aging and age-related disease: from mechanisms to therapy. *Nat Med*. 2015;21(12):1424-1435. doi 10.1038/nm.4000

Cohn R.L., Gasek N.S., Kuchel G.A., Xu M. The heterogeneity of cellular senescence: insights at the single-cell level. *Trends Cell Biol*. 2023;33(1):9-17. doi 10.1016/j.tcb.2022.04.011

Di Micco R., Krizhanovsky V., Baker D., d'Adda di Fagagna F. Cellular senescence in ageing: from mechanisms to therapeutic opportunities. *Nat Rev Mol Cell Biol*. 2021;22(2):75-95. doi 10.1038/s41580-020-00314-w

Domínguez Conde C., Xu C., Jarvis L.B., Rainbow D.B., Wells S.B., Gomes T., Howlett S.K., … Sims P.A., Farber D.L., Saeb-Parsy K., Jones J.L., Teichmann S.A. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*. 2022;376(6594): eabl5197. doi 10.1126/science.abl5197

Franceschi C., Garagnani P., Parini P., Giuliani C., Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol*. 2018;14(10):576-590. doi 10.1038/s41574-018-0059-4

Ge M.X., Yu Q., Li G.H., Yang L.Q., He Y., Li J., Kong Q.P. Multiple time-series expression trajectories imply dynamic functional changes during cellular senescence. *Comput Struct Biotechnol J*. 2022;20:4131-4137. doi 10.1016/j.csbj.2022.08.005

Hense J.D., Isola J.V.V., Garcia D.N., Magalhães L.S., Masternak M.M., Stout M.B., Schneider A. The role of cellular senescence in ovarian aging. *NPJ Aging*. 2024;10(1):35. doi 10.1038/s41514-024-00157-1

Kim S., Kim C. Transcriptomic analysis of cellular senescence: one step closer to senescence atlas. *Mol Cells*. 2021;44(3):136-145. doi 10.14348/molcells.2021.2239

Korsunsky I., Millard N., Fan J., Slowikowski K., Zhang F., Wei K., Baglaenko Y., Brenner M., Loh P.R., Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296. doi 10.1038/s41592-019-0619-0

Li J., Xiao C., Li C., He J. Tissue-resident immune cells: from defining characteristics to roles in diseases. *Signal Transduct Target Ther*. 2025;10(1):12. doi 10.1038/s41392-024-02050-5

Liao Z., Yeo H.L., Wong S.W., Zhao Y. Cellular senescence: mechanisms and therapeutic potential. *Biomedicines*. 2021;9(12):1769. doi 10.3390/biomedicines9121769

Lorenzo E.C., Torrance B.L., Keilich S.R., Al-Naggar I., Harrison A., Xu M., Bartley J.M., Haynes L. Senescence-induced changes in CD4 T cell differentiation can be alleviated by treatment with senolytics. *Aging Cell*. 2022;21(1):e13525. doi 10.1111/acel.13525

Matveeva K., Vasilieva M., Minskaia E., Rybtsov S., Shevyrev D. T-cell immunity against senescence: potential role and perspectives. *Front Immunol*. 2024;15:1360109. doi 10.3389/fimmu.2024.1360109

Regulski M.J. Cellular senescence: what, why, and how. *Wounds*. 2017; 29(6):168-174

Reynolds G., Vegh P., Fletcher J., Poyner E.F.M., Stephenson E., Goh I., Botting R.A., … Rajan N., Reynolds N.J., Teichmann S.A., Watt F.M., Haniffa M. Developmental cell programs are co-opted in inflammatory skin disease. *Science*. 2021;371(6527):eaba6500. doi 10.1126/science.aba6500

Sanborn M.A., Wang X., Gao S., Dai Y., Rehman J. Unveiling the cell-type-specific landscape of cellular senescence through single-cell transcriptomics using *SenePy*. *Nat Commun*. 2025;16:1884. doi 10.1038/s41467-025-57047-7

Shin S.H., Lee Y.H., Rho N.K., Park K.Y. Skin aging from mechanisms to interventions: focusing on dermal aging. *Front Physiol*. 2023; 14:1195272. doi 10.3389/fphys.2023.1195272

Song P., An J., Zou M.-H. Immune clearance of senescent cells to combat ageing and chronic diseases. *Cells*. 2020;9(3):671. doi 10.3390/cells9030671

Song S., Kirkland J.L., Sun Y., Tchkonia T., Jiang J. Targeting senescent cells for a healthier aging: challenges and opportunities. *Adv Sci*. 2020;7(23):2002611. doi 10.1002/advs.202002611

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545-15550. doi 10.1073/pnas.0506580102

Witham M.D., Granic A., Miwa S., Passos J.F., Richardson G.D., Sayer A.A. New Horizons in cellular senescence for clinicians. *Age Ageing*. 2023;52(7):afad127. doi 10.1093/ageing/afad127

Wolf F.A., Angerer P., Theis F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi 10.1186/s13059-017-1382-0

Wolock S.L., Lopez R., Klein A.M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;8(4):281-291.e9. doi 10.1016/j.cels.2018.11.005

Yang D., Sun B., Li S., Wei W., Liu X., Cui X., Zhang X., Liu N., Yan L., Deng Y., Zhao X. NKG2D-CAR T cells eliminate senescent cells in aged mice and nonhuman primates. *Sci Transl Med*. 2023;15(709):eadd1951. doi 10.1126/scitranslmed.add1951

Yousefzadeh M.J., Melos K.I., Angelini L., Burd C.E., Robbins P.D., Niedernhofer L.J. Mouse models of accelerated cellular senescence. In: Demaria M. (Ed.) Cellular Senescence. Methods in Molecular Biology. Vol. 1896. New York: Humana Press, 2019;203-230. doi 10.1007/978-1-4939-8931-7_17

Yu G., Wang L.G., Han Y., He Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012; 16(5):284-287. doi 10.1089/omi.2011.0118

Zhang W., Zhang K., Shi J., Qiu H., Kan C., Ma Y., Hou N., Han F., Sun X. The impact of the senescent microenvironment on tumorigenesis: insights for cancer therapy. *Aging Cell*. 2024;23(5):e14182. doi 10.1111/acel.14182

Zhang X., Zhang R., Yu J. New understanding of the relevant role of LINE-1 retrotransposition in human disease and immune modulation. *Front Cell Dev Biol*. 2020;8:657. doi 10.3389/fcell.2020.00657

Zhang X., Ng Y.E., Chini L.C.S., Heeren A.A., White T.A., Li H., Huang H., Doolittle M.L., Khosla S., LeBrasseur N.K. Senescent skeletal muscle fibroadipogenic progenitors recruit and promote M2 polarization of macrophages. *Aging Cell*. 2024;23(3):e14069. doi 10.1111/acel.14069

Conflict of interest. The authors declare no conflict of interest.

Received July 30, 2025. Revised September 25, 2025. Accepted September 26, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(7):1145-1154

doi 10.18699/vjgb-25-119

OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning

E.V. Malyugin[1], D.A. Afonnikov[2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
evgeny.malyugin98@gmail.com

Abstract. In recent years, the rapid growth of sequencing data has exacerbated the problem of functional annotation of protein sequences, as traditional homology-based methods face limitations when working with distant homologs, making it difficult to accurately determine protein functions. This paper introduces the OrthoML2GO method for protein function prediction, which integrates homology searches using the USEARCH algorithm, orthogroup analysis based on OrthoDB version 12.0, and a machine learning algorithm (gradient boosting). A key feature of our approach is the use of orthogroup information to account for the evolutionary and functional similarity of proteins and the application of machine learning to refine the assigned GO terms for the target sequence. To select the optimal algorithm for protein annotation, the following approaches were applied sequentially: the k-nearest neighbors (KNN) method; a method based on the annotation of the orthogroup most represented in the k-nearest homologs (OG); a method of verifying the GO terms identified in the previous stage using machine learning algorithms. A comparison of the prediction accuracy of GO terms using the OrthoML2GO method with the Blast2GO and PANNZER2 annotation programs was performed on sequence samples from both individual organisms (humans, Arabidopsis) and a combined sample represented by different taxa. Our results demonstrate that the proposed method is comparable to, and by some evaluation metrics outperforms, these existing methods in terms of the quality of protein function prediction, especially on large and heterogeneous samples of organisms. The greatest performance improvement is achieved by combining information about the closest homologs and orthogroups with verification of terms using machine learning methods. Our approach demonstrates high performance for large-scale automatic protein annotation, and prospects for further development include optimizing machine learning model parameters for specific biological tasks and integrating additional sources of structural and functional information, which will further improve the method's accuracy and versatility. In addition, the introduction of new bioinformatics tools and the expansion of the annotated protein database will contribute to the further improvement of the proposed approach.

Key words: protein function prediction; gene ontology; homology; orthogroup; machine learning

For citation: Malyugin E.V., Afonnikov D.A. OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1145-1154. doi 10.18699/vjgb-25-119

Funding. The work was supported by the Kurchatov Genomic Center of ICG SB RAS under agreement with the Ministry of Science and Higher Education of the Russian Federation No. 075-15-2019-1662, and by the state budget project No. FWNR-2022-0020.

Acknowledgements. This research was supported in part through computational resources of HPC facilities at collaborative center "Bioinformatics" ICG SB RAS.

OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения

Е.В. Малюгин[1], Д.А. Афонников[2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
evgeny.malyugin98@gmail.com

Аннотация. В последние годы быстрый рост объемов данных секвенирования обострил проблему функциональной аннотации белковых последовательностей, поскольку традиционные методы, основанные на гомологии, сталкиваются с ограничениями при работе с отдаленными гомологами, что затрудняет наиболее точное

определение функций белков. В нашей работе представлен метод предсказания функций белков OrthoML2GO, который интегрирует поиск гомологичных последовательностей с помощью алгоритма USEARCH, анализ ортогрупп на базе OrthoDB 12-й версии и алгоритм машинного обучения (градиентный бустинг). Ключевая особенность подхода заключается в использовании информации об ортогруппах для учета эволюционного и функционального сходства белков и применения машинного обучения для дальнейшего уточнения терминов Gene Ontology (GO) для анализируемой последовательности. Для выбора оптимального алгоритма аннотации белков были поэтапно применены следующие подходы: метод *k* ближайших соседей (KNN); метод на основе аннотации ортогруппы, наиболее представленной у *k* ближайших гомологов (OG); метод верификации выявленных на предыдущем этапе терминов GO с помощью алгоритмов машинного обучения. Проведено сравнение точности предсказания терминов GO методом OrthoML2GO с программами аннотации Blast2GO и PANNZER2 на выборках последовательностей как отдельных организмов (человек, арабидопсис), так и на комбинированной выборке последовательностей, представленных разными таксонами. Результаты показали, что предложенный метод не уступает, а по некоторым показателям превосходит их по качеству предсказания функций белков, особенно на больших и разнородных выборках организмов, а наибольший прирост точности достигается за счет комбинации информации о ближайших гомологах и ортогруппах в сочетании с верификацией терминов методами машинного обучения. Разработанный подход демонстрирует высокую эффективность для крупномасштабной автоматической аннотации белков. Перспективы дальнейшего развития включают оптимизацию параметров моделей машинного обучения под конкретные биологические задачи и интеграцию дополнительных источников структурно-функциональной информации, что позволит еще больше повысить точность и универсальность метода. Кроме того, внедрение новых инструментов биоинформатики и расширение базы данных аннотированных белков будут способствовать дальнейшему совершенствованию предложенного подхода.
Ключевые слова: предсказание функций белка; генная онтология; гомология; ортогруппа; машинное обучение

Introduction

The introduction of next-generation sequencing (NGS) technologies has led to exponential growth in the volume of data on DNA, RNA, and protein sequences (Goodwin et al., 2016). The primary sources of these data are large-scale and numerous projects in genomics, transcriptomics, and proteomics (Cheng et al., 2018; Lewin et al., 2018). However, the function of a significant proportion of the sequences identified in such projects remains unknown (Galperin, Koonin, 2010).

Expert gene annotation requires substantial time to search for gene function information in the literature, and although it is the most reliable method, it is impractical to apply it to the vast number of newly predicted genes. Therefore, for most new amino acid sequences (hereafter referred to as sequences for brevity), the development of effective automatic annotation methods is necessary to determine their molecular functions, roles in cellular processes, and cellular localization. Given the widespread use of the Gene Ontology (GO) database for functional annotation (Ashburner et al., 2000; Du Plessis et al., 2011; Gene Ontology Consortium, 2023), the task reduces to automatically assigning these terms to sequences.

Most methods for predicting protein function, based on sequence or three-dimensional structure analysis, rely on a fundamental principle: function can be predicted by establishing reliable structural or evolutionary similarity with a protein, the function of which is already known (Benso et al., 2013). A crucial task here is deciphering the relationship between the detected structural or sequence similarity and the actual level of functional relatedness (Pearson, 2013). Among these methods, homology-based function prediction methods are widely regarded for their broad applicability and relative simplicity. Homology-based methods assign GO terms to the analyzed protein based on the similarity of its amino acid sequence to the primary structures of proteins with known functions. In other words, the function of a protein can be deciphered by analyzing its similarity to other proteins for which the function has been reliably determined (Eisenberg et al., 2000; Pearson, 2013).

The BLAST method (Altschul et al., 1990) is widely used for comparing the amino acid sequences and identifying homologous regions. However, new tools for searching homologous sequences in databases have recently emerged, such as GHOSTX (Suzuki et al., 2014), DIAMOND (Buchfink et al., 2015), MMseqs2 (Steinegger, Söding, 2017), and others. Their characteristic feature is high processing speed, orders of magnitude faster than BLAST, achieved primarily through more efficient processing of matched sequence fragments.

The concept of homology is fundamental for drawing conclusions about the evolutionary processes of gene formation and function. In the early 1970s, Walter Fitch (Fitch, 1970) proposed classifying homologous proteins into orthologs and paralogs according to their origin. Orthologs originate from the evolutionary divergence of genes in different taxa during speciation. Paralogs are formed through gene duplications. It is assumed that orthologs retain the function of the ancestral gene from the ancestral species, while paralogs may acquire new functions after duplication events (Fitch, 2000; Kuzniar et al., 2008; Altenhoff et al., 2019). Given the immense importance of orthologs for comparative genomics and functional annotation, information on orthologous genes and their families is accumulated in several specialized databases, which are crucial for identifying and analyzing orthologous groups of genes (orthogroups) (Jensen et al., 2008; Kriventseva et al., 2008). It should be noted that methods involving machine learning algorithms are successfully used to solve gene function prediction problems, allowing for increased accuracy compared to earlier approaches (Sanderson et al., 2023; Yuan et al., 2023).

This work investigates the possibility of predicting protein functions based on searching for homologous sequences,

Table 1. List of organisms included in the study

Species name	Number of sequences	Annotation source
Arabidopsis thaliana	27,655	TAIR (Reiser et al., 2024)
Homo sapiens	19,763	EBI Gene Ontology Annotation Database (Huntley et al., 2015)
Drosophila melanogaster	28,543 (includes isoforms)	FlyBase (Öztürk-Çolak et al., 2024)
Solanum tuberosum	40,722 (includes isoforms)	SpudDB (Hamilton et al., 2025a)
Danio rerio	33,428 (includes isoforms)	ZFIN (Bradford et al., 2022)
Chlamydomonas reinhardtii	16,090	PhycoCosm (Grigoriev et al., 2021)
Oryza sativa	34,226 (includes isoforms)	RGAP (Hamilton et al., 2025b)

considering their orthologs, and employing machine learning methods. A step-by-step analysis of the influence of these three factors on the accuracy of GO term prediction was performed. It is shown that among machine learning methods, the gradient boosting algorithm demonstrates the highest prediction accuracy. Based on this, the OrthoML2GO prediction algorithm was implemented. Its accuracy was compared with the Blast2GO and PANNZER2 methods. It is shown that the proposed method provides higher accuracy, especially on large and heterogeneous datasets.

Materials and methods

Amino acid sequence data. The lists of organism species and amino acid sequences used in the work are presented in Table 1. They include organisms with varying degrees of genome annotation completeness (Table S1)[1], representing different taxa of both plants and animals: dicots, monocots, unicellular algae, vertebrates, arthropods (Table 1).

OrthoDB as a source of homologous sequences, annotations, and orthology information. The OrthoDB v 12.0 database (https://www.orthodb.org/) (Tegenfeldt et al., 2025) was used as a source of homologous sequences, their GO term annotations, and orthology data. The database includes information on 5,827 eukaryotic species, 17,551 bacteria, 607 archaea, and 7,962 viruses. It contains over 162 million sequences classified into over 10 million orthogroups. The database also includes GO annotation for part of the sequences and thus represents a convenient source for their classification into orthologs and GO annotation. Furthermore, this database provides classification of protein sequences into orthologous families, for which generalized functional annotations of proteins in GO terms are also provided.

Search for homologous sequences. The search for homologs was performed using the USEARCH v 11.0.667 algorithm (https://drive5.com/usearch/) (Edgar, 2010) with the usearch_local command. It performs searches for high-identity matches orders of magnitude faster than BLAST. During the search for homologous sequences, it was inevitable that the list of homologs included the query sequence itself. For an objective evaluation, identical sequences were excluded from the search results.

[1] Supplementary Tables S1–S12 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Malugin_Engl_29_7.pdf

General sequence annotation scheme. The GO term annotation pipeline was implemented using Linux bash scripts and the R programming language using the computational resources of the "Bioinformatics" collective use center at ICG SB RAS. Three algorithms for annotating protein functions based on the OrthoDB database were developed (Fig. 1).

On the left (Fig. 1*a*), the OrthoDB v 12.0 database (Tegenfeldt et al., 2025) is schematically shown in a large oval with representatives of orthologous groups (orthogroups) OG1...OG3 (Sequences of orthologous families are shown as rectangles of the same color). The first, basic sequence prediction algorithm is based on the search for *k*-nearest homologs and is denoted as KNN. Using the USEARCH program, homologous sequences are searched for the analyzed sequence in the OrthoDB database and ranked by similarity level. They can include representatives of both the same orthogroup and others (shown in different colors). The analyzed sequence is assigned the GO terms of the *k* most similar sequences (Fig. 1*b*).

The second method is based on the principle of orthology and is denoted as OG. For each of the *k*-nearest homologs of the analyzed sequence, its orthogroup in the OrthoDB database is determined. The orthogroup to which the analyzed sequence belongs is determined by a voting method: it is the orthogroup with the highest frequency of occurrence among all *k*-nearest homologs (Fig. 1*c*). GO terms for sequences from this orthogroup are assigned to the analyzed sequence (Fig. 1*d*).

The third approach, denoted as KNN+OG (Fig. 1*e*), involves combining the GO terms obtained from the KNN and OG algorithms for the query sequence (Fig. 1*f*). This list of GO terms is compared with the reference (true) annotation using measures such as: precision, recall (sensitivity), accuracy, and F-score (F-measure), which was the resulting measure (Fig. 1*g* and "Verification of terms using machine learning methods" section).

Methods for annotating the analyzed sequence with GO terms. *K-nearest homologs method* (KNN). The *k*-nearest homologs by similarity level are determined as a result of searching the OrthoDB database with the USEARCH program with the following parameters: identity (amino acid sequence identity) = 50 %, coverage (coverage of the analyzed sequence by the found homolog) = 70 %, *e*-value (statistical significance of the found match) = 10^{-6}, which is justified by the goal of reducing false positives at the homolog search stage.

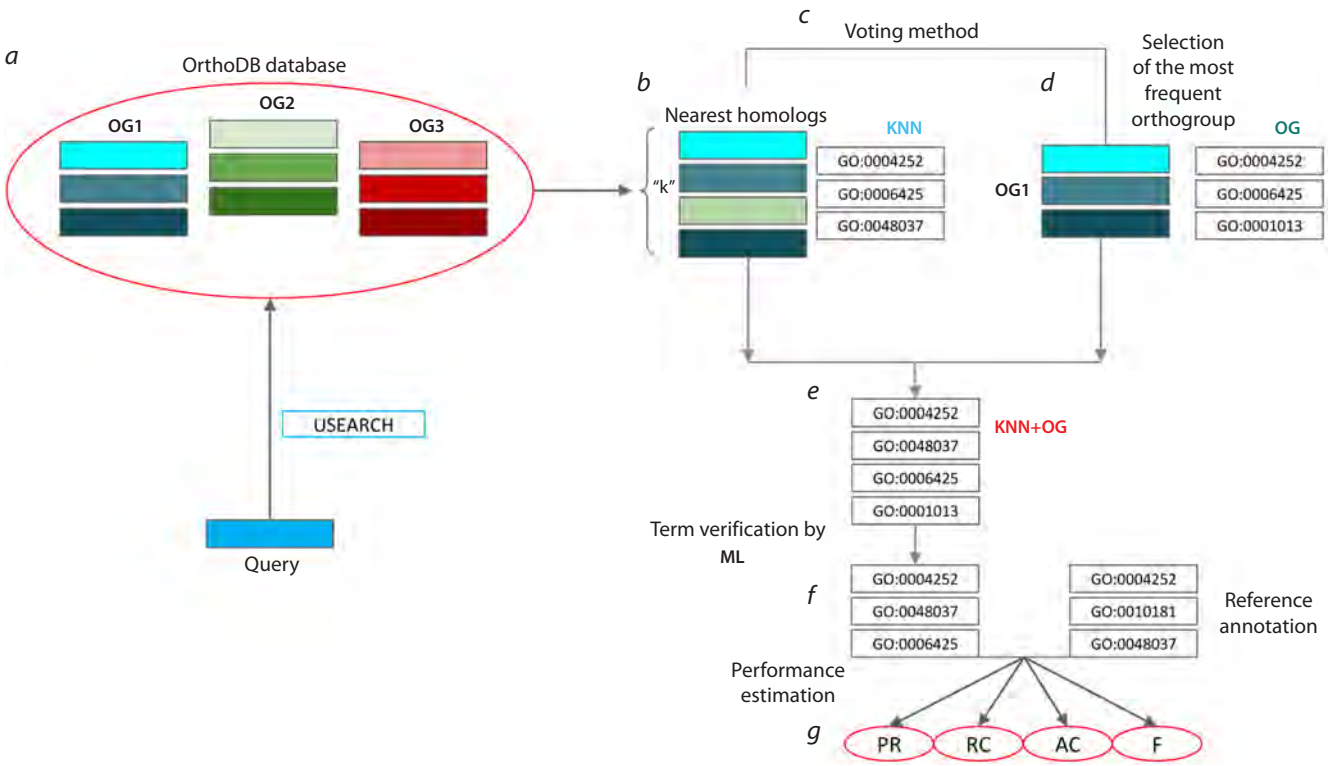


Fig. 1. General scheme of sequence annotation and its evaluation. Sequences belonging to the same orthogroup are represented by different shades of the same color: blue, green, or red.

a – OrthoDB database with orthogroups; *b* – assignment of GO terms from the *k*-nearest homologs (KNN method); *c* – determination of the most frequent orthogroup by voting; *d* – assignment of GO terms associated with the selected orthogroup (OG method); *e* – combination of GO terms obtained by the KNN and OG methods (KNN+OG method); *f* – verification of the combined term list using machine learning; *g* – comparison of predicted terms with the reference annotation and calculation of metrics.

The analyzed sequence was assigned the GO terms of the *k* most similar sequences available in the OrthoDB database. The value of parameter *k* can vary (Kharsikar et al., 2007; Dongardive, Abraham, 2016). Therefore, the optimal value within the interval *k* = 1–30 with a step of 5 was determined based on the highest accuracy in term identification using the OrthoDB annotation (Tables S4–S9).

***Using orthologous groups* (OG).** In this method, for each of the *k*-nearest homologs identified by the KNN method, the orthologous group corresponding to the most ancient ancestral taxon was selected using the OrthoDB annotation. Then, the orthogroup with the highest frequency among the *k*-nearest homologs was determined and assigned to the analyzed sequence. GO annotation terms for sequences from this orthogroup in the OrthoDB database were assigned to the analyzed sequence. The KNN+OG method combines GO terms (excluding duplicates) obtained separately by the KNN and OG methods described above.

Verification of terms using machine learning methods. To refine the list of predicted GO terms at the third stage of analysis (Fig. 1*f*), three machine learning (ML) algorithms were employed: logistic regression (LR), gradient boosting (XGB), and random forest (RF). Note that this stage does not allow adding new terms to the annotation. Instead, it filters out terms for which the similarity parameters between the analyzed sequence and its homologs do not meet the specified criteria.

The logistic regression method (LR) is implemented in the built-in stats package (R Core Team, 2013) via the function glm (family = binomial). Logistic regression predicts the probability of an object belonging to a class (e. g., "spam" or "not spam"). It predicts the probability of an object belonging to a class based on a weighted sum of features and passes it through a logistic (sigmoid) function, which normalizes the result to a number (probability) between 0 and 1. Gradient boosting (XGB – eXtreme Gradient Boosting) was used in the variant implemented in the xgboost package (Chen, Guestrin, 2016), function xgb.train. The random forest method (RF) was applied in the version from the randomForest package (Liaw, Wiener, 2002), function randomForest. Both gradient boosting and random forest are ensemble algorithms based on decision trees. This means that the final prediction is the result of the collective work of many individual decision trees. The parameters of the gradient boosting and random forest algorithms are specified in the Table S12.

Parameters for the models were selected during training, and in each method, their set was the same for all GO terms, analyzed sequences, and their homologs. These are terms reflecting the level of similarity, amino acid composition, and frequency of GO terms (Table S2). If a GO term in a homolog

was present in the annotation of the analyzed sequence in the training set, the prediction function value in the machine learning method was 1, otherwise, 0.

To evaluate the accuracy of machine learning methods, amino acid sequences of *Arabidopsis thaliana* and *Homo sapiens* proteins were used (Table 1). The set of sequences for each of these two species was divided into two parts: 80 % for training and 20 % for testing. Additionally, a combined sample of proteins from the organisms listed in Table 1 was formed: from the combined sample, 50,000 sequences were randomly selected for training, and 20,000 non-overlapping sequences were selected for testing the machine learning models (Table S3).

Evaluation metrics. Annotation accuracy evaluation was performed in R using the dplyr package (Wickham et al., 2025). For this, two lists were formed: (a) a reference list with amino acid sequences annotated with GO terms from databases for model organisms (more details in Table S1) and (b) a list obtained through functional annotation using various annotation methods (Fig. 1). To assess the accuracy of the annotation obtained by each of the methods described above, they were compared with the reference annotation. Hereafter, True Positive (TP) refers to GO terms present in both lists; False Positive (FP) refers to terms present in the predicted annotation list but absent in the reference (true) list; False Negative (FN) refers to terms present in the reference list but absent in the predicted annotation list.

The following metrics were used to evaluate protein annotation: Precision (PR), Recall (RC), Accuracy (AC), as well as the F-score metric, which was the resulting measure (Note. Here, "Accuracy (AC)" is a defined metric, distinct from the general concept of prediction accuracy):

Precision (PR) – the proportion of true positive predictions among all positive predictions of the method:

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100. \quad (1)$$

Recall (RC) – the proportion of true positive predictions among all true terms in the reference annotation:

$$\mathrm{RC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100. \quad (2)$$

Accuracy (AC) is defined as the arithmetic mean of Precision and Recall:

$$\mathrm{AC} = \frac{\mathrm{PR}+\mathrm{RC}}{2} \times 100. \quad (3)$$

F-score (F-measure) represents the harmonic mean between Precision and Recall. This metric approaches zero if either Precision or Recall approaches zero:

$$F1 = 2\,\frac{\mathrm{PR}\times\mathrm{RC}}{\mathrm{PR}+\mathrm{RC}} \times 100. \quad (4)$$

Since machine learning algorithms (LR, XGB, RF) estimate the probability of a GO term belonging to the analyzed sequence, and not a binary decision, it is necessary to choose a cutoff threshold (t) above which the term will be considered predicted. To account for data imbalance and to choose an optimal threshold independent of its specific value, the $F_{\max}$ metric was calculated for the cutoff threshold $t \in (0; 1)$ with a step of 0.1. A GO term was considered correctly predicted (positive class) if its predicted probability exceeded threshold t. $F_{\max}$ is defined as the maximum value of F-score(t) across all thresholds:

$$F_{\max} = \max\left\{2\,\frac{\mathrm{PR}(t)\times\mathrm{RC}(t)}{\mathrm{PR}(t)+\mathrm{RC}(t)}\right\} \times 100. \quad (5)$$

In GO term prediction tasks, where the distribution of terms by frequency of occurrence is extremely imbalanced (some terms are very common, others are extremely rare), and classification is multi-label (one protein can correspond to many terms), the $F_{\max}$ metric is often used. It is calculated for the entire set of predictions by varying the cutoff threshold (t), above which a term predicted by the ML model is considered positive. $F_{\max}$ shows the maximum quality that the model can achieve in the ideal case of threshold selection. Unlike the *F1*-score, which is calculated for a fixed threshold, $F_{\max}$ evaluates the quality of ranking terms by probability.

Comparison with other methods. To validate the developed OrthoML2GO method, it was compared with the Blast2GO (Conesa et al., 2005) and PANNZER2 (Törönen et al., 2018) methods. BLAST homology searches were launched on the computational complex of the "Bioinformatics" collective use center at ICG SB RAS. The launch parameters for Blast2GO and PANNZER2 were run with default parameters.

Results and discussion

Impact of orthogroup information on GO term prediction performance

To assess the influence of orthogroup information on function prediction performance, a comparison of the *F1*-score was conducted for three annotation methods with three algorithms (KNN, OG, and KNN+OG) depending on the number of nearest homologs for *A. thaliana* sequences (Fig. 2).

As shown in Figure 2, the *F1*-score depends on the parameter k for all three annotation variants. However, the nature of these dependencies is different: OG demonstrates the lowest performance (*F1* < 41 %). For the OG method, as for the other methods, a maximum is observed at k = 15. Moreover, increasing the parameter k results in a gradual, albeit slight, decrease in the *F1*-score. For the most accurate prediction, determining the correct orthologous group of the protein, which can be identified even at small values of k, is sufficient. A further increase in k only adds noise to the prediction due to an increase in false positive GO terms from orthogroups to which the protein in question does not actually belong.

The KNN method shows a pronounced dependence of performance on the parameter k. At small values (k = 5), the *F1*-score is the lowest (~40 %) and lower than the OG and KNN+OG methods, which is probably due to an insufficient number of homologs for reliable statistical inference and high sensitivity to noise and potential annotation errors of individual sequences. When k increases to 15, *F1* grows to a maximum value (~52 %); however, a further increase in k leads to a gradual decrease in performance, as distant homologs which may carry functionally irrelevant information for the target sequence (false positive GO terms) begin to enter the sample.

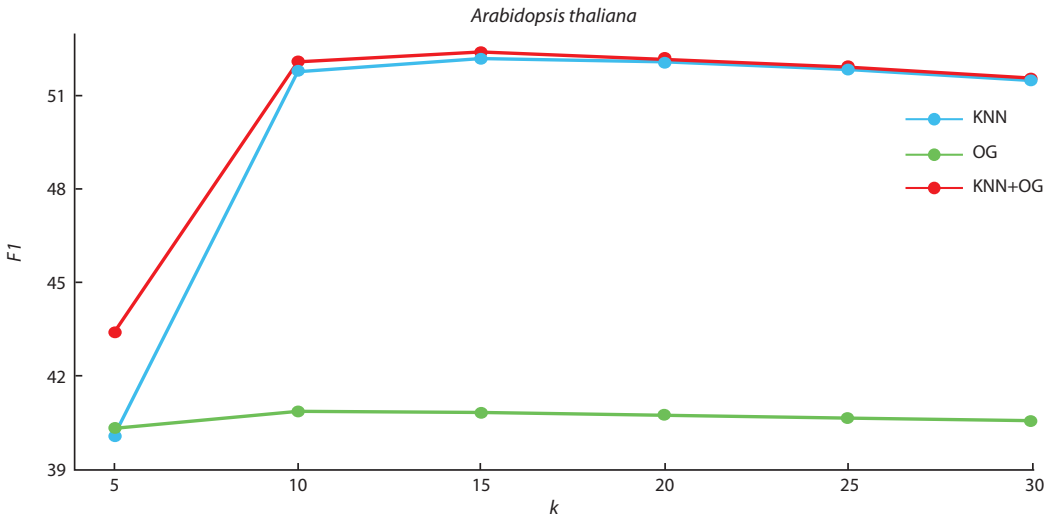


Fig. 2. Dependence of the *F1*-score in *A. thaliana* proteins on the parameter *k* (number of nearest homologs) for three annotation variants.

The *X*-axis shows the values of *k*. The *Y*-axis shows the values of the *F1*-score (in percent). Lines of different colors correspond to different annotation algorithms of our method: KNN – blue line; OG – green line; KNN+OG – red line.

Note that combining the KNN and OG methods (KNN+OG) leads to an increase in the *F1*-score for all values of the parameter *k*, and the greatest increase (more than 3 % in absolute value) is observed precisely at $k = 5$. This can be explained by the fact that with small *k*, the list of homologs may be unstable and statistically unreliable. Incorporating orthogroup information, which aggregates data on the function of a whole group of evolutionarily related genes, stabilizes the prediction and compensates for the insufficiency of data from a small number of nearest neighbors.

It is worth noting that the *F1*-score value in the range of 40–52 % represents a competitive result for the task of protein function prediction, as confirmed by comparison with other popular methods (see section "Comparison of the performance of KNN, KNN+OG, and OrthoML2GO with the Blast2GO and PANNZER2 tools"). This is due to the rather complex nature of the task: firstly, as mentioned earlier, GO annotation is multiple, i. e., one protein corresponds to many terms, and the prediction is considered correct only if all correct terms are found and no extra ones are added. Secondly, the distribution of GO terms is extremely imbalanced: some terms are very common, others are extremely rare, which further complicates achieving high accuracy. Thus, the absolute value of the *F1*-score should be interpreted in the context of the task's complexity and in comparison, to alternative approaches.

Results for other organisms are shown in the Supplementary materials (Tables S4–S9). Combining the KNN and OG methods (KNN+OG) allows us to obtain an integrated prediction that demonstrates the greatest gain in accuracy at small values of the parameter *k* for all organisms except *Chlamydomonas reinhardtii*. For example, for *Danio rerio* proteins at $k = 5$, the KNN+OG method surpasses the basic KNN by more than 13 % in absolute value of the *F1*-score (74.66 *vs.* 61.37 %). This is explained by the fact that with small *k*, the list of homologs may be statistically unreliable and sensitive to noise in the annotations of individual sequences. Integrating orthogroup data mitigates the statistical unreliability associated with a small number of nearest homologs. Thus, the hybrid KNN+OG approach not only demonstrates the best performance at the peak (at $k = 15$) but also significantly reduces the dependence of prediction accuracy on the parameter *k*, making the method more robust.

Thus, combining the KNN and OG variants (KNN+OG) allows obtaining an integrated prediction, giving a better estimate compared to each of them individually for all values of the parameter *k* for most organisms, and it will be used for machine learning.

Verification of GO terms by various machine learning algorithms

To verify false positive GO terms obtained at the previous stage, machine learning algorithms such as logistic regression (LR), gradient boosting (XGB), and random forest (RF) were used (see section "Verification of terms using machine learning methods"). A comparison of the accuracy of machine learning methods using the F_{max} measure (see section "Evaluation metrics") on test data of *A. thaliana*, *H. sapiens*, and a combined sample of 20,000 sequences from different organisms is presented in Table 2.

Logistic regression demonstrates significantly lower F_{max} values compared to gradient boosting and random forest methods, with the difference reaching over 25 %. This is likely due to the fact that ensemble methods (XGB and RF), unlike the linear LR model, are capable of capturing complex nonlinear relationships between features. Furthermore, these methods are more robust to noise in the data due to bagging (RF) and boosting (XGB) procedures, which average the predictions of many individual decision trees, reducing the influence of outliers and incorrect annotations of individual proteins. Gradient boosting (XGB) demonstrates the best results on *Arabidopsis* sequences and the general sample of all organisms, but it only slightly trails the random forest method

Table 2. Comparison of the F_{max} measure on test data for different machine learning algorithms, %

Dataset	LR	XGB	RF
Arabidopsis thaliana	53.20	**68.95**	66.86
Homo sapiens	71.92	**83.92**	84.02
Combined sample	52.25	**79.55**	78.32

on human proteins (with an F_{max} difference of only 0.1 %). Thus, for the final version of the OrthoML2GO method, the gradient boosting (XGB) machine learning method was chosen, as it showed the best results on the test samples.

Comparison of the performance of KNN, KNN+OG, and OrthoML2GO with the Blast2GO and PANNZER2 tools

For a comprehensive assessment of the developed method's effectiveness, its performance was compared with two widely used automatic functional annotation tools – Blast2GO and PANNZER2. The comparison was performed on three test datasets: individual proteomes of *A. thaliana* and *H. sapiens*, as well as a combined sample including sequences of all organisms listed in Table 1. As the resulting metric for methods not using machine learning (KNN, KNN+OG, Blast2GO), the *F1*-score was applied, while for OrthoML2GO and PANNZER2, which output a probabilistic estimate, the F_{max} metric was used, allowing us to evaluate the maximum achievable quality of the model with an ideal choice of cutoff threshold (Table 3).

Analysis of the results demonstrates that the developed OrthoML2GO method, integrating homology search, orthogroup analysis, and verification of GO terms using gradient boosting, shows a statistically significant advantage in performance over all compared methods on all test samples. Thus, for *A. thaliana*, OrthoML2GO achieved an F_{max} of 68.95 %. This represents an 18.21 % increase over PANNZER2 (F_{max} = 50.74 %) and a 14.65 % increase over the *F1*-score of Blast2GO (54.30 %). On human proteins, compared to PANNZER2, OrthoML2GO performed significantly better – 83.92 *vs.* 75.14 %, while for the Blast2GO method, the *F1* value was 54.95 %. On the combined sample of all organisms, an improvement in the F-measure indicator of more than 30 % was observed compared to all other methods.

Notably, the hybrid KNN+OG approach, which underlies OrthoML2GO, demonstrates a small but consistent improvement compared to the basic KNN on all samples, confirming the usefulness of integrating orthogroup information. However, the main gain in accuracy is provided by gradient boosting (XGB), which effectively verifies false positive predictions arising from annotation noise.

A key factor contributing to the success of the OrthoML2GO method is its integration of evolutionary information from homologs and orthogroups within the OrthoDB database, combined with subsequent verification of GO terms using gradient boosting. In contrast to PANNZER2 and Blast2GO, our method incorporates orthogroup information and verifies GO terms using decision tree ensembles, adaptively selecting the most informative features. Ultimately, this allowed reducing the proportion of false positive annotations and increasing accuracy from 8 % (on human protein sequences) to 30 % (on the combined sample) compared to analogues.

It is important to note a potential limitation in the comparison: our machine learning models were trained on a sample of sequences from OrthoDB, while Blast2GO and PANNZER2 rely on broader datasets derived from UniProt. This difference in training data may introduce a bias in the comparative accuracy estimates.

Assessment of prediction performance for different GO aspects

For a more detailed analysis of the method's performance, a comparative analysis of the prediction accuracy of GO terms for the three main aspects (ontologies) of Gene Ontology was performed: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). The evaluation results on the combined sample for various machine learning algorithms used at the verification stage are presented in Table 4.

The results show that all machine learning algorithms demonstrate a similar trend: the highest prediction accuracy is achieved for the Cellular Component (CC) aspect, followed by Molecular Function (MF), and the accuracy is somewhat lower for Biological Process (BP). This is consistent with the generally accepted view in bioinformatics: predicting cellular localization (CC) is often the easiest task, as it strongly correlates with the presence of specific signal peptides and domains. Prediction of molecular function (MF) also largely depends on conserved functional domains. At the same time, prediction of involvement in biological processes (BP) is the most complex, as the same protein can participate in several processes, and the processes themselves are defined by complex interactions of many proteins, which is more difficult to deduce solely from homology and orthology data.

The XGB method, chosen for OrthoML2GO, demonstrated the best results among all tested algorithms across all three

Table 3. Comparison of the methods KNN, KNN+OG, OrthoML2GO (XGB), PANNZER2 and Blast2GO on three datasets, %

Dataset	KNN*	KNN+OG*	OrthoML2GO (XGB)	PANNZER2	Blast2GO*
Arabidopsis thaliana	51.54	51.68	68.95	50.74	54.30
Homo sapiens	71.72	72.18	83.92	75.14	54.95
Combined sample	47.29	47.35	79.55	49.14	42.11

Note. For methods marked with an asterisk *, the *F1*-score is reported; for OrthoML2GO and PANNZER2, the F_{max} metric is used.

Table 4. Comparison of prediction performance for Gene Ontology aspects on the combined sample, %

Algorithm	BP Biological Process	MF Molecular Function	CC Cellular Component
LR	50.9	48.5	56.8
RF	78.4	77.0	82.9
XGB (OrthoML2GO)	78.8	79.8	83.6

Table 5. Estimates of GO term annotation accuracy for different aspects by various methods according to literature data, %

Method	BP	MF	CC	Reference
PANNZER2	78.4	85.8	85.3	Törönen et al., 2018
DeepGOPlus	58.5	47.4	69.9	Kulmanov, Hoehndorf, 2020
GOLabeler	58.6	37.2	69.1	You et al., 2018
NetGO 2.0	66.6	36.6	66.3	Yao et al., 2021
TALE+	66.7	45.9	67.7	Cao, Shen, 2021

aspects, further confirming its suitability as the final classifier. The performance of our method is competitive with the accuracy estimates of other methods reported in the literature (Table 5). The comparison was performed using the F_{max} metric for individual Gene Ontology aspects: BP – biological processes, MF – molecular functions, CC – cellular components.

It can be noted that the OrthoML2GO method (Table 4) demonstrated competitive results: 78.8 % (BP), 79.8 % (MF), and 83.6 % (CC) on a sample of 20,000 sequences from seven heterogeneous organisms – both plants and animals. Upon comparison, it is evident that OrthoML2GO surpasses most of the studied methods in all aspects. However, PANNZER2 showed higher values for MF (85.8 %) and CC (85.3 %), albeit on a smaller and less diverse sample (5,000 sequences from Swiss-Prot).

It is worth noting that direct quantitative comparison with other methods may be complicated by methodological differences. Firstly, test samples differ significantly: most methods use the UniProt/Swiss-Prot database, while our combined sample includes both plants and animals, which may affect the comparability of results. Secondly, the version of Gene Ontology is critically important: OrthoML2GO relies on the latest version of OrthoDB v12 annotation (GO 2025), which may lead to difficulties in comparing quality metrics.

To demonstrate the applicability of the OrthoML2GO method to poorly studied organisms, the proteome of the green alga *Ostreococcus lucimarinus* was annotated (Tables S10 and S11). The method predicted functions for 5,273 out of 7,603 protein sequences. The analysis revealed a predominance of such biological processes as phosphorylation (GO:0016310) and translation (GO:0006412). Among molecular functions, ATP binding (GO:0005524) and nucleotide binding (GO:0000166) were the most frequent, and among cellular components, membrane (GO:0016020) and nucleus (GO:0005634). These results demonstrate the method's ability to annotate poorly studied proteomes and identify functional profiles characteristic of non-model organisms.

Conclusion

The developed method, OrthoML2GO, which integrates homology searches and orthogroup analysis from the OrthoDB database with gradient boosting, demonstrated high efficiency on test samples. One of the main results is a significant improvement in annotation accuracy due to the combined approach, which combines the *k*-nearest neighbors method and information about orthologous groups (KNN+OG). This hybrid method surpassed the individual KNN and OG approaches, especially at small values of the parameter *k*. Verification of GO terms using machine learning algorithms, particularly gradient boosting (XGB), allowed for a further increase in accuracy through effective filtering of false positive predictions arising from distant homologs and orthogroups.

The obtained results confirm that the use of evolutionary information contained in the OrthoDB orthogroups, combined with machine learning algorithms, is an effective strategy for automatic prediction of protein sequence functions. The proposed OrthoML2GO method can be a good alternative to existing methods. It is worth noting that further improvement in accuracy is possible by optimizing machine learning parameters, as well as by including additional sources of biological information. As prospects for further research, the following directions are outlined: evaluation of the model's transferability to poorly annotated proteomes and comparative analysis with other methods using machine learning, including neural network-based ones.

References

Altenhoff A.M., Glover N.M., Dessimoz C. Inferring orthology and paralogy. *Methods Mol Biol*. 2019;1910:149-175. doi 10.1007/978-1-4939-9074-0_5

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi 10.1016/S0022-2836(05)80360-2

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., … Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1): 25-29. doi 10.1038/75556

Benso A., Di Carlo S., Ur Rehman H., Politano G., Savino A., Suravajhala P. A combined approach for genome wide protein function annotation/prediction. *Proteome Sci*. 2013;11(Suppl. 1):S1. doi 10.1186/1477-5956-11-S1-S1

Bradford Y.M., Van Slyke C.E., Ruzicka L., Singer A., Eagle A., Fashena D., Howe D.G., Frazer K., Martin R., Paddock H., Pich C., Ramachandran S., Westerfield M. Zebrafish information network, the knowledgebase for *Danio rerio* research. *Genetics*. 2022;220(4): iyac016. doi 10.1093/genetics/iyac016

Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59-60. doi 10.1038/nmeth.3176

Cao Y., Shen Y. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics*. 2021; 37(18):2825-2833. doi 10.1093/bioinformatics/btab198

Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD ‘16. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016;785-794. doi 10.1145/2939672.2939785

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.M., Li F.W., … Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K. 10KP: A phylodiverse genome sequencing plan. *Gigascience*. 2018;7(3):1-9. doi 10.1093/gigascience/giy013

Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18): 3674-3676. doi 10.1093/bioinformatics/bti610

Dongardive J., Abraham S. Protein Sequence Classification Based on N-Gram and K-Nearest Neighbor Algorithm. In: Behera H., Mohapatra D. (Eds). Computational Intelligence in Data Mining. Vol. 2. Advances in Intelligent Systems and Computing. Vol. 411. Springer, New Delhi, 2016;163-171 doi 10.1007/978-81-322-2731-1_15

du Plessis L., Skunca N., Dessimoz C. The what, where, how and why of gene ontology – a primer for bioinformaticians. *Brief Bioinform*. 2011;12(6):723-735. doi 10.1093/bib/bbr002

Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-2461. doi 10.1093/bioinformatics/btq461

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O. Protein function in the post-genomic era. *Nature*. 2000;405(6788):823-826. doi 10.1038/35015694

Fitch W.M. Distinguishing homologous from analogous proteins. *Syst Biol*. 1970;19(2):99-113. doi 10.2307/2412448

Fitch W.M. Homology a personal view on some of the problems. *Trends Genet*. 2000;16(5):227-231. doi 10.1016/s0168-9525(00)02005-9

Galperin M.Y., Koonin E.V. From complete genome sequence to ‘complete’ understanding? *Trends Biotechnol*. 2010;28(8):398-406. doi 10.1016/j.tibtech.2010.05.006

Gene Ontology Consortium; Aleksander S.A., Balhoff J., Carbon S., Cherry J.M., Drabkin H.J., Ebert D., ... Ponferrada V., Zorn A., Ramachandran S., Ruzicka L., Westerfield M. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224(1):iyad031. doi 10.1093/genetics/iyad031

Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi 10.1038/nrg.2016.49

Grigoriev I.V., Hayes R.D., Calhoun S., Kamel B., Wang A., Ahrendt S., Dusheyko S., Nikitin R., Mondo S.J., Salamov A., Shabalov I., Kuo A. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res*. 2021;49(D1):1004-1011. doi 10.1093/nar/gkaa898

Hamilton J.P., Brose J., Buell C.R. SpudDB: a database for accessing potato genomic data. *Genetics*. 2025a;229(3):iyae205. doi 10.1093/genetics/iyae205

Hamilton J.P., Li C., Buell C.R. The rice genome annotation project: an updated database for mining the rice genome. *Nucleic Acids Res*. 2025b;53(1):1614-1622. doi 10.1093/nar/gkae1061

Huntley R.P., Sawford T., Mutowo-Meullenet P., Shypitsyna A., Bonilla C., Martin M.J., O’Donovan C. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res*. 2015; 43(D1):1057-1063. doi 10.1093/nar/gku1113

Jensen L.J., Julien P., Kuhn M., von Mering C., Muller J., Doerks T., Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2008;36(Database issue): 250-254. doi 10.1093/nar/gkm796

Kharsikar S., Mugler D., Sheffer D., Moore F., Duan Z.H. A weighted k-nearest neighbor method for gene ontology based protein function prediction. In: Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS ‘07). IEEE Computer Society, USA, 2007;25-31. doi 10.1109/IMSCCS.2007.13

Kriventseva E.V., Rahman N., Espinosa O., Zdobnov E.M. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res*. 2008;36(Database issue):271-275. doi 10.1093/nar/gkm845

Kulmanov M., Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*. 2020;36(2):422-429. doi 10.1093/bioinformatics/btz595

Kuzniar A., van Ham R.C., Pongor S., Leunissen J.A. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24(11):539-551. doi 10.1016/j.tig.2008.08.009

Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R., …van Sluys M.A., Soltis P.S., Xu X., Yang H., Zhang G. Earth BioGenome project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*. 2018;115(17):4325-4333. doi 10.1073/pnas.1720115115

Liaw A., Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22. doi 10.32614/CRAN.package.randomForest

Öztürk-Çolak A., Marygold S.J., Antonazzo G., Attrill H., Goutte-Gattat D., Jenkins V.K., Matthews B.B., Millburn G., Dos Santos G., Tabone C.J.; FlyBase Consortium. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024;227(1):iyad211. doi 10.1093/genetics/iyad211

Pearson W.R. An introduction to sequence similarity (“homology”) searching. *Curr Protoc Bioinformatics*. 2013;42(3):3.1.1-3.1.8. doi 10.1002/0471250953.bi0301s42

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2013. Available: http://www.R-project.org/

Reiser L., Bakker E., Subramaniam S., Chen X., Sawant S., Khosa K., Prithvi T., Berardini T.Z. The Arabidopsis Information Resource in 2024. *Genetics*. 2024;227(1):iyae027. doi 10.1093/genetics/iyae027

Sanderson T., Bileschi M.L., Belanger D., Colwell L.J. ProteInfer, deep neural networks for protein functional inference. *eLife*. 2023;12: e80942. doi 10.7554/eLife.80942

Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026-1028. doi 10.1038/nbt.3988

Suzuki S., Kakuta M., Ishida T., Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and

a database suffix array. *PLoS One*. 2014;9(8):e103833. doi 10.1371/journal.pone.0103833

Tegenfeldt F., Kuznetsov D., Manni M., Berkeley M., Zdobnov E.M., Kriventseva E.V. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res*. 2025; 53(D1):D516-D522. doi 10.1093/nar/gkae987

Törönen P., Medlar A., Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res*. 2018;46(W1):W84-W88. doi 10.1093/nar/gky350

Wickham H., François R., Henry L., Müller K., Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2025. doi 10.32614/CRAN.package.dplyr

Yao S., You R., Wang S., Xiong Y., Huang X., Zhu S. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res*. 2021;49(W1):W469-W475. doi 10.1093/nar/gkab398

You R., Zhang Z., Xiong Y., Sun F., Mamitsuka H., Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*. 2018;34(14):2465-2473. doi 10.1093/bioinformatics/bty130

Yuan Q., Xie J., Xie J., Zhao H., Yang Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief Bioinform*. 2023;24(3): bbad117. doi 10.1093/bib/bbad117

Conflict of interest. The authors declare no conflict of interest.

Received July 24, 2025. Revised September 10, 2025. Accepted September 15, 2025.

Прием статей через электронную редакцию на сайте http://vavilov.elpub.ru/index.php/jour
Предварительно нужно зарегистрироваться как автору, затем в правом верхнем углу страницы выбрать «Отправить рукопись». После завершения загрузки материалов обязательно выбрать опцию «Отправить письмо», в этом случае редакция автоматически будет уведомлена о получении новой рукописи.

«Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» до 2011 г. выходил под названием «Информационный вестник ВОГиС»/ "The Herald of Vavilov Society for Geneticists and Breeding Scientists".

Сетевое издание «Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» – реестровая запись СМИ Эл № ФС77-85772, зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 14 августа 2023 г.

Издание включено ВАК Минобрнауки России в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, Russian Science Citation Index, Российский индекс научного цитирования, ВИНИТИ, Web of Science CC, Scopus, PubMed Central, DOAJ, ROAD, Ulrich's Periodicals Directory, Google Scholar.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте https://vavilovj-icg.ru/
и платформе Научной электронной библиотеки, elibrary.ru/title_about.asp?id=32440

англоязычная версия – на сайте vavilov.elpub.ru/index.php/jour
и платформе PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/journals/3805/

✉ email: vavilov_journal@bionet.nsc.ru

Издатель: Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук»,
проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Адрес редакции: проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Секретарь по организационным вопросам С.В. Зубова. Тел.: (383)3634977.
Издание подготовлено информационно-издательским отделом ИЦиГ СО РАН. Тел.: (383)3634963*5218.
Начальник отдела: Т.Ф. Чалкова. Редакторы: В.Д. Ахметова, И.Ю. Ануфриева. Дизайн: А.В. Харкевич.
Компьютерная графика и верстка: Т.Б. Коняхина, О.Н. Савватеева.

Дата выхода в свет 01.12.2025. Формат 60 × 84 1/8. Уч.-изд. л. 32.9