2024 • 28 • 8

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-24-88

2024 • 28 • 8

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Published 8 times annually
doi 10.18699/vjgb-24-88

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 • 2024 • 28 • 8

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

CONTENTS • 2024 • 28 • 8

Biomedicine

Вавиловский журнал генетики и селекции. 2024;28(8):807
doi 10.18699/vjgb-24-109

Уважаемые коллеги, дорогие читатели!

Представляем Вашему вниманию очередной выпуск «Вавиловского журнала генетики и селекции», посвященный вопросам биоинформатики и системной компьютерной биологии. С появлением массового секвенирования геномов началась разработка обширного арсенала методов биоинформатики для анализа структурно-функциональной организации геномов, генов и кодируемых ими РНК и белков, а также подходов системной компьютерной биологии, ориентированных на реконструкцию, анализ и моделирование генных сетей, контролирующих формирование фенотипических признаков организмов на основе информации, закодированной в геномах, и функционирующих на различных иерархических уровнях организации живых систем (начиная с геномов, генов, белков, метаболических путей и генных сетей, включая клетки и ткани, и заканчивая целостными организмами).

Методы биоинформатики и системной компьютерной биологии прочно вошли в арсенал исследовательских инструментов, используемых во всех областях наук о жизни. Однако в последнее десятилетие их значимость еще более возросла: в связи со стремительным развитием омиксных технологий (геномики, транскриптомики, протеомики, метаболомики) и других высокопроизводительных методов экспериментального исследования молекулярно-генетических систем и процессов в генетике произошел информационный взрыв. Она стала главным источником больших данных, перегнав по темпам роста не только все другие науки и технологии, но и мировые социальные сети. Огромные объемы и сложность накапливаемых в настоящее время больших генетических данных требуют создания информационно-программных комплексов, основанных на новом поколении методов биоинформатики и системной компьютерной биологии, использующих вычислительные конвейеры, реализующих сложные сценарии анализа и интегрирующих большое количество разнообразных программных продуктов и баз данных, полученных в том числе с помощью методов искусственного интеллекта.

Новая эра больших данных, в которую входят науки о жизни, требует трансформации базовых подходов биоинформатики и системной компьютерной биологии. В чем это проявляется? Во-первых, в широком применении методов искусственного интеллекта и их интеграции с классическими методами биоинформатики и системной компьютерной биологии. Во-вторых, в разработке на этой основе нового поколения информационно-программных систем для планирования экспериментов по проверке результатов компьютерных предсказаний, полученных при анализе больших генетических данных. Движение в этом направлении будет означать фактически смену базовой парадигмы исследований: от науки, направляемой гипотезами, к науке, направляемой большими данными. Биоинформатика и системная компьютерная биология, которым посвящен данный выпуск журнала, находятся на острие этого движения к новой биологии.

Научный редактор выпуска
академик Н.А. Колчанов,
научный руководитель ФИЦ ИЦиГ СО РАН

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):808-821

doi 10.18699/vjgb-24-89

Candidate SNP markers of changes in the expression levels of the human *SCN9A* gene as a hub gene for pain generation, perception, response and anesthesia

P.A. Dotsenko [1, 2, 3], K.A. Zolotareva [1], R.A. Ivanov[1], I.V. Chadaeva [1, 3], N.L. Podkolodnyy [1, 4], V.A. Ivanisenko [1, 2, 3], P.S. Demenkov [1], S.A. Lashin [1, 2], M.P. Ponomarenko [1, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[4] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
pon@bionet.nsc.ru

Abstract. In this work, we for the first time performed a comprehensive bioinformatics analysis of 568 human genes that, according to the NCBI Gene database as on September 15, 2024, were associated with pain generation, perception and anesthesia. The *SCN9A* gene encoding the sodium voltage-gated channel α subunit 9 and expressed in sensory neurons for transferring signals to the central nervous system about tissue damage was the only one involved in all the processes of interest at once as a hub gene. First, with our tool called OrthoWeb, we estimated the phylostratigraphic age indices (PAIs) for each of the genes, that is, identified the taxon of the most recent common ancestor of the organisms for which that gene has been sequenced. The mean PAI for all genes under study, including *SCN9A* as a hub gene for pain generation, perception, response and anesthesia, was '4'. On the evolutionary scale by the Kyoto Encyclopedia of Genes and Genomes (KEGG), the ancestor is the phylum Chordata, some of the most ancient of which evolved the central and the peripheral nervous system. Next, with our tool called ANDSystem, we found that phosphorylation of ion channels is a centerpiece in pain generation, perception, response and anesthesia, on which the efficiency of signal transduction from the peripheral to the central system depends. This conclusion was consistent with literature data on a key role an efficient signal transduction from the peripheral to the central system from the peripheral to the central system for adjusting the human circadian rhythm through detection of a change from the dark of night to the light of day and for identification of the direction of the source of sound by auditory brainstem nuclei, for generating the response to cold stress and for physical coordination. 21 candidate SNP marker of significant SCN9A over- and underexpression. Finally, the ratio of SCN9A upregulating to downregulating SNPs was compared to that for all known human genes estimated by the 1000 Genomes Project Consortium. It was found that *SCN9A* as a hub gene for pain generation, perception, pain response and anesthesia is acted on by natural selection against its downregulation, to keep the nervous system highly informed on the status of the organism and the environment.

Key words: human; TBP; SNP; promoter; hub gene; *SCN9A*; expression change; pain generation; pain perception; pain response; anesthesia.

For citation: Dotsenko P.A., Zolotareva K.A., Ivanov R.A., Chadaeva I.V., Podkolodnyy N.L., Ivanisenko V.A., Demenkov P.S., Lashin S.A., Ponomarenko M.P. Candidate SNP markers of changes in the expression levels of the human *SCN9A* gene as a hub gene for pain generation, perception, response and anesthesia . *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):808-821. doi 10.18699/vjgb-24-89

Funding. This work was funded by the Russian Federal Science and Technology Program for the Development of Genetic Technologies. The authors are grateful to the Shared Center "Bioinformatics" for access to computing resources under Budget Project FWNR-2022-0020.

Кандидатные SNP-маркеры изменения экспрессии гена *SCN9A* человека в качестве интегратора генерации, чувства, ответа на боль и анестезии

П.А. Доценко [1, 2, 3], К.А. Золотарева [1], Р.А. Иванов[1], И.В. Чадаева [1, 3], Н.Л. Подколодный [1, 4], В.А. Иванисенко [1, 2, 3], П.С. Деменков [1], С.А. Лашин [1, 2], М.П. Пономаренко [1, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[4] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
pon@bionet.nsc.ru

Аннотация. В настоящей работе впервые проведен комплексный биоинформатический анализ генов человека, связанных с генерацией, чувством и ответом на боль наряду с обезболиванием, которые были представлены 568 генами человека согласно базе данных NCBI Gene (дата обращения 15.09.2024). Ген *SCN9A* человека (sodium voltage-gated channel α subunit 9) передачи сигналов о повреждении тканей от сенсорных нейронов в центральную нервную систему был единственным среди исследуемых 568 генов, который вовлечен во все анализируемые процессы как ген-интегратор для них. Сначала с использованием созданного нами ранее инструмента OrthoWeb для каждого гена оценили таксон ближайшего общего предка всех организмов, у которого расшифрована ДНК этого гена (т. е. индекс филостратиграфического возраста, PAI). Среднеарифметическая оценка PAI для всех анализируемых генов, а также его значение для гена *SCN9A*, интегратора генерации, чувства и ответа на боль наряду с анестезией, оказались равными 4. На эволюционной шкале Киотской энциклопедии генов и геномов (KEGG) это соответствует таксону Chordata, у одних из самых древних представителей которого произошла специализация центральной и периферической нервной системы. Далее с помощью созданной нами системы ANDSystem мы выявили фосфорилирование ионных каналов как краеугольного камня в генерации, чувстве, ответе на боль и обезболивание, которое определяет эффективность передачи сигналов из периферической в центральную нервную систему. Этот вывод согласуется с литературными данными о ключевой роли эффективной передачи сигналов периферической нервной системы в центральную при коррекции циркадного ритма человека через фактическую детекцию фоторецепторами смены ночной темноты на дневное освещение, а также при определении направления на источник звука слуховыми ядрами мозга, формировании ответа на холодовой стресс и при координации движений у человека. Затем с использованием ранее созданной нами базы данных Human_SNP_TATAdb был предложен 21 кандидатный SNP-маркер значимого увеличения и уменьшения экспрессии гена *SCN9A* человека. Наконец, отношение встречаемости этих SNP-маркеров сравнили с полногеномным отношением, которое было оценено консорциумом «1000 геномов». В результате обнаружено, что *SCN9A* как ген-интегратор генерации, чувства, ответа на боль наряду с анестезией подвержен естественному отбору против снижения его экспрессии для поддержания высокого уровня контроля состояния организма и параметров внешней среды.

Ключевые слова: человек; TBP; SNP; промотор; ген-интегратор; *SCN9A*; изменение экспрессии; генерация боли; чувство боли; ответ на боль; анестезия.

Introduction

In 2020, the Council of the International Association for the Study of Pain (IASP) unanimously accepted the definition of pain as “An unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage” (Raja et al., 2020). Six accompanying notes were accepted to ensure the proper use of the term pain depending on the context (Raja et al., 2020). It was recommended that pain be conceived as an individual’s unpleasant emotional experience enhanced by biological, psychological and social factors. In addition, pain is not the same as the pulsing activity of the peripheral and the central nervous system’s sensitive nervous fibers excited by diverse stimuli and called “nociception”, “nociperception” or – in a narrower sense – physiological pain. The individuals develop the concept of pain as part of their personal experience. The IASP Council also recommended that the patients’ opinion about the pain they sense be considered. Although pain serves an adaptive role, it may have an adverse effect on social and psychological well-being as well as on the function of the human organism. Finally, the verbal description of pain is one of the many ways the individual can express this feeling and if he fails, chances are he may be is experiencing it nonetheless.

Considering the above, we focused on physiological pain, to be termed just “pain” throughout for brevity and because the term pain is used in this narrow sense by such renowned sources of scientific data as NCBI Gene (Brown et al., 2015) and Gene Ontology (Gene Ontology Consortium, 2015), on which we rely in this work.

Here we are for the first time conducting a comprehensive bioinformatics study of pain and anesthesia as a practical service in applied medicine, when patient treatment requires that both self-consciousness and awareness of the environment be reduced or eliminated by use of anesthetic drugs essential for organismic homeostasis, according to recommendations from the Association of Anaesthetists’ (Klein et al., 2021; Lucas et al., 2021). The need to explore further is so high that 49,305 and 3,782 original scientific papers related to pain and anesthesia, respectively, were collected in PubMed (Lu, 2011) as on September 15, 2024. With this in mind, we used our freely available web services OrthoWeb (Mustafin et al., 2020) and ANDSystem (Ivanisenko et al., 2015), and the Human_SNP_TATAdb database (Filonov et al., 2023) and analyzed 568 human genes associated with pain generation, perception, response and anesthesia, according to NCBI Gene (Brown et al., 2015) as on September 15, 2024. We verified our results against data from the independent web services PANTHER (Mi et al., 2021), DAVID (Sherman et al., 2022), STRING (Szklarczyk et al., 2023), Metascape (Zhou et al., 2019) and GeneMania (Warde-Farley et al., 2010), the ClinVar database (Landrum et al., 2014) and similar whole-genome results coming from the 1000 Genomes Project Consortium

(1000 Genomes Project Consortium et al., 2012), with Haldane's dilemma (Haldane, 1957) and the neutral theory of molecular evolution (Kimura, 1968) factored in.

Materials and methods

The human genes. A total of 568 human genes (n = 568) were studied. The list of the genes was generated by querying "*Homo sapiens*" *AND* "[gene key word]" in NCBI Gene (Brown et al., 2015) accessed on September 15, 2024. The activated filters were *Protein-coding genes*, *Genomic*, *Annotated genes*, *Ensembl* and *Current*, to return the most completely annotated protein-coding human genes.

The Phylostratigraphic Age Index (PAI) of the human genes. With OrthoWeb (Mustafin et al., 2020), we identified for each of the 568 genes all the biological species that had freely available orthologs to this gene and thus identified the most recent common ancestor of these species (Samet, 1985; Sun et al., 2008; Morozova et al., 2020), whose age served as the phylostratigraphic age indices (PAI) of the gene according to KEGG, the Kyoto Encyclopedia of Genes and Genomes (Kanehisa, Goto, 2000).

The associative network for pain generation, perception, response and anesthesia was reconstructed using ANDSystem (Ivanisenko et al., 2015). The results obtained were verified against the independent web services PANTHER (Mi et al., 2021), DAVID (Sherman et al., 2022), STRING (Szklarczyk et al., 2023), Metascape (Zhou et al., 2019) and GeneMania (Warde-Farley et al., 2010). The amount of consistency between the results coming from these web service and ANDSystem (Ivanisenko et al., 2015) was inferred by searching for the corresponding publications in PubMed (Lu, 2011).

Supervised annotation of the effects of changes in human gene expression levels on pain generation, perception, response and anesthesia. The effects of changes in *SCN9A* expression levels on pain generation, perception, response and anesthesia were explored by searching for the corresponding publications in PubMed (Lu, 2011).

The effects of single-nucleotide polymorphism (SNP) variants in the human gene promoters on the expression levels of these genes. The estimates of the statistical significance of the decrease or increase in the expression levels of the human genes for the minor *vs.* reference alleles of the SNP in the promoters of these genes were taken from the Human_SNP_TATAdb knowledge base (Filonov et al., 2023).

Verification of the estimations of the effects of SNPs in the human gene promoters on the expression levels of these genes. Selective verification of the *in silico* estimates of the effects of SNPs in the human gene promoters on the expression levels of these genes was performed using ClinVar (Landrum et al., 2014), PubMed (Lu, 2011) and literature data by the 1000 Genomes Project Consortium (Lowy-Gallego et al., 2019) for assessing the prevalence of such SNPs in the entire reference human genome, with Haldane's dilemma (Haldane, 1957) and the neutral theory of molecular evolution (Kimura, 1968) factored in.

Statistical analysis. The statistical criteria for the Kolmogorov–Smirnov test and the binomial distribution were tested using STATISTICA (Statsoft™, USA).

Results

SCN9A as a hub gene for pain generation, perception, response and anesthesia

We have herein worked on 568 human genes selected with NCBI Gene (Brown et al., 2015) (see Materials and methods). Of them, 553 were associated with pain; 231, with pain generation; 84, with pain perception; 39, with pain response; and 28, with anesthesia (Fig. 1*A*). The gene that is in red color font on the Venn diagram showing all possible overlaps between the gene groups (Fig. 1*A*) is *SCN9A*, the only gene shared by these groups. *SCN9A* encodes the sodium voltage-gated channel α subunit 9 and is expressed in sensory neurons for transferring signals to the central nervous system about tissue damage. Thus it was decided to consider *SCN9A* to be a hub gene for pain generation, perception, response and anesthesia.

The differences in PAI between the pain-generation-specific, perception-specific, response-specific and anesthesia-specific groups of genes do not reach statistical significance

We estimated the phylostratigraphic age index (PAI) for each of the 568 human genes. The histogram with the number of the genes being worked with within each of the 16 time intervals on the PAI scale according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Goto, 2000) is shown in the Figure 1*B*. The evolutionary estimates of the PAIs of the human genes associated with pain generation, perception, response and anesthesia, statistically significantly meet a normal distribution (Kolmogorov–Smirnov test: K = 1.03, $p < 0.05$). In line with the Central Limit Theorem (Kwak, Kim, 2017), which may imply that the PAI estimates reflect an integration of a great diversity of critical pain criteria. Considering this, we focused on *SCN9A* as a human hub gene for pain generation, perception, response and anesthesia.

The hypothetical link between the PAIs of the human genes associated with either pain generation or perception or response or anesthesia was verified using a box-and-whisker diagram for the overlaps between these groups of genes (Fig. 1*C*). The difference in PAI between the overlapped portions of the gene groups does not reach statistical significance, nor does it the difference between them and *SCN9A* as a hub gene for pain criteria in humans (Fig. 1*A*). That fact reinforced our confidence that *SCN9A* is worthy of our commitment.

The associative network for pain generation, perception, response and anesthesia

The associative network for *SCN9A* (Fig. 2) was constructed with ANDSystem (Ivanisenko et al., 2015). In the upper central part is the human gene *SCN9A*; in the lower central part, its encoded protein; in the middle central part, phosphorylation as a molecular-genetic process that is most mentioned in relation to this gene, as ANDSystem (Ivanisenko et al., 2015) suggests.

In the left-hand central part of the Figure 2 is *DPYSL2*, the only human gene associated with *SCN9A* itself, its encoded protein and phosphorylation. Additionally, in the left-hand bottom corner are four genes and their encoded proteins that interact with SCN9A, and in the left-hand upper corner

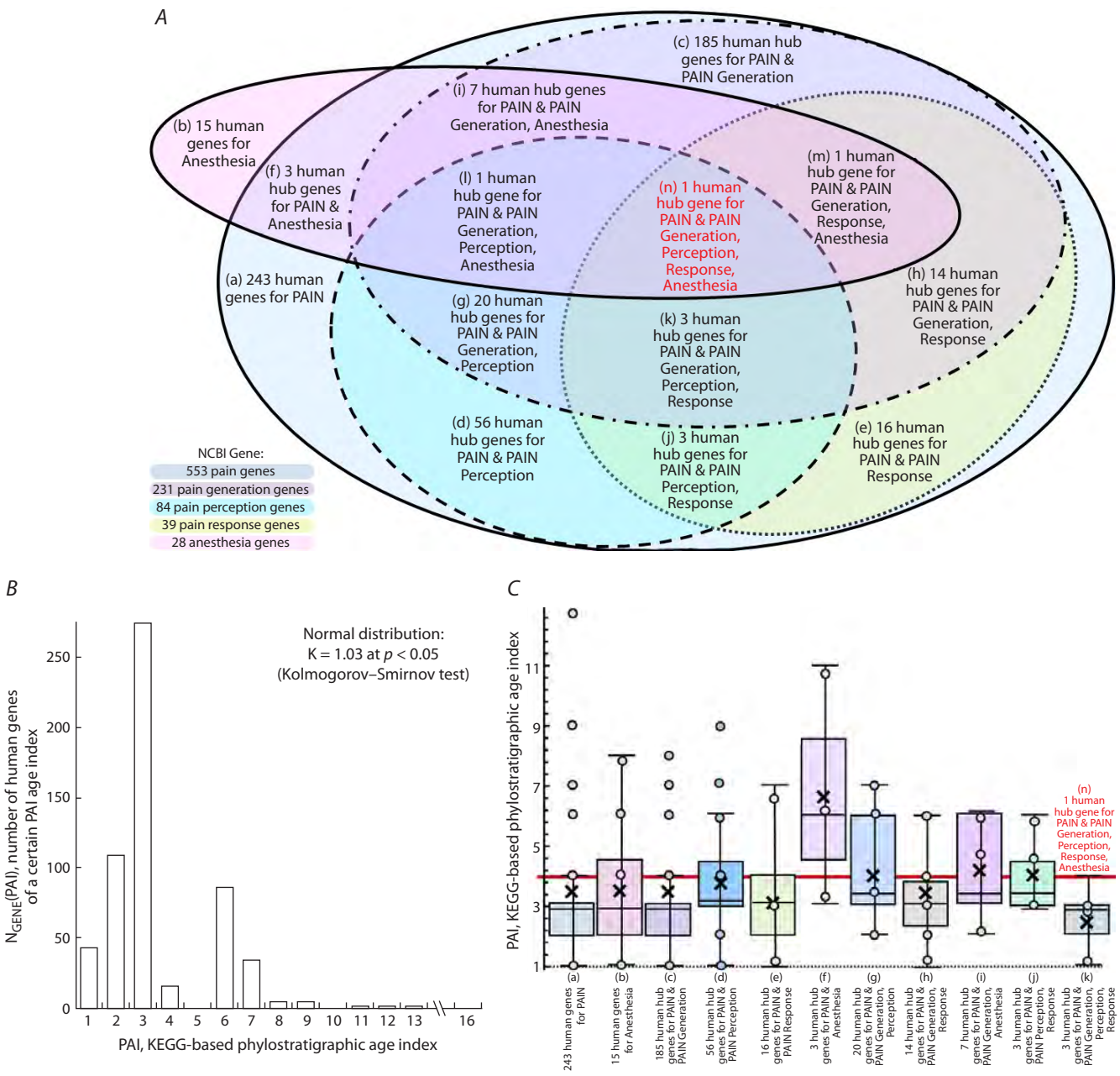


Fig. 1. The human genes returned by querying "*Homo sapiens" AND* "[gene key word]" in NCBI Gene (Brown et al., 2015) with *Protein-coding*, *Genomic*, *Annotated genes*, *Ensembl* and *Current* as activated filters.

A – Venn diagram for the 568 human genes: "[gene key word]" = "Pain" returned 553; "Generation of pain" – 231; "Perception of pain" – 84; "Response to pain" – 39; and "Anesthesia" – 28. *SCN9A*, the only human hub gene for pain generation, perception, response and anesthesia, is in red color font. *B* – the genes' phylostratigraphic age index (PAI) meeting a normal distribution (the Kolmogorov–Smirnov test K = 1.03, $p < 0.05$). *C* – the box-and-whisker diagram, where its height is its range from the 25 to the 75 % quartile, IQR; the line is the median, the 50 % quartile; the cross is the mean; the error bar "I" is the 95 % confidence interval; the circles are the genes. The PAI scale: 1 = Cellular organism, 4,100 Ma (Bell et al., 2015), 2 = Eukaryota, 1,850 Ma (Leander, 2020), 3 = Metazoa, 665 Ma (Maloof et al., 2010a), 4 = Chordata, 541 Ma (Maloof et al., 2010b), 5 = Craniata, 535 Ma (Maloof et al., 2010b), 6 = Vertebrata, 525 Ma (Shu et al., 1999), 7 = Euteleostomi, 420 Ma (Diogo, 2007), 8 = Mammalia, 225 Ma (Datta, 2005), 9 = Eutheria, 160 Ma (Luo et al., 2011), 10 = Euarchontoglires, 65 Ma(Kumar et al., 2013), 11 = Primates, 55 Ma (Chatterjee et al., 2009), 12 = Haplorrhini, 50 Ma (Dunn et al., 2016), 13 = Catarrhini, 44 Ma (Harrison, 2013), 14 = Hominidae, 17 Ma (Hey, 2005), 15 = *Homo*, 2.8 Ma (Schrenk et al., 2014), 16 = *Homo sapiens*, 0.35 Ma (Scerry et al., 2018).

are 11 human genes and their encoded proteins that interact with *SCN9A* and are involved in phosphorylation. The other 25 genes and their proteins interact with SCN9A and are involved in phosphorylation, too (Fig. 2, right). In total, Figure 2 shows 42 human genes, of which 14 were among the 568 genes associated with pain generation, perception, response and anesthesia (Fig. 1).

The overlap between the lists of 42 and 568 genes is statistically significant in terms of the reference human genome, which contains 19,424 annotated protein-coding genes, as suggested by NCBI Gene (Brown et al., 2015) as on August 20, 2024, with *Ensembl*, *Current*, *Protein-coding genes*, *Genomic* and *Annotated genes* as the activated filters: the binomial distribution at $p < 10^{-6}$.

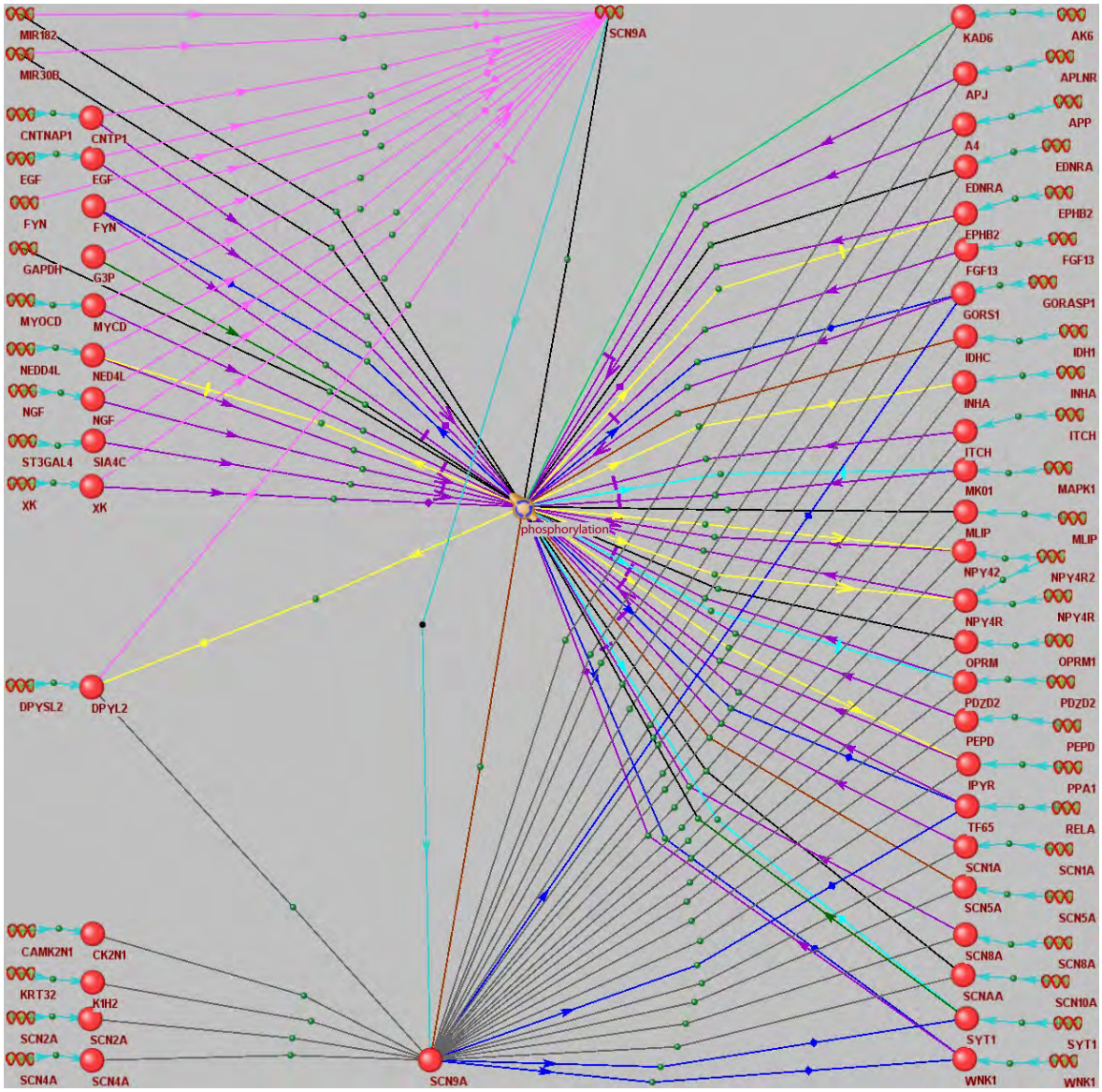


Fig. 2. The associative network of *SCN9A*, its encoded protein and their closest partners in the human organism. The network was constructed with ANDSystem (Ivanisenko et al., 2015) by automated analysis of freely available texts and database entries returned by querying "[list of genes] [immediate associations only] *Genes Proteins Pathway*" for [list of genes] = "*SCN9A*".

Legend: – gene; ● – protein; – phosphorylation as the most statistically significant biological process involving all the genes and proteins found ($P_{ADJ} < 10^{-13}$, Fisher's *Z* with the Bonferroni correction for multiple comparisons). Arrows: sharp-headed – activation; blunt-headed – inhibition; head-free – involvement; yellow – activity; dark-blue – transport; black – contact; purple – function; red – regulation; turquoise – expression.

This implies that ANDSystem (Ivanisenko et al., 2015) fed with *SCN9A* alone as a hub gene for pain generation, perception, response and anesthesia (Fig. 1) statistically significantly reconstructed the list of human genes (Fig. 2) that are associated with these processes in NCBI Gene (Brown et al., 2015).

Verification of the ANDSystem result against those on Gene Ontology term enrichment for the groups of genes by independent web services

A comparison between the result by ANDSystem (Ivanisenko et al., 2015) suggesting that phosphorylation is the most statistically significant biological process for pain generation, perception, response and anesthesia (Fig. 2) and the results by independent web services on Gene Ontology term enrichment for the groups of genes (Gene Ontology Consortium, 2015) is given in Table 1.

For example, as the upper row of that table suggests, for 42 human genes in the Figure 2, the web service PANTHER (Mi et al., 2021) revealed "GO:0086002 ~ cardiac muscle cell action potential involved in contraction" as the most statistically significant biological process involving these 42 genes ($P_{ADJ} < 10^{-9}$, statistical significance with a correction for multiple comparisons). The rightmost cell of this row contains a citation from an overview by V. Iyer et al. (2007): "Phosphorylation of the calcium channel augments Ca^{2+}

influx, which triggers a corresponding increase in Ca^{2+} release from the sarcoplasmic reticulum, thereby enhancing the force of contraction". In its sense, the results from ANDSystem (Ivanisenko et al., 2015) and PANTHER (Mi et al., 2021) for the 42 genes in Figure 2 are consistent.

In total, Table 1 shows 11 similar consistencies between the results coming from ANDSystem and five independent web services (PANTHER, DAVID, STRING, Metascape and GeneMania) about GO term enrichment for the groups of gene (Gene Ontology Consortium, 2015).

Table 1. A comparison between the result by ANDSystem (Ivanisenko et al., 2015) and the results by other web services on Gene Ontology term enrichment for the groups of genes (Gene Ontology Consortium, 2015)

	The most enriched GO term			The association between phosphorylation found by ANDSystem (Ivanisenko et al., 2015) and the best GO term found independently
	Web service	GO: ID	P_{ADJ}	
		Biological process		
1	PANTHER (Mi et al., 2021)	GO:0086002 ~ cardiac muscle cell action potential involved in contraction	10^{-9}	According to a comprehensive overview by V. Iyer et al. (2007), "phosphorylation of the calcium channel augments Ca^{2+} influx, which triggers a corresponding increase in Ca^{2+} release from the sarcoplasmic reticulum, thereby enhancing the force of contraction"
2	DAVID (Sherman et al., 2022)	GO:0086010 ~ membrane depolarization during action potential	10^{-9}	In a cellular model of pain using the human cell line HEK293T (Kerth et al., 2021): the I848T substitution in SCN9A creates a novel phosphorylation site, improving neuronal sensitivity and excitability due to an increased range (potential) of depolarization of the nerons' membrane
3	STRING (Szklarczyk et al., 2023)	GO:0043269 ~ regulation of ion transport	10^{-11}	In a biomedical tissue model of pain using a rat DRG culture (Stamboulian et al., 2010): Scn9A phosphorylation regulates ion transport by varying the activation threshold and the duration of inactivation of voltage-gated potassium channel
4	Metascape (Zhou et al., 2019)	GO:0044057 ~ regulation of system process	10^{-9}	Meta-analysis of freely available information resources and databases for traditional Chinese medicine (Shuyuan, Haoyu, 2023) pointed at "GO:0042327 ~ positive regulation of phosphorylation" and "GO:0044057 ~ regulation of system process" among the best GO terms characterizing the treatment of premature ventricular contractions by use of *Nardostachys jatamansi* radix and rhizoma
5	GeneMania (Warde-Farley et al., 2010)	GO:0034706 ~ sodium channel complex	10^{-18}	In a biomedical tissue model of pain using cerebellar Purkinje neurons acutely isolated from two-week-old mice (Grieco et al., 2002): constitutive phosphorylation of the sodium channel complex is required for making the blocking element functional for producing resurgent sodium current
		Molecular function		
6	PANTHER (Mi et al., 2021)	GO:0005248 ~ voltage-gated sodium channel activity	10^{-10}	In a subcellular model of pain using the human cell line HEK293T (Sokolov et al., 2018): *SCN9A* phosphorylation increases the conductance of this voltage-gated sodium channels for Na^+ ions
7	DAVID (Sherman et al., 2022)		10^{-9}	
8	STRING (Szklarczyk et al., 2023)		10^{-9}	
		Cellular component		
9	PANTHER (Mi et al., 2021)	GO:0001518 ~ voltage-gated sodium channel complex	10^{-12}	In a subcellular model of pain using the human cell line HEK293T (Sokolov et al., 2018): *SCN9A* phosphorylation promotes the association of the β3 subunit shifting the steady-state inactivation of the voltage-gated sodium channel to a more rapid recovery from inactivation within their complexes
10	DAVID (Sherman et al., 2022)		10^{-10}	
11	STRING (Szklarczyk et al., 2023)		10^{-10}	

Note. P_{ADJ} is the statistical significance of GO term enrichment for the groups of genes, with a correction for multiple comparisons used in the web service as indicated.

Table 2. Clinical implications of *SCN9A* downregulation and upregulation for pain generation, perception, response and anesthesia according to PubMed (Lu, 2011)

#	Process	Change in *SCN9A* expression levels	
		Downregulation	Upregulation
1	Pain generation	In a model of neuropathic pain using C57BL/6 mice (Palomes-Borrajo et al., 2021): treatment of an injured nerve with drug JQ1 reduced the pain generation frequency by downregulating *SCN9A*, which reduced the excitability of sensory neurons	According to a comprehensive overview by M.D. Baker and M.A. Nassar (2020): the mutation-induced growth in SCN9A activity increases the pain generation frequency due to an increased excitability of sensory neurons
2	Pain perception	In a biomedical model of pain using *Scn9a* KO mice (Shields et al., 2018): a reduction in the excitability of small- to medium-diameter sensory neurons due to a decrease in sodium TTX-sensitive channels in them	According to a comprehensive overview by S.D. Dib-Hajj et al. (2007): the mutation-induced growth in SCN9A activity reduces the activation threshold and slows down deactivation of voltage-gated sodium channels, which increases the excitability of sensory neurons and leads to erythromelalgia and paroxysmal extreme pain disorder
3	Pain response	In a model of neuropathic pain using C57BL/6 mice (Palomes-Borrajo et al., 2021): treatment of an injured nerve with drug JQ1 increased the response time to painful stimulus against the control with underexpressed *SCN9A*	In a model of spontaneous pain using transgenic CRISPR/Cas9 mice with the R185H mutation as a clinical marker of small fiber neuropathy (Xue et al., 2022): less time elapsed between exposure of the paw or tail to noxious heat and the animal's response to it
4	Anesthesia	In a biomedical model of pain using *Scn9a* KO mice (Shields et al., 2018): a reduction in *SCN9A* expression levels and inhibition of its encoded proteins may have a painkilling effect	In a meta-analysis of tumor transcriptomes compared to adjacent non-tumor tissues (Garate et al., 2021): *SCN9A* overexpression is a clinical marker of tumor reflecting a specific type of tumor pain and suggesting the need for analgesic therapy alongside traditional antitumor therapy (Cui et al., 2011)

The effects of changes in the expression levels of SCN9A as a hub gene on pain generation, perception, response and anesthesia

At this stage of our work, we sent text-based queries to PubMed (Lu, 2011) and thus performed a supervised annotation of *SCN9A* down- and upregulation by comparing them with literature data on the clinical manifestations of the changes in pain generation, perception, response and anesthesia (Table 2).

In Human_SNP_TATAdb (Filonov et al., 2023), we found 21 candidate SNP marker of a significant change in TBP affinity for the promoters of this gene and, consequently, a change in the expression levels of this gene (Table 3). Four of the 21 SNP marker of the significant change in *SCN9A* expression levels have known clinical implications (Table 3), as ClinVar (Landrum et al., 2014) suggests. It was demonstrated, with one of the four clinical SNP markers of pain, rs201905758:T as an example, (Fig. 3), how this SNP marker was detected by the web service SNP_TATA_Comparator (Ponomarenko et al., 2015) run in automated mode using the BioPerl library (Stajich et al., 2002) for access to Ensembl (Zerbino et al., 2015) and dbSNP (Day, 2010), the official repository of the reference human genome and the reference human variome, respectively. According to ClinVar (Landrum et al., 2014), four of the 21 SNPs were clinically proven markers of paroxysmal extreme pain disorder (PEPD), small fiber neuropathy (SFN), primary erythromelalgia (PE) and channelopathy-associated congenital insensitivity to pain (CIP) (Table 3).

As can be seen from the rightmost column "Δ" of Table 3, any of these four clinically proven markers of the *SCN9A* gene increases its expression levels as a hub gene for pain generation, perception, response and anesthesia. This encouraged us to perform a supervised PubMed-based annotation of the effects of *SCN9A* overexpression on pain generation, perception, response and anesthesia (Table 4). According to the many clinical overviews that have been written, for example (Dabby, 2012; Bennett, Woods, 2014; Shields et al., 2018; Taub, Woolf, 2024), SCN9A excess in PEPD, SFN and PE increases pain generation, perception and response, while low-molecular-weight inhibitors of *SCN9A* are anesthetics.

As far as CIP is concerned, according to clinical observations (Kim et al., 2015), secondary insensitivity to pain alternates with episodes of hypersensitivity to pain in PEPD, SFN and PE due to excessive SCN9A, this hypersensitivity being primary to insensitivity. It looks as if, because there were too many voltage-gated sodium channels in *SCN9A*, neural hyperexcitability depleted their battery now it needs to be recharged – to recover the previous levels of pain generation, perception and response. In this sense, all the *in silico* estimates of *SCN9A* overexpression with all clinically proven SNP markers of pain in PEPD, SFN, PE and CIP are consistent with the manifestation of excessive *SCN9A* in patients with these pathologies.

Comparison of the prevalence of the candidate SNP markers of changes in SCN9A expression levels against the whole-genome frequency of such SNPs

In conclusion, we compared the prevalence of the candidate SNP markers of changes in SCN9A expression levels (Table 3) with the frequency of such SNPs across the human genome

Table 3. Candidate SNP markers in the 90-bp proximal regions of the promoters before the transcription start sites of *SCN9A*, a human gene for pain integration, generation, perception, response and anesthesia, according to our *in silico* analysis as shown in the Figure 3 and documented in the Human_SNP_TATAdb database (Filonov et al., 2023)

#	Candidate SNP marker				K_D, nM, *in silico*		Significance			
	dbSNP ID:min (Day, 2010)	5′ flanking region	WT →min	3 flanking region	WT	min				
					$M_0 \pm SEM$	$M_0 \pm SEM$	Z	*p*	ρ	Δ
1	rs1341944281:G	gttttctaat	A→G	gttgatttcc	3.32±0.34	6.41±0.52	10.04	10^{-6}	A	
2	rs1470018720:C	ccgggcgcgc	T→C	ggggtgggga	86.89±6.63	104.57±8.03	3.42	10^{-3}	B	
3	rs1477103793:C	gcgcgctggg	A→C	tggggacccg	86.89±6.63	120.27±8.57	6.23	10^{-6}	A	↓
4	rs1559004384:G	atttcctgtt	T→G	tcattgtgtt	3.32±0.34	3.86±0.42	2.01	0.05	D	
5	rs933017443:C	gcggggctgc	T→C	ccctcgggga	56.13±5.20	120.27±8.57	13.04	10^{-6}	A	
6	rs1028575943:A	cgcgctggga	G→A	ggggacccgg	86.89±6.63	66.72±4.61	5.13	10^{-6}	A	
7	rs1038516207:A	gagtggagga	G→A	cgcgctggga	86.89±6.63	69.11±4.82	4.43	10^{-3}	B	
8	rs1282480960:G	ctaatattaa	C→G	tttcctgttt	3.32±0.34	2.67±0.27	3.03	10^{-2}	C	
9	rs1284056769:A	gagggagcaa	G→A	agggagggag	86.89±6.63	75.42±5.45	2.70	10^{-2}	C	
10	rs1343738748:T	gggagcaagg	G→T	ggagggaggg	86.89±6.63	63.02±5.10	5.78	10^{-6}	A	
11	rs1410144156:A	gctgggagga	G→A	gacccgggcg	86.89±6.63	71.32±4.53	3.98	10^{-3}	B	
12	rs1697331114:A	tgattattat	C→A	taagcaaaca	3.32±0.34	2.37±0.26	4.45	10^{-3}	B	
13	rs1700681124:T	gggctgctac	C→T	tcggggaggc	56.13±5.20	35.36±3.15	7.20	10^{-6}	A	
14	rs1700681309:A	gggaggcggg	G→A	agctgccctc	86.89±6.63	40.86±3.84	12.47	10^{-6}	A	
15	rs1700683197:A	agtggaggag	G→A	gcgctgggag	86.89±6.63	30.32±2.29	19.64	10^{-6}	A	↑
16	rs1700683375:A	gggaggagtg	G→A	ccgggcgcgc	86.89±6.63	60.80±4.59	6.65	10^{-6}	A	
17	rs890040570:A	cggcgcagct	G→A	aggaggcaaa	86.89±6.63	64.68±7.23	4.36	10^{-3}	B	
	ClinVar (Landrum et al., 2014): clinical SNP markers of paroxysmal extreme pain disorder, small fiber neuropathy, primary erythromelalgia and channelopathy-associated congenital insensitivity to pain									
18	rs148362057:A	gcagtctgct	T→A	gcaggagggg	91.71±6.30	41.82±3.93	13.50	10^{-6}	A	
19	rs1881440:T	gccctggcag	G→T	tccacgggcg	91.71±6.30	41.75±3.64	14.19	10^{-6}	A	
20	rs201905758:A	gctacctcca	C→A	gaggcggggc	56.13±5.20	47.62±4.43	2.51	0.05	D	
21	rs201905758:T	gctacctcca	C→T	gaggcggggc	56.13±5.20	43.97±4.65	3.48	10^{-3}	B	

Note. WT and min are the ancestral (norm) and the minor (pathology) alleles of the SNP, respectively; K_D is the equilibrium dissociation constant of the TBP–promoter complex expressed in nanomoles per liter (nM); M_0 and SEN are the context-dependent *in silico* estimate and its standard error, respectively; *Z*, *p* and ρ are the Fisher *Z* value and the level of its statistical significance as well as the heuristic prioritization of the *in silico* estimates from the best (A) to the worst (D) in alphabetic order; Δ – increase (↑) or decrease (↓) in *SCN9A* expression levels.

according to 1000 Genomes Project (Table 5). Individual human genomes possess an average of 1,000 SNPs each, of which 200 and 800 correspond, respectively, to an increase and a decrease in TBP-promoter affinity and eventually и to an increase and a decrease in the expression levels of human genes with these SNPs (Kasowski et al., 2010; 1000 Genomes Project Consortium et al., 2012).

In terms of Haldane's dilemma (Haldane, 1957) and the neutral theory of molecular evolution (Kimura, 1968), this prevalence of deleterious over beneficial regulatory SNPs signifies a neutral drift event, whish is statistically significantly different from the prevalence of 21 candidate SNP marker of changes in *SCN9A* expression levels ($p < 10^{-6}$, binomial distribution) (Table 5). This result implies that *SCN9A* is under natural selection against its downregulation, to keep the nervous system highly informed on the status of the organism and the environment.

Discussion

In this work, we for the first time performed a comprehensive bioinformatics analysis of 568 human genes that, according to the NCBI Gene database as on September 15, 2024, were

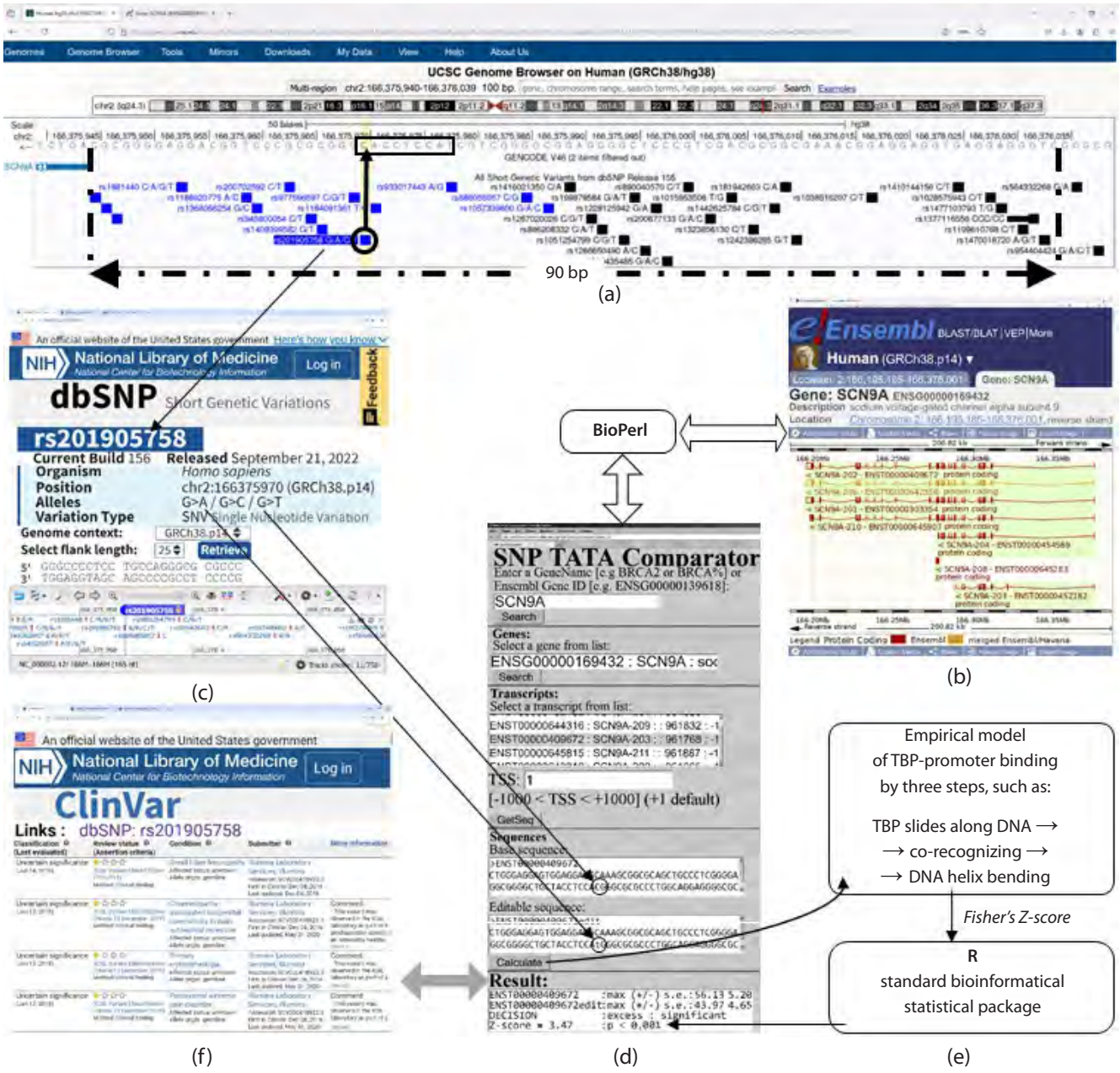


Fig. 3. An example: analysis of the candidate SNP marker rs201905758:T in the 90-bp proximal region (a two-headed dash-and-dot arrow in pane (a) before the start site of transcript SCN9A-203 from *SCN9A*, according to Ensembl (Zerbino et al., 2015), using SNP_TATA_Comparator (Ponomarenko et al., 2015).

Legend: (a) – visualization of the promoter being analyzed with the web service UCSC Genome Browser (Raney et al., 2024); (b) – the Ensembl database (Zerbino et al., 2015); (c) – description of SNP rs201905758 in the dbSNP database (Day, 2010); (d) and (e) – the use of SNP_TATA_Comparator and the principle of its operation, respectively (Ponomarenko et al., 2015); (f) – description of rs201905758:G→t, a clinically proven SNP marker for pain sensing pathology, according to ClinVar (Landrum et al., 2014).

associated with pain generation, perception and anesthesia. Our effort was strongly enabled by our freely available developments OrthoWeb (Mustafin et al., 2020), ANDSystem (Ivanisenko et al., 2015) and Human_SNP_TATAdb (Filonov et al., 2023). As a result, we identified *SCN9A* as being a hub gene for these biological processes (Fig. 1). Its Phylostratigraphic Age Index PAI = 4, according to the KEGG scale (Kanehisa, Goto, 2000), was not statistically different from the PAIs of the human genes associated with any of the combinations of the pain-related conditions in question and corresponded to the phylum Chordata, some of the most ancient of which evolved the central and the peripheral nervous system (Holland L.Z., Holland N.D., 1999).

Phosphorylation was found to be a key molecular genetic process in pain generation, response and anesthesia (Fig. 2). This result is consistent, first of all, with experimental data for a biomedical model of pain using the human cell line HEK293T (Kerth et al., 2021). C.M. Kerth and the co-workers found that the I → T substitution at position 848 of human protein SCN9A creates a novel phosphorylation site of this

Table 4. The effect of *SCN9A* overexpression on pain generation, perception and anesthesia in paroxysmal extreme pain disorder, primary erythromelalgia, small-fiber neuropathy and channelopathy-associated congenital insensitivity to pain, according to PubMed (Lu, 2011)

Process	Paroxysmal extreme pain disorder (PEPD)	Small-fiber neuropathy (SFN)	Primary erythromelalgia (PE)	Channelopathy-associated congenital insensitivity to pain (CIP)
Pain generation	According to a comprehensive overview (Drenth, Waxman, 2007): mutations leading to *SCN9A* gain-of-function in PEPD patients induce prolonged action potentials and repetitive neuron firing in response to exposure to cold or stretching	According to a comprehensive overview (Hoeijmakers et al., 2012), mutations leading to *SCN9A* gain-of-function in SFN patients have peripheral small-diameter axons generate pain and end up degenerated	Analysis of the pedigree of a Chinese family with PE (Wu et al., 2017) revealed the F826Y substitution leading to SCN9A gain-of-function and eventually to pain hypergeneration and insensitivity to painkillers	According to a clinical case at Centre Hospitalier Universitaire Sainte-Justine (Montreal), a 6-year-old girl born to healthy non-consanguineous French Canadian parents was found to have the I234T mutation enhancing the *SCN9A* function, of which the primary manifestation were PEPD and PE with multiple daily episodes of pain erythema affecting extremities and hidrosis and secondary CIP between these episodes because the voltage-gated sodium channels exceed the threshold polarization number in neuronal hyperexcitability, as if their "battery ran out of charge" (Kim et al., 2015), while all these symptoms were successfully managed by anesthesia against PEPD and PE. Such a paradoxical comorbidity of secondary CIP, on the one hand, and, on the other hand, PEPD, PE and SFN, primary to it, may have an extremely dangerous clinical manifestation in myotonia, such as paralysis following muscle hypercontraction (Kim et al., 2015). As to CIP as a primary pathology, it is clinically observed in patients with total loss of function in sodium ion channels, including this channel in *SCN9A* (Dabby, 2012; Bennett, Woods, 2014; Shields et al., 2018)
Pain perception	According to a comprehensive overview (Dabby, 2012): one of the most prevalent forms of clinical manifestation with mutations leading the *SCN9A* gain-of-function is a growth of pain perception in PEPD patients	According to a comprehensive overview (Taub, Woolf, 2024): with mutations leading to *SCN9A* gain-of-function, SFN patients feel ardor, tingling, heat and allodynia in the extremities. The prevalence grows with each year (Dabby, 2012)	According to a comprehensive overview (Dabby, 2012): with mutations leading to *SCN9A* gain-of-function, PE patients experience enhanced pain sensation as one of the most frequent forms of clinical manifestation of such mutations	
Pain response	According to a comprehensive overview (Stephenson, 2013): infants with PEPD are observed to be myotonic and have skin flushing with harlequin color change	In a biomedical model of SFN using transgenic fish *Danio rerio* with the artificial mutation I228M or G856D for Scn9A overexpression (Eijkenboom et al., 2019): larval activity grows with a rise in the environmental temperature	According to an overview (Renthal, 2020): ardor, body temperature rising, physical loading, tight cloths and footwear, hot and spicy food provoke episodes of ardor, heat and erythema in the PE patients' extremities and faces. The more severe PE, the more frequent the events	
Anesthesia	According to an overview (Hisama et al., 2020): the most efficient treatment in PEPD is with carbamazepine, a sodium channel blocker	In a comprehensive experimental and bioinformatics study of SFN (Shao et al., 2016), a context analysis of miRNA-30b showed that *SCN9A* mRNA may be its target, and the use of rats confirmed that miRNA-30b overproduction in an injured nerve decreases pain	In a pharmaceutical model of PE using human cell line HEK293A (Cregg et al., 2014): in low doses, mexiletine as a sodium channel blocker can normalize pain generation, perception and response	

protein, which is accompanied by an increase in neuronal sensitivity and excitability due to an increased range (potential) of depolarization of the neurons' membrane.

Additionally, the conclusion made about the importance of ion channel phosphorylation for pain generation, response and anesthesia is consistent to (Table 1) literature data about the importance of calcium channel phosphorylation in the myocardial cells (Iyer et al., 2007) and the importance of sodium channel phosphorylation in the cerebellar Purkinje neurons for physical coordination (Grieco et al., 2002).

Another example was found in PubMed (Lu, 2011): a cellular model of circadian rhythm using chick photoreceptors; under this model, increased phosphorylation of the ion channels in retinal cones in response to increased illumination the day offers after the dark of the night was the main event of the circadian rhythm in this model animal (Chae et al., 2007). The study of ophthalmic pathologies in rats revealed that phosphorylation of the ion channels in the optic nerve regulates visual system pathways (Ogata et al., 2022). Additionally, phosphorylation of the potassium channel in auditory

Table 5. A comparison between the prevalence of the identified candidate SNP markers of increased and decreased affinity of TBP to the *SCN9A* promoters (Fig. 3, Table 3) against whole-genome estimates according to the 1000 Genomes Project

Reference human genome: assembly GRCh38/hg38 (Lowy-Gallego et al., 2019), dbSNP build 155 (Day, 2010)		Number of objects in focus					Neutral drift (Haldane, 1957; Kimura, 1968)
		N_{GENE}	N_{SNP}	N_{Δ}	$N_{\downarrow}$	$N_{\uparrow}$	H_0 hypothesis: $N_{\downarrow} \geq 4N_{\uparrow}$ (Kasowski et al., 2010) binomial distribution, *p*
Prevalence of SNP markers of significant increase or decrease in TBP-promoter affinity	Whole-genome estimate for all human genes (1000 Genomes Project Consortium et al., 2012)	30,000	100,000	1,000	800	200	>0.50
	Partial estimate for *SCN9A* alone (this work)	1	37#	21*	5*	16*	$<10^{-6}$

Note. N_{GENE} – number of genes being worked with; N_{SNP} – number of SNPs being worked with; N_{Δ} – number of SNPs with ability to increase ($N_{\downarrow}$) and to decrease ($N_{\uparrow}$) TBP affinity to human gene promoters. # – see Figure 3, *A*; * – see Table 2.

neurons is basic to the ability to identify the direction of the source of sound due to microsecond delays in registering signals from it by auditory brainstem nuclei (Song et al., 2005). The phosphorylation levels of the SNAP-25 channel in the amygdala, cortex and hippocampus increased with the growth in the intensity of cold stress in mouse studies (Yamamori et al., 2014). Together, this provides a solid piece of evidence about a key role that ion channel phosphorylation has in the specialization into the central and the peripheral nervous system in general and during pain generation, perception, response and anesthesia.

At the final step, we used the Human_SNP_TATAdb database (Filonov et al., 2023) and proposed 21 candidate SNP marker of a significant change in the expression levels of *SCN9A* which encodes the sodium voltage-gated channel α subunit 9 and is expressed in sensory neurons for transferring signals to the central nervous system about tissue damage (Table 3). In ClinVar (Landrum et al., 2014), we found the descriptions of clinical *in vivo* manifestations for four of the 21 predicted SNP markers that were consistent with our *in silico* estimates (Table 4). A comparison between the prevalence of the SNPs identified in the *SCN9A* promoters and the whole-genome estimates according to the 1000 Genomes Project Consortium in 2012 leads to the conclusion that natural selection acts against *SCN9A* downregulation (Table 5), which indicates an adaptive role of pain and its perception as well as response to pain and anesthesia (Raja et al., 2020).

Overall, the results obtained are consistent with the independent authors', and in some cases refine and summarize them.

Conclusion

We have for the first time performed a comprehensive bioinformatics analysis of 568 human genes, which according to the NCBI Gene database (Brown et al., 2015) were associated with pain and anesthesia. From among them, we singled out *SCN9A*, the gene encoding the sodium voltage-gated channel α subunit 9 and expressed in sensory neurons for transferring signals to the central nervous system about tissue damage was the only one involved in all the processes of interest at once as a hub gene. With OrthoWeb (Mustafin et al., 2020), we estimated the phylostratigraphic age index (PAI) for *SCN9A*. It was "4", which corresponds to the phylum Chordata, some of the most ancient of which evolved the central and the peripheral nervous system (Holland L.Z., Holland N.D., 1999). The associative network of *SCN9A* was reconstructed using ANDSystem (Ivanisenko et al., 2015), where ion channel phosphorylation in *SCN9A* is a factor on which the efficiency of signal transduction from the peripheral to the central nervous system depends and which is a centerpiece in pain generation, perception, response and anesthesia. Finally, the search of the Human_SNP_TATAdb database (Filonov et al., 2023) revealed 21 candidate SNP marker of a significant change in *SCN9A* expression levels. The ratio of SCN9A upregulating to downregulating SNPs was compared to that for all known human genes (1000 Genomes Project Consortium et al., 2012). As a result, we for the first time obtained *in silico* whole-genome evidence that pain generation, perception, response and anesthesia (Raja et al., 2020) have an adaptive role, and their efficiency is controlled by natural selection.

References

1000 Genomes Project Consortium; Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., McVean G.A. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi 10.1038/nature11632

Baker M.D., Nassar M.A. Painful and painless mutations of *SCN9A* and *SCN11A* voltage-gated sodium channels. *Pflugers Arch*. 2020; 472(7):865-880. doi 10.1007/s00424-020-02419-9

Bell E.A., Boehnke P., Harrison T.M., Mao W.L. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. USA*. 2015;112(47):14518-14521. doi 10.1073/pnas.1517557112

Bennett D.L., Woods C.G. Painful and painless channelopathies. *Lancet Neurol*. 2014;13(6):587-599. doi 10.1016/S1474-4422(14)70024-9

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

Chae K.S., Ko G.Y., Dryer S.E. Tyrosine phosphorylation of cGMP-gated ion channels is under circadian control in chick retina photoreceptors. *Invest. Ophthalmol. Vis. Sci*. 2007;48(2):901-906. doi 10.1167/iovs.06-0824

Chatterjee H.J., Ho S.Y., Barnes I., Groves C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol*. 2009;9:259. doi 10.1186/1471-2148-9-259

Cregg R., Cox J.J., Bennett D.L., Wood J.N., Werdehausen R. Mexiletine as a treatment for primary erythromelalgia: normalization of biophysical properties of mutant L858F $Na_V1.7$ sodium channels. *Br. J. Pharmacol*. 2014;171(19):4455-4463. doi 10.1111/bph.12788

Cui J.H., Kim W.M., Lee H.G., Kim Y.O., Kim C.M., Yoon M.H. Antinociceptive effect of intrathecal cannabinoid receptor agonist WIN 55,212-2 in a rat bone tumor pain model. *Neurosci. Lett*. 2011; 493(3):67-71. doi 10.1016/j.neulet.2010.12.052

Dabby R. Pain disorders and erythromelalgia caused by voltage-gated sodium channel mutations. *Curr. Neurol. Neurosci. Rep.* 2012;12(1): 76-83. doi 10.1007/s11910-011-0233-8

Datta P.M. Earliest mammal with transversely expanded upper molar from the Late Triassic (Carnian) Tiki formation, South Rewa Gondwana Basin, India. *J. Vertebr. Paleontol*. 2005;25(1):200-207. doi 10.1671/0272-4634(2005)025[0200:EMWTEU]2.0.CO;2

Day I.N. dbSNP in the detail and copy number complexities. *Hum. Mutat*. 2010;31(1):2-4. doi 10.1002/humu.21149

Dib-Hajj S.D., Cummins T.R., Black J.A., Waxman S.G. From genes to pain: $Na_v1.7$ and human pain disorders. *Trends Neurosci*. 2007; 30(11):555-463. doi 10.1016/j.tins.2007.08.004

Diogo R. The Origin of Higher Clades: Osteology, Myology, Phylogeny and Evolution of Bony Fishes and the Rise of Tetrapods. NY: CRC Press, 2007. doi 10.1201/9780429063978

Drenth J.P., Waxman S.G. Mutations in sodium-channel gene *SCN9A* cause a spectrum of human genetic pain disorders. *J. Clin. Invest*. 2007;117(12):3603-3609. doi 10.1172/JCI33297

Dunn R.H., Rose K.D., Rana R.S., Kumar K., Sahni A., Smith T. New euprimate postcrania from the early Eocene of Gujarat, India, and the strepsirrhine-haplorhine divergence. *J. Hum. Evol*. 2016;99: 25-51. doi 10.1016/j.jhevol.2016.06.006

Eijkenboom I., Sopacua M., Otten A.B.C., Gerrits M.M., Hoeijmakers J.G.J., Waxman S.G., Lombardi R., Lauria G., Merkies I.S.J., Smeets H.J.M., Faber C.G., Vanoevelen J.M.; PROPANE Study Group. Expression of pathogenic *SCN9A* mutations in the zebrafish: a model to study small-fiber neuropathy. *Exp. Neurol*. 2019;311: 257-264. doi 10.1016/j.expneurol.2018.10.008

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 Journal of Genetics and Breeding*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85

Garate J., Maimo-Barcelo A., Bestard-Escalas J., Fernandez R., Perez-Romero K., Martinez M.A., Payeras M.A., Lopez D.H., Fernandez J.A., Barcelo-Coblijn G. A drastic shift in lipid adducts in colon cancer detected by MALDI-IMS exposes alterations in specific K+ channels. *Cancers* (*Basel*). 2021;13(6):1350. doi 10.3390/cancers13061350

Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43(D1):D1049-1056. doi 10.1093/nar/gku1179

Grieco T.M., Afshari F.S., Raman I.M. A role for phosphorylation in the maintenance of resurgent sodium current in cerebellar purkinje neurons. *J. Neurosci*. 2002;22(8):3100-3107. doi 10.1523/jneurosci.22-08-03100.2002

Haldane J.B.S. The cost of natural selection. *J. Genet*. 1957;55:511-524. doi 10.1007/bf02984069

Harrison T. Catarrhine origins. In: A Companion to Paleoanthropology. NY: Blackwell Publ. Ltd., 2013;376-396

Hey J. The ancestor's tale A pilgrimage to the dawn of evolution. *J. Clin. Invest*. 2005;115(7):1680. doi 10.1172/JCI25761

Hisama F.M., Dib-Hajj S.D., Waxman S.G. SCN9A Neuropathic pain syndromes. 2006 May 6 [Updated 2020 Jan 23]. In: GeneReviews®. [Internet]. Seattle (WA): Univ. of Washington, Seattle; 1993-2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1163/

Hoeijmakers J.G., Merkies I.S., Gerrits M.M., Waxman S.G., Faber C.G. Genetic aspects of sodium channelopathy in small fiber neuropathy. *Clin. Genet*. 2012;82(4):351-358. doi 10.1111/j.1399-0004.2012.01937.x

Holland L.Z., Holland N.D. Chordate origins of the vertebrate central nervous system. *Curr. Opin. Neurobiol*. 1999;9(5):596-602. doi 10.1016/S0959-4388(99)00003-3

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

Iyer V., Edelman E.R., Lilly L.S. Basic cardiac structure and function. In: Pathophysiology of Heart Disease. Baltimore: Lippincott Williams & Wilkins, 2007;1-28

Kanehisa M., Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi 10.1093/nar/28.1.27

Kasowski M., Grubert F., Heffelfinger C., Hariharan M., Asabere A., Waszak S.M., Habegger L., Rozowsky J., Shi M., Urban A.E., Hong M.Y., Karczewski K.J., Huber W., Weissman S.M., Gerstein M.B., Korbel J.O., Snyder M. Variation in transcription factor binding among humans. *Science*. 2010;328(5975):232-235. doi 10.1126/science.1183621

Kerth C.M., Hautvast P., Korner J., Lampert A., Meents J.E. Phosphorylation of a chronic pain mutation in the voltage-gated sodium channel Nav1.7 increases voltage sensitivity. *J. Biol. Chem*. 2021; 296:100227. doi 10.1074/jbc.RA120.014288

Kim D.T., Rossignol E., Najem K., Ospina L.H. Bilateral congenital corneal anesthesia in a patient with *SCN9A* mutation, confirmed primary erythromelalgia, and paroxysmal extreme pain disorder. *J. AAPOS*. 2015;19(5):478-479. doi 10.1016/j.jaapos.2015.05.015

Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968; 217(5129):624-626. doi 10.1038/217624a0

Klein A.A., Meek T., Allcock E., Cook T.M., Mincher N., Morris C., Nimmo A.F., Pandit J.J., Pawa A., Rodney G., Sheraton T., Young P. Recommendations for standards of monitoring during anaesthesia and recovery 2021: Guideline from the Association of Anaesthetists. *Anaesthesia*. 2021;76(9):1212-1223. doi 10.1111/anae.15501

Kumar V., Hallstrom B.M., Janke A. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PLoS One*. 2013;8(4):e60019. doi 10.1371/journal.pone.0060019

Kwak S.G., Kim J.H. Central limit theorem: the cornerstone of modern statistics. *Korean J. Anesthesiol*. 2017;70(2):144-156. doi 10.4097/kjae.2017.70.2.144

Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., Maglott D.R. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(D1):D980-D985. doi 10.1093/nar/gkt1113

Leander B.S. Predatory protists. *Curr. Biol*. 2020;30(10):R510-R516. doi 10.1016/j.cub.2020.03.052

Lowy-Gallego E., Fairley S., Zheng-Bradley X., Ruffier M., Clarke L., Flicek P.; 1000 Genomes Project Consortium. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res*. 2019;4:50. doi 10.12688/wellcomeopenres.15126.2

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

Lucas D.N., Russell R., Bamber J.H., Elton C.D. Recommendations for standards of monitoring during anaesthesia and recovery 2021. *Anaesthesia*. 2021;76(10):1426-1427. doi 10.1111/anae.15528

Luo Z.X., Yuan C.X., Meng Q.J., Ji Q. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature*. 2011; 476(7361):442-445. doi 10.1038/nature10291

Maloof A.C., Porter S.M., Moore J.L., Dudas F.O., Bowring S.A., Higgins J.A., Fike D.A., Eddy M.P. The earliest Cambrian record of animals and ocean geochemical change. *Geol. Soc. Am. Bull*. 2010a;122(11-12):1731-1774. doi 10.1130/B30346.1

Maloof A.C., Rose C.V., Beach R., Samuels B.M., Calmet C.C., Erwin D.H., Poirier G.R., Yao N., Simons F.J. Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nat. Geosci*. 2010b;3:653-659. doi 10.1038/ngeo934

Mi H., Ebert D., Muruganujan A., Mills C., Albou L.P., Mushayamaha T., Thomas P.D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. 2021;49(D1):D394-D403. doi 10.1093/nar/gkaa1106

Morozova O.V., Alekseeva A.E., Sashina T.A., Brusnigina N.F., Epifanova N.V., Kashnikov A.U., Zverev V.V., Novikova N.A. Phylodynamics of G4P[8] and G2P[4] strains of rotavirus A isolated in Russia in 2017 based on full-genome analyses. *Virus Genes*. 2020; 56(5):537-545. doi 10.1007/s11262-020-01771-3

Mustafin Z., Mukhin A., Afonnikov D., Matushkin Y., Lashin S. OrthoWeb – web application for macro- and microevolutionary analysis of genes. In: Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020). Novosibirsk, 2020; 228-229. doi 10.18699/BGRS/SB-2020-145

Ogata G., Partida G.J., Fasoli A., Ishida A.T. Calcium/calmodulin-dependent protein kinase II associates with the K^+ channel isoform Kv4.3 in adult rat optic nerve. *Front. Neuroanat*. 2022;16:958986. doi 10.3389/fnana.2022.958986

Palomes-Borrajo G., Badia J., Navarro X., Penas C. Nerve excitability and neuropathic pain is reduced by BET protein inhibition after spared nerve injury. *J. Pain*. 2021;22(12):1617-1630. doi 10.1016/j.jpain.2021.05.005

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed. Res. Int*. 2015; 2015:359835. doi 10.1155/2015/359835

Raja S.N., Carr D.B., Cohen M., Finnerup N.B., Flor H., Gibson S., Keefe F.J., Mogil J.S., Ringkamp M., Sluka K.A., Song X.J., Stevens B., Sullivan M.D., Tutelman P.R., Ushida T., Vader K. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain*. 2020;161(9): 1976-1982. doi 10.1097/j.pain.0000000000001939

Raney B.J., Barber G.P., Benet-Pagès A., Casper J., Clawson H., Cline M.S., Diekhans M., Fischer C., Navarro Gonzalez J., Hickey G., Hinrichs A.S., Kuhn R.M., Lee B.T., Lee C.M., Le Mercier P., Miga K.H., Nassar L.R., Nejad P., Paten B., Perez G., Schmelter D., Speir M.L., Wick B.D., Zweig A.S., Haussler D., Kent W.J., Haeussler M. The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res*. 2024;52(D1):D1082-D1088. doi 10.1093/nar/gkad987

Renthal W. Pain genetics. In: Rosenberg R.N., Pascual J.M. (Eds) Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease. Amsterdam: Elsevier, 2020;397-410. doi 10.1016/b978-0-12-813866-3.00023-0

Samet H. A top-down quadtree traversal algorithm. *IEEE Trans. Pattern Anal. Mach. Intell*. 1985;7(1):94-98. doi 10.1109/tpami.1985.4767622

Scerri E.M.L., Thomas M.G., Manica A., Gunz P., Stock J.T., Stringer C., Grove M., Groucutt H.S., Timmermann A., Rightmire G.P., d'Errico F., Tryon C.A., Drake N.A., Brooks A.S., Dennell R.W., Durbin R., Henn B.M., Lee-Thorp J., deMenocal P., Petraglia M.D., Thompson J.C., Scally A., Chikhi L. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol. Evol*. 2018;33(8):582-594. doi 10.1016/j.tree.2018.05.005

Schrenk F., Kullmer O., Bromage T. The earliest putative homo fossils. In: Henke W., Tattersall I. (Eds) Handbook of Paleoanthropology. Berlin: Springer, 2014;1-19. doi 10.1007/978-3-540-33761-4_52

Shao J., Cao J., Wang J., Ren X., Su S., Li M., Li Z., Zhao Q., Zang W. MicroRNA-30b regulates expression of the sodium channel Nav1.7 in nerve injury-induced neuropathic pain in the rat. *Mol. Pain*. 2016; 12:1744806916671523. doi 10.1177/1744806916671523

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

Shields S.D., Deng L., Reese R.M., Dourado M., Tao J., Foreman O., Chang J.H., Hackos D.H. Insensitivity to pain upon adult-onset deletion of Nav1.7 or its blockade with selective inhibitors. *J. Neurosci*. 2018;38(47):10180-10201. doi 10.1523/jneurosci.1049-18.2018

Shu D.-G., Luo H.-L., Conway-Morris S., Zhang X.-L., Hu S.-X., Chen L., Han J., Zhu M., Li Y., Chen L.-Z. Lower Cambrian vertebrates from south China. *Nature*. 1999;402(6757):42-46. doi 10.1038/46965

Shuyuan L., Haoyu C. Mechanism of Nardostachyos Radix et Rhizoma–Salidroside in the treatment of premature ventricular beats based on network pharmacology and molecular docking. *Sci. Rep*. 2023;13(1):20741. doi 10.1038/s41598-023-48277-0

Sokolov M.V., Henrich-Noack P., Raynoschek C., Franzen B., Larsson O., Main M., Dabrowski M. Co-expression of β subunits with the voltage-gated sodium channel $Na_V1.7$: the importance of subunit association and phosphorylation and their effects on channel pharmacology and biophysics. *J. Mol. Neurosci*. 2018;65(2):154-166. doi 10.1007/s12031-018-1082-6

Song P., Yang Y., Barnes-Davies M., Bhattacharjee A., Hamann M., Forsythe I.D., Oliver D.L., Kaczmarek L.K. Acoustic environment determines phosphorylation state of the Kv3.1 potassium channel in auditory neurons. *Nat. Neurosci*. 2005;8(10):1335-1342. doi 10.1038/nn1533

Stajich J.E., Block D., Boulez K., Brenner S.E., Chervitz S.A., Dagdigian C., Fuellen G., Gilbert J.G., Korf I., Lapp H., Lehvaslaiho H., Matsalla C., Mungall C.J., Osborne B.I., Pocock M.R., Schattner P., Senger M., Stein L.D., Stupka E., Wilkinson M.D., Birney E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 2002;12(10):1611-1618. doi 10.1101/gr.361602

Stamboulian S., Choi J.S., Ahn H.S., Chang Y.W., Tyrrell L., Black J.A., Waxman S.G., Dib-Hajj S.D. ERK1/2 mitogen-activated protein kinase phosphorylates sodium channel $Na_v1.7$ and alters its gating properties. *J. Neurosci*. 2010;30(5):1637-1647. doi 10.1523/jneurosci.4872-09.2010

Stephenson J.B. Syncopes and other paroxysmal events. *Handb. Clin. Neurol*. 2013;112:861-866. doi 10.1016/B978-0-444-52910-7.00007-6

Sun G.L., Shen W., Wen J.F. Triosephosphate isomerase genes in two trophic modes of euglenoids (euglenophyceae) and their phylogenetic analysis. *J. Eukaryot. Microbiol*. 2008;55(3):170-177. doi 10.1111/j.1550-7408.2008.00324.x

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

Taub D.G., Woolf C.J. Age-dependent small fiber neuropathy: mechanistic insights from animal models. *Exp. Neurol*. 2024;377:114811. doi 10.1016/j.expneurol.2024.114811

Warde-Farley D., Donaldson S.L., Comes O., Zuberi K., Badrawi R., Chao P., Franz M., Grouios C., Kazi F., Lopes C.T., Maitland A., Mostafavi S., Montojo J., Shao Q., Wright G., Bader G.D., Morris Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38(W2):W214-W220. doi 10.1093/nar/gkq537

Wu B., Zhang Y., Tang H., Yang M., Long H., Shi G., Tang J., Shi X. A novel SCN9A mutation (F826Y) in primary erythromelalgia alters the excitability of Nav1.7. *Curr. Mol. Med*. 2017;17(6):450-457. doi 10.2174/1566524017666171009105029

Xue Y., Kremer M., Muniz Moreno M.D.M., Chidiac C., Lorentz R., Birling M.C., Barrot M., Herault Y., Gaveriaux-Ruff C. The human *SCN9A*^{R185H} point mutation induces pain hypersensitivity and spontaneous pain in mice. *Front. Mol. Neurosci*. 2022;15:913990. doi 10.3389/fnmol.2022.913990

Yamamori S., Sugaya D., Iida Y., Kokubo H., Itakura M., Suzuki E., Kataoka M., Miyaoka H., Takahashi M. Stress-induced phosphorylation of SNAP-25. *Neurosci. Lett*. 2014;561:182-187. doi 10.1016/j.neulet.2013.12.044

Zerbino D.R., Wilder S.P., Johnson N., Juettemann T., Flicek P.R. The ensemble regulatory build. *Genome Biol*. 2015;16(1):56. doi 10.1186/s13059-015-0621-5

Zhou Y., Zhou B., Pache L., Chang M., Khodabakhshi A.H., Tanaseichuk O., Benner C., Chanda S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun*. 2019;10(1):1523. doi 10.1038/s41467-019-09234-6

Conflict of interest. The authors declare no conflict of interest.
Received September 23, 2024. Revised November 13, 2024. Accepted November 13, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):822-833

doi 10.18699/vjgb-24-90

MetArea: a software package for analysis of the mutually exclusive occurrence in pairs of motifs of transcription factor binding sites based on ChIP-seq data

V.G. Levitsky [1, 2], A.V. Tsukanov [1], T.I. Merkulova [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
levitsky@bionet.nsc.ru

Abstract. ChIP-seq technology, which is based on chromatin immunoprecipitation (ChIP), allows mapping a set of genomic loci (peaks) containing binding sites (BS) for the investigated (target) transcription factor (TF). A TF may recognize several structurally different BS motifs. The multiprotein complex mapped in a ChIP-seq experiment includes target and other "partner" TFs linked by protein-protein interactions. Not all these TFs bind to DNA directly. Therefore, both target and partner TFs recognize enriched BS motifs in peaks. A *de novo* search approach is used to search for enriched TF BS motifs in ChIP-seq data. For a pair of enriched BS motifs of TFs, the co-occurrence or mutually exclusive occurrence can be detected from a set of peaks: the co-occurrence reflects a more frequent occurrence of two motifs in the same peaks, while the mutually exclusive means their more frequent detection in different peaks. We propose the MetArea software package to identify pairs of TF BS motifs with the mutually exclusive occurrence in ChIP-seq data. MetArea was designed to predict the structural diversity of BS motifs of the same TFs, and the functional relation of BS motifs of different TFs. The functional relation of the motifs of the two distinct TFs presumes that they are interchangeable as part of a multiprotein complex that uses the BS of these TFs to bind directly to DNA in different peaks. MetArea calculates the estimates of recognition performance pAUPRC (partial area under the Precision–Recall curve) for each of the two input single motifs, identifies the "joint" motif, and computes the performance for it too. The goal of the analysis is to find pairs of single motifs A and B for which the accuracy of the joint A&B motif is higher than those of both single motifs.

Key words: *de novo* motif search; PR curve; area under curve; structural variants of transcription factor binding site motifs; cooperative action of transcription factors.

For citation: Levitsky V.G., Tsukanov A.V., Merkulova T.I. MetArea: a software package for analysis of the mutually exclusive occurrence in pairs of motifs of transcription factor binding sites based on ChIP-seq data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):822-833. doi 10.18699/vjgb-24-90

Funding. The work was supported by the Russian government project No. FWNR-2022-0020, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences.

Acknowledgements. The bioinformatics 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 budget project No. FWNR-2022-0020.

Программный комплекс MetArea для анализа взаимоисключающей встречаемости в парах мотивов сайтов связывания транскрипционных факторов по данным ChIP-seq

В.Г. Левицкий [1, 2], А.В. Цуканов [1], Т.И. Меркулова [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
levitsky@bionet.nsc.ru

Аннотация. Технология ChIP-seq, основанная на иммунопреципитации хроматина (ChIP), позволяет картировать набор геномных локусов (пиков), содержащих сайты связывания (СС) для исследуемого (целевого) транскрипционного фактора (ТФ). ТФ может распознавать несколько структурно различных мотивов СС. Мультибелковый комплекс, картируемый в эксперименте ChIP-seq, включает целевой и другие «партнерские» ТФ, связанные белок-белковыми взаимодействиями. Не все из этих ТФ связываются с ДНК напрямую. Поэтому и целевой, и партнерские ТФ распознают обогащенные мотивы СС в пиках. Для поиска обогащенных мотивов по

данным ChIP-seq применяется подход *de novo* поиска. Для пары обогащенных мотивов СС ТФ в наборе пиков может быть обнаружена совместная или взаимоисключающая встречаемость: совместная отражает более частое нахождение двух мотивов СС ТФ в одних пиках, а взаимоисключающая – в разных пиках. Мы предлагаем программный комплекс (ПК) MetAreа для выявления пар мотивов СС ТФ со взаимоисключающей встречаемостью по данным ChIP-seq. ПК MetAreа предназначен для предсказания структурного разнообразия мотивов СС одного ТФ и функциональной связи мотивов СС разных ТФ. Функциональная связь мотивов двух разных ТФ предполагает, что они взаимозаменяемы в составе мультибелкового комплекса, который использует СС этих ТФ для прямого связывания с ДНК в различных пиках. ПК MetAreа рассчитывает оценки точности распознавания pAUPRC (частичная площадь под кривой Precision–Recall) для каждого из двух входных одиночных мотивов, определяет их «объединенный» мотив и оценивает точность для него. Целью анализа является поиск пар одиночных мотивов A и B, для которых точность объединенного мотива A&B выше точностей обоих одиночных мотивов.

Ключевые слова: *de novo* поиск мотивов; кривая PR; площадь под кривой; структурные варианты мотивов сайтов связывания транскрипционных факторов; кооперативное действие транскрипционных факторов.

Introduction

Transcription factors (TFs) are proteins that have the ability to specifically bind DNA and thereby regulate gene transcription. About 1,600 human proteins are TFs (Lambert et al., 2018). TF binding sites (BSs) in eukaryotic genomic DNA are short regions, typically 6 to 20 base pairs (bp) in length (Vorontsov et al., 2024). TFs are usually able to bind not to a single DNA sequence, but to many similar ones. The TF BS motif in DNA is a general representation of the available diversity of such similar sequences (D'haeseleer, 2006). It is very difficult to establish clear patterns that determine the affinity of nucleotide sequences of genomic DNA to TFs. Only a few nucleotide positions are at least moderately conserved in TF BS motifs, i. e. they are unchanged in most natural BSs. Typically, the number of such positions is much less than a half of a motif length. The diversity of TF BS motifs *in vivo* is still very poorly studied because of the great variety of TF binding mechanisms to DNA. They include, in addition to direct binding, binding by other TFs or through them as intermediaries, use of the spatial structure of DNA within the nucleosome for binding, etc. (Morgunova, Taipale, 2017; Levitsky et al., 2020; Zeitlinger, 2020).

The most popular model of TF BS motifs is the traditional positional weight matrix (PWM) (Wasserman, Sandelin, 2004; Tognon et al., 2023). The PWM estimates the affinity of a site as the sum of the contributions (weights) of all its positions, where the weight of each position is defined by its nucleotide type. Alternative motif models are able to complement the predictions of the PWM model (Levitsky et al., 2007; Siebert, Söding, 2016; Tsukanov et al., 2022), i. e. to predict TF BSs in such genomic loci where the PWM model does not. The common difference between all alternative motif models and the traditional PWM model is the assessment of site affinity through the contribution of nucleotide frequency dependences between different motif positions.

DNA-binding domains (DBDs) provide TFs the ability to interact with DNA. The structure of a TF's DBD determines the variants of its BS motifs (Wingender, 2013; Lambert et al., 2018; Nagy G., Nagy L., 2020). Hierarchical classification of TFs based on the DBD structure in the TFClass database (Wingender, 2013; Wingender et al., 2013, 2015, 2018) defines classes of TFs based on their DBD structure. For example, the Hocomoco database (Vorontsov et al., 2024) annotates the BS motifs of 949 different human TFs. These TFs belong to 34 classes, but ten classes with at least ten TFs account for 858 TFs (more than 90 % of all 949 TFs), and the three largest classes, C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1}, and Basic helix-loop-helix factors (bHLH) {1.2} include 373, 184, and 76 TFs, respectively. The alignment of TF DBD sequences defines families and subfamilies of TFs below the classes in the hierarchy.

TFs of eukaryotes interact with DNA *in vivo* as part of multiprotein complexes including several TFs. TFs in such complexes are called "partner TFs", as there are protein-protein interactions between them. The common (cooperative) action of several TFs on the regulatory region of a gene is able to change the local environment of chromatin and regulate gene transcription (Morgunova, Taipale, 2017; Zeitlinger, 2020; Georgakopoulos-Soares et al., 2023). Many classes of TFs are characterized by the ability of TFs to bind to completely structurally different BSs (Rogers et al., 2019; Vorontsov et al., 2024). For example, TFs of the "Nuclear receptors with C4 zinc fingers {2.1}" class can bind as monomers and dimers. In the dimer case, the BS includes two half-sites; the spacer between them and the DNA strands of half-sites can vary. TFs of the "Basic leucine zipper factors (bZIP) {1.1}" class bind only as dimers, two half-sites are always located in the same DNA strand and the spacer is almost unchanged (Nagy G., Nagy L., 2020). Hereinafter, indices in curly brackets are labelled according to the TFClass database (Wingender et al., 2013, 2015, 2018). There are several types of DBDs of eukaryotic TFs that can function as dimers including pairs of closely related TFs (Amoutzias et al., 2008). TFs similar in DBD structure often recognize similar TF BS motifs (Lambert et al., 2018; Ambrosini et al., 2020), with the only clear exception to this rule being the BS motifs of TFs from the "C2H2 zinc finger factors {2.3}" class.

The identification of TF BSs in genomes has advanced significantly in the last 15 years with the advent of high-throughput massive sequencing methods, in particular, the experimental ChIP-seq technology. This technology gives for the target TF a set of genomic loci (peaks), usually several hundred bp in length, where the binding of the multiprotein complex of many TFs, including the target TF, has been experimentally mapped. Therefore, two types of peaks are responsible for direct and indirect binding of the target TF to genomic DNA. Direct binding means that the target TF is bound to DNA directly, and indirect binding means that the

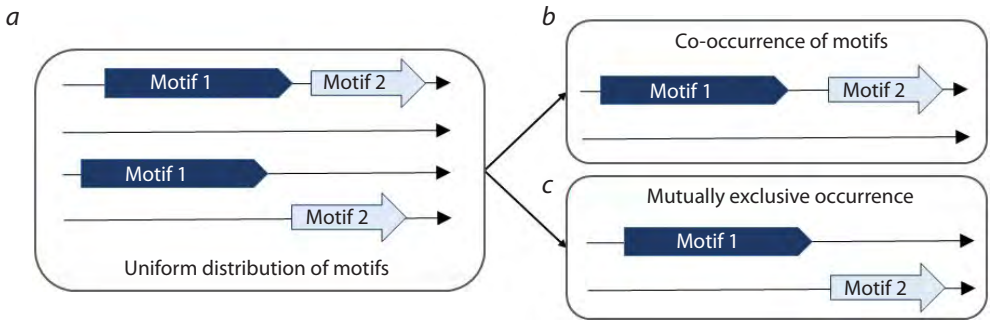


Fig. 1. Schema of the distinction between the terms of co-occurrence and mutually exclusive occurrence of TF BS motifs.

Let the frequency of occurrence of each of the two motifs in a peak be 50 %. *a* – the two motifs appear in the peaks independently of each other, there are four equally likely cases of motif mapping in the peaks; *b* – co-occurrence means that both motifs are in the same peak or neither of them is present; *c* – mutually exclusive occurrence denotes that only one of two motifs can be found in a peak. The arrows from panel *a* to panels *b* and *c* indicate that the four cases of panel *a* are exactly separated into two groups of two cases in panels *b* and *c*.

target TF is bound only by protein-protein interactions with one or more partner TFs, which in turn are bound to DNA directly. The presence of direct/indirect binding implies that the BS motifs of the target/partner TFs are enriched in the peaks, and the motifs of the target TFs are present only in part of the peaks. The term "enrichment" is used to reflect the increased content of TF BS motifs in genomic loci obtained from ChIP-seq massive sequencing data, i. e. increased content of TF BS motifs compared to their expected content due to random reasons. The negative set of DNA sequences is applied to estimate this expected motif content. We have shown that for ChIP-seq peaks, it is more efficient to select random genome loci matching the peaks in G/C-content into the negative set than to use synthetic sequences obtained from the peaks by nucleotide shuffling (Raditsa et al., 2024).

Once enriched BS motifs have been identified for a given ChIP-seq dataset of peaks, the analysis of statistical patterns of motif occurrences in pairs can identify the mechanisms of action of TFs. The concepts of synergy and antagonism of motifs within composite elements (CEs), as stable pairs of motifs, have been previously proposed (Kel et al., 1995). Synergy means that the result of the action of a pair of TFs is notably superior to that of each of them separately. Antagonism, on the contrary, implies that TFs impede each other. For example, one of two TFs is an activator and the other is a repressor, so that one displaces the other. Unfortunately, the concepts of synergy and antagonism refer to a stable pair of two motifs occurring in DNA, and these two cases cannot be distinguished by the frequencies of co-occurrence in the pair of motifs.

More than 15 years have passed since the era of massive sequencing of TF BS began (Jonhson et al., 2007); today, the role of bioinformatics analysis of whole-genome data in understanding the mechanisms of TF's action cannot be overestimated. In the case of ChIP-seq data, bioinformatics analysis does not deal with individual loci in the genome, but with a set of hundreds or even thousands of such loci where both direct and indirect binding of the target TFs can be observed. In moving from separate consideration of the frequencies of two TF BS motifs in a set of ChIP-seq peaks to observation of statistical patterns in their pairs, it is reasonable to consider two possibilities for these two motifs:

- they co-occur more frequently in the same peaks than it is expected by chance and less frequently occur separately in different peaks;
- they occur more often in different peaks and less often co-occur in the same peaks.

Therefore, we propose the terms of co-occurrence and mutually exclusive occurrence for the pair of TF BS motifs (Fig. 1).

Co-occurrence in a pair of motifs reflects the presence of a CE, a pair of closely located TF BS motifs in DNA, a small spacer between them, or they overlap (Kel et al., 1995; Levitsky et al., 2019). Mutually exclusive occurrence in a pair can have two explanations. Either it represents two structural types of the BS of the same TF (it binds differently in various peaks), or these two BSs belong two distinct TFs. Assuming that the two BS motifs correspond to two distinct TFs within the same multiprotein complex, we can propose that one TF interacting directly with DNA is replaced by another TF. Therefore, the trend of divergence of BS motifs of two TFs into different peaks may indicate a functional relationship of these motifs, in the simplest case representing the aforementioned substitution. For a co-occurrence, in the case of both synergy and antagonism, the two TFs bind to DNA in close proximity to each other (at least for some time they may be in contact even in antagonism), most likely they are within the same multiprotein complex. In the case of mutually exclusive occurrence, on the contrary, the BS motifs and the corresponding TFs are in distant DNA regions (different peaks). Therefore, we assume that the two motifs represent alternative traces of one common molecular function of TFs:

- the same TF recognizes two BS motifs of different structure, or
- binding to DNA occurs through distinct TFs and their BS motifs; these TFs are in the same multiprotein complex.

Figure 2 shows both these possibilities.

The AUC ROC (Area Under Curve) is the traditional quantitative measure of the accuracy of a binary classifier.

The term ROC stands for Receiver Operating Characteristic curve. For the TF BS motif, the ROC curve is defined as the dependence of the fraction of predicted sequences from the positive set (TPR, True Positive Rate) on the fraction of predicted sequences from the negative set (FPR, False Positive Rate). However, for TF BS motif recognition models in ChIP-seq data, it is more efficient to measure FPR as the expected frequency of a motif in the negative sequence set, but not as the fraction of predicted sequences for this set. This provides higher accuracy of assessment of motif model predictions at stringent and even medium recognition thresholds (Tsukanov et al., 2022). For the TF BS motif recognition model, the recognition accuracy can be calculated as the partial area under the ROC curve (pAUC ROC) (Tsukanov et al., 2022). The pAUC ROC value is equal to the fraction of the area under the curve bounded by the maximum allowable expected frequency of a motif. The area under the ROC curve integrates the fraction of peaks having the predicted TF BSs (the fraction of correctly predicted peaks, *Y* axis) over a wide range of recognition thresholds, calculated as the frequency of the motif in the negative set (*X* axis).

In this study, we propose the MetArea approach, which considers two separate "single" motifs as well as a "joint" motif, meaning the occurrence of either of the two single motifs. To predict a joint motif in a DNA sequence, it is sufficient to predict at least one of the two single motifs in it at a given threshold of expected motif frequency. Calculating the frequency of such a joint motif exactly even for a single DNA sequence poses an obstacle due to the huge variety of possible overlaps between single motifs. Therefore, to assess the accuracy of a motif model, we developed and applied the measure of accuracy "Partial area under the PR curve (Precision–Recall)". To calculate it we need only to track the number of recognized sequences in the positive and negative sets.

The PR curve is the dependence of the Precision measure (the ratio of the number of predicted sequences in the positive set to the number of predicted sequences in the positive and negative sets) on the Recall measure (the ratio of the number of predicted sequences in the positive set to the total number of sequence in this set). The PR curve is an alternative to the more popular ROC curve (Davis, Goadrich, 2006; Keilwagen et al., 2019). The advantage of the area under the PR curve measure over the area under the ROC curve measure is the ratio between the contributions of the mild and stringent recognition thresholds corresponding to the predicted sites of low and high affinity. Compared to the ROC curve, the PR curve provides greater contributions from high-affinity sites than from low-affinity sites. The ROC curve does the opposite. According to the PR curve, the contributions from sites with a low affinity may even tend to zero if such sites do not contain a specific nucleotide context. This is due to equal probabilities of site recognition in the positive and negative sets (Saito, Rehmsmeier, 2015).

We developed the MetArea software package (SP) to identify pairs of TF BS motifs with mutually exclusive occurrence. The MetArea SP calculates the partial area under the PR curve (pAUPRC) accuracy estimates for each of the two input single motifs as well as for their combination, the "joint motif". This allows the detection of mutually exclusive occurrence of these two input motifs.

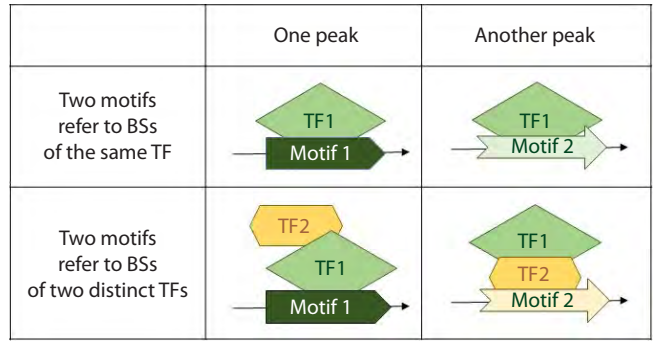


Fig. 2. Presumed origin of the mutually exclusive occurrence of two TF BS motifs in a set of ChIP-seq peaks.

The two columns represent two different peaks. Mutually exclusive occurrence in a pair of motifs could mean that either the pair of motifs represents two structurally distinct motifs of the same TF (this TF recognize these two motifs in different peaks), or the pair of motifs corresponds to BSs of different TFs. In this case, we assume that one TF interacting directly with DNA is replaced by another TF in some multiprotein complex (TF1 by TF2).

Materials and methods

ChIP-seq data from the GTRD database were used in the analysis (Kolmykov et al., 2021). For each ChIP-seq experiment, a set of 1,000 best quality peaks was analyzed according to preprocessing with the MACS2 tool (Zhang et al., 2008). In this study, enriched motifs obtained from the results of *de novo* motif search and mouse *Mus musculus* TF BS motifs from the Hocomoco database (https://hocomoco12.autosome.org/) (Vorontsov et al., 2024) were used in the analyses. *De novo* search for motifs of the traditional PWM and alternative SiteGA models of TF BS motifs was performed using STREME https://meme-suite.org/meme/tools/streme (Bailey, 2021) and https://github.com/parthian-sterlet/sitega (Tsukanov et al., 2022). The significance of similarity of the enriched motifs from the results of *de novo* search (STREME motifs) with the motifs of known TFs from the Hocomoco, Cis-BP (Weirauch et al., 2014) and JASPAR (Rauluseviciute et al., 2024) databases was assessed by the TomTom tool https://meme-suite.org/meme/tools/tomtom (Gupta et al., 2007). The MetArea SP also allows motifs from the Hocomoco and JASPAR databases to be selected for analysis according to the previously used approach (MCOT SP) (Levitsky et al., 2019). The best hit of a motif model has an expected frequency of at least 2E-5 in the set of promoters of all protein-coding genes of the genome. The best hit is given by the predicted site with the highest possible value of the recognition function of a motif model.

In total, the MetArea SP includes 1,420/1,142 motifs for 942/713 human/mouse TFs from the Hocomoco database, and 556/151 motifs for 555/148 plant/insect TFs from the JASPAR database. The MetArea SP is available at https://github.com/parthian-sterlet/metarea. For a detailed description of the MetArea SP algorithm, see the Results section below. The

MetArea SP implements the approach from the MCOT SP (Levitsky et al., 2019) to assess the similarity of the analyzed motifs of the PWM model (nucleotide frequency matrices).

Results

General description of the MetArea SP

The MetArea SP allows analyzing both pairs of motifs of the traditional PWM model and pairs of motives of the traditional PWM and alternative SiteGA models (Levitsky et al., 2007; Tsukanov et al., 2022). Figure 3 presents the general scheme of the MetArea SP pipeline.

The input data and parameters of the MetArea SP are listed below:

- Two motifs: (1) a combination of two motifs of the PWM model given by two nucleotide frequency matrices (NFMs), or (2) a combination of a motif of the PWM model given by an NFM and a motif of the SiteGA model given by its weight matrix, see https://github.com/parthian-sterlet/sitega (Tsukanov et al., 2022).
- Positive set in FASTA format (the set of ChIP-seq peaks, NF sequences, Number of Foreground sequences).
- Negative set in FASTA format (NB sequences, Number of Background sequences); it is recommended to prepare it in advance from the positive set and the whole genome by the AntiNoise SP (Raditsa et al., 2024), https://github.com/parthian-sterlet/antinoise. For each sequence of the positive set, several sequences of the negative set are selected randomly in the whole genome by its length and G/C-content. Further in the analysis, NF/NB = 5.
- The set of promoters of all genes of the genome is required to determine recognition thresholds based on the calculation of the Table 'Threshold vs. ERR' ("Recognition function threshold vs. Motif frequency in the set of all genome promoters") for each of the input motifs.
- The ERR_{MAX} threshold for the maximum expected motif frequency (Expected Recognition Rate, ERR) for each input motif.
- Tables 'Threshold vs. ERR' for each input motif.

The maximum motif frequency of 0.01 means that BS specificity corresponds to one site per one hundred nucleotide positions. The recommended range for the threshold of expected motif frequency ERR_{MAX} is 0.001 to 0.01. The ERR_{MAX} value of 0.002 is used below. We have previously used the 'Threshold vs. ERR' tables to set recognition thresholds across motifs (Levitsky et al., 2019; Tsukanov et al., 2021, 2022). Each motif and its 'Threshold vs. ERR' table are presented in a binary-format file generated by the MetArea SP components to calculate the expected motif frequencies for the PWM and SiteGA motif models.

The outputs of the MetArea SP are:

- A text file with PR curves for each of the input motifs as well as their joint motif.
- A text file with the values of pAUPRC recognition accuracy estimates for each of the input motifs, as well as for their joint motif, the value of the ratio of areas under the curves (see below), and the estimate of motifs' similarity (for pairs of PWM motifs only).

Definition of recognition thresholds for different motifs

The recognition function thresholds of each of the two input motifs, according to pre-calculated 'Threshold vs. ERR' tables, are transformed into a common scale of expected motif frequency, ERR (Levitsky et al., 2019; Tsukanov et al., 2021, 2022). This is necessary to construct the PR curve of the joint motif. The expected motif frequency ERR for the input motifs is calculated up to the threshold ERR_{MAX}, so that all expected frequencies satisfy the criterion: $ERR < ERR_{MAX}$.

The expected motif frequency in the promoter set was calculated as follows. The values of the motif recognition function

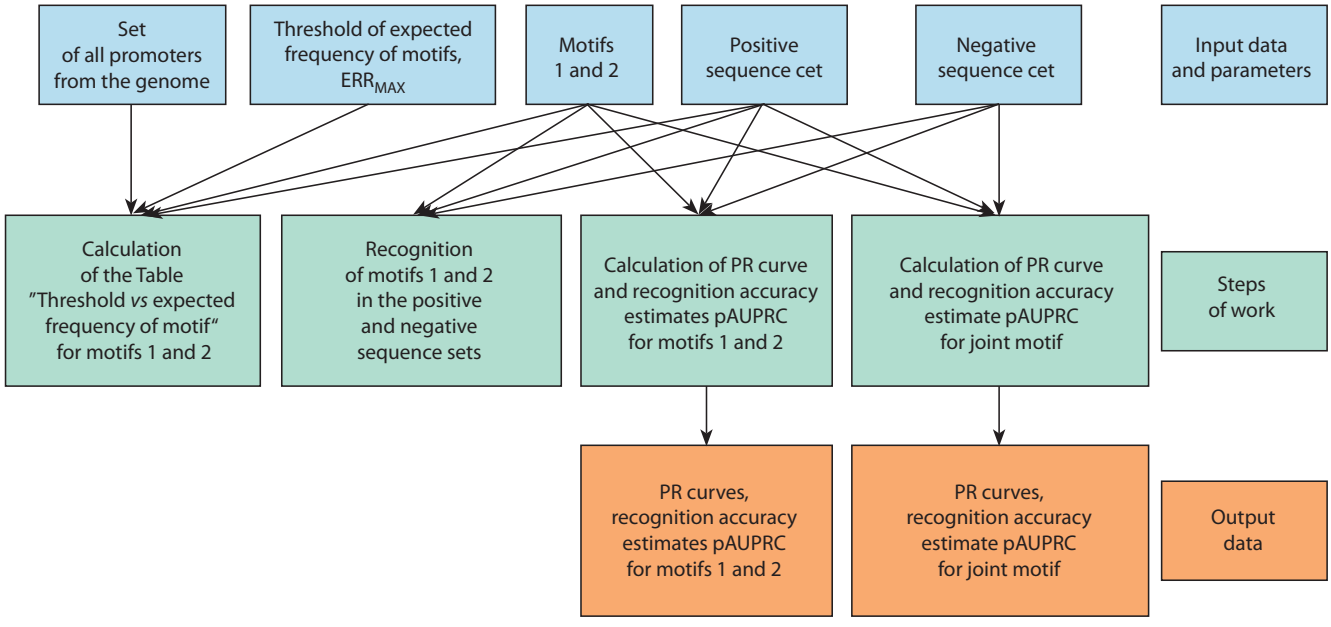


Fig. 3. General scheme of the MetArea SP pipeline.

for each predicted site in the set at each position and DNA strand were determined. Then, for each recognition threshold, the expected motif frequency was calculated as the ratio of the number of predicted BSs with the recognition function values equal to or higher than the recognition threshold to the total number of positions available for such BSs in the set in both DNA strands.

Statistical metrics and the PR curve

The PR curve (Davis, Goadrich, 2006) for the TF BS motif model can be defined as follows: the *X* axis means the ratio of the number of sequences from the positive set (peaks) with predicted sites to the number of all peaks (TPR, True Positive Rate, Recall, REC):

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \quad (1)$$

Here, TP/FN (True Positives/False Negatives) is the number of correctly/incorrectly predicted sequences from the positive set (TP + FN = NF).

The *Y* axis of the PR curve implies the ratio of the number of predicted sequences in the positive set to the number of all predicted sequences in positive and negative sets (Precision, PREC), according to (Davis, Goadrich, 2006):

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \quad (2)$$

Here, FP (False Positives) is the number of predicted sequences in the negative set. Taking into account the difference in the number of sequences between the positive (NF) and negative (NB) sets, we corrected the calculation of the Precision value as follows:

$$\mathrm{PREC} = \frac{\mathrm{TPR}}{\mathrm{TPR}+\mathrm{FPR}} = \frac{\mathrm{TP/NF}}{\mathrm{TP/NF}+\mathrm{FP/NB}} = \frac{\mathrm{TP}}{\mathrm{TP}+(\mathrm{NF/NB})\times\mathrm{FP}}. \quad (3)$$

Here, TPR and FPR are the fractions of predicted sequences in the positive and negative sets. The NF/NB coefficient takes into account the difference between the sizes of negative (NB) and positive (NF) sets. The expected numbers of predicted sequences of positive (TP) and negative (FP) sets due to random reasons are proportional to the set sizes, NF and NB, respectively. Hence, we introduce the NF/NB coefficient to unify the behavior of the PR curve for different ratios of positive and negative set sizes.

Partial area under the PR curve and the ratio of areas under curves

The MetArea algorithm uses the tables "Recognition function threshold vs. Motif frequency in the set of all genome promoters" described above, and performs recognition of two input single motifs in the positive and negative sets. Next, the pAUPRC measure is calculated for the single motifs as well as for the joint motif. The calculation of the partial area under the curve PR (pAUPRC) is limited by the criteria imposed on the Recall (*X* axis) and Precision (*Y* axis) measures, that is, the area is partial on both the *X* axis and the *Y* axis. The example in Figure 4 explains the choice of the partial area in both axes.

The criterion for the partial area under the PR curve on the *X* axis is the participation in the calculation of the pAUPRC measure of a part of the whole range of the Recall measure from 0 to 1. This criterion means that not all peaks with predicted sites are involved, but only those peaks, the best hits of which have an expected frequency below the threshold, $\mathrm{ERR} < \mathrm{ERR}_{\mathrm{MAX}}$ (Fig. 4). Here, we chose the milder threshold of the expected frequency ($\mathrm{ERR}_{\mathrm{MAX}} = 0.002$) than the one previously used to analyze the motifs of target TFs ($\mathrm{ERR}_{\mathrm{MAX}} = 0.001$) (Tsukanov et al., 2022). We previously analyzed the motifs of target TFs of ChIP-seq experiments, and the MetArea SP analyzes the BS motifs of both target TFs and less conservative ones of partner TFs.

The criterion for the partial area under the PR curve on the *Y* axis subtracts from each value of the Precision measure its expected value $\mathrm{PREC}_{\mathrm{EXP}}$ (Fig. 4) (Saito, Rehmsmeier, 2015). For a model that is equally likely to recognize sequences from the positive and negative set, the PR curve is a horizontal line:

$$\mathrm{PREC}_{\mathrm{EXP}} = \frac{\mathrm{NF}}{\mathrm{NF}+\mathrm{NB}} = 0.5. \quad (4)$$

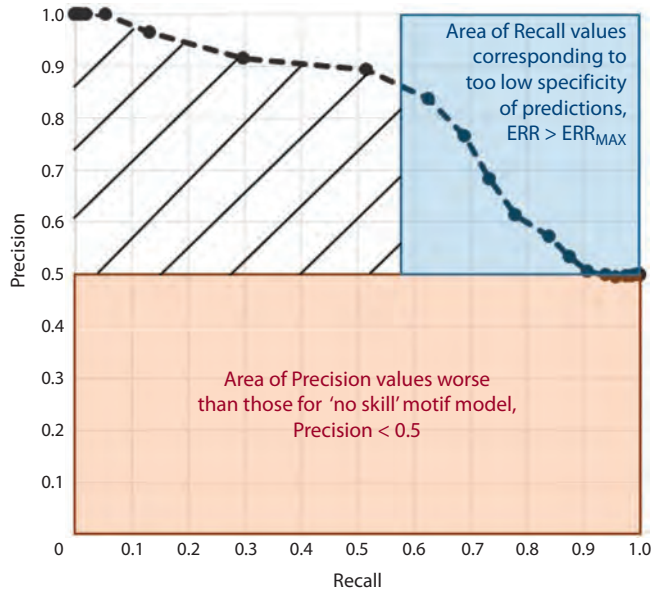


Fig. 4. Scheme of calculation of the partial area under the PR curve.

The *X* axis is the Recall measure (the probability of predicting the positive set sequence, Recall = TPR = TP/NF), formula (1). The *Y* axis is the Precision measure, the ratio of the probability of predicting the positive set sequence to the sum of the probabilities of predicting the positive and negative set sequences, Precision = TPR / (TPR + FPR), formula (3). The pink area marks Precision < 0.5 values corresponding to predictions worse than those of a "no skill" model equally likely to predict sequences in the positive and negative sets. The criteria Precision > 0.5/Precision < 0.5 mark areas of selection towards the positive/negative sets. The blue area shows the area of predicted sequences of the positive set with very low specificity. They correspond to the expected frequency of the motif greater than the threshold, $\mathrm{ERR} > \mathrm{ERR}_{\mathrm{MAX}}$. The normal distribution with the mean and standard deviation (μ_N, σ_N) = (5, 2.5) was taken to generate the data of the negative set example, and the positive set was a mixture of 50 %/50 % normal distributions (μ_{P1}, σ_{P1}) = (10, 1) and (μ_{P2}, σ_{P2}) = (5.5, 4). These distributions model sites passing and failing to pass the threshold $\mathrm{ERR}_{\mathrm{MAX}}$ of the expected motif frequency. The shading denotes the area determining the metric pAUPRC as the partial area under the curve.

This ratio is constant and equal to 0.5 because the FP value was normalized above, so the set sizes in this formula can already be considered equal. Hence, the partial area under the PR curve in the MetArea SP is calculated as the following sum:

$$\text{pAUPRC} = \frac{2}{\text{NF}} \times \sum_{i=1}^{\text{NI}} \left[\left\{ \frac{\text{PREC}(i) + \text{PREC}(i-1)}{2} - \text{PREC}_{\text{EXP}} \right\} \times \right.$$
$$\left. \times \left\{ \text{REC}(i) - \text{REC}(i-1) \right\} \right]. \quad (5)$$

Here, NI is the mildest threshold, determined as described above from the expected frequencies and the input parameter ERR_{MAX}. The 2/NF factor is required to normalize the value of pAUPRC to the maximum value of 1. The maximum value of the first multiplier under the sum, $\{(\text{PREC}(i) + \text{PREC}(i-1))/2 - \text{PREC}_{\text{EXP}}\}$, is 0.5 since the maximum Precision value is 1; and the maximum value of the sums of the second multipliers, $\{\text{REC}(i) - \text{REC}(i-1)\}$, is NF, the size of the positive set.

The criterion for predicting the functional relation of motifs reflects the increase in the accuracy estimate of the joint motif compared to the accuracy estimates of single motifs. This criterion quantitatively assesses mutually exclusive occurrence in pairs of motifs. For a pair of motifs A and B, the criterion requires a higher value of the accuracy estimate pAUPRC(A&B) of the joint motif A&B compared to the values of the accuracy estimates of both single motifs, pAUPRC(A) and pAUPRC(B). Calculated as follows, the Ratio of Areas Under Curves (RAUC) should exceed one:

$$\text{RAUC(A, B)} = \frac{\text{pAUPRC(A\&B)}}{\text{Max}\{\text{pAUPRC(A)}, \text{pAUPRC(B)}\}} > 1. \quad (6)$$

Application options of the MetArea SP

MetArea SP inputs can be TF BS motifs with expected enrichment in the positive *vs* negative set, e. g., such motifs are the results of a *de novo* motif search (Bailey, 2021). Separate applications of SP implement massive analyses of the collections of TF BS motifs from the Hocomoco and JASPAR databases. Analysis of multiple pairs of motifs allows identification of pairs that reveal a larger increase in pAUPRC recognition accuracy estimates when motifs are combined. The MetArea SP allows several application options, implemented as separate programs. The following application options consider the PWM motif model:

- two given motifs;
- several given motifs, for K motifs all possible $\{K \times (K-1)/2\}$ pairs are checked;
- a given motif *vs* all M motifs of BS of known TFs from the database. For a given motif, all its M pairs with the motifs from the Hocomoco (human, mouse) or JASPAR (plants, insects) collections are checked;
- all BS motifs of known TFs from the database are checked. From all M motifs of known TFs from the Hocomoco or JASPAR collection, K motifs with the highest pAUPRC accuracy scores are selected and all $\{K \times (K-1)/2\}$ possible pairs of these motifs are tested.

The application options for the PWM and SiteGA motif models:

- motif PWM and motif SiteGA.

Next, we provide examples of the results of ChIP-seq data analysis for different application options of the MetArea SP.

Analysis of several given motifs of the PWM model

Consider the ChIP-seq dataset for the BHLHA15 TF (Hess et al., 2016) (GTRD PEAKS039234, GEO GSE86289) for mouse pancreas. Application of a *de novo* search (STREME tool) (Bailey, 2021) showed that among the five motifs with the highest enrichment, four had significant similarity ($p < 0.001$) (Gupta et al., 2007) to known BHLHA15 TF BS motifs from the Hocomoco. The motifs #1/#5 and #2/#4 are similar to BHA15.H12CORE.0.P.B and BHA15.H12CORE.1.SM.B, respectively (Fig. 5*a*). These motifs correspond to the consensus E-box CAnnTG with spacers GC and AT, so they are labelled BHLHA15_GC_1/BHLHA15_GC_2, and BHLHA15_AT_1/BHLHA15_AT_2, respectively. Motif #3 has significant similarity ($p < 0.001$) to the BS motif of the CTCF TF (CTCF.H12CORE.0.P.B) (Fig. 5*a*).

Analysis of the values of the pAUPRC recognition accuracy estimates for single motifs and their pairwise joint motifs (Fig. 5*b*) is based on the corresponding RAUC values for pairs of motifs (Fig. 5*c*), the similarity assessment of pairs of motifs is required to control for significantly similar motifs (Fig. 5*d*). High RAUCs are found for the pairs of motifs BHLHA15_GC_1/BHLHA15_TA2 and BHLHA15_GC_1/BHLHA15_TA1, the PR curves for them are shown in Figure 5*e*, *f*. The CTCF motif has high RAUCs with BHLHA15_GC1 and BHLHA15_TA2 motifs (Fig. 5*c*). The pair of BHLHA15_TA2 and CTCF motifs was found to have the maximum RAUC of 1.48 (Fig. 5*c*). Overall, our results are consistent with the ability of the TF BHLHA15 to bind to DNA only as part of the dimer of two bHLH TFs (Amoutzias et al., 2008). The trend towards divergence of BSs of various structure of the BHLHA15 TF into different peaks could mean that (1) the dimer may comprise different TFs from the bHLH class (including BHLHA15 TF), and (2) the binding of the dimer is influenced by other partner TFs, that form multiprotein complexes with the BHLHA15 TF. Hence, the DBD of the BHLHA15 TF adopts various conformations, e. g., TF CTCF, the BS motif of which is also enriched (Fig. 5*a*), may be a partner TF. According to experimental data: (1) several TFs from the bHLH class have protein-protein interactions with the CTCF TF (BIOGRID database, https://thebiogrid.org/); (2) analysis of partner TFs by genomic co-localization (Hu et al., 2020) confirms that several TFs from the bHLH class are co-localized with CTCF TFs at the same genomic loci *in vivo*.

Analysis of all BS motifs of known TFs from the database

Consider the ChIP-seq dataset for TF AR (Androgene Receptor) for the mouse prostate (Chen et al., 2013) (GTRD PEAKS035588, GEO GSM1145307). Figure 6 for this ChIP-seq dataset shows the matrix of the pairwise RAUC values for the 15 most enriched TF BS motifs according to the pAUPRC measure out of all 1,142 mouse TF BS motifs from the Hocomoco database. Among these 15 motifs, seven motifs belong to the TF AR BS and its homologues from the same subfamily GR-like (NR3C) {2.1.1.1.1} of the Steroid hormone receptors {2.1.1} family of the Nuclear receptors with C4 zinc fingers {2.1} class. This family defines the target TF AR, and the likely motifs of its BS. The other eight motifs out of 15 belong to BS of TFs from the subfamilies FOXA {3.3.1.1}, FOXJ {3.3.1.10}, FOXM {3.3.1.13} and FOXP {3.3.1.16}.

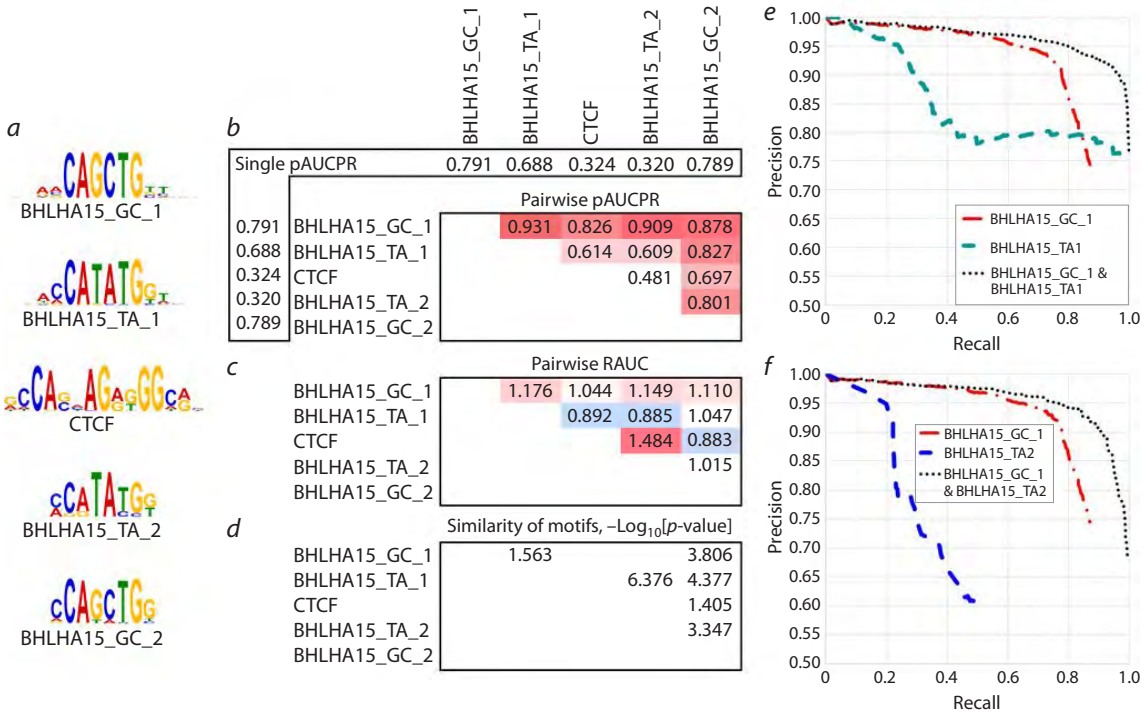


Fig. 5. Analysis of the five most enriched motifs from the *de novo* motif search results (STREME) (Bailey, 2021) for the ChIP-seq dataset for mouse BHLHA15 TF (Hess et al., 2016) (GTRD PEAKS039234, GEO GSM2299654/GSM2299655).

a – sequence logos for five motifs, sorted by enrichment significance obtained from the STREME tool; BHLHA15 TF BS motifs are labelled according to the dinucleotide in their spacer in the CAnnTG consensus; *b* – table of pairwise values of pAUPRC accuracy estimates of the joint motifs constructed from pairwise combinations of motifs, headers indicate pAUPRC values for single motifs, shades of red mark the maximum pAUPRC values of the joint motifs; *c* – table of pAUPRC values in pairs of motifs, shades of red and blue mark values greater than and less than one; *d* – table of significances of motifs similarity, $-\mathrm{Log}_{10}[p\text{-value}]$; *e*, *f* – PR curves for single motifs and their pairwise joint motifs BHLHA15_GC_1/BHLHA15_TA1 and BHLHA15_GC_1/BHLHA15_TA2.

They comprise the same FOX family {3.3.1} from the class Fork head/winged helix factors {3.3}. TFs of this family are putative partner TFs for AR TFs, e. g. Foxa1 TF is known for the same prostate tissue (Yang, Yu, 2015).

The pAUPRC values are greater than 1 for almost all pairs of GR-like/FOX motifs. For example, the RAUC value of 1.03 for the ANDR.H12CORE.0.P.B (pAUPRC rank 1) and FOXA2.H12CORE.0.PSM.A (rank 5) pair corresponds to the maximum value pAUPRC = 0.853 among all pairs of GR-like/FOX motifs. The pAUPRC values for pairs of GR-like/GR-like motifs exceed the value of 1 only for some pairs of motifs. The ANDR.H12CORE.2.P.B motif (rank 7) has a distinct consensus among all other GR-like motifs (AAACA instead of GNACA, see the Logo column, Figure 6); it has high RAUC values, and this is the only motif with RAUC values above 1 in all pairs with other GR-like and FOX motifs. In particular, among pairs of GR-like/GR-like motifs, the maximum pAUPRC value of 0.876 with a RAUC of 1.06 is achieved for the pair of motifs ANDR.H12CORE.0.P.B (rank 1) and ANDR.H12CORE.2.P.B (rank 7). Also, high RAUC values in pairs of GR-like/GR-like motifs were found for the MCR.H12CORE.1.SM.B motif, but it has the lowest pAUPRC rank of 15. This motif is a monomer-binding motif, not a dimer. Among the FOX/FOX motif pairs, there are almost no RAUC values greater than 1.

Overall, the high RAUC values of many pairs of GR-like/GR-like motifs suggest that the AR TF binds in different peaks using distinct structural types of GR-like motifs. A similar assumption can be made on the binding of a TF dimer consisting of AR and a TF from the FOX family according to the high RAUC values for pairs of GR-like/FOX motifs. The results obtained for ChIP-seq data for the AR TF imply the following. (1) Binding of AR TF to DNA occurs in the AR/AR and AR/Foxa1 dimers (if it is the Foxa1 TF that binds to FOX motifs under experimental conditions), and (2) both TFs allow a large variety of different structural types of BSs, so various pairs of motifs diverge in different peaks.

Analysis of the pair of motifs of the PWM and SiteGA models

Consider the ChIP-seq dataset for the E2F4 TF for primary innate immunity dendritic cells derived from mouse bone marrow stimulated with the pathogenic component lipopolysaccharide for 120 minutes (Garber et al., 2012) (GTRD PEAKS035857, GEO GSM881061). Figure 7 shows the PR curves for the PWM, SiteGA, and their joint PWM & SiteGA motifs calculated by the MetArea SP. The pAUPRC values for the PWM, SiteGA, and the joint PWM & SiteGA motifs are 0.457, 0.358, and 0.47, respectively; the pAUPRC value of the joint motif is 1.028.

The PWM and SiteGA motif models are based on very different methodological principles (Levitsky et al., 2007). The PWM model represents high-affinity sites defined by the most conserved positions and the most frequent nucleotides in them.

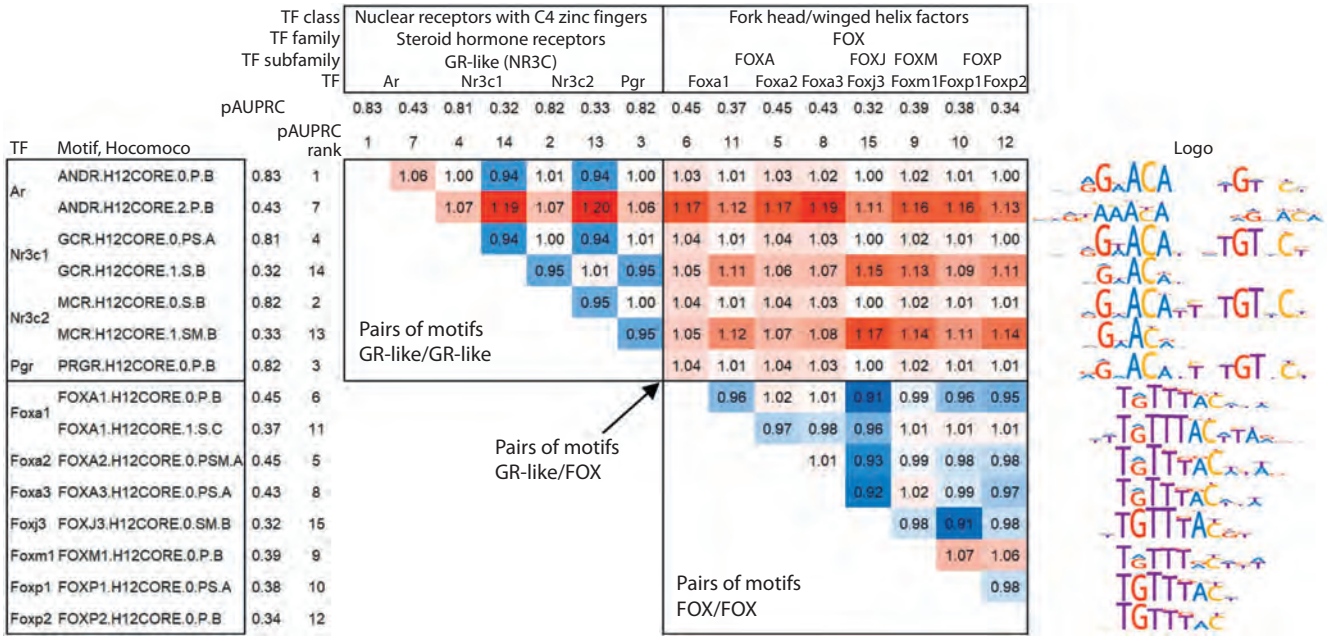


Fig. 6. Results of the analysis of BS motifs of known TFs from the Hocomoco database for the ChIP-seq dataset for AR TF in mouse prostate (Chen et al., 2013).

The 15 most enriched motifs according to the pAUPRC accuracy estimates are included in the analysis, headers of rows and columns show values and ranks of the pAUPRC metrics and the names of TFs from the Hocomoco database. Row headers indicate motif identifiers from Hocomoco, and column headers indicate the names of the TF class, family, and subfamily. In the table, shades of red/blue indicate changes in RAUC up/down from the neutral value of 1. The rightmost column shows the sequence logos of the motifs from the Hocomoco database. Black rectangles mark GR-like and FOX motifs in row and column headers, and in the table, pairs of BS TF motifs GR-like/GR-like, GR-like/FOX and FOX/FOX.

The SiteGA model comprises sites containing dependencies of different positions that presumably originate from the common actions of at least two TFs in cooperative binding to DNA (Morgunova, Taipale, 2017; Levitsky et al., 2020). Predicted sites of the SiteGA model are markedly less conserved than those of the PWM model; the SiteGA model is able to predict low affinity sites better than the PWM model (Tsukanov et al., 2022). Combining the PWM and SiteGA models improves recognition of low-affinity sites, as reflected by the greater extent of the PR curve of the joint PWM & SiteGA motif on the *X* axis (Recall), compared to each of the single PWM and SiteGA motifs. Although the joint motif has smaller Precision values (Fig. 7, *Y* axis) than the PWM model, the wider range of Recall values (*X* axis) determines the increase in the pAUPRC measure of the joint motif. Single motifs up to the threshold of expected motif frequency $ERR_{MAX} = 0.002$ recognize 73.2 % (PWM) and 63.3 % (SiteGA) of peaks, the joint motif recognizes 79.9 %.

The hypothesis that the PWM and SiteGA models represent different structural types of the E2F4 TF BS is confirmed by the TomTom motif comparison tool (p-value < 0.05) (Gupta et al., 2007). To prove this, for the PWM model, we used its nucleotide frequency matrix, and for the SiteGA model, as previously (Tsukanov et al., 2022), the nucleotide frequency matrix constructed from the predicted sites. The ability of the E2F4 TF to bind to different structural types of BSs is also indicated by the experiment of M. Garber et al. (2012), where the genomic binding loci of 25 TFs were determined under the same conditions. The loci of E2F4 TFs were shown to overlap

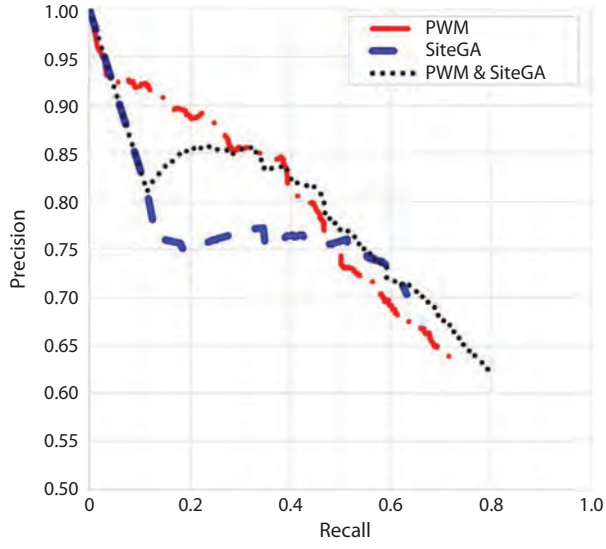


Fig. 7. Results of analysis of the motif pair of the PWM and SiteGA models by the MetArea SP.

Red, blue, and black represent PR curves for the PWM, SiteGA motifs, and the joint motif PWM & SiteGA. The ChIP-seq dataset for TF E2F4 (GTRD PEAKS035857, GEO GSM881061) was used in the analysis.

significantly with the loci of five TFs: EGR2, EGR1, IRF2, ETS2 and E2F1. Consequently, it can be assumed that the TF E2F4 is part of the same multiprotein complexes with these TFs. Therefore, in different TF loci, E2F4 has to change its BSs to a greater or lesser extent to adapt to the BSs of partner TFs.

Discussion

In our study, we propose the novel MetArea approach for detecting mutually exclusive occurrence in pairs of TF BS motifs based on analyses of single ChIP-seq datasets. If two motifs are structurally distinct BS motifs of the same TF in various peaks, then the mutually exclusive occurrence is due to the preferences of this TF to bind to either one or the other structural type of BS in the peaks, but it is less common to observe two BSs of different structures in the same peaks. If the BS motifs belong to two different TFs, mutually exclusive occurrence can result from the participation of both TFs in the same multiprotein complexes, but in different peaks one or another TF binds to DNA directly, but it is less common to observe BSs of both TFs in the same peak.

During the development of the MetArea SP, we abandoned the use of the metric of the partial area under the ROC curve (pAUC ROC) (Levitsky, Tsukanov, 2024) and used the metric of the area under the PR curve (Davis, Goadrich, 2006) to determine the metric of the partial area under the PR curve. It had been previously proposed (Davis, Goadrich, 2006) that the application of the area under the AUC ROC curve cannot be correct if the actual recognition thresholds of a binary classifier should be quite stringent. Therefore, we should take note if the advantage of one motif relative to another is recruited in the interval of mild recognition thresholds (at the right tail of the ROC curve). To correctly compare two motifs in this case, instead of the metric of the area under the AUC ROC curve, we previously used the metric "Partial Area Under the ROC Curve, pAUC". Instead of the full-size range of the False Positive Rate (FPR, the fraction of recognized objects from the negative set, *X* axis of the ROC curve) from 0 to 1, this metric uses only a certain left part of it, discarding the range of too large *FPR* values. We implemented this approach to compare the recognition accuracy of TF BS motifs of the PWM, BaMM and SiteGA models (Tsukanov et al., 2022). There, we used the criterion on the Expected Recognition Rate, ERR < 0.001, to restrict the recognition thresholds of motifs in order to compute the pAUC ROC accuracy estimates.

Unfortunately, this approach is not suitable to compute the accuracy of the joint motif required in the implementation of the MetArea approach. The rationale for this is the necessity to count the frequency of the joint motif, i. e. the number of its hits. It is possible for non-overlapping single motifs, and in the case of their overlapping, the frequency of the joint motif should be reduced in some way. An alternative way to get rid of the overestimation of accuracy given by the AUC ROC measure is to switch from the ROC curve to the PR curve and calculate the area under the PR curve (Davis, Goadrich, 2006; Keilwagen, et al., 2019).

Several approaches have been previously proposed to identify the occurrence of different TF BS motifs or different sets of motifs in various peak fractions of a single set of ChIP-seq peaks. The DIVERSITY tool (Mitra et al., 2018) partitions a set of ChIP-seq peaks into several non-overlapping groups, so that each group is represented by its enriched motif from *de novo* search results. Later, the authors allowed that each group of peaks is not represented by a single motif, but by a combination of several motifs. The cisDIVERSITY tool (Biswas, Narlikar, 2021) for the set of peaks performs a *de novo* search for enriched motifs using the PWM model, and then distributes the found motifs into several non-overlapping groups of peaks so that all groups make up the entire set of peaks. Each of the motifs has different frequencies across groups, e. g., some groups have higher frequencies than other groups, while other groups may not have a motif. The tasks of the DIVERSITY/cisDIVERSITY and MetArea tools are similar in that different motifs are separated into certain fractions of peaks. However, the DIVERSITY/cisDIVERSITY tools: (1) identify the entire variety of motifs and divide all peaks into groups in order to find distinct motifs or combinations of them for different groups; (2) consider only the traditional PWM motif model. The MetArea SP (1) considers only pairs of motifs, to find pairs of motifs that better complement each other by maximizing the accuracy measure pAUPRC for the joint motif; (2) considers both the traditional PWM model and alternative models of the TF BS motif.

Conclusion

We have developed the MetArea SP. It uses a single set of ChIP-seq peaks to calculate the "Partial Area Under the PR Curve" (pAUPRC) accuracy measure for the two input single TF BS motifs, determines the joint motif from them, and also calculates the pAUPRC measure for it. Creating a joint motif from the two single motifs and calculating a pAUPRC accuracy estimate for it allows comparing two single motifs and their overall effect on a uniform scale. The excess of accuracy estimates of the joint motif over those of both single motifs indicates their mutually exclusive occurrence. The results of the MetArea analysis allow predicting the functional relationship of the two motifs, and hence their corresponding TFs. In particular, the MetArea SP can offer substantial arguments for or against the hypothesis that the two motifs are structural types of the BS of a single TF. Similarly, support or rejection are proposed for the hypothesis that the BS motifs represent two TFs together involved in the regulation of gene transcription as part of a single multiprotein complex. In summary, the MetArea SP predicts for a given ChIP-seq dataset (1) structural diversity of BSs of a single TF and (2) pairs of BS motifs of different TFs acting to regulate gene transcription as part of single multiprotein complexes of many TFs.

References

Ambrosini G., Vorontsov I., Penzar D., Groux R., Forne O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Buche P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol*. 2020;21:114. 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;37:2834-2840. doi 10.1093/bioinformatics/btab203

Biswas A., Narlikar L. A universal framework for detecting cis-regulatory diversity in DNA regions. *Genome Res*. 2021;31(9):1646-1662. doi 10.1101/gr.274563.120

Chen Y., Chi P., Rockowitz S., Iaquinta P.J., Shamu T., Shukla S., Gao D., Sirota I., Carver B.S., Wongvipat J., Scher H.I., Zheng D., Sawyers C.L. ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nat. Med*. 2013;19(8):1023-1029. doi 10.1038/nm.3216

Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. New York: Assoc. for Computing Machinery, 2006;233-240. doi 10.1145/1143844.1143874

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

Garber M., Yosef N., Goren A., Raychowdhury R., Thielke A., Guttman M., Robinson J., Minie B., Chevrier N., Itzhaki Z., Blecher-Gonen R., Bornstein C., Amann-Zalcenstein D., Weiner A., Friedrich D., Meldrim J., Ram O., Cheng C., Gnirke A., Fisher S., Friedman N., Wong B., Bernstein B.E., Nusbaum C., Hacohen N., Regev A., Amit I. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell*. 2012;47(5):810-822. doi 10.1016/j.molcel.2012.07.030

Georgakopoulos-Soares I., Deng C., Agarwal V., Chan C.S.Y., Zhao J., Inoue F., Ahituv N. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nat. Commun*. 2023;14:2333. doi 10.1038/s41467-023-37960-5

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

Hess D.A., Strelau K.M., Karki A., Jiang M., Azevedo-Pouly A.C., Lee A.H., Deering T.G., Hoang C.Q., MacDonald R.J., Konieczny S.F. MIST1 links secretion and stress as both target and regulator of the unfolded protein response. *Mol. Cell. Biol*. 2016;36(23): 2931-2944. doi 10.1128/MCB.00366-16

Hu G., Dong X., Gong S., Song Y., Hutchins A.P., Yao H. Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions. *Nucleic Acids Res*. 2020;48(17):9606-9620. doi 10.1093/nar/gkaa705

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

Keilwagen J., Posch S., Grau J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol*. 2019;20(1):9. doi 10.1186/s13059-018-1614-y

Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res*. 1995;23(20):4097-4103. doi 10.1093/nar/23.20.4097

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

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinformatics*. 2007;8(1):481. doi 10.1186/1471-2105-8-481

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:e139. doi 10.1093/nar/gkz800

Levitsky V., Oshchepkov D., Zemlyanskaya E., Merkulova T. Asymmetric conservation within pairs of co-occurred motifs mediates weak direct binding of transcription factors in ChIP-Seq data. *Int. J. Mol. Sci*. 2020;21(17):E6023. doi 10.3390/ijms21176023

Levitsky V.G., Tsukanov A.V. MetArea tool for predicting structural variability and cooperative binding of transcription factors in ChIP-seq data. In: 14th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2024). 2024;136-138. doi 10.18699/bgrs2024-1.2-17

Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol*. 2018;14(4):e1006090. doi 10.1371/journal.pcbi.1006090

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

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 Genom. 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., Lucas J., Chèneby J., Baranasic D., Khan A., Fornes O., Gundersen S., Johansen M., Hovig E., 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

Rogers J.M., Waters C.T., Seegar T.C.M., Jarrett S.M., Hallworth A.N., Blacklow S.C., Bulyk M.L. Bispecific forkhead transcription factor FoxN3 recognizes two distinct motifs with different DNA shapes. *Mol. Cell*. 2019;74(2):245-253.e6. doi 10.1016/j.molcel.2019.01.019

Saito T., Rehmsmeier M. The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi 10.1371/journal.pone.0118432

Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res*. 2016;44:6055-6069. doi 10.1093/nar/gkw521

Tognon M., Giugno R., Pinello L. A survey on algorithms to characterize transcription factor binding sites. *Brief. Bioinform*. 2023;24(3): bbad156. doi 10.1093/bib/bbad156

Tsukanov A.V., Levitsky V.G., Merkulova T.I. Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. *Vavilov J. Genet. Breed*. 2021;25(1):7-17. doi 10.18699/VJ21.002

Tsukanov A.V., Mironova V.V., Levitsky V.G. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front. Plant Sci*. 2022;13:938545. doi 10.3389/fpls.2022.938545

Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A., Kamenets V., Kasianova A., Kolmykov S., Yevshin I.S., Favorov A., 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., Lambert S.A., Mann I., Cook K., Zheng H., Goity A., van Bakel H., Lozano J.C., Galli M., Lew-

sey M.G., Huang E., Mukherjee T., Chen X., Reece-Hoyes J.S., Govindarajan S., Shaulsky G., Walhout A.J.M., 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. 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

Yang Y.A., Yu J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes Dis*. 2015;2(2): 144-151. doi 10.1016/j.gendis.2015.01.003

Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code. *Curr. Opin. Syst. Biol*. 2020;23:22-31. doi 10.1016/j.coisb.2020.08.002

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9: R137. doi 10.1186/gb-2008-9-9-r137

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

Received October 19, 2024. Revised November 20, 2024. Accepted November 21, 2024.

GENOMICS AND TRANSCRIPTOMICS
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):834-842

doi 10.18699/vjgb-24-91

Computer analysis shows differences between mitochondrial miRNAs and other miRNAs

P.S. Vorozheykin[1], I.I. Titov[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] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

pavel.vorozheykin@gmail.com

Abstract. A subclass of miRNAs with as yet unknown specific functions is mitomiRs – mitochondrial miRNAs that are mainly derived from nuclear DNA and are imported into mitochondria; moreover, changes in the expression levels of mitomiRs are associated with some diseases. To identify the most pronounced characteristics of mitochondrial miRNAs that distinguish them from other miRNAs, we classified mitomiR sequences using the Random Forest algorithm. The analysis revealed, for the first time, a significant difference between mitomiRs and other microRNAs by the following criteria (in descending order of importance in the classification): mitomiRs are evolutionarily older (have a lower phylostratigraphic age index, PAI); have more targets and disease associations, including mitochondrial ones (two-sided Fisher's exact test, average *p*-values 1.82×10^{-89}/1.13×10^{-96} for all mRNA/diseases and 6.01×10^{-22}/1.09×10^{-9} for mitochondria, respectively); and are in the class of "circulating" miRNAs (average *p*-value 1.20×10^{-56}). The identified differences between mitomiRs and other miRNAs may help uncover the mode of miRNA delivery into mitochondria, indicate the evolutionary conservation and importance of mitomiRs in the regulation of mitochondrial function and metabolism, and generally show that mitomiRs are not randomly encountered miRNAs. Information on 1,312 experimentally validated mitomiR sequences for three organisms (*Homo sapiens*, *Mus musculus* and *Rattus norvegicus*) is collected in the mitomiRdb database (https://mitomiRdb.org).
Key words: mitomiR; mitochondria; miRNA; evolution; database.

For citation: Vorozheykin P.S., Titov I.I. Computer analysis shows differences between mitochondrial miRNAs and other miRNAs. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):834-842. doi 10.18699/vjgb-24-91

Funding. Supported by Budget Project No. FWNR-2022-0020.

Компьютерный анализ показывает отличия митохондриальных микроРНК от остальных микроРНК

П.С. Ворожейкин[1], И.И. Титов[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] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

pavel.vorozheykin@gmail.com

Аннотация. Одним из подклассов микроРНК с до сих пор неизвестными специальными функциями являются митомиРы (mitomiRs) – митохондриальные микроРНК, которые в основном происходят из ядерной ДНК и импортируются в митохондрии, при этом изменение уровня их экспрессии ассоциировано с рядом заболеваний. Для выявления характерных особенностей митохондриальных микроРНК, отличающих их от остальных микроРНК, мы провели классификацию этих последовательностей с помощью метода случайного леса. Проведенный анализ впервые выявил достоверные различия между митомиРами и микроРНК по следующим характеристикам (по убыванию степени их важности в классификации): митомиРы имеют достоверно больший эволюционный возраст (низкий индекс филостратиграфического возраста, PAI), большее количество мишеней и ассоциаций с болезнями, в том числе митохондриальными (двусторонний точный тест Фишера, средние *p*-значения 1.82×10^{-89}/1.13×10^{-96} для всех мРНК/болезней и 6.01×10^{-22}/1.09×10^{-9} для митохондриальных); принадлежат к классу «циркулирующих» (среднее *p*-значение 1.20×10^{-56}). Обнаруженные различия между митомиРами и остальными микроРНК могут помочь раскрыть способ доставки микроРНК в митохондрии, свидетельствуют об эволюционной консервативности и важности митомиРов в регулировании функций и метаболизма митохондрий, а в целом говорят о том, что митомиРы не являются случайными микроРНК. Информация о 1312 экспериментально подтвержденных последовательностях митомиРов для трех организмов (*Homo sapiens*, *Mus musculus* и *Rattus norvegicus*) собрана в базе mitomiRdb (https://mitomiRdb.org).
Ключевые слова: митомиР; митохондрия; микроРНК; эволюция; база данных.

Introduction

Mitochondria engage in extensive bidirectional inter-compartmental crosstalk to regulate their proteome, overall cellular fitness and organismal health. To date, it is well known that the fundamental pathways of the miRNA biogenesis start in the nucleus and end in the cytoplasm (Bartel, 2018; Salim et al., 2022; Ziętara et al., 2023). However, there is evidence that these short non-coding RNA sequences are also present in organelles, in particular, in mitochondria (Lung et al., 2006; Kren et al., 2009). In many cases, mitochondrial microRNAs (the so-called mitomiRs) are more abundant in the mitochondria than in the cytoplasm. These observations suggest a nucleus miRNA translocation into mitochondria and/or the existence of a complete miRNA maturation process within mitochondria.

The existence of a transport mechanism is supported by the detection of the so-called circulating miRNAs (Pozniak et al., 2022). There are also arguments in favor of the second option: first, the miRNA machinery proteins AGO2 and Dicer, which are involved in the canonical pathway of microRNA biogenesis, have been found in mitochondria (Bandiera et al., 2011; Wang W.-X. et al., 2015); second, mitochondrial gene expression can be regulated by mitochondrial miRNAs and this regulation inevitably manifests itself in mitochondria-related diseases (Li et al., 2012; Tomasetti et al., 2014; Zhang et al., 2014; Lin, Chu, 2021; Erturk et al., 2022; Gohel, Singh, 2022). Since the composition of miRISC (miRNA-induced silencing complex) varies at different development stages, this suggests the possibility of a mitochondria-specific miRNA origin and biogenesis, as well as potentially unknown functions of nuclear miRNAs within mitochondria. This highlights mitochondrial miRNAs as a new subclass of miRNAs with significant implications for scientific research. Nevertheless, the specific functions and biogenesis pathways of mitomiRs remain unexplored, and it is still unclear whether mitomiRs are merely typical microRNAs that happen to be observed in mitochondria by chance.

To reveal specific features of this new miRNA class, we analyzed all miRNA sequences using the Random Forest algorithm and determined the most important criteria for miRNA classification (listed in descending order of their importance): the phylostratigraphic age index (PAI) of the miRNA; the presence of miRNA targets, and whether the miRNA belongs to the "circulating" class of miRNAs. Based on the obtained data, we drew conclusions regarding the age of mitomiRs, their possible appearance in mitochondria, and their significance for the organism functioning.

The explored mitomiRs have been collected in the mitomiRdb database (https://mitomiRdb.org) – a manually curated repository of experimentally discovered mitochondrial miRNAs. This database stores information about mitomiRs for three mammals: *Homo sapiens, Mus musculus* and *Rattus norvegicus*. There are 1,312 annotated sequences with details such as identifiers, nucleotide sequences, and secondary structures of precursors. Additionally, the database provides references to publications with supporting experiments and evidences of experimentally validated miRNA-mRNA and miRNA-disease associations, including those related to mitochondria. All collected data are available online and can be freely downloaded for further computational analysis.

Materials and methods

Mature miRNA sequences were downloaded from the miRBase database (https://miRBase.org, releases 10–22.1) (Kozomara et al., 2019). The latest release of the database contains 48,885 annotated miRNA sequences from 285 species. The total number of *H. sapiens, M. musculus* and *R. norvegicus* miRNAs – 5,398 sequences, of which 2,274 were marked as "high confidence" by database curators (those miRNAs, the reads of which align with the canonical pre-miRNA processing patterns by Drosha/Dicer complexes).

To study the relationship of mitomiRs with mRNA, we used the miRTarBase database (https://mirtarbase.cuhk.edu.cn, release 8.0) (Huang et al., 2020) – a manually curated repository of experimentally validated microRNA-target interactions from scientific publications with experimental evidence of direct interactions. The total number of annotated entries of microRNA-mRNA interactions for human, mouse and rat miRNAs is equal to 553,118. Among these, 13,311 entries are noted as "supported by strong experimental evidence", while the remaining 539,807 entries are based on "weak" proof.

Data on experimentally validated microRNA-disease associations were obtained from the RNADisease database (http://www.rnadisease.org, release 4.0, "Experimental data" section, miRNA-disease information entries) (Chen et al., 2022). Each association was manually curated from publications, with particular attention being paid to experimental evidence of the miRNA role in regulation and pathogenesis of diseases as well as the analysis of miRNA-mRNA complementary binding and its involvement in disease progression. The total number of annotated entries for the three considered species (human, mouse and rat) amounts to 211,150.

The following mtDNA reference sequences were used to determine the localization of mitochondrial miRNAs and the mitochondrial genes: *H. sapiens* (NC_012920.1), *M. musculus* (NC_005089.1), and *R. norvegicus* (NC_001665.2) (Sayers et al., 2022). To explore the evolution of the let-7a-5p binding site, the following mitochondrial genomes of primates were used: *Gorilla gorilla* (NC_001645.1), *Pan paniscus* (NC_001644.1), *Pongo pygmaues* (NC_001646.1), *Pan troglodytes* (NC_001643.1), and *Symphalangus syndactylus* (NC_014047.1) (Sayers et al., 2022).

To calculate the phylogenetic age index (PAI) of miRNAs, we took the taxonomic lineages from the NCBI server (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump, data as of July 12, 2022) (Sayers et al., 2022). For each miRNA sequence from 285 organisms, all its homologous sequences were identified to determine the distribution of similar microRNAs across the species. Two nucleotide sequences were considered homologous (phylogenetically related) if the Hamming distance of their globally aligned sequences was less than 10 % of the alignment length. Alignment parameters: a match score of 5.0, a mismatch penalty of –4.0, an initial insertion/deletion penalty of –10.0, and an extending insertion/deletion penalty of –0.5. To calculate the PAI of the miRNA sequence, we used a set of organisms in which miRNA homologues appeared.

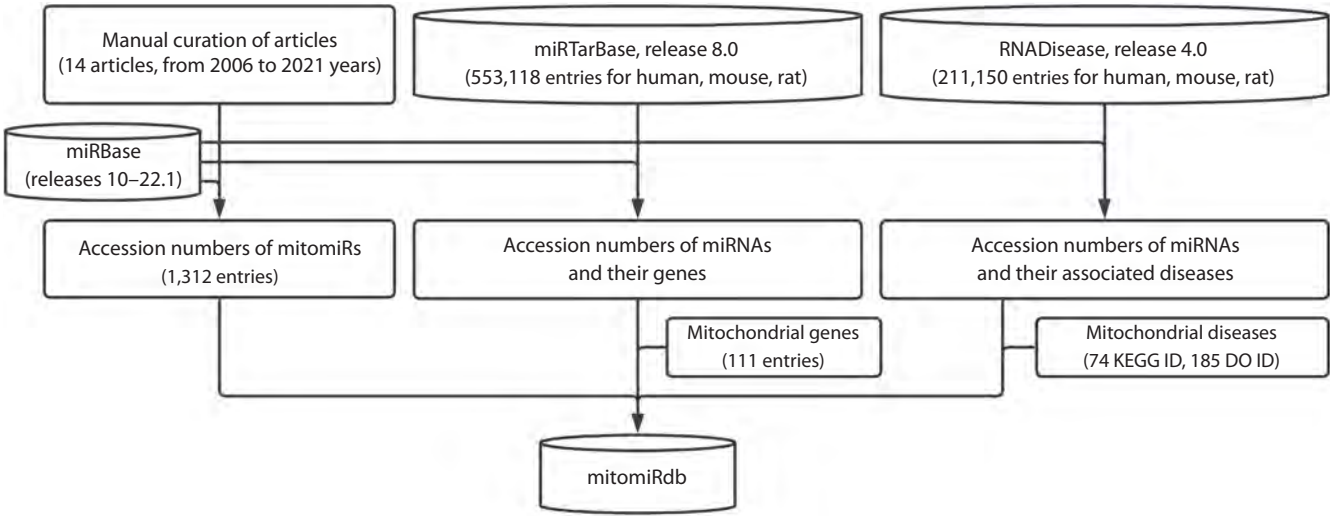


Fig. 1. A schematic workflow for collecting information about mitomiRs to form the mitomiRdb database.

According to taxonomic lineages, the PAI value represents the serial number of the most common taxon of this set (numbering from zero) (Mustafin et al., 2019).

DO- and KEGG-identifiers and names of diseases were obtained from the Disease Ontology Project (Schriml et al., 2022) and the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2017). With that information, we have compiled a list of mitochondria-associated diseases. For this purpose, we took names and identifiers of diseases that are included in the supergroup (and its subgroups) KEGG H01427 (Mitochondrial diseases) and those explicitly mentioning mitochondria in their names. Furthermore, a list of mitochondria-associated diseases with the total number of 74 KEGG and 185 DO ID entries was made (data provided in Supplementary Material 1)[1].

Information on "circulating" miRNAs was retrieved from the miRandola database (release on February 2017, 606 miRNAs) and the plasmiR database (release on June 17, 2021, 251 miRNAs) (Russo et al., 2018; Tastsoglou et al., 2021). These extracellular miRNAs are detected in traceable quantities in blood and other body fluids. The total number of circulating miRNAs from two databases is 628 (590 human, 18 mouse, and 20 rat miRNAs).

Figure 1 presents a schematic workflow outlining the process of gathering information about mitomiRs and establishing connections between mitomiRs and their targets or associated diseases. First, we selected the papers, which explore mitochondria-located miRNAs or contain references to the term "mitomiR". Out of them, we took 14 articles (published between 2006 and 2021) that reported the experimentally verified presence of miRNA sequences within mitochondria isolated from three mammal species (*H. sapiens, M. musculus, R. norvegicus*) for different cell types and tissues (Lung et al., 2006; Kren et al., 2009; Bian et al., 2010; Bandiera et al., 2011; Barrey et al., 2011; Mercer et al., 2011; Das et al., 2012; Sripada et al., 2012; Dasgupta et al., 2015; Jagannathan et al., 2015; Wang W.-X. et al., 2015; Wang X. et al., 2017; Fan et al., 2019; Zheng et al., 2021). In these studies, miRNAs are mentioned by their names (e. g., hsa-miR-1), which may have changed over time in the miRBase database. To ensure consistency, based on the miRBase annotation history, we matched each miRNA name with its corresponding unique accession number (MIMAT number) from the miRBase. The accession number allows the unambiguous identification of a mitomiR sequence across database releases.

During the matching process, we found that some previously annotated mitomiRs had been excluded from the recent miRBase releases. To ensure comprehensive coverage, we extended our dataset of mitomiRs to include 40 additional miRNAs and their 41 precursors that had been previously annotated in the miRBase database (Supplementary Material 2). As a result, we compiled accession numbers, sequences, secondary structures of precursors, and other additional information for 1,312 mitomiRs that were enriched in mitochondria.

To characterize mitochondria-associated miRNAs, in addition to the mitomiRs sample, we compiled a dataset of 4,126 sequences (referred to as non-mitomiRs), which contains human, mouse and rat miRNAs from the miRBase database (release 22.1) except all identified mitomiRs.

Then, using the naming history of the miRBase sequences, for each miRTarBase and RNADisease entry, its miRNA identifier (ID) was matched to the unique miRBase accession number for further possibility of unambiguous association of microRNAs with targeted genes and related diseases. For some entries it was not possible to clearly identify an accession number of miRNA due to incomplete or inconsistent database information (e. g., hsa-miR-b5539 and hsa-miRPlus-C1100 are not identifiable miRBase IDs; the database entry contains pre-miRNA name hsa-let-7a-1, which cannot be unequivocally matched to a single miRNA in the miRNA-miRNA duplex).

[1] Supplementary Materials 1–4 are available at:
https://vavilovj-icg.ru/download/pict-2024-28/appx27.xlsx

By removing these ambiguous entries, we established reliable references between identified mitomiRs and their respective targets and diseases.

Finally, having the set of mitomiR-target and mitomiR-disease associations, we compared how mitomiRs are related to known mitochondrial genes and diseases. We have included in the database the information on relationships between mitomiRs and 111 known mitochondrial genes (which encode for rRNAs, tRNAs, and protein subunits) for three (human, mouse, rat) examined mtDNAs. Each entry in the RNADisease database provides the name of disease and one or several disease identifiers: Disease Ontology (DO) ID (Schriml et al., 2022), MeSH ID (Sayers et al., 2022), and KEGG ID (Kanehisa et al., 2017; Schriml et al., 2022). Therefore, for each mitomiR sequence, we additionally indicated its connection (or a lack of connection) to the prepared list of mitochondrial diseases (Supplementary Material 3) as well as to all diseases.

To identify and rank the most powerful characteristics of mitomiRs in comparison to other miRNAs (non-mitomiRs), we analyzed miRNA sequences using the Random Forest algorithm (Breiman, 2001). Four binary and one numerical criteria were chosen for classification. Binary criteria: (1) whether the miRNA sequence is "circulating"; (2) whether the miRNA is "confident" according to the miRBase declaration; (3) whether the miRNA has a validated target; and (4) whether the miRNA is associated with a disease. The numerical criterion was the PAI value of the miRNA sequence.

The Random Forest algorithm was carried out 100 times on specific datasets: each dataset consists of all mitomiRs and non-mitomiRs except all homologous sequences but one (randomly selected) member from each homologue group. In each iteration, a randomly generated subset (one third of the total dataset) serves as a test dataset, while the remaining part is used for model training. Statistical estimates and significance levels for the criteria were averaged over all tests.

Results

Statistics. Considering the papers with the data on mitochondria-located miRNAs for three mammal species (*Homo sapiens, Mus musculus,* and *Rattus norvegicus*), we obtained information on 1,312 accession numbers of the mitomiR sequences. Among them, there were sequences that had been excluded from the miRBase database for various reasons. For example, miRNAs hsa-miR-1974, hsa-miR-1977 and hsa-miR-1978 overlap with mitochondrial tRNAs; hsa-miR-6723-5p has a reads pattern from RNA-seq experiments that does not support its annotation as a miRNA in the miRBase; mmu-miR-2145 is a fragment of 5S rRNA; and other entries are suspected of being transcriptional noise or products of non-canonical maturation process. Approximately 66.6 % (874) of the discovered mitomiRs correspond to human miRNAs, while the rest belong to mouse (30.6 %, 401) and rat (2.8 %, 37).

By comparing mitomiR sequences and names, we found 16 (out of possible 37, based on the number of mitomiRs in *R. norvegicus*) conserved mitomiRs, i. e. those detected in mitochondria across all of the three considered species. Additionally, 30.6 % of all mitomiRs have been described in more than one publication, which may represent their higher credibility as mitochondrial. Only nine of the mitomiR sequences (hsa-miR-1973, hsa-miR-1974, hsa-miR-1977, hsa-miR-1978, hsa-miR-4461, hsa-miR-4463, hsa-miR-4284, hsa-miR-4485-3p, mmu-miR-805) are fully mapped to mitochondrial DNA (three to tRNAs and rRNAs, two to protein-coding regions, and one to the D-loop). Notably, among them, only five mitomiRs (hsa-miR-1973, hsa-miR-1974, hsa-miR-1977, hsa-miR-1978, hsa-miR-4485-3p) have been additionally validated by RT-PCR/RT-qPCR/qRT-PCR analysis or observed in mitochondria in greater abundance than in the cytoplasm.

Criteria. The selected characteristics of mitomiRs (in comparison with the rest of miRNAs) do not allow classifying mitomiRs by one of the chosen criteria (Fig. 2*a*). However, the Random Forest classification algorithm ranks criteria by their influence, highlighting the most important characteristics of mitomiRs. With the considered criteria, the Random Forest model achieved average prediction errors (the fraction of incorrectly classified samples) of 0.20 ± 0.003 for the training dataset, and 0.22 ± 0.006, for the test dataset. The most influential criteria for the classification were: (1) the PAI value of the miRNA, (2) the presence of miRNA targets, and (3) whether the miRNA is classified as circulating (Fig. 2*b*). In contrast, the least important criteria were the miRNA's association with disease and its confidence level (as defined by miRBase); miRNA confidence plays the minimal role in the classification.

The evolutionary characteristic PAI (phylostratigraphic age index, describes the age of a mitomiR) appeared to be the most significant criterion for mitomiR classification. PAI denotes the serial number of a taxon (node of the phylostratigraphic tree) furthest from its root and occurring in taxonomic lineages of the microRNA sequence and its homologues. According to the PAI values, mitomiR sequences generally show, on average, greater evolutionary conservation than non-mitomiRs. The minimum value of mitomiRs' PAI is 4 (Fig. 3). Only four mitomiRs (hsa-miR-99a-5p, mmu-miR-99a-5p, hsa-miR-100-5p, and mmu-miR-100-5p) and two non-mitomiRs (rno-miR-99a-5p, rno-miR-100-5p) have this PAI due to their homologs being found in *Nematostella vectensis*, which testifies to an ancient history of these miRNAs' origin (Grimson et al., 2008).

However, in the evolutionarily distant species, miRNA sequence homology, even with the presence of a hairpin structure, does not guarantee the existence of a real miRNA (Grimson et al., 2008). All the aforementioned mitomiRs belong to the abundant miRBase family mir-10. This family also contains a set of human and mouse mitomiRs (miR-10a, miR-10b, miR-125b from the 5p-branch of precursors) with the PAI of 5. In contrast, the rat miRNAs from this family correspond to non-mitomiRs, which is probably due to the limited number of mitomiRs found in the rat. It is known that abundant miRNA families, such as mir-10, tend to be older, more efficient, target more genes, and are more likely to be associated with diseases. All of these factors point out

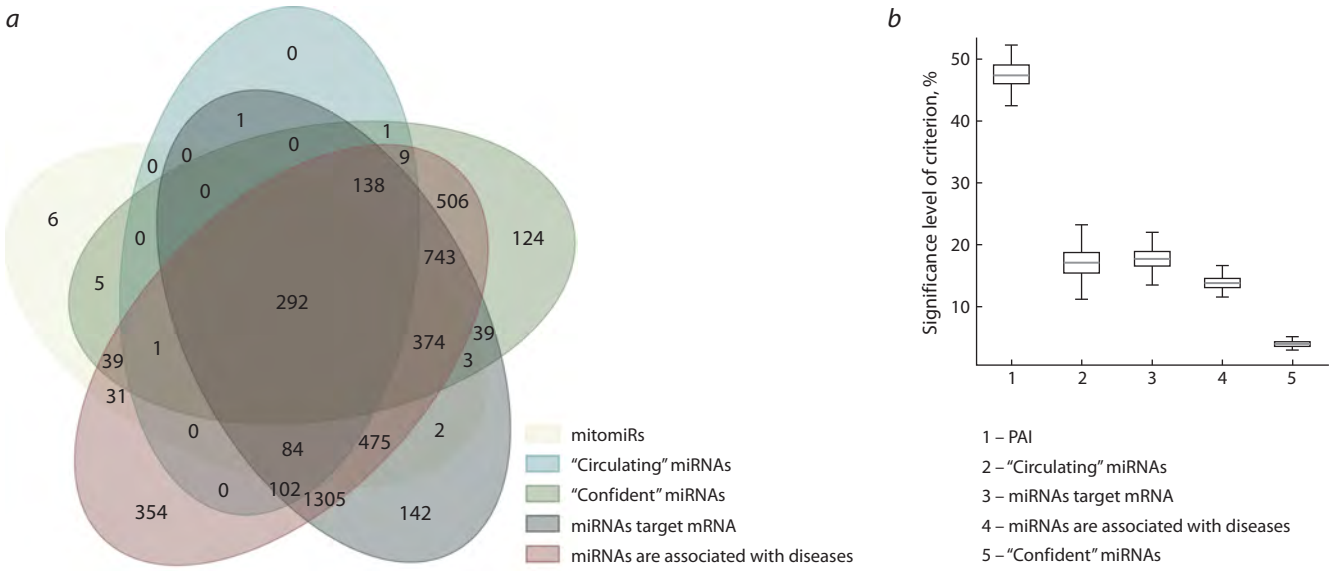


Fig. 2. Comparison of microRNA class criteria.

a, the Venn diagram illustrates the distribution of miRNA sequences across five classes: mitomiR sequences, "confident" miRNAs, "circulating" miRNAs and the miRNA sequences with known target and disease associations; *b*, the significance levels of criteria for mitomiR prediction, as determined by the Random Forest algorithm (averaged values over 100 tests). The horizontal line indicates the average significance level, the "whiskers" extend from the box to the farthest data point lying within 1.5 times the inter-quartile range from the box.

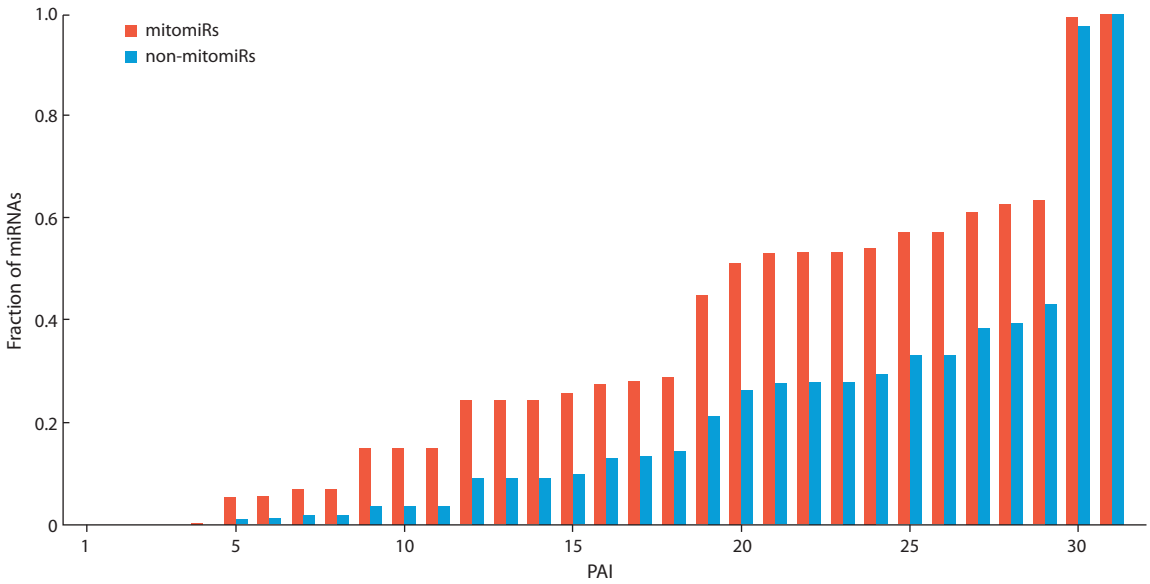


Fig. 3. Cumulative distribution of mitomiR and non-mitomiR sequences by PAI values. The fraction of mitomiRs with a PAI value less than 16 exceeds the corresponding fraction of non-mitomiRs, with the significance level of $1.10 \times 10^{-41} \pm 5.97 \times 10^{-41}$.

The minimum PAI value is 4, which corresponds to four mitomiRs (hsa-miR-99a-5p, mmu-miR-99a-5p, hsa-miR-100-5p, mmu-miR-100-5p) and two non-mitomiRs (rno-miR-99a-5p, rno-miR-100-5p) from the abundant miRNA family mir-10.

the importance of mitomiRs in mitochondrial function and metabolism.

The next important criteria for the mitomiR classification are the presence of miRNA-target associations and whether the miRNA is classified as circulating. The total number of the mitomiR-mRNA interactions is 23,151, which includes 3,318 entries with "strong" evidence of interactions and 19,833 entries with "weak" evidence (Supplementary Material 4). It should be noted that the considered subset of miRTarBase does not contain entries with interspecies interactions, meaning there are no observations where the species of the miRNA does not match the species of the targeted mRNA. Notably, a significantly fewer number of mitomiRs (in contrast to the number of non-mitomiRs) are associated with approximately the same number of mRNAs. Two-sided Fisher's exact test (average *p*-value

Characteristics, for which significant differences between mitomiRs and other miRNAs are observed

Characteristic	mitomiRs	non-mitomiRs	*p*-value (100 tests)
Total number of sequences	1,312	4,126	–
Fraction of miRNAs that target mRNA (miRTarBase)	0.94	0.60	$1.82\times10^{-89}\pm7.73\times10^{-89}$
Fraction of miRNAs that target mitochondrial mRNA (miRTarBase)	0.05	0.0002	$1.20\times10^{-56}\pm7.44\times10^{-56}$
Fraction of miRNAs that are associated with diseases (RNADisease)	0.99	0.77	$1.13\times10^{-96}\pm3.89\times10^{-96}$
Fraction of miRNAs that are associated with mitochondrial diseases (RNADisease)	0.03	0.002	$1.09\times10^{-9}\pm2.14\times10^{-9}$
Fraction of miRNAs that are circulating	0.29	0.06	$1.20\times10^{-56}\pm7.44\times10^{-56}$
Fraction of miRNAs with PAI less than 16	0.26	0.01	$1.10\times10^{-41}\pm5.97\times10^{-41}$

Note. The significance of differences between the characteristics was evaluated by averaging results over 100 iterations, each involving the random selection of one miRNA sequence from each group of homologous.

$1.82\times10^{-89}\pm7.73\times10^{-89}$) demonstrates a significant connection between the type of miRNA (mitomiR or non-mitomiR) and its association with mRNA (see the Table). All of this may indicate the important regulatory role of mitochondrial miRNAs.

Further, having a set of mitomiR-mRNA associations, we considered only mitochondrial genes and their connections to mitomiRs. The total number of such genes (which encode for rRNAs, tRNAs, and protein subunits) across the three examined mtDNAs is 111. A notable feature of the miRTarBase database is that it provides information only on protein-coding mitochondrial genes and does not cover RNA-coding genes. Additionally, we found that 65 mitomiRs target mitochondrial mRNA, while 1,247 do not. Moreover, sequences of all targeting mitomiRs are not mapped to the mtDNA, meaning that these mitomiRs are external to mitochondria. The mitomiRs target 12 mitochondrial mRNAs: *ND1*, *ND2*, *ND3*, *ND4*, *ND4L*, *ND5*, *ND6*, *COX1*, *COX2*, *COX3*, *CYTB*, *ATP6*. The largest number of mitomiRs (more than 15) target only two human mRNAs (*ATP6* and *COX1*), while the lowest number of mitomiRs (fewer than 5) is associated with the human mRNAs *ND3* and *ND4L*, as well as the mouse and rat mRNAs *COX1*. Furthermore, among 4,126 non-mitomiRs, only one (hsa-miR-15a-3p) targets mitochondrial mRNA (*ND4L*). Two-sided Fisher's exact test (average *p*-value $6.01\times10^{-22}\pm2.30\times10^{-21}$) demonstrates a significant association between the type of microRNA (mitomiR or non-mitomiR) and its interaction with mitochondrial mRNA.

The significance of the "circulating" miRNA criterion may reflect a specific mode of mitomiR transportation into mitochondria. Comparison between mitomiRs and non-mitomiRs (Fig. 2*a*) shows that mitomiRs are more prevalent among circulating miRNAs than non-mitomiRs (377 mitomiRs *vs* 251 non-mitomiRs). Two-sided Fisher's exact test (carried out on miRNA sets purified from homologous sequences) yielded an average *p*-value of $1.20\times10^{-56}\pm7.44\times10^{-56}$.

Although association with diseases demonstrates lower significance for classification than the previously mentioned criteria (due to its similarity with the "presence of targets" criterion, see the Table and Supplementary Material 4), it remains essential for understanding mitomiR functions. For each mitomiR and non-mitomiR entry, we indicated its connection (or a lack of connection) to mitochondrial diseases. Using entries from the RNADisease database, we discovered 36 mitomiRs (out of 1,312) associated with mitochondrial diseases based on their names or identifiers. On the other hand, only 9 out of 4,126 non-mitomiRs had the same association, which may be due to the targeting of nuclear mRNAs producing mitochondria-localized product. Both mitomiRs and non-mitomiRs showed associations with the disease group "Mitochondrial disease" and MNGIE-syndrome (Supplementary Material 3). Two-sided Fisher's exact test (average *p*-values $1.13\times10^{-96}\pm3.89\times10^{-96}$/$1.09\times10^{-9}\pm2.14\times10^{-9}$ for all and for mitochondrial diseases, respectively) confirmed that mitomiRs are more closely related to diseases (including mitochondrial) than non-mitomiRs. The associations with mitochondrial mRNA and diseases suggest an important role of mitomiRs in mitochondria activity.

Data. The mitomiRdb database (https://mitomiRdb.org) offers a web-based user interface for accessing mitomiR data and for performing information extraction. The database includes the following data (according to miRBase): unique identifier (MIMAT), name, nucleotide sequence, and the organism in which the mitochondrial miRNA was observed. In addition, a mitomiR's confidence flag highlights entries which are associated with mitochondrial mRNA or diseases and those mapped to the mitochondrial genome. The database provides additional information about the secondary structure of miRNA precursors, references to the supporting publications, and the list of associated diseases and genes. For entries classified as "confident" mitomiRs, a list of associated mitochondrial genes and diseases is provided, along with a note indicating the presence of the mitomiR sequence in mtDNA. All the data presented are available for download in SQLite format for further computational analysis (doi.org/10.6084/m9.figshare.22592380).

Discussion

To date, numerous microRNAs have been detected in mitochondria. It is still unknown whether the presence of these miRNAs is due to their functional roles or it is simply a coincidence that random miRNAs have been observed in these organelles. If the former is true, mitochondrial miRNAs may have special features of biogenesis and specific regulation of the expression of genes, including mitochondrial ones. In this study, we analyzed the characteristics of miRNAs to identify the factors that distinguish mitomiRs from other microRNAs and to confirm the fact that the observation of this miRNA class is not by chance.

The most significant feature of mitomiRs is the phylostratigraphic age index (PAI), which characterizes the evolutionary age of miRNA sequences. A smaller PAI for mitomiRs indicates that, on average, mitomiRs are older than non-mitomiRs. Like most old miRNAs, they are more frequently involved in a greater number of important regulatory processes, including those related to mitochondrial functions.

A significant association of mitomiRs with mRNAs (including mitochondrial ones) has been revealed based on experimentally determined interactions of microRNAs with targets. This suggests that the presence of mitomiRs within mitochondria is not by chance, and underscores the importance of mitomiRs for the functioning of the entire organism and the mitochondria in particular. Importantly, miRNA-mRNA interactions do not necessarily result in gene silencing. Approximately half of the Argonaute mRNA crosslinks involve miRNA-mRNA bindings that lack a contiguous match to miRNA seed nucleotides (Grosswendt et al., 2014), which are most critical for target association (Chandradoss et al., 2015; Salomon et al., 2015). These non-canonical binding sites, although identified by crosslinking (CLIP-methods), do not always mediate gene expression (Agarwal et al., 2015). Therefore, evolutionary conservation may serve as useful evidence of site functionality. To test this hypothesis for a single mitomiR example, we selected the only site where the crosslinking study aligned with the computer prediction (Khorsandi et al., 2018). This site is responsible for targeting *mt-ND5* by miRNA hsa-let-7a, it resides between positions 13,418–13,439 of human mtDNA and in roughly the same position (1,081 bp from the *ND5* start) in other primates. However, the site appears in the human due to a synonymous nucleotide substitution (C>T) in the site position that corresponds to the second seed nucleotide of let-7a. Meanwhile, in this position, there is a backward SNP (T>C) (rs386829181, 7×10^{-4} allele frequency) (Sherry, 2001), which has been associated with cranial meningiomas in Chinese patients.

The next factor that distinguishes mitomiRs from other microRNAs is their assignment as circulating miRNAs. Circulating miRNAs are a type of extracellular RNAs that are observed in sufficient quantities in various body fluids. The importance of this factor for the mitomiR classification suggests a potential similarity between the mechanisms of free miRNA transfer out of the cell and the translocation of mitomiRs within mitochondria.

The criterion based on mitomiR-disease associations appears less important for classification, possibly due to the overlapping associations of both mitomiRs and non-mitomiRs with targets and diseases (Supplementary Material 4). Despite this, the criterion demonstrates a significant connection between diseases and mitomiRs, which may imply that mitomiRs play an important role in regulating various biological processes, including those related to mitochondria.

The least important factor is the "confidence" of miRNA, as defined by miRBase standards. This may indicate the existence of an unknown maturation pathway for mitomiRs, which forms a microRNA-microRNA duplex with non-canonical overhanging ends, rather than the canonical 2-nucleotide overhangs that arise from Dicer and Drosha cleavage of pre-miRNA.

Despite the presence of mitochondria in all cell types of the studied mammals, the contribution of tissue specificity factor and miRNA expression levels to the difference between mitomiRs and non-mitomiRs cannot yet be assessed. This limitation arises from the fact that existing experimental observations of mitomiRs cover a small number of tissues and provide insufficient information about the expression of mitomiRs.

Conclusion

The following characteristics of mitochondrial miRNAs allow to separate mitomiRs from other microRNAs (in descending order of importance): phylostratigraphic age index (PAI), the presence of microRNA targets, and the classification of microRNAs as "circulating". These identified characteristics may help to shed light on the origin, processing and function of mitomiRs.

All experimentally investigated mitomiRs have been collected in the mitomiRdb database (https://mitomiRdb.org). The database may be useful for a more comprehensive study of microRNAs and their subclass of mitomiRs.

References

Agarwal V., Bell G.W., Nam J.-W., Bartel D.P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015;4: e05005. doi 10.7554/eLife.05005

Bandiera S., Rüberg S., Girard M., Cagnard N., Hanein S., Chrétien D., Munnich A., Lyonnet S., Henrion-Caude A. Nuclear outsourcing of RNA interference components to human mitochondria. *PLoS One*. 2011;6(6):e20746. doi 10.1371/journal.pone.0020746

Barrey E., Saint-Auret G., Bonnamy B., Damas D., Boyer O., Gidrol X. Pre-microRNA and mature microRNA in human mitochondria. *PLoS One*. 2011;6(5):e20220. doi 10.1371/journal.pone.0020220

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

Bian Z., Li L.-M., Tang R., Hou D.-X., Chen X., Zhang C.-Y., Zen K. Identification of mouse liver mitochondria-associated miRNAs and their potential biological functions. *Cell Res.* 2010;20(9):1076-1078. doi 10.1038/cr.2010.119

Breiman L. Random forests. *Mach. Learn.* 2001;45(1):5-32. doi 10.1023/A:1010933404324

Chandradoss S.D., Schirle N.T., Szczepaniak M., MacRae I.J., Joo C. A dynamic search process underlies microRNA targeting. *Cell*. 2015;162(1):96-107. doi 10.1016/j.cell.2015.06.032

Chen J., Lin J., Hu Y., Ye M., Yao L., Wu L., Zhang W., Wang M., Deng T., Guo F., Huang Y., Zhu B., Wang D. RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. *Nucleic Acids Res.* 2022;51(D1):D1397-D1404. doi 10.1093/nar/gkac814

Das S., Ferlito M., Kent O.A., Fox-Talbot K., Wang R., Liu D., Raghavachari N., Yang Y., Wheelan S.J., Murphy E., Steenbergen C. Nuclear miRNA regulates the mitochondrial genome in the heart. *Circ. Res.* 2012;110(12):1596-1603. doi 10.1161/CIRCRESAHA.112.267732

Dasgupta N., Peng Y., Tan Z., Ciraolo G., Wang D., Li R. miRNAs in mtDNA-less cell mitochondria. *Cell Death Discov.* 2015;1(1):15004. doi 10.1038/cddiscovery.2015.4

Erturk E., Enes Onur O., Akgun O., Tuna G., Yildiz Y., Ari F. Mitochondrial miRNAs (mitomiRs): their potential roles in breast and other cancers. *Mitochondrion*. 2022;66:74-81. doi 10.1016/j.mito.2022.08.002

Fan S., Tian T., Chen W., Lv X., Lei X., Zhang H., Sun S., Cai L., Pan G., He L., Ou Z., Lin X., Wang X., Perez M.F., Tu Z., Ferrone S., Tannous B.A., Li J. Mitochondrial miRNA determines chemoresistance by reprogramming metabolism and regulating mitochondrial transcription. *Cancer Res.* 2019;79(6):1069-1084. doi 10.1158/0008-5472.CAN-18-2505

Gohel D., Singh R. Different platforms for mitomiRs in mitochondria: emerging facets in regulation of mitochondrial functions. *Mitochondrion*. 2022;66:67-73. doi 10.1016/j.mito.2022.08.003

Grimson A., Srivastava M., Fahey B., Woodcroft B.J., Chiang H.R., King N., Degnan B.M., Rokhsar D.S., Bartel D.P. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*. 2008;455(7217):1193-1197. doi 10.1038/nature07415

Grosswendt S., Filipchyk A., Manzano M., Klironomos F., Schilling M., Herzog M., Gottwein E., Rajewsky N. Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol. Cell*. 2014;54(6):1042-1054. doi 10.1016/j.molcel.2014.03.049

Huang H.-Y., Lin Y.-C.-D., Li J., Huang K.-Y., Shrestha S., Hong H.-C., Tang Y., Chen Y.-G., Jin C.-N., Yu Y., Xu J.-T., Li Y.-M., Cai X.-X., Zhou Z.-Y., Chen X.-H., Pei Y.-Y., Hu L., Su J.-J., Cui S.-D., Wang F., Xie Y.-Y., Ding S.-Y., Luo M.-F., Chou C.-H., Chang N.-W., Chen K.-W., Cheng Y.-H., Wan X.-H., Hsu W.-L., Lee T.-Y., Wei F.-X., Huang H.-D. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 2020;48(D1):D148-D154. doi 10.1093/nar/gkz896

Jagannathan R., Thapa D., Nichols C.E., Shepherd D.L., Stricker J.C., Croston T.L., Baseler W.A., Lewis S.E., Martinez I., Hollander J.M. Translational regulation of the mitochondrial genome following redistribution of mitochondrial microRNA in the diabetic heart. *Circ. Cardiovasc. Genet.* 2015;8(6):785-802. doi 10.1161/CIRCGENETICS.115.001067

Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(D1):D353-D361. doi 10.1093/nar/gkw1092

Khorsandi S.E., Salehi S., Cortes M., Vilca-Melendez H., Menon K., Srinivasan P., Prachalias A., Jassem W., Heaton N. An in silico argument for mitochondrial microRNA as a determinant of primary non function in liver transplantation. *Sci. Rep.* 2018;8(1):3105. doi 10.1038/s41598-018-21091-9

Kozomara A., Birgaoanu M., Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 2019;47(D1):D155-D162. doi 10.1093/nar/gky1141

Kren B.T., Wong P.Y.-P., Sarver A., Zhang X., Zeng Y., Steer C.J. MicroRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis. *RNA Biol.* 2009;6(1):65-72. doi 10.4161/rna.6.1.7534

Li P., Jiao J., Gao G., Prabhakar B.S. Control of mitochondrial activity by miRNAs. *J. Cell. Biochem.* 2012;113(4):1104-1110. doi 10.1002/jcb.24004

Lin H.-Y., Chu P.-Y. Advances in understanding mitochondrial microRNAs (mitomiRs) on the pathogenesis of triple-negative breast cancer (TNBC). *Oxid. Med. Cell. Longev.* 2021;2021:5517777. doi 10.1155/2021/5517777

Lung B., Zemann A., Madej M.J., Schuelke M., Techritz S., Ruf S., Bock R., Hüttenhofer A. Identification of small non-coding RNAs from mitochondria and chloroplasts. *Nucleic Acids Res.* 2006;34(14):3842-3852. doi 10.1093/nar/gkl448

Mercer T.R., Neph S., Dinger M.E., Crawford J., Smith M.A., Shearwood A.-M.J., Haugen E., Bracken C.P., Rackham O., Stamatoyannopoulos J.A., Filipovska A., Mattick J.S. The human mitochondrial transcriptome. *Cell*. 2011;146(4):645-658. doi 10.1016/j.cell.2011.06.051

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

Pozniak T., Shcharbin D., Bryszewska M. Circulating microRNAs in medicine. *Int. J. Mol. Sci.* 2022;23(7):3996. doi 10.3390/ijms23073996

Russo F., Di Bella S., Vannini F., Berti G., Scoyni F., Cook H.V., Santos A., Nigita G., Bonnici V., Laganà A., Geraci F., Pulvirenti A., Giugno R., De Masi F., Belling K., Jensen L.J., Brunak S., Pellegrini M., Ferro A. miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Res.* 2018;46(D1):D354-D359. doi 10.1093/nar/gkx854

Salim U., Kumar A., Kulshreshtha R., Vivekanandan P. Biogenesis, characterization, and functions of mirtrons. *WIREs RNA.* 2022;13(1):e1680. doi 10.1002/wrna.1680

Salomon W.E., Jolly S.M., Moore M.J., Zamore P.D., Serebrov V. Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. *Cell*. 2015;162(1):84-95. doi 10.1016/j.cell.2015.06.029

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2022;50(D1):D20-D26. doi 10.1093/nar/gkab1112

Schriml L.M., Munro J.B., Schor M., Olley D., McCracken C., Felix V., Baron J.A., Jackson R., Bello S.M., Bearer C., Lichenstein R., Bisordi K., Dialo N.C., Giglio M., Greene C. The Human Disease Ontology 2022 update. *Nucleic Acids Res.* 2022;50(D1):D1255-D1261. doi 10.1093/nar/gkab1063

Sherry S.T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311. doi 10.1093/nar/29.1.308

Sripada L., Tomar D., Prajapati P., Singh Rochika, Singh A.K., Singh Rajesh. Systematic analysis of small RNAs associated with human mitochondria by deep sequencing: detailed analysis of mitochondrial associated miRNA. *PLoS One*. 2012;7(9):e44873. doi 10.1371/journal.pone.0044873

Tastsoglou S., Miliotis M., Kavakiotis I., Alexiou A., Gkotsi E.C., Lambropoulou A., Lygnos V., Kotsira V., Maroulis V., Zisis D., Skoufos G., Hatzigeorgiou A.G. PlasmiR: a manual collection of circulating microRNAs of prognostic and diagnostic value. *Cancers*. 2021;13(15):3680. doi 10.3390/cancers13153680

Tomasetti M., Santarelli L., Neuzil J., Dong L. MicroRNA regulation of cancer metabolism: role in tumour suppression. *Mitochondrion*. 2014;19:29-38. doi 10.1016/j.mito.2014.06.004

Wang W.-X., Visavadiya N.P., Pandya J.D., Nelson P.T., Sullivan P.G., Springer J.E. Mitochondria-associated microRNAs in rat hippocampus following traumatic brain injury. *Exp. Neurol.* 2015;265:84-93. doi 10.1016/j.expneurol.2014.12.018

Wang X., Song C., Zhou X., Han X., Li J., Wang Z., Shang H., Liu Y., Cao H. Mitochondria associated microRNA expression profiling of heart failure. *BioMed Res. Int.* 2017;2017:4042509. doi 10.1155/2017/4042509

Zhang X., Zuo X., Yang B., Li Z., Xue Y., Zhou Y., Huang J., Zhao X., Zhou J., Yan Y., Zhang H., Guo P., Sun H., Guo L., Zhang Y., Fu X.-D. MicroRNA directly enhances mitochondrial translation during muscle differentiation. *Cell*. 2014;158(3):607-619. doi 10.1016/j.cell.2014.05.047

Zheng H., Liu J., Yu J., McAlinden A. Expression profiling of mitochondria-associated microRNAs during osteogenic differentiation of human MSCs. *Bone*. 2021;151:116058. doi 10.1016/j.bone.2021.116058

Ziętara K.J., Lejman J., Wojciechowska K., Lejman M. The importance of selected dysregulated microRNAs in diagnosis and prognosis of childhood B-cell precursor acute lymphoblastic leukemia. *Cancers*. 2023;15(2):428. doi 10.3390/cancers15020428

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

Received September 18, 2024. Revised November 6, 2024. Accepted November 7, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):843-853

doi 10.18699/vjgb-24-92

A novel approach to analyzing the evolution of SARS-CoV-2 based on visualization and clustering of large genetic data compactly represented in operative memory

A.Yu. Palyanov [1, 2, 3], N.V. Palyanova [2]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Virology, Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
palyanov@iis.nsk.su

Abstract. SARS-CoV-2 is a virus for which an outstanding number of genome variants were collected, sequenced and stored from sources all around the world. Raw data in FASTA format include 16.8 million genomes, each ≈29,900 nt (nucleotides), with a total size of $\approx 500 \cdot 10^9$ nt, or 465 Gb. We suggest an approach to data representation and organization, with which all this can be stored losslessly in the operative memory (RAM) of a common PC. Moreover, just ≈330 Mb will be enough. Aligning all genomes versus the initial Wuhan-Hu-1 reference sequence allows each to be represented as a data structure containing lists of point mutations, deletions and insertions. Our implementation of such data representation resulted in a 1:1500 compression ratio (for comparison, compression of the same data with the popular WinRAR archiver gives only 1:62) and fast access to genomes (and their metadata) and comparisons between different genome variants. With this approach implemented as a C++ program, we performed an analysis of various properties of the set of SARS-CoV-2 genomes available in NCBI Genbank (within a period from 24.12.2019 to 24.06.2024). We calculated the distribution of the number of genomes with undetermined nucleotides, 'N's, vs the number of such nucleotides in them, the number of unique genomes and clusters of identical genomes, and the distribution of clusters by size (the number of identical genomes) and duration (the time interval between each cluster's first and last genome). Finally, the evolution of distributions of the number of changes (editing distance between each genome and reference sequence) caused by substitutions, deletions and insertions was visualized as 3D surfaces, which clearly show the process of viral evolution over 4.5 years, with a time step = 1 week. It is in good correspondence with phylogenetic trees (usually based on 3–4 thousand of genome variant representatives), but is built over millions of genomes, shows more details and is independent of the type of lineage/clade classification.

Key words: coronavirus; SARS-CoV-2; genome; variants; evolution; software system; big data; compact representation of data; analysis; visualization.

For citation: Palyanov A.Yu., Palyanova N.V. A novel approach to analyzing the evolution of SARS-CoV-2 based on visualization and clustering of large genetic data compactly represented in operative memory. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):843-853. doi 10.18699/vjgb-24-92

Funding. This research was funded by RSF, grant number 23-64-00005.

Новый подход к анализу эволюции SARS-CoV-2, основанный на визуализации и кластеризации больших объемов генетических данных, компактно представленных в оперативной памяти

А.Ю. Пальянов [1, 2, 3], Н.В. Пальянова [2]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт вирусологии, Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
palyanov@iis.nsk.su

Аннотация. Коронавирус SARS-CoV-2 – это вирус, для которого было собрано, секвенировано и сохранено рекордное количество вариантов генома из источников по всему миру. Нуклеотидные последовательности в формате FASTA включают 16.8 млн геномов, каждый длиной ≈29 900 нт (нуклеотидов), общим размером $\approx 500 \cdot 10^9$ нт, или 466 Гб. Мы предлагаем способ представления данных, позволяющий разместить без потерь всю эту ин-

формацию в оперативной памяти (RAM) обычного персонального компьютера. Более того, будет достаточно всего ≈330 Мб. Выравнивание их всех относительно исходной референсной последовательности Wunah-Hu-1 позволяет представить каждый геном как структуру данных, содержащую списки точечных мутаций, делеций и вставок. Наша реализация такого представления данных привела к коэффициенту сжатия 1:1500 (для сравнения, упаковка данных с помощью популярного архиватора WinRAR дает степень сжатия только 1:62) и обеспечила возможность быстрого вычисления редакционного расстояния между различными вариантами генома. С помощью этого подхода, реализованного в виде программы на C++, мы провели анализ различных свойств набора геномов SARS-CoV-2, содержащихся в NCBI Genbank, собранных за 4.5 года (с 24.12.2019 по 24.06.2024). Были рассчитаны распределение числа геномов от числа неопределенных нуклеотидов "N" в них, число уникальных геномов и кластеров из идентичных геномов, а также распределение кластеров по размеру (числу идентичных геномов) и продолжительности (длине временного интервала между первым и последним геномом каждого кластера). Наконец, эволюция распределений числа изменений (редакционное расстояние между каждым геномом и референсной последовательностью), вызванных заменами, делециями и вставками, была визуализирована в виде 3D поверхностей, наглядно изображающих процесс вирусной эволюции в течение 4.5 лет, с интервалом в одну неделю. Такая визуализация хорошо соотносится с филогенетическими деревьями (обычно рассчитываемыми по 3–4 тыс. представителей вариантов генома), но строится на основе миллионов геномов, отображает больше деталей и не зависит от типа классификации линий/клад.

Ключевые слова: коронавирус; SARS-CoV-2; геном; варианты; эволюция; программная система; большие данные; компактизация; анализ; визуализация.

Introduction

SARS-CoV-2 coronavirus, the very first sample of which, named Wuhan/Hu-1/2019, was collected on 24 December 2019 (Wu et al., 2020), caused the largest pandemic in the last 100 years (since the Spanish flu of 1918–1920). 4.5 years later, it is still persisting, evolving and being detected in people around the world, albeit in much smaller numbers than during the peak of the pandemic and with less severe consequences. However, usually infection rates rise again with the arrival of autumn, and 2024 is no exception. According to the World Health Association (https://data.who.int/dashboards/covid19/cases, section "COVID-19 cases, country level trends"), by mid-September 2024 many countries have already started to experience an increase in the incidence of the disease. For example, in Russia, 26.7 thousand cases of SARS-CoV-2 infection were registered in July 2024, 24.7 thousand – in August, and already 62.2 thousand in the first half of September. In different countries there are certain features of the dynamics of the number of infections, depending on many factors, the analysis of interrelationships between which, in particular, we studied in (Palyanova et al., 2022, 2023).

SARS-CoV-2 virus samples obtained worldwide are sequenced and uploaded to databases, the largest of which are GISAID (gisaid.org) and NCBI Genbank (www.ncbi.nlm.nih.gov/sars-cov-2/) – as of 06.2024 they contain more than $16.7 \cdot 10^6$ and more than $8.6 \cdot 10^6$ SARS-CoV-2 genome samples, respectively. In comparison, human influenza virus, the earliest samples of which date back to 1905 in GISAID, has been represented by approximately $5.22 \cdot 10^5$ genomes over more than a century. Given that the typical genome size of SARS-CoV-2 is 29.9 kb, the total volume of the genomes of this virus represented in GISAID is about $500 \cdot 10^9$ nt (or 465 Gb), and in Genbank, about $258 \cdot 10^9$ nt (241 Gb). All these data will not fit into the RAM of an average modern PC (16...64 Gb), while working with them directly from files located on a hard disc (HDD) or solid-state drive (SSD) will be much slower than from RAM. Read speeds from modern HDD/SDD/RAM have typical values of about 0.2, 3 and 50 Gb/s, respectively, so for significant data volumes and computational loads, working specifically with RAM is highly desirable.

Despite vaccination and drug treatment, there is currently no way to completely eliminate SARS-CoV-2 (Cui et al., 2023), so it is likely to remain with mankind for a long time, adding to the numerous list of more than 200 acute respiratory infections, including influenza, respiratory syncytial virus, rhinovirus, coronavirus, adenovirus, and other infections that cause catarrhal inflammation of the respiratory tract.

The longer a virus exists, the more changes accumulate in its genome – each new generation is obtained as a result of replication of viruses of the previous generation, in the process of which errors/changes may occur. Gene mutations can result in substitutions, deletions and insertions of one or more nucleotides, as well as in translocations, duplications and inversions of different parts of the gene. For example, point mutations occur spontaneously with frequencies of 10^{-8}–10^{-6} for DNA viruses, and 10^{-6}–10^{-4}, for RNA viruses (Sanjuán, Domingo-Calap, 2016), the molecular machinery for replication of which (RNA polymerase) lacks an error-correcting mechanism (exonuclease). Coronaviruses and toroviruses, which do have it (Campagnola et al., 2022), are exceptions, as they have some of the largest genomes for RNA viruses, and too rapid accumulation of errors in them is apparently not desirable and does not favour virus survival.

According to (Amicone et al., 2022), the error rate during SARS-CoV-2 replication is $1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ substitutions per position per infectious cycle of cell infection (i. e. from virus entry into the cell to the exit of new virions from the cell). At the same time, the rate of evolutionary changes in the SARS-CoV-2 genome is estimated to be $8.9 \cdot 10^{-4}$ substitutions per position per year (Sonnleitner et al., 2022).

In addition to the above-mentioned mechanisms that can affect a single genome, there are also those that can create new combinations based on the genetic material of different genome variants. When two different variants of the same virus infect the same organism simultaneously (e. g., Delta and Omicron infection in the case of SARS-CoV-2 (Bolze et al., 2022)), they may interact during replication (Simon-Loriere,

Holmes, 2011), resulting in recombinants or reassortants (in the case of viruses with a segmented genome).

Regardless of which mechanism caused a particular change, the Levenshtein distance (also called edit distance) can be calculated for any pair of genomes of the considered virus, defined as the minimum number of single-character operations (substitutions, deletions, insertions) that need to be performed in the first genome to produce the second genome (or in the second genome to produce the first one – the result is the same). In other words, the Levenshtein distance sets a metric that defines the difference between two sequences of symbols. Thus, each variant of the SARS-CoV-2 genome out of the millions available can be compared to the original Wuhan-Hu-1 reference genome. For this purpose, it is necessary to perform a global alignment of all available sequences with respect to the reference, which we performed using the NextAlign/NextClade software (https://github.com/nextstrain/nextclade) (Aksamentov et al., 2021). As a result, for each considered viral genome sequence, we calculated a list of changes (deletions, insertions, or point substitutions) that distinguish it from the reference genome sequence.

For a virus with a genome size of 30,000 nt, a single point substitution could occur at each of 30,000 positions and result in a change of an existing nucleotide (A, T, G, or C) to one of the three others, giving rise to $30{,}000 \cdot 3 = 90{,}000$ different variants. A single insertion can be made at 30,001 positions – added either at the beginning or end of the sequence, or in any of the 29,999 spaces between the available nucleotides. It may contain any of the four letters of the alphabet, i. e. there are 120,004 different variations of such insertions. Finally, a deletion can occur in any of 30,000 positions, resulting in a number of variants equal to the number of positions. However, the deletions and insertions that leave the virus viable most often occur in blocks that are multiples of three, because otherwise such a change would result in a shift in the reading frame, which in the vast majority of cases makes the genome non-viable. Thus, even one single change can be carried out in more than 240,000 different ways, although a significant part of them (especially those corresponding to deletions and insertions) will make the genome non-viable.

The combination of two arbitrary point substitutions is already $(240{,}000)^2 = 5.8 \cdot 10^{10}$ and three – $(240{,}000)^3 = 1.4 \cdot 10^{16}$ variants, and this time among them there will be those with no reading frame shift (the result of changes – deletion or insertion of one triplet, i. e. three subsequent nucleotides). At the same time, the number of differences between some modern variants of SARS-CoV-2 and the reference genome already exceeds 200, and, for scale, the editorial distance between SARS-CoV-2 and the nearest genome of another virus – bat coronavirus, RaTG13 – is 1,136 (96.1 % of nucleotides match) (Zhou et al., 2020; Temmam et al., 2022). A number of questions about the space of variants of SARS-CoV-2 genetic sequences are discussed in more detail in (Palyanov, Palyanova, 2023), where, in particular, it is shown that thenumber of already realized variants of the virus is a negligible fraction of those that are potentially possible. Thus, both the continued monitoring of new SARS-CoV-2 variants and the analysis of the millions of genomes already accumulated over the past 4.5 years are of interest both from a practical point of view and for obtaining new fundamental knowledge in virology and epidemiology.

Materials and methods

The results presented in this paper were obtained using a software package that we created to analyze the evolution of viruses. The C++ programming language and the development environment "Microsoft Visual Studio Community 2019" were used. One third-party software module required to perform global alignment of viral genomes was used – NextAlign by NextClade (https://github.com/nextstrain/nextclade/releases). Workstation based on Intel Core i7-10700K, 3.8 GHz, 8 cores, 32 Gb DDR4 operative memory was used for all computations.

The data for the analysis – the genetic sequences of SARS-CoV-2. The data used in this paper are the complete set of SARS-CoV-2 genomes contained in the Genbank database (www.ncbi.nlm.nih.gov/sars-cov-2/) on day 24.06.2024 (4.5 years since the collection of the first sample of this virus, Wuhan-Hu-1, 24.12.2019). The number of genomes is 8,641,740 and their total size is 242 Gb. The SARS-CoV-2 reference genome, which is 29,903 nt long, consists of a 5′ UTR (265 nt long), a CDS (which is 29,409 nt long and encodes 29 proteins (Bai et al., 2022)), and a 3′ UTR (229 nt long) (UTR is the untranslated region, CDS is the coding sequence). This dataset, which continues to grow over time, was analyzed to investigate evolutionary changes in SARS-CoV-2. Another source of data is the GISAID database, which includes a significant amount of the Genbank data; other genomes from it have yet to be analyzed and compared with the results for the Genbank genomes.

Data quality, preliminary data analysis and filtering. One of the first issues that arise when working with a set of nucleotide sequences of viral genomes is their quality. In particular, the sequences may contain not only letters encoding nucleotides (A, T, G, C), but also "N"s indicating unidentified, unknown nucleotides in the corresponding positions. The larger the number of "N"s, the greater the uncertainty, and the worse for the results of the analysis and their validity. In this regard, it is of course helpful to know how many such sequences there are in the dataset under study, and how many "N"s are present in them.

Our calculations showed that out of the full sequence set (8,641,740), unidentified "N" nucleotides occur in 6,609,933 genomes (76.5 %) and are missing only in the remaining 2,031,807 (23.5 %). However, if we consider only CDS, the number of genomes without "N" almost doubles, reaching 3,742,117 (43.3 %). In addition, we plotted the dependence of the frequency of "N" occurrence on the position in the genome on the basis of the full set of sequences for which the global alignment was performed (Fig. 1).

As can be seen, there are two most significant peaks, at the beginning and at the end of the genome, corresponding to the non-coding regions of 5′ UTR and 3′ UTR, the total length of which is 1.65 % of the length of the whole genome. It is also known that in the genetic sequences of SARS-CoV-2 from GISAID and Genbank, the non-translational regions have a high variation in their lengths (Palyanov, Palyanova, 2023). Considering that the number of genomes in which "N" occurs in UTRs and does not occur in CDS is 22.5 % of all genomes, excluding UTR regions from consideration will almost double

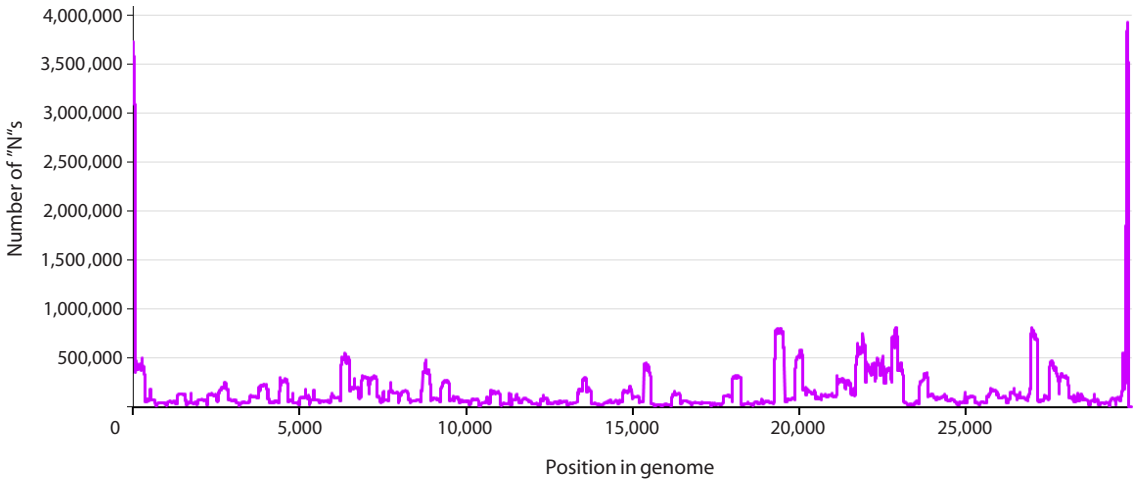


Fig. 1. Frequency dependence of "N" occurrence rate as a function of position in the genome (abscissa axis), obtained by summarization over the full set of SARS-CoV-2 genetic sequences from Genbank in the interval from 24.12.2019 to 24.06.2024

the set of data suitable for analysis (23.5 % of sequences in which "N" does not occur at all, neither in CDS nor in UTRs, will be supplemented with another 22.5 % in which "N" is present only in UTRs).

Depending on what is the distribution of genomes by the number of "N"s contained in their CDS, we can either use those genomes with only a few "N"s (in comparison with editing distance values of the order of 100 point substitutions, this is an insignificant value, although their presence introduces some uncertainty), or use only those genomes with no "N"s in the CDS. Having plotted the mentioned distribution (Fig. 2), we found that exactly one "N" is present in 1.8 % of genomes, two and three – in 0.8 and 0.9 %, respectively, and the number of "N"s between 1 and 10 per genome – in 5.4 %. As a result, at this stage it was decided to work only with genomes in which "N" is absent in CDS, and to use only CDS in calculations, excluding 5′ UTRs and 3′ UTRs.

Methods, algorithms and data structures. To construct global alignments of all genomes (using the Wuhan-Hu-1 genome as a reference), we used the console version of NextAlign (running in multithreaded mode), called with the necessary parameters being passed from our software system. This happens during the first run or when the alignments need to be recalculated (e. g., if a different genome dataset is used). On the full dataset mentioned above, consisting of 8.6 million SARS-CoV-2 genomes, the calculation of alignments takes about one day on a workstation with an Intel Core i7-10700K @ 3.8 GHz processor (8 cores, 16 threads) and 32 Gb of RAM (DDR4, 3,600 GHz). The output of NextAling for all calculated alignments is stored in files on the hard drive in the working directory of the program, which are then used by our system as input data used for analysis. The files are composed as tables with several dozen columns, including various genome characteristics and metadata, as well as lists of mutations, deletions and insertions that distinguish the considered genome from the reference genome.

As the data are read, a list of structures is dynamically generated in the computer's RAM, each of which includes the virus name, collection date, geographical data, and a complete set of changes that distinguish the current variant from the reference genome:

- a list of point mutations (single position substitutions), each element of which contains the position number in the genome corresponding to the mutation and a letter encoding the nucleotide that appears at that position as a result of the substitution (the previous nucleotide that was present before the mutation is not stored – it can always be read from the corresponding position in the reference genome if necessary);
- a list of deletions, each of which is defined by two numbers – the positions of the beginning and the end of the deletion;
- a list of insertions, each of which is defined by the position in the genome immediately after which the insertion took place, as well as by the inserted sequence.

This organization of the data allows two arbitrary genomes to be compared quickly and easily. It is especially quick to determine whether they are identical or not. Instead of comparing each of the 29,409 positions of the first and second genomes, it is enough to simply compare the number of elements in their lists of point mutations, deletions and insertions – at least one difference makes it clear that the genomes are different. However, in this way it is possible to obtain not only the result of genome comparison, but also to calculate the editorial distance between them. Matching elements of the difference lists do not contribute to the difference between the genomes, whereas each element of difference from the reference, present in one genome and absent in the other, adds a corresponding amount of difference. Each substitution that occurred at the same position in both genomes but resulted in substitutions to different nucleotides also, of course, adds +1 to the edit distance. Given that the list sizes are quite small, the comparison is much faster than comparing two genomes without prior alignment.

Our proposed method of compact representation of nucleotide sequences of related genomes in computer memory has much in common with the compression method that represents sequences in the form of a phylogenetic tree with substitutions on edges. Moreover, the very representation of each genome

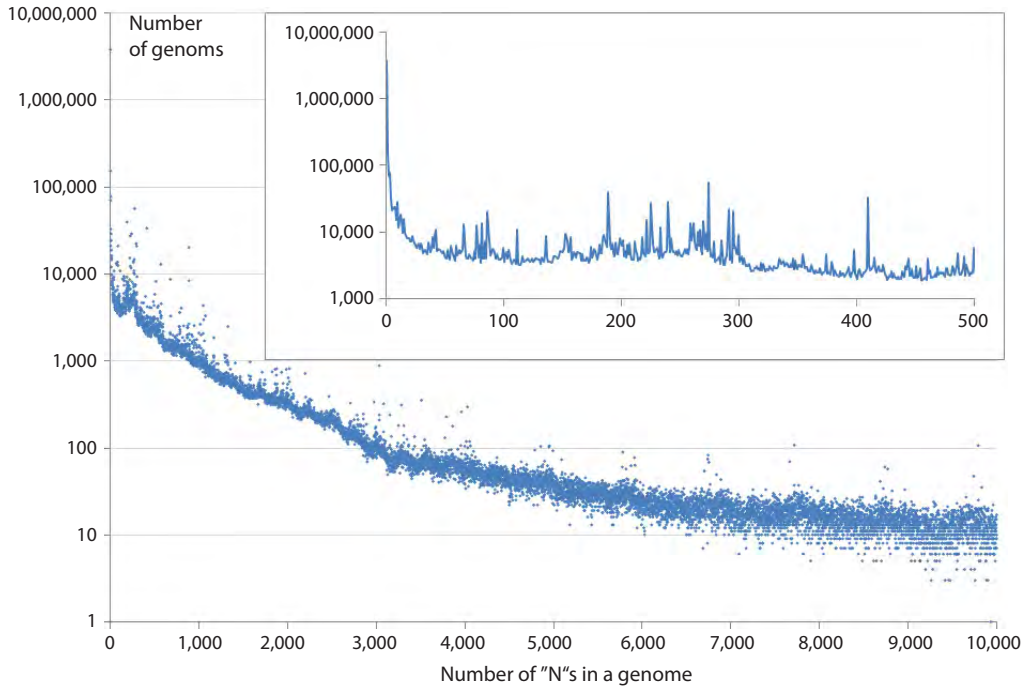


Fig. 2. Distribution of genomes by the number of unidentified "N" nucleotides in them.

Calculated over the full set of SARS-CoV-2 genetic sequences from Genbank between 24.12.2019 and 24.06.2024. The inset shows the same relationship, but with higher resolution, for the number of "N"s in the genome between 0 and 500.

as a set of changes that need to be made to go from the reference to the genome under consideration is based on the same data representing the structure of the phylogenetic tree built on the basis of multiple alignments of the sequences under consideration.

The peculiarity of our implementation is that the data structure in a PC operative memory, representing the set of sequences under consideration, is not a phylogenetic tree, but instead is a list of its "leaves" sorted in chronological order, by the date of obtaining samples. For such tasks as analyzing not just the available spectrum of virus variants, but its evolutionary changes taking into account the time of their emergence, our approach provides a significant advantage in the speed of data access. The point is that it allows us to move along the time axis simply by increasing or decreasing the index of an array element consisting of time-ordered genomes. And in the tree representation, the search for all genomes corresponding to a certain year, month and day, in general case, may require traversing the whole tree, and so for each genome variant being processed. At the same time, each "leaf" in our approach contains all the information about its "branch" of the tree, which allows one to easily and quickly calculate the editing distance for any pair of genome variants.

Results

Cluster structure of the genomes dataset

In the course of the study, we noticed that among the genomes under consideration, there are quite often genomes with CDSs that are 100 % identical to each other, while the date of sample collection, geographical data, and other metadata most often are different. By adding a function to our software system to identify all genomes with identical CDSs (and combine them into "clusters"), we have divided the entire set of genomes into such groups. The statistics on them turned out to be as follows from the Table.

Also, we calculated the relationship between cluster size and the number of clusters of a particular size (Fig. 3). At the same time, there is no obvious dependence between cluster sizes and their lifetime; the distribution is a cloud of points, most of which is concentrated in the region from 1 to 1,000 on the "cluster size" axis and from 1 to 500 on the "cluster lifetime" axis (Fig. 4).

We also plotted (Fig. 5) all clusters of size ≥ 200 on the time axis and represented them as lines, with the beginning and end corresponding to the clusters' existence intervals. In addition, we added all clusters of size 100–199, the end of the existence interval of which has a value ≥ 1,000 days since the first SARS-CoV-2 genome was obtained. This set of clusters covers the entire timeline, although there are clearly other clusters in the interval between 3 and 3.5 years, but all of them are smaller than those shown in the Figure 5. The individual line "19A", the longest in the Figure 5, corresponds to the cluster with the longest lifetime (1,539 days or 4.2 years) mentioned in Table. In this regard, the genetic line 19A, which has survived for such a long time, appears to be quite interesting. This genome variant was detected quite stably both at the beginning of the pandemic and in 2023–2024.

A novel approach to visualizing evolutionary changes in SARS-CoV-2

Having obtained the ability to quickly calculate the editorial distance between a pair of any nucleotide sequence variants, we did it for the whole set of SARS-CoV-2 genomes from

Statistical data on clusters composed of genomes with completely identical CDSs, including size and extent in time

The total number of SARS-CoV-2 genomes from Genbank in the interval from 24.12.2019 to 24.06.2024	8,641,740
The number of genomes from the full dataset, the CDS of which doesn't contain unidentified/unknown nucleotides "N"	3,742,117
The number of genomes (in the set of genomes with CDS without "N"s) with a unique nucleotide sequence found nowhere else in other genomes	1,690,699
The number of genomes composing clusters with a size = 2 or larger (in the set of genomes with CDS without "N"s)	2,051,448
The number of clusters with a size ≥ 2	461,511
The number of clusters for which the time of existence (the interval between the earliest and latest date of genome sample collection dates among the cluster members) is more than 1 day	366,427
Maximal size of the cluster (the number of genomes which compose it)	12,824
Maximal time of cluster existence (lineage 19A)	1,539 days
Average cluster size	4.4
Average cluster existence time	14.8 days
Average cluster existence time (excluding those which exist only 1 day)	18.4 days

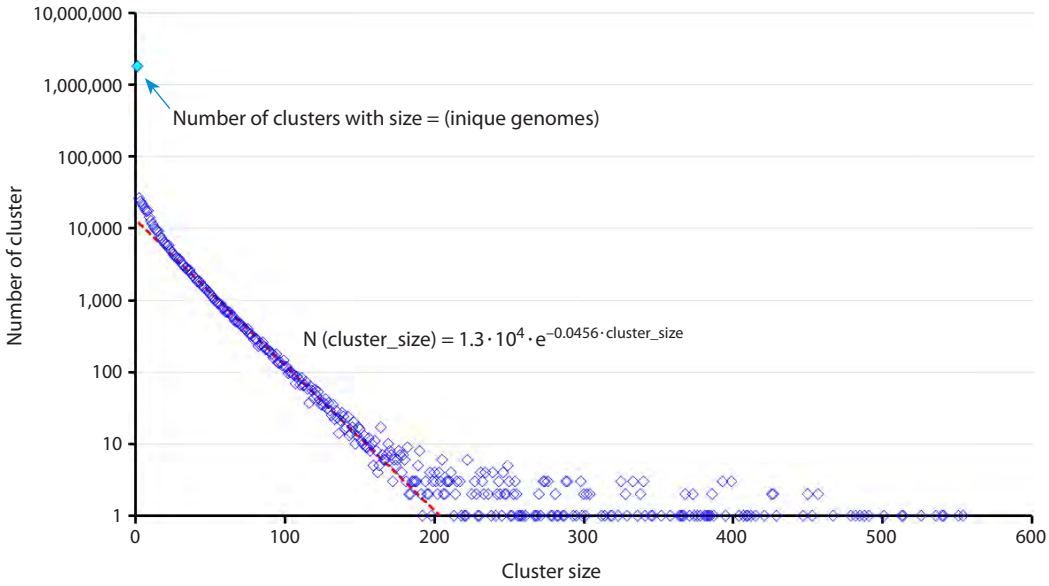


Fig. 3. The dependence between the number of clusters and their size for the set of SARS-CoV-2 genomes from Genbank in the interval from 24.12.2019 to 24.06.2024.

In the interval of cluster size values from 20 to 200, it is well approximated by an exponent with the parameters indicated in the Figure.

Genbank for 4.5 years. Thus, for each genome, there is a pair of numbers – the collection date of the genome sample and the editing distance between it and the reference. Sometimes different variants of genomes appear to possess the same pair of values "date + editorial distance", because the same value of the editing distance can be a result of, for example, a deletion of length 30, an insertion of the same length, or 30 point mutations scattered throughout the genome. If we introduce a third value – the number of cases in which the genome variant has a certain editing distance between it and the reference and a certain date of sample collection – then we can calculate triples of these values on the basis of the complete set of SARS-CoV-2 genomes and display them as a surface, which we did (Fig. 6).

In Figure 6, we have marked a number of interesting landscape elements with blue dots, for each of which the corresponding spectrum of variants has been calculated. For many of them, it was possible to place this information in the figure. The landscape shows regions with different features – narrow extended "mountain ranges" with a beginning, an end and a characteristic slope angle (which has close values for most of them), apparently related to the rate of accumulation of changes in the genome appearing due to point nucleotide substitutions.

There are also regions in which the editing distance for the entire set of variants existing at a given point in time changes rapidly and significantly in the mean value or experiences branching, splitting into several parallel, visually distinguis-

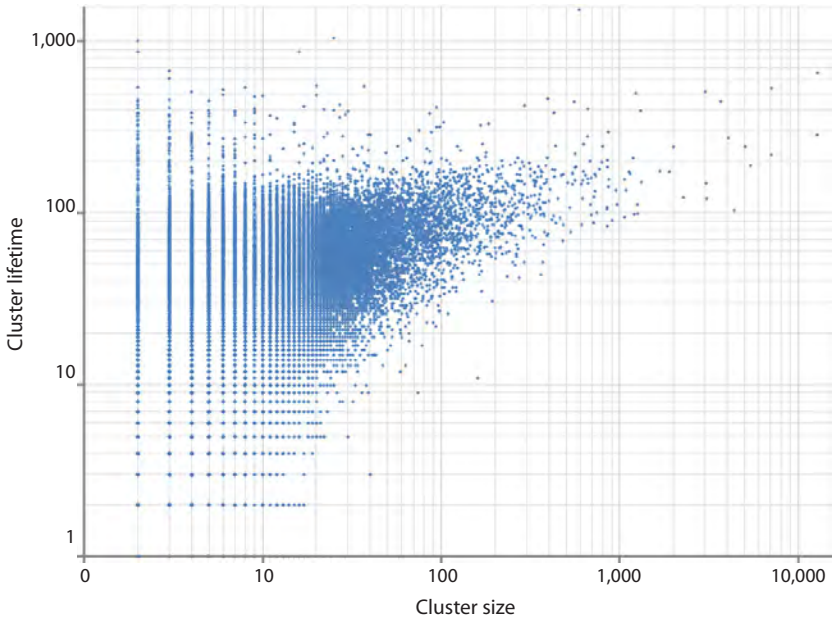


Fig. 4. Dot cloud representing the set of SARS-CoV-2 genomes from Genbank (between 24.12.2019 and 24.06.2024) using their "cluster size" and "cluster lifetime" features.

hable paths. Assuming that such abrupt and significant changes could be due to deletions, insertions, or recombination events, we constructed three more figures similar to Figure 6, for which we used not the full value of the editing distance, but its three separate contributions – from a set of point substitutions (Fig. 7), deletions (Fig. 8), and insertions (Fig. 9). Of the total number of genomes (with no "N" in CDS), 3,742,117, the number of genomes with mutations relative to the reference was 3,741,518, the number of genomes with deletions was 3,520,077, and the number of genomes with insertions was 528,414.

As can be seen in Figures 7–9, the dynamics of evolutionary changes introduced into SARS-CoV-2 genomes by different evolutionary mechanisms differ quite significantly for substitutions, deletions, and insertions. From Figure 7 we can conclude that the accumulation of the number of point mutations (substitutions) increases linearly on a large time scale (especially in the case of following the upper boundary of the evolutionary pathway region). In 4.5 years, about 130 point mutations were accumulated, i. e., the growth rate was about 29 nt per year (≈2.4 per month or ≈0.08 per day). The impact of deletions is also significant, and their number also grows near linearly with time, but at a slower rate of about 50 nt in 4.5 years, i. e., about 11 per year or slightly less than one per month. And finally, insertions, as can be seen from Figure 8, make a noticeably smaller contribution than substitutions and deletions, which, with the exception of the first year, practically does not grow with time – it stays at the level of +20 nt relative to the reference genome (although the content of these insertions, in principle, can change over time, from year to year).

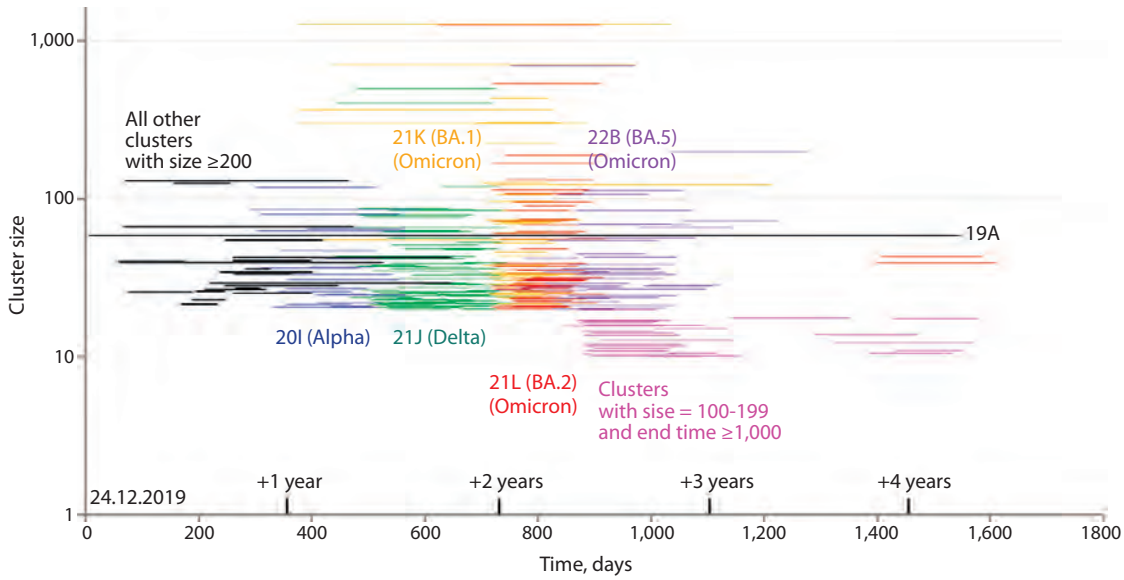


Fig. 5. The largest clusters (size ≥200) and their existence intervals (lines connecting the day of the first appearance of the genome variant representing this cluster and the day when the last sample with the same genome was taken).

All clusters of size ≥200 depicted in the same colour belong to the same genetic lineage, the name of which is also displayed with that colour. All genomes belonging to the same line are equal, and genomes of different lines differ between each other. The exceptions are "all other clusters with size ≥200" depicted in black, which represent a collection of different genetic lineages (19A, 20A, 20B, 20C, 20E and 20F), and clusters of size 100–199 depicted in magenta (which have an end-of-life date ≥1,000 days from the date of collection of the first SARS-CoV-2 genome, 24.12.2019).

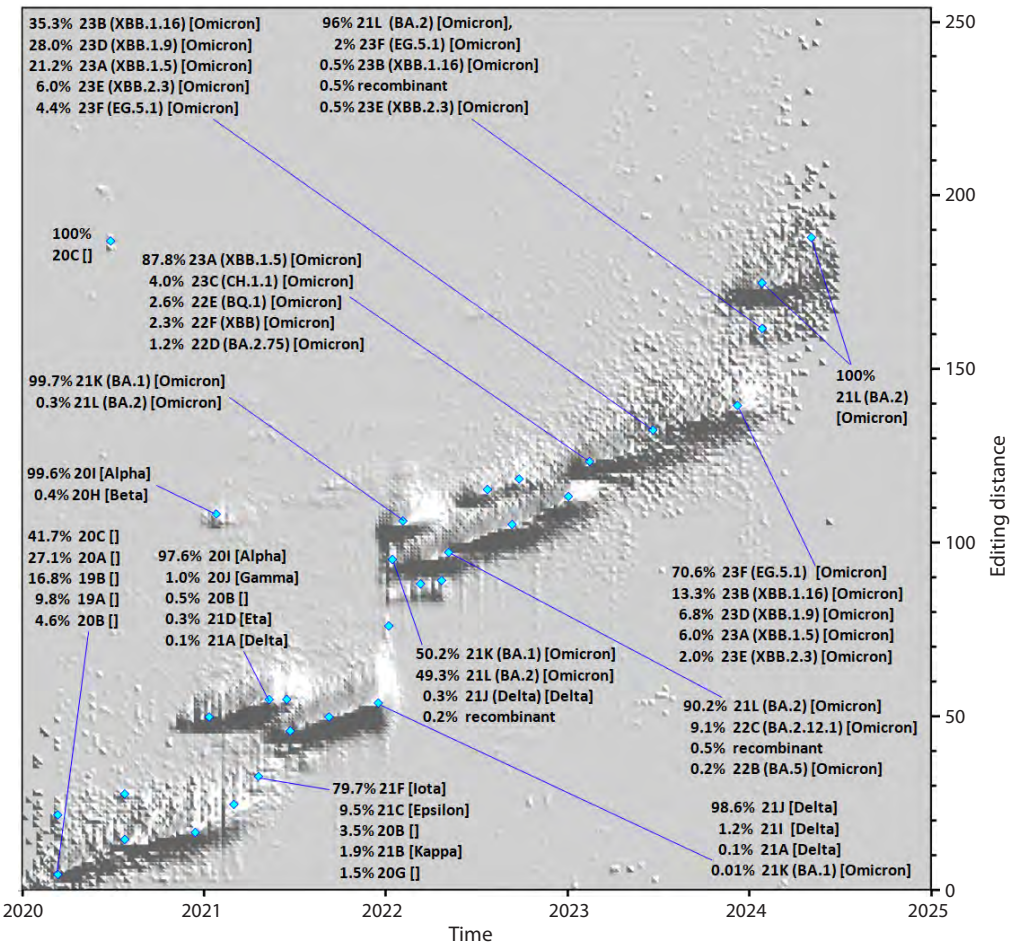


Fig. 6. The landscape of the space of SARS-CoV-2 variants "visited" by the virus variants during the period from 24.12.2019 to 24.06.2024, projected on three axes: *OX* – time (sampled at 1 week), *OY* – difference (editing distance) between a point on the landscape and a reference genome, *OZ* – fraction of genomes corresponding to a point with certain *X* and *Y* values, referred to the total number of genomes collected in week *X*. Such normalization is necessary due to the fact that the dependence of the number of genome samples collected in one or another week all over the world has significant changes over time, and without the proposed normalization the weekly distribution in case of, for example, 100 genomes will be completely invisible in comparison with some other week represented by 10,000 genomes, whereas even for 100 genomes distributions are quite informative.

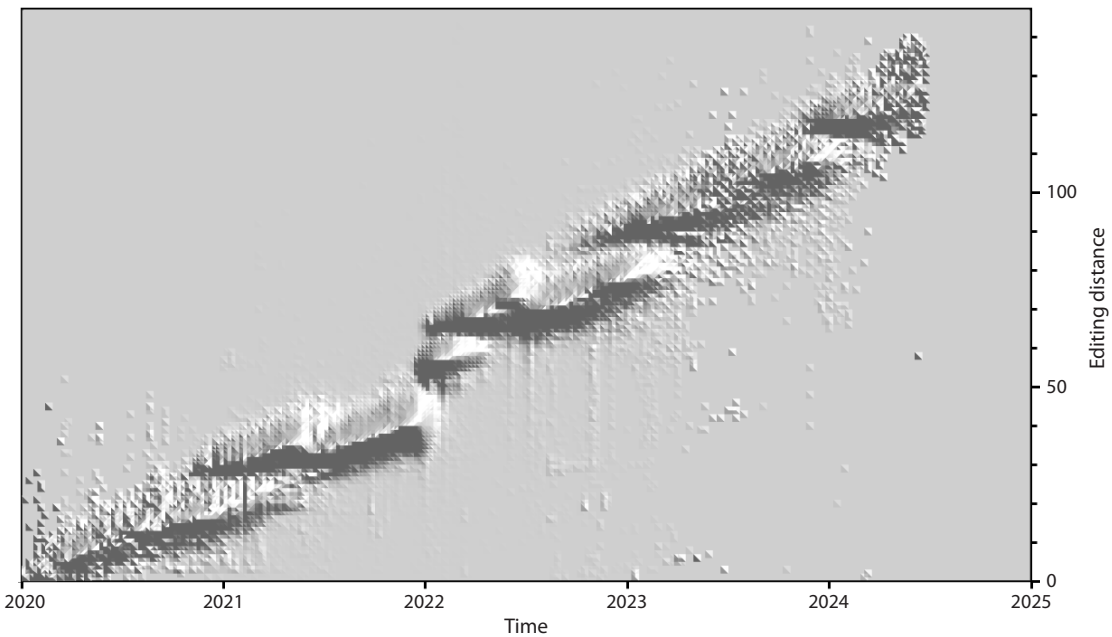


Fig. 7. The landscape of evolutionary changes based on contributions from point substitutions only.

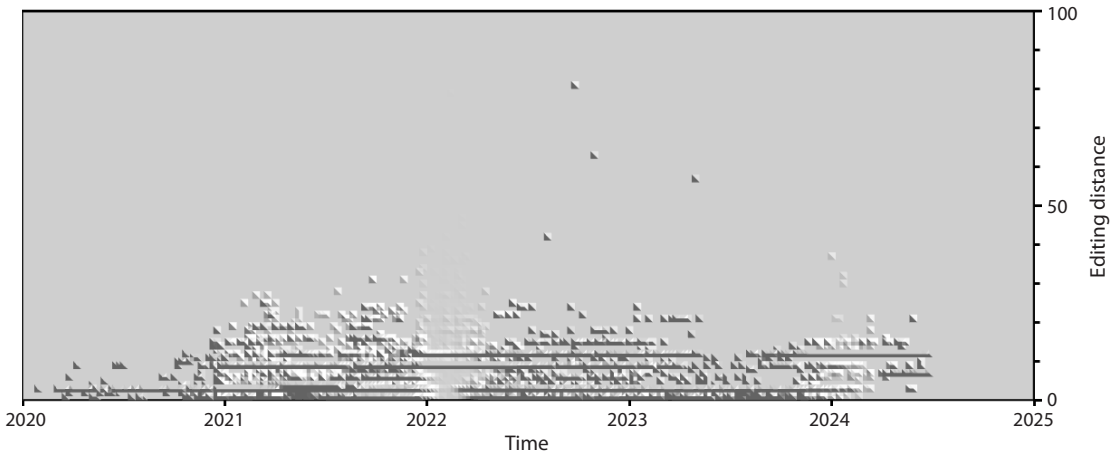


Fig. 8. The landscape of evolutionary changes based on contributions from point insertions only.

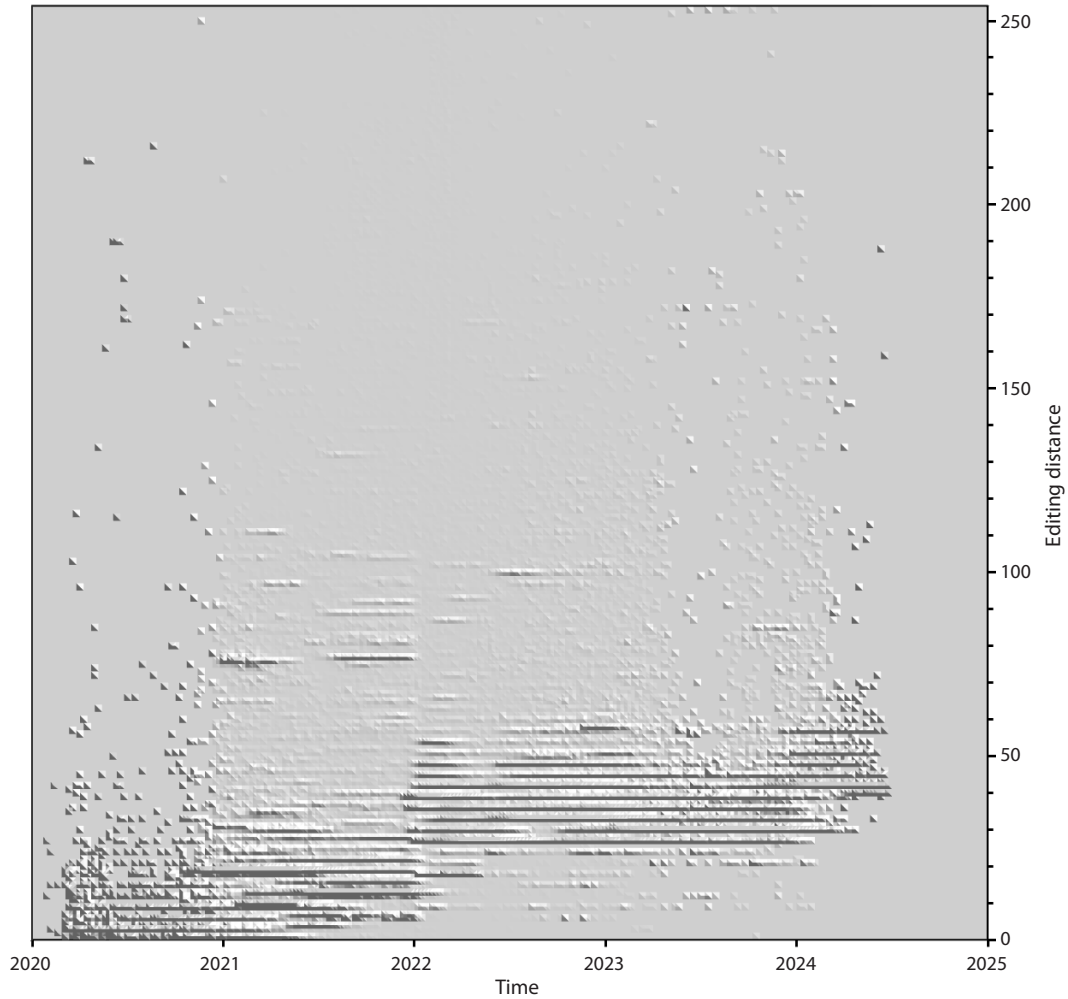


Fig. 9. The landscape of evolutionary changes based on contributions from point deletions only.

Discussion

We performed a number of evaluations and calculations, mainly using software tools developed by us, to improve our understanding of which features and trends related to the evolution of SARS-CoV-2 coronavirus genetic sequences can be found and effectively used. The proposed method of visualizing landscapes of evolutionary changes allowed us to display many details and features that are not visible, for example, on a phylogenetic tree. At the same time, rapid changes in the evolutionary trajectory accompanied by stepwise changes in the value of the editing distance, as can be seen from Figure 6, are usually accompanied by a change of the dominant virus

variant in the population. Thus, for example, for one of such "jumps" on the evolutionary landscape, we were able to see the genetic lineages "Iota", "Epsilon", and "Kappa" in the corresponding spectrum of variants (approximately in the first quarter of 2022).

The observed non-growing contribution to the editing distance of inserts, previously mentioned for the results related to Figure 8, may be due to the fact that too many inserts may compromise the stability of the virus. Increasing the number of inserts increases the physical size of the genome and thus may impair its ability to fit inside the protein envelope, which is presumably designed to contain an object of a certain size. Thus, in the course of evolution, the number of insertions relative to the reference genome may increase, but not exceeding 20–30 nt. If some of the new variants turn out to be more adapted than their predecessors, they may soon displace them. It can be seen that only at the beginning of 2024, variants with no insertions at all are disappearing – apparently, due to the fact that during 4 years of evolution, such insertions were found, which turned out to be noticeably more adaptable than variants with no insertions at all and became established in the population. It can also be seen in Figures 8 and 9 that the values of contributions to the editing distance from deletions and insertions in most cases have a length multiple of three, which has an obvious explanation – other length variants will lead to a shift of the reading frame during the synthesis of proteins encoded in the genome and, in most cases, to non-viable copies of genomes, the virions of which cannot be assembled.

Conclusion

As a result of this work we have obtained the following main results:

- a method of representing nucleotide sequences of virus genomes, which provides their extremely compact representation in computer memory, has been proposed and implemented as a computer program. On the example of SARS-CoV-2 coronavirus it is has been shown that compression of ≈1,500 times is provided. Using it for transmission of genetic data over the internet could reduce the load on servers and network traffic by a corresponding number of times (especially when transmitting large datasets);
- for the complete set of SARS-CoV-2 genomes (without "N"s in CDS), the presence of clusters of completely identical genomes has been investigated. It has been found that their size can exceed 10,000, and their lifetime can cover up to several hundred days;
- a new way of displaying the evolutionary dynamics of viruses in the form of a landscape visualizing the projection of the space of virus genome variants on three axes – time (T), editing distance to the reference genome (D), and the fraction of genomes (P) at each point (T, D) in the total number of genomes corresponding to a given T is proposed;
- it is also shown that the landscape constructed for D (calculated as the sum of contributions from point mutations, deletions and insertions) can be divided into three separate landscapes calculated separately for each of the contributions. Each of them has a different character, allowing the contribution and impact of each of the mentioned mechanisms on virus evolution to be estimated. The constants characterizing each of the mechanisms and the rate of changes acquired due to it have been calculated;
- the fact that the lineage 19A has existed for the longest time compared to the other clusters, covering the entire pandemic period, allows us to propose to create new vaccines against SARS-CoV-2 on the basis of this lineage, as it retains the greatest competitiveness compared to the other variants, and thus contains the most characteristic features of this virus that can be recognized by the immune system.

Our further plans include investigation of the possibilities of the proposed method of evolution visualization in more detail, but we can already state that it seems to be useful, has the potential for further use and development, and can be applied not only to SARS-CoV-2, but also to other viruses. The same can be said about the proposed method of compact representation of viral genomes, the application of which in all areas related to the storage, network transmission, processing and analysis of a large number of variants of related genomes (both viruses and living organisms) will provide significant advantages.

References

Aksamentov I., Roemer C., Hodcroft B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Software*. 2021;6(67):3773. doi 10.21105/joss.03773

Amicone M., Borges V., Alves M.J., Isidro J., Zé-Zé L., Duarte S., Vieira L., Guiomar R., Gomes J.P., Gordo I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health*. 2022;10(1):142-155. doi 10.1093/emph/eoac010

Bai C., Zhong Q., Gao G.F. Overview of SARS-CoV-2 genome-encoded proteins. *Sci. China Life Sci*. 2022;65(2):280-294. doi 10.1007/s11427-021-1964-4

Bolze A., Basler T., White S., Rossi A.D., Wyman D., Dai H., Roychoudhury P., Greninger A.L., Hayashibara K., Beatty M., Shah S., Stous S., McCrone J.T., Kil E., Cassens T., Tsan K., Nguyen J., Ramirez J., Carter S., Cirulli E.T., Barrett K.S., Washington N.L., Belda-Ferre P., Jacobs S., Sandoval E., Becker D., Lu J.T., Isaksson M., Lee W., Luo S. Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. *Med*. 2022;3(12):848-859. doi 10.1016/j.medj.2022.10.002

Campagnola G., Govindarajan V., Pelletier A., Canard B., Peersen O.B. The SARS-CoV-2 nsp12 polymerase active site is tuned for large-genome replication. *J. Virol*. 2022;96(16):e0067122. doi 10.1128/jvi.00671-22

Cui X., Wang Y., Zhai J., Xue M., Zheng C., Yu L. Future trajectory of SARS-CoV-2: Constant spillover back and forth between humans and animals. *Virus Res*. 2023;328:199075. doi 10.1016/j.virusres.2023.199075

Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):839-850. doi 10.18699/VJGB-23-97

Palyanova N.V., Sobolev I.A., Alekseev A., Glushenko A., Kazachkova E., Markhaev A., Kononova Y., Gulyaeva M., Adamenko L., Kurskaya O., Bi Y., Xin Y., Sharshov K., Shestopalov A. Genomic and epidemiological features of COVID-19 in the Novosibirsk region during the beginning of the pandemic. *Viruses*. 2022;14(9):2036. doi 10.3390/v14092036

Palyanova N.V., Sobolev I.A., Palyanov A.Yu., Kurskaya O.G., Komissarov A.B., Danilenko D.M., Fadeev A.V., Shestopalov A.M. The development of the SARS-CoV-2 epidemic in different regions of

Siberia in the 2020–2022 period. *Viruses*. 2023;15(10):2014. doi 10.3390/v15102014

Sanjuán R., Domingo-Calap P. Mechanisms of viral mutation. *Cell. Mol. Life Sci*. 2016;73(23):4433-4448. doi 10.1007/s00018-016-2299-6

Simon-Loriere E., Holmes E.C. Why do RNA viruses recombine? *Nat. Rev. Microbiol*. 2011;9(8):617-626. doi 10.1038/nrmicro2614

Sonnleitner S.T., Prelog M., Sonnleitner S., Hinterbichler E., Halbfurter H., Kopecky D.B.C., Almanzar G., Koblmüller S., Sturmbauer C., Feist L., Horres R., Posch W., Walde G. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun*. 2022;13(1):2560. doi 10.1038/s41467-022-30163-4

Temmam S., Vongphayloth K., Baquero E., Munier S., Bonomi M., Regnault B., Douangboubpha B., Karami Y., Chrétien D., Sanamxay D., Xayaphet V., Paphaphanh P., Lacoste V., Somlor S., Lakeomany K., Phommavanh N., Pérot P., Dehan O., Amara F., Donati F., Bigot T., Nilges M., Rey F.A., van der Werf S., Brey P.T., Eloit M. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature*. 2022;604(7905):330-336. doi 10.1038/s41586-022-04532-4

Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.-G., Hu Y., Tao Z.-W., Tian J.-H., Pei Y.-Y., Yuan M.-L., Zhang Y.-L., Dai F.-H., Liu Y., Wang Q.-M., Zheng J.-J., Xu L., Holmes E.C., Zhang Y.-Z. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. doi 10.1038/s41586-020-2008-3

Zhou P., Yang X.L., Wang X.G., Hu B., Zhang L., Zhang W., Si H.-R., Zhu Y., Li B., Huang C.-L., Chen H.-D., Chen J., Luo Y., Guo H., Jiang R.-D., Liu M.-Q., Chen Y., Shen X.-R., Wang X., Zheng X.-S., Zhao K., Chen Q.-J., Deng F., Liu L.-L., Shi Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. doi 10.1038/s41586-020-2012-7

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

Received September 15, 2024. Revised October 23, 2024. Accepted October 24, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):854-863

doi 10.18699/vjgb-24-93

Search for and functional annotation of multi-domain PLA2 family proteins in flatworms

M.E. Bocharnikova [1, 2, 3], I.I. Turnaev [1, 3], D.A. Afonnikov [1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
bocharnikova@bionet.nsc.ru

Abstract. The phospholipase A2 (PLA2) is a superfamily of hydrolases that catalyze the hydrolysis of phospholipids and play a key role in many molecular processes in the cells and the organism as a whole. This family consists of 16 groups divided into six main types. PLA2 were first isolated from venom toxins and porcine pancreatic juice. The study of these enzymes is currently of great interest, since it has been shown that a number of PLA2 are involved in the processes of carcinogenesis. PLA2 enzymes were characterized in detail in model organisms and humans. However, their presence and functional role in non-model organisms is poorly understood. Such poorly studied taxa include flatworms, a number of species of which are human parasites. Several PLA2 genes have previously been characterized in parasitic flatworms and their possible role in parasite-host interaction has been shown. However, no systematic identification of the PLA2 genes in this taxon has been carried out. The paper provides a search for and a comparative analysis of PLA2 sequences encoded in the genomes of flatworms. 44 species represented by two free-living and 42 parasitic organisms were studied. The analysis was based on identification of orthologous groups of protein-coding genes, taking into account the domain structure of proteins. In flatworms, 12 of the 13 known types of animal A2 phospholipases were found, represented by 11 orthologous groups. Some phospholipases of several types fell into one orthologous group, some types split into several orthogroups in accordance with their domain structure. It has been shown that phospholipases A2 of the calcium-independent type, platelet-activating phospholipases from group G8 and lysosomal phospholipases from group G15 are represented in all large taxa of flatworms and the vast majority of the species studied by us. In free-living flatworms PLA2 genes have multiple copies. In parasitic flatworms, on the contrary, loss of genes occur specifically in individual taxa specifically for groups or subfamilies of PLAs. An orthologous group of secreted phospholipases has been identified, which is represented only in Digenea and this family has undergone duplications in the genomes of opisthorchids. Interestingly, a number of experimental studies have previously shown the effect of *Clonorchis sinensis* proteins of this orthogroup on the cancer transformation of host cells. Our results made it possible for the first time to systematically identify PLA2 sequences in flatworms, and demonstrated that their evolution is subject to gene loss processes characteristic of parasite genomes in general. In addition, our analysis allowed us to identify taxon-specific processes of duplication and loss of PLA2 genes in parasitic organisms, which may be associated with the processes of their interaction with the host organism.

Key words: phospholipase A2; flatworms; multi-domain proteins; parasitism; phylogeny; domain structure.

For citation: Bocharnikova M.E., Turnaev I.I., Afonnikov D.A. Search for and functional annotation of multi-domain PLA2 family proteins in flatworms. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):854-863. doi 10.18699/vjgb-24-93

Funding. The work was supported by budget project No. FWNR-2022-0020. The data analysis was performed using the computational resources of the "Bioinformatics" ICG SB RAS Joint Computational Center.

Поиск и функциональная аннотация многодоменных белков семейства ФА2 у плоских червей

М.Е. Бочарникова [1, 2, 3], И.И. Турнаев [1, 3], Д.А. Афонников [1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
bocharnikova@bionet.nsc.ru

Аннотация. Фосфолипазы типа А2 (ФА2) – это семейство гидролаз, которые катализируют процесс гидролиза фосфолипидов, играя ключевую роль во многих молекулярных процессах при функционировании клеток и организма в целом. Данное семейство подразделяется на 16 групп, объединенных в шесть основных типов.

Впервые ФА2 были выделены как цитотоксины яда у змей и ферменты панкреатического сока у свиней. Изучение этих ферментов в настоящее время вызывает большой интерес, поскольку было показано, что ряд ФА2 участвует в процессах канцерогенеза. Наиболее хорошо изучены ферменты ФА2 у модельных организмов и человека. Однако их наличие и функциональная роль у немодельных организмов изучены слабо. К таким малоизученным таксонам относятся плоские черви, ряд видов которых является паразитами человека. У паразитических плоских червей ранее было охарактеризовано несколько генов ФА2 и показана их возможная роль во взаимодействии «паразит–хозяин». Но систематической идентификации генов ФА2 у этого таксона не проведено. В работе осуществлены поиск и сравнительный анализ последовательностей ФА2, кодируемых в геномах плоских червей. Исследовано 44 вида, представленных 2 свободноживущими и 42 паразитическими организмами. Анализ выполнен на основе поиска ортологических групп белок-кодирующих генов с учетом доменной структуры белков. У плоских червей обнаружено 12 из 13 известных типов фосфолипаз А2, имеющихся в 11 ортологических группах. Часть фосфолипаз нескольких типов попала в одну ортологическую группу, часть типов распалась на несколько ортогрупп в соответствии с особенностями доменной структуры. Показано, что ФА2 кальций-независимого типа, ФА2 тромбоцитарно-активирующего типа групп G8 и лизосомальные ФА2 группы G15 представлены во всех крупных таксонах плоских червей и в большинстве изученных нами видов. Для генов, кодирующих ферменты у свободноживущих червей, наблюдается множественное число копий. У паразитических плоских червей, наоборот, происходит потеря основной части генов специфически по отношению как к отдельным таксонам, так и к отдельным группам/подсемействам ФА2. Обнаружена ортологическая группа секретируемых фосфолипаз, которая среди паразитов имеется только у дигенетических сосальщиков, при этом в геномах описторхид это семейство подверглось дупликациям. Интересно, что ранее в ряде экспериментальных работ показано влияние белков *Clonorchis sinensis* этой ортогруппы на раковую трансформацию клеток организма-хозяина. Наши результаты дали возможность впервые систематически идентифицировать последовательности ФА2 у плоских червей и продемонстрировали, что их эволюция подвержена процессам потерь генов, характерных в целом для геномов паразитов. Кроме того, наш анализ позволил выявить таксон-специфические процессы дупликаций и потерь генов ФА2 у паразитических организмов, которые могут быть связаны с процессами их взаимодействия с организмом хозяина.

Ключевые слова: фосфолипаза А2; плоские черви; многодоменные белки; паразитизм; филогения; структура доменов.

Introduction

The protein family of phospholipases A2 (PLA2) is a group of hydrolases that catalyze the hydrolysis of phospholipids, playing a key role in the functioning of cells and the organism as a whole (Filkin et al., 2020; Murakami et al., 2020). Phospholipases A2 are known to be the main components of venom toxins in snakes (Bitar et al., 2021), insects (Bitar et al., 2021), predatory invertebrates, for example, arachnids (Salabi, Jafari, 2024) or mollusks (McIntosh et al., 1995). Phospholipases A2 from snake venom hydrolyze phospholipids of cell membranes, which leads to cell destruction, release of arachidonic acid and activation of inflammatory processes. Their effects can also lead to more serious pathogenic effects, including damage to the nervous system (Bitar et al., 2021), which demonstrates the multiplicity of their functions (Gutiérrez, Lomonte, 2013).

The PLA2 family is divided into 16 groups (Dennis et al., 2011), united into six main types: secreted, cytosolic, calcium-independent, platelet-activating factors, lysosomal and adipospecific (Murakami et al., 2020). The main molecular functions of PLA2 include lipid cleavage, fatty acid remodeling, and interaction with phospholipids of lysosomes and adipose tissue (Mouchlis, Dennis, 2022). In animals, these enzymes are involved in a large number of important processes related to antibacterial, antiviral, immune and anti-inflammatory activities (Dennis et al., 2011).

The antiparasitic properties of phospholipases A2 are also known (Teixeira et al., 2022). Currently, these proteins are of great interest due to the fact that the impairment of lipid metabolism regulated by PLA2 often leads to various diseases, including carcinogenesis (Turnaev et al., 2022). Secreted PLA2 have increased expression in malignant tumors of organs such as the stomach (Scott et al., 2010), lungs (Park et al., 2012), intestines (Murase et al., 2017) and liver (Shang et al., 2017).

PLA2 are ancient genes and are found in all taxa of living organisms – bacteria, protists, archaea, animals, fungi and plants (Nevalainen et al., 2012). Their evolutionary analysis allows to consider in more detail the functional features of these proteins, to clarify their role in the most important biological processes (Murakami et al., 2020; Turnaev et al., 2022). PLA2 enzymes have been most well studied in model organisms and humans. However, their presence and functional role in non-model organisms have been poorly studied. Such poorly studied taxa include flatworms, a number of species of which are human parasites.

Flatworms (Platyhelminthes) are one of the oldest groups of multicellular animals. Their origin goes back to the early stages of the evolution of multicellular organisms. Studies by B. Egger et al. (2015) show that flatworms appeared more than 500 million years ago, during the Cambrian period, making them one of the first animals with an organized tissue structure. Along with mollusks (Mollusca) and annelids, they belong to a broader group, Lophotrochozoa (Egger et al., 2015; Laumer et al., 2015). At the same time, flatworms are often considered as a sister group to mollusks (Laumer et al.,

2015), which emphasizes their close evolutionary relationship. The importance of studying the biology of flatworms is due to the fact that most of their species are parasites – the main agents of helminthic diseases transmitted through infected fish, affecting a significant number of people[1]. Numerous studies have shown that long-term infections such as opisthorchiasis, schistosomiasis and similar helminthiasis can lead to serious consequences for the health of the host organism (Carbonell et al., 2021; Ogorodova et al., 2015; Pakharukova et al., 2019a), including the development of cancer (Pakharukova et al., 2019b; Mordvinov et al., 2021).

In parasitic flatworms, phospholipase A2 is widely present in excretory secretory products (ESP), which are secreted to affect the host (Wang et al., 2014), indicating the potential pathogenic effects of these enzymes on the host body. For example, a number of studies have experimentally shown that phospholipases A, C, and D of the parasitic flatworm *Clonorchis sinensis* are associated with fibrosis in the host (Hu et al., 2009). It has also been shown that phospholipases A2 of group 3 of *C. sinensis* are involved in the processes of carcinogenesis in host cells (Shang et al., 2017). However, currently there is only scattered information about phospholipases A2 in flatworms and their representation in genomes. Their functions in parasites are poorly described. This highlights the need for a deeper analysis and annotation of the functions of phospholipases A2 in flatworms, including parasitic worms, in order to better understand their role in pathogenesis and develop effective methods to combat helminthic infections.

In this work the structure, functions and evolution of phospholipases A2 in flatworms were studied. Identification of the phospholipase A2 protein sequences in flatworms was performed, and they were divided into orthogroups. Phylogenetic analysis of sequences from the orthogroups was carried out. Domain strictures and putative functions of PLA2 enzymes were analyzed.

Materials and methods

The OrthoDom computational pipeline for the identification of orthologous groups of proteins taking into account the domain structure. To identify PLA2 orthologous groups in flatworms, taking into account the domain structure, we used information on reference sequences of well-annotated PLA2 in model animals and the OrthoDom computational pipeline. The scheme is shown in Figure 1.

The OrthoDom pipeline allows to search for sequences of families of multidomain proteins in protein sequences encoded in the genomes of organisms under study based on orthology and domain analysis. As input data (marker 1 in Figure 1), sequences of the family of multidomain proteins with high-quality annotation (as a rule, identified and annotated in model organisms) are used (reference sequences). For reference sequences, lists of functional domains that they include are specified. For these domains, the corresponding HMM profiles (marker 2) are extracted from the Pfam 33.1 database (Mistry et al., 2021). Further, using the hmmsearch program of the HMMer 3.3.2 package (Eddy, 2011), validation of reference proteins is carried out for the presence of these domains (marker 3), since for some of them domains may be fragmented or absent.

Another set of input data is the amino acid sequences (proteomes) of the studied organisms (usually non-model ones), in which it is required to determine the orthologs of the reference proteins (marker 4). Orthologous groups for amino acid sequences of reference proteins and proteins of the studied organisms were determined by the OrthoFinder v. 2.5.4 program (Emms, Kelly, 2019). The orthologous groups of interest are identified (marker 5) by the presence of reference sequences. Sequences were additionally checked for the presence of specified domains. The sequences of orthologs of reference proteins identified in this way (marker 6) were further processed for phylogeny reconstruction by the IQ-TREE program (Nguyen et al., 2015). Phylogenetic trees were visualized using the web version of the iTOL program (Letunic, Bork, 2024).

Reference sequences of phospholipases A2 and their functional domains. To identify phospholipases A2 in flatworms, we used well annotated sequences of vertebrate phospholipases classified by type in a number of previous works. These proteins were considered as reference and were used to determine the type of phospholipases in orthologous groups of flatworm proteins. The sample of reference proteins is based on the PLA2 sequences from the work (Huang et al., 2015) (9 types of phospholipases in humans and some vertebrates). These sequences were supplemented with sequences from the NCBI database identified on the basis of homology using BLASTP (Turnaev et al., 2022). According to the classification of phospholipases A2 proposed by M. Murakami et al. (2020), out of a total of 16 groups of phospholipases of living organisms, the reference sample included phospholipases of 13 groups, since groups of phospholipases A2 11, 13 and 14 are present only in plants (Murakami et al., 2020). As a result, the reference sample of phospholipases A2 included 13 groups of PLA2 from 15 vertebrate taxa. The list of reference sequences from the articles by I.I. Turnaev et al. (2022) and Q. Huang et al. (2015), the type of phospholipase, the species name of the organism, and the identifier used in this work are given in Supplementary Material 1[2]. The list of key domains of these proteins and their HMM models is given in Supplementary Material 2.

Since phospholipases A2 contain not only PLA2 domains, but also other characteristic domains (Dennis et al., 2011), they were also identified after the identification of orthologs. The list of protein domains considered is provided in Supplementary Material 3.

Sources of genomic data. We studied the sequences of protein-coding genes from the genomes of flatworms of 44 species represented by two free-living and 42 parasitic organisms. The amino acid sequences encoded by mRNAs of the corresponding genes presented in the Wormbase

[1] On the state of sanitary and epidemiological welfare of the population in the Russian Federation in 2014: a state report. Moscow: Rospotrebnadzor, 2015, vol. 206.

[2] Supplementary Materials 1–6 are available at: http://vavilov.elpub.ru/jour/manager/files/Suppl_Bocharnikova_Engl_28_8.pdf

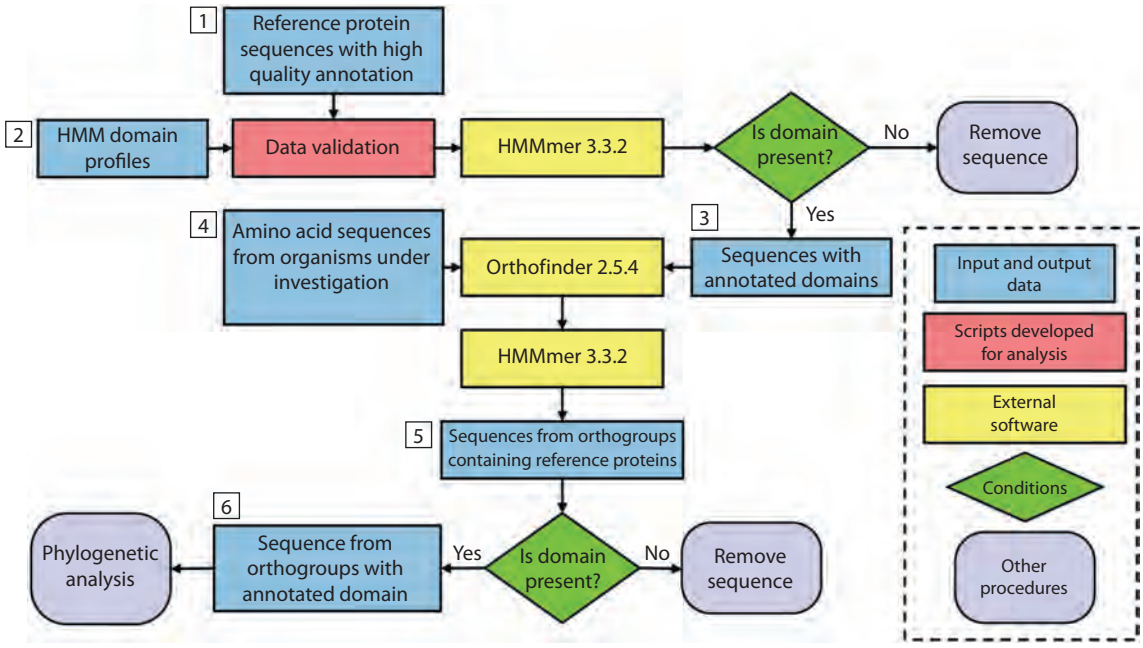


Fig. 1. Block diagram of the OrthoDom computing pipeline.
Block designations are shown in a dotted rectangle on the right.

Parasite 18.0 database (Howe et al., 2017) were analyzed. These species include the main taxa of flatworms: the class of digenetic flukes (Digenea), the class of tapeworms (Cestoda), the class of monogenetic flukes (Monogenea) and the class of ciliated worms (Turbellaria) (Brusa et al., 2020). Among the listed classes, the latter is a class of free-living worms, representatives of all other classes are obligate parasites, and monogenetic flukes are entoparasites, and digenetic flukes and cestodes are endoparasites. As an external group in the analysis, we used mollusk sequences from the genomes of the Pacific oyster (*Crassostrea gigas*), the sea saucer (*Lottia gigantea*) and the Philippine mussel (*Modiolus philippinarum*), since it is known that mollusks are a sister group to flatworms (Bernhard et al., 2015; Laumer et al., 2015). The amino acid sequences of mollusks were taken from the MolluskDB 2.0 database (Caurcel et al., 2021). The genome identifiers of flatworms and mollusks, species names of organisms and their types, and lifestyle are presented in Supplementary Material 4.

Statistical processing of the results. To assess the presence of phospholipases of various orthologous groups in flatworms, for large taxa (Digenea, Cestoda, Monogenea and Turbellaria), we estimated the average number of phospholipase sequences for the orthogroup in the genome (n) and the standard deviation (σ). The average number n of sequences in each orthogroup by taxa shows how common phospholipase sequences are in the studied organisms. The standard deviations σ show a variation in the values of the number of sequences around the average. The greater the standard deviation, the greater the diversity in the number of sequences across taxa. Additionally, we evaluated the parameter f (representation, %), the fraction of organisms in a large taxon that contain at least one of the orthogroup sequences. If it is equal to 100 %, then all organisms of the taxon contain at least one sequence from the orthogroup. If some organisms do not contain any sequence from the phospholipase orthogroup, then the f value is less than 100 %.

Results

As a result of the analysis carried out using the OrthoDom pipeline, 11 orthogroups were identified in flatworms, which contain reference sequences of phospholipases A2. Note that of all the groups of phospholipases A2, the sequences of which were used as a reference, only the sequences of group 9 did not show homology in the proteomes of mollusks and flatworms (they were not included in any of the orthogroups defined for these organisms). Thus, according to the classification of M. Murakami et al. (2020), 12 out of the 13 known groups of animal phospholipases A2 fell into the PLA2 orthogroups of mollusks and flatworms.

The Table shows the distribution of the identified orthogroups containing phospholipase sequences of flatworms and mollusks, and a number of statistical characteristics for them in terms of representation in the five main taxa. It can be seen that the correspondence between orthogroups and known types of phospholipases is non-exclusive. The Table shows that some orthogroups include several types of phospholipases. For example, the OG0003047 orthogroup contains sequences of phospholipase groups 1, 2, 5, 10. On the other hand, some types of phospholipases were represented by several orthogroups. For example, the sequences of phospholipase A2 group 6 split into orthogroups OG0000019,

Characteristics of the occurrence of phospholipase A2 orthogroup genes in mollusks and large flatworm taxa

Orthogroup ID	PLA2 type	PLA2 group	Mollusca (3)			Tricladida (2)			Monogenea (1)			Digenea (22)			Cestoda (19)		
			n	σ	*f*, %	*n*	σ	*f*, %	*n*	σ	*f*, %	*n*	σ	*f*, %	*n*	σ	*f*, %
OG0003047	Secreted	G1, 2, 5,10	2.0	1.0	100	5.0	2.8	100	1.0	0.0	100	0.0	0.0	0	0.0	0.0	0
OG0003722	Secreted	G3	4.7	3.8	100	3.0	1.4	100	0.0	0.0	0	1.0	1.9	41	0.0	0.0	0
OG0007610	Secreted	G12	1.0	0.0	100	0.5	0.7	50	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0
OG0000019*	Calcium-independent	G6	2.3	0.6	100	2.0	1.4	100	2.0	0.0	100	1.0	0.4	91	1.3	0.9	95
OG0000217*	Calcium-independent	G6	1.0	0.0	100	6.5	7.8	100	2.0	0.0	100	1.1	0.4	95	1.2	0.6	95
OG0000961*	Calcium-independent	G6	1.3	0.6	100	1.5	0.7	100	1.0	0.0	100	1.0	0.0	100	0.5	0.6	47
OG0007914	PAF	G7	1.0	0.0	100	4.5	6.4	50	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0
OG0004972*	PAF	G8	1.0	0.0	100	1.0	0.0	100	1.0	0.0	100	1.0	0.5	86	1.1	0.7	95
OG0000127	Cytosolic	G4	2.7	2.1	100	0.5	0.7	50	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0
OG0000135*	Lysosomal	G15	1.7	1.2	100	**8.5**	0.7	100	**3.0**	0.0	100	**4.7**	2.0	100	**2.1**	2.5	89
OG0007915	Adipo-specific	G16	3.0	1.0	100	1.0	1.4	50	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0

Note. The rows correspond to different orthogroups of phospholipases. The columns include: the orthogroup ID, the type and groups of phospholipases represented in it; statistics for the studied large taxa (average values (*n*), standard deviations (σ) of the number of sequences in orthogroups by taxa, the representation (*f*) of sequences in different species). The number of species is given in parentheses next to the name of the taxon. The maximum average values of the *n* number of sequences in orthogroups by taxa are shown in bold, the minimum values are underlined. The largest *n* values for ortogroups are highlighted by gray background. The complete table is presented in Supplementary Material 5.

*Orthogroups, the sequences of which are represented in all large taxa of flatworms.

OG0000217, OG0000961. In other cases, each orthogroup corresponded to one type and group of PLA2.

First, it should be noted that orthogroups differ in the number of sequences they are represented by. Thus, in the orthogroup OG0000135, which represents the only group of lysosomal phospholipases group G15, the average number of orthologs per proteome in each taxon of flatworms is the largest, compared with other orthogroups (from 2.1 in cestodes to 8.5 in triclads). Note that in mollusks, this group of phospholipases is not the largest one: the average number of sequences per proteome is 1.7. This taxon has the most numerous OG00003722 group: the average number of sequences is 4.7 (secreted PLA2 G3).

The Table also shows that a high average number of proteome sequences assigned to different phospholipases is characteristic of Tricladida, which are free-living, in contrast to the other taxa, which are parasitic. Only in the case of orthogroup OG00004972 (type PAF, group G8), the average number of sequences per proteome in free-living worms (1) is less than in cestodes (1.11), but this number is not less than in the other taxa.

The Table also demonstrates that the orthogroups we have identified are unevenly represented in various taxa. First, the PLA2 groups, which are found in all large taxa of flatworms. These are calcium-independent type PLA2, namely orthogroups OG0000019, OG0000217, OG0000961 (the sixth group of PLA2). At the same time, proteins of the first two orthogroups are represented by the vast majority of species from large taxa (more than 90 %). Orthogroup OG0000961 is characterized by the absence of orthologs for half of the cestode species. For one of these groups (OG0000217), the average number of proteins in free-living worms (6.5) is several times higher than that in parasitic worms (1.1–2). Proteins of this group in cestodes are represented in only half of the studied species (the average number of PLA2 per proteome is 0.5, the standard deviation is 0.6).

Another orthogroup, the representatives of which are found in all taxa of flatworms, is OG0004972 (the eighth group of platelet-activating type PLA2). In all major taxa, these proteins are present in more than 95 % of species, except for the digenetic flukes, in which this proportion is 87 %. These genes have 1–2 copies per proteome.

Another orthogroup represented in all major taxa is OG0000135, which includes lysosomal PLA2 of group G15. The sequences of this group are represented by more than one copy per proteome, and are characterized by the largest number of copies compared to others (see above).

Secondly, in the Table, orthogroups specific to free-living worms can be distinguished, the genes of which are completely absent in parasitic worms. These orthogroups were divided into four types: secreted, PAF, cytosolic and adipo-specific (OG0007610, OG0007914, OG0000127, OG0007915, respectively). Proteins in all these orthogroups are present in at least one of the two free-living species studied by us.

Thirdly, the Table demonstrates the presence of orthogroups specific to individual parasitic taxa. For example, orthogroup OG0003047 (phospholipases of groups G1, G2, G5, G10) is found only in Monogenea (in all species). Orthogroup OG0003722 is found only in Digenea (about half of the species). At the same time, cestodes have the smallest number of phospholipase orthogroups, in particular, all secreted phospholipases are missing.

Thus, the results allow us to conclude that most of the animal PLA2 groups (12 out of 13) are found in free-living worms, and most of them have a large number of copies. The number of genes in orthogroups and the number of orthogroups in parasitic worms is reduced in comparison with the free-living ones. Monogenea have one orthogroup including secreted proteins, all calcium-independent, one orthogroup including PAF, and one including lysosomal phospholipases A2. In Digenea, proteins from an orthogroup other than Monogenea and orthogroups including PAF and lysosomal phospholipases are present. All calcium-independent PLA2, PAF, and lysosomal phospholipases A2 are present in cestodes, but the secreted ones are completely absent. Various taxa of parasitic worms have phospholipases common to all of them, as well as specific ones.

The structural diversity of phospholipases

The domain organization for a number of phospholipases is shown in Figures 2 and 3. Figure 2 shows the domain structure of phospholipases from the OG0003047 orthogroup, which includes the reference proteins of the PLA2 groups G1, G2, G5 and G10.

Figure 2 shows that the sequences of secreted phospholipase A2 orthogroup OG0003047 have a length of approximately 200–250 amino acids. The phospholipase domain occupies more than 80 % of the total protein. Thus, the primary structure of secreted PLA2 in flatworms shows high similarity with human PLA2 structures of the corresponding types (Turnaev et al., 2022). Note that this orthogroup is represented only in free-living organisms.

The domain organization of the sequences of orthogroups OG0000019, OG0000217 and OG0000961 is shown in Figure 3. These are enzymes that belong to group 6. Despite the fact that group 6 PLA2 has been divided into three specified orthogroups, all of them contain a patatin domain key to this group (Fig. 3). The domain structure of the sequences of the OG0000019 orthogroup corresponds to the subgroup *A* typical for group 6 PLA2, which is characterized by a patatin domain and seven ankyrin domains. The composition of the domains of the OG0000217 orthogroup proteins corresponds to the typical group 6 subgroup *C* PLA2, which in addition to the patatin domain has three cNMP domains. The composition of the sequence domains of the OG0000961 orthogroup is similar to subgroups D and E typical for group 6 PLA2, which are characterized only by a patatin-like phospholipase domain located at the N-end of the sequence (Turnaev et al., 2022).

Thus, the analysis of the functional domains of phospholipases shows that proteins belonging to phospholipases of different types, but having a similar domain composition, form a common orthogroup, and sequences with different domain compositions of phospholipases of even the same type break down into different orthogroups.

Fig. 2. Domain structure of sequences of orthogroup OG0003047, phospholipase A2 of the secreted type.

The scale corresponding to 100 amino acids is shown on the right, the phospholipase domain is marked in red. The figure shows 10 sequences randomly selected among all the sequences of the OG0003047 orthogroup.

Phylogenetic analysis of flatworm phospholipases

For orthogroups, the domain structure of which is presented in Figures 2 and 3, we reconstructed phylogenetic trees.

Sequences of the OG0003047 orthogroup were found in free-living flatworms and one representative of Monogenea (Fig. 4). In species of free-living flatworms, the number of sequences belonging to this orthogroup is high (see the Table), in a representative of Monogenea species, *Protopolystoma xenopodis* (short designation ProXen), only one gene encoding a phospholipase of this type is observed.

Phylogenetic trees of orthogroups containing PLA2 of group 6 are presented in Supplementary Material 6. In the figures of calcium-independent PLA2 of group 6 (Supplementary Material 6, Fig. 1–3), similar patterns can be seen. It is worth noting that protein sequences of parasitic flatworms are highly conservative. In Figure 3, it can be seen that the domain structure of the sequences is similar among representatives of different parasitic taxa. This allows us to conclude that group 6 PLA2 is a conservative protein that plays a key role in the basic processes of life of parasitic flatworms.

Secreted phospholipases A2, which belong to orthogroup OG0003722, are worth noting. This orthogroup is characterized by the fact that in parasitic worms only the Digenea contains sequences of this orthogroup. The phylogenetic tree of sequences belonging to this orthogroup is shown in Figure 5.

Figure 5 shows that several copies of the PLA2 gene of this orthogroup are found in free-living worms. Digenetic flukes also have several copies of this gene, which are distributed in different clades. This suggests that duplications of the PLA2 gene of the G3 group are characteristic of these organisms. As a rule, the molecular evolution of parasites proceeds much faster compared to representatives of free-living organisms (Trouvé et al., 1998). The analysis of phylogenetic trees confirms this statement for phospholipase A2, where longer

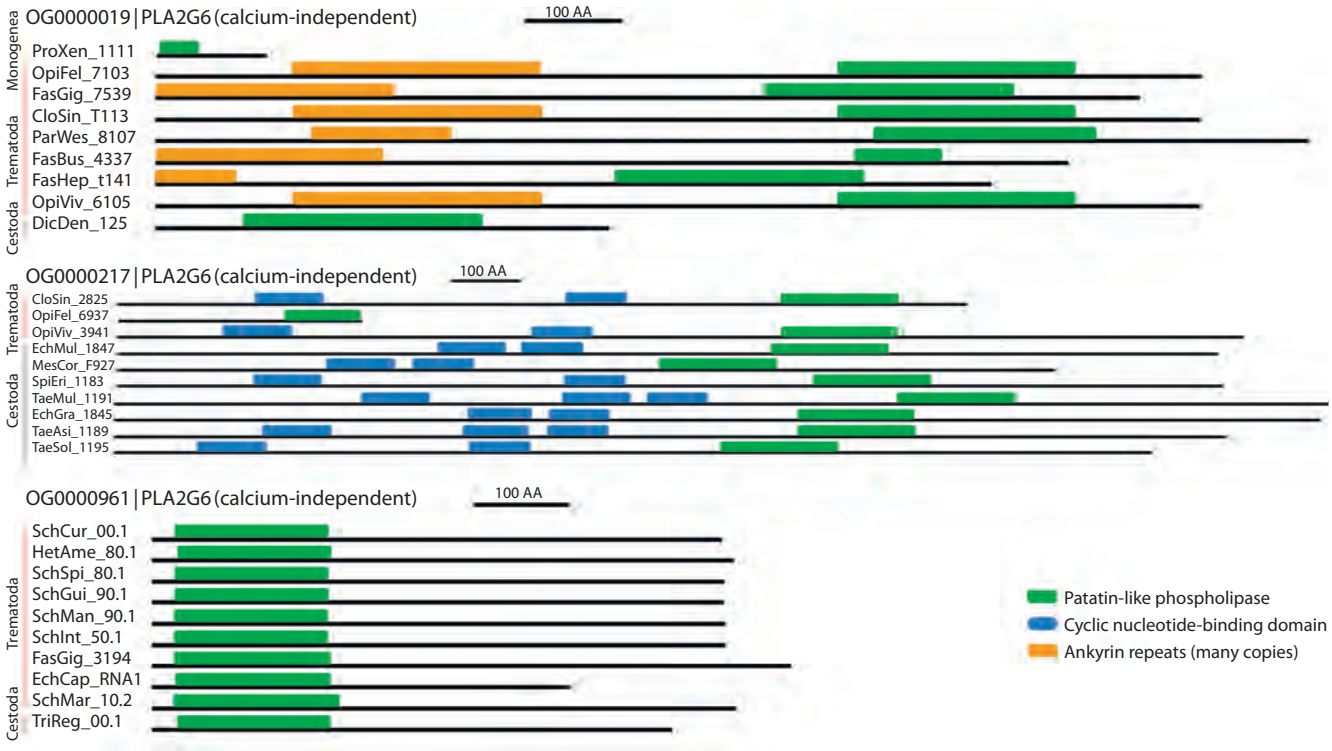


Fig. 3. Domain structure of sequences of orthogroups OG0000019, OG0000217, OG0000961, calcium-independent phospholipase A2.

The scale corresponding to 100 amino acids is shown on the right, the patatin-like phospholipase domain is marked in green, the cNMP domain is blue, and ankyrin repeats are orange. The figure shows 10 sequences (from 30 in total) randomly selected among all the sequences of orthogroups OG0000019, OG0000217, OG0000961.

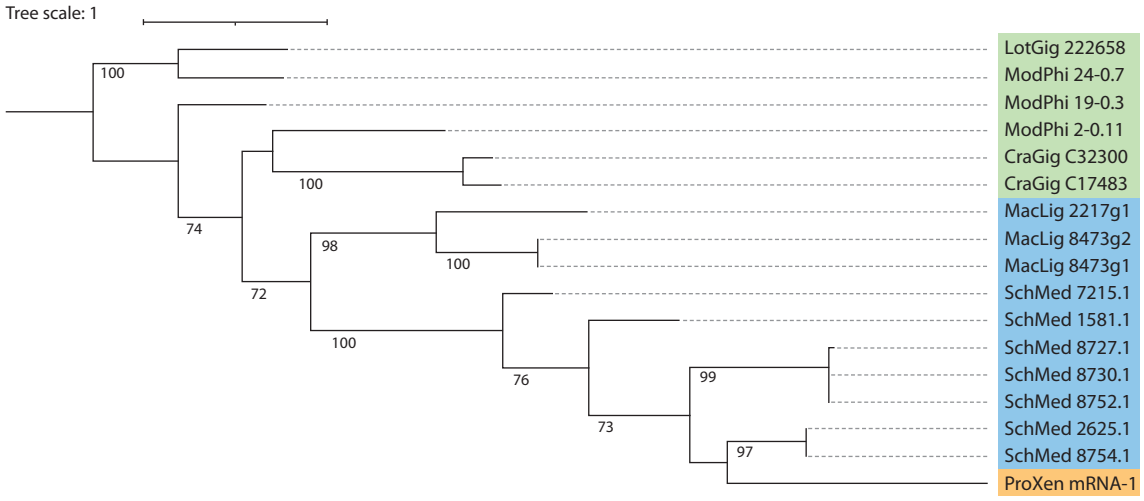


Fig. 4. Phylogenetic tree of phospholipase A2 orthogroup OG0003047 (PLA2G1, 2, 5, 10, secreted).

In the figure, the sequences of mollusks (Mollusca) are highlighted in green, free-living worms (Turbellaria), in blue, and monogenea (Monogenea), in orange.

branches are observed in parasites, which indicates a high rate of evolution of these molecules.

Discussion

Despite the fact that phospholipases of various types, PLA2 among them, are components of ESP of parasitic flatworms (Wang et al., 2014) and that an association with carcinogenesis in the host has been demonstrated for a number of them (Hu et al., 2009; Shang et al., 2017), they are still insufficiently studied for the Platyhelminthes taxon (Dennis et al., 2011). Here, almost all known groups of phospholipases in flatworms were identified. The OrthoDom pipeline allowed to split them into orthogroups, taking into account the domain structure. These results are consistent with the

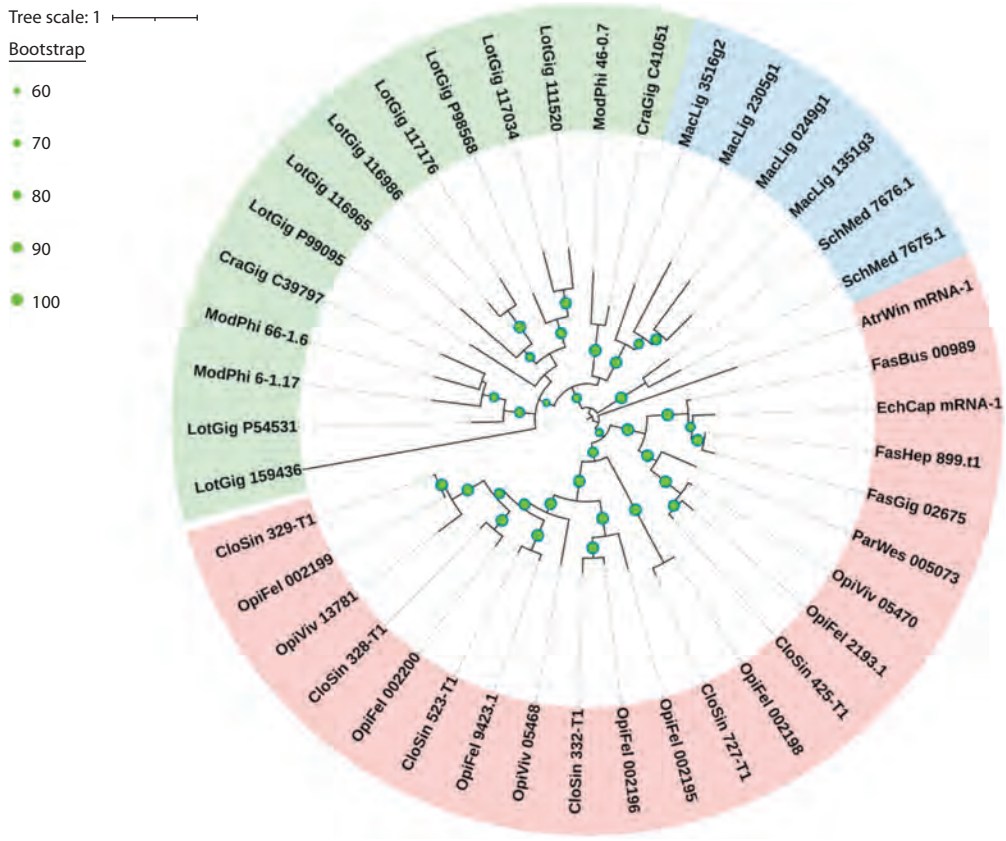


Fig. 5. Phylogenetic tree of sequences of phospholipases A2 of orthogroup OG0003722 (PLA2 G3, secreted).
In the Figure, mollusks (Mollusca) are highlighted in green, free-living flatworms (Turbellaria), in blue, and digenetic flukes (Digenea), in red.

classification of phospholipases A2 and their domain organization, presented in the works of E.A. Dennis et al. (2011). The method of identifying orthologs based on the domain structure has demonstrated its effectiveness in isolating orthogroups of proteins, taking into account the differences in the composition of their domains.

Our analysis made it possible to identify them and showed that in the evolution of A2 phospholipases in flatworms, peculiarities can be identified that are characteristic of the evolution of parasite genomes, for example, gene loss due to a parasitic lifestyle (Langleib et al., 2024). Indeed, our study demonstrated that some PLA2 groups are reduced in parasitic flatworms, and most genes are represented by a single copy. There are groups of phospholipases lost in some large taxa.

Comparative analysis of orthogroups of PLA2 genes shows that a relatively high degree of duplication is observed among PLA2 in free-living worms, with an average number of paralogs per species reaching five. This phenomenon implies the presence of significant adaptive capabilities, which may be due to a variety of environmental factors. Free-living organisms exposed to higher levels of environmental competition can use this diversity of PLA2 genes to increase viability and resistance to environmental changes. However, in orthogroup OG0000135 containing lysosomal type PLA2, genes are duplicated even in parasitic flatworms. What caused this anomaly remains to be seen.

A number of experimental studies have shown that some phospholipases A2 can participate in carcinogenesis, contributing to the activation of a number of cellular signaling pathways and interaction with the host body's immune system. For example, chronic infection caused by *C. sinensis* leads to liver fibrosis and cholangiocarcinoma (Shang et al., 2016). Moreover, *C. sinensis* uses group 3 phospholipases A2 as an ESP, which plays an important role in host kidney pathogenesis (Wu et al., 2021). As a result of the study, it was found that among the secreted phospholipases of digenetic flukes, only phospholipase A2 of group 3 is present, whereas in cestodes there are no secreted phospholipases A2. Given that parasitic flatworms are able to manipulate the metabolism of their hosts by using phospholipases to extract the necessary resources, it can be assumed that similar mechanisms may work in cancer cells.

Conclusion

Phospholipases A2 are a family of hydrolases that catalyze the hydrolysis of phospholipids, playing a key role in many molecular processes in the functioning of cells and the body as a whole. Their diversity in flatworms has been poorly studied, and in our work, we conducted such an analysis for

the first time. We found that 12 out of the 13 known types of phospholipases A2 are present in free-living worms. These organisms have an increased number of gene copies compared to parasitic worms. Unique features of some orthogroups have been identified, which may probably be associated with carcinogenesis in the host caused by a parasitic infection.

References

Bitar L., Jundi D., Rima M., Al Alam J., Sabatier J.M., Fajloun Z. Bee venom PLA2 *versus* snake venom PLA2: Evaluation of structural and functional properties. *Venoms Toxins.* 2021;2(1):22-33. doi 10.2174/2666121701999210101225032

Brusa F., Leal-Zanchet A.M., Noreña C., Damborenea C. Phylum Platyhelminthes. In: Thorp and Covich's Freshwater Invertebrates. Ch. 5. Academic Press, 2020;101-120. doi 10.1016/B978-0-12-804225-0.00005-8

Carbonell C., Rodríguez-Alonso B., López-Bernús A., Almeida H., Galindo-Pérez I., Velasco-Tirado V., Belhassen-García M. Clinical spectrum of schistosomiasis: an update. *J. Clin. Med.* 2021;10(23): 5521. doi 10.3390/jcm10235521

Caurcel C., Laetsch D.R., Challis R., Kumar S., Gharbi K., Blaxter M. MolluscDB: a genome and transcriptome database for molluscs. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 2021;376(1825):20200157. doi 10.1098/rstb.2020.0157

Dennis E.A., Cao J., Hsu Y.-H., Magrioti V., Kokotos G. Phospholipase A_2 enzymes: physical structure, biological function, disease implication, chemical inhibition, and therapeutic intervention. *Chem. Rev.* 2011;111(10):6130-6185. doi 10.1021/cr200085w

Eddy S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* 2011;7(10):e1002195. doi 10.1371/journal.pcbi.1002195

Egger B., Lapraz F., Tomiczek B., Müller S., Dessimoz C., Girstmair J., Telford M.J. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr. Biol.* 2015;25(10):1347-1353. doi 10.1016/j.cub.2015.03.034

Emms D.M., Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. doi 10.1186/s13059-019-1832-y

Filkin S.Yu., Lipkin A.V., Fedorov A.N. Phospholipase superfamily: structure, functions, and biotechnological applications. *Uspekhi Biologicheskoi Khimii = Biochemistry* (*Moscow*). 2020;85(Suppl.1): S177S195. DOI 10.1134/S0006297920140096

Gutiérrez J.M., Lomonte B. Phospholipases A_2: unveiling the secrets of a functionally versatile group of snake venom toxins. *Toxicon.* 2013; 62:27-39. doi 10.1016/j.toxicon.2012.09.006

Howe K.L., Bolt B.J., Shafie M., Kersey P., Berriman M. WormBase ParaSite − a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.* 2017;215:2-10. doi 10.1016/j.molbiopara.2016.11.005

Hu F., Hu X., Ma C., Zhao J., Xu J., Yu X. Molecular characterization of a novel *Clonorchis sinensis* secretory phospholipase A_2 and investigation of its potential contribution to hepatic fibrosis. *Mol. Biochem. Parasitol.* 2009;167(2):127-134. doi 10.1016/j.molbiopara.2009.05.003

Huang Q., Wu Y., Qin C., He W., Wei X. Phylogenetic and structural analysis of the phospholipase A_2 gene family in vertebrates. *Int. J. Mol. Med.* 2015;35(3):587-596. doi 10.3892/ijmm.2014.2047

Langleib M., Calvelo J., Costábile A., Castillo E., Tort J.F., Hoffmann F.G., Iriarte A. Evolutionary analysis of species-specific duplications in flatworm genomes. *Mol. Phylogenet. Evol.* 2024;199: 108141. doi 10.1016/j.ympev.2024.108141

Laumer C.E., Hejnol A., Giribet G. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife.* 2015;4:e05503. doi 10.7554/eLife.05503

Letunic I., Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 2024;52(W1):W78-W82. doi 10.1093/nar/gkae268

McIntosh J.M., Ghomashchi F., Gelb M.H., Dooley D.J., Stoehr S.J., Giordani A.B., Olivera B.M. Conodipine-M, a novel phospholipase A_2 isolated from the venom of the marine snail *Conus magus*. *J. Biol. Chem.* 1995;270(8):3518-3526. doi 10.1074/jbc.270.8.3518

Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L., Bateman A. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419. doi 10.1093/nar/gkaa913

Mordvinov V.A., Minkova G.A., Kovner A.V., Ponomarev D.V., Lvova M.N., Zaparina O., Pakharukova M.Y. A tumorigenic cell line derived from a hamster cholangiocarcinoma associated with *Opisthorchis felineus* liver fluke infection. *Life Sci.* 2021;277:119494. doi 10.1016/j.lfs.2021.119494

Mouchlis V.D., Dennis E.A. Membrane association allosterically regulates phospholipase A_2 enzymes and their specificity. *Acc. Chem. Res.* 2022;55(23):3303-3311. doi 10.1021/acs.accounts.2c00497

Murakami M., Sato H., Taketomi Y. Updating phospholipase A_2 biology. *Biomolecules.* 2020;10(10):1457. doi 10.3390/biom10101457

Murase R., Taketomi Y., Miki Y., Nishito Y., Saito M., Fukami K., Murakami M. Group III phospholipase A_2 promotes colitis and colorectal cancer. *Sci. Rep.* 2017;7(1):12261. doi 10.1038/s41598-017-12434-z

Nevalainen T.J., Cardoso J.C., Riikonen P.T. Conserved domains and evolution of secreted phospholipases A_2. *FEBS J.* 2012;279(4): 636-649. doi 10.1111/j.1742-4658.2011.08453.x

Nguyen L.T., Schmidt H.A., Von Haeseler A., Minh B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 2015;32(1):268-274. doi 10.1093/molbev/msu300

Ogorodova L.M., Fedorova O.S., Sripa B., Mordvinov V.A., Katokhin A.V., Keiser J.; TOPIC Consortium. Opisthorchiasis: an overlooked danger. *PLoS Negl. Trop. Dis.* 2015:9(4):e0003563. doi 10.1371/journal.pntd.0003563

Pakharukova M.Y., Zaparina O.G., Kapushchak Y.K., Baginskaya N.V., Mordvinov V.A. *Opisthorchis felineus* infection provokes time-dependent accumulation of oxidative hepatobiliary lesions in the injured hamster liver. *PLoS One.* 2019a;14(5):e0216757. doi 10.1371/journal.pone.0216757

Pakharukova M.Y., da Costa J.M.C., Mordvinov V.A. The liver fluke *Opisthorchis felineus* as a group III or group I carcinogen. *4open.* 2019b;2:23. doi 10.1051/fopen/2019016

Park J.B., Lee C.S., Jang J.H., Ghim J., Kim Y.J., You S., Ryu S.H. Phospholipase signalling networks in cancer. *Nat. Rev. Cancer.* 2012;12(11):782-792. doi 10.1038/nrc3379

Salabi F., Jafari H. Whole transcriptome sequencing reveals the activity of the PLA_2 family members in *Androctonus crassicauda* (Scorpionida: Buthidae) venom gland. *FASEB J.* 2024;38(10):e23658. doi 10.1096/fj.202400178RR

Scott K.F., Sajinovic M., Hein J., Nixdorf S., Galettis P., Liauw W., Russell P.J. Emerging roles for phospholipase A_2 enzymes in cancer. *Biochimie.* 2010;92(6):601-610. doi 10.1016/j.biochi.2010.03.019

Shang M., Xie Z., Tang Z., He L., Wang X., Wang C., Li X. Expression of *Clonorchis sinensis* GIIIsPLA$_2$ protein in baculovirus-infected insect cells and its overexpression facilitating epithelial-mesenchymal transition in Huh7 cells via AKT pathway. *Parasitol. Res.* 2017; 116:1307-1316. doi 10.1007/s00436-017-5409-y

Teixeira S.C., da Silva M.S., Gomes A.A.S., Moretti N.S., Lopes D.S., Ferro E.A.V., de Melo Rodrigues V. Panacea within a Pandora's box: the antiparasitic effects of phospholipases A_2 (PLA_2s) from snake venoms. *Trends Parasitol.* 2022;38(1):80-94. doi 10.1016/j.pt.2021.07.004

Trouvé S., Sasal P., Jourdane J., Renau F., Morand S. The evolution of life-history traits in parasitic and free-living platyhelminthes: a new perspective. *Oecologia*. 1998;115:370-378. doi 10.1007/s004420050530

Turnaev I.I., Bocharnikova M.E., Afonnikov D.A. Human phospholipases A2: a functional and evolutionary analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):787-797. doi 10.18699/VJGB-22-95

Wang X., Hu F., Hu X., Chen W., Huang Y., Yu X. Proteomic identification of potential *Clonorchis sinensis* excretory/secretory products capable of binding and activating human hepatic stellate cells. *Parasitol. Res.* 2014;113:3063-3071. doi 10.1007/s00436-014-3972-z

Wu Y.J., He Q., Shang M., Yin Y.X., Li Y., Du X., Li X.R. The NF-κB signalling pathway and TM7SF3 contribute to liver fibrosis caused by secreted phospholipase A2 of Clonorchis sinensis. *Parasit. Vectors*. 2021;14:1-9. doi 10.1186/s13071-021-04663-z

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

Received October 21, 2024. Revised November 25, 2024. Accepted November 26, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):864-873

doi 10.18699/vjgb-24-94

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

A.D. Mikhailova[1], S.A. Lashin [1, 2, 3], V.A. Ivanisenko [1, 2, 3], P.S. Demenkov [1, 2, 3], E.V. Ignatieva [1, 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
[3] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
eignat@bionet.nsc.ru

Abstract. Cholesterol is an essential structural component of cell membranes and a precursor of vitamin D, as well as steroid hormones. Humans and other animal species can absorb cholesterol from food. Cholesterol is also synthesized *de novo* in the cells of many tissues. We have previously reconstructed the gene network regulating intracellular cholesterol levels, which included regulatory circuits involving transcription factors from the SREBP (Sterol Regulatory Element-Binding Proteins) subfamily. The activity of SREBP transcription factors is regulated inversely depending on the intracellular cholesterol level. This mechanism is implemented with the participation of proteins SCAP, INSIG1, INSIG2, MBTPS1/S1P and MBTPS2/S2P. This group of proteins, together with the SREBP factors, is designated as "cholesterol sensor". An elevated cholesterol level is a risk factor for the development of cardiovascular diseases and may also be observed in obesity, diabetes and other pathological conditions. Systematization of information about the molecular mechanisms controlling the activity of SREBP factors and cholesterol biosynthesis in the form of a gene network and building new knowledge about the gene network as a single object is extremely important for understanding the molecular mechanisms underlying the predisposition to diseases. With a computer tool, ANDSystem, we have built a gene network regulating cholesterol biosynthesis. The gene network included data on: (1) the complete set of enzymes involved in cholesterol biosynthesis; (2) proteins that function as part of the "cholesterol sensor"; (3) proteins that regulate the activity of the "cholesterol sensor"; (4) genes encoding proteins of these groups; (5) genes whose transcription is regulated by SREBP factors (SREBP target genes). The gene network was analyzed and feedback loops that control the activity of SREBP factors were identified. These feedback loops involved the *PPARG*, *NR0B2*/*SHP1*, *LPIN1*, and *AR* genes and the proteins they encode. Analysis of the phylostratigraphic age of the genes showed that the ancestral forms of most human genes encoding the enzymes of cholesterol biosynthesis and the proteins of the "cholesterol sensor" may have arisen at early evolutionary stages (*Cellular organisms* (the root of the phylostratigraphic tree) and the stages of *Eukaryota* and *Metazoa* divergence). However, the mechanism of gene transcription regulation in response to changes in cholesterol levels may only have formed at later evolutionary stages, since the phylostratigraphic age of the genes encoding the transcription factors SREBP1 and SREBP2 corresponds to the stage of *Vertebrata* divergence.

Key words: cholesterol biosynthesis; transcription factors; SREBP; gene networks; feedback loops; evolution; phylostratigraphy; gene age.

For citation: 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 Journal of Genetics and Breeding*. 2024;28(8):864-873. doi 10.18699/vjgb-24-94

Funding. The work was supported by the publicly funded project № FWNR-2022-0020 of the Federal Research Center ICG SB RAS.

Acknowledgements. The authors express their gratitude to Z.S. Mustafin, who provided data on PAI values for human protein-coding genes.

Реконструкция и компьютерный анализ структурно-функциональной организации генной сети регуляции биосинтеза холестерина у человека и эволюционная характеристика участвующих в ней генов

А.Д. Михайлова[1], С.А. Лашин[1, 2, 3], В.А. Иванисенко[1, 2, 3], П.С. Деменков[1, 2, 3], Е.В. Игнатьева[1, 2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

eignat@bionet.nsc.ru

Аннотация. Холестерин – это незаменимая структурная компонента клеточных мембран, предшественник витамина D и стероидных гормонов. У человека и других видов животных холестерин поступает в организм с пищей, а также синтезируется в клетках многих тканей *de novo*. Ранее нами была реконструирована генная сеть регуляции внутриклеточного уровня холестерина, включавшая регуляторные контуры, функционирующие при участии транскрипционных факторов подсемейства SREBP (sterol regulatory element-binding proteins). Активность транскрипционных факторов подсемейства SREBP регулируется в обратной зависимости от уровня холестерина в клетке. Этот механизм реализуется при участии белков «холестеринового сенсора», включающего белки SCAP, INSIG1, INSIG2, MBTPS1/S1P, MBTPS2/S2P и транскрипционные факторы подсемейства SREBP. Повышенный уровень холестерина является фактором риска сердечно-сосудистых заболеваний, а также сопутствующим фактором многих патологических состояний. Систематизация сведений о молекулярных механизмах, контролирующих активность факторов подсемейства SREBP и биосинтез холестерина, в формате генной сети и получение новых знаний о генной сети как едином объекте чрезвычайно важны в контексте понимания молекулярных механизмов развития заболеваний. Средствами компьютерной системы ANDSystem нами построена генная сеть регуляции биосинтеза холестерина в клетке. Генная сеть включает данные: (1) о ферментах, осуществляющих биосинтез холестерина; (2) белках, функционирующих в составе «холестеринового сенсора»; (3) белках, регулирующих активность белков «холестеринового сенсора»; (4) генах, кодирующих белки этих групп; (5) генах, транскрипция которых регулируется при участии транскрипционных факторов подсемейства SREBP (генах-мишенях). Проведен анализ генной сети и выявлены замкнутые регуляторные контуры, контролирующие активность транскрипционных факторов подсемейства SREBP. Эти контуры реализуются с участием генов *PPARG*, *NR0B2/SHP1*, *LPIN1*, *AR* и кодируемых ими белков. Исследование филостратиграфического возраста генов показало, что предковые формы большинства генов человека, кодирующих ферменты биосинтеза холестерина и белки «холестеринового сенсора», могли возникнуть на достаточно ранних эволюционных этапах (*Cellular organisms* (корень филостратиграфического дерева) и этапах дивергенции *Eukaryota* и *Metazoa*). Однако механизм регуляции транскрипции генов в ответ на изменение уровня холестерина мог сформироваться только на более поздних эволюционных этапах, поскольку филостратиграфический возраст генов транскрипционных факторов подсемейства SREBP соответствует более позднему этапу эволюции (стадии дивергенции *Vertebrata*).

Ключевые слова: биосинтез холестерина; транскрипционные факторы; SREBP; генные сети; регуляторые обратные связи; эволюция; филостратиграфия; возраст гена.

Introduction

Cholesterol is an important substance in the animal body. It is present in all tissues as part of cell membranes, stabilizing the membrane structure (Koolman, Roehm, 2005). With an increase in cholesterol content, the membrane becomes more densely packed, contains fewer cavities, due to which its permeability to small molecules, including oxygen, decreases. This mechanism contributed to the adaptation of organisms to an oxygen-rich atmosphere, and, as a result, the protection of cells from oxidative stress (Zuniga-Hertz, Patel, 2019). It is noteworthy that cholesterol is not synthesized in fungi and plants, and the cell membrane of these organisms contains compounds similar in structure – ergosterol (in fungi) and β-sitosterol and stigmasterol (in plants) (Desmond, Gribaldo, 2009; Ferrer et al., 2017; Choy et al., 2023).

In animals, cholesterol has other important functions. This substance is a precursor of bile acids and steroid hormones (progesterone, estradiol, testosterone, calcitriol, cortisol) (Luo et al., 2020; Schade et al., 2020).

In humans and other animal species, cholesterol enters the body with food, and is also synthesized in the cells of many tissues *de novo* (Luo et al., 2020). The initial metabolites for cholesterol synthesis are acetyl-CoA and acetoacetyl-CoA, and more than 20 enzymes are involved in the biosynthesis process (Desmond, Gribaldo, 2009; Nes, 2011). Intermediate metabolites of the cholesterol biosynthesis pathway, such as geranylgeranyl pyrophosphate and farnesyl pyrophosphate, can also play an important role in animal cells. These metabolites are substrates in prenylation reactions. Prenylation is a common covalent post-translational modification of vari-

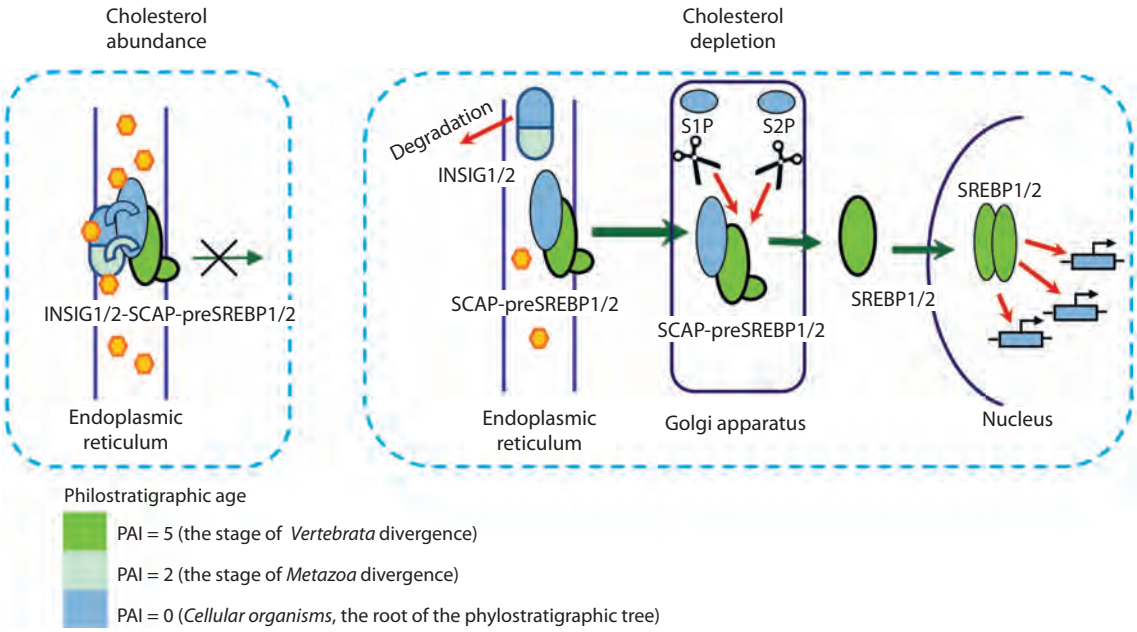


Fig. 1. The functioning of the "cholesterol sensor".

Yellow hexagons represent cholesterol molecules; INSIG1/2 – endoplasmic reticulum anchor proteins INSIG1 and INSIG2; SREBP1/2 – transcription factors SREBP1 and SREBP2; preSREBP1/2 – preSREBP1 and preSREBP2, which are inactive precursor proteins of SREBP1 and SREBP2; SCAP – SREBF chaperone protein interacting with preSREBP1 and preSREBP2; S1P and S2P proteins are proteases that are encoded by the *MBTPS1* and *MBTPS2* genes (respectively). The colors of the objects correspond to the phylostratigraphic age of the genes, which was estimated based on the PAI (the procedure for calculating PAI is described in the "Materials and methods" section). At high cholesterol levels (the left part of the Figure), cholesterol stabilizes the structure of INSIG1 and INSIG2 (designated as INSIG1/2), increasing its affinity for SCAP. The anchor proteins INSIG1 and INSIG2 help the SCAP-preSREBP1/2 complex to be preserved on the ER membrane. In cholesterol-deprived cells (the right part of the Figure), the reduction of sterol leads to ubiquitination and rapid degradation of INSIG1/2. The binding of SCAP to INSIG1/2 is destabilized. This gives the SCAP-preSREBP1/2 complex an opportunity to escape ER. The SCAP-preSREBP1/2 complex is transported to the Golgi apparatus, where the preSREBP1/2 proteins are cleaved by the S1P and S2P proteases. As a result of cleavage of the preSREBP1 and preSREBP2 proteins, active transcription factors SREBP1 and SREBP2 (designated as SREBP1/2) are formed. The description of the scheme is based on publications (DeBose-Boyd, Ye, 2018; Jiang et al., 2020).

ous proteins. Proteins that undergo prenylation include, for example, Ras and small GTP-binding proteins (GTPases). Such post-translational prenylation is important for the proper localization and activation of proteins (Waller et al., 2019).

Earlier, a gene network regulating intracellular cholesterol level was built, and four feedback loops involving transcription factors from the sterol regulatory element-binding protein subfamily (SREBP1 and SREBP2) were identified (Kolchanov et al., 2013; Merkulova et al., 2013). In the cells of animal organisms, there is a mechanism regulating the activity of transcription factors from the SREBP subfamily depending on cholesterol level (DeBose-Boyd, Ye, 2018; Jiang et al., 2020). This mechanism involves a number of proteins, which, in combination with transcription factors from the SREBP subfamily, will be further referred to as the "cholesterol sensor". A diagram showing how the "cholesterol sensor" functions is given in Figure 1.

The fuctioning of SREBPs can also be regulated in response to external signals affecting the cell, for example, insulin and growth factors (Sundqvist et al., 2005; Arito et al., 2008; Peterson et al., 2011). Due to regulation of this kind, fine-tuning of the SREBPs activity is carried out depending on the state of the cell and the organism as a whole. In turn, SREBPs control the expression of proteins involved in the regulation of a large number of cellular functions, integrating local gene networks that control various biological processes (Jeon, Osborne, 2012).

Elevated cholesterol levels are a risk factor for the development of cardiovascular diseases (atherosclerosis, coronary heart disease) (VargasAlarcon et al., 2019; Macvanin et al., 2024), and can also act as a concomitant factor in obesity (Kim et al., 2010), diabetes (Zhang F. et al., 2018), non-alcoholic fatty liver disease, non-alcoholic steatohepatitis (Li et al., 2023), hepatocarcinoma (Paul et al., 2022), tumor processes (Jiang et al., 2020) and inflammation (Shimano, Sato, 2017). Obtaining new knowledge about the gene network regulating cholesterol biosynthesis, as a single object, is extremely important in the context of understanding the connection of this system with diseases.

The aim of this study is to systematize data on the molecular mechanisms controlling the activity of transcription factors of SREBP subfamily and mechanisms controlling cholesterol biosynthesis using the format of a gene network and subsequent analysis of the structural and functional organization of the network and analysis of the evolutionary characteristics of the genes involved in it.

Materials and methods

Lists of genes used for building the gene network. The list comprising 24 human genes encoding enzymes of cholesterol biosynthesis (Supplementary Material 1)[1] was compiled based on data from WikiPathways (Agrawal et al., 2024).

The list, which included seven genes encoding proteins of the "cholesterol sensor" (Supplementary Material 2) was formed based on the description of the mechanism regulating activity of SREBP1 and SREBP2 according to data given in publications (DeBoseBoyd, Ye, 2018; Jiang et al., 2020).

The list containing 31 human genes, the transcription of which is regulated by factors of the SREBP subfamily (SREBP1 or SREBP2 target genes), was formed based on data from TRRD (Kolchanov et al., 2002) and TRRUST (https://www.grnpedia.org/trrust/) (Han et al., 2018). The final version of the list of SREBP target genes (Supplementary Material 3) included genes for which data on associations with SREBP1 or SREBP2 were found in ANDSystem (Ivanisenko et al., 2019).

The list of genes encoding proteins regulating the activity of proteins and genes of the "cholesterol sensor" ("regulatory proteins") (Supplementary Material 4) was formed using ANDSystem (Ivanisenko et al., 2019). "Regulatory proteins" were found using ANDVisio (ANDSystem software component) with the help of the built-in Pathway wizard tool. The associations between the "regulatory proteins" and proteins or genes of the "cholesterol sensor" obtained in this way were verified manually.

Building the gene network regulating cholesterol biosynthesis. The construction of the gene network was carried out using ANDSystem (Ivanisenko et al., 2019). In the first step, we built gene networks that included small groups of genes (hereinafter referred to as "small gene networks"). The procedures for building "small networks" are described in Supplementary Material 5. The number of objects in the networks is given in Supplementary Material 6. These "small networks" were then merged together in the ANDVisio tool applying the "Union of graphs" command. We merged "small networks" that included the following associations: (1) between the "regulatory proteins" and genes and proteins of the "cholesterol sensor"; (2) between SREBPs and target genes, and between target genes and the encoded proteins; (3) between proteins encoded by SREBP target genes, and genes and proteins of the "cholesterol sensor"; (4) between genes or proteins of the "cholesterol sensor" (with the exception of SREBPs) and the *SREBF1*, *SREBF2* genes and the encoded proteins; (5) between enzymes of cholesterol biosynthesis and cholesterol.

Search for feedback loops. The feedback loops that included 3, 4 or 5 objects, among which were factors SREBP1 and SREBP2, were found with the help of the ANDVisio built-in Pathway wizard tool. The search was performed based on the templates presented in Supplementary Material 7. According to the length of the template (which was equal to the number of objects involved in feedback loops), the number and types of intermediate objects were specified. The pathways found in this way were expanded by adding interactions between genes and the encoded proteins ("expression" type interactions), thus obtaining closed regulatory circuits.

Identification of tissues where the functioning of feedback loops may be observed. We used data from the GTEx project (GTEx Consortium, 2020) extracted from the Expression atlas (https://www.ebi.ac.uk/gxa/home). Examples of tissues or organs where the expression level of each gene involved in a particular feedback loop was at least 10 TPM were selected.

Analysis of the evolutionary characteristics of genes. The evolutionary characteristics of genes were evaluated using phylostratigraphic age index (PAI). PAI values were calculated for 19,556 human protein-coding genes using the Orthoscape software tool (Mustafin et al., 2017) as was described in (Mustafin et al., 2021).

Results and discussion

The gene network regulating cholesterol biosynthesis

At the first step, the so-called "small gene networks" were built using the ANDVisio program (as was described in "Materials and methods" and Supplementary Material 5). Next, the "small gene networks" were merged using the ANDVisio program. Thus, a gene network regulating cholesterol biosynthesis was constructed (Fig. 2). This network included: (1) the *SREBF1* and *SREBF2* genes and the proteins encoded by them; (2) five proteins regulating the activity of the SREBP1 and SREBP2 factors (INSIG1, INSIG2, SCAP, MBTPS1/S1P MBTPS2/S2P), and the genes encoding them ("cholesterol sensor"); (3) 62 proteins regulating the activity of genes and proteins of the "cholesterol sensor" ("regulatory proteins"); (4) 31 SREBP target genes (including *SREBF2* itself) and the proteins encoded by them; (5) 243 interactions between objects (Fig. 2).

Feedback loops involving transcription factors from the SREBP subfamily

Feedback loops involving transcription factors from the SREBP subfamily with length 2, 3, and 4. These feedback loops are shown in Figure 3. The factors from the SREBP subfamily are indicated in Figure 3 as SRBP1 and SRBP2. One of the three feedback loops shown in Figure 3 is positive and two feedbacks are negative.

SREBP2 (protein) → *SREBF2* (gene) → SREBP2 (protein).

The shortest feedback loop, which included two objects (Fig. 3*a*), was revealed when examining the list of SREBP target genes (Supplementary Material 3). According to R. Sato and co-authors, the promoter of the human *SREBF2* contains SREBP2 binding site (Sato et al., 1996), mediating positive autoregulation of *SREBF2* gene expression.

The search for feedback loops involving SREBPs was based on templates No. 1–4 presented in Supplementary Material 7. As a result, two feedbacks involving SREBP1 were found (Fig. 3*b*, *c*). No loops involving SREBP2 were found.

SREBP1 (protein) → *LPIN1* (gene) → LPIN1 (protein) → SREBP1 (protein) (Fig. 3*b*). This is a negative feedback loop involving the *LPIN1* gene (lipin 1) and the encoded

[1] Supplementary Materials 1–9 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Mikhailova_Engl_28_8.pdf

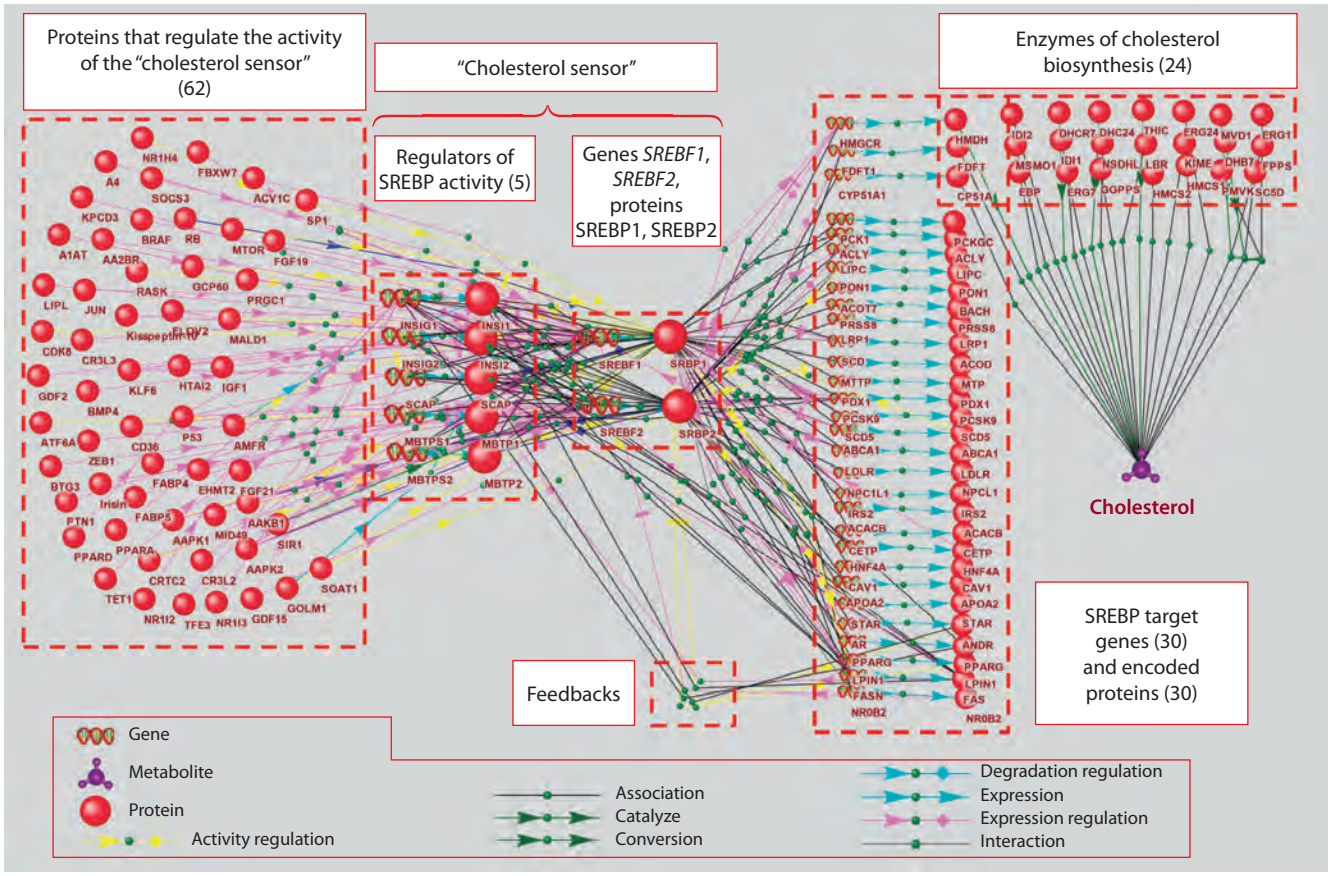


Fig. 2. The gene network regulating cholesterol biosynthesis, visualized by ANDVisio. The ANDVisio program designates SREBP1 and SREBP2 as SRBP1 and SRBP2.

Lists of genes from each functional group are presented in Supplementary Materials 1-4. Supplementary Material 3 contains one more target gene (i. e. 31 genes), in the Figure this 31st gene (*SREBF2*) is placed in the group of objects designated as the "cholesterol sensor".

protein. The promoter of the human *LPIN1* contains the sterol regulatory element, and this element is responsible for the transcription activation of *LPIN1*, mediated by SREBP1 (in the Figure it is indicated as SRBP1) (Ishimoto et al., 2009). The LPIN1 protein suppresses the activity of SREBP1, preventing SREBP1 from binding to regulatory regions of its target genes, including the *LPIN1* gene itself (Mateus et al., 2021). This mechanism is realized by regulating the SREBP1 transport inside the nucleus by the LPIN1 protein. LPIN1 promotes SREBP1 translocation to the nuclear lamina, where SREBP1 is inactivated (Peterson et al., 2011). The activity of LPIN1 is controlled by the mTOR kinase, which is involved in the response to growth factors (Peterson et al., 2011). Thus, the existence of a feedback loop involving LPIN1 indicates that the amplitude of transcriptional response to SREBP1 may be affected by growth factors.

SREBP1 (protein) → *NR0B2/SHP1* (gene) → NR0B2/SHP1 (protein) → *SREBF1* (gene) → SREBP1 (protein) (Fig. 3*c*). This feedback loop involves the *NR0B2/SHP1* gene and the encoded protein (SHP1, small heterodimer partner). The human *NR0B2/SHP1* gene transcription is activated by SREBP1 (in the Figure it is indicated as SRBP1) (Kim et al., 2004). According to the UniProt Knowledge base (UniProt_ID = NR0B2_HUMAN), SHP1 is a transcription corepressor, it interacts with a number of transcription factors, preventing their activation by ligands. Thus, ligand-dependent transcription factors LRH-1, LXR and RXR may activate *SREBF1* gene transcription, but the SHP1 protein prevents this activatory effect (Watanabe et al., 2004). Thus, the existence of a regulatory loop involving *NR0B2/SHP1* and the encoded protein indicates that the transcriptional response to decreased cholesterol levels may be affected by other low molecular weight hydrophobic substances, which are ligands of transcription factors LRH-1, LXR, RXR and corepressor NR0B2/SHP1.

Feedback loops with length 5 involving factors from the SREBP subfamily, as well as proteins functioning within the "cholesterol sensor". We identified three regulatory circuits involving proteins functioning within the "cholesterol sensor", which, in turn, affect the activity of SREBPs (Fig. 4). These feedback loops matched templates No. 7 and No. 8 presented in Supplementary Material 7. Two feedbacks included SREBP1 (indicated as SRBP1) (Fig. 4*a*, *c*) and one feedback loop included SREBP2 (indicated as SRBP2) (Fig. 4*b*). Two of the three regulatory loops are negative, and one is positive.

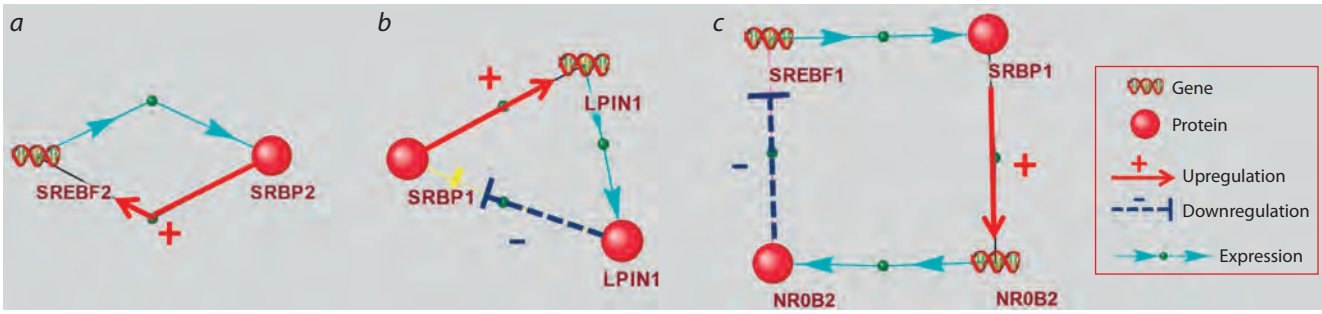


Fig. 3. Feedback loops involving factors from the SREBP subfamily (indicated as SRBP1 and SRBP2).

a – positive autoregulation of *SREBF2* gene expression; *b* – a feedback loop involving the *LPIN1* gene and the encoded protein; *c* – a feedback loop involving the *NR0B2/SHP1* gene and the encoded protein.

SREBP1 (protein) → *PPARG* (gene) → PPARG (protein) → *INSIG1* (gene) → INSIG1 (protein) → SREBP1 (protein) (Fig. 4*a*).

SREBP2 (protein) → *PPARG* (gene) → PPARG (protein) → *INSIG1* (gene) → INSIG1 (protein) → SREBP2 (protein) (Fig. 4*b*).

Two regulatory loops were found involving factors from the SREBP subfamily, as well as the *PPARG* and *INSIG1* genes and encoded proteins. SREBP1 and SREBP2 (in Figures 4*a* and *b* these proteins are designated as SRBP1 and SRBP2) can interact with binding sites in the human *PPARG* promoter increasing transcriptional activity of *PPARG* (Fajas et al., 1999). PPARG is a transcription factor that can interact with the binding site (PPRE1) in the human *INSIG1* promoter and activate transcription of the *INSIG1* gene (Kast-Woelbern et al., 2004). This leads to increased expression of the INSIG1 protein, which retains preSREBP1 and preSREBP2 on the membrane of the endoplasmic reticulum, thereby suppressing translocation of preSREBPs to the Golgi apparatus, where SREBPs are activated by proteolytic processing (Roth et al., 2008).

SREBP1 (protein) → *AR* (gene) → ANDR (protein) → *SCAP* (gene) → SCAP (protein) → SREBP1 (protein) (Fig. 4*c*).

The promoter region of the human *AR* gene encoding the androgen receptor (in the Figure this protein is designated as ANDR) contains SREBP1 binding site. SREBP1 (in Figure 4*c* this protein is designated as SRBP1) binds to this regulatory element and activates the transcription of *AR* (Huang et al., 2010). The ANDR protein binds to the androgen response element in intron 8 of the human *SCAP* gene. This interaction leads to increased expression of SCAP (Heemers et al., 2004). In turn, SCAP escorts preSREBPs from endoplasmic reticulum to the Golgi apparatus where the SREBPs are activated (Guo et al., 2019). Thus, this is a positive feedback loop.

An examination of gene expression data from the GTEx project (GTEx Consortium, 2020) showed that the regulatory loops we found (Fig. 3 and 4) can function in a wide range of tissues. Examples of such tissues are given in Supplementary Materials 8 and 9.

The phylostratigraphic age of genes encoding enzymes of cholesterol biosynthesis and proteins functioning within the "cholesterol sensor"

The phylostratigraphic age index (PAI) was used to estimate the phylostratigraphic age of the genes. The PAI value indicates the evolutionary stage corresponding to the divergence stage of certain taxa. The PAI index takes values from 1 to 15 (Mustafin et al., 2021). The greater the PAI value of the studied gene, the younger the gene is.

Genes encoding enzymes of cholesterol biosynthesis. Figure 5 shows distributions by PAI values for all human protein-coding genes (black columns, control group) and 24 genes encoding enzymes of the cholesterol biosynthesis pathway (green columns). PAI values for genes encoding enzymes of the cholesterol biosynthesis pathway are presented

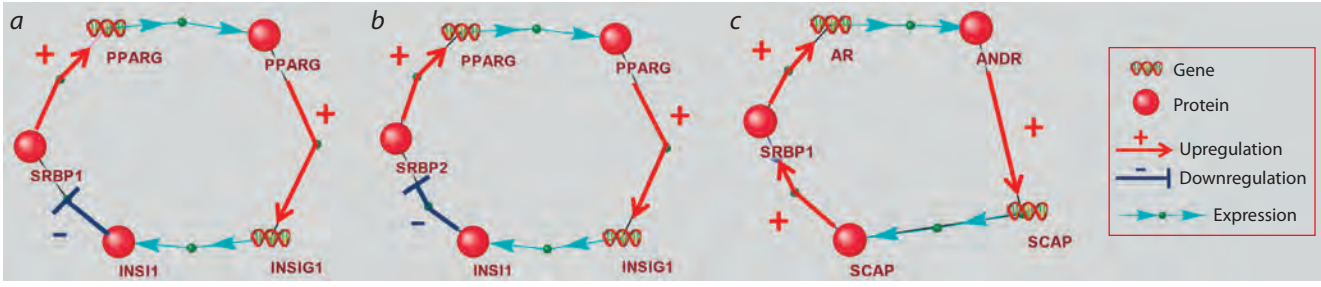
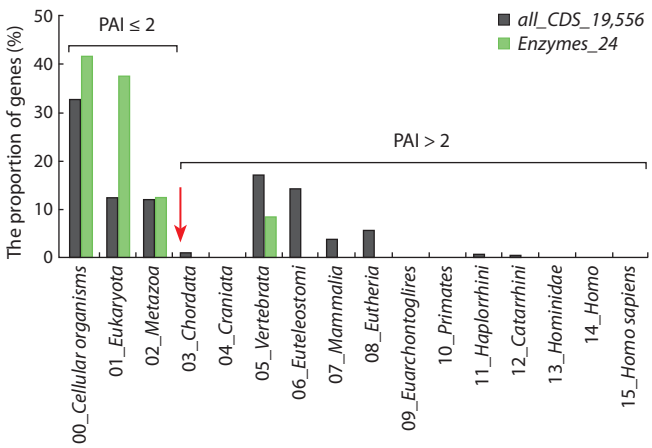


Fig. 4. Feedback loops involving factors from the SREBP subfamily (designated as SRBP1 and SRBP2) and other genes and proteins functioning within the "cholesterol sensor".

a – feedback involving the SREBP1, *PPARG* and *INSIG1* genes, as well as the encoded proteins; *b* – feedback involving the SREBP2, *PPARG* and *INSIG1* genes, as well as the encoded proteins; *c* – feedback involving the SREBP1, *AR* and *SCAP* genes, as well as the encoded proteins.

a



b

	all_CDS_19,556	*Enzymes_24*		χ^2 test
		Observed number	Expected number	*p*-value
PAI ≤ 2	6,397	22	7.85	
PAI > 2	13,159	2	16.15	
Total number	19,556	24	24	$p < 0.001$

Fig. 5. Phylostratigraphic age of human genes encoding enzymes of the cholesterol biosynthesis.

a – distribution of PAI values (indicated on the *X* axis) for all human protein-coding genes (control group of genes, designated as *all_CDS_19,556*, black columns) and genes encoding enzymes of the cholesterol biosynthesis (this group of genes is designated as *Enzymes_24*, green columns); *b* – according to the Chi-square criterion, the observed numbers of genes encoding enzymes and having PAI ≤ 2 differ from the expected numbers ($p < 0.001$).

in Supplementary Material 1. PAI values for the genes of the control group (designated as *all_CDS_19,556*) are unevenly distributed (Fig. 5*a*, black columns). Approximately one third of the genes (~33 %) had a PAI equal to zero (*Cellular organisms*, the root of the phytostratigraphic tree). And almost one fifth (17 %) of all protein-coding genes had a PAI value equal to 5 (the stage of *Vertebrata* divergence).

When considering the distribution of PAI values for a set of human genes encoding enzymes of cholesterol biosynthesis (Supplementary Material 1), it was found that 22 genes out of 24 (i. e. 92 %) had a PAI value ≤2 (*Cellular organisms* (the root of the phylostratigraphic tree) and the stages of *Eukaryota* and *Metazoa* divergence) (Fig. 5*a*, green columns). This number was different ($p < 0.001$) from the expected number (7.85) calculated based on the distribution obtained for a set of all human protein-coding genes containing 19,556 genes (Fig. 5*b*).

Thus, it turned out that the genes encoding enzymes of cholesterol biosynthesis are characterized by lower values of the PAI index compared to the set of all human protein-coding genes, that is, they are on average more "ancient". This is in good agreement with the already known concepts. Firstly, cholesterol is found in ancient sedimentary rocks, and its derivatives are used as biological markers of past life on Earth (Simoneit, 2002). Secondly, it was found that the genes encoding enzymes of cholesterol biosynthesis were inherited by multicellular organisms from their last common eukaryotic ancestor (Zhang T. et al., 2019). In addition, it has been shown that enzymes involved in amino acid, carbohydrate and energy metabolism (including lipid metabolism) are highly conservative (Peregrín-Alvarez et al., 2009). This is due to the fact that the role of the enzyme is to interact with the substrate molecule, that is, the three-dimensional structures of the enzyme and the substrate must spatially fit each other. Therefore, as a rule, it is not the protein-coding, but the regulatory region of the gene encoding the enzyme that undergoes evolutionary changes.

Genes encoding proteins functioning within the "cholesterol sensor". As mentioned above and shown in Figure 1, the "cholesterol sensor" is a set of proteins providing the regulation of the transcription of genes depending on the intracellular cholesterol level. The set of genes encoding proteins of this group includes: (1) the *SREBF1* and *SREBF2* genes encoding transcription factors; (2) the *SCAP*, *INSIG1*, and *INSIG2* genes encoding proteins that change their conformational properties in response to changes in cholesterol levels, thereby regulating the rate of formation of active SREBPs; (3) the *MBTPS1* and *MBTPS2* genes encoding S1P and S2P proteases that cleave precursor proteins preSREBP1 and preSREBP2 (DeBose-Boyd, Ye, 2018; Jiang et al., 2020). The phylostratigraphic age of these genes indicates the ancient origin of their ancestral forms (see the color designations of objects in Figure 1, as well as Supplementary Material 2).

Four genes (*SCAP, INSIG1, MBTPS1/S1P* and *MBTPS2/S2P*) have a PAI value equal to zero (*Cellular organisms*, the root of the phylostratigraphic tree). *INSIG2* has a PAI value equal to 2 (the stage of *Metazoa* divergence). However, the *SREBF1 и SREBF2* genes are younger. They have PAI values equal to 5 (the stage of *Vertebrata* divergence). Thus, although cholesterol was synthesized even in the most ancient organisms (Simoneit, 2002; Zhang T. et al., 2019), the molecular mechanism controlling intracellular cholesterol level could have been formed at a later stage of evolution. This could have happened no earlier than the first vertebrates appeared.

The stage of *Vertebrata* divergence is characterized by a more complex organization of a number of physiological systems (Fig. 6). The formation of the backbone was accompanied by musculoskeletal system development and made it possible to move faster. As a result, the oxygen demand of muscles and other tissues increased. A two-chamber heart was formed in vertebrates, which provided more efficient blood pumping and oxygen supply (Stephenson et al., 2017). At this stage of evolution, the respiratory system was being improved, and specialized oxygen-carrying blood cells (erythrocytes) arose (Snyder, Sheafor, 1999; Svoboda, Bartunek, 2015). The increased oxygen supply, on the one hand, contributed to the intensification of metabolic processes; on the other hand, it could cause oxidative stress.

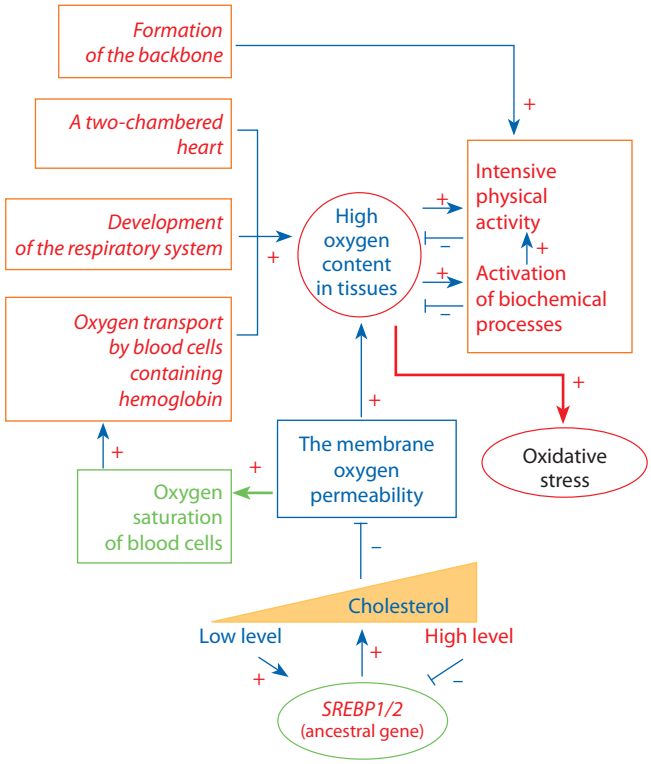


Fig. 6. Characteristic features of the musculoskeletal, circulatory and respiratory systems, formed in animals at the evolutionary stage of *Vertebrata* divergence (shown in italics), and the significant role of cholesterol as a factor reducing oxygen permeability of the cell membrane.

The cell membrane cholesterol content affects the permeability of the membrane to oxygen: when cholesterol content is high, the membrane becomes more solid leading to reduced oxygen permeability (Zuniga-Hertz, Patel, 2019). This, on the one hand, protects cells from oxidative stress, but, on the other hand, inhibits the transport of oxygen to red blood cells and negatively affects the biochemical processes occurring with oxygen consumption. Thus, it became necessary to maintain the intracellular cholesterol level in an appropriate range. Since a certain evolutionary stage, this control was carried out by transcription factors from the SREBP subfamily.

Conclusion

This paper presents a gene network regulating cholesterol biosynthesis in human cells. The gene network systematizes data on: (1) the set of enzymes that carry out cholesterol biosynthesis; (2) proteins functioning within the "cholesterol sensor" (including transcription factors from the SREBP subfamily), this sensor is involved in the regulation of gene expression depending on the intracellular cholesterol level; (3) proteins regulating the activity of proteins functioning within the "cholesterol sensor"; (4) genes encoding proteins of these groups; (5) SREBP target genes. Feedback loops have been identified that control the activity of transcription factors from the SREBP subfamily, indicating the complex nature of the molecular genetic mechanisms that regulate cholesterol biosynthesis. In the future, we plan to expand the network by including higher-level regulatory effects ("regulators of regulators"). Such an extension will help to identify additional feedback loops controlling cholesterol biosynthesis.

The analysis of the phylostratigraphic age of genes has shown that the ancestral forms of most human genes encoding enzymes of cholesterol biosynthesis and proteins of the "cholesterol sensor" could have been formed at early evolutionary stages (*Cellular organisms* (the root of the phylostratigraphic tree), as well as the stages of *Eukaryota and Metazoa* divergence). However, the phytostratigraphic age of genes encoding transcription factors of the SREBP subfamily corresponds to the stage of *Vertebrata* divergence. This fact indicates that the mechanism of gene transcription regulation in accordance with changes in cholesterol levels could have been formed at later evolutionary stages, that is, not earlier than the stage of *Vertebrata* divergence.

References

Agrawal A., Balci H., Hanspers K., Coort S.L., Martens M., Slenter D.N., Ehrhart F., Digles D., Waagmeester A., Wassink I., Abbassi-Daloii T., Lopes E.N., Iyer A., Acosta J.M., Willighagen L.G., Nishida K., Riutta A., Basaric H., Evelo C.T., Willighagen E.L., Kutmon M., Pico A.R. WikiPathways 2024: next generation pathway database. *Nucleic Acids Res*. 2024;52(D1):D679-D689. doi 10.1093/nar/gkad960

Arito M., Horiba T., Hachimura S., Inoue J., Sato R. Growth factor-induced phosphorylation of sterol regulatory element-binding proteins inhibits sumoylation, thereby stimulating the expression of their target genes, low density lipoprotein uptake, and lipid synthesis. *J. Biol. Chem*. 2008;283(22):15224-15231. doi 10.1074/jbc.M800910200

Choy H.L., Gaylord E.A., Doering T.L. Ergosterol distribution controls surface structure formation and fungal pathogenicity. *mBio*. 2023;14(4):e0135323. doi 10.1128/mbio.01353-23

DeBose-Boyd R.A., Ye J. SREBPs in lipid metabolism, insulin signaling, and beyond. *Trends Biochem. Sci*. 2018;43(5):358-368. doi 10.1016/j.tibs.2018.01.005

Desmond E., Gribaldo S. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol. Evol*. 2009;10(1):364-381. doi 10.1093/gbe/evp036

Fajas L., Schoonjans K., Gelman L., Kim J.B., Najib J., Martin G., Fruchart J.C., Briggs M., Spiegelman B.M., Auwerx J. Regulation of peroxisome proliferator-activated receptor gamma expression by adipocyte differentiation and determination factor 1/sterol regulatory element binding protein 1: implications for adipocyte differentiation and metabolism. *Mol. Cell. Biol*. 1999;19(8):5495-503. doi 10.1128/MCB.19.8.5495

Ferrer A., Altabella T., Arró M., Boronat A. Emerging roles for conjugated sterols in plants. *Prog. Lipid Res*. 2017;67:27-37. doi 10.1016/j.plipres.2017.06.002

GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330. doi 10.1126/science.aaz1776

Guo D., Wang Y., Wang J., Song L., Wang Z., Mao B., Tan N. RA-XII suppresses the development and growth of liver cancer by inhibition of lipogenesis via SCAP-dependent SREBP supression. *Molecules*. 2019;24(9):1829. doi 10.3390/molecules24091829

Han H., Cho J.W., Lee S., Yun A., Kim H., Bae D., Yang S., Kim C.Y., Lee M., Kim E., Lee S., Kang B., Jeong D., Kim Y., Jeon H.N., Jung H., Nam S., Chung M., Kim J.H., Lee I. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46(D1):D380-D386. doi 10.1093/nar/gkx1013

Heemers H., Verrijdt G., Organe S., Claessens F., Heyns W., Verhoeven G., Swinnen J.V. Identification of an androgen response element in intron 8 of the sterol regulatory element-binding protein

cleavage-activating protein gene allowing direct regulation by the androgen receptor. *J. Biol. Chem*. 2004;279(29):30880-30887. doi 10.1074/jbc.M401615200

Huang W.C., Zhau H.E., Chung L.W.K. Androgen receptor survival signaling is blocked by anti-β2-microglobulin monoclonal antibody via a MAPK/lipogenic pathway in human prostate cancer cells. *J. Biol. Chem*. 2010;285(11):7947-7956. doi 10.1074/jbc.M109.092759

Ishimoto K., Nakamura H., Tachibana K., Yamasaki D., Ota A., Hirano K.I., Tanaka T., Hamakubo T., Sakai J., Kodama T., Doi T. Sterol-mediated regulation of human lipin 1 gene expression in hepatoblastoma cells. *J. Biol. Chem*. 2009;284(33):22195-22205. doi 10.1074/jbc.M109.028753

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

Jeon T.I., Osborne T.F. SREBPs: metabolic integrators in physiology and metabolism. *Trends Endocrinol. Metab*. 2012:23(2):65-72. doi 10.1016/j.tem.2011.10.004

Jiang T., Zhang G., Lou Z. Role of the sterol regulatory element binding protein pathway in tumorigenesis. *Front. Oncol*. 2020;10:1788. doi 10.3389/fonc.2020.01788

Kast-Woelbern H.R., Dana S.L., Cesario R.M., Sun L., de Grandpre L.Y., Brooks M.E., Osburn D.L., Reifel-Miller A., Klausing K., Leibowitz M.D. Rosiglitazone induction of Insig-1 in white adipose tissue reveals a novel interplay of peroxisome proliferator-activated receptor gamma and sterol regulatory element-binding protein in the regulation of adipogenesis. *J. Biol. Chem*. 2004;279(23):23908-23915. doi 10.1074/jbc.M403145200

Kim H.J., Kim J.Y., Kim J.Y., Park S.K., Seo J.H., Kim J.B., Lee I.K., Kim K.S., Choi H.S. Differential regulation of human and mouse orphan nuclear receptor small heterodimer partner promoter by sterol regulatory element binding protein-1. *J. Biol. Chem*. 2004;279(27):28122-228131. doi 10.1074/jbc.M313302200

Kim H., Hiraishi A., Tsuchiya K., Sakamoto K. (-) Epigallocatechin gallate suppresses the differentiation of 3T3-L1 preadipocytes through transcription factors FoxO1 and SREBP1c. *Cytotechnology*. 2010;62(3):245-255. doi 10.1007/s10616-010-9285-x

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*. 2002;30(1):312-317. doi 10.1093/nar/30.1.312

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850 (in Russian)

Koolman J., Roehm K.H. (Eds). Color Atlas of Biochemistry. Stuttgart; New York: Thieme, 2005

Li N., Li X., Ding Y., Liu X., Diggle K., Kisseleva T., Brenner D.A. SREBP regulation of lipid metabolism in liver disease, and therapeutic strategies. *Biomedicines*. 2023;11(12):3280. doi 10.3390/biomedicines11123280

Luo J., Yang H., Song B.L. Mechanisms and regulation of cholesterol homeostasis. *Nat. Rev. Mol. Cell Biol*. 2020;21(4):225-245. doi 10.1038/s41580-019-0190-7

Macvanin M.T., Gluvic Z.M., Klisic A.N., Manojlovic M.S., Suri J.S., Rizzo M., Isenovic E.R. The link between miRNAs and PCKS9 in atherosclerosis. *Curr. Med. Chem*. 2024;31(42):6926-6956. doi 10.2174/0109298673262124231102042914

Mateus T., Martins F., Nunes A., Herdeiro M.T., Rebelo S. Metabolic alterations in myotonic dystrophy type 1 and their correlation with lipin. *Int. J. Environ. Res. Public. Health*. 2021;18(4):1794. doi 10.3390/ijerph18041794

Merkulova T.I., Ananko E.A., Ignatieva E.V., Kolchanov N.A. Transcription regulatory codes of eukaryotic genomes. *Russ. J. Genet*. 2013;49(1):29-45. doi 10.1134/S1022795413010079

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., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nes W.D. Biosynthesis of cholesterol and other sterols. *Chem. Rev*. 2011;111(10):6423-6451. doi 10.1021/cr200021m

Paul B., Lewinska M., Andersen J.B. Lipid alterations in chronic liver disease and liver cancer. *JHEP Rep*. 2022;4(6):100479. doi 10.1016/j.jhepr.2022.100479

Peregrín-Alvarez J.M., Sanford C., Parkinson J. The conservation and evolutionary modularity of metabolism. *Genome Biol*. 2009;10:R63. doi 10.1186/gb-2009-10-6-r63

Peterson T.R., Sengupta S.S., Harris T.E., Carmack A.E., Kang S.A., Balderas E., Guertin D.A., Madden K.L., Carpenter A.E., Finck B.N., Sabatini D.M. mTOR complex 1 regulates lipin 1 localization to control the SREBP pathway. *Cell*. 2011;146(3):408-420. doi 10.1016/j.cell.2011.06.034

Roth A., Looser R., Kaufmann M., Blättler S.M., Rencurel F., Huang W., Moore D.D., Meyer U.A. Regulatory cross-talk between drug metabolism and lipid homeostasis: constitutive androstane receptor and pregnane X receptor increase Insig-1 expression. *Mol. Pharmacol*. 2008;73(4):1282-1289. doi 10.1124/mol.107.041012

Sato R., Inoue J., Kawabe Y., Kodama T., Takano T., Maeda M. Sterol-dependent transcriptional regulation of sterol regulatory element-binding protein-2. *J. Biol. Chem*. 1996;271(43):26461-26464. doi 10.1074/jbc.271.43.26461

Schade D.S., Shey L., Eaton R.P. Cholesterol review: a metabolically important molecule. *Endocr. Pract*. 2020;26(12):1514-1523. doi 10.4158/EP-2020-0347

Shimano H., Sato R. SREBP-regulated lipid metabolism: convergent physiology – divergent pathophysiology. *Nat. Rev. Endocrinol*. 2017;13(12):710-730. doi 10.1038/nrendo.2017.91

Simoneit B.R. Molecular indicators (biomarkers) of past life. *Anat. Rec*. 2002;268(3):186-195. doi 10.1002/ar.10153

Snyder G.K., Sheafor B. Red blood cells: Centerpiece in the evolution of the vertebrate circulatory system. *Integr. Comp. Biol*. 1999;39(2):189-198. doi 10.1093/icb/39.2.189

Stephenson A., Adams J.W., Vaccarezza M. The vertebrate heart: an evolutionary perspective. *J. Anat*. 2017;231(6):787-797. doi 10.1111/joa.12687

Sundqvist A., Bengoechea-Alonso M.T., Ye X., Lukiyanchuk V., Jin J., Harper J.W., Ericsson J. Control of lipid metabolism by phosphorylation-dependent degradation of the SREBP family of transcription factors by SCF(Fbw7). *Cell Metab*. 2005;1(6):379-391. doi 10.1016/j.cmet.2005.04.010

Svoboda O., Bartunek P. Origins of the vertebrate erythro/megakaryocytic system. *Biomed. Res. Int*. 2015;2015:632171. doi 10.1155/2015/632171

Vargas-Alarcon G., Gonzalez-Pacheco H., Perez-Mendez O., Posadas-Sanchez R., Cardoso-Saldaña G., Ramirez-Bello J., Escobedo G., Nieto-Lima B., Fragoso J.M. *SREBF1c* and *SREBF2* gene polymorphisms are associated with acute coronary syndrome and blood lipid levels in Mexican population. *PLoS One*. 2019;14(9):e0222017. doi 10.1371/journal.pone.0222017

Waller D.D., Park J., Tsantrizos Y.S. Inhibition of farnesyl pyrophosphate (FPP) and/or geranylgeranyl pyrophosphate (GGPP) biosynthesis and its implication in the treatment of cancers. *Crit. Rev. Biochem. Mol. Biol*. 2019;54(1):41-60. doi 10.1080/10409238.2019.1568964

Watanabe M., Houten S.M., Wang L., Moschetta A., Mangelsdorf D.J., Heyman R.A., Moore D.D., Auwerx J. Bile acids lower triglyceride levels via a pathway involving FXR, SHP, and SREBP-1c. *J. Clin. Invest*. 2004;113(10):1408-1418. doi 10.1172/JCI21025

Zhang F., Sun W., Chen J., Jiang L., Yang P., Huang Y., Gong A., Liu S., Ma S. SREBP-2, a new target of metformin? *Drug Des. Devel. Ther*. 2018;12:4163-4170. doi 10.2147/DDDT.S190094

Zhang T., Yuan D., Xie J., Lei Y., Li J., Fang G., Tian L., Liu J., Cui Y., Zhang M., Xiao Y., Xu Y., Zhang J., Zhu M., Zhan S., Li S. Evolution of the cholesterol biosynthesis pathway in animals. *Mol. Biol. Evol*. 2019;36(11):2548-2556. doi 10.1093/molbev/msz167

Zuniga-Hertz J.P., Patel H.H. The evolution of cholesterol-rich membrane in oxygen adaption: The respiratory system as a model. *Front. Physiol*. 2019;10:1340. doi 10.3389/fphys.2019.01340

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

Received November 6, 2024. Revised November 15, 2024. Accepted November 18, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):874-881

doi 10.18699/vjgb-24-95

Orthoweb: a software package for evolutionary analysis of gene networks

R.A. Ivanov [1] ✉, A.M. Mukhin [1, 2], F.V. Kazantsev [1, 2], Z.S. Mustafin [2], D.A. Afonnikov [1, 2], Y.G. Matushkin [1], S.A. Lashin [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
✉ ivanovromanart@bionet.nsc.ru

Abstract. This article introduces Orthoweb (https://orthoweb.sysbio.cytogen.ru/), a software package developed for the calculation of evolutionary indices, including phylostratigraphic indices and divergence indices (K_a/K_s) for individual genes as well as for gene networks. The phylostratigraphic age index (PAI) allows the evolutionary stage of a gene's emergence (and thus indirectly the approximate time of its origin, known as "evolutionary age") to be assessed based on the analysis of orthologous genes across closely and distantly related taxa. Additionally, Orthoweb supports the calculation of the transcriptome age index (TAI) and the transcriptome divergence index (TDI). These indices are important for understanding the dynamics of gene expression and its impact on the development and adaptation of organisms. Orthoweb also includes optional analytical features, such as the ability to explore Gene Ontology (GO) terms associated with genes, facilitating functional enrichment analyses that link evolutionary origins of genes to biological processes. Furthermore, it offers tools for SNP enrichment analysis, enabling the users to assess the evolutionary significance of genetic variants within specific genomic regions. A key feature of Orthoweb is its ability to integrate these indices with gene network analysis. The software offers advanced visualization tools, such as gene network mapping and graphical representations of phylostratigraphic index distributions of network elements, ensuring intuitive interpretation of complex evolutionary relationships. To further streamline research workflows, Orthoweb includes a database of pre-calculated indices for numerous taxa, accessible via an application programming interface (API). This feature allows the users to retrieve pre-computed phylostratigraphic and divergence data efficiently, significantly reducing computational time and effort.
Key words: gene networks; evolution; phylostratigraphy.

For citation: 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 Journal of Genetics and Breeding*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Funding. This work was supported by State Budgetary Project No. FWNR-2022-0020.

Orthoweb: программный комплекс для эволюционного анализа генных сетей

Р.А. Иванов [1] ✉, А.М. Мухин [1, 2], Ф.В. Казанцев [1, 2], З.С. Мустафин [2], Д.А. Афонников [1, 2], Ю.Г. Матушкин [1], С.А. Лашин [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
✉ ivanovromanart@bionet.nsc.ru

Аннотация. В данной статье описывается Orthoweb (https://orthoweb.sysbio.cytogen.ru/) – программный комплекс, разработанный для вычисления эволюционных индексов, включая филостратиграфические индексы и индексы дивергенции (K_a/K_s) как отдельных генов, так и генных сетей. Индекс филостратиграфического возраста (PAI) позволяет оценить эволюционную стадию появления гена (при этом косвенно оценив приблизительное время его возникновения – так называемый эволюционный возраст) на основе анализа ортологичных генов у близкородственных и дальнородственных таксонов. Кроме того, Orthoweb поддерживает расчет индексов возраста транскриптома (TAI) и дивергенции транскриптома (TDI). Эти индексы важны для понимания динамики экспрессии генов и ее последствий для развития и адаптации организмов. Orthoweb содержит также дополнительные аналитические функции, такие как возможность анализа терминов Gene Ontology (GO), что позволяет проводить функциональное обогащение и связывать эволюционное происхождение генов с биологическими процессами. Помимо этого, доступна возможность анализа обогащения по однонуклеотидным полиморфиз-

мам (SNP), который помогает исследовать эволюционное значение генетических вариантов в конкретных геномных регионах. Одной из ключевых особенностей Orthoweb является интеграция перечисленных индексов с анализом генетических сетей. Программный пакет предлагает расширенные средства визуализации, такие как картирование генетических сетей и графическое представление распределения филостратиграфических индексов элементов сетей, что облегчает интуитивную интерпретацию сложных эволюционных связей. Для упрощения рабочих процессов в Orthoweb включена база данных с предварительно рассчитанными индексами для множества таксонов, доступная через API. Эта функция позволяет эффективно получать готовые данные по филостратиграфическим индексам и индексам дивергенции, значительно сокращая время вычислений.
Ключевые слова: генные сети; эволюция; филостратиграфия.

Introduction

The evolutionary analysis of gene networks allows the study of the origin and development of biological systems in the context of evolution. One of the key aspects of this analysis is the study of gene age indices, which allows us to determine the temporal framework for the emergence and diversification of genes across different phylogenetic lineages. Phylostratigraphy, a methodology based on estimating the evolutionary age of genes, provides an opportunity to identify ancient and recently emerged genes as well as to understand their functional significance in biological processes (Domazet-Lošo, Tautz, 2008; Tautz, Domazet-Lošo, 2011; Šestak et al., 2013; Xie et al., 2017). The aim of phylostratigraphic analysis is to determine the age of a founder gene by assessing the distribution of its homologous genes in the genomes of organisms belonging to different taxonomic groups. The Phylostratigraphic Age Index (PAI) is used in phylostratigraphy to estimate the time of origin of genes and corresponds to the oldest phylostratum that includes homologous sequences of the target gene.

The search for genes with homology restricted to specific taxa is particularly interesting from an evolutionary biology perspective, as several studies have shown that novel genes can play an important role in the emergence of new evolutionary traits and may be associated with the appearance of new morphological features in land plants (Bowles et al., 2020) and multicellular animals (Paps, Holland, 2018). It has also been shown that evolutionarily novel genes are involved in organ development cascades, particularly in brain tissue development (An et al., 2023), and that taxon-specific genes are overrepresented in stress response systems and the immune system (Dornburg, Yoder, 2022). Some researchers have also suggested that taxon-specific genes are associated with ecological specialisation in various taxa (Baalsrud et al., 2018).

However, the classical approach to phylostratigraphy faces several limitations due to the increasing volume of genomic data and the insufficient accuracy of the BLASTP algorithm in identifying homologs. These factors, together with high computational complexity, result in phylostratigraphic analyses of whole genome data using BLASTP taking up to several weeks (Buchfink et al., 2021). Consequently, there is a growing need for the development of new software solutions for phylostratigraphic analysis.

Modern software tools such as fagin (Arendsee et al., 2019), GenEra (Barrera-Redondo et al., 2023) and oggmap (Ullrich, Glytnasi, 2023) offer alternative approaches to phylostratigraphic analysis, allowing researchers to overcome some of the limitations of classical methods. The fagin program, written in R, uses a homology search approach based on identifying syntenic regions in the target genome and then searching for homology in both amino acid and nucleotide sequences. The developers of the GenEra software package have introduced several modifications to the classical method of homology detection in phylostratigraphy by replacing the traditional BLASTP search method with the DIAMOND v2 algorithm. This substitution improves the identification of distant homologs by removing restrictions on the number of top sequence matches during alignment. In addition, GenEra's developers have incorporated features to assess homology detection error and taxonomic representativeness – a metric that considers the presence of gene homologs in at least one representative species at each intermediate taxonomic level between the most distantly related genus and the target species. The oggmap program (Ullrich, Glytnasi, 2023), implemented as a Python package, is designed to generate orthology maps (orthomaps), or, in other words, phylostratigraphic index values for the age of specified ortholog groups, based on the results of tools such as OrthoFinder (Emms, Kelly, 2019) and eggNOG (Huerta-Cepas et al., 2019). Unlike classical phylostratigraphy, this approach does not include a step for ortholog detection using alignment tools. Instead, it relies on precomputed orthology search results in the form of orthomaps, which are then used to estimate gene age. These orthomaps contain information about the ages of genes within each ortholog group.

However, for comprehensive evolutionary analysis, these tools and approaches require knowledge of programming languages. In addition, most of these software solutions rely on alignment algorithms such as BLAST, the runtime of which can significantly slow down the analysis in certain cases. Finally, the existing implementations for calculating phylostratigraphic indices are currently unable to perform a comprehensive and rapid evolutionary analysis of gene network components. In this article, we present Orthoweb – a software package for the evolutionary analysis of gene networks and individual genes – implemented as a web application and available at https://orthoweb.sysbio.cytogen.ru.

Materials and methods

Orthoweb has been developed in Java using the Spring framework to implement server-side functionality and the Vue.JS and webix frameworks for the client side. A set of cytoscape.js libraries is used for network visualization. MongoDB is used as the database management system (DBMS) to store data from the KEGG database (taxa, list of orthologs, coding sequences, etc.) and intermediate analysis results, which

significantly increases the speed of subsequent work with these data.

A database based on the PostgreSQL DBMS is used to store the calculated indices. Access to the data is provided through REST API technology implemented with the FLASK library (flask.palletsprojects.com). This programmatic interface allows data retrieval from various engineering modelling environments (e. g. Matlab, Octave, Statistica) or standard libraries of scripting programming languages (e. g. R, Python).

Results

Functionality of Orthoweb

Calculation of evolutionary age indices of single genes. The primary function of Orthoweb is the estimation of phylostratigraphic age indices (PAI) of genes.

Orthoweb implements two methods to determine PAI: 1) based on the analysis of homology sequence identity metrics and 2) using the classification of proteins into orthologous groups from the KEGG database (KEGG Orthology – KO). Using the KO information from the KEGG database (Kanehisa et al., 2016), Orthoweb allows the identification of orthologs for each protein sequence and determines the species in the genomes of which these orthologs have been found. The taxonomic lineages of the identified species are sequentially compared to the lineage of the studied species to determine their evolutionary ancestry and to determine the most recent common ancestor for a given gene. The position of this ancestor, measured as its distance from the root of the taxonomic tree, is calculated as the PAI (Fig. 1). The taxonomic lineages of orthologs have already been curated in the KEGG database, requiring minimal additional configuration by the user. The calculated PAI indices are stored in a regularly updated database, which is discussed in more detail in the chapter "Database for storing results". As KEGG orthogroup data is frequently updated, Orthoweb also allows to calculate PAI indices directly from KEGG orthogroups to ensure access to the most up-to-date information. However, such data are not available for all genes. For example, in humans, only about two-thirds of the genes represented in KEGG are associated with KO groups.

The second method for calculating PAI involves using the Best Similarity Table, which is available for the vast majority of genes represented in KEGG (Kanehisa et al., 2016). This method allows users to select homologous genes based on parameters such as the amino acid sequence identity of the proteins encoded by the genes and the results of the Smith–Waterman local sequence alignment algorithm.

Calculation of divergence indices. Orthoweb also supports the calculation of the ratio of nonsynonymous to synonymous substitutions (the d_N/d_S ratio) between the sequence of the gene under study and each of its homologs in closely related species, reflected in the Divergence Index (DI). This index allows researchers to determine the type of selection acting on a gene. The index is calculated based on the d_N/d_S ratio (also referred to as K_a/K_s in the literature), where d_N represents the proportion of nonsynonymous substitutions in the sequences of the gene under study and its homologs (i. e. substitutions that result in a change in the amino acid encoded by the codon)

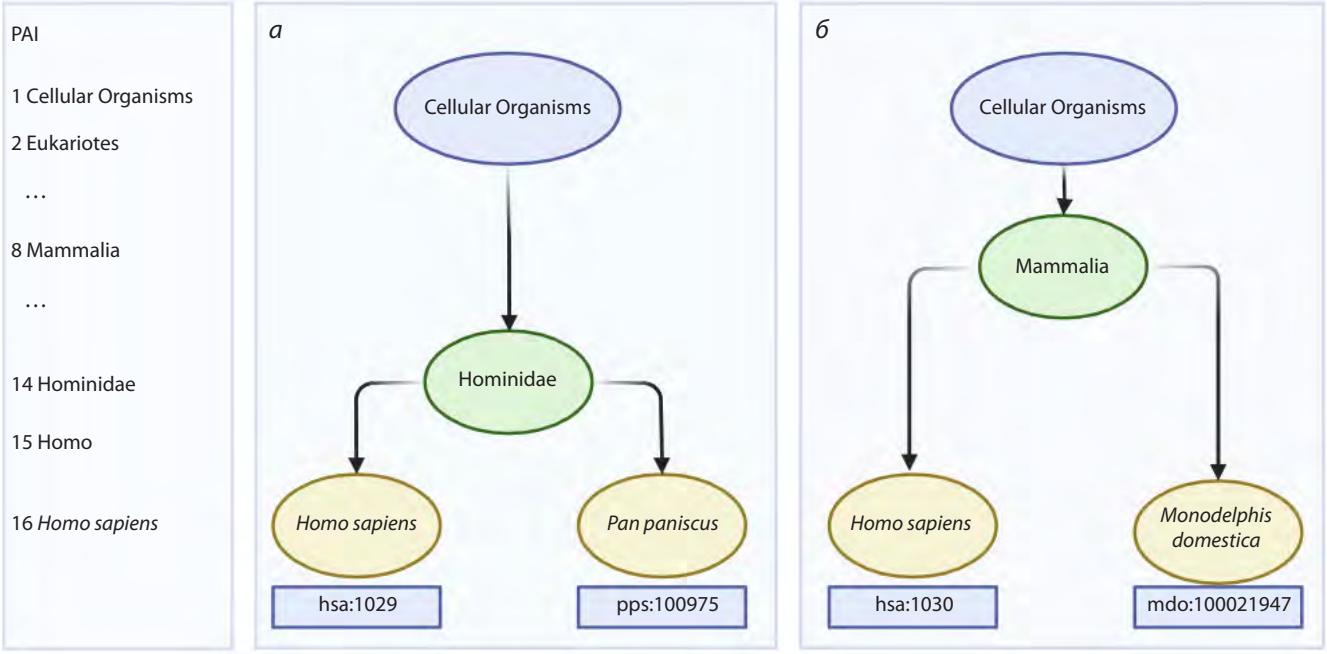


Fig. 1. Example of a PAI calculation for two *Homo sapiens* genes.

a – example of an evolutionarily younger gene hsa:1029 (CDKN2A), where the most distantly related organism with an identified ortholog of this gene is *Pan paniscus* (bonobo chimpanzee); *b* – example of an evolutionarily older gene hsa:1030 (CDKN2B), where the most distantly related organism with an identified ortholog of this gene is *Monodelphis domestica* (grey short-tailed opossum). It can be concluded that the gene in example (*a*) is evolutionarily younger than the gene in example (*b*). The scale on the left indicates the PAI index, which corresponds to the depth of the taxonomic tree node. Adapted from (Mustafin et al., 2021).

and d_S represents the proportion of synonymous substitutions (i. e. those that do not result in a change in the encoded amino acid). It is generally accepted that DI values less than 1 indicate that the gene is under purifying selection, values close to 1 suggest neutral evolution, and values greater than 1 imply positive selection (Yang, Nielsen, 2000).

When comparing a single homologous sequence, DI is equivalent to d_N/d_S. In cases where multiple homologs are present, DI is equal to the average d_N/d_S value across all comparisons. When calculating the DI index, Orthoweb users can select the taxonomic depth of analysis to account for the evolutionary variability of the gene between organisms with varying evolutionary distances. The calculation of the d_N/d_S ratio is performed using the PAML software package (Yang, 2007).

Calculation of gene enrichment with single nucleotide polymorphisms and Gene Ontology term analysis. Orthoweb also integrates information on Gene Ontology (GO) terms associated with genes and the enrichment of the studied genes with single nucleotide polymorphisms (SNPs). To retrieve information on Gene Ontology terms, Orthoweb uses the resource available at http://geneontology.org/ (Ashburner et al., 2000; Carbon et al., 2021). Data retrieval is performed using the API (application programming interface) provided. For example, a query for the TBP gene is constructed as follows: http://api.geneontology.org/api/bioentity/gene/NCBIGene:6908/function, specifying the database and the gene identifier within it. Orthoweb provides this information autonomously, relying on associated databases for most model organisms (e. g. TAIR for *Arabidopsis thaliana*, FlyBase for *Drosophila melanogaster*, etc.), while for other organisms, it uses the UniProt database. If Gene Ontology contains data for the gene under study and KEGG provides the required identifier – which is true for nearly all well-characterised genes – then identifiers and names of GO terms associated with the gene will be retrieved.

To obtain data on the enrichment of target genes with single nucleotide polymorphisms, an automated query system for the NCBI SNP database (Sayers et al., 2022) is implemented. The query is constructed based on the gene identifier. For example, for the TBP gene with the identifier hsa:6908, the query would take the following form: https://www.ncbi.nlm.nih.gov/snp/?term=6908. As a result of this query, the user will be provided with the number of SNPs found. It should be noted that in the current version of Orthoweb, the SNP search is only implemented for human genes.

Calculation of evolutionary indices of a group of genes. Orthoweb also supports the input of gene expression data for the calculation of phylotranscriptomic indices. Phylotranscriptomic index analysis is an approach that integrates information on the evolutionary age of genes with data on their expression levels. This analysis enables the study of the relationship between the PAI index of genes and changes in their activity in the context of different physiological states, adaptive responses or developmental stages of organisms. Using phylotranscriptomic analysis, it is possible to uncover how the evolutionary features of the genome relate to the transcriptional regulation and functional dynamics of genes in different biological contexts. Phylotranscriptomic indices include two evolutionary indices: Transcriptome Age Index (Domazet-Lošo, Tautz, 2010) and Transcriptome Divergence Index (Quint et al., 2012)

The Transcriptome Age Index (TAI) represents the weighted average age of the transcriptome in a given biological process. Expression data serve as an additional multiplier and are used to normalise the result so that the higher the final TAI/TDI value, the greater the contribution of evolutionarily younger/more variable genes. The formulas used to calculate these indices are as follows:

$$\mathrm{TAI} = \frac{\sum_{i=1}^{n} ps_i e_i}{\sum_{i=1}^{n} e_i},$$

where ps_i is an integer representing the PAI for gene i, e_i is the expression level derived from transcriptomic data for gene i, and n is the total number of genes.

The Transcriptome Divergence Index (TDI) measures transcriptome divergence and reflects the degree of conservation of a transcriptome in a particular process. This can be used to identify biological processes or development stages in which more conserved, or younger, genes are more highly expressed.

$$\mathrm{TDI} = \frac{\sum_{i=1}^{n} \mathrm{DI}_i e_i}{\sum_{i=1}^{n} e_i}.$$

where DI_i is the divergence index for gene i, e_i is the expression level for gene i, n is the total number of genes.

Orthoweb usage examples

To illustrate how Orthoweb works, we will describe its workflow and give examples of its use in phylostratigraphic analysis.

Analysis of individual gene characteristics. When analysing evolutionary indices for single genes, Orthoweb accepts several input file formats: a list of genes entered via a web form, a list of genes uploaded from a file, or a file containing interactions between elements of a gene network in .txt or .tsv format. Users can select the desired input data format in the corresponding form labelled *Choose the type of input data* (Fig. 2). For accurate analysis in Orthoweb, KEGG gene identifiers must be provided.

The next step involves selecting the analysis mode in the form titled *The type of orthology*. In this form, you can choose one of two options: calculating phylostratigraphic indices using ortholog family and KO group analysis (the *KEGG Orthology groups* option) or using homologous sequence analysis (the *Best Similarity Table* option).

When selecting the KEGG Orthology groups mode, it is also necessary to decide whether to include paralogous genes in the analysis by configuring the *KO groups filtering* option.

When selecting the mode for calculating phylostratigraphic indices of genes based on homologous sequence analysis, it is necessary to specify the thresholds for amino acid sequence identity (set to 0.5 by default) and for the Smith–Waterman algorithm score used to filter homologous genes in the *The thresholds to filter orthologous genes* option.

In the *Additional parameters* section, several additional analysis options can be selected: calculation of the divergence index (DI) in the *DI analysis* option, assessment of enrichment

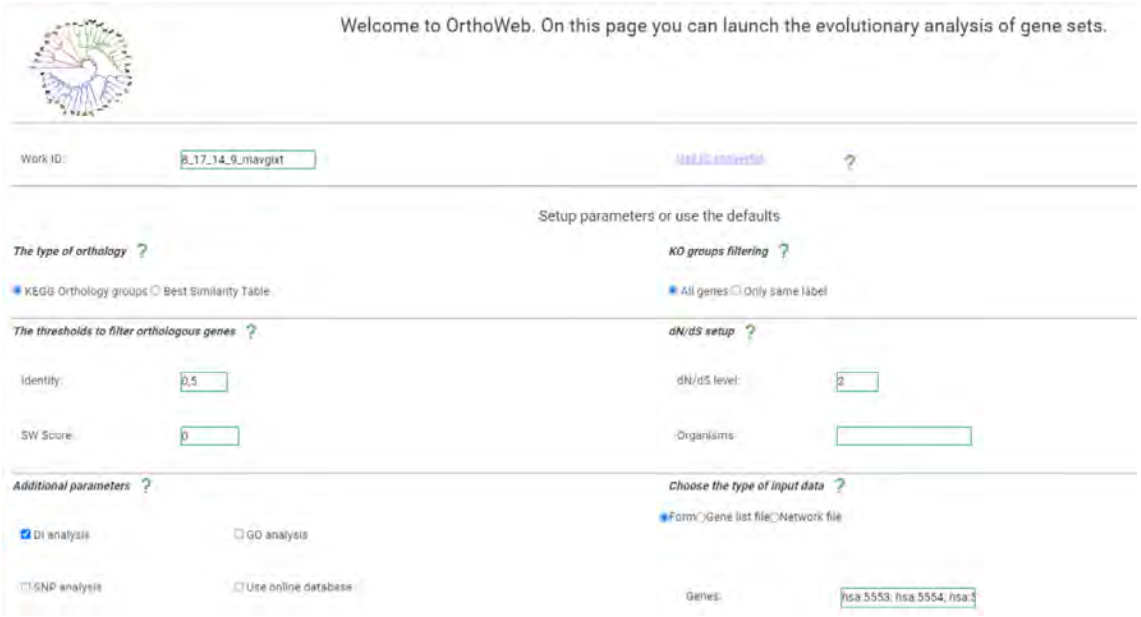


Fig. 2. The Start Page of the Orthoweb Web Service.

with single nucleotide polymorphisms (SNPs) and identification of Gene Ontology terms. For DI calculation, it is also possible to configure the groups of organisms for which the index is calculated in the d_N/d_S *setup* window. This option provides two configurations for the analysis. The first parameter, d_N/d_S *level*, defines the taxonomic level at which the d_N/d_S analysis is performed. This type of analysis is primarily used to compare sequences of closely related organisms. A value of 1 limits the analysis to organisms within a single genus. For example, when analysing human genes, a value of 2 indicates that the d_N/d_S will be calculated relative to other organisms in the Hominidae family. The second field, *Organisms*, allows you to enter specific species codes from the KEGG database. For example, to compare the sequence of a studied human gene not with all hominids but only with gorillas, the code "*ggo*" should be entered in this field.

The output of Orthoweb for these analysis modes will be an archive file containing a tabular text file with the following data columns: Gene – KEGG gene identifiers, Label – Entrez gene identifiers, PAI – phylostratigraphic age index values; additional columns with values from supplementary analysis modes: DI, SNP and GO label.

Analysis of gene group characteristics. To calculate the Transcriptome Age Index (TAI) and the Transcriptome Divergence Index (TDI), it is necessary to select the input data format option *Network file – Use expression*. In this mode, the user must provide a tab-delimited text file containing one column of gene names and several columns of normalised expression values, labelled according to the experimental conditions under which the expression analysis was performed. The input file can be either a gene network file or simply a list of genes.

As output, the Orthoweb program generates a tab-delimited text file with three columns: Data – with the names of the conditions specified in the input file, TAI – with the transcriptome age index values for the selected set of genes, and TDI – with the transcriptome divergence index values under the given conditions.

Gene network analysis. In addition to the analysis of indices for individual genes and gene lists, Orthoweb implements phylostratigraphic analysis and visualization of gene networks. Users can analyse networks imported from the KEGG Pathway (Kanehisa et al., 2017) and WikiPathways databases, as well as networks uploaded from text files. Access to network analysis from these databases is provided via the following link: https://orthoweb.sysbio.cytogen.ru/pathway.html.

Orthoweb supports import and analysis of networks from two major databases. The first supported database, KEGG Pathway, contains numerous gene networks and pathways classified according to various criteria such as metabolism, organismal system functions and human diseases. To start the analysis, the user must specify the pathway code and the organism for which the network is to be imported. As an output of network analysis from KEGG Pathway, Orthoweb will generate a gene network where the nodes display PAI values determined based on the KO groups present in the network. Since all elements in KEGG networks are described in the KEGG database itself, importing and analysing such networks is very convenient for Orthoweb, which retrieves most of the information needed for analysis directly from KEGG.

As an example of this mode in Orthoweb, we analysed the Wnt/β-catenin signalling cascade network (Fig. 3). The Wnt/β-catenin signalling pathway is involved in the regulation of the cell cycle, adhesion, migration and differentiation. Activation of the pathway begins with the binding of WNT ligands to Frizzled and LRP receptors on the cell surface. This leads to the stabilisation and accumulation of β-catenin in the cytoplasm and its subsequent translocation to the nucleus, where it interacts with transcription factors and stimulates the

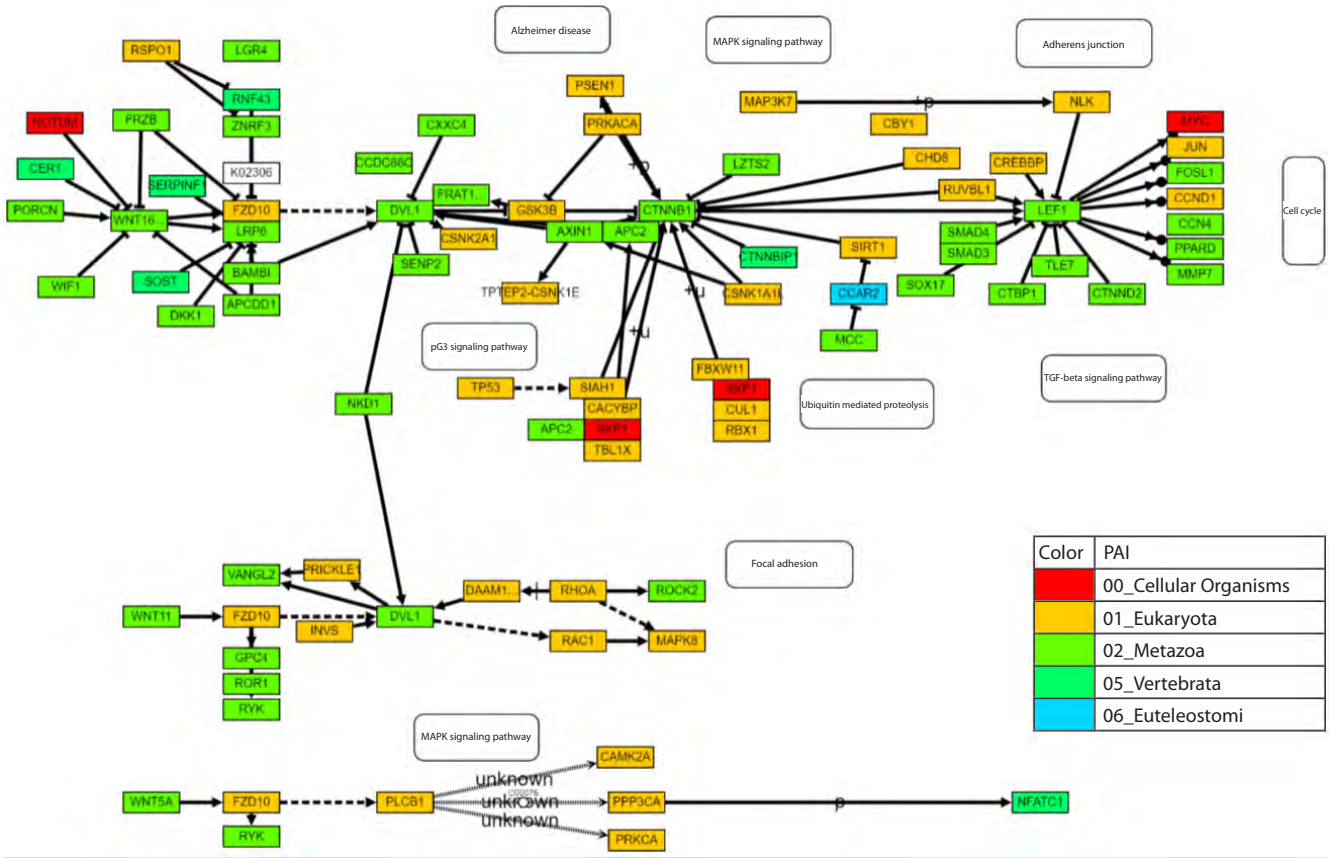


Fig. 3. Example of network visualization from the KEGG Pathway database for the "Wnt signalling pathway" (https://www.kegg.jp/pathway/hsa04310), analysed using Orthoweb.

The color of each node corresponds to the PAI index of the gene (white elements represent pathways and chemical compounds). By default, the standard network structure is imported with preserved element coordinates, but the network scale can be adjusted by the user and each element can be interacted with.

expression of target genes (Davidson et al., 2009). Dysregulation of this pathway has been implicated in the development of several cancer types (Zhan et al., 2017). This signalling cascade is one of the most ancient signalling pathways and predominantly involves genes that originated during the emergence of multicellular organisms and eukaryotes (PAI = 1, 2).

The second database available for network import is WikiPathways. The networks presented in WikiPathways contain more details, entities and interaction variants compared to KEGG, which makes their complete import more difficult and requires the consideration of identifiers from several different databases.

Orthoweb provides a step-by-step process for importing and analysing user-generated gene networks. Users can first import a network in TSV format (a tab-delimited text file) and then interact with it, e. g. rearrange elements, before starting the analysis. This format is compatible with the widely used STRING tool (von Mering et al., 2005), ensuring seamless integration of STRING data into Orthoweb without additional processing. For networks imported from STRING, the combined_score column contains the reliability of identified interactions, with weights ranging from 0 to 1. Upon completion of the analysis, the gene colours are updated to reflect their PAI values (Fig. 4). If additional analysis modes described earlier in the text are enabled, they will also be reflected in the visualization.

Database for storing results

To speed up index calculations and avoid redundant recalculations, Orthoweb includes a database containing tables for organisms, genes, pre-calculated PAI indices, DI indices, Gene Ontology terms (identifiers and names), SNPs and PAI indices determined based on KO groups. In addition to its use in interactive mode, this database can also be accessed via an API (Application Programming Interface) for integration with modelling environments or common scripting languages (Matlab, Octave, R, Python, etc.). This provides access to all available information on calculated PAI and DI indices for genes of specific organisms, allowing users to build data selection and visualization workflows. The API allows database queries to be made via specially structured URLs. Query results are returned as a structured text file in JSON format. A description of the API query keys and an example query to the database can be found in the Supplementary Material[1].

[1] Supplementary Material is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Ivanov_Engl_28_8.pdf

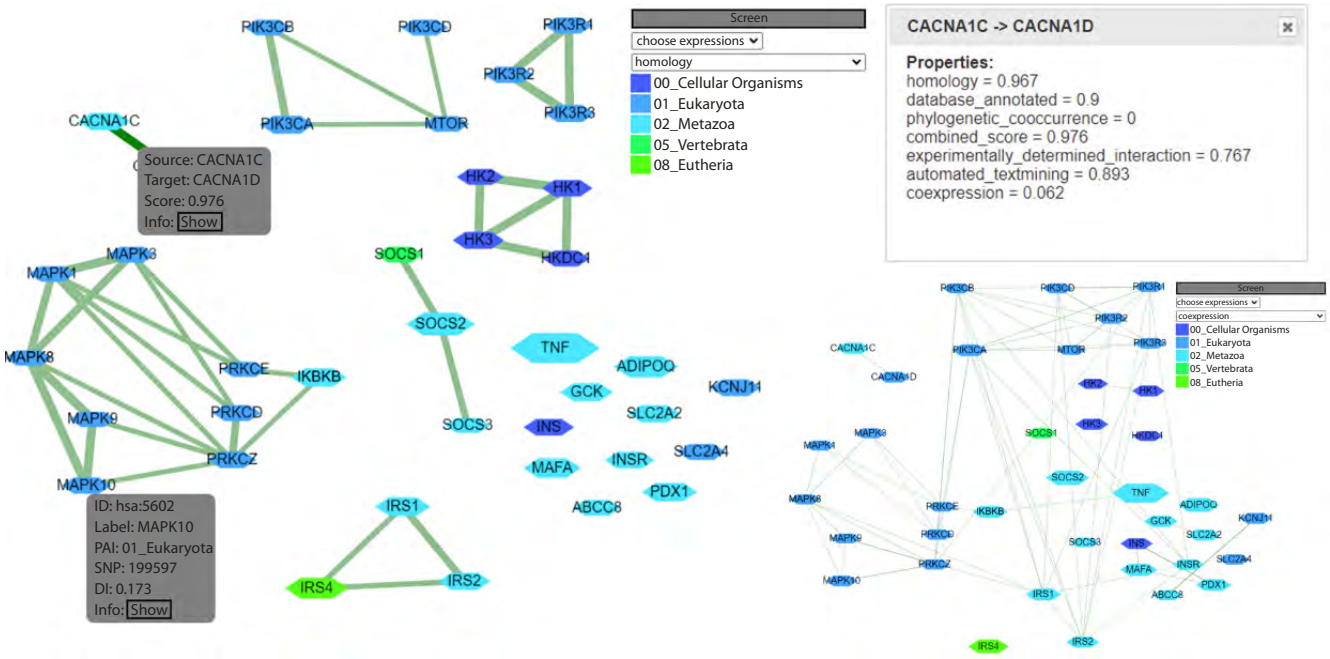


Fig. 4. Example of a network imported from the STRING tool, where the color of each node corresponds to its PAI index and the thickness of the edges represents the combined_score value from STRING. By selecting a specific interaction within the network, information about the confidence levels of that interaction in STRING is provided.

Conclusion

In this article, we present Orthoweb – a software platform designed for the analysis of phylostratigraphic and divergence indices for both individual genes and gene networks. Orthoweb also allows the integration of evolutionary index values with gene expression data under different conditions.

One of the key features of Orthoweb is its advanced data visualization capabilities. The tools for mapping evolutionary indices onto gene networks greatly simplify the interpretation of complex evolutionary relationships, making the results of analysis accessible to a wide range of researchers.

References

An N.A., Zhang J., Mo F., Luan X., Tian L., Shen Q.S., Li X., Li C., Zhou F., Zhang B., Ji M., Qi J., Zhou W.-Z., Ding W., Chen J.-Y., Yu J., Zhang L., Shu S., Hu B., Li C.-Y. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* 2023;7(2):264-278. doi 10.1038/s41559-022-01925-6

Arendsee Z., Li J., Singh U., Bhandary P., Seetharam A., Wurtele E.S. fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics*. 2019;20(1):440. doi 10.1186/s12859-019-3023-y

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., 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

Baalsrud H.T., Tørresen O.K., Solbakken M.H., Salzburger W., Hanel R., Jakobsen K.S., Jentoft S. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* 2018;35(3):593-606. doi 10.1093/molbev/msx311

Barrera-Redondo J., Lotharukpong J.S., Drost H.-G., Coelho S.M. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biol.* 2023;24(1):54. doi 10.1186/s13059-023-02895-z

Bowles A.M.C., Bechtold U., Paps J. The origin of land plants is rooted in two bursts of genomic novelty. *Curr. Biol.* 2020;30(3):530-536.e2. doi 10.1016/j.cub.2019.11.090

Buchfink B., Reuter K., Drost H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*. 2021;18(4):366-368. doi 10.1038/s41592-021-01101-x

Carbon S., Douglass E., Good B.M., Unni D.R., Harris N.L., Mungall C.J., Basu S., Chisholm R.L., Dodson R.J., Hartline E., … Stein L., Howe D.G., Toro S., Westerfield M., Jaiswal P., Cooper L., Elser J. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021;49(D1):D325-D334. doi 10.1093/nar/gkaa1113

Davidson G., Shen J., Huang Y.-L., Su Y., Karaulanov E., Bartscherer K., Hassler C., Stannek P., Boutros M., Niehrs C. Cell cycle control of Wnt receptor activation. *Dev. Cell*. 2009;17(6):788-799. doi 10.1016/j.devcel.2009.11.006

Domazet-Lošo T., Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* 2008;25(12):2699-2707. doi 10.1093/molbev/msn214

Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*. 2010;468(7325):815-819. doi 10.1038/nature09632

Dornburg A., Yoder J.A. On the relationship between extant innate immune receptors and the evolutionary origins of jawed vertebrate adaptive immunity. *Immunogenetics*. 2022;74(1):111-128. doi 10.1007/s00251-021-01232-7

Emms D.M., Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. doi 10.1186/s13059-019-1832-y

Huerta-Cepas J., Szklarczyk D., Heller D., Hernández-Plaza A., Forslund S.K., Cook H., Mende D.R., Letunic I., Rattei T., Jensen L.J., von Mering C., Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on

5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1): D309-D314. doi 10.1093/nar/gky1085

Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):D457-D462. doi 10.1093/nar/gkv1070

Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(D1):D353-D361. doi 10.1093/nar/gkw1092

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

Paps J., Holland P.W.H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat. Commun.* 2018; 9(1):1730. doi 10.1038/s41467-018-04136-5

Quint M., Drost H.G., Gabel A., Ullrich K.K., Bönn M., Grosse I. A transcriptomic hourglass in plant embryogenesis. *Nature*. 2012; 490(7418):98-101. doi 10.1038/nature11394

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2022;50(D1):D20-D26. doi 10.1093/nar/gkab1112

Šestak M.S., Božičević V., Bakarić R., Dunjko V., Domazet-Lošo T. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front. Zool.* 2013;10(1):18. doi 10.1186/1742-9994-10-18

Tautz D., Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 2011;12(10):692-702. doi 10.1038/nrg3053

Ullrich K.K., Glytnasi N.E. oggmap: a Python package to extract gene ages per orthogroup and link them with single-cell RNA data. *Bioinformatics*. 2023;39(11):btad657. doi 10.1093/bioinformatics/btad657

von Mering C., Jensen L.J., Snel B., Hooper S.D., Krupp M., Foglierini M., Jouffre N., Huynen M.A., Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 2005;33(D1):D433-D437. doi 10.1093/nar/gki005

Xie L., Draizen E.J., Bourne P.E. Harnessing big data for systems pharmacology. *Annu. Rev. Pharmacol. Toxicol.* 2017;57(1):245-262. doi 10.1146/annurev-pharmtox-010716-104659

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591. doi 10.1093/molbev/msm088

Yang Z., Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 2000;17(1):32-43. doi 10.1093/oxfordjournals.molbev.a026236

Zhan T., Rindtorff N., Boutros M. Wnt signaling in cancer. *Oncogene*. 2017;36(11):1461-1473. doi 10.1038/onc.2016.304

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

Received November 8, 2024. Revised November 21, 2024. Accepted November 22, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):882-896

doi 10.18699/vjgb-24-96

Investigation of metabolic features of glioblastoma tissue and the peritumoral environment using targeted metabolomics screening by LC-MS/MS and gene network analysis

N.V. Basov[1, 2 §], A.V. Adamovskaya[2, 3 §], A.D. Rogachev[1, 2], E.V. Gaisler[2], P.S. Demenkov[2, 3], T.V. Ivanisenko[2, 3], A.S. Venzel[2, 3], S.V. Mishinov[4], V.V. Stupak[4], S.V. Cheresiz[2], O.S. Oleshko[2], E.A. Butikova[2, 5], A.E. Osechkova[1, 6], Yu.S. Sotnikova[1, 2, 6], Y.V. Patrushev[2, 6], A.S. Pozdnyakov[7], I.N. Lavrik[3], V.A. Ivanisenko[2, 3, 8], A.G. Pokrovsky[2]

[1] N.N. Vorozhtsov Novosibirsk Institute of Organic Chemistry of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Novosibirsk Research Institute of Traumatology and Orthopedics named after Ya.L. Tsivyan of the Ministry of Health of the Russian Federation, Novosibirsk, Russia
[5] Research Institute of Clinical and Experimental Lymphology – Branch of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[6] Boreskov Institute of Catalysis of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[7] A.E. Favorsky Irkutsk Institute of Chemistry of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
[8] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

basov@nioch.nsc.ru

Abstract. The metabolomic profiles of glioblastoma and surrounding brain tissue, comprising 17 glioblastoma samples and 15 peritumoral tissue samples, were thoroughly analyzed in this investigation. The LC-MS/MS method was used to analyze over 400 metabolites, revealing significant variations in metabolite content between tumor and peritumoral tissues. Statistical analyses, including the Mann–Whitney and Cucconi tests, identified several metabolites, particularly ceramides, that showed significant differences between glioblastoma and peritumoral tissues. Pathway analysis using the KEGG database, conducted with MetaboAnalyst 6.0, revealed a statistically significant overrepresentation of sphingolipid metabolism, suggesting a critical role of these lipid molecules in glioblastoma pathogenesis. Using computational systems biology and artificial intelligence methods implemented in a cognitive platform, ANDSystem, molecular genetic regulatory pathways were reconstructed to describe potential mechanisms underlying the dysfunction of sphingolipid metabolism enzymes. These reconstructed pathways were integrated into a regulatory gene network comprising 15 genes, 329 proteins, and 389 interactions. Notably, 119 out of the 294 proteins regulating the key enzymes of sphingolipid metabolism were associated with glioblastoma. Analysis of the overrepresentation of Gene Ontology biological processes revealed the statistical significance of 184 processes, including apoptosis, the NF-kB signaling pathway, proliferation, migration, angiogenesis, and pyroptosis, many of which play an important role in oncogenesis. The findings of this study emphasize the pivotal role of sphingolipid metabolism in glioblastoma development and open new prospects for therapeutic approaches modulating this metabolism.

Key words: glioblastoma; peritumoral tissue; markers; metabolomics; LC-MS/MS; sphingolipids; metabolic pathways; gene networks; cognitive system ANDSystem.

For citation: Basov N.V., Adamovskaya A.V., Rogachev A.D., Gaisler E.V., Demenkov P.S., Ivanisenko T.V., Venzel A.S., Mishinov S.V., Stupak V.V., Cheresiz S.V., Oleshko O.S., Butikova E.A., Osechkova A.E., Sotnikova Yu.S., Patrushev Y.V., Pozdnyakov A.S., Lavrik I.N., Ivanisenko V.A., Pokrovsky A.G. Investigation of metabolic features of glioblastoma tissue and the peritumoral environment using targeted metabolomics screening by LC-MS/MS and gene network analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):882-896. doi 10.18699/vjgb-24-96

Funding. The production of monolithic columns for LC was made possible by the financial support of the FWUR-2024-0032 project. The selection and preparation of samples and their subsequent analysis by LC-MS/MS were supported by the funds of the state task No. FSUS-2020-0035. The bioinformatics analysis was funded by the budget project FWNR-2022-0020.

§ The authors contributed equally to this study.

Изучение особенностей метаболизма тканей глиобластомы и перитуморального пространства при использовании таргетированного метаболомного скрининга методом ВЭЖХ-МС/МС и генных сетей

Н.В. Басов[1, 2 §], А.В. Адамовская [2, 3 §], А.Д. Рогачев [1, 2], Е.В. Гайслер[2], П.С. Деменков [2, 3], Т.В. Иванисенко [2, 3], А.С. Вензель[2, 3], С.В. Мишинов [4], В.В. Ступак[4], С.В. Чересиз[2], О.С. Олешко[2], Е.А. Бутикова[2, 5], А.Е. Осечкова [1, 6], Ю.С. Сотникова[1, 2, 6], Ю.В. Патрушев[2, 6], А.С. Поздняков[7], И.Н. Лаврик[3], В.А. Иванисенко [2, 3, 8], А.Г. Покровский [2]

[1] Новосибирский институт органической химии им. Н.Н. Ворожцова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Новосибирский научно-исследовательский институт травматологии и ортопедии им. Я.Л. Цивьяна Министерства здравоохранения Российской Федерации, Новосибирск, Россия
[5] Научно-исследовательский институт клинической и экспериментальной лимфологии – филиал Федерального исследовательского центра Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[6] Федеральный исследовательский центр Институт катализа им. Г.К. Борескова Сибирского отделения Российской академии наук, Новосибирск, Россия
[7] Федеральный исследовательский центр «Иркутский институт химии им. А.Е. Фаворского Сибирского отделения Российской академии наук», Иркутск, Россия
[8] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
basov@nioch.nsc.ru

Аннотация. В ходе исследования проведен комплексный анализ метаболомных профилей глиобластомы и прилегающей ткани головного мозга, включавший 17 образцов глиобластомы и 15 образцов перитуморальной ткани. С использованием метода ВЭЖХ-МС/МС было проанализировано более 400 метаболитов, что позволило выявить значимые различия в их содержании между опухолевой и перитуморальной тканями. Статистический анализ, включавший тесты Манна–Уитни и Куккони, показал, что существенное количество метаболитов, в частности церамиды, значимо различается в тканях глиобластомы и перитуморального пространства. Анализ метаболических путей из базы данных KEGG, выполненный с помощью MetaboAnalyst 6.0, выявил статистически значимую перепредставленность метаболизма сфинголипидов, что указывает на важную роль этих липидных молекул в патогенезе глиобластомы. С использованием методов компьютерной системной биологии и искусственного интеллекта, реализованных в когнитивной системе ANDSystem, реконструированы молекулярно-генетические регуляторные пути, описывающие потенциальные механизмы нарушения функции ферментов метаболизма сфинголипидов. Реконструированные пути были объединены в регуляторную генную сеть. Данная сеть включала 15 генов, 329 белков и 389 взаимодействий, при этом 119 из 294 белков, регулирующих ключевые ферменты сфинголипидного метаболизма, оказались ассоциированы с глиобластомой. Анализ перепредставленности биологических процессов Gene Ontology показал статистическую значимость 184 процессов, в том числе апоптоза, сигнального пути NF-κB, пролиферации, миграции, ангиогенеза и пироптоза, многие из которых играют важную роль в онкогенезе. Результаты исследования подчеркивают ключевую роль метаболизма сфинголипидов в развитии глиобластомы и открывают новые перспективы для разработки терапевтических подходов, направленных на его модуляцию.

Ключевые слова: глиобластома; перитуморальная ткань; маркеры; метаболомика; ВЭЖХ-МС/МС; сфинголипиды; метаболические пути; генные сети; когнитивная система ANDSystem.

Introduction

Glioblastoma (GBM) is the most prevalent primary brain tumor in adults, with its aggressiveness primarily dictated by its invasive nature – active infiltration of individual or clustered malignant cells into the brain parenchyma surrounding the tumor (Vollmann-Zwerenz et al., 2020). The World Health Organization (WHO) classifies gliomas based on cell type and aggressiveness: grade I includes benign tumors, while grade IV encompasses the most aggressive tumor types, including glioblastomas (Louis et al., 2021). Poor survival rates among GBM patients, even after the most radical surgeries to remove the primary tumor accompanied by multimodal chemoradiotherapy (Omuro, DeAngelis, 2013), are linked to the reappearance of malignant growths. These often occur directly within the postoperative cavity, in its 2-cm marginal zone, or as distant and multiple recurrent tumor foci. Such recurrent tumors are believed to form from GBM cells in the peritumoral zone that re-migrate back into the primary tumor cavity or to distant areas of the brain.

Despite established approaches to disease verification, the challenge of predicting tumor growth and sensitivity to treatment remains unresolved. In 2016, the WHO introduced a new classification system for brain tumors, incorporating genetic markers such as IDH1/IDH2, O-6-methylguanine DNA methyltransferase (MGMT), and epidermal growth factor receptors (EGFR) (Louis et al., 2021). This system enables clinicians to differentiate tumors not only by cell type and aggressiveness, as was possible with previous methods,

but also by the genetic phenotype of neoplastic cells, offering a stronger correlation with tumor prognosis (Jaroch et al., 2021). Molecular biomarkers have become an essential component of glial tumor evaluation, influencing clinical decisions in various glioma subtypes, including treatment strategies. The potential for glioma classification based on molecular markers continues to be explored, promising better implementation of personalized therapeutic approaches (Siegal, 2015). Additionally, the use of omics technologies, such as metabolomic screening, represents an exciting avenue of contemporary research aimed at identifying disease biomarkers.

Like most malignancies, glioblastoma exhibits a unique bioenergetic state of aerobic glycolysis, known as the Warburg effect (Siegal, 2015), in which aerobic glycolysis serves as the primary source of ATP for cancer cells (Warburg, 1956). Although the understanding of cancer cell metabolism is continually evolving, the specific advantages that cancer cells gain from metabolic transformation remain unclear (Koppenol et al., 2011). Additionally, the mechanisms by which hypoxia influences the metabolic reprogramming of tumor cells are not yet fully understood. Recent discoveries of the connections between oncogenes and metabolic processes have reignited interest in Warburg's findings (Poteet et al., 2013). A growing body of evidence suggests that the adaptation of aerobic glycolysis in cancer cells may contribute to biomass accumulation, thereby promoting the proliferation of malignant cells (Heiden et al., 2009). The study of tumor cell metabolism is essential for developing models that accurately reflect the composition of the tumor microenvironment (Liberti, Locasale, 2016) and for identifying new, effective therapeutic strategies. The metabolomic approach to studying glioblastoma has gained significant attention, not only as a diagnostic tool but also as a means to investigate GBM metabolism. Insights from such studies can aid in the development of novel therapeutic interventions (Pandey et al., 2017; Zhou, Wahl, 2019). In some respects, metabolomic analysis surpasses gene expression analysis, as gene function can be influenced by epigenetic modifications and post-translational changes. In contrast, metabolites act as direct indicators of enzymatic activity and biochemical processes within the cell (Pandey et al., 2017).

The analysis of metabolic differences between various regions of glioblastoma, particularly between the central region of the tumor and the peritumoral zone, is considered one of the most reliable methods for studying the tumor's metabolic characteristics (Wolf et al., 2010; Chinnaiyan et al., 2012). These metabolic differences can be assessed using samples obtained intraoperatively during tumor resection (Youngblood et al., 2021). However, only few such studies have been reported in the literature.

Various methods based on the analysis of metabolic pathways and gene networks are employed to identify molecular genetic mechanisms underlying the observed metabolomic data. Gene networks provide valuable insights into the genetic regulation of the identified metabolic pathways, forming a foundation for integrating metabolomic and genomic data (Kolchanov et al., 2013). We have previously developed ANDSystem, a software and information system designed for automated extraction of biological and medical knowledge from scientific publications using artificial intelligence methods (Demenkov et al., 2011; Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022). This software enables users to reconstruct, expand, and graphically visualize gene networks, apply data filtering, and search for regulatory pathways in the global gene network using templates. ANDSystem has been utilized to analyze molecular genetic mechanisms across a wide range of diseases, including comorbid conditions, organismal responses to stress, identification of pharmacological targets, and other research objectives (Bragina et al., 2014, 2016, 2023; Popik et al., 2016; Saik et al., 2016, 2018a, b, 2019; Zolotareva et al., 2019; Antropova et al., 2022; Demenkov et al., 2023).

Metabolomic and proteomic data have been analyzed using ANDSystem (Pastushkova et al., 2013, 2019; Binder et al., 2014; Larina et al., 2015; Rogachev et al., 2021; Ivanisenko V.A. et al., 2022, 2023). For instance, metabolomic analysis of blood plasma from COVID-19 patients identified the role of non-structural viral proteins in metabolic disorders associated with the disease, which contributed to changes in the metabolomic profile (Ivanisenko V.A. et al., 2022). Additionally, analysis of metabolomic profiles from patients with postoperative delirium, conducted using ANDSystem, helped identify potential markers represented by several sphingolipids and revealed molecular genetic mechanisms underlying their metabolic disruptions (Ivanisenko V.A. et al., 2023).

In this study, a targeted screening of a broad spectrum of metabolites was conducted in glioblastoma and peritumoral tissue. Statistical analysis of the screening data identified metabolites involved in sphingolipid metabolism, with significantly different levels observed between tumor and peritumoral tissue. Using gene network reconstruction, genes with the greatest regulatory influence on the function and expression of key enzymes in sphingolipid metabolism were identified. These included both established tumor markers (p53, TNF-α, VEGF, etc.) and promising candidate markers (KLF4, E2F4, etc.). Disruption of these genes' functions in glioblastoma may explain the observed alterations in the metabolomic profile.

Materials and methods

Reagents and materials. Methanol and acetonitrile used for sample preparation and analysis were of gradient HPLC grade and were purchased from Khimmed (Moscow, Russia). Purified water was prepared using a Sartorius arium 611DI system (Göttingen, Germany). Eluent A was prepared according to the protocol described by Li et al. (2017).

Patients – study participants. Tumor tissue was obtained from patients who underwent surgical treatment in the neurosurgical department of the Ya.L. Tsivyan Novosibirsk Research Institute of Traumatology and Orthopedics (Novosibirsk, Russia) for first-diagnosed GBM between 2019 and

2022. The cohort included patients with grade IV gliomas hospitalized for surgical tumor resection. Diagnoses were confirmed by MRI and histopathological examination of excisional biopsy specimens. The final diagnosis was established based on histological analysis and the consensus of two pathomorphologists, following the WHO classification. Tumor samples were collected intraoperatively and anonymized for the investigators. The clinical study identifier was NCT03865355. A total of 17 patients were included in the study. Below is their gender and age distribution:

Sex (M/F)	Age, years old				
	Min	Max	Average	Median	Std. deviation
8/9	28.2	69.6	54.9	63.22	15.9

Collection of tumor tissue samples from patients. Tumor tissue samples were obtained during cytoreductive surgical interventions. After collection, the samples were immediately placed in RPMI 1640 cell culture medium without additives and stored at +4 °C until processing. Tumor sections from different regions (tumor center and peritumoral tissue) were separated using surgical instruments into fragments ranging in size from 2×2×2 mm to 5×5×5 mm. These fragments were wrapped in sterile foil bags, frozen in liquid nitrogen, and subsequently stored in a low-temperature freezer at –80 °C. In total, glioblastoma samples from 17 patients and peritumoral tissue samples from 15 patients were included in the study.

Compliance with ethical standards. The study was reviewed and approved by the Ethics Committee of the Zelman Institute of Medicine and Psychology at Novosibirsk State University (meeting minutes dated January 4, 2018). All experimental protocols were approved, and all procedures involving human participants adhered to the ethical standards of the institutional research committee, the 1964 Declaration of Helsinki, and its subsequent amendments or equivalent ethical standards. Written informed consent was obtained from each participant prior to inclusion in the study. Additionally, the study was approved by the Local Ethical Committee of the Ya.L. Tsivyan Novosibirsk Research Institute of Traumatology and Orthopedics (meeting minutes dated September 11, 2017, No. 050/17), and informed voluntary consent was obtained from all participants.

Sample preparation of glioblastoma and peritumoral tissue samples. Metabolite extraction from glioblastoma and peritumoral tissue samples was carried out simultaneously using a modified protocol based on (Yuan et al., 2012; Li et al., 2017). In a 1.5 mL tube, 250 μL of chilled 80 % methanol (vol/vol) was added per 10 mg of tissue to a sample weighing between 9 and 33 mg. The samples were homogenized for 2 minutes using a Bertin Minilys tissue homogenizer (Rockville, Maryland, USA), with granite chips added to enhance sample disintegration. This was followed by incubation at –70 °C for 24 hours. After incubation, the samples were vortexed and centrifuged at +4 °C and 16,000 g for 15 minutes. The resulting supernatant was carefully transferred to a new polypropylene tube. An equal volume of chilled 80 % methanol (vol/vol) was then added to the remaining precipitate, vortexed for 1 minute, incubated at –70 °C for 30 minutes, and centrifuged under the same conditions. The supernatants from both extractions were combined, and a 500 μL aliquot was taken and evaporated to dryness using a SpeedVac concentrator vacuum centrifuge (Thermo Fisher Scientific/Savant, Waltham, USA). The dried samples were reconstituted in 20 μL of MilliQ water and subjected to analysis.

High-performance liquid chromatography with mass spectrometric detection. Samples were analyzed by high-performance liquid chromatography with tandem mass spectrometric detection (LC-MS/MS), following the procedure described by Basov, Rogachev et al. (2024). Chromatographic separation was performed using an LC-20AD Prominence chromatograph (Shimadzu, Japan) equipped with a SIL 20AC autosampler (Shimadzu, Japan) maintained at 10 °C. The injection volume was 2 μL. Eluent A consisted of 5 % acetonitrile in 20 mM ammonium carbonate $(NH_4)_2CO_3$ aqueous solution, adjusted to pH 9.8 with aqueous ammonia solution, and eluent B was 100 % acetonitrile. Each sample was analyzed twice, in hydrophilic interaction liquid chromatography (HILIC) and reversed-phase chromatography (RP LC) modes. The HILIC gradient was as follows: 0 min – 98 % B, 2 min – 98 % B, 6 min – 0 % B, 10 min – 0 % B, followed by column equilibration for 4 min. The RP LC gradient was 0 min – 0 % B, 1 min – 0 % B, 6 min – 98 % B, 16 min – 98 % B, with column equilibration for 3 min. The flow rate for both analyses was 300 μL/min. Chromatographic analyses were performed using a monolithic column based on 1-vinyl-1,2,4-triazole (2 × 60 mm), synthesized as described by Patrushev et al. (2018) through copolymerization of styrene, divinylbenzene, and 1-vinyl-1,2,4-triazole in a volume ratio of 10:50:40, respectively, within a glass tube with an inner diameter of 2 mm.

Mass spectrometric detection. Detection of 489 metabolites was performed in multiple reaction monitoring (MRM) mode as positive and negative ions using an API 6500 QTRAP mass spectrometer (AB SCIEX, USA) equipped with an electrospray ionization source operating in the positive/negative switch mode. The primary mass spectrometric parameters were as follows: ion spray voltage (IS) was set at 5500 V for positive ionization mode and –4500 V for negative ionization mode; the ion source temperature was at 475 °C; CAD gas was set as "Medium"; GS1, GS2 and curtain gas were 33, 33 and 30 psi, respectively. The declustering potential (DP) was ±91 V, the entrance potential (EP) was ±10 V, and the collision cell exit potential (CXP) was ±9 V. In addition, the polarity switching (settling) time was set at 5 ms, and dwell time was 3 ms for each MRM transition. Precursor and fragment ion transitions, metabolite names, dwell times, and the appropriate collision energies for both positive and negative ion modes were adapted from the studies: Yuan et al. (2012) and Li et al. (2017) (Supplementary Material 1)[1]. Device control and data acquisition were collected using Analyst 1.6.3 software (AB SCIEX),

[1] Supplementary Materials 1–5 are available at: https://vavilovj-icg.ru/download/pict-2024-28/appx31.xlsx

Table 1. Templates of molecular genetic pathways regulating enzymes in metabolic pathways by human proteins

No.	Template name	Scheme of regulatory pathway template
P1	Protein-protein interactions	Hp – protein-protein interactions → Kp
P2	Regulation of protein function	Hp – regulation of activity/degradation/proteolysis/transportation → Kp
P3	Regulation of expression	Hp – regulation of expression → Kg → expression → Kp

Note. Hp – human proteins; Kg – genes encoding enzymes of the KEGG metabolic pathway; Kp – enzymes of the KEGG metabolic pathway.

while chromatograms were processed using Skyline 24.0 software (Adams et al., 2020).

Pre-processing and statistical analysis of the data. Statistical analysis of the metabolomic screening results for glioblastoma and peritumoral tissue was conducted using the Mann–Whitney and Cucconi tests, implemented in Python packages (SciPy and Nonparstat). The Mann–Whitney test was employed to identify significant differences between groups, while the Cucconi test provided additional validation of the identified differences under conditions of sample heterogeneity. Outlier correction was performed as follows: an outlier was defined as any value outside the 1.5 interquartile range (IQR). Identified outliers were replaced with adjusted values calculated as $1.5 \times IQR \pm 10^{-5}$ (subtracted for upper outliers and added for lower outliers).

The online platform MetaboAnalyst 6.0 (http://www.metaboanalyst.ca/) (Pang et al., 2021) and its Enrichment Analysis tool were used to identify overrepresented metabolic pathways based on highly significant metabolites.

Reconstruction of gene networks. Gene network reconstruction was performed using the ANDVisio graphical user interface within the ANDSystem software and information system (http://www-bionet.sscc.ru/andvisio/). In the Pathway Wizard module of ANDVisio, templates of regulatory pathways for enzymes involved in the identified metabolic pathways were created using human protein data (Table 1). The list of human protein identifiers was obtained from the SwissProt database (https://www.uniprot.org/).

The list of proteins associated with glioblastoma was retrieved from the ANDSystem knowledge base. Overrepresentation analysis of Gene Ontology biological processes was performed using the DAVID web service (https://david.ncifcrf.gov/) with default settings.

Results

Samples of glioblastoma and brain tissue adjacent to the tumor (17 glioblastoma and 15 peritumoral tissue samples) were collected and analyzed as part of the study. Metabolomic analysis was performed using the LC-MS/MS approach developed previously (Basov et al., 2024). The chromatograms were processed by integrating the peak area of each metabolite, and the resulting signals were compared between glioblastoma and peritumoral tissue samples. Peak area values for 446 metabolites were obtained from the analysis.

The Mann–Whitney test, with a critical value of $p < 0.05$, was used as the primary method for statistical analysis of the metabolomic screening results. The nonparametric Cucconi test was employed as an additional method for group comparisons. Metabolite lists satisfying each test at $p < 0.05$ were compared, resulting in a subset of metabolites that met the criteria for both tests (Table 2).

The overrepresentation of KEGG metabolic pathways was analyzed for the identified set of metabolites using MetaboAnalyst 6.0 (Table 3). This analysis revealed sphingolipid metabolism as a statistically significant overrepresented metabolic pathway. Another marker-enriched pathway, the KEGG metabolic pathway "Pantothenic acid and CoA biosynthesis", had a *p*-value of 0.012; however, after correction for multiple comparisons, the *p*-value exceeded the significance threshold of 0.05, resulting in a corrected *p*-value of 0.46.

Among the metabolites identified as potential markers (Table 2), 5 out of 22 (~23 %) belonged to ceramides, a class of lipid molecules that are key components of cell membranes. Additionally, 3 metabolites – 4-phosphopantothenic acid, malonyl-CoA, and coenzyme A – were identified as major precursors in *de novo* lipid biosynthesis. The ceramide content was at least twofold higher in tumor tissue compared to peritumoral tissue. Furthermore, the variance was significantly greater in the glioblastoma samples, indicating higher heterogeneity within this group (Fig. 1).

The levels of metabolites in the pantothenic acid and CoA synthesis pathway were significantly lower in tumor tissues (Fig. 2), suggesting their active utilization in lipid biosynthesis.

In the next phase of our study, we investigated potential mechanisms underlying the dysfunction of sphingolipid metabolism enzymes. To achieve this, molecular genetic regulatory pathways were reconstructed using ANDSystem with the templates presented in Table 1. These templates represent the potential regulation of enzymes involved in sphingolipid metabolism by human proteins (Supplementary Materials 2–4). The starting point of the reconstructed regulatory pathways included all human proteins, while the endpoint comprised sphingolipid metabolism enzymes

Table 2. Metabolites with significant differences ($p < 0.05$) between glioblastoma and peritumoral samples

No.	Metabolite	*p*-value (Cucconi test)	*p*-value (Mann–Whitney test)
1	Malonyl-CoA	9.99×10^{-4}	1.17×10^{-4}
2	SM (d18:1/22:0 OH)	3.00×10^{-3}	2.52×10^{-3}
3	Octanoylcarnitine	8.99×10^{-3}	4.10×10^{-3}
4	Pyroglutamic acid	2.00×10^{-3}	4.62×10^{-3}
5	Ceramide (d18:1/16:0 OH)*	3.00×10^{-3}	4.62×10^{-3}
6	3-Phosphoglyceric acid	1.20×10^{-2}	5.20×10^{-3}
7	THC 18:1/20:0	1.60×10^{-2}	7.33×10^{-3}
8	Hexose Disaccharide Pool	1.30×10^{-2}	8.20×10^{-3}
9	4-Phosphopantothenate	2.80×10^{-2}	8.20×10^{-3}
10	2-Octenoylcarnitine	2.80×10^{-2}	4.10×10^{-3}
11	Ceramide (d18:1/16:0)*	4.00×10^{-3}	9.17×10^{-3}
12	Ceramide (d18:1/22:0)*	8.99×10^{-3}	1.14×10^{-2}
13	Coenzyme A	4.00×10^{-2}	1.27×10^{-2}
14	Pyridoxal	3.10×10^{-2}	1.41×10^{-2}
15	N-carbamoyl-L-aspartate	4.40×10^{-2}	1.74×10^{-2}
16	Citrulline	2.50×10^{-2}	1.92×10^{-2}
17	Decanoylcarnitine	4.70×10^{-2}	2.12×10^{-2}
18	GC (18:2/16:0)	2.40×10^{-2}	2.35×10^{-2}
19	Purine	2.00×10^{-2}	3.45×10^{-2}
20	Ceramide (d18:1/16:2)*	3.20×10^{-2}	4.14×10^{-2}
21	Ceramide (d18:1/16:1 OH)*	2.50×10^{-2}	4.53×10^{-2}
22	Glycerophosphocholine	3.00×10^{-2}	4.95×10^{-2}

* Metabolites belonging to the ceramide class.

Table 3. Overrepresented KEGG metabolic pathways for a set of metabolomic markers

Metabolic pathway (KEGG)	*p*-value	FDR*
Metabolism of sphingolipids	7.95×10^{-5}	6.36×10^{-3}
Biosynthesis of pantothenic acid and CoA	11.5×10^{-3}	4.6×10^{-1}

* FDR (False Discovery Rate) represents correction for multiple comparisons.

from the KEGG database involved in the metabolism of ceramide, sphingomyelin (SM), glucosylceramide (GC), and trihexosylceramide (THC). For the purposes of this study, these enzymes are referred to as key enzymes of sphingolipid metabolism. The regulatory pathways considered included interactions such as protein-protein interactions, regulation of gene expression, and regulation of protein activity, degradation, or transport.

The reconstructed regulatory pathways were integrated into a unified gene network (Fig. 3). This regulatory gene network comprised 15 genes, 329 proteins (including 35 enzymes involved in sphingolipid metabolism), and 389 interactions among them.

According to the ANDSystem knowledge base, evidence from the literature indicates dysfunction in glioblastoma for 119 out of 294 gene network proteins regulating key enzymes of sphingolipid metabolism. A subnetwork of the regulatory gene network, illustrating the interactions of these proteins with key enzymes of sphingolipid metabolism, is presented in Figure 4.

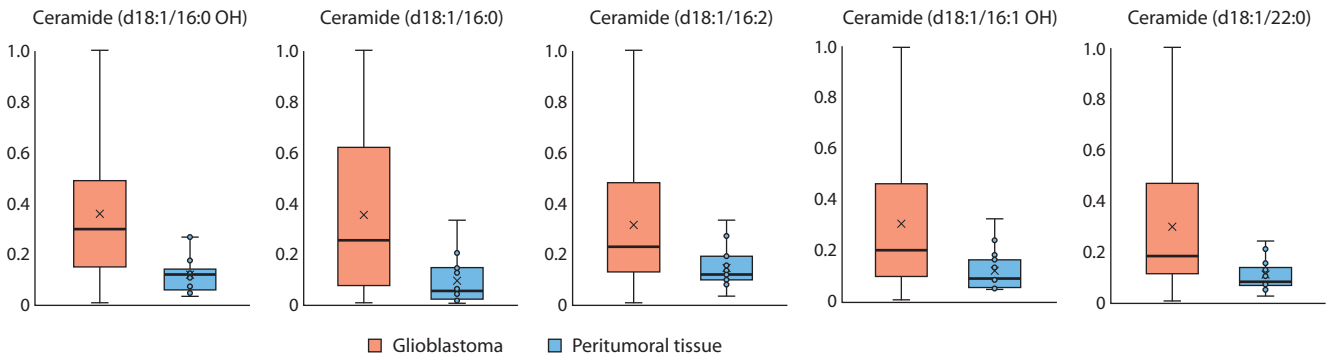


Fig. 1. Levels of marker ceramides in tumor and peritumoral tissues.

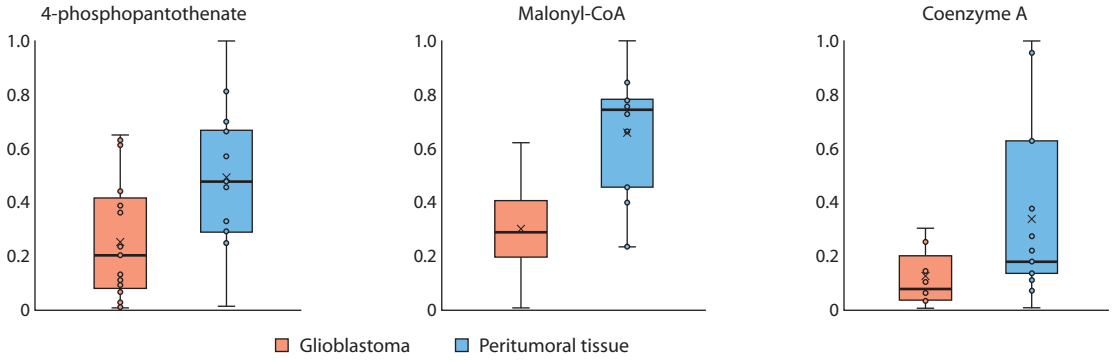


Fig. 2. Levels of CoA and related metabolites in tumor and peritumoral tissues.

In total, the ANDSystem knowledge base contains association information for 2,393 human glioblastoma-related proteins, 119 of which were included in the regulatory gene network. Based on a hypergeometric test, the reconstructed gene network is statistically significantly associated with glioblastoma (p-value < 10^{-35}).

Gene Ontology overrepresentation analysis of biological processes for genes in the resulting gene network identified 184 statistically significant processes. These include apoptosis, the NF-κB signaling pathway, proliferation, migration, and angiogenesis, which are commonly dysregulated in many cancers, as well as pyroptosis – a process, the role of which in glioblastoma is currently under active investigation (Supplementary Material 5).

Discussion

Susceptibility of glioblastoma cells to changes in coenzyme A metabolite levels

Our study identified reduced levels of CoA and malonyl-CoA in glioblastoma tissues compared to peritumoral tissues (Fig. 2). For *de novo* fatty acid synthesis, glioblastoma cells must produce cytosolic acetyl-CoA, which can be generated either from citrate via ATP-citrate lyase or from acetate via acetyl-CoA synthetase (Santos, Schulze, 2012). Mashimo et al. (2014) demonstrated that brain tumors of various cellular origins have the ability to oxidize injected acetate. The authors suggest that acetate oxidation is facilitated by the activation of acetyl-CoA synthetase isoform ACSS2, achieved through upregulated expression. The higher expression of ACSS2 in glioblastoma compared to lower-grade gliomas supports the hypothesis that enzyme activation is associated with increased acetate oxidation by the tumor. Furthermore, ACSS2 deficiency in mouse models of hepatocellular carcinoma has been shown to reduce tumor burden and inhibit tumor growth (Comerford et al., 2014).

Malonyl-CoA level determines the direction of fatty acid metabolism, specifically whether it supports triglyceride synthesis or oxidation (Clarke S.D., Nakamura, 2004). Previous studies reported that inhibition of β-oxidation in human glioblastoma cells by etomoxir, a carnitine palmitoyltransferase-1 inhibitor, significantly reduces ATP, NADPH, and reduced glutathione levels, thereby impairing cell viability (Pike et al., 2011). These findings suggest that β-oxidation contributes to oxidative stress resistance in glioblastoma cells, and our results support this hypothesis. Additionally, malonyl-CoA level has been shown to influence the response to various chemotherapeutic agents. For instance, in a study on a breast cancer cell model, malonyl-CoA levels significantly increased following fatty acid synthase inhibition and decreased upon inhibition of acetyl-CoA carboxylase (Pizer et al., 2000). Key metabolic pathways,

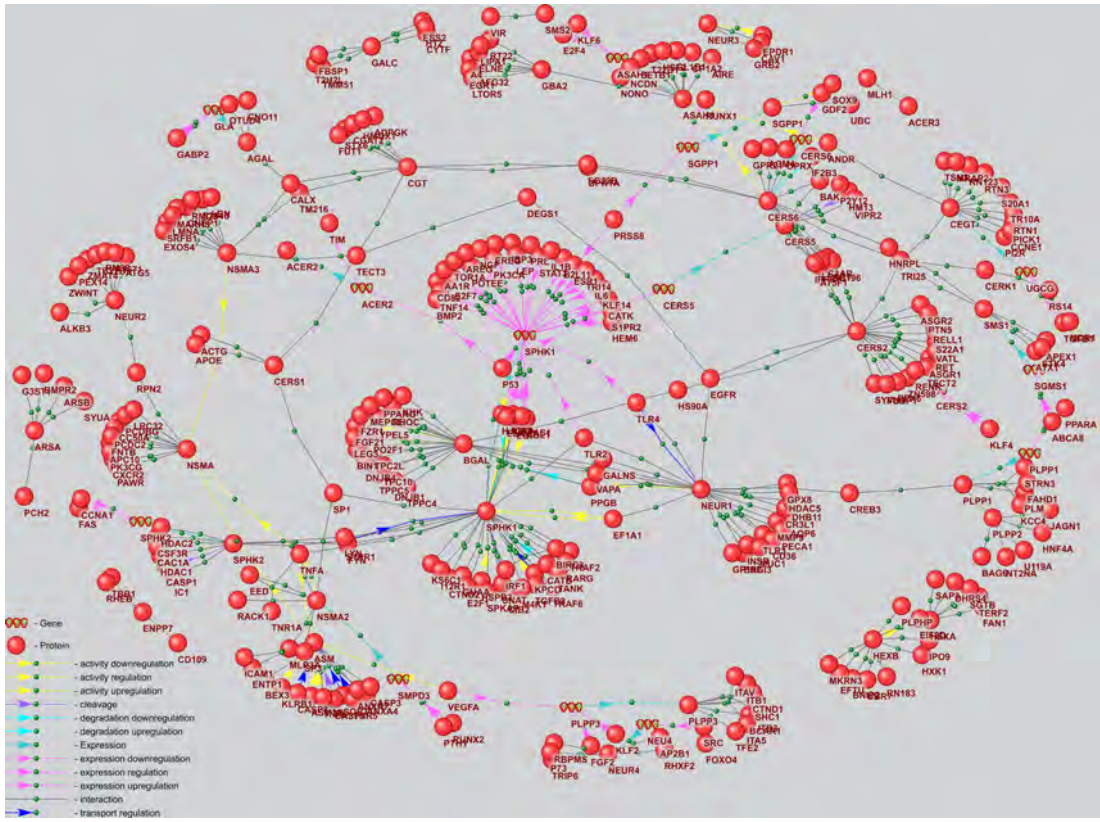


Fig. 3. Gene regulatory network of key enzymes in sphingolipid metabolism, reconstructed by integrating regulatory pathways based on three types of templates.

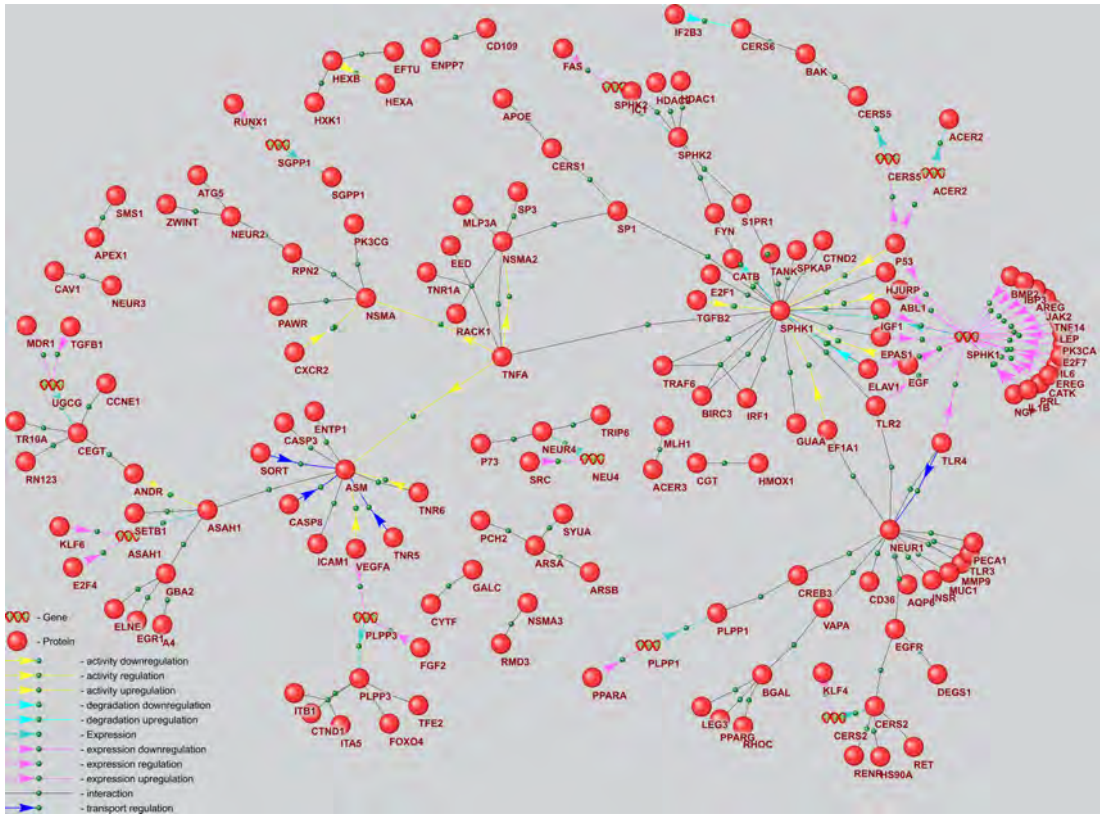


Fig. 4. Subnetwork of the gene regulatory network for key enzymes of sphingolipid metabolism regulated by glioblastoma genetic markers.

such as glycolysis and the tricarboxylic acid (TCA) cycle, are regulated by multiple microRNAs that control specific steps within these pathways. Cancer cells predominantly rely on aerobic glycolysis instead of the TCA cycle, enabling them to sustain high ATP levels to meet biosynthetic demands (Chan et al., 2015).

Our study revealed an increase in the level of 4-phosphopantothenate, consistent with the observed changes in the lipid profile. The synthesis of this metabolite is catalyzed by pantothenate kinase, the first enzyme in the CoA biosynthetic pathway. The role of pantothenate kinase in glioblastoma has been extensively discussed in the literature. For instance, Poli et al. (2010) reported that silencing pantothenate kinase-2 significantly reduced the growth of the U373 glioma cell line. Acetyl-CoA and lipid levels may also be regulated by the microRNAs miR-103 and miR-107 (Wilfred et al., 2007). Additionally, evidence suggests that miR-103 suppresses glioblastoma cell proliferation and migration (Chen L.P. et al., 2018), while miR-107 inhibits glioblastoma angiogenesis by upregulating its expression (Chen L. et al., 2016).

Linking ceramide biosynthesis to tumor growth

Ceramides, lipid mediators of the sphingolipid class, play a role in signaling pathways that regulate cell proliferation, differentiation, and cell death (Riboni et al., 2002). Our study demonstrates that the levels of ceramides (16:0), (16:0 OH), (16:2), (16:1 OH), and (22:0) – derivatives of sphingomyelin (18:1) – are elevated in tumor tissue compared to peritumoral tissue (Fig. 1). Peritumoral tissue was used as a control because its collection during surgery does not compromise the treatment prognosis for patients. The observed increase in ceramide levels in tumor tissue suggests alterations in the enzymatic systems responsible for ceramide biosynthesis and degradation, potentially contributing to tumor growth and the evasion of apoptosis by tumor cells.

Ceramide formation occurs via three main pathways. The sphingomyelinase pathway involves the action of sphingomyelinase, an enzyme that cleaves sphingomyelin in the cell membrane to release ceramides. In the *de novo* synthesis pathway, ceramides are produced from simpler precursor molecules through a series of enzymatic reactions. The salvage pathway reutilizes sphingolipids by cleaving them into sphingosine, which is subsequently realkylated to form ceramide.

The key enzyme in the sphingomyelinase pathway is sphingomyelinase (SMase), which catalyzes the hydrolysis of sphingomyelin. As sphingomyelin is one of the most abundant phospholipids in the cell membrane, this pathway's significance lies in its role in targeting the cell membrane for extracellular signals that trigger programmed cell death and cellular stress (Haimovitz-Friedman et al., 1994). SMase exists in three main types: acidic (aSMase), neutral (nSMase), and alkaline (alk-SMase). Stimulation of SMase activity can be induced by various factors, including antitumor drugs. Sphingomyelinase inhibitors, such as perphenazine and fluphenazine – classified as functional inhibitors of acidic sphingomyelinase (FIASMA) – show potential in cancer therapy, though further studies are needed to validate their efficacy (Kornhuber et al., 2010). Recent research has identified an inhibitor, Arc39, that blocks lysosomal and secretory aSMase *in vitro* in L929, HepG2, and B16 cells (Naser et al., 2020), as well as a light-inducible PCAI inhibitor capable of inhibiting aSMase (Prause et al., 2020). Additionally, sphingomyelinase plays a critical role in sphingolipid metabolism, which may influence cancer development (Clarke C.J. et al., 2011). Its inhibition and subsequent effects on exosomes are of growing interest for oncology and the development of therapeutic strategies (Lin M. et al., 2018).

The sphingomyelin synthase (SMS) family, comprising three members – SMS1, SMS2, and SMS-related protein (SMSr) (Chen Y., Cao, 2017) – catalyzes the synthesis of sphingomyelins from ceramides (Cer) and phosphatidylcholine, releasing diacylglycerol as a byproduct. Selective inhibition of SMS has been shown to increase ceramide concentration in the endoplasmic reticulum, triggering autophagy in hippocampal neurons (Gulbins et al., 2018). In glioblastoma, treatment with 2-hydroxyoleic acid, an antitumor drug, was observed to enhance SMS activity. Activation of SMS2 decreases ceramide levels and promotes cell proliferation via the transforming growth factor-β (TGF-β)/Smad signaling pathway. Conversely, inhibition of SMS2 by specific miRNAs led to ceramide accumulation and accelerated cell death (Zheng et al., 2019). Recent research has shown that SMS2 is activated in breast cancer, inducing macrophage polarization and promoting tumor progression (Deng et al., 2021). Notably, SMS2 knockdown reduced the release of cytokines that drive macrophage polarization into M2 macrophages, thereby suppressing tumor growth (Deng et al., 2021). Furthermore, downregulation of SMS1 has been reported in patients with metastatic melanoma, where it is associated with worse prognosis due to an imbalance between sphingomyelin and glucosylceramide levels (Bilal et al., 2019).

Serine palmitoyltransferase (SPT) is a three subunits heteromeric enzyme that catalyzes the first step of *de novo* ceramide synthesis by condensing L-serine and palmitoyl-coenzyme A to form 3-ketosfinganine. Increased SPT activity has been observed in response to chemotherapy and radiotherapy across various cancers. Several SPT inhibitors that block tumor growth have been identified. For instance, myriocin (ISP-1), a potent SPT inhibitor (Glaros et al., 2007), has been shown to suppress the growth of breast cancer cells (Ogretmen, 2018) and B16F10 melanoma cells by arresting the G2/M phase (Lee et al., 2011). Similar effects have been observed in human lung adenocarcinoma (HCC4006) cells, where SPT inhibition correlates with growth suppression (Sano et al., 2017). Furthermore, SPT inhibition by myriocin or specific miRNAs reduced U87MG glioblastoma cell proliferation by suppressing intracellular S1P levels (Bernhart et al., 2015). This antitumor activity is believed to result from increased levels of pro-apoptotic ceramides. In some cases, SPT activation

contributes to therapeutic efficacy; for example, fenretinide, a synthetic retinoid, elevates desaturated ceramide levels, inducing apoptosis in neuroblastoma cells (Maurer et al., 1999).

Ceramide synthase (CerS) plays a role in both *de novo* ceramide synthesis and the salvage pathway. The CerS family comprises six isoforms, each synthesizing ceramides with specific fatty acyl-CoA chain lengths, which determine their biological activity. For instance, CerS1 produces ceramides (18:0), which inhibit tumor growth (Wang Z. et al., 2017), while CerS5 and CerS6 generate ceramides (16:0), which are associated with anti-apoptotic effects in head and neck squamous cell carcinoma (Moro et al., 2019). The CerS1-specific inhibitor P053 reduces ceramide (18:0) levels in HEK 293 cells (Turner et al., 2018). Fingolimod-derived analogues (FTY720) selectively inhibit specific CerS isoforms, with inhibitors such as ST1058 and ST1074 targeting CerS2 and CerS4, while ST1072 blocks CerS4 and CerS6 activity, and ST1060 inhibits CerS2 (Schiffmann et al., 2012).

Dihydroceramide desaturase (Des1, DEGS1) is the final enzyme in *de novo* ceramide synthesis, converting dihydroceramide into ceramide by introducing a trans double bond at the C_4-C_5 position. Knockdown of Des1 by miRNAs results in cell cycle arrest in neuroblastoma cells (Kraveka et al., 2007). Resveratrol, a polyphenol with antioxidant properties, inhibits Des1 and induces autophagy in HGC27 gastric cancer cells (Signorelli et al., 2009). Other Des1 inhibitors, such as γ-tocotrienol, phenoxodiol, and celecoxib, promote autophagy by causing dihydroceramide accumulation in glioblastoma cell lines (T98G and U87MG) through Des1 inhibition (Signorelli et al., 2009). A specific Des1 inhibitor, N-[(1R,2S)-2-hydroxy-1-hydroxymethyl-2-(2-tridecyl-1-cyclopropenyl)ethyl]octanamide, effectively activates autophagy and apoptosis in U87MG glioblastoma cells. Additionally, treatment with tetrahydrocannabinol alters the lipid composition of the endoplasmic reticulum, leading to dihydroceramide accumulation and stimulating autophagy and apoptosis in U87MG cells through reduced Des1 expression (Hernández-Tiedra et al., 2016).

Glucosylceramide synthase (GCS) is a lysosomal enzyme that glycosylates ceramides to form glycosylceramides. Elevated GCS levels have been observed in various cancers and are associated with resistance to antitumor therapies (Madigan et al., 2020).

Ceramidase is an enzyme that hydrolyzes ceramides, removing fatty acid residues to produce sphingosines. Overexpression of acidic ceramidase (ASAH1) has been detected in melanomas and is likely linked to chemotherapy resistance. ASAH1 has also been implicated in mitochondrial function and cellular autophagy in melanoma cells (Lai M. et al., 2021). Alk-SMase has been reported to play a significant role in tumor cell growth, migration, and invasion (Zhang et al., 2020). Structural analogues of ceramides have shown efficacy as selective ceramide synthase inhibitors, inhibiting cell growth and emerging as promising candidates for antitumor treatments (Steiner et al., 2016).

Disruption of genetic regulation of sphingolipid metabolism in glioblastoma

The application of ANDSystem enabled the reconstruction of a gene network describing the regulation of key enzymes involved in sphingolipid metabolism (Fig. 3). Analysis of this regulatory network revealed that 119 of its proteins are associated with glioblastoma, confirming the significant connection between the reconstructed network and this disease (p-value < 10^{-35}). Among the most extensively studied glioblastoma-associated proteins included in the network are p53, TNF-α, TGF-β, VEGF, KLF4, and E2F4.

It is well-established that p53 is involved in numerous intracellular processes, and its dysfunction is commonly observed in various cancers. Notably, p53 plays a critical role in sphingolipid metabolism, regulating the activity of five key enzymes (CerS5, CerS6, SMPD3, ACER2, SPHK1) out of the 35 enzymes represented in the reconstructed gene network. According to Lacroix et al. (2020), p53 in tumor cells increases the expression of ceramide synthases 5 (CerS5) and 6 (CerS6) and neutral sphingomyelinase 2 (SMPD3), which are ceramide-synthesizing enzymes. Additionally, the induction of alk-SMase-2 transcription by p53 was investigated in studies by Wang Y. et al. (2017) and Xu et al. (2018).

According to the regulatory gene network, tumor necrosis factor-alpha (TNF-α) stimulates the activity of three enzymes involved in sphingolipid metabolism: acidic sphingomyelinase (ASM), neutral sphingomyelinase (NSMA), and neutral sphingomyelinase 2 (NSMA2). By enhancing the activity of these enzymes, TNF-α may facilitate sphingomyelin hydrolysis and promote ceramide formation.

Transforming growth factor-beta (TGF-β) plays a crucial role in various cell types. It initiates cellular signaling cascades that activate downstream substrates and regulatory proteins, ultimately inducing the transcription of multiple target genes. Within the regulatory gene network, TGF-β2 has been shown to enhance the activity of sphingosine kinase-1 (SPHK1), a finding supported by Ren et al. (2009).

Vascular endothelial growth factor (VEGF) has been identified as a component of the tumor microenvironment with the capacity to activate endothelial cells. VEGF signaling operates through tyrosine kinase receptors VEGFR1 and VEGFR2, promoting endothelial cell migration, survival, proliferation, and differentiation. This process initiates angiogenesis, tumor growth, and metastasis. Within the regulatory gene network, VEGF is involved in suppressing acidic sphingomyelinase (ASM) activity and regulating the expression of phospholipid phosphatase 3 (PLPP3). Glioblastoma is characterized by a high degree of vascularization and VEGF overexpression, making this gene a compelling target for glioblastoma therapy (Tea et al., 2020).

Kruppel-like factor 4 (KLF4) is involved in regulating proliferation, differentiation, apoptosis, and somatic cell reprogramming. Evidence also indicates that KLF4 functions as a tumor suppressor in certain cancers (El-Karim et al., 2013). Within the regulatory gene network, KLF4 modulates the expression of the ceramide synthase 2 (CerS2) gene.

Chromatin immunoprecipitation analysis demonstrated that KLF4 directly binds to the promoter region of *CerS2*, activating its expression (Fan et al., 2015).

According to the reconstructed connections in the gene network, the E2F4 protein regulates the expression of ASAH1. Literature evidence indicates that E2F4 functions as a transcriptional repressor, playing a crucial role in suppressing genes associated with proliferation. Mutations and overexpression of the E2F4 gene have been linked to human cancers. By binding to the promoter region of the ASAH1 gene, E2F4 suppresses its expression (Melland-Smith et al., 2015).

Significant biological processes associated with the gene network

The overrepresented biological processes involving participants of the regulatory gene network (Supplementary Material 5) can be grouped into several categories, including programmed cell death, cell mobility, angiogenesis, and proliferation – all of which are well-documented in the context of cancer (Hanahan, Weinberg, 2000). Among these, programmed cell death via pyroptosis has garnered particular interest in recent years due to its potential role in the development and progression of glioblastoma (Lin J. et al., 2022). In the gene network, pyroptosis is represented by several caspases (CASP1, CASP3, and CASP8) and neutrophil elastase, which, under specific conditions, cleaves Gasdermin D (GSDMD) to activate pyroptosis or cleaves GSDMB, thereby inhibiting pyroptosis (Kambara et al., 2018; Oltra et al., 2023). These pyroptosis-related proteins in the gene network play a significant role in regulating this process (Rao et al., 2022) (Supplementary Material 5).

Angiogenesis is essential for providing nutrients and oxygen to glioblastoma, supporting tumor growth (Lara-Velazquez et al., 2017). Key members of the gene network, including vascular endothelial growth factor A (VEGFA), epidermal growth factor (EGF), and the catalytic subunit A of phosphatidylinositol-4,5-bisphosphate-3-kinase (PIK3CA), have been identified as important genetic markers of glioblastoma involved in angiogenesis. These genes are significant drivers of the angiogenic process (Danielsen, Rofstad, 1998). Glioblastoma is also characterized by a high capacity for invasion, with tumor cells infiltrating surrounding brain tissue, making complete surgical removal challenging and often impossible (Vollmann-Zwerenz et al., 2020). The regulatory gene network highlights proteins such as thyroid receptor-interacting protein 6 (TRIP6), which is overexpressed in glioblastoma and promotes tumor cell invasion (Lai Y.-J. et al., 2010), as well as TGF-β1, integrin alpha-V (ITAV), and cyclic AMP-responsive element-binding protein 3 (CREB3). These proteins are associated with cell migration and contribute to glioblastoma's invasive properties.

Conclusion

A targeted metabolomic screening of glioblastoma and peritumoral tissues from cancer patients was conducted using the LC-MS/MS method. Bioinformatic analysis of the resulting metabolic profiles, employing statistical methods and gene network reconstruction, provided valuable insights into the mechanisms underlying glioblastoma development and progression. The study revealed altered metabolism of coenzyme A (CoA) and related metabolites in glioblastoma tissues, distinguishing them from peritumoral cells. Reduced levels of CoA and malonyl-CoA in glioblastoma tissues suggest increased β-oxidation of fatty acids and enhanced resistance to oxidative stress in glioblastoma cells.

Additionally, elevated ceramide levels in tumor tissue indicate potential modifications in the enzymatic activity involved in ceramide synthesis and degradation, which may be linked to tumor growth. These findings suggest that disruptions in lipid metabolism, particularly involving CoA and ceramide pathways, play a crucial role in glioblastoma pathogenesis. Such alterations highlight potential therapeutic targets for developing novel treatments aimed at the disrupted metabolic pathways in tumor cells. In particular, inhibition of key enzymes, such as serine palmitoyltransferase and sphingomyelinase, emerges as a promising strategy to reduce cell viability and potentially prevent further growth of glioblastoma cells.

Thus, the findings of this study enhance our understanding of the metabolic characteristics of glioblastoma and offer new opportunities for developing targeted therapeutic strategies focused on disrupting lipid metabolism in tumor cells. Future research on specific metabolic alterations across different glioblastoma subtypes, alongside the development and evaluation of inhibitors targeting key enzymes, could contribute significantly to advancing treatment options for this disease.

References

Adams K.J., Pratt B., Bose N., Dubois L.G., St. John-Williams L., Perrott K.M., Ky K., Kapahi P., Sharma V., Maccoss M.J., Moseley M.A., Colton C.A., Maclean B.X., Schilling B., Thompson J.W. Skyline for small molecules: a unifying software package for quantitative metabolomics. *J. Proteome Res*. 2020;19(4):1447-1458. doi 10.1021/acs.jproteome.9b00640

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Lavrik I.N., Ivanisenko V.A. Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. doi 10.18699/VJGB-22-89

Basov N.V., Rogachev A.D., Aleshkova M.A., Gaisler E.V., Sotnikova Y.S., Patrushev Y.V., Tolstikova T.G., Yarovaya O.I., Pokrovsky A.G., Salakhutdinov N.F. Global LC-MS/MS targeted metabolomics using a combination of HILIC and RP LC separation modes on an organic monolithic column based on 1-vinyl-1,2,4-triazole. *Talanta*. 2024;267:125168. doi 10.1016/j.talanta.2023.125168

Bernhart E., Damm S., Wintersperger A., Nusshold C., Brunner A.M., Plastira I., Rechberger G., Reicher H., Wadsack C., Zimmer A., Malle E., Sattler W. Interference with distinct steps of sphingolipid synthesis and signaling attenuates proliferation of U87MG glioma cells. *Biochem. Pharmacol*. 2015;96(2):119-130. doi 10.1016/j.bcp.2015.05.007

Bilal F., Montfort A., Gilhodes J., Garcia V., Riond J., Carpentier S., Filleron T., Colacios C., Levade T., Daher A., Meyer N., Andrieu-

Abadie N., Ségui B. Sphingomyelin synthase 1 (SMS1) downregulation is associated with sphingolipid reprogramming and a worse prognosis in melanoma. *Front. Pharmacol*. 2019;10:443. doi 10.3389/fphar.2019.00443

Binder H., Wirth H., Arakelyan A., Lembcke K., Tiys E.S., Ivanisenko V.A., Kolchanov N.A., Kononikhin A., Popov I., Nikolaev E.N., Pastushkova L.K., Larina I.M. Time-course human urine proteomics in space-flight simulation experiments. *BMC Genomics*. 2014; 15(Suppl.12):S2. doi 10.1186/1471-2164-15-S12-S2

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

Bragina E.Y., Gomboeva D.E., Saik O.V., Ivanisenko V.A., Freidin M.B., Nazarenko M.S., Puzyrev V.P. Apoptosis genes as a key to identification of inverse comorbidity of Huntington's disease and cancer. *Int. J. Mol. Sci*. 2023;24(11):9385. doi 10.3390/ijms 24119385

Chan B., Manley J., Lee J., Singh S.R. The emerging roles of microRNAs in cancer metabolism. *Cancer Lett*. 2015;356(2, Part A): 301-308. doi 10.1016/j.canlet.2014.10.011

Chen L., Li Z.Y., Xu S.Y., Zhang X.J., Zhang Y., Luo K., Li W.P. Upregulation of miR-107 inhibits glioma angiogenesis and VEGF expression. *Cell. Mol. Neurobiol*. 2016;36(1):113-120. doi 10.1007/s10571-015-0225-3

Chen L.P., Zhang N.N., Ren X.Q., He J., Li Y. miR-103/miR-195/miR-15b regulate SALL4 and inhibit proliferation and migration in glioma. *Molecules*. 2018;23(11):2938. doi 10.3390/molecules23112938

Chen Y., Cao Y. The sphingomyelin synthase family: proteins, diseases, and inhibitors. *Biol. Chem*. 2017;398(12):1319-1325. doi 10.1515/hsz-2017-0148

Chinnaiyan P., Kensicki E., Bloom G., Prabhu A., Sarcar B., Kahali S., Eschrich S., Qu X., Forsyth P., Gillies R. The metabolomic signature of malignant glioma reflects accelerated anabolic metabolism. *Cancer Res*. 2012;72(22):5878-5888. doi 10.1158/0008-5472.CAN-12-1572-T

Clarke C.J., Cloessner E.A., Roddy P.L., Hannun Y.A. Neutral sphingomyelinase 2 (nSMase2) is the primary neutral sphingomyelinase isoform activated by tumour necrosis factor-α in MCF-7 cells. *Biochem. J*. 2011;435(2):381-390. doi 10.1042/BJ20101752

Clarke S.D., Nakamura M.T. Fatty acid synthesis and its regulation. In: Encyclopedia of Biological Chemistry. Acad. Press, 2004;99-103. doi 10.1016/B0-12-443710-9/00224-6

Comerford S.A., Huang Z., Du X., Wang Y., Cai L., Witkiewicz A.K., Walters H., Tantawy M.N., Fu A., Manning H.C., Horton J.D., Hammer R.E., Mcknight S.L., Tu B.P. Acetate dependence of tumors. *Cell*. 2014;159(7):1591-1602. doi 10.1016/j.cell.2014.11.020

Danielsen T., Rofstad E.K. VEGF, bFGF and EGF in the angiogenesis of human melanoma xenografts. *Int. J. Cancer*. 1998;76(6): 836-841. doi 10.1002/(sici)1097-0215(19980610)76:6<836::aid-ijc12>3.0.co;2-0

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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):784-793. doi 10.18699/VJGB-23-91

Deng Y., Hu J.C., He S.H., Lou B., Ding T.B., Yang J.T., Mo M.G., Ye D.Y., Zhou L., Jiang X.C., Yu K., Dong J.B. Sphingomyelin synthase 2 facilitates M2-like macrophage polarization and tumor progression in a mouse model of triple-negative breast cancer. *Acta Pharmacol. Sin*. 2021;42(1):149-159. doi 10.1038/s41401-020-0419-1

El-Karim E.A., Hagos E.G., Ghaleb A.M., Yu B., Yang V.W. Krüppel-like factor 4 regulates genetic stability in mouse embryonic fibroblasts. *Mol. Cancer*. 2013;12:89. doi 10.1186/1476-4598-12-89

Fan S.H., Wang Y.Y., Wu Z.Y., Zhang Z.F., Lu J., Li M.Q., Shan Q., Wu D.M., Sun C.H., Hu B., Zheng Y.L. AGPAT9 suppresses cell growth, invasion and metastasis by counteracting acidic tumor microenvironment through KLF4/LASS2/V-ATPase signaling pathway in breast cancer. *Oncotarget*. 2015;6(21):18406-18417. doi 10.18632/oncotarget.4074

Glaros E.N., Kim W.S., Wu B.J., Suarna C., Quinn C.M., Rye K.A., Stocker R., Jessup W., Garner B. Inhibition of atherosclerosis by the serine palmitoyl transferase inhibitor myriocin is associated with reduced plasma glycosphingolipid concentration. *Biochem. Pharmacol*. 2007;73(9):1340-1346. doi 10.1016/j.bcp.2006.12.023

Gulbins A., Schumacher F., Becker K.A., Wilker B., Soddemann M., Boldrin F., Müller C.P., Edwards M.J., Goodman M., Caldwell C.C., Kleuser B., Kornhuber J., Szabo I., Gulbins E. Antidepressants act by inducing autophagy controlled by sphingomyelin-ceramide. *Mol. Psychiatry*. 2018;23(12):2324-2346. doi 10.1038/s41380-018-0090-9

Haimovitz-Friedman A., Kan C.C., Ehleiter D., Persaud R.S., McLoughlin M., Fuks Z., Kolesnick R.N. Ionizing radiation acts on cellular membranes to generate ceramide and initiate apoptosis. *J. Exp. Med*. 1994;180(2):525-535. doi 10.1084/jem.180.2.525

Hanahan D., Weinberg R.A. The hallmarks of cancer. *Cell*. 2000; 100(1):57-70. doi 10.1016/s0092-8674(00)81683-9

Heiden M.G.V., Cantley L.C., Thompson C.B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science*. 2009;324(5930):1029-1033. doi 10.1126/science.1160809

Hernández-Tiedra S., Fabriàs G., Dávila D., Salanueva Í.J., Casas J., Montes L.R., Antón Z., García-Taboada E., Salazar-Roa M., Lorente M., Nylandsted J., Armstrong J., López-Valero I., McKee C.S., Serrano-Puebla A., García-López R., González-Martínez J., Abad J.L., Hanada K., Boya P., Goñi F., Guzmán M., Lovat P., Jäättelä M., Alonso A., Velasco G. Dihydroceramide accumulation mediates cytotoxic autophagy of cancer cells via autolysosome destabilization. *Autophagy*. 2016;12(11):2213-2229. doi 10.1080/15548627.2016.1213927

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 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., 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 Yu.I., Voevoda S.M., Karpenko T.N., Velichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolo-

mics and gene regulatory networks analysis reveal the role of non-structural 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., 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 Journal of Genetics and Breeding*. 2023;27(7):768-775. doi 10.18699/VJGB-23-89

Jaroch K., Modrakowska P., Bojko B. Glioblastoma metabolomics – in vitro studies. *Metabolites*. 2021;11(5):315. doi 10.3390/metabo11050315

Kambara H., Liu F., Zhang X., Liu P., Bajrami B., Teng Y., Zhao L., Zhou S., Yu H., Zhou W., Silberstein L.E., Cheng T., Han M., Xu Y., Luo H.R. Gasdermin D exerts anti-inflammatory effects by promoting neutrophil death. *Cell Rep*. 2018;22(11):2924-2936. doi 10.1016/j.celrep.2018.02.067

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850 (in Russian)

Koppenol W.H., Bounds P.L., Dang C.V. Otto Warburg's contributions to current concepts of cancer metabolism. *Nat. Rev. Cancer*. 2011;11(5):325-337. doi 10.1038/nrc3038

Kornhuber J., Tripal P., Reichel M., Mühle C., Rhein C., Muehlbacher M., Groemer T.W., Gulbins E. Functional inhibitors of acid sphingomyelinase (FIASMAs): a novel pharmacological group of drugs with broad clinical applications. *Cell Physiol. Biochem*. 2010;26(1):9-20. doi 10.1159/000315101

Kraveka J.M., Li L., Szulc Z.M., Bielawski J., Ogretmen B., Hannun Y.A., Obeid L.M., Bielawska A. Involvement of dihydroceramide desaturase in cell cycle progression in human neuroblastoma cells. *J. Biol. Chem*. 2007;282(23):16718-16728. doi 10.1074/jbc.M700647200

Lacroix M., Riscal R., Arena G., Linares L.K., Le Cam L. Metabolic functions of the tumor suppressor p53: implications in normal physiology, metabolic disorders, and cancer. *Mol. Metab*. 2020;33:2-22. doi 10.1016/j.molmet.2019.10.002

Lai M., La Rocca V., Amato R., Freer G., Costa M., Spezia P.G., Quaranta P., Lombardo G., Piomelli D., Pistello M. Ablation of acid ceramidase impairs autophagy and mitochondria activity in melanoma cells. *Int. J. Mol. Sci*. 2021;22(6):3247. doi 10.3390/ijms22063247

Lai Y.-J., Lin V.T.G., Zheng Y., Benveniste E.N., Lin F.-T. The adaptor protein TRIP6 antagonizes fas-induced apoptosis but promotes its effect on cell migration. *Mol. Cell. Biol*. 2010;30(23):5582-5596. doi 10.1128/MCB.00134-10

Lara-Velazquez M., Al-Kharboosh R., Jeanneret S., Vazquez-Ramos C., Mahato D., Tavanaiepour D., Rahmathulla G., Quinone-Hinojosa A. Advances in brain tumor surgery for glioblastoma in adults. *Brain Sci*. 2017;7(12):166. doi 10.3390/brainsci7120166

Larina I.M., Pastushkova L.K., 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

Lee Y.S., Choi K.M., Choi M.H., Ji S.Y., Lee S., Sin D.M., Oh K.W., Lee Y.M., Hong J.T., Yun Y.P., Yoo H.S. Serine palmitoyltransferase inhibitor myriocin induces growth inhibition of B16F10 melanoma cells through G_2/M phase arrest. *Cell Prolif*. 2011;44(4):320-329. doi 10.1111/j.1365-2184.2011.00761.x

Li K., Naviaux J.C., Bright A.T., Wang L., Naviaux R.K. A robust, single-injection method for targeted, broad-spectrum plasma metabolomics. *Metabolomics*. 2017;13(10):122. doi 10.1007/s11306-017-1264-1

Liberti M.V., Locasale J.W. The Warburg effect: how does it benefit cancer cells? *Trends Biochem. Sci*. 2016;41(3):211-218. doi 10.1016/j.tibs.2015.12.001

Lin J., Lai X., Liu X., Yan H., Wu C. Pyroptosis in glioblastoma: a crucial regulator of the tumour immune microenvironment and a predictor of prognosis. *J. Cell. Mol. Med*. 2022;26(5):1579-1593. doi 10.1111/jcmm.17200

Lin M., Liao W., Dong M., Zhu R., Xiao J., Sun T., Chen Z., Wu B., Jin J. Exosomal neutral sphingomyelinase 1 suppresses hepatocellular carcinoma via decreasing the ratio of sphingomyelin/ceramide. *FEBS J*. 2018;285(20):3835-3848. doi 10.1111/febs.14635

Louis D.N., Perry A., Wesseling P., Brat D.J., Cree I.A., Figarella-Branger D., Hawkins C., Ng H.K., Pfister S.M., Reifenberger G., Soffietti R., Von Deimling A., Ellison D.W. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology*. 2021;23(8):1231-1251. doi 10.1093/neuonc/noab106

Madigan J.P., Robey R.W., Poprawski J.E., Huang H., Clarke C.J., Gottesman M.M., Cabot M.C., Rosenberg D.W. A role for ceramide glycosylation in resistance to oxaliplatin in colorectal cancer. *Exp. Cell Res*. 2020;388(2):111860. doi 10.1016/j.yexcr.2020.111860

Mashimo T., Pichumani K., Vemireddy V., Hatanpaa K.J., Singh D.K., Sirasanagandla S., Nannepaga S., Piccirillo S.G., Kovacs Z., Foong C., Huang Z., Barnett S., Mickey B.E., Deberardinis R.J., Tu B.P., Maher E.A., Bachoo R.M. Acetate is a bioenergetic substrate for human glioblastoma and brain metastases. *Cell*. 2014;159(7):1603-1614. doi 10.1016/j.cell.2014.11.025

Maurer B.J., Metelitsa L.S., Seeger R.C., Cabot M.C., Reynolds C.P. Increase of ceramide and induction of mixed apoptosis/necrosis by N-(4-hydroxyphenyl)-retinamide in neuroblastoma cell lines. *J. Natl. Cancer Inst*. 1999;91(13):1138-1146. doi 10.1093/jnci/91.13.1138

Melland-Smith M., Ermini L., Chauvin S., Craig-Barnes H., Tagliaferro A., Todros T., Post M., Caniggia I. Disruption of sphingolipid metabolism augments ceramide-induced autophagy in preeclampsia. *Autophagy*. 2015;11(4):653-669. doi 10.1080/15548627.2015.1034414

Moro K., Nagahashi M., Gabriel E., Takabe K., Wakai T. Clinical application of ceramide in cancer treatment. *Breast Cancer*. 2019;26(4):407-415. doi 10.1007/s12282-019-00953-8

Naser E., Kadow S., Schumacher F., Mohamed Z.H., Kappe C., Hessler G., Pollmeier B., Kleuser B., Arenz C., Becker K.A., Gulbins E., Carpinteiro A. Characterization of the small molecule ARC39, a direct and specific inhibitor of acid sphingomyelinase in vitro. *J. Lipid Res*. 2020;61(6):896-910. doi 10.1194/jlr.RA120000682

Ogretmen B. Sphingolipid metabolism in cancer signalling and therapy. *Nat. Rev. Cancer*. 2018;18(1):33-50. doi 10.1038/nrc.2017.96

Oltra S.S., Colomo S., Sin L., Pérez-López M., Lázaro S., Molina-Crespo A., Choi K.H., Ros-Pardo D., Martínez L., Morales S., González-Paramos C., Orantes A., Soriano M., Hernández A., Lluch A., Rojo F., Albanell J., Gómez-Puertas P., Ko J.K., Sarrió D., Moreno-Bueno G. Distinct GSDMB protein isoforms and protease cleavage processes differentially control pyroptotic cell death and mitochondrial damage in cancer cells. *Cell Death Differ*. 2023;30(5):1366-1381. doi 10.1038/s41418-023-01143-y

Omuro A., DeAngelis L.M. Glioblastoma and other malignant gliomas: a clinical review. *JAMA*. 2013;310(17):1842-1850. doi 10.1001/jama.2013.280319

Pandey R., Caflisch L., Lodi A., Brenner A.J., Tiziani S. Metabolomic signature of brain cancer. *Mol. Carcinog*. 2017;56(11):2355-2371. doi 10.1002/mc.22694

Pang Z., Chong J., Zhou G., de Lima Morais D.A., Chang L., Barrette M., Gauthier C., Jacques P.-É., Li S., Xia J. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res*. 2021;49(W1):W388-W396. doi 10.1093/nar/gkab382

Pastushkova L.K., Kireev K.S., Kononikhin A.S., Tiys E.S., Popov I.A., Starodubtseva N.L., Dobrokhotov I.V., Ivanisenko V.A.,

Larina I.M., Kolchanov N.A., Nikolaev E.N. Detection of renal tissue and urinary tract proteins in the human urine after space flight. *PLoS One*. 2013;8(8):e71652. doi 10.1371/journal.pone.0071652

Pastushkova L.K., Kashirina D.N., Brzhozovskiy A.G., Kononikhin A.S., Tiys E.S., Ivanisenko V.A., Koloteva M.I., Nikolaev E.N., Larina I.M. Evaluation of cardiovascular system state by urine proteome after manned space flight. *Acta Astronaut*. 2019;160:594-600. doi 10.1016/j.actaastro.2019.02.015

Patrushev Y., Yudina Y., Sidelnikov V. Monolithic rod columns for HPLC based on divinylbenzene-styrene copolymer with 1-vinylimidazole and 4-vinylpyridine. *J. Liq. Chromatogr. Relat. Technol*. 2018;41(8):458-466. doi 10.1080/10826076.2018.1455149

Pike L.S., Smift A.L., Croteau N.J., Ferrick D.A., Wu M. Inhibition of fatty acid oxidation by etomoxir impairs NADPH production and increases reactive oxygen species resulting in ATP depletion and cell death in human glioblastoma cells. *Biochim. Biophys. Acta*. 2011;1807(6):726-734. doi 10.1016/j.bbabio.2010.10.022

Pizer E.S., Thupari J., Han W.F., Pinn M.L., Chrest F.J., Frehywot G.L., Townsend C.A., Kuhajda F.P. Malonyl-coenzyme-A is a potential mediator of cytotoxicity induced by fatty-acid synthase inhibition in human breast cancer cells and xenografts. *Cancer Res*. 2000;60(2):213-218

Poli M., Derosas M., Luscieti S., Cavadini P., Campanella A., Verardi R., Finazzi D., Arosio P. Pantothenate kinase-2 (Pank2) silencing causes cell growth reduction, cell-specific ferroportin upregulation and iron deregulation. *Neurobiol. Dis.* 2010;39(2):204-210. doi 10.1016/j.nbd.2010.04.009

Popik O.V., Petrovskiy E.D., Mishchenko E.L., Lavrik I.N., Ivanisenko V.A. Mosaic gene network modelling identified new regulatory mechanisms in HCV infection. *Virus Res*. 2016;218:71-78. doi 10.1016/j.virusres.2015.10.004

Poteet E., Choudhury G.R., Winters A., Li W., Ryou M.G., Liu R., Tang L., Ghorpade A., Wen Y., Yuan F., Keir S.T., Yan H., Bigner D.D., Simpkins J.W., Yang S.H. Reversing the Warburg effect as a treatment for glioblastoma. *J. Biol. Chem*. 2013;288(13):9153-9164. doi 10.1074/jbc.M112.440354

Prause K., Naseri G., Schumacher F., Kappe C., Kleuser B., Arenz C. A photocaged inhibitor of acid sphingomyelinase. *Chem. Commun*. 2020;56(94):14885-14888. doi 10.1039/d0cc06661c

Rao Z., Zhu Y., Yang P., Chen Z., Xia Y., Qiao C., Liu W., Deng H., Li J., Ning P., Wang Z. Pyroptosis in inflammatory diseases and cancer. *Theranostics*. 2022;12(9):4310-4329. doi 10.7150/thno.71086

Ren S., Babelova A., Moreth K., Xin C., Eberhardt W., Doller A., Pavenstädt H., Schaefer L., Pfeilschifter J., Huwiler A. Transforming growth factor-B2 upregulates sphingosine kinase-1 activity, which in turn attenuates the fibrotic response to TGF-B2 by impeding CTGF expression. *Kidney Int*. 2009;76(8):857-867. doi 10.1038/ki.2009.297

Riboni L., Campanella R., Bassi R., Villani R., Gaini S.M., Martinelli-Boneschi F., Viani P., Tettamanti G. Ceramide levels are inversely associated with malignant progression of human glial tumors. *Glia*. 2002;39(2):105-113. doi 10.1002/glia.10087

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. doi 10.3390/metabo11030133

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., 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*. 2018a;15(4):20180054. doi 10.1515/jib-2018-0054

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofestaedt R., Lavrik I.N., Rogaev E.I., Ivanisenko V.A. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018b;11(Suppl.1):15. doi 10.1186/s12920-018-0331-4

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

Sano O., Kazetani K.I., Adachi R., Kurasawa O., Kawamoto T., Iwata H. Using a biologically annotated library to analyze the anticancer mechanism of serine palmitoyl transferase (SPT) inhibitors. *FEBS Open Bio*. 2017;7(4):495-503. doi 10.1002/2211-5463.12196

Santos C.R., Schulze A. Lipid metabolism in cancer. *FEBS J*. 2012;279(15):2610-2623. doi 10.1111/j.1742-4658.2012.08644.x

Schiffmann S., Hartmann D., Fuchs S., Birod K., Ferreirs N., Schreiber Y., Zivkovic A., Geisslinger G., Grösch S., Stark H. Inhibitors of specific ceramide synthases. *Biochimie*. 2012;94(2):558-565. doi 10.1016/j.biochi.2011.09.007

Siegal T. Clinical impact of molecular biomarkers in gliomas. *J. Clin. Neurosci*. 2015;22(3):437-444. doi 10.1016/j.jocn.2014.10.004

Signorelli P., Munoz-Olaya J.M., Gagliostro V., Casas J., Ghidoni R., Fabriàs G. Dihydroceramide intracellular increase in response to resveratrol treatment mediates autophagy in gastric cancer cells. *Cancer Lett*. 2009;282(2):238-243. doi 10.1016/j.canlet.2009.03.020

Steiner R., Saied E.M., Othman A., Arenz C., Maccarone A.T., Poad B.L.J., Blanksby S.J., Von Eckardstein A., Hornemann T. Elucidating the chemical structure of native 1-deoxysphingosine. *J. Lipid Res*. 2016;57(7):1194-1203. doi 10.1194/jlr.M067033

Tea M.N., Poonnoose S.I., Pitson S.M. Targeting the sphingolipid system as a therapeutic direction for glioblastoma. *Cancers* (*Basel*). 2020;12(1):111. doi 10.3390/cancers12010111

Turner N., Lim X.Y., Toop H.D., Osborne B., Brandon A.E., Taylor E.N., Fiveash C.E., Govindaraju H., Teo J.D., McEwen H.P., Couttas T.A., Butler S.M., Das A., Kowalski G.M., Bruce C.R., Hoehn K.L., Fath T., Schmitz-Peiffer C., Cooney G.J., Montgomery M.K., Morris J.C., Don A.S. A selective inhibitor of ceramide synthase 1 reveals a novel role in fat metabolism. *Nat. Commun*. 2018;9(1):3165. doi 10.1038/s41467-018-05613-7

Vollmann-Zwerenz A., Leidgens V., Feliciello G., Klein C.A., Hau P. Tumor cell invasion in glioblastoma. *Int. J. Mol. Sci*. 2020;21(6):1932. doi 10.3390/ijms21061932

Wang Y., Zhang C., Jin Y., Wang S., He Q., Liu Z., Ai Q., Lei Y., Li Y., Song F., Bu Y. Alkaline ceramidase 2 is a novel direct target of p53 and induces autophagy and apoptosis through ROS generation. *Sci. Rep*. 2017;7:44573. doi 10.1038/srep44573

Wang Z., Wen L., Zhu F., Wang Y., Xie Q., Chen Z., Li Y. Overexpression of ceramide synthase 1 increases C18-ceramide and leads to lethal autophagy in human glioma. *Oncotarget*. 2017;8(61):104022-104036. doi 10.18632/oncotarget.21955

Warburg O. On the origin of cancer cells. *Science*. 1956;123(3191):309-314. doi 10.1126/science.123.3191.309

Wilfred B.R., Wang W.X., Nelson P.T. Energizing miRNA research: a review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways. *Mol. Genet. Metab*. 2007;91(3):209-217. doi 10.1016/j.ymgme.2007.03.011

Wolf A., Agnihotri S., Guha A. Targeting metabolic remodeling in glioblastoma multiforme. *Oncotarget*. 2010;1(7):552-562. doi 10.18632/oncotarget.190

Xu R., Garcia-Barros M., Wen S., Li F., Lin C.L., Hannun Y.A., Obeid L.M., Mao C. Tumor suppressor p53 links ceramide metabolism to DNA damage response through alkaline ceramidase 2. *Cell Death Differ*. 2018;25(5):841-856. doi 10.1038/s41418-017-0018-y

Youngblood M.W., Stupp R., Sonabend A.M. Role of resection in glioblastoma management. *Neurosurg. Clin. N. Am*. 2021;32(1):9-22. doi 10.1016/j.nec.2020.08.002

Yuan M., Breitkopf S.B., Yang X., Asara J.M. A positive/negative ion – switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat. Protoc*. 2012;7(5):872-881. doi 10.1038/nprot.2012.024

Zhang S., Huang P., Dai H., Li Q., Hu L., Peng J., Jiang S., Xu Y., Wu Z., Nie H., Zhang Z., Yin W., Zhang X., Lu J. TIMELESS regulates sphingolipid metabolism and tumor cell growth through Sp1/ACER2/S1P axis in ER-positive breast cancer. *Cell Death Dis*. 2020;11(10):892. doi 10.1038/s41419-020-03106-4

Zheng K., Chen Z., Feng H., Chen Y., Zhang C., Yu J., Luo Y., Zhao L., Jiang X., Shi F. Sphingomyelin synthase 2 promotes an aggressive breast cancer phenotype by disrupting the homoeostasis of ceramide and sphingomyelin. *Cell Death Dis*. 2019;10(3):157. doi 10.1038/s41419-019-1303-0

Zhou W., Wahl D.R. Metabolic abnormalities in glioblastoma and metabolic strategies to overcome treatment resistance. *Cancers* (*Basel*). 2019;11(9):1231. doi 10.3390/cancers11091231

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 September 20, 2024. Revised October 30, 2024. Accepted November 2, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):897-903

doi 10.18699/vjgb-24-97

A software module to assess the metabolic potential of mutant strains of the bacterium *Corynebacterium glutamicum*

F.V. Kazantsev [1, 2, 3], M.F. Trofimova[2], T.M. Khlebodarova[1, 2], Yu.G. Matushkin [1, 2, 3], S.A. Lashin [1, 2, 3]

[1] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
kazfdr@bionet.nsc.ru

Abstract. Technologies for the production of a range of compounds using microorganisms are becoming increasingly popular in industry. The creation of highly productive strains whose metabolism is aimed to the synthesis of a specific desired product is impossible without complex directed modifications of the genome using mathematical and computer modeling methods. One of the bacterial species actively used in biotechnological production is *Corynebacterium glutamicum*. There are already 5 whole-genome flux balance models for it, which can be used for metabolism research and optimization tasks. The paper presents fluxMicrobiotech, a software module developed at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, which implements a series of computational protocols designed for high-performance computer analysis of *C. glutamicum* whole-genome flux balance models. The tool is based on libraries from the opencobra community (https://opencobra.github.io) within the Python programming language (https://www.python.org), using the Pandas (https://pandas.pydata.org) and Escher (https://escher.readthedocs.io) libraries . It is configured to operate on a 'file-in/file-out' basis. The model, environmental conditions, and model constraints are specified as separate text table files, which allows one to prepare a series of files for each section, creating databases of available test scenarios for variations of the model. Or vice versa, allowing a single model to be tested under a series of different cultivation conditions. Post-processing tools for modeling data are set up, providing visualization of summary charts and metabolic maps.

Key words: flux models; bacterial metabolism; metabolic optimization; rational metabolic engineering.

For citation: Kazantsev F.V., Trofimova M.F., Khlebodarova T.M., Matushkin Yu.G., Lashin S.A. A software module to assess the metabolic potential of mutant strains of the bacterium *Corynebacterium glutamicum*. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):897-903. doi 10.18699/vjgb-24-97

Funding. This work was supported by the projects of the Kurchatov Genomic Centre of ICG SB RAS No. 075-15-2019-1662.

Программный модуль для оценки метаболического потенциала мутантных штаммов бактерии *Corynebacterium glutamicum*

Ф.В. Казанцев [1, 2, 3], М.Ф. Трофимова[2], Т.М. Хлебодарова[1, 2], Ю.Г. Матушкин [1, 2, 3], С.А. Лашин [1, 2, 3]

[1] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
kazfdr@bionet.nsc.ru

Аннотация. Технологии производства различных соединений с применением микроорганизмов приобретают все большую популярность в промышленном производстве. Создание современных высокопродуктивных штаммов, метаболизм которых ориентирован на синтез конкретного целевого продукта, невозможно без комплексной направленной модификации генома с применением методов математического и компьютерного моделирования. Одним из видов бактерий, активно используемых в биотехнологическом производстве, является *Corynebacterium glutamicum*. Для него существует уже пять полногеномных потоковых моделей, которые можно использовать для задач исследования и оптимизации метаболизма. В работе представлен программный модуль развиваемого в Институте цитологии и генетики СО РАН инструмента FluxMicrobiotech, в рамках которого реализована серия вычислительных протоколов, предназначенных для массового компьютерного анализа потоковых моделей *C. glutamicum* на высокопроизводительных вычислительных компьютерах. Программный модуль реализован на языке Python с применением библиотек Pandas, cobraPy и Escher и

настроен на работу по принципу «файл на вход/файл на выход». Модель, условия среды и ограничения модели задаются как отдельные текстовые табличные файлы, что позволяет заготовить серию файлов для каждого из разделов, создавая базы доступных сценариев испытаний для вариаций модели. Или, наоборот, позволяет испытывать одну модель в серии разных условий культивирования. Настроены инструменты постобработки данных моделирования, обеспечивающие визуализацию сводных диаграмм и метаболических карт.
Ключевые слова: потоковые модели; метаболизм бактерии; оптимизация метаболизма; рациональная метаболическая инженерия.

Introduction

Technologies for the production of a range of compounds using microorganisms are becoming increasingly popular in the industry. Creation of modern highly productive microorganism strains, the metabolism of which is focused on synthesis of a specific target product, is impossible without complex directed genome modifications. To date, a wide range of rational and systemic metabolic engineering methods have been developed to increase the production of target substances (Sheremetieva et al., 2023, 2024), the use of which, together with computer modelling approaches, will make it possible to more accurately assess the impact of genome changes on the dynamics of the system and the yield of the final product (Ananda et al., 2024). Implementation of the flux-based mathematical modelling methods for molecular genetic and metabolic systems within the computational modelling frameworks (Mendoza et al., 2019; Mao et al., 2023) and creation of whole-genome flux-based mathematical models allow *in silico* prediction of genetic modifications required to increase culture growth rate and target product yield under optimal conditions on different substrates (Gu et al., 2019; Mao et al., 2023).

One of the bacterial species actively used in biotechnological production is *Corynebacterium glutamicum*. Since its discovery in 1956 (Kinoshita et al., 1957) until now, the main application of this bacterial species has been the production of amino acids and their derivatives (Tsuge, Matsuzawa, 2021), which is currently the second most economically important process in industrial biotechnology (Barcelos et al., 2018). *C. glutamicum* are non-pathogenic, GC-rich, Gram-positive soil bacteria. They do not form spores, grow rapidly, do not require special conditions for growth, do not secrete proteases, have a relatively stable genome and are resistant to high concentrations of potentially toxic substances, making this microorganism an ideal platform for the development of industrially relevant strains based on it (Wendisch et al., 2016).

The main approaches for modifying the genome of biotechnologically relevant bacterial strains include: 1) gene knockouts (switching off); 2) insertion of additional genes leading to the creation of new metabolic reaction chains; 3) insertion of mutations both in the regulatory regions of genes and in the structure of genes in order to decrease/increase gene expression and activity of their products, respectively; 4) other modern methods of *C. glutamicum* genome editing, without which it is impossible to realize a large number of directed modifications necessary for the implementation of rational and systemic metabolic engineering approaches (Sheremetieva et al., 2023, 2024). Effective planning, execution and control of such modifications are difficult without the use of mathematical and computational modelling techniques.

The paper is dedicated to the development of a software module within the framework of the FluxMicrobiotech toolkit created at the Institute of Cytology and Genetics SB RAS. The toolkit was created to assess the metabolic potential of a bacterium using flux modelling methods, including a set of computational protocols configured for massive computational analysis of the metabolism of target bacterial strains when cultivated on different nutrient media and under different environmental conditions (aerobic/anaerobic).

Materials and methods

The developed computational protocols are based on the open source flux modelling methods library opencobra (opencobra.github.io) within the Python programming language (https://www.python.org/). The protocols are designed as "notebooks" in the Jupyter programming environment (https://jupyter.org/). This structure allows combining computational blocks with stages of results analysis. The approach of organizing computations using "notebooks" has become a familiar tool in big data analysis methodology, implying the creation of computational pipelines and their regular adjustment to changing objective conditions. Control of the correct use is gained by a powerful toolkit of annotations to the calculation stage. The cobraPy (https://opencobra.github.io/cobrapy/) and Pandas (https://pandas.pydata.org/) libraries are used to solve optimization problems. The yEd Graph Editor (https://yworks.com/products/yed) is used for the raw visualization of gene networks. Creation of metabolic maps and plotting of solutions on them during modelling is implemented in the Escher toolkit (escher.github.io/). The developed protocols support high-performance computing methods and require memory to store the results. Thus, it is recommended to carry out the work on high-performance computers.

The flow modelling techniques (the alternative term is FBA – Flux Balance Analysis) used in this paper belong to the linear programming problem domain. It is to address the challenges of metabolic research that a series of computational FBA| method libraries are being developed within the opencobra community (https://opencobra.github.io). The basis of this methodology is the representation of the metabolic pathway as a graph given by an adjacency matrix with the rows corresponding to metabolites, and the columns, to metabolic reactions and processes. Matrix elements are stoichiometric coefficients specifying the proportion of a metabolite and its role in the selected reaction (reagent or reaction product). Such matrices can be constructed manually by carefully describing the target metabolic pathways, or automatically by generating a matrix from genomic information. Using a well-annotated bacterial genome sequence and various bioinformatics tools, potential metabolic pathways and the bacterium's ability to

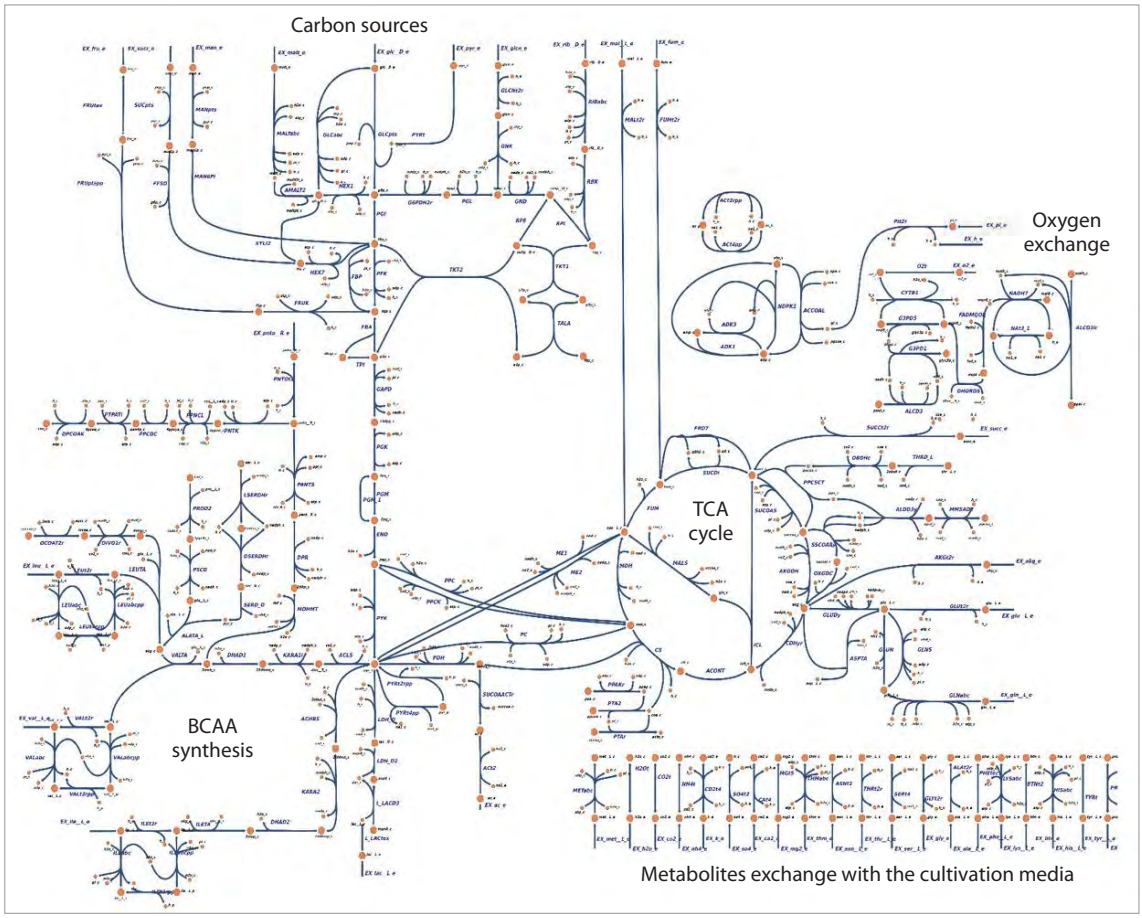


Fig. 1. Metabolic map focused on metabolic pathways for the synthesis of branched-chain amino acids (BCAAs).
The visualization was done in the Escher tool as an extended network of the iCGB21FR model.

synthesize target metabolites can be identified. It is this information that is processed by software tools for generating Whole Genome Flux Models (the alternative term is GSM – genome-scale metabolic models) (Machado et al., 2018; Kulyashov et al., 2023).

A flux model constructed in the manner mentioned above is a starting point in the task of assessing the metabolism of a bacterium and can contain several thousand reactions describing the full set of functionalities available in the genome. There is the BiGG database (http://bigg.ucsd.edu/), which is positioned as a central point for storing and reusing flux models. This resource contains the largest collection of whole-genome mathematical models developed for different organisms, and in addition is being developed as a database of reference biochemical reactions for these types of models as well. Within BiGG, the Escher metabolic network visualization tool (King et al., 2015) is being developed in parallel, allowing the same metabolic maps to be reused for models of different organisms. The BiGG database contains 108 published and manually validated whole-genome metabolic models for 40 different organisms (Norsigian et al., 2019).

Thus, the bundling of genome data, tools for building and annotating whole-genome flux models, and their integration within the BiGG approach provide the basis for high-throughput computational analyses of bacterial metabolism. While the model is whole-genome, only a subset of the metabolic pathway reactions for key metabolites are used when displaying the metabolic map as a graph (Fig. 1), assuming that pathways not included in the visualization are also involved in the analysis.

Results

Flux model of *Corynebacterium glutamicum*

To date, several mathematical models describing the metabolism of the bacterium *C. glutamicum* have been created and published: iEZ482, iCW773, iCGB21FR, ecCGL1, iJM658 (Kjeldsen, Nielsen, 2009; Zelle et al., 2015; Mei et al., 2016; Zhang et al., 2017; Feierabend et al., 2021; Niu et al., 2022). These models are based on whole-genome data and have been verified on experimental data on bacterial growth, ability to synthesize amino acids on different carbon sources and under different cultivation medium conditions. The models were used to analyse the production of glutamate (Mei et al., 2016; Feierabend et al., 2021), isoleucine (Zhang et al., 2017) and lysine (Kjeldsen, Nielsen, 2009; Zhang et al., 2017; Niu et al., 2022).

The iEZ482 model was presented in 2015 and describes the metabolism of strain ATCC 13032. It contains 475 metabolic reactions and 408 metabolites. The model was validated by

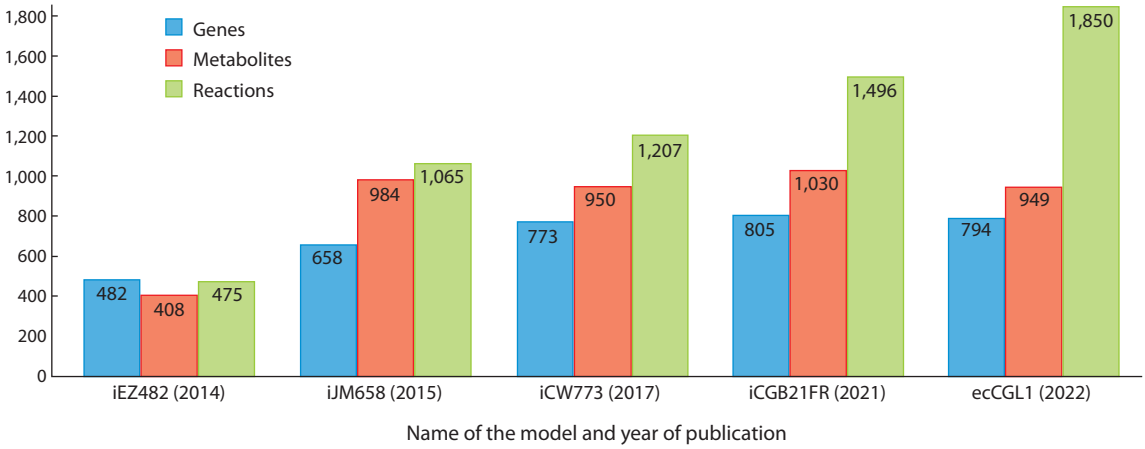


Fig. 2. Mathematical models of *C. glutamicum* metabolism and their main characteristics.

the authors using experimental data on the ability to excrete 20 amino acids. The iCW773 model published in 2017 contains 1,207 reactions and 950 metabolites. Based on iCW773, the ecCGL1 model was published in 2022. It provides a mathematical description of the metabolism of the bacterium *C. glutamicum* strain ATCC 13032 with enzymatic constraints, in which not only metabolites and reactions are specified, but also constraints on the maximum concentration of enzymes in the bacterium are incorporated. The iJM658 model was built for strain S9114, published in 2016, and contains 658 genes, 984 metabolites and 1,065 reactions. Further development of whole-genome modelling for *C. glutamicum* ATCC 13032 led to the iCGB21FR model, released in 2021. The model contains 1,496 reactions, 1,030 metabolites, 805 genes and 3 compartments: extracellular space, cytosol and periplasm. Validation of the model was performed by the authors on the metabolism of L-glutamate, which in turn is a precursor for the synthesis of a series of amino acids. Characteristics of the found models are presented in Figure 2.

The iCGB21FR model was chosen as the base model for setting up computational protocols, building metabolic maps and data post-processing tools, as it describes the metabolism of *C. glutamicum* bacteria in the most complete and up-to-date way. It can also serve as a benchmark for model annotation, as it covers most of the recommendation points in the systems biology model design standard, including references to existing databases and ontologies. The iCGB21FR model is freely available in the BioModels database (https://www.ebi.ac.uk/biomodels, model identifier MODEL2102050001). The model demonstrates the ability of the bacterium to grow on different carbon sources under aerobic and anaerobic conditions on three different culture media: minimal M9 medium, minimal CGXII medium, and complete lysogenic broth (LB) medium. These conditions differ in the quantity and quality (availability of additional carbon or amino acid sources) of metabolites that the model can consume from the culture medium for processing into metabolic products.

Computational protocols

The developed software module contains a series of basic computational scripts, the data flow of which is schematically represented in Figure 3. This is a prepared Jupyter lab notebook in which the calculation parameters are set.

The starting conditions for all protocols are the same:

1) it is necessary to specify the flux model (*.json file), which describes the basic structure and constraints of the model. This model can be obtained from the BIGG databases or created using the cobraPy software toolkit;
2) set the cultivation medium parameters as a tabular text file (*.csv);
3) set additional constraints on model fluxes as a tabular text file (*.csv).

Then, depending on the task to be solved, the calculation parameters are set up. Jupyter lab notebook as a computational protocol allows users to quickly modify each block of calculations if necessary. As a result, the computational protocol is actually specified through a set of files: model, cultivation medium, additional constraints. This provides the ability to prepare a series of files for each section, creating databases of available test scenarios for variations of a model or, conversely, testing a single model under a series of different cultivation conditions.

The result of the protocol is the vector of resulting velocities over the entire model structure (or a set of such vectors in the form of a rectangular matrix). For post-processing tasks, a toolkit has been set up to display data both as result diagrams and as a visualization of flows on a metabolic map (Fig. 1). The task of exporting the results as a series of interactive metabolic maps was done using the Escher toolkit (https://escher.readthedocs.io).

Scenario for estimating biomass growth

The bacterial cultivation medium plays a major role in biotechnological production. The media can be of minimal biochemical composition or rich in amino acids, so that the bacterium can consume them from the medium rather than spending internal resources to synthesize amino acids and other metabolites. In order to estimate metabolic parameters of strains using modelling, it is necessary to set the cultivation conditions as precisely as possible.

The first test of model adequacy is its ability to predict biomass growth on given substrates in accordance with

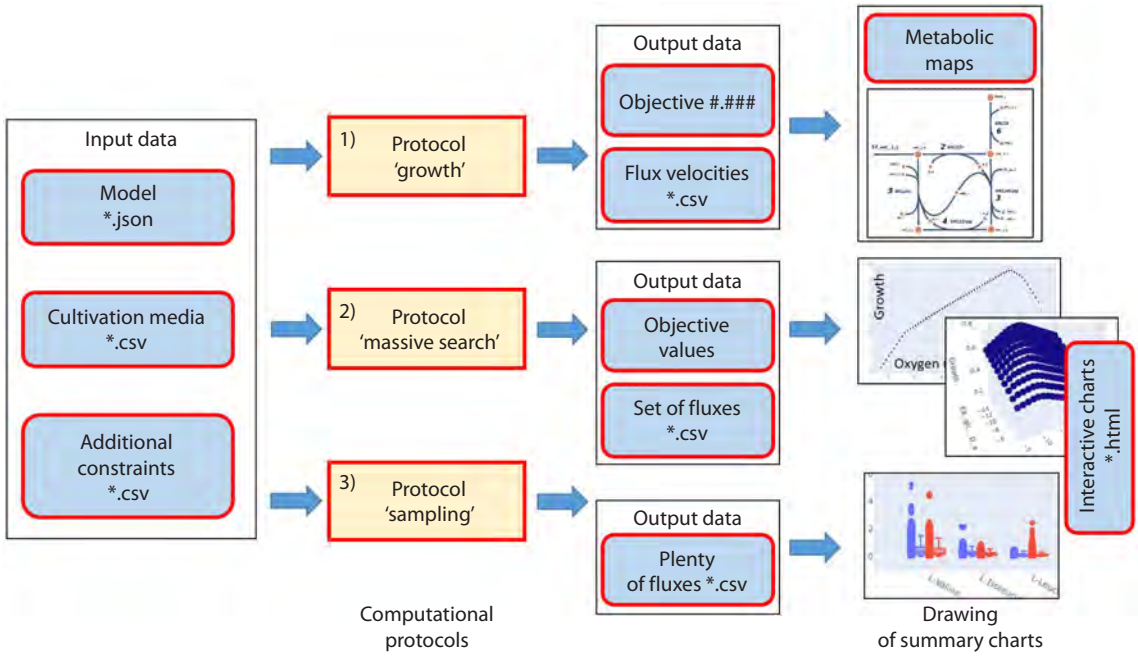


Fig. 3. A data flow diagram of computational protocols.

experimental data. This parameter is usually not difficult to investigate experimentally: there is plenty of data on strain growth rates and substrate uptake rates or lack of growth on selected carbon sources. Comparison of these values is a key step in the basic evaluation of the model for correctness. Specifically, the iCGB21FR model was tested for completeness on multiple media for its ability to synthesize amino acids under both aerobic and anaerobic conditions. By varying the conditions of the cultivation medium, the limiting substrates in the biomass production reaction can be evaluated. This scenario is also suitable for assessing the ability to achieve the selected reactions under given cultivation medium conditions, i. e. to test the sufficiency of metabolites in the medium to potentially complete the targeted metabolic reactions.

Scenario for evaluating the optimization of the space of feasible solutions

The previous scenario tested the implementation of targeted pathways from the point of substrate uptake to specific metabolic reactions. The next aspect of the study of such models is to assess the ability of the bacterium to operate under given conditions, i. e. the ability to synthesize a series of metabolites on a given substrate under the applied constraints in principle. Sampling methods for estimating the feasible solution space are helpful in this task. The solution in the "sampling" method is a vector of flux rates through all metabolic reactions that satisfies the balance conditions and user-applied constraints on the boundaries of the selected reaction rates. In contrast to the flux balance analysis method, "sampling" generates a set of possible feasible solutions of the reaction system in the model without specifying target characteristics, which makes this method convenient for evaluating ways to optimize reactions (Herrmann et al., 2019).

For a more accurate representation of the space of possible solutions, it is necessary to generate a sufficiently large number of samples with sizes of dozens/hundreds of thousands of points in the solution space (taking into account that each point in this space is described by hundreds or even thousands of numerical values of flow velocities). As a result, one can obtain a set of points in the solution space that can indicate the most frequent solutions under given conditions. The method uniformly selects points covering the solution space. By mapping the points to the coordinates of the target velocities, the expected distribution of values can be obtained. Thus, we do not get a specific distribution of fluxes on the metabolic map, but a series of solutions (a series of resultant fluxes/cloud of points). Each point in this series of solutions can be mapped onto the rate axis of selected reactions of the metabolic network. This approach allows comparing flux distributions of both several models under the same conditions and one model under different conditions/constraints (Fig. 4).

In particular, a series of computational experiments on the effect of gene knockouts on metabolite excretion identified the *atpB* gene (KEGG cgb:cg1362), the synthesis product of which is involved in the ATP phosphorylation reaction (Fig. 4). Knockout of *atpB* provides potentially greater excretion of L-valine. Indirect evidence for the importance of this gene comes from the study (Jensen et al., 1993), which has shown that mutations in the ATP synthase operon in *Escherichia coli* can lead to a higher growth rate on glucose.

Running the calculations for 10 thousand solutions/points generates about 200 Mb of data in one run. Calculations and post-processing of such data are recommended to be performed on high-performance computational machines.

Conclusion

The largest database of whole-genome models, BIGG (http://bigg.ucsd.edu/models), has 108 models for 40 different or-

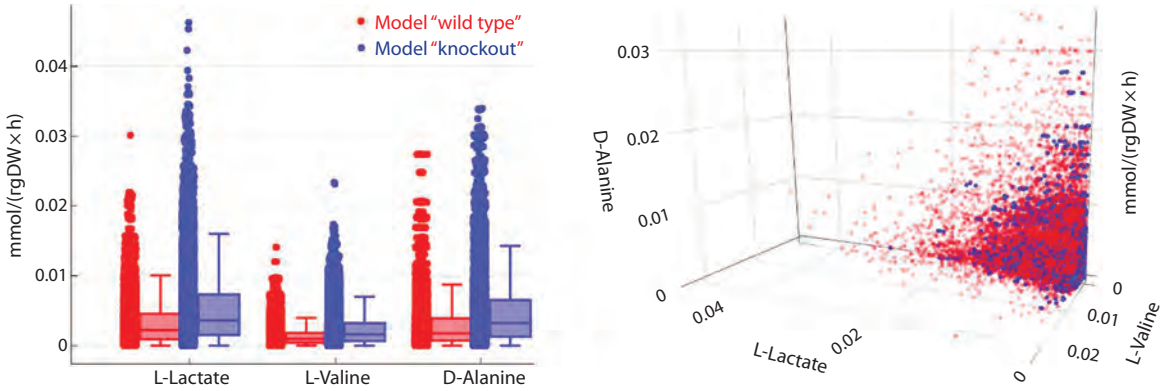


Fig. 4. Comparison result of two variants of the iCGB21FR model: a baseline ("wild type") model and a "knockout" model where a knockout of the periplasmic ATP synthase (*atpB*) gene is introduced.

On the left – representation of lactate, valine and alanine excretion rate values; on the right – representation of the same values in one three-dimensional space (projections of 10,000 solution points on L-valine, D-alanine and L-lactate axes). Reaction flux rates in the model are expressed in mmol per gram of biomass dry weight per hour (mmol/(gDW × h)).

ganisms. We found at least five whole-genome mathematical models on *C. glutamicum*, indicating a great interest in the object of study. The methodology of whole-genome modelling itself is still in the development stage and requires manual customization of tools for each new object. This gives a wide space for the development of mathematical and computational modelling techniques within the systems biologists/rational metabolic engineers' community. Studies are now underway to incorporate transcriptomic and proteomic data into these types of models, leading to higher predictive power than simpler flux models.

Although *C. glutamicum* has been studied since 1956 (Kinoshita et al., 1957), gathering public information on strains of the bacterium is a challenge in itself. There are many strains for which the data is commercially available and may not be in the public domain. The development of computational pipelines will allow them to be applied to the metabolism of other strains in the future.

The proposed software module in the form of a series of computational protocols is configured for mass analysis of *C. glutamicum* strain models on cultivation on different nutrient media and under different environmental conditions (aerobic/anaerobic). The protocols are configured to run on a file-as-input/file-as-output basis, where the model, environment conditions, and model constraints are specified as separate files. Methods for visualization of simulation results have been set up, in particular for displaying data on a series of user-prepared metabolic maps. The specifics of algorithm execution require the use of high-performance computers and access to large amounts of data storage. The module is a part of the FluxMicrobiotech tool being developed at ICG SB RAS.

References

Ananda R., Daud K.M., Zainudin S. A review of advances in integrating gene regulatory networks and metabolic networks for designing strain optimization. *J. King Saud Univ. Comput. Inf. Sci.* 2024; 36(6):102120. doi 10.1016/j.jksuci.2024.102120

Barcelos M.C.S., Lupki F.B., Campolina G.A., Nelson D.L., Molina G. The colors of biotechnology: general overview and developments of white, green and blue areas. *FEMS Microbiol. Lett.* 2018;365(21): fny239. doi 10.1093/femsle/fny239

Feierabend M., Renz A., Zelle E., Nöh K., Wiechert W., Dräger A. High-quality genome-scale reconstruction of *Corynebacterium glutamicum* ATCC 13032. *Front. Microbiol.* 2021;12:750206. doi 10.3389/fmicb.2021.750206

Gu C., Kim G.B., Kim W.J., Kim H.U., Lee S.Y. Current status and applications of genome-scale metabolic models. *Genome Biol.* 2019; 20(1):121. doi 10.1186/s13059-019-1730-3

Herrmann H.A., Dyson B.C., Vass L., Johnson G.N., Schwartz J.-M. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ Syst. Biol. Appl.* 2019;5(1):32. doi 10.1038/s41540-019-0109-0

Jensen P.R., Michelsen O., Westerhoff H.V. Control analysis of the dependence of *Escherichia coli* physiology on the H^+-ATPase. *Proc. Natl. Acad. Sci. USA.* 1993;90(17):8068-8072. doi 10.1073/pnas.90.17.8068

King Z.A., Dräger A., Ebrahim A., Sonnenschein N., Lewis N.E., Palsson B.O. Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput. Biol.* 2015;11(8):e1004321. doi 10.1371/journal.pcbi.1004321

Kinoshita S., Udaka S., Shimono M. Studies on the amino acid fermentation. *J. Gen. Appl. Microbiol.* 1957;3(3):193-205. doi 10.2323/jgam.3.193

Kjeldsen K.R., Nielsen J. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.* 2009;102(2):583-597. doi 10.1002/bit.22067

Kulyashov M.A., Kolmykov S.K., Khlebodarova T.M., Akberdin I.R. State-of the-art constraint-based modeling of microbial metabolism: from basics to context-specific models with a focus on methanotrophs. *Microorganisms*. 2023;11(12):2987. doi 10.3390/microorganisms11122987

Machado D., Andrejev S., Tramontano M., Patil K.R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* 2018;46(15):7542-7553. doi 10.1093/nar/gky537

Mao Z., Yuan Q., Li H., Zhang Y., Huang Y., Yang C., Wang R., Yang Y., Wu Y., Yang S., Liao X., Ma H. CAVE: a cloud-based platform for analysis and visualization of metabolic pathways. *Nucleic Acids Res.* 2023;51(W1):W70-W77. doi 10.1093/nar/gkad360

Mei J., Xu N., Ye C., Liu L., Wu J. Reconstruction and analysis of a genome-scale metabolic network of *Corynebacterium glutamicum* S9114. *Gene*. 2016;575(2):615-622. doi 10.1016/j.gene.2015.09.038

Mendoza S.N., Olivier B.G., Molenaar D., Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* 2019;20(1):158. doi 10.1186/s13059-019-1769-1

Niu J., Mao Z., Mao Y., Wu K., Shi Z., Yuan Q., Cai J., Ma H. Construction and analysis of an enzyme-constrained metabolic model of *Corynebacterium glutamicum*. *Biomolecules*. 2022;12(10):1499. doi 10.3390/biom12101499

Norsigian C.J., Pusarla N., McConn J.L., Yurkovich J.T., Dräger A., Palsson B.O., King Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* 2019;48(D1):D402-D406. doi 10.1093/nar/gkz1054

Sheremetieva M.E., Anufriev K.E., Khlebodarova T.M., Kolchanov N.A., Yanenko A.S. Rational metabolic engineering of *Corynebacterium glutamicum* to create a producer of L-valine. *Vavilov J. Genet. Breed.* 2023;26(8):743-757. doi 10.18699/VJGB-22-90

Sheremetieva M.E., Khlebodarova T.M., Derbikov D.D., Rozantseva V.V., Kolchanov N.A., Yanenko A.S. Systems metabolic engineering of *Corynebacterium glutamicum* to create a producer of L-valine. *Biotekhnologiya = Biotechnology*. 2024;40(3):3-23. doi 10.56304/S0234275824030025 (in Russian)

Tsuge Y., Matsuzawa H. Recent progress in production of amino acid-derived chemicals using *Corynebacterium glutamicum*. *World J. Microbiol. Biotechnol.* 2021;37(3):49. doi 10.1007/s11274-021-03007-4

Wendisch V.F., Jorge J.M.P., Pérez-García F., Sgobba E. Updates on industrial production of amino acids using *Corynebacterium glutamicum*. *World J. Microbiol. Biotechnol.* 2016;32(6):105. doi 10.1007/s11274-016-2060-1

Zelle E., Nööh K., Wiechert W. Growth and production capabilities of *Corynebacterium glutamicum*: interrogating a genome-scale metabolic network model. In: Burkovski A. (Ed.) *Corynebacterium glutamicum*: From Systems Biology to Biotechnological Applications. Caister Acad. Press, 2015;39-56. doi 10.21775/9781910190050.04

Zhang Yu, Cai J., Shang X., Wang B., Liu S., Chai X., Tan T., Zhang Yun, Wen T. A new genome-scale metabolic model of *Corynebacterium glutamicum* and its application. *Biotechnol. Biofuels*. 2017;10(1):169. doi 10.1186/s13068-017-0856-3

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

Received September 17, 2024. Revised November 20, 2024. Accepted November 21, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):904-917

DOI 10.18699/vjgb-24-98

Reconstruction and computational analysis of the microRNA regulation gene network in wheat drought response mechanisms

M.A. Kleshchev [1, 4], A.V. Maltseva [1, 3, 4], E.A. Antropova [1, 4], P.S. Demenkov [1, 2, 3, 4], T.V. Ivanisenko [1, 2, 3, 4], Y.L. Orlov [5, 6], H. Chao [7], M. Chen [7], N.A. Kolchanov [1, 2, 3], V.A. Ivanisenko [1, 2, 3, 4]

[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
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Research Center in the Field of Artificial Intelligence of Novosibirsk State University, Novosibirsk, Russia
[5] Agrarian and Technological Institute, Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, Russia
[6] Digital Health Center, I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, Russia
[7] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China

salix@bionet.nsc.ru

Abstract. Drought is a critical factor limiting the productivity of bread wheat (*Triticum aestivum* L.), one of the key agricultural crops. Wheat adaptation to water deficit is ensured by complex molecular genetic mechanisms, including the coordinated work of multiple genes regulated by transcription factors and signaling non-coding RNAs, particularly microRNAs (miRNAs). miRNA-mediated regulation of gene expression is considered one of the main mechanisms of plant resistance to abiotic stresses. Studying these mechanisms necessitates computational systems biology methods. This work aims to reconstruct and analyze the gene network associated with miRNA regulation of wheat adaptation to drought. Using the ANDSystem software and the specialized Smart crop knowledge base adapted for wheat genetics and breeding, we reconstructed a wheat gene network responding to water deficit, comprising 144 genes, 1,017 proteins, and 21 wheat miRNAs. Analysis revealed that miRNAs primarily regulate genes controlling the morphogenesis of shoots and roots, crucial for morphological adaptation to drought. The key network components regulated by miRNAs are the MYBa and WRKY41 family transcription factors, heat-shock protein HSP90, and the RPM1 protein. These proteins are associated with phytohormone signaling pathways and calcium-dependent protein kinases significant in plant water deficit adaptation. Several miRNAs (*MIR7757*, *MIR9653a*, *MIR9671* and *MIR9672b*) were identified that had not been previously discussed in wheat drought adaptation. These miRNAs regulate many network nodes and are promising candidates for experimental studies to enhance wheat resistance to water deficiency. The results obtained can find application in breeding for the development of new wheat varieties with increased resistance to water deficit, which is of substantial importance for agriculture in the context of climate change.

Key words: microRNA; bread wheat; drought; genes; genetic regulation; associative gene networks; plant bioinformatics; Smart crop knowledge base; ANDSystem computer tool.

For citation: Kleshchev M.A., Maltseva A.V., Antropova E.A., Demenkov P.S., Ivanisenko T.V., Orlov Y.L., Chao H., Chen M., Kolchanov N.A., Ivanisenko V.A. Reconstruction and computational analysis of the microRNA regulation gene network in wheat drought response mechanisms. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):904-917. DOI 10.18699/vjgb-24-98

Funding. The work of MAK, AVM, EAA, PSD, TVI, YLO, NAK, and VAI was supported by the Russian-Chinese grant from the Russian Science Foundation No. 23-44-00030. The work of MCh and HCh was supported by the National Natural Science Foundation of China (No. 32261133526).

Реконструкция и компьютерный анализ генной сети, отражающей роль микроРНК в регуляции ответа пшеницы на засуху

М.А. Клещев [1, 4], А.В. Мальцева [1, 3, 4], Е.А. Антропова [1, 4], П.С. Деменков [1, 2, 3, 4], Т.В. Иванисенко [1, 2, 3, 4], Ю.Л. Орлов [5, 6], Х. Чао [7], М. Чэнь [7], Н.А. Колчанов [1, 2, 3], В.А. Иванисенко [1, 2, 3, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

[4] Исследовательский центр в сфере искусственного интеллекта Новосибирского национального исследовательского государственного университета, Новосибирск, Россия
[5] Аграрно-технологический институт Российского университета дружбы народов им. Патриса Лумумбы, Москва, Россия
[6] Центр цифровой медицины, Первый Московский государственный медицинский университет им. И.М. Сеченова Минздрава России (Сеченовский Университет), Москва, Россия
[7] Отдел биоинформатики, Колледж естественных наук, Чжэцзянский университет, Ханчжоу, Китай
salix@bionet.nsc.ru

Аннотация. Недостаток влаги – критический фактор, ограничивающий продуктивность мягкой пшеницы (*Triticum aestivum* L.), одной из ключевых сельскохозяйственных культур. Адаптация пшеницы к водному дефициту обеспечивается комплексными молекулярно-генетическими механизмами, включающими согласованную работу множества генов, регулируемых транскрипционными факторами и сигнальными некодирующими РНК, в частности микроРНК. микроРНК – опосредованная регуляция экспрессии генов – рассматривается как один из основных механизмов устойчивости растений к абиотическим стрессам. Изучение этих сложных молекулярно-генетических механизмов требует применения методов компьютерной системной биологии. Цель данной работы – реконструкция и компьютерный анализ генной сети, связанной с микроРНК-регуляцией адаптации мягкой пшеницы к условиям недостаточного увлажнения. Для достижения этой цели использованы программно-информационная система ANDSystem и специализированная база знаний Smart crop, адаптированная для области генетики и селекции пшеницы. Нами была реконструирована генная сеть ответа пшеницы на водный дефицит, включающая 144 гена, 1017 белков и 21 микроРНК пшеницы. Анализ сети выявил, что микроРНК преимущественно регулируют гены, контролирующие процессы морфогенеза побегов и корней растений, что играет важную роль в морфологических адаптациях к засухе. Ключевыми компонентами генной сети, регулируемыми микроРНК, оказались транскрипционные факторы семейств MYB и WRKY, а также белок теплового шока HSP90 и белок RPM1. Эти белки связаны с сигнальными путями фитогормонов и кальций-зависимыми протеинкиназами, играющими существенную роль в адаптации растений к водному дефициту. Было идентифицировано несколько микроРНК (*MIR7757*, *MIR9653a*, *MIR9671*, *MIR9672b*), ранее не обсуждавшихся в контексте адаптации пшеницы к засухе, которые являются кандидатами для дальнейших экспериментальных исследований, направленных на усиление устойчивости пшеницы к недостатку влаги. Полученные результаты могут быть полезными для создания новых сортов пшеницы с повышенной устойчивостью к водному дефициту, что имеет существенное значение для сельского хозяйства в условиях изменения климата.
Ключевые слова: микроРНК; мягкая пшеница; дефицит влаги; гены; генетическая регуляция; ассоциативные генные сети; биоинформатика растений; база знаний Smart crop; программно-информационная система ANDSystem.

Introduction

The productivity of bread wheat (*Triticum aestivum L.*) – a crucial agricultural crop – depends on many environmental factors, including micronutrient availability, temperature, moisture, and soil salinity. Water deficiency is the most important factor limiting wheat productivity (Pakul et al., 2018; Jeyasri et al., 2021). Therefore, studying the physiological and molecular genetic mechanisms of wheat adaptation to water deficiency is an urgent task, the solution of which is necessary for developing new drought-resistant varieties (Langridge, Reynolds, 2021) and improving agricultural technologies.

Plant resistance to insufficient moisture conditions is ensured by several physiological and morphological adaptations, which include enhanced apical growth and inhibition of lateral root growth, leaf abscission, changes in development rate, maintenance of tissue osmotic pressure, reduced transpiration through changes in stomatal apparatus functioning, and activation of cellular antioxidant defense. The functioning of these physiological mechanisms is provided by the coordinated work of numerous genes. It has been shown that water deficiency causes changes in the expression of genes activated by abscisic acid, genes encoding glutathione S-transferase (GST), and the dehydrin protein family (Ferdous et al., 2015).

Signal perception by receptors on the cell wall and cell membrane leads to the activation of intracellular signaling cascades, mainly due to increased levels of reactive oxygen species (ROS) and changes in calcium ion levels. Additionally, important mediators coordinating the initiation of signaling cascades are phytohormones such as abscisic acid (ABA), jasmonic acid (JA), salicylic acid (SA), and ethylene (ET). Stress-activated signaling cascades include, in particular, mitogen-activated protein kinase (MAPK) and calcium-dependent protein kinase (CDPK) signaling pathways. Kinases and phosphatases activate or suppress various transcription factors, which in turn regulate the activity of genes controlling adaptation to adverse conditions (Baillo et al., 2019).

Currently, five gene families are known to encode transcription factors regulating adaptation processes to water deficiency: bZIP (mainly AREB/ABF), DREB (AP2/EREBP), MYB/MYC, NAC, and WRKY (Gahlaut et al., 2016). Literature analysis shows that modification of these transcription factors through genetic engineering methods can enhance plant resistance to adverse environmental factors. For example, transgenic wheat plants containing the Arabidopsis (*Arabidopsis thaliana*) *DREB1A* gene showed increased resistance to drought and salt stress without yield reduction (Pellegrineschi et al., 2004). C.F. Niu and colleagues (2012) obtained transgenic wheat plants with increased expression of the *TaWRKY2* and *TaWRKY19* genes. These plants demonstrated improved resistance to drought and oxidative stress.

Besides transcription factors, gene expression can also be regulated by signaling non-coding RNA molecules. These include circular RNAs (circRNAs), as well as linear long non-coding RNAs (lncRNAs) and microRNAs (Li N. et al.,

2022). These signaling molecules can regulate the expression of any genes involved in stress response, including transcription factors, and the expression of genes encoding signaling RNAs can also change in response to stress, providing an additional level of regulation.

MicroRNAs are single-stranded non-coding RNA molecules 20–25 nucleotides in length that regulate gene activity in plants by binding to the target gene's messenger RNA, leading to its degradation and translation inhibition (Ma, Hu, 2023). It has been revealed that microRNA expression changes in plants in response to water deficiency, which has been shown for many plant species, including wheat. In Triticeae species under drought conditions, the expression of microRNA genes *miR159*, *miR1137*, *miR1318*, *miR168*, and others changed, with the direction of expression changes depending on tissue type, plant developmental stage, and the duration and intensity of exposure (Alptekin et al., 2017). In response to water deficiency in wheat root tissues, there were changes in the expression of microRNA *miR1119*, its target – transcription factor *MYC2*, as well as changes in the expression of numerous stress-response genes, increased abscisic acid content, and cellular antioxidant system activity (Shamloo-Dashtpagerdi et al., 2023).

Thus, microRNA impact on transcription factors can lead to activity changes in entire gene sets. Therefore, microRNAs can be considered master regulators of gene networks that form regulatory modules together with transcription factors and their target genes, including those ensuring plant adaptation to abiotic stress (Zhang et al., 2022) and plant growth and development (Liebsch, Palatnik, 2020). Consequently, targeting microRNAs and their regulatory module activity could become a tool for genetic manipulation of agricultural crops to achieve optimal growth and development parameters (Wang H., Wang H., 2015).

Bioinformatic methods for integrating and analyzing large omics data, including gene network reconstruction methods, are particularly important in marker-assisted breeding (Chao et al., 2023). Bioinformatic analysis of gene networks can help identify regulatory modules involved in plant adaptation to adverse environmental factors and understand its molecular mechanisms.

Previously, the ANDSystem software and information system was developed for reconstructing gene networks based on information obtained from factographic databases and collected through automatic analysis of scientific publication texts (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022). ANDSystem has been applied to solve problems in various areas of biology and biomedicine, including research on molecular genetic mechanisms of asthma development (Bragina et al., 2014; Saik et al., 2018; Zolotareva et al., 2019), lymphedema (Saik et al., 2019), tuberculosis (Bragina et al., 2016), hepatitis C (Saik et al., 2016), coronavirus infection (Ivanisenko V.A. et al., 2022), Huntington's disease (Bragina et al., 2023), glioma (Rogachev et al., 2021), post-operative delirium (Ivanisenko V.A. et al., 2023), hepatocellular carcinoma (Antropova et al., 2023), and study of the proteomic profile of cosmonauts (Larina et al., 2015; Pastushkova et al., 2019).

In the field of plant biology, ANDSystem has been used for reconstruction and analysis of the regulatory gene network of cell wall functioning in *A. thaliana* L. leaves in response to insufficient moisture (Volyanskaya et al., 2023). Based on ANDSystem, the SOLANUM TUBEROSUM knowledge base was created, containing information about genetic regulation of potato metabolic pathways (Ivanisenko T.V. et al., 2018), and prioritization of potato genes involved in the formation of agronomically valuable plant traits was conducted (Demenkov et al., 2019). It should also be noted that the ANDSystem software and information system was previously used for reconstructing gene networks describing microRNA regulation of the external apoptosis pathway (Khlebodarova et al., 2023).

The aim of this work is to reconstruct and analyze the gene network that regulates wheat adaptation to insufficient moisture conditions through microRNAs.

Materials and methods

Search for information about drought response genes. Information about bread wheat genes experimentally proven to be associated with plant adaptation to drought conditions was extracted from full-text experimental and review articles indexed in PubMed (https://pubmed.ncbi.nlm.nih.gov/) as of September 2024. The search was conducted using keywords "wheat", "Triticum aestivum", "drought", "drought tolerance", "gene", "genetic", "regulation" and their combinations.

Additionally, information about genes related to water deficit response was extracted from the AmiGO gene ontology database for the term "response to water deprivation" (term ID GO:0009414). Furthermore, genes associated with the term "response to water deficiency" in the ANDSystem software and information system were included in the list of drought response genes. As a result, a list of genes shown to be involved in wheat adaptation to water deficit was compiled. This list was used as input data for gene network reconstruction.

Smart crop Knowledge Base. This work utilized the Smart crop knowledge base, which is a specialized version of the ANDSystem software and information system focused on rice and wheat genetics and breeding. Three key modules of ANDSystem were customized for the subject area:

Domain-specific ontology module. This module contains expanded dictionaries covering various research objects, such as genes, proteins, metabolites, non-coding RNAs/microRNAs, biological processes, genetic biomarkers, QTL polymorphisms, plant varieties, breeding-significant qualities, phenotypic traits, diseases, pathogens, pests, resistance markers to plant protection products, molecular targets for chemical plant protection products, biotic and abiotic factors, plant protection products (herbicides), and others. Various databases and ontologies were used to form the dictionaries, including NCBI Gene, ChEBI, MirBase, Gene Ontology, Wheat Ontology, Rice Ontology, Wheat Trait and Phenotype Ontology, The International Herbicide-Resistant Weed Database, and others. The dictionaries were supplemented with synonyms and spelling variants of the names to improve object recognition in texts.

Information extraction module from factographic databases. This module performs automated data extraction from various sources, including relational databases (e. g., ChEBI), ontologies in OBO and OWL formats (using the ROBOT tool), text

files in tabular formats (CSV, TSV), and PSI-MI XML 2.5 formats. Specialized extractor programs were created to process information from databases such as NCBI Gene, ToppGene, GrainGenes, IntAc, and others.

Text mining module using semantic linguistic templates. This module is designed to extract knowledge from text sources (scientific articles, patents) using semantic linguistic templates.

The development of new templates and adaptation of existing ones in ANDSystem allowed for effective identification and extraction of various types of interactions between objects. The templates cover such interaction types as associations, regulation of gene and protein expression and activity, physical interactions, catalytic reactions, participation in biological processes, marker relationships, and others. In total, more than 2,000 templates were developed and used, significantly improving the accuracy and completeness of information extraction.

Customizing ANDSystem for the field of rice and wheat breeding and genetics allowed for the integration of data from various sources and ensured effective extraction and analysis of knowledge necessary for research in this subject area.

Gene network reconstruction and analysis. Gene network reconstruction and analysis were performed using the Query Master of the ANDVisio software module (Demenkov et al., 2012), which serves as the user interface in the ANDSystem and Smart crop systems.

Proteins and genes important for the reconstructed gene network functioning were identified using the "NetworkConnectivity" indicator, which characterizes the number of connections between a given network node and other nodes. Genes and proteins were then ranked according to this indicator to find the most significant nodes in the network. Functional annotation of the gene set (analysis of overrepresentation of Gene Ontology terms and KEGG pathways) represented in the network was conducted using the Database for Annotation, Visualization and Integrated Discovery (DAVID version 2021; https://david.ncifcrf.gov/) with default settings (statistical significance was considered at p-value < 0.05 with Bonferroni correction).

Results and discussion

Analysis of published literature (Nagy et al., 2013; Gupta et al., 2014; Liu et al., 2015; Gahlaut et al., 2016; Shojaee et al., 2022) revealed 130 genes involved in wheat adaptation to moisture deficiency. Additionally, 15 genes were associated with the Gene Ontology term "response to water deprivation" (term ID GO:0009414). Further, using the Smart crop knowledge base of the ANDSystem software and information system, 59 genes involved in wheat adaptation to moisture deficiency were discovered. The resulting list of 204 genes shown to be involved in wheat adaptation to insufficient moisture (drought response genes) is provided in Supplementary Material 1[1]. Using this gene list as input data, we reconstructed an associative gene network, to which we added microRNAs that, according to the Smart crop knowledge base, directly regulate at least one network component. This associative network (Fig. 1) included 75 genes, 98 proteins, and 14 wheat microRNAs, as well as 695 interactions between network components. Of these, the following connection types were represented: 594 connections – "association", 39 – "expression regulation", 21 – "interaction", 18 – "expression", 12 – "activity regulation", 7 – "catalysis", 2 – "expression enhancement", and 1 connection each for "expression suppression" and "coexpression" types. The list of microRNAs and their target genes included in the drought response network, established according to Smart crop data, is shown in Table 1.

These microRNAs primarily target genes encoding transcription factors from the GAMYB (*MybA*, *Ta-GAMyb*, *MYB3R1*), WRKY (*WRKY41*) families, auxin response factor (*ARF22*, *LOC123121554*, *LOC123181091*), MADS-box transcription factor (*WM30*), and SQUAMOSA family transcription factor (*LOC123151797*, *LOC123159884*, *SBP16*).

GAMYB transcription factors, which have highly conserved binding sites with *MIR159a* (Millar et al., 2019), participate in gibberellin-mediated activation of hydrolase gene expression in the seed aleurone layer (Woodger et al., 2003). In vegetative plant parts, *MIR159* suppresses the expression of GAMYB transcription factor, which is a growth inhibitor ensuring normal plant development (Millar et al., 2019). *MIR159* expression changes in response to drought, along with changes in *GAMyb* gene expression in potato (Yang J. et al., 2014) and bread wheat (Liu et al., 2015). Additionally, the *MybA* gene product regulates peroxidase gene expression (Wei et al., 2021), contributing to plant adaptation to adverse environmental factors. *MIR160* targets genes encoding ARF transcription factor, a key component ensuring plant response to auxins (Li Y. et al., 2023) – phytohormones that, in particular, stimulate apical dominance, promoting root length growth, which is a morphological adaptation of plants to moisture deficiency.

Besides transcription factors, another microRNA target in the drought response gene network is the RLK serine/threonine kinase gene, which interacts with calmodulins and participates in plant adaptation to abiotic stress (Virdi et al., 2015).

Thus, analysis of the gene network, which includes genes and proteins, the role of which in drought response has been experimentally shown, identified several microRNAs regulating important nodes of this gene network (transcription factors), with some microRNAs (*MIR1120*, *MIR1120c*, *MIR1130a*, *MIR444a*, *MIR444b*, *MIR7757*, *MIR9674a*, *MIR9677a*, *MIR9773*) not having been previously discussed in literature in connection with wheat adaptation to drought, which may be promising for further research.

However, it should be noted that microRNAs often have many target genes, which may also be components of the drought response gene network, although their role is not currently experimentally established. Additionally, microRNAs can regulate genes controlling stress response not only directly but also through intermediaries. Therefore, using the Smart crop knowledge base, the initial gene network was supplemented with the following components: 1) all predicted, according to Smart crop data, targets of those 14 microRNAs that directly regulate known drought response genes and are listed in Table 1; 2) genes and proteins directly connected to drought response genes, as well as their regulating microRNAs.

[1] Supplementary Materials 1–4 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Kleshchev_Engl_28_8.xlsx

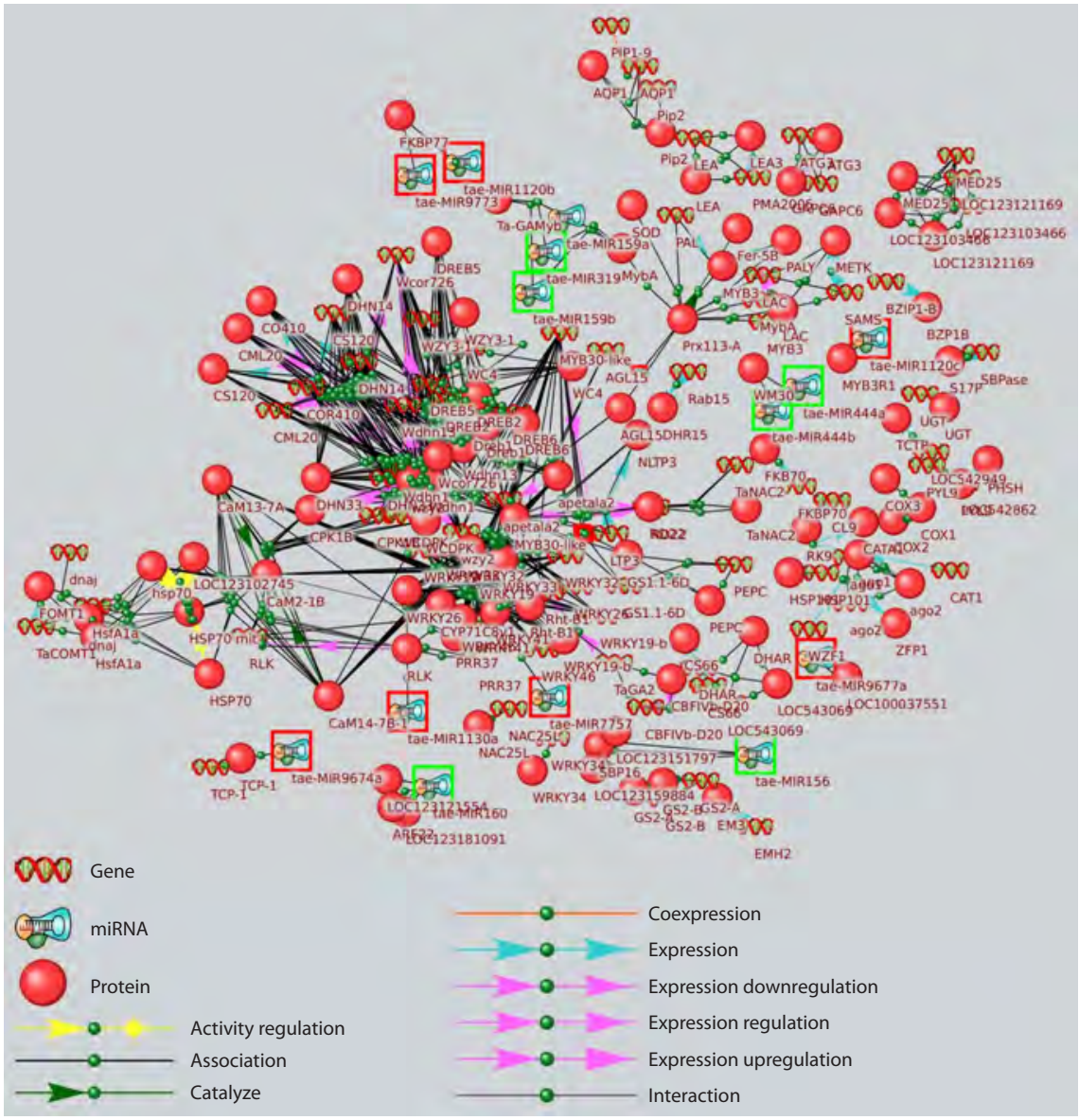


Fig. 1. Associative network of genes and proteins experimentally proven to be involved in wheat adaptation to moisture deficit, supplemented with microRNAs directly regulating them.

Green frames indicate microRNAs with data linking them to drought, red frames indicate microRNAs without such data.

The resulting associative network is presented in Supplementary Material 2. The list of genes and proteins included in this network is provided in Supplementary Material 3. The network includes 144 genes, 1,017 proteins, and 21 wheat microRNAs, as well as 5,188 connections between network components. Of these, 4,158 connections correspond to the "association" type, 372 connections, to "interaction", 329 connections, to "catalysis", 180 connections, to "expression regulation", 42 connections, to "activity regulation", 24 connections, to "cleavage", 21 connections, to "expression", 15 connections, to "expression suppression", 12 connections, to "expression enhancement", and 7 connections, to "coexpression".

Functional annotation of all components (genes and proteins) of the expanded associative gene network is shown in Table 2. As seen from Table 2, gene network components are significantly enriched with terms characterizing biological processes related to centriole assembly, shoot morphogenesis (regulation of morphogenesis of a branching structure), delayed post-embryonic development, response to abiotic and biotic stress factors, and response to abscisic acid. Additionally, gene network components are involved in mitogen-dependent protein kinase and phosphatidylinositol signaling pathways.

Interestingly, the expanded gene network includes genes involved not only in adaptation to water deficit (Gene Ontology term GO:0009414, "response to water deprivation") but also in plant response to other adverse factors, including cold adaptation (Gene Ontology term GO:0009631, "cold acclimation") and interaction with pathogens (KEGG pathway taes04626,

Table 1. List of wheat microRNAs directly regulating drought response genes

microRNA	Target gene	References
MIR160	*ARF22*, *LOC123121554*, *LOC123181091*	Kumar et al., 2015
MIR319	*MybA*, *Ta-GAMyb*	Li Y.-F. et al., 2013
MIR159a	*MybA*, *Ta-GAMyb*	Liu et al., 2015
MIR159b	*MybA*, *Ta-GAMyb*	Liu et al., 2015
MIR1120b	*FKBP77*	–
MIR1120c	*MYB3R1*	–
MIR1130a	*RLK*	–
MIR156	*LOC123151797*, *LOC123159884*, *SBP16*	Singroha et al., 2021
MIR444a	*WM30*	–
MIR444b	*WM30*	–
MIR7757	*WRKY41*	–
MIR9674a	*TCP-1*	–
MIR9677a	*LOC100037551*	–
MIR9773	*FKBP77*	–

Note. microRNAs that have been experimentally shown to change expression in response to moisture deficiency are highlighted in bold. References to the corresponding literature sources are provided for these microRNAs.

Table 2. Functional annotation of the expanded wheat drought response gene network

Term	Number of genes	FE	*p*-value
Biological processes			
GO:0098534~centriole assembly	4	68.3	0.0065
GO:2000032~regulation of secondary shoot formation	10	23.0	0.0000
GO:0060688~regulation of morphogenesis of a branching structure	10	23.0	0.0000
GO:0048581~negative regulation of post-embryonic development	5	20.3	0.0390
GO:0009631~cold acclimation	10	12.6	0.0000
GO:0009414~response to water deprivation	14	11.4	0.0000
GO:0009737~response to abscisic acid	25	6.2	0.0000
GO:0009891~positive regulation of biosynthetic process	32	2.6	0.0013
GO:0045935~positive regulation of nucleobase-containing compound metabolic process	32	2.6	0.0018
GO:0098542~defense response to other organism	89	2.5	0.0000
GO:0051252~regulation of RNA metabolic process	139	1.7	0.0000
GO:0010556~regulation of macromolecule biosynthetic process	155	1.6	0.0000
Molecular functions			
GO:0043531~ADP binding	95	3.2	0.0000
GO:0043565~sequence-specific DNA binding	103	3.1	0.0000
GO:0000976~transcription cis-regulatory region binding	59	3.0	0.0000
GO:0003690~double-stranded DNA binding	60	2.2	0.0000
KEGG Pathways			
taes04016:MAPK signaling pathway – plant	33	4.4	0.0000
taes04626:Plant-pathogen interaction	32	2.9	0.0000
taes04070:Phosphatidylinositol signaling system	12	4.0	0.0161

Note. FE – fold enrichment; *p*-value – statistical significance indicator of gene and protein enrichment in the associative network with Bonferroni correction.

"plant-pathogen interaction"). This is likely due to the fact that products of the same genes can participate in plant response to various stress factors, ensuring plant adaptation to a complex of adverse factors. In particular, genes in our gene network associated with the term "cold acclimation" (GO:0009631) belong to the families of dehydrins and cold-shock proteins.

It is known that proteins of the dehydrin family, by participating in cell membrane stabilization, contribute to plant adaptation to various abiotic stress factors, including moisture deficiency, temperature reduction, and soil salinity (Szlachtowska, Rurek, 2023). On the other hand, cold-shock proteins, which are crucial participants in plant cold adaptation, can also play a certain role in plant response to moisture deficit by regulating the activity of genes, the products of which participate in cellular antioxidant defense (Yu T.F. et al., 2017; Li C. et al., 2021a). Additionally, according to literature, such components of the drought response gene network as calmodulins (Cheval et al., 2013) and WRKY transcription factors (Wani et al., 2021) can also participate in regulating plant immunity and protecting plants from pathogens.

In the expanded associative gene network, the highest number of connections with other network components (Network Connectivity) was found for MYB30-like transcription factor, calmodulin proteins (CaM13-7A, CaM14-7B-1, CaM2-1B), APETALA2-like protein, which is a member of the APETALA2 (AP2) subfamily of AP2/Ethylene Responsive Factor (ERF) transcription factors, as well as RHT1 protein, WRKY41 transcription factor, and cytochrome P450 (CYP71C8v1). Genes encoding these proteins have already been discussed in literature as controlling plant response to moisture deficit.

MYB transcription factors are among the most common families of transcription factors in plants that participate in plant development and response to various adverse environmental factors, including moisture deficiency. MYB transcription factors, by binding to MYB cis-elements in promoters of multiple target genes, regulate a number of biological processes, particularly flavonoid biosynthesis, which is necessary for protection against oxidative stress. Additionally, MYB transcription factors activate genes controlling epicuticular wax formation, which reduces moisture evaporation from plant leaves (Wang X. et al., 2021).

It is known that calcium is a crucial secondary messenger, the concentration of which changes in response to various adverse factors, including moisture deficiency. Calmodulins and calmodulin-like proteins, by binding to calcium ions, change their conformation and modulate the activity of numerous other proteins, including kinases, transcription factors, transporters, and enzymes of various metabolic pathways that ensure plant adaptation to the environment (Ranty et al., 2016). In particular, increased expression of a gene encoding one of the calmodulin family proteins in wheat was observed in response to moisture deficiency and increased salinity, and expression of this gene in transgenic Arabidopsis plants increased their resistance to these adverse factors (Li Y. et al., 2022).

Proteins of the APETALA2 (AP2) subfamily belong to the AP2/Ethylene Responsive Factor (ERF) family of transcription factors, which regulate the expression of genes providing adaptation to adverse environmental conditions, including drought (Park S.Y., Grabau, 2016; Srivastava, Kumar, 2018). Expression of genes encoding AP2 subfamily proteins, *TaAP2-1-1A*, *TaAP2-1-1D*, was increased in response to drought in wheat (Yu Y. et al., 2022).

Cytochrome P450 family proteins are enzymes involved in multiple metabolic pathways for the synthesis of plant secondary metabolites, phytohormones, and antioxidants, which play an important role in plant adaptation to the environment (Pandian et al., 2020). In the study (Li Y., Wei, 2020), it was shown that in wheat, in response to drought, there were changes in the expression of 77 genes encoding cytochrome P450s, which participate in the biosynthesis of abscisic acid, an important mediator activating various signaling cascades in plant stress responses, as well as cytochrome P450s involved in the synthesis of flavonoids, which play an important role in plant cell antioxidant defense.

Among the intermediary proteins connected to experimentally found drought response genes, the gene *LOC123186119*, encoding the disease resistance protein RPM1, had the highest number of connections with other network components. It is connected to all WRKY family transcription factors represented in the network, as well as to calcium-dependent protein kinases 7 and 19. Additionally, the RPM1 protein is a target of microRNA *MIR7757*. The list of 21 microRNAs associated with components of the expanded gene network is shown in Table 3. The complete list of 984 predicted microRNA targets according to the Smart crop database is presented in Supplementary Material 4.

The results of functional annotation of microRNA target genes in the associative network are shown in Table 4. As seen from Table 4, microRNA targets in the drought response gene network are involved in morphogenesis processes of plant lateral shoots and roots, as well as plant immunity, purine transport and metabolism, and transcription factor functioning. Genes controlling shoot morphogenesis processes in the expanded gene network (see Supplementary Material 2) mainly include targets of microRNA *miR319*, encoding the TEOSINTE BRANCHED/CYCLOIDEA/PCF (TCP) transcription factor family, which is involved in forming plant shoot and root architecture (Tokizawa et al., 2023), including root hair formation (Wang M.Y. et al., 2013), which is an important morphological adaptation of plants to moisture deficiency.

The involvement of TEOSINTE BRANCHED/CYCLOIDEA/PCF (TCP) family transcription factors in response to insufficient moisture is discussed in literature (Manna et al., 2021), although their participation in moisture deficit response has not been shown for wheat. Knockout of miR319 family members *IbmiR319a* and *IbmiR319c* in transgenic sweet potato plants led to increased sensitivity to moisture deficiency, increased number of stomata, decreased lignin content, and disruption of hormonal regulation of plant growth (Ren et al., 2022). The authors suggest that these morphological changes are caused by changes in the expression of transcription factor TCP11/17, which is a target of *IbmiR319a* and *IbmiR319c*.

Among the 21 microRNAs in the expanded gene network (see Table 3), 14 were directly connected to genes, the role of which in wheat adaptation to moisture deficiency has been experimentally proven. Seven microRNAs (*MIR9668*, *MIR1121*,

Table 3. List of microRNAs and their target genes with the highest number of connections in the network

microRNA	MirBaseID	Number of microRNA targets contained in the network	Target with maximum number of connections represented in the network	Number of connections the target has in the network
MIR1120b	MI0030404	149	***FKBP77***	4
MIR1130a	MI0006192	148	*LOC123051594*, *LOC123091557*, *LOC123096508*	5
MIR159b	MI0006171	143	***MybA***	11
MIR7757	MI0030410	102	*LOC123186119* (*RPM1*) и *WRKY41*	49
MIR1120c	MI0030409	98	*LOC123078649*	3
MIR444a	MI0006178	58	*LOC123078649*	2
MIR444b	MI0016467	58	*LOC100037552*	2
MIR9773	MI0031525	54	***FKBP77***	4
MIR9674a	MI0030403	47	***TCP-1***	1
MIR159a	MI0006170	36	***MybA***	11
MIR156	MI0016450	24	***SBP16***	1
MIR319	MI0016453	22	***MybA***	11
MIR160	MI0006172	20	***ARF22***	1
MIR9677a	MI0030414	18	***LOC100037551***	1
MIR9668	MI0030392	2	*LOC543328*	7
MIR1121	MI0006183	1	*UCRIA*	3
MIR395b	MI0016464	1	*LOC123190485*	3
MIR9653a	MI0030370	1	*LOC543111*	25
MIR9671	MI0030395	1	*LOC543244*	36
MIR9672b	MI0031526	1	*LOC543244*	36
MIR9679	MI0030418	1	*LOC123114245*	5

Note. MicroRNAs and target genes with known significance in drought adaptation in wheat are highlighted in bold. MicroRNAs that directly regulate known water deficit response genes are underlined.

Table 4. Functional annotation of wheat microRNA target genes in the drought response gene network

Term	Number of genes	FE	*p*-value
GO:2000032~regulation of secondary shoot formation	10	32.03	0.0000
GO:1905428~regulation of plant organ formation	10	32.03	0.0000
GO:0060688~regulation of morphogenesis of a branching structure	10	32.03	0.0000
GO:0098542~defense response to other organism	87	3.36	0.0000
GO:0015211~purine nucleoside transmembrane transporter activity	6	11.89	0.0216
GO:0043531~ADP binding	95	4.34	0.0000
GO:0043565~sequence-specific DNA binding	48	1.95	0.0027
GO:0003700~DNA-binding transcription factor activity	60	1.81	0.0006
GO:0032559~adenyl ribonucleotide binding	161	1.48	0.0000
GO:0017076~purine nucleotide binding	172	1.42	0.0000
GO:0032553~ribonucleotide binding	166	1.41	0.0000

Note. FE – foldenrichment, *p*-value – statistical significance indicator of enrichment with Bonferroni correction.

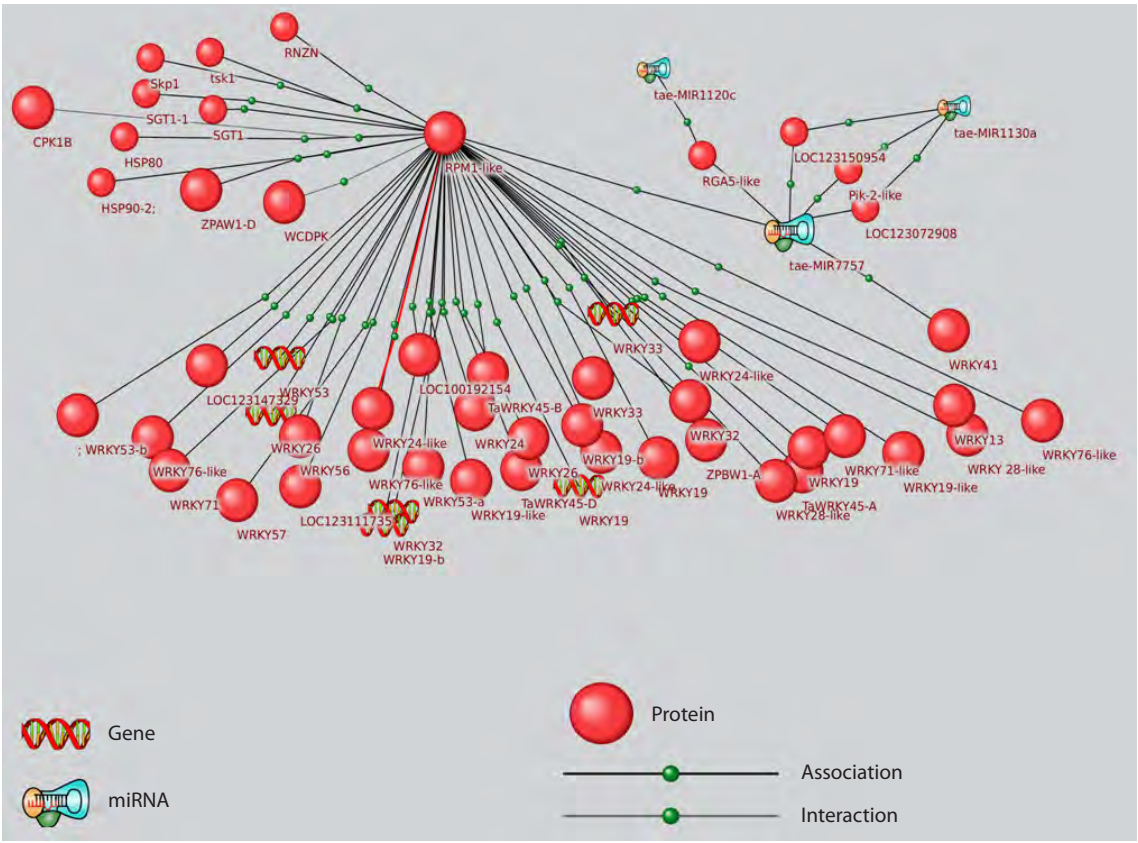


Fig. 2. Associative network of microRNA *MIR7757*, its targets, and intermediaries connected to the targets.
Large spheres indicate proteins experimentally shown to be involved in drought response.

MIR395b, *MIR9653a*, *MIR9671*, *MIR9672b*, *MIR9679*) were connected to drought response genes through an intermediary. MicroRNAs *MIR1120b*, *MIR1130a*, *MIR159b*, *MIR7757* и *MIR1120c* had the highest number of connections with other network components.

In particular, it is interesting to note that not only did microRNA *MIR7757* have connections with many network nodes (102), but its target, the *LOC123186119* gene encoding disease resistance protein RPM1, was connected to the highest number (49) of other network nodes. These nodes include a set of WRKY family transcription factors, as well as calcium-dependent protein kinases 7 (WCDPK) and 19 (CPK 1B), SKP1 and SGT1 proteins, and heat shock protein HSP80 (Fig. 2).

Numerous data obtained from different plant species indicate that WRKY family transcription factors play a crucial role in adaptation to various stress factors, including moisture deficit. Increased expression of WRKY transcription factors contributes to reduced ion loss, activation of leaf stomatal apparatus, decreased moisture loss, and reduced reactive oxygen species content (Khoso et al., 2022).

It is known that WRKY transcription factors modulate the activity of signaling pathways of phytohormones – salicylic acid, ethylene, abscisic acid, jasmonic acid, mitogen-activated protein kinase MAPK (Jiang et al., 2017), as well as calmodulins, including through physical interaction with the calcium domain in calmodulins (Park C.Y. et al., 2005). The activity of WRKY transcription factors is controlled by various signaling pathways and phytohormones, including ethylene (Li J. et al., 2006), abscisic acid (Chen et al., 2010), and MAPK signaling pathway (Mao et al., 2011), which ensures changes in WRKY activity depending on environmental conditions. Thus, WRKY transcription factors are a crucial regulatory link in plant stress response, affecting the activity of multiple genes regulating adaptation, while WRKY activity can change depending on the nature of the impact, providing flexible plant adaptation to changing environmental conditions.

Calmodulins and calcium-dependent protein kinases, by binding to calcium ions, the concentration of which increases in response to stress factors, change the functioning of abscisic acid signaling pathways, which in turn causes changes in seed maturation rate, stomatal closure, and reduced reactive oxygen species content (Asano et al., 2012).

The SKP1 protein is part of the SCF (Skp1-Cullin 1-F-box) complex, which is a ubiquitin ligase playing an important role in hormonal signal transmission, circadian rhythm regulation, plant growth and development (Hong et al., 2012), and adaptation to adverse factors (Saxena et al., 2023). Thus, *MIR7757* may be a crucial master regulator of the moisture deficit response gene network, acting both directly on the WRKY41 transcription factor and through an intermediary – RPM1-like protein, coordinating phytohormone signaling pathways, MAPK, and calcium-dependent protein kinases. This protein plays an important role in plant immunity; however, its signifi-

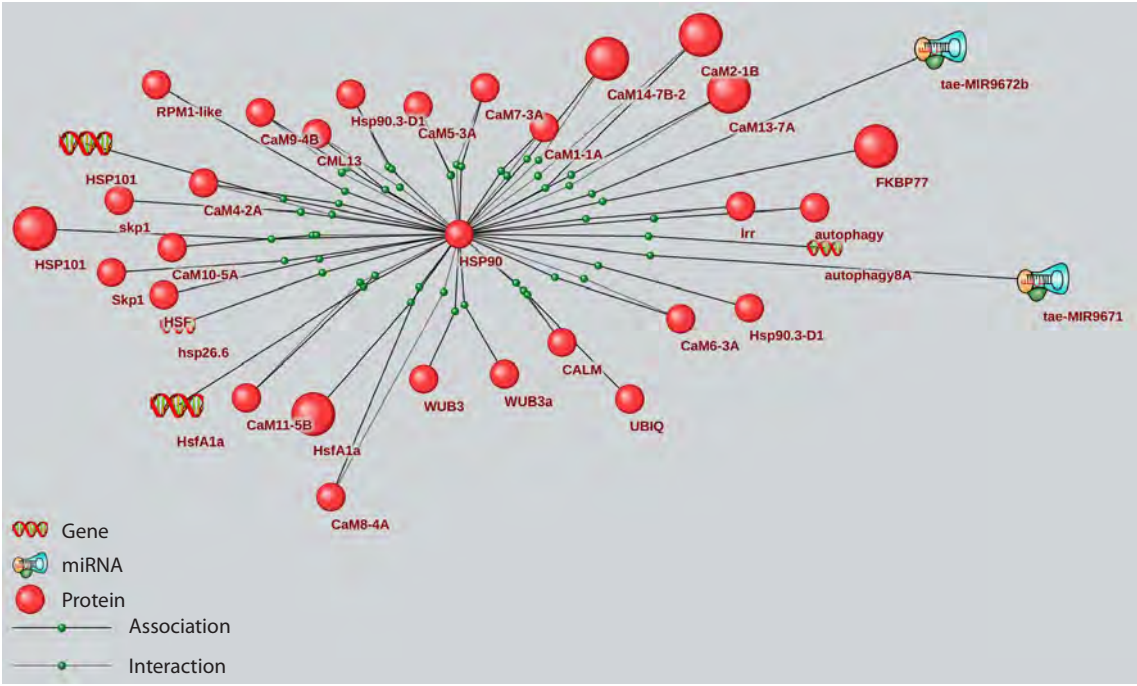


Fig. 3. Associative network of microRNAs *MIR9671*, *MIR9672b*, their targets, and proteins connected to the targets.
Large spheres indicate proteins experimentally shown to be involved in drought response.

cance in wheat response to water deficit is unknown, although it was reported that *PRM1* gene expression was increased in grape leaves in response to moisture deficiency (Haider et al., 2017). Additionally, there is no data on changes in wheat *MIR7757* microRNA expression under moisture deficiency; therefore, this microRNA, as well as other microRNAs with a high number of network node connections and their target genes, are promising candidates for experimental investigation of microRNA regulation of wheat response to water deficit.

The target of two other microRNAs, *MIR9671* and *MIR9672b*, heat shock protein 90, encoded by the *LOC543244* gene, also has extensive (n = 36) connections with other gene network nodes, namely calmodulins (CaM14-7B-1, LOC123104984, etc.), heat shock protein 101, SKP1 and RPM1 proteins discussed above, heat stress transcription factor HSf1a, and polyubiquitin UBIQ (Fig. 3).

It is known that the HSP90 protein, a highly conserved chaperone, is a crucial component of eukaryotic cell homeostasis and participates in plant adaptation to various types of abiotic stress, modulation of plant growth and development by interacting with auxin and jasmonic acid signaling pathways. The HSP90 protein, together with its co-chaperones, stabilizes the auxin receptor complex under conditions of increased air temperature (an environmental factor that often accompanies moisture deficiency) and promotes physiological and morphological adaptations induced by auxin, particularly root elongation (di Donato, Geisler, 2019). Additionally, HSP90, by interacting with protein ligases, assists in the removal of damaged proteins.

It should be noted that numerous calmodulin proteins, by binding to calcium ions during stress, not only activate calcium-dependent protein kinase signaling pathways but also activate HSP90 expression (Virdi et al., 2011), providing additional heat shock protein-mediated activation of the plant hormonal system. Thus, microRNAs *MIR9671* and *MIR9672b*, through their target HSP90, can modulate hormonal signaling of auxin and jasmonic acid, as well as the functioning of the protein ubiquitination system during abiotic stress.

Considering the important role of the HSP90 protein in response to abiotic stress, it can be hypothesized that enhancing its expression by artificially weakening the activity or expression of microRNAs *MIR9671*, *MIR9672b* may increase wheat plant resistance to moisture deficiency. However, it should be noted that HSP90 has a pleiotropic effect, affecting a significant number of cell signaling pathways (di Donato, Geisler, 2019), therefore microRNA-mediated weakening of its expression may be necessary for adaptive changes in some signaling pathways at a certain stage of plant development or during environmental changes.

Thus, microRNAs *MIR9671*, *MIR9672b*, along with *MIR7757*, which were not previously discussed in literature in connection with wheat response to drought, may be promising for further experimental investigation of microRNA regulation of bread wheat response to water deficit.

Several experiments conducted on various plant species have shown that artificial modulation of microRNA expression allows changing regulatory gene network functioning, affecting the expression of genes responsible for adaptation to adverse environmental conditions or the formation of certain economically valuable traits. Modern genetic engineering technologies – RNA interference, creation of special vectors expressing specific microRNAs, as well as genome editing methods such as CRISPR/Cas9 and Transcription activator-

like effector nucleases (TALEN) – make it possible to enhance or weaken microRNA expression and activity depending on whether the products of microRNA target genes have a stimulating or weakening effect on target biological processes (Abbas et al., 2022; Raza et al., 2023). For example, an artificial increase in *miR319* expression using special vectors, as well as an artificial decrease in expression of its targets, TEOSINTE BRANCHED/CYCLOIDEA/PCF (TCP) transcription factors, led to increased cold resistance in rice plants (Yang C. et al., 2013). In another study (Ni et al., 2013), suppression of *miR169* expression increased expression of its target – transcription factor *GmNFYA3*, which improved soybean plant drought resistance through changes in expression of genes related to water deficit adaptation. An artificial increase in *miR172b-3p* expression in transgenic potato led to weakened expression of the *ERFRAP2-7-like* gene and enhanced carbon fixation by plants (Raza et al., 2023).

Additionally, patents have been obtained for transgenic plants, created using microRNAs, that have increased productivity and resistance to adverse environmental factors, demonstrating successful practical use of microRNAs for creating new plant varieties. For example, transgenic rice with increased expression of *Osa-miR393* microRNA and enhanced tillering was obtained (patent CN102533760A (Wang S., Zhang, 2011)). Transgenic tomato plants with suppressed expression of *miR156e-3p* microRNA and increased resistance to low temperatures were created (patent CN111705077B (Zhou et al., 2020)).

Thus, modulation of microRNA activity using genetic engineering may become a promising method of modern biotechnology aimed at increasing plant resistance to adverse environmental conditions, including moisture deficiency, and ultimately their productivity.

Conclusion

Using the Smart crop knowledge base of the ANDSystem software and information system, reconstruction of the gene network of microRNA regulation of bread wheat adaptation to moisture deficiency was performed. Genes in the network regulate root and shoot morphogenesis processes, response to abiotic and biotic stress factors, and are involved in signaling pathways of abscisic acid and calcium-dependent protein kinases.

Twenty-one microRNAs regulating the wheat drought response gene network were identified, the targets of which are mainly involved in controlling plant morphogenesis processes. The most significant nodes in this network regulated by microRNAs are MYBa and WRKY41 transcription factors, HSP90 heat shock protein, and RPM1 protein, which is connected to WRKY family transcription factor proteins, calcium-dependent protein kinases, and phytohormone signaling pathways – auxin, jasmonic acid, and abscisic acid, which are crucial in controlling plant adaptation to moisture deficiency. Several microRNAs that were not previously discussed in literature in connection with drought adaptation (*MIR7757*, *MIR9671*, *MIR9672b*) regulate a significant number of network nodes and therefore may be promising for further experimental investigation of microRNA regulation of bread wheat response to water deficit.

References

Abbas A., Shah A.N., Tanveer M., Ahmed W., Shah A.A., Fiaz S., Waqas M.M., Ullah S. MiRNA fine tuning for crop improvement: using advance computational models and biotechnological tools. *Mol. Biol. Rep*. 2022;49(6):5437-5450. doi 10.1007/s11033-022-07231-5

Alptekin B., Langridge P., Budak H. Abiotic stress miRNomes in the *Triticeae*. *Funct. Integr. Genomics*. 2017;17(2-3):145-170. doi 10.1007/s10142-016-0525-9

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

Asano T., Hayashi N., Kikuchi S., Ohsugi R. CDPK-mediated abiotic stress signaling. *Plant Signal Behav.* 2012;7(7):817-821. doi 10.4161/psb.20351

Baillo E.H., Kimotho R.N., Zhang Z., Xu P. Transcription factors associated with abiotic and biotic stress tolerance and their potential for crops improvement. *Genes* (*Basel*). 2019;10(10):771. doi 10.3390/genes10100771

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

Bragina E.Y., Gomboeva D.E., Saik O.V., Ivanisenko V.A., Freidin M.B., Nazarenko M.S., Puzyrev V.P. Apoptosis genes as a key to identification of inverse comorbidity of huntington's disease and cancer. *Int. J. Mol. Sci*. 2023;24(11):9385. doi 10.3390/ijms24119385

Chao H., Zhang S., Hu Y., Ni Q., Xin S., Zhao L., Ivanisenko V.A., Orlov Y.L., Chen M. Integrating omics databases for enhanced crop breeding. *J. Integr. Bioinform*. 2023;20(4):20230012. doi 10.1515/jib-2023-0012

Chen H., Lai Z., Shi J., Xiao Y., Chen Z., Xu X. Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress. *BMC Plant Biol*. 2010;10:281. doi 10.1186/1471-2229-10-281

Cheval C., Aldon D., Galaud J.P., Ranty B. Calcium/calmodulin-mediated regulation of plant immunity. *Biochim. Biophys. Acta*. 2013;1833(7):1766-1771. doi 10.1016/j.bbamcr.2013.01.031

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

Demenkov P.S., Saik O.V., Ivanisenko T.V., Kolchanov N.A., Kochetov A.V., Ivanisenko V.A. Prioritization of potato genes involved in the formation of agronomically valuable traits using the SOLANUM TUBEROSUM knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2019;23(3):312-319. doi 10.18699/VJ19.501

di Donato M., Geisler M. HSP90 and co-chaperones: a multitaskers' view on plant hormone biology. *FEBS Lett*. 2019;593(13):1415-1430. doi 10.1002/1873-3468.13499

Ferdous J., Hussain S.S., Shi B.J. Role of microRNAs in plant drought tolerance. *Plant Biotechnol. J*. 2015;13(3):293-305. doi 10.1111/pbi.12318

Gahlaut V., Jaiswal V., Kumar A., Gupta P.K. Transcription factors involved in drought tolerance and their possible role in developing

drought tolerant cultivars with emphasis on wheat (*Triticum aestivum* L.). *Theor. Appl. Genet*. 2016;129(11):2019-2042. doi 10.1007/s00122-016-2794-z

Gupta O.P., Meena N.L., Sharma I., Sharma P. Differential regulation of microRNAs in response to osmotic, salt and cold stresses in wheat. *Mol. Biol. Rep*. 2014;41(7):4623-4629. doi 10.1007/s11033-014-3333-0

Haider M.S., Kurjogi M.M., Khalil-Ur-Rehman M., Fiaz M., Pervaiz T., Jiu S., Haifeng J., Chen W., Fang J. Grapevine immune signaling network in response to drought stress as revealed by transcriptomic analysis. *Plant Physiol. Biochem*. 2017;121:187-195. doi 10.1016/j.plaphy.2017.10.026

Hong M.J., Kim D.Y., Kang S.Y., Kim D.S., Kim J.B., Seo Y.W. Wheat F-box protein recruits proteins and regulates their abundance during wheat spike development. *Mol. Biol. Rep*. 2012;39(10):9681-9696. doi 10.1007/s11033-012-1833-3

Ivanisenko T.V., Sayk O.V., Demenkov P.S., Khlestkin V.K., Khlestkina E.K., Kolchanov N.A., Ivanisenko V.A. The SOLANUM TUBEROSUM knowledge base: the section on molecular-genetic regulation of metabolic pathways. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018;22(1):8-17. doi 10.18699/VJ18.325 (in Russian)

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 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., 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., Karpenko T.N., Velichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolomics and gene regulatory networks analysis reveal the role of non-structural 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., 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 Journal of Genetics and Breeding*. 2023;27(7):768-775. doi 10.18699/VJGB-23-89

Jeyasri R., Muthuramalingam P., Satish L., Pandian S.K., Chen J.T., Ahmar S., Wang X., Mora-Poblete F., Ramesh M. An overview of abiotic stress in cereal crops: negative impacts, regulation, biotechnology and integrated omics. *Plants* (*Basel*). 2021;10(7):1472. doi 10.3390/plants10071472

Jiang J., Ma S., Ye N., Jiang M., Cao J., Zhang J. WRKY transcription factors in plant responses to stresses. *J. Integr. Plant Biol*. 2017;59(2):86-101. doi 10.1111/jipb.12513

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

Khoso M.A., Hussain A., Ritonga F.N., Ali Q., Channa M.M., Alshegaihi R.M., Meng Q., Ali M., Zaman W., Brohi R.D., Liu F., Manghwar H. WRKY transcription factors (TFs): Molecular switches to regulate drought, temperature, and salinity stresses in plants. *Front. Plant Sci*. 2022;13:1039329. doi 10.3389/fpls.2022.1039329

Kumar R.R., Pathak H., Sharma S.K., Kala Y.K., Nirjal M.K., Singh G.P., Goswami S., Rai R.D. Novel and conserved heat-responsive microRNAs in wheat (*Triticum aestivum* L.). *Funct. Integr. Genomics*. 2015;15(3):323-348. doi 10.1007/s10142-014-0421-0

Langridge P., Reynolds M. Breeding for drought and heat tolerance in wheat. *Theor. Appl. Genet*. 2021;134(6):1753-1769. doi 10.1007/s00122-021-03795-1

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., Hou N., Fang N., He J., Ma Z., Ma F., Guan Q., Li X. Cold shock protein 3 plays a negative role in apple drought tolerance by regulating oxidative stress response. *Plant Physiol. Biochem*. 2021a;168:83-92. doi 10.1016/j.plaphy.2021.10.003

Li C., Li L., Reynolds M.P., Wang J., Chang X., Mao X., Jing R. Recognizing the hidden half in wheat: root system attributes associated with drought tolerance. *J. Exp. Bot*. 2021b;72(14):5117-5133. doi 10.1093/jxb/erab124

Li J., Brader G., Kariola T., Palva E.T. WRKY70 modulates the selection of signaling pathways in plant defense. *Plant J*. 2006;46(3):477-491. doi 10.1111/j.1365-313X.2006.02712.x

Li N., Liu T., Guo F., Yang J., Shi Y., Wang S., Sun D. Identification of long non-coding RNA-microRNA-mRNA regulatory modules and their potential roles in drought stress response in wheat (*Triticum aestivum* L.). *Front. Plant Sci*. 2022;13:1011064. doi 10.3389/fpls.2022.1011064

Li Y., Wei K. Comparative functional genomics analysis of cytochrome P450 gene superfamily in wheat and maize. *BMC Plant Biol*. 2020;20(1):93. doi 10.1186/s12870-020-2288-7

Li Y., Zhang H., Dong F., Zou J., Gao C., Zhu Z., Liu Y. Multiple roles of wheat calmodulin genes during stress treatment and TaCAM2-D as a positive regulator in response to drought and salt tolerance. *Int. J. Biol. Macromol*. 2022;220:985-997. doi 10.1016/j.ijbiomac.2022.08.124

Li Y., Han S., Qi Y. Advances in structure and function of auxin response factor in plants. *J. Integr. Plant Biol*. 2023;65(3):617-632. doi 10.1111/jipb.13392

Li Y.-F., Zheng Y., Jagadeeswaran G., Sunkar R. Characterization of small RNAs and their target genes in wheat seedlings using sequencing-based approaches. *Plant Sci*. 2013;203-204:17-24. doi 10.1016/j.plantsci.2012.12.014

Liebsch D., Palatnik J.F. MicroRNA miR396, GRF transcription factors and GIF co-regulators: a conserved plant growth regulatory module with potential for breeding and biotechnology. *Curr. Opin. Plant Biol*. 2020;53:31-42. doi 10.1016/j.pbi.2019.09.008

Liu J., Feng L., Li J., He Z. Genetic and epigenetic control of plant heat responses. *Front. Plant Sci*. 2015;6:267. doi 10.3389/fpls.2015.00267

Ma Z., Hu L. MicroRNA: a dynamic player from signalling to abiotic tolerance in plants. *Int. J. Mol. Sci*. 2023;24(14):11364. doi 10.3390/ijms241411364

Manna M., Thakur T., Chirom O., Mandlik R., Deshmukh R., Salvi P. Transcription factors as key molecular target to strengthen the

drought stress tolerance in plants. *Physiol. Plant.* 2021;172(2):847-868. doi 10.1111/ppl.13268

Mao G., Meng X., Liu Y., Zheng Z., Chen Z., Zhang S. Phosphorylation of a WRKY transcription factor by two pathogen-responsive MAPKs drives phytoalexin biosynthesis in *Arabidopsis*. *Plant Cell*. 2011;23(4):1639-1653. doi 10.1105/tpc.111.084996

Millar A.A., Lohe A., Wong G. Biology and function of miR159 in plants. *Plants* (*Basel*). 2019;8(8):255. doi 10.3390/plants8080255

Nagy Z., Németh E., Guóth A., Bona L., Wodala B., Pécsváradi A. Metabolic indicators of drought stress tolerance in wheat: glutamine synthetase isoenzymes and Rubisco. *Plant Physiol. Biochem*. 2013; 67:48-54. doi 10.1016/j.plaphy.2013.03.001

Ni Z., Hu Z., Jiang Q., Zhang H. *GmNFYA3*, a target gene of miR169, is a positive regulator of plant tolerance to drought stress. *Plant Mol. Biol*. 2013;82(1-2):113-129. doi 10.1007/s11103-013-0040-5

Niu C.F., Wei W., Zhou Q.Y., Tian A.G., Hao Y.J., Zhang W.K., Ma B., Lin Q., Zhang Z.B., Zhang J.S., Chen S.Y. Wheat *WRKY* genes *TaWRKY2* and *TaWRKY19* regulate abiotic stress tolerance in transgenic *Arabidopsis* plants. *Plant Cell Environ*. 2012;35(6):1156-1170. doi 10.1111/j.1365-3040.2012.02480.x

Pakul A.L., Lapshinov N.A., Bozhanova G.V., Pakul V.N. The main factors influencing efficiency of spring common wheat agrocenosis. *Sibirskii Vestnik Sel'skokhozyajstvennoi Nauki = Siberian Herald of Agricultural Science*. 2018;48(6):21-29. doi 10.26898/0370-8799-2018-6-3 (in Russian)

Pandian B.A., Sathishraj R., Djanaguiraman M., Prasad P.V.V., Jugulam M. Role of cytochrome P450 enzymes in plant stress response. *Antioxidants* (*Basel*). 2020;9(5):454. doi 10.3390/antiox9050454

Park C.Y., Lee J.H., Yoo J.H., Moon B.C., Choi M.S., Kang Y.H., Lee S.M., Kim H.S., Kang K.Y., Chung W.S., Lim C.O., Cho M.J. WRKY group IId transcription factors interact with calmodulin. *FEBS Lett*. 2005;579(6):1545-1550. doi 10.1016/j.febslet.2005.01.057

Park S.Y., Grabau E. Differential isoform expression and protein localization from alternatively spliced Apetala2 in peanut under drought stress. *J. Plant Physiol*. 2016;206:98-102. doi 10.1016/j.jplph.2016.09.007

Pastushkova L., Kashirina D.N., Brzhozovskiy A.G., Kononikhin A.S., Tiys E.S., Ivanisenko V.A., Koloteva M.I., Nikolaev E.N., Larina I.M. Evaluation of cardiovascular system state by urine proteome after manned space flight. *Acta Astronaut.* 2019;160:594-600. doi 10.1016/j.actaastro.2019.02.015

Pellegrineschi A., Reynolds M., Pacheco M., Brito R.M., Almeraya R., Yamaguchi-Shinozaki K., Hoisington D. Stress-induced expression in wheat of the *Arabidopsis thaliana* DREB1A gene delays water stress symptoms under greenhouse conditions. *Genome*. 2004; 47(3):493-500. doi 10.1139/g03-140

Ranty B., Aldon D., Cotelle V., Galaud J.P., Thuleau P., Mazars C. Calcium sensors as key hubs in plant responses to biotic and abiotic stresses. *Front. Plant Sci*. 2016;7:327. doi 10.3389/fpls.2016.00327

Raza A., Charagh S., Karikari B., Sharif R., Yadav V., Mubarik M.S., Habib M., Zhuang Y., Zhang C., Chen H., Varshney R.K., Zhuang W. miRNAs for crop improvement. *Plant Physiol. Biochem*. 2023;201: 107857. doi 10.1016/j.plaphy.2023.107857

Ren L., Zhang T., Wu H., Ge X., Wan H., Chen S., Li Z., Ma D., Wang A. Blocking IbmiR319a impacts plant architecture and reduces drought tolerance in sweet potato. *Genes* (*Basel*). 2022;13(3): 404. doi 10.3390/genes13030404

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. doi 10.3390/metabo11030133

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., 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

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

Saxena H., Negi H., Sharma B. Role of F-box E3-ubiquitin ligases in plant development and stress responses. *Plant Cell Rep*. 2023; 42(7):1133-1146. doi 10.1007/s00299-023-03023-8

Shamloo-Dashtpagerdi R., Shahriari A.G., Tahmasebi A., Vetukuri R.R. Potential role of the regulatory miR1119-MYC2 module in wheat (*Triticum aestivum* L.) drought tolerance. *Front. Plant Sci*. 2023;14: 1161245. doi 10.3389/fpls.2023.1161245

Shojaee S., Ravash R., Shiran B., Ebrahimie E. Meta-analysis highlights the key drought responsive genes in genes: *PEPC* and *TaSAG7* are hubs response networks. *J. Genet. Eng. Biotechnol.* 2022;20(1): 127. doi 10.1186/s43141-022-00395-4

Srivastava R., Kumar R. The expanding roles of APETALA2/Ethylene Responsive Factors and their potential applications in crop improvement. *Brief. Funct. Genomics*. 2018;18(4):240-254. doi 10.1093/bfgp/elz001

Szlachtowska Z., Rurek M. Plant dehydrins and dehydrin-like proteins: characterization and participation in abiotic stress response. *Front. Plant Sci*. 2023;14:1213188. doi 10.3389/fpls.2023.1213188

Tokizawa M., Enomoto T., Chandnani R., Mora-Macías J., Burbridge C., Armenta-Medina A., Kobayashi Y., Yamamoto Y.Y., Koyama H., Kochian L.V. The transcription factors, STOP1 and TCP20, are required for root system architecture alterations in response to nitrate deficiency. *Proc. Natl. Acad. Sci. USA*. 2023;120(35):e2300446120. doi 10.1073/pnas.2300446120

Virdi A.S., Pareek A., Singh P. Evidence for the possible involvement of calmodulin in regulation of steady state levels of Hsp90 family members (Hsp87 and Hsp85) in response to heat shock in sorghum. *Plant Signal. Behav*. 2011;6(3):393-399. doi 10.4161/psb.6.3.13867

Virdi A.S., Singh S., Singh P. Abiotic stress responses in plants: roles of calmodulin-regulated proteins. *Front. Plant Sci.* 2015;6:809. doi 10.3389/fpls.2015.00809

Volyanskaya A.R., Antropova E.A., Zubairova U.S., Demenkov P.S., Venzel A.S., Orlov Y.L., Makarova A.A., Ivanisenko T.V., Gorshkova T.A., Aglyamova A.R., Kolchanov N.A., Chen M., Ivanisenko V.A. Reconstruction and analysis of the gene regulatory network for cell wall function in *Arabidopsis thaliana* L. leaves in response to water deficit. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(8):1031-1041. doi 10.18699/VJGB-23-118

Wang H., Wang H. The miR156/SPL module, a regulatory hub and versatile toolbox, gears up crops for enhanced agronomic traits. *Mol. Plant.* 2015;8(5):677-688. doi 10.1016/j.molp.2015.01.008

Wang M.Y., Zhao P.M., Cheng H.Q., Han L.B., Wu X.M., Gao P., Wang H.Y., Yang C.L., Zhong N.Q., Zuo J.R., Xia G.X. The cotton transcription factor TCP14 functions in auxin-mediated epidermal cell differentiation and elongation. *Plant Physiol.* 2013;162(3): 1669-1680. doi 10.1104/pp.113.215673

Wang S., Zhang M. Small-molecule ribonucleic acid (RNA) OsamiR393 for improving rice tillering and application. 2011. https://patents.google.com/patent/CN102533760A/en?oq=CN102533760A

Wang X., Niu Y., Zheng Y. Multiple functions of MYB transcription factors in abiotic stress responses. *Int. J. Mol. Sci*. 2021;22(11): 6125. doi 10.3390/ijms22116125

Wani S.H., Anand S., Singh B., Bohra A., Joshi R. WRKY transcription factors and plant defense responses: latest discoveries and future prospects. *Plant Cell Rep*. 2021;40(7):1071-1085. doi 10.1007/s00299-021-02691-8

Wei T., Guo D., Liu J. PtrMYB3, a R2R3-MYB transcription factor from *Poncirus trifoliata*, negatively regulates salt tolerance and hydrogen peroxide scavenging. *Antioxidants* (*Basel*). 2021;10(9):1388. doi 10.3390/antiox10091388

Woodger F.J., Gubler F., Pogson B.J., Jacobsen J.V. A Mak-like kinase is a repressor of GAMYB in barley aleurone. *Plant J*. 2003;33(4):707-717. doi 10.1046/j.1365-313x.2003.01663.x

Yang C., Li D., Mao D., Liu X., Ji C., Li X., Zhao X., Cheng Z., Chen C., Zhu L. Overexpression of microRNA319 impacts leaf morphogenesis and leads to enhanced cold tolerance in rice (*Oryza sativa* L.). *Plant Cell Environ*. 2013;36(12):2207-2218. doi 10.1111/pce.12130

Yang J., Zhang N., Mi X., Wu L., Ma R., Zhu X., Yao L., Jin X., Si H., Wang D. Identification of miR159s and their target genes and expression analysis under drought stress in potato. *Comput. Biol. Chem*. 2014;53(Part B):204-213. doi 10.1016/j.compbiolchem.2014.09.009

Yu T.F., Xu Z.S., Guo J.K., Wang Y.X., Abernathy B., Fu J.D., Chen X., Zhou Y.B., Chen M., Ye X.G., Ma Y.Z. Improved drought tolerance in wheat plants overexpressing a synthetic bacterial cold shock protein gene *SeCspA*. *Sci Rep*. 2017;7:44050. doi 10.1038/srep44050

Yu Y., Yu M., Zhang S., Song T., Zhang M., Zhou H., Wang Y., Xiang J., Zhang X. Transcriptomic identification of wheat AP2/ERF transcription factors and functional characterization of *TaERF-6-3A* in response to drought and salinity stresses. *Int. J. Mol. Sci*. 2022;23(6):3272. doi 10.3390/ijms23063272

Zhang F., Yang J., Zhang N., Wu J., Si H. Roles of microRNAs in abiotic stress response and characteristics regulation of plant. *Front. Plant Sci*. 2022;13:919243. doi 10.3389/fpls.2022.919243

Zhou Y., Zhang L., Yu J. Application of tomato miR156e-3p gene in improvement of tomato low-temperature resistance and plant overexpression vector. 2020. https://patents.google.com/patent/CN111705077B/en?oq=CN111705077B

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 September 22, 2024. Revised October 31, 2024. Accepted November 2, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):918-926

doi 10.18699/vjgb-24-99

Root cap border cells as regulators of rhizosphere microbiota

N.A. Omelyanchuk[1], V.A. Cherenko[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. A rhizosphere (a narrow area of soil around plant roots) is an ecological niche, within which beneficial microorganisms and pathogens compete with each other for organic carbon compounds and for the opportunity to colonize roots. The roots secrete rhizodeposits into the rhizosphere, which include border cells, products of root cell death and liquids secreted by living cells (root exudates). Border cells, which have their name due to their location in the soil next to the root (at the border of the root and soil), represent terminal differentiation of columella and adjacent lateral root cap cells. Border cells can detach from the root cap surface both as single cells and as cell layers. Border cells are constantly supplied to the soil throughout plant life, and the type and intensity of border cells' sloughing depend on both plant species and soil conditions. Currently, data on the factors that control the type of border cells' release and its regulation have been described in different plant species. Border cells are specialized for interaction with the environment, in particular, they are a living barrier between soil microbiota and roots. After separation of border cells from the root tip, transcription of primary metabolism genes decreases, whereas transcription of secondary metabolism genes as well as the synthesis and secretion of mucilage containing these metabolites along with extracellular DNA, proteoglycans and other substances increase. The mucilage that the border cells are embedded in serves both to attract microorganisms promoting plant growth and to protect plants from pathogens. In this review, we describe interactions of border cells with various types of microorganisms and demonstrate their importance for plant growth and disease resistance.

Key words: root; border cells; biotic stress; plant defense against pathogens; soil symbionts.

For citation: Omelyanchuk N.A., Cherenko V.A., Zemlyanskaya E.V. Root cap border cells as regulators of rhizosphere microbiota. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):918-926. doi 10.18699/vjgb-24-99

Funding. The work was funded by the budget project FWNR-2022-0005.

Пограничные клетки корневого чехлика как регулятор ризосферной микробиоты

Н.А. Омельянчук[1], В.А. Черенко[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

Аннотация. Ризосфера (почва, окружающая корни растения) – это экологическая ниша, внутри которой полезные микроорганизмы и патогены конкурируют друг с другом за органические углеродные соединения и возможность колонизации корней. Для взаимодействия с микробиотой корни выделяют в ризосферу ризодепозиты, к которым относят пограничные клетки, продукты гибели клеток корня и секретируемые живыми клетками жидкости (корневые экссудаты). Пограничные клетки, получившие свое название ввиду их локализации в почве рядом с корнем (на границе корня и почвы), представляют собой конечный этап дифференцировки клеток корневого чехлика. Слущивание пограничных клеток с поверхности корневого чехлика может происходить как одиночными клетками, так и рядами клеток. Пограничные клетки постоянно поставляются в почву на протяжении всей жизни растения, а тип и интенсивность слущивания пограничных клеток определяются как видом растений, так и почвенными условиями. В настоящее время появились данные о факторах, контролирующих тип слущивания, а также исследования этого процесса и его регуляции у разных видов растений. Пограничные клетки специализированы для взаимодействия с внешней средой, в частности, они служат живым барьером между корнем и почвенной микробиотой. После отделения от кончика корня в пограничных клетках снижается уровень первичного метаболизма и повышается число транскриптов генов вторичного метаболизма, усиливаются синтез компонентов и выделение слизи, содержащей вторичные метаболиты, внеклеточную ДНК, протеогликаны и другие вещества. Слизь, в которую пограничные клетки оказываются погруженными, служит

как для привлечения микроорганизмов, способствующих росту растения, так и для защиты корня от патогенов. В настоящем обзоре описаны взаимодействия пограничных клеток с различными видами микроорганизмов и продемонстрирована их важность для роста растений и их устойчивости к болезням. Эти аспекты могут быть использованы в генной инженерии и селекции для усиления полезных функций пограничных клеток, что, в свою очередь, откроет новые горизонты для повышения урожайности и устойчивости сельскохозяйственных культур.
Ключевые слова: корень; пограничные клетки; биотический стресс; защита растений от патогенов; почвенные симбионты.

Introduction

Plant roots are surrounded by a large number of microorganisms: in the rhizosphere (the narrow soil zone directly contacting roots), one gram of soil contains ~10^8–10^9 bacteria, 10^5–10^6 fungi, and 10^3–10^5 algae and protozoa (Mendes et al., 2013). This metabolically active microbiota modifies soil properties and influences both root and overall plant growth. In turn, the root system penetrates deeply into the soil, altering it by releasing rhizodeposits, living and dead cells, and various organic compounds that affect the composition and abundance of microbial populations. A substantial part of rhizodeposits consists of cells regularly sloughed from the surface of the root cap, a small organ located at the very tip of the root (Hawes et al., 2011). These sloughed cells, called border cells, are named for their position at the root-soil boundary (Hawes, Lin, 1990). Border cells are living cells that secrete mucilage containing polysaccharides, proteins, and a range of other substances (Driouich et al., 2021). This mucilage forms a matrix, in which the border cells become embedded. As the root grows, border cells interact with the cells located above the root cap and can be found at considerable distances from the root tip, where they originated from (Hawes, Lin, 1990; Driouich et al., 2019).

Border cells have been described in ferns, gymnosperms, and angiosperms (Vermeer, McCully, 1982; Hawes et al., 2003; Forino et al., 2012). The number of viable border cells per root depends on the plant family and also varies with root growth. In young roots (up to 2 cm), this number ranges from 800 in *Bromus carinatus* and 11,000 in *Cucumis sativus* to 17,000 in *Zea mays*, with a significant reduction in roots longer than 9 cm, to 70, 300, and 150 cells, respectively (Odell et al., 2008; Darshan et al., 2020). The number of border cells can even vary among different ecotypes of the same species and depends on growth conditions (Zhao et al., 2000; Iijima et al., 2003; Pankievicz et al., 2022). For example, when pea plants are exposed to high levels of carbon dioxide, the production of border cells doubles compared to normal conditions (Zhao et al., 2000).

Border cells are “renewable”, i. e. they are constantly supplied to the soil and have a definite lifespan (Driouich et al., 2019). For example, the root system of a single pea plant produces approximately 3,000–4,000 border cells per day. The duration for which border cells remain viable after being sloughed from the root cap surface varies among plant species, ranging from several days in *Arabidopsis* (Vicré et al., 2005; Plancot et al., 2013) to several weeks in maize (Vermeer, McCully, 1982). In many angiosperm families (such as grasses, legumes, and cucurbits), the outermost layer of the root cap detaches as individual viable border cells, with no connections between them (Driouich et al., 2007). In contrast, in some other families, such as Brassicaceae (including the model species *Arabidopsis thaliana* L.), living cells are sloughed off as a single layer (Fendrych et al., 2014). Therefore, these cells are classified as a distinct group, “border-like cells” (Vicré et al., 2005; Driouich et al., 2007; Plancot et al., 2013). Additionally, an alternative term has been proposed to encompass both border cells and border-like cells: “root associated, cap-derived cells” (root AC-DC) (Driouich et al., 2019).

At present, new data have emerged on factors controlling the sloughing mode of the outer root cap cells and functions of border cells in different plant species. According to these data, border cells can be defined as living cells sloughed off from the root cap into the environment as individual cells, layers of cells, or multilayered aggregates and serving specialized functions in supporting plant growth and defense responses (Darshan et al., 2020). Accordingly, we will use the general term “border cells” regardless of their sloughing type.

In this review, we examine in detail the factors determining the sloughing type of border cells, describe the differences between border cells and other root tip cells, their secretory function, and the formation of rhizosphere microbiota under the influence of border cell secretions.

Border cell differentiation and sloughing modes in diverse plant species

In *A. thaliana*, the root cap consists of two distinct parts: the centrally located columella and the lateral root cap (LRC), which surrounds the columella and root meristem located above (Dolan et al., 1993). In the transition zone, the outer LRC cells undergo programmed cell death with rapid autolysis, and these processes progress toward the root tip (Fendrych et al., 2014). In contrast, the outer columella cells together with a few adjacent outer LRC cells detach as a single layer of living cells (Vicré et al., 2005; Durand et al., 2009). Initially, a gap is formed in the outer LRC layer slightly above the quiescent center, followed by detachment of cells in this layer, culminating in separation of the outer columella cell layer (Fig. 1*a*) (Shi et al., 2018). The entire process, from the initial gap to the complete detachment, takes approximately 18 hours, with another 18 hours passing before the new outermost layer begins to slough off. It is important to note that the cells sloughing from the root cap fit the original definition of border cells – they are located at the boundary between the root and the soil (Hawes and Lin, 1990). Moreover, in *A. thaliana*, up to 12 % of roots of Columbia ecotype seedlings produce individual, isolated border cells (Karve et al., 2016).

The primary components of middle lamellae – the parts of the cell walls that “glue” neighboring cells together – are pectins (polygalacturonans composed of homogalacturonans,

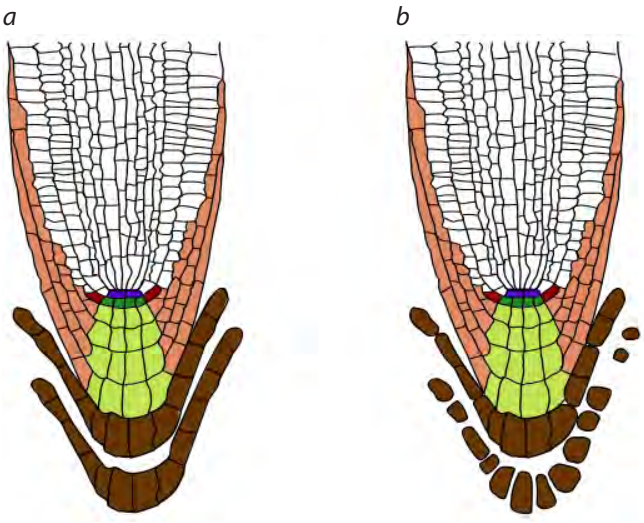


Fig. 1. Sloughing of border cells as a single layer (*a*) and as individual cells (*b*) in *A. thaliana* seedlings.

a – root tip of a wild-type seedling; *b* – root tip of an *nlp7* mutant. Blue indicates quiescent center, dark green represents columella initials, light green denotes columella, light brown indicates lateral root cap (LRC), red depicts epidermis/LRC initials, and dark brown indicates border cells. The schematic representations are based on data from (Karve et al., 2016).

rhamnogalacturonans, and substituted galacturonans) (Caffall, Mohnen, 2009; Albersheim et al., 2010). Pectins are synthesized within the cell and subsequently secreted to the cell wall predominantly in a methyl-esterified form (Atmodjo et al., 2013). In the cell wall, pectin methylesterases remove methyl groups, generating free carboxyl groups on galacturonic acid residues of polygalacturonans. This leads to a local pH decrease, and acidification that promotes the activity of polygalacturonases, which hydrolyze polygalacturonans (Moustacas et al., 1991; Micheli, 2001). This mechanism explains disintegration of the border cell layer into individual cells in *A. thaliana* in response to low pH stress (Karve et al., 2016). The role of pectins in border cell sloughing has also been demonstrated in pea (Wen et al., 1999). When the expression of a gene encoding pectin methylesterase is inhibited, border cells fail to detach from the root. In *A. thaliana* mutants for the *QUASIMODO 1/2* genes, which exhibit reduced production of one component of pectin – homogalacturonan (a linear polymer of galacturonic acid) – root cap cells slough off as individual cells (Durand et al., 2009).

In *A. thaliana*, NIN-LIKE PROTEIN7 (NLP7) transcription factor regulates sloughing of the border cells as a whole layer (Karve et al., 2016). Loss-of-function mutation *nlp7* enhances sloughing of individual border cells from the root cap surface (Fig. 1*b*) (Karve et al., 2016). While only 12 % of wild type roots exhibited release of individual border cells, it was observed in 44 % of roots in *nlp7* mutants. In these mutants, the levels of cellulose and pectin are reduced, and genes encoding cellulase (CEL5) and pectin lyases – the enzymes that weaken the cell wall – are activated. In *A. thaliana*, *CEL5* inactivation results in a decreased rate of border cell sloughing (Del Campillo et al., 2004). Similarly, individual border cell sloughing occurs upon loss of function of *AUTOPHAGY 5* (*ATG5*), one of the key regulators of autophagy (Goh et al., 2022). In *atg5* mutants, border cells fail to form autophagosomes and a central vacuole.

There is significant diversity in the modes of border cells' sloughing. For example, in *Acacia mangium*, a tropical tree of the legume family, LRC-derived border cells slough off acropetally (towards the root apex) from the root transition zone as sheets composed of several cell rows, while columella cells slough as separate border cells (Endo et al., 2011). Among three leguminous tree species native to sub-Saharan Africa, *Balanites aegyptiaca* exhibits separate sloughing of root cap cells, whereas in *Acacia raddiana* and *Tamarindus indica*, sloughing occurs both as individual cells and in layers (Carreras et al., 2020). In *Pinus densiflora*, individual elongated border cells are released from the central region of the root cap, while sheath-shaped long layers of cells slough from the lateral sides (Shirakawa et al., 2023).

In soybean, three morphotypes of border cells have been identified: spherical, intermediate, and elongated (Ropitaux et al., 2020). Spherical border cells are predominantly localized near the root cap, intermediate cells surround the root in the meristematic zone, while elongated cells encircle the root in the elongation and differentiation zones (Fig. 2). Elongated cells constitute more than 30 % of border cells and can occur either as single cells or as groups of tens or several dozen cells tightly attached to one another. Approximately 80 % of elongated cells and 50 % of spherical border cells are viable. In maize, spherical cells detach from the columella, whereas the LRC produces elongated cells (Guinel, McCully, 1987). In banana, elongated (ellipsoidal) cells make up 92 % of border cells, with the remaining 8 % being spherical cells containing amyloplasts (Wuyts et al., 2006). In potato, small spherical border cells were observed in the root cap region, whereas elongated cells were primarily localized around the elongation zone (Koroney et al., 2016). Both cell types contained starch.

Thus, outer root cap cells can be removed from its surface via programmed cell death and subsequent rapid autolysis, as well as through the detachment of interconnected or separated living cells. Subsequently, death of border cells in the soil produces cellular debris, which serves as a nutrient source for the microbiota. Compared to root tip cells, border cells exhibit reduced primary metabolism and increased expression of secondary metabolism genes, which encode proteins for the synthesis of wax, phenylpropanoids, lignin, phenolic compounds, and flavonoids (Watson et al., 2015).

Large starch reserves in the border cells provide energy and carbon source necessary for secondary metabolite synthesis. Additionally, border cells synthesize a unique set of proteins: 13 % of proteins produced in border cells are not detectable in the root tip (Brigham et al., 1995). Thus, border cells represent the final stage of the root cap cells differentiation. Taken together, it is evident that differentiation and sloughing of the border cells is an energy-consuming process. This raises the question: for what significant purposes do plants release a large number of living cells from the root cap periphery in a regulated manner. Undoubtedly, this implies the crucial role of border cells in interactions with the root environment.

Composition and functions of mucilage secreted by border cells

The process of how precursors of the border cells acquire the ability to secrete mucilage has been described in detail for columella cells in *Arabidopsis* (Maeda et al., 2019). Provided that columella initials are designated as the c1 layer, when cells transit from c5 to c6, mucilage begins accumulating along the lateral cell walls, while the shootward cell walls start degrading (Fig. 3). In c7 cells, most of the mucilage is released into the intercellular space between the c6 and c7 layers. In parallel, a vacuole develops, and amyloplasts undergo degradation. After the border cells' separation, the mucilage from the intercellular space passes into the rhizosphere, while border cells continue its secretion. Thus, border cells become surrounded by dense, fibrillar mucilage (Ropitaux et al., 2020). The Golgi apparatus, essential for secretion, develops in the peripheral columella cells before they separate and become border cells (Poulsen et al., 2008). Golgi-derived vesicles, including those fusing with the plasma membrane, are characteristic of border cells (Driouich et al., 2007; Wang et al., 2017). In soybeans, spherical border cells produce the largest quantity of mucilage, whereas elongated border cells produce the least (Ropitaux et al., 2020).

In most plant species, approximately 94 % of the soluble mucilage fraction consists of neutral and acidic polysaccharides, with the remaining 6 % being proteins (Carminati, Vetterlein, 2013). 25 % of the proteins synthesized by border cells are immediately released into the environment (Brigham et al., 1995). Similarly, the majority of metabolites produced in the border cells are secreted promptly after their synthesis. The root cap mucilage in 3- to 4-day-old maize seedlings contains 2,848 distinct proteins, of which a substantial proportion (25 %) is involved in metabolism. The remaining proteins are functionally related to the cell wall, reactive oxygen species, nutrient acquisition, and stress response (Ma et al., 2010). Interestingly, 85–94 % of the mucilage proteins in *A. thaliana* and rapeseed have homologs present in maize mucilage. This indicates a certain conservation in the protein composition of mucilage between monocotyledons and dicotyledons.

Acidic (pectic) polysaccharides impart gel-like properties to mucilage, i. e. make it a gel with a porous structure. The mucilage secreted by border cells can retain water up to 1,000 times its weight (Guinel, McCully, 1986). In soybeans, the primary component of the fibrous structure within mucilage is the neutral polysaccharide xyloglucan (Ropitaux et al., 2019). Xyloglucan and cellulose form molecular cross-bridges connecting border cells. It is known that the primary cell wall of dicotyledons consists of cellulose and xyloglucan polysaccharides embedded in a matrix of pectins, glycoproteins and proteoglycans (Driouich et al., 2012); thus, border cells secrete cell wall polysaccharides and proteoglycans, which form the matrix and internal structure of mucilage (Castilleux et al., 2018; Driouich et al., 2019).

Among the protein components of the border cells' exudate, hydroxyproline-rich glycoproteins, such as extensin and arabinogalactan proteins, are prominent (Vicré et al., 2005; Plancot et al., 2013). Arabinogalactan proteins have been identified in the mucilage of pea, *Arabidopsis*, rapeseed, and potato (Knee et al., 2001; Durand et al., 2009; Cannesan et al., 2012; Koroney et al., 2016). In addition to these components, mucilage contains phenolic acids, phospholipids, antimicrobial peptides/proteins (defensins, pathogenesis-related proteins, and others), phytoalexins, histone H4, enzymes, extracellular DNA, reactive oxygen species (ROS) toxic to pathogens, and ROS-producing enzymes (Wen et al., 2007, 2017; Carminati, Vetterlein, 2013; Plancot et al., 2013; Weiller et al., 2017).

The mucilage secreted by the border cells and the border cells themselves form a complex known as the "Root Extracellular Trap (RET)" (Driouich et al., 2013). RET shares many features with extracellular traps of animals, produced by phagocytic immune cells (neutrophils, macrophages, mast cells, eosinophils, heterophils) upon stimulation (Driouich et al., 2019, 2021). In both plants and animals, extracellular traps exhibit nonspecific activity against a wide range of

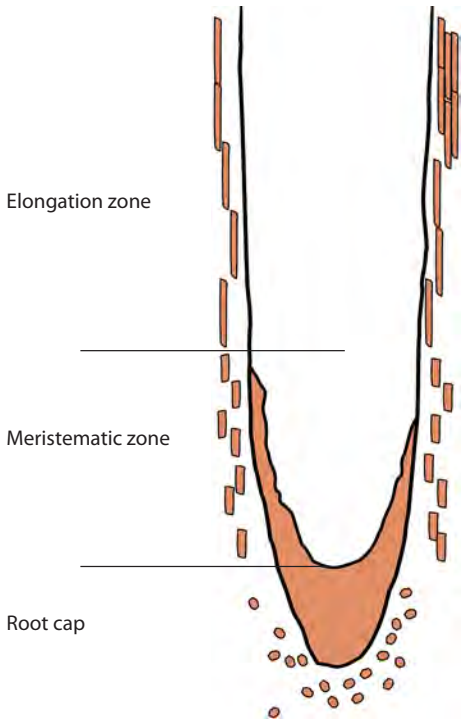


Fig. 2. Three morphotypes of soybean border cells.

The root cap, along with spherical, intermediate, and elongated border cells, are depicted in brown. The scheme was prepared based on data published by (Ropitaux et al., 2020).

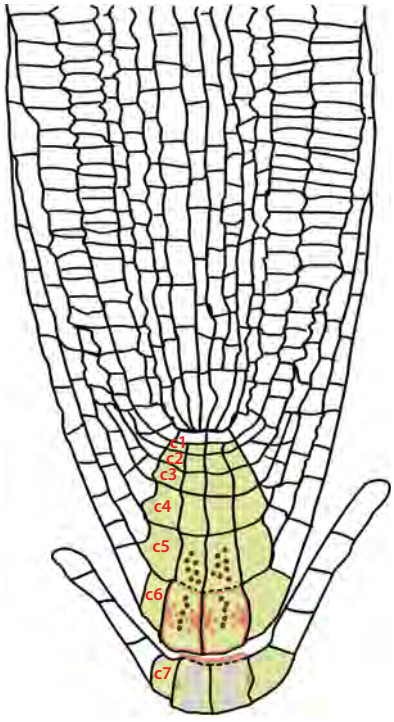


Fig. 3. Differentiation of border cells in the columella of *A. thaliana*.

Columella cells are shown in light green. The columella cell layers are numbered sequentially from c1 (columella initials) to c7. Starch granules are represented by brown dots, mucilage, by red, and vacuoles, by gray.

microbial and fungal pathogens. These traps contain similar defensive components (antimicrobial proteins and extracellular DNA) and perform the same functions – capturing, immobilizing, and destroying pathogens, thereby limiting the spread of microbes to other tissues.

The mechanism of action of extracellular DNA secreted by border cells remains unclear (Monticolo et al., 2020). However, the degradation of extracellular DNA in the border cell exudate with DNase treatment resulted in a loss of root resistance to pathogenic fungi (Wen et al., 2009). Mutations in genes encoding secreted DNases in phytopathogenic bacteria and fungi led to a decrease in the infectivity of these pathogens for plant roots (Hawes et al., 2016; Tran et al., 2016). DNase secretion has been reported in numerous soilborne pathogenic fungal species and certain bacterial species (Darshan et al., 2020). Border cells of pea and tomato secrete extracellular DNA in response to pathogenic bacteria, whereas nonpathogenic bacteria do not induce DNA secretion (Tran et al., 2016).

Human histone H4, which shares 97 % homology with pea histone H4 secreted by border cells, is lethal for *Ralstonia solanacearum*, a bacterium infecting pea roots. The toxic activity of histone H4 is neutralized when the roots are treated with antibodies against this protein (Tran et al., 2016).

Border cells shape microbiota in the rhizosphere

Border cells protect plants and promote their growth by preventing root infection with pathogens or stimulating associations with beneficial microbiota. Co-cultivation of border cells embedded in mucilage with various bacterial species on agar surfaces revealed various bacterial responses to border cells and their exudate (Gochnauer et al., 1990). The observed effects included strong growth inhibition (*Rhizobium sp.* and *Escherichia coli*), strong stimulation (*Pseudomonas fluorescens*), no effect (*Streptomyces sp.* and *Cytophaga sp.*) or initial inhibition followed by strong stimulation and subsequent spore formation (*Bacillus spp.*).

Thus, the composition of the bacterial community in the rhizosphere is determined by the ability of bacterial species to respond to the compounds in the border cell exudate. It can be assumed that, through this mechanism, border cells actively control not only bacteria but also fungi, protists, etc. Besides, the exudate of border cells influences the microbiome composition due to different responses of microbe species to the carbon sources it contains (Knee et al., 2001; Benizri et al., 2007).

Rhizospheric bacteria that are beneficial to plants are classified into a special group called plant growth-promoting rhizobacteria (PGPR) (Hasan et al., 2024). PGPR are diverse in species composition and include representatives of *Agrobacterium*, *Arthrobacter*, *Azotobacter*, *Azospirillum*, *Burkholderia*, *Caulobacter*, *Chromobacterium*, *Erwinia*, *Flavobacterium*, *Micrococcous*, *Pseudomonas*, *Rhizobium*, *Serratia* and other genera. By interacting with roots, these bacteria enhance plant resistance to biotic and abiotic stresses, increase the availability of various elements (iron, potassium, phosphorus, etc.) in the soil, synthesize phytohormones and other metabolites that influence plant growth, and contribute to soil detoxification from many harmful contaminants. Many PGPRs inhibit growth of pathogenic organisms by producing antibiotics (Ulloa-Ogaz et al., 2015).

Actinomycetes not only promote plant growth by themselves, some of their isolates enhance growth and spore germination of arbuscular mycorrhizal fungi beneficial for plants, thereby acting also as mycorrhiza helper bacteria (Franco-Correa et al., 2010). Other actinomycete isolates have demonstrated strong activity against plant pathogenic fungi (Lee, Hwang, 2002). The bacteria *Herbaspirillum seropedicae* forms nitrogen-fixing associations with roots of maize and other cereals (Chubatsu et al., 2012). Notably, humic acids increase both host border cell sloughing and the density of these bacteria in the root tip region (Canellas, Olivares, 2017).

Living border cells are the primary producers of mucilage, which contains substances that attract plant-beneficial microorganisms (Hawes et al., 1998). Border cells secrete compounds, which either stimulate branching of mycorrhizal hyphae or trigger biofilm formation in several beneficial bacteria (Nagahashi, Douds, 2004; Beauregard et al., 2013). The degradation of arabinogalactan proteins by specific agents reduces the colonization of border cells and root tips by *Rhizobium* bacteria (Vicré et al., 2005). In *Pinus densiflora*, during the early stages of root development (prior to mycorrhiza formation), rhizobacteria contacting with border cells and their exudate protects host roots by inhibiting pathogen growth (Shirakawa et al., 2023).

Arbuscular mycorrhizae, widespread soil fungi, form symbiotic associations with many angiosperms, including most agricultural crops (Khaliq et al., 2022). Mycorrhiza improves water and nutrient uptake by plants, especially phosphorus, while plants provide the fungi with 10–20 % of their photosynthates. Moreover, the number of border cells produced by different plant species positively correlates with their ability to form mycorrhizal associations (Niemira et al., 1996; Arriola et al., 1997). One strain of the ascomycete fungus *Trichoderma*, when colonizing border cells of wheat seedlings, caused approximately a 40 % increase in stem biomass and suppressed the growth of pathogenic *Fusarium* species by more than 90 % (Jaroszuk-Ściseł et al., 2019).

It is now evident that a new field in agricultural biotechnology is emerging – rhizosphere microbiome bioengineering, which aims to populate the rhizosphere predominantly with plant-beneficial microorganisms (Mohanram, Kumar, 2019). For instance, bacterial genera such as *Bacillus* and *Pseudomonas* are currently used as biofertilizers and for biological plant protection, including the production of biopreparations against pathogens, serving as natural enemies of pathogens or as inducers of systemic resistance in plants (Hasan et al., 2024). Another promising approach for engineering the rhizosphere microbiome is modification of border cells (Mohanram, Kumar, 2019). The effectiveness of this approach has been demonstrated through the transformation of *Arabidopsis* and potato plants with a gene encoding a peptide-based nematode repellent under the control of the *Arabidopsis MDK4-20* gene promoter (Lilley et al., 2011). This promoter is specifically expressed in root cap cells and border cells, and the transformation resulted in transgenic plants that are resistant to nematodes.

Border cells interact with soil pathogens

The release of border cells, which secrete various compounds into the soil, represents one of the mechanisms utilized by plants to combat pathogens (Hawes et al., 2000). We have previously mentioned antimicrobial functions of the mucilage, mediated by certain proteins, secondary metabolites, and extracellular DNA, which provide protection against some fungi and bacteria (Wen et al., 2009; Cannesan et al., 2011; Koroney et al., 2016; Tran et al., 2016). However, the interaction of border cells with pathogens is not limited to the bactericidal and fungicidal properties of their secreted mucilage. Border cells can perceive specific pathogen signals, known as pathogen-associated molecular patterns (MAMPs/PAMPs), and respond to them with typical MAMP-induced primary immune responses, including the production of reactive oxygen species and reinforcement of cell walls through the accumulation and modification of extensins and the deposition of callose (Plancot et al., 2013).

Pathogen attack can enhance border cells' sloughing, stimulate mucilage production by these cells, or alter its composition (Cannesan et al., 2011; Koroney et al., 2016). For example, treatment of roots with an elicitor derived from *Pectobacterium atrosepticum*, a soilborne potato pathogen, modifies the mucilage composition, including the profile of arabinogalactan proteins (Koroney et al., 2016). The oomycete *Aphanomyces euteiches* causes up to 80 % yield loss in peas by invading their roots, which leads to root growth arrest and plant death (Cannesan et al., 2011). Inoculation of pea roots with *A. euteiche*s increases the number of border cells, and this increase correlates with the quantity of oospores used for inoculation. In response to inoculation, border cells induce the synthesis of pisatin, a phenolic phytoalexin that, at certain concentrations, inhibits hyphal growth and zoospore production *in vitro*.

Thus, enhanced synthesis of this compound may contribute to increased pea root resistance against this infection. Moreover, border cells attract the oomycete via chemotaxis and subsequently neutralize it using antimicrobial components of the mucilage (Hawes et al., 2016). Specifically, arabinogalactan proteins, which are the components of the mucilage and cell walls of the border cells, have been shown to induce encystment and prevent germination of the pathogen's zoospores (Cannesan et al., 2012). Consequently, border cells and their exude prevent zoospore colonization of root tips by blocking their entry into root tissues and inducing their lysis (Ropitaux et al., 2020).

Border cells of rye seedlings neutralize a pathogenic strain of the fungus *Fusarium culmorum* by stimulating spore germination into macroconidia and forming compact clusters with them around the root cap, referred to as mantle-like structures, whereas non-pathogenic strains do not form such structures (Jaroszuk-Ściseł et al., 2009). In addition to well-known mechanisms for suppressing fungal infection (inhibition of spore germination, suppression of fungal pathogenesis gene activity, enhancement of plant defense gene expression), the formation of mantle-like structures on the root tip represents another type of root–pathogen interaction, where the border cells' exude, conversely, induces rapid spore germination followed by border cells death and suppression of fungal growth (Gunawardena et al., 2005).

The formation of mantle-like structures on the root tip was also observed during inoculation of pea roots with the pathogenic fungus *Nectria haematococca*, with most of the root tips remaining intact beneath the mantle-like structure (Gunawardena, Hawes, 2002). In this infection, only about 4 % of the root tips are damaged, whereas in the case of proteolytic degradation of the border cell secretion, all root tips are affected (Wen et al., 2007).

Conclusion

Thus, border cells are viable components of the root system that play a key role in root interactions with rhizosphere microorganisms. After detaching from the root tip, border cells alter their metabolism, synthesizing and releasing hydrated mucilage containing proteoglycans, secondary metabolites, antimicrobial proteins, and extracellular DNA. This mucilage acts as an active agent for attracting beneficial microorganisms that promote plant growth. At the same time, border cells serve as a barrier to pathogens. They secrete various antimicrobial substances, and their primary immune response is triggered by different elicitors. All these aspects can be targeted through genetic engineering and breeding to enhance the beneficial functions of border cells for plants.

References

Albersheim P., Darvill A., Roberts K., Sederoff R., Staehelin A. Plant Cell Walls. From Chemistry to Biology. New York: Garland Science, 2010

Arriola L., Niemira B.A., Safir G.R. Border cells and arbuscular mycorrhizae in four Amaranthaceae species. *Phytopathology*. 1997; 87(12):1240-1242. doi 10.1094/PHYTO.1997.87.12.1240

Atmodjo M.A., Hao Z., Mohnen D. Evolving views of pectin biosynthesis. *Annu. Rev. Plant Biol.* 2013;64:747-779. doi 10.1146/annurev-arplant-042811-105534

Beauregard P.B., Chai Y., Vlamakis H., Losick R., Kolter R. *Bacillus subtilis* biofilm induction by plant polysaccharides. *Proc. Natl. Acad. Sci. USA*. 2013;110(17):1621-1630. doi 10.1073/pnas.1218984110

Benizri E., Nguyen C., Piutti S., Slezack-Deschaumes S., Philippot L. Additions of maize root mucilage to soil changed the structure of the bacterial community. *Soil Biol. Biochem*. 2007;39(5):1230-1233. doi 10.1016/j.soilbio.2006.12.026

Brigham L.A., Woo H.H., Nicoll S.M., Hawes M.C. Differential expression of proteins and mRNAs from border cells and root tips of pea. *Plant Physiol*. 1995;109(2):457-463. doi 10.1104/pp.109.2.457

Caffall K.H., Mohnen D. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydr. Res*. 2009;344: 1879-1900. doi 10.1016/j.carres.2009.05.021

Canellas L.P., Olivares F.L. Production of border cells and colonization of maize root tips by *Herbaspirillum seropedicae* are modulated by humic acid. *Plant Soil*. 2017;417:403-413. doi 10.1007/s11104-017-3267-0

Cannesan M.A., Gangneux C., Lanoue A., Giron D., Laval K., Hawes M., Driouich A., Vicré-Gibouin M. Association between border cell responses and localized root infection by pathogenic *Aphanomyces euteiches*. *Ann. Bot.* 2011;108(3):459-469. doi 10.1093/aob/mcr177

Cannesan M.A., Durand C., Burel C., Gangneux C., Lerouge P., Ishii T., Laval K., Follet-Gueye M.L., Driouich A., Vicré-Gibouin M. Effect of arabinogalactan proteins from the root caps of pea and *Brassica napus* on *Aphanomyces euteiches* zoospore chemotaxis and germi-

nation. *Plant Physiol.* 2012;159(4):1658-1670. doi 10.1104/pp.112.198507

Carminati A., Vetterlein D. Plasticity of rhizosphere hydraulic properties as a key for efficient utilization of scarce resources. *Ann. Bot.* 2013;112(2):277-290. doi 10.1093/aob/mcs262

Carreras A., Bernard S., Durambur G., Gügi B., Loutelier C., Pawlak B., Boulogne I., Vicré M., Driouich A., Goffner D., Follet-Gueye M.L. In vitro characterization of root extracellular trap and exudates of three Sahelian woody plant species. *Planta*. 2020;251(1):19. doi 10.1007/s00425-019-03302-3

Castilleux R., Plancot B., Ropitaux M., Carreras A., Leprince J., Boulogne I., Follet-Gueye M.L., Popper Z.A., Driouich A., Vicré M. Cell wall extensins in root – microbe interactions and root secretions. *J. Exp. Bot*. 2018;69(18):4235-4247. doi 10.1093/jxb/ery238

Chubatsu L.S., Monteiro R.A., de Souza E.M., de Oliveira M.A.S., Yates M.G., Wassem R., Bonatto A.C., Huergo L.F., Steffens M.B.R., Rigo L.U., Pedrosa F.D.O. Nitrogen fixation control in *Herbaspirillum seropedicae*. *Plant Soil*. 2012;356:197-207. doi 10.1007/s11104-011-0819-6

Darshan K., Singh J., Yadav S., Venugopala K.M., Aggarwal R. Root border cells: A pioneer's of plant defence in rhizosphere. *Indian J. Agric. Sci*. 2020;90(10):1850-1855. doi 10.56093/ijas.v90i10.107884

Del Campillo E.D., Abdel-Aziz A., Crawford D., Patterson S.E. Root cap specific expression of an endo-β-1,4-D-glucanase (cellulase): a new marker to study root development in *Arabidopsis*. *Plant Mol. Biol*. 2004;56(2):309-323. doi 10.1007/s11103-004-3380-3

Dolan L., Janmaat K., Willemsen V., Linstead P., Poethig S., Roberts K., Scheres B. Cellular organisation of the *Arabidopsis thaliana* root. *Development*. 1993;119(1):71-84. doi 10.1242/dev.119.1.71

Driouich A., Durand C., Vicre-Gibouin M. Formation and separation of root border cells. *Trends Plant Sci.* 2007;12:14-19. doi 10.1016/j.tplants.2006.11.003

Driouich A., Follet-Gueye M.L., Bernard S., Kousar S., Chevalier L., Vicré-Gibouin M., Lerouxel O. Golgi-mediated synthesis and secretion of matrix polysaccharides of the primary cell wall of higher plants. *Front Plant Sci*. 2012;3:79. doi 10.3389/fpls.2012.00079

Driouich A., Follet-Gueye M.L., Vicré-Gibouin M., Hawes M. Root border cells and secretions as critical elements in plant host defense. *Curr. Opin. Plant Biol*. 2013;16(4):489-495. doi 10.1016/j.pbi.2013.06.010

Driouich A., Smith C., Ropitaux M., Chambard M., Boulogne I., Bernard S., Follet-Gueye M.L., Vicré M., Moore J. Root extracellular traps versus neutrophil extracellular traps in host defence, a case of functional convergence? *Biol. Rev. Camb. Philos. Soc.* 2019;94(5):1685-1700. doi 10.1111/brv.12522

Driouich A., Gaudry A., Pawlak B., Moore J.P. Root cap-derived cells and mucilage: a protective network at the root tip. *Protoplasma*. 2021;258(6):1179-1185. doi 10.1007/s00709-021-01660-y

Durand C., Vicré-Gibouin M., Follet-Gueye M.L., Duponchel L., Moreau M., Lerouge P., Driouich A. The organization pattern of root border-like cells of Arabidopsis is dependent on cell wall homogalacturonan. *Plant Physiol*. 2009;150(3):1411-1421. doi 10.1104/pp.109.136382

Endo I., Tange T., Osawa H. A cell-type-specific defect in border cell formation in the *Acacia mangium* root cap developing an extraordinary sheath of sloughed-off cells. *Ann. Bot*. 2011;108(2):279-290. doi 10.1093/aob/mcr139

Fendrych M., Hautegem T.V., Durme M.V., Olvera-Carrillo Y., Huysmans M., Karimi M., Lippens S., Guérin C.J., Krebs M., Schumacher K., Nowack M.K. Programmed cell death controlled by ANAC033/SOMBRERO determines root cap organ size in *Arabidopsis*. *Curr. Biol*. 2014;24:931. doi 10.1016/j.cub.2014.03.025

Forino L.M.C., Castiglione M.R., Bartoli G., Balestri M., Andreucci A., Tagliasacchi A.M. Arsenic-induced morphogenic response in roots of arsenic hyperaccumulator fern *Pteris vittata*. *J. Hazard. Mater*. 2012;235-236:271-278. doi 10.1016/j.jhazmat.2012.07.051

Franco-Correa M., Quintana A., Duque C., Suarez C., Rodríguez M.X., Barea J.M. Evaluation of actinomycete strains for key traits related with plant growth promotion and mycorrhiza helping activities. *Appl. Soil Ecol*. 2010;45(3):209-217. doi 10.1016/j.apsoil.2010.04.007

Gochnauer M.B., Sealey L.J., McCully M.E. Do detached root-cap cells influence bacteria associated with maize roots? *Plant Cell Environ.* 1990;13(8):793-801. doi 10.1111/j.1365-3040.1990.tb01095.x

Goh T., Sakamoto K., Wang P., Kozono S., Ueno K., Miyashima S., Toyokura K., Fukaki H., Kang B.H., Nakajima K. Autophagy promotes organelle clearance and organized cell separation of living root-ap cells in *Arabidopsis thaliana*. *Development*. 2022;149(11):dev200593. doi 10.1242/dev.200593

Guinel F.C., McCully M.E. Some water-related physical properties of maize root-cap mucilage. *Plant Cell Environ*. 1986;9(8):657-666. doi 10.1111/J.1365-3040.1986.TB01624.X

Guinel F.C., McCully M.E. The cells shed by the root cap of *Zea*: their origin and some structural and physiological properties. *Plant Cell Environ*. 1987;10(7):565-578. doi 10.1111/1365-3040.EP11604101

Gunawardena U., Hawes M.C. Tissue specific localization of root infection by fungal pathogens: role of root border cells. *Mol. Plant Microbe Interact*. 2002;15(11):1128-1136. doi 10.1094/MPMI.2002.15.11.1128

Gunawardena U., Rodriguez M., Straney D., Romeo J.T., VanEtten H.D., Hawes M.C. Tissue-specific localization of pea root infection by *Nectria haematococca*. Mechanisms and consequences. *Plant Physiol*. 2005;137(4):1363-1374. doi 10.1104/pp.104.056366

Hasan A., Tabassum B., Hashim M., Khan N. Role of plant growth promoting rhizobacteria (PGPR) as a plant growth enhancer for sustainable agriculture: A review. *Bacteria*. 2024;3(2):59-75. doi 10.20944/preprints202310.1504.v1

Hawes M., Allen C., Turgeon B.G., Curlango-Rivera G., Minh Tran T., Huskey D.A., Xiong Z. Root border cells and their role in plant defense. *Annu. Rev. Phytopathol*. 2016;54:143-161. doi 10.1146/annurev-phyto-080615-100140

Hawes M.C., Lin H.J. Correlation of pectolytic enzyme activity with the programmed release of cells from root caps of pea (*Pisum sativum*). *Plant Physiol*. 1990;94(4):1855-1859. doi 10.1104/pp.94.4.1855

Hawes M.C., Brigham L.A., Wen F., Woo H.H., Zhu Y. Function of root border cells in plant health: Pioneers in the rhizosphere. *Annu. Rev. Phytopathol*. 1998;36:311-327. doi 10.1146/annurev.phyto.36.1.311

Hawes M.C., Gunawardena U., Miyasaka S., Zhao X. The role of root border cells in plant defense. *Trends Plant Sci*. 2000;5(3):128-133. doi 10.1016/s1360-1385(00)01556-9

Hawes M.C., Bengough G., Cassab G., Ponce G. Root caps and rhizosphere. *J. Plant Growth Regul*. 2003;21:352-367. doi 10.1007/s00344-002-0035-y

Hawes M.C., Curlango-Rivera G., Wen F., White G.J., VanEtten H.D., Xiong Z. Extracellular DNA: the tip of root defenses? *Plant Sci*. 2011;180(6):741-745. doi 10.1016/j.plantsci.2011.02.007

Iijima M., Barlow P.W., Bengough A.G. Root cap structure and cell production rates of maize (*Zea mays*) roots in compacted sand. *New Phytol*. 2003;160(1):127-134. doi 10.1046/j.1469-8137.2003.00860.x

Jaroszuk-Ściseł J., Kurek E., Rodzik B., Winiarczyk K. Interactions between rye (*Secale cereale*) root border cells (RBCs) and pathogenic and nonpathogenic rhizosphere strains of *Fusarium culmorum*. *Mycol. Res*. 2009;113(10):1053-1061. doi 10.1016/j.mycres.2009.07.001

Jaroszuk-Ściseł J., Tyśkiewicz R., Nowak A., Ozimek E., Majewska M., Hanaka A., Tyśkiewicz K., Pawlik A., Janusz G. Phytohormones

(auxin, gibberellin) and ACC deaminase in vitro synthesized by the mycoparasitic *Trichoderma* DEMTkZ3A0 strain and changes in the level of auxin and plant resistance markers in wheat seedlings inoculated with this strain conidia. *Int. J. Mol. Sci*. 2019;20(19):4923. doi 10.3390/ijms20194923

Karve R., Suárez-Román F., Iyer-Pascuzzi A.S. The transcription factor NIN-LIKE PROTEIN7 controls border-like cell release. *Plant Physiol*. 2016;171(3):2101-2111. doi 10.1104/pp.16.00453

Khaliq A., Perveen S., Alamer K.H., Zia Ul Haq M., Rafique Z., Alsudays I.M., Althobaiti A.T., Saleh M.A., Hussain S., Attia H. Arbuscular mycorrhizal fungi symbiosis to enhance plant – soil interaction. *Sustainability*. 2022;14(13):7840. doi 10.3390/su14137840

Knee E.M., Gong F.C., Gao M., Teplitski M., Jones A.R., Foxworthy A., Mort A.J., Bauer W.D. Root mucilage from pea and its utilization by rhizosphere bacteria as a sole carbon source. *Mol. Plant Microbe Interact*. 2001;14(6):775-784. doi 10.1094/MPMI.2001.14.6.775

Koroney A.S., Plasson C., Pawlak B., Sidikou R., Driouich A., Menu-Bouaouiche L., Vicré-Gibouin M. Root exudate of *Solanum tuberosum* is enriched in galactose-containing molecules and impacts the growth of *Pectobacterium atrosepticum*. *Ann. Bot*. 2016;118(4):797-808. doi 10.1093/aob/mcw128

Lee J.Y., Hwang B.K. Diversity of antifungal actinomycetes in various vegetative soils of Korea. *Can. J. Microbiol*. 2002;48(5):407-417. doi 10.1139/w02-025

Lilley C.J., Wang D., Atkinson H.J., Urwin P.E. Effective delivery of a nematode-repellent peptide using a root-cap-specific promoter. *Plant Biotechnol. J*. 2011;9(2):151-161. doi 10.1111/j.1467-7652.2010.00542.x

Ma W., Muthreich N., Liao C., Franz-Wachtel M., Schütz W., Zhang F., Hochholdinger F., Li C. The mucilage proteome of maize (*Zea mays* L.) primary roots. *J. Proteome Res*. 2010;9(6):2968-2976. doi 10.1021/pr901168v

Maeda K., Kunieda T., Tamura K., Hatano K., Hara-Nishimura I., Shimada T. Identification of periplasmic root-cap mucilage in developing columella cells of *Arabidopsis thaliana*. *Plant Cell Physiol*. 2019;60(6):1296-1303. doi 10.1093/pcp/pcz047

Mendes R., Garbeva P., Raaijmakers J.M. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol. Rev*. 2013;37(5):634-663. doi 10.1111/1574-6976.12028

Micheli F. Pectin methylesterases: cell wall enzymes with important roles in plant physiology. *Trends Plant Sci*. 2001;6(9):414-419. doi 10.1016/s1360-1385(01)02045-3

Mohanram S., Kumar P. Rhizosphere microbiome: revisiting the synergy of plant-microbe interactions. *Ann. Microbiol*. 2019;69(3):307-320. doi 10.1007/s13213-019-01448-9

Monticolo F., Palomba E., Termolino P., Chiaiese P., De Alteriis E., Mazzoleni S., Chiusano M.L. The role of DNA in the extracellular environment: a focus on NETs, RETs and biofilms. *Front. Plant Sci*. 2020;11:589837. doi 10.3389/fpls.2020.589837

Moustacas A.M., Nari J., Borel M., Noat G., Ricard J. Pectin methylesterase, metal ions and plant cell-wall extension. The role of metal ions in plant cell-wall extension. *Biochem. J*. 1991;279(2):351-354. doi 10.1042/bj2790343

Nagahashi G., Douds D.D. Isolated root caps, border cells, and mucilage from host roots stimulate hyphal branching of the arbuscular mycorrhizal fungus, *Gigaspora gigantea*. *Mycol. Res*. 2004;108(9):1079-1088. doi 10.1017/s0953756204000693

Niemira B.A., Safir G.R., Hawes M.C. Arbuscular mycorrhizal colonization and border cell production: a possible correlation. *Phytopathology*. 1996;86(6):563-565

Odell R.E., Dumlao M.R., Samar D., Silk W.K. Stage-dependent border cell and carbon flow from roots to rhizosphere. *Am. J. Bot*. 2008;95(4):441-446. doi 10.3732/ajb.95.4.441

Pankievicz V.C.S., Delaux P.M., Infante V., Hirsch H.H., Rajasekar S., Zamora P., Jayaraman D., Calderon C.I., Bennett A., Ané J.M. Nitrogen fixation and mucilage production on maize aerial roots is controlled by aerial root development and border cell functions. *Front. Plant Sci*. 2022;13:977056. doi 10.3389/fpls.2022.977056

Plancot B., Santaella C., Jaber R., Kiefer-Meyer M.C., Follet-Gueye M.L., Leprince J., Gattin I., Souc C., Driouich A., Vicré-Gibouin M. Deciphering the responses of root border-like cells of Arabidopsis and flax to pathogen-derived elicitors. *Plant Physiol*. 2013;163(4):1584-1597. doi 10.1104/pp.113.222356

Poulsen L.R., López-Marqués R.L., McDowell S.C., Okkeri J., Licht D., Schulz A., Pomorski T., Harper J.F., Palmgren M.G. The *Arabidopsis* P_4-ATPase ALA3 localizes to the Golgi and requires a β-subunit to function in lipid translocation and secretory vesicle formation. *Plant Cell*. 2008;20(3):658-676. doi 10.1105/tpc.107.054767

Ropitaux M., Bernard S., Follet-Gueye M.L., Vicré M., Boulogne I., Driouich A. Xyloglucan and cellulose form molecular cross-bridges connecting root border cells in pea (*Pisum sativum*). *Plant Physiol. Biochem*. 2019;139:191-196. doi 10.1016/j.plaphy.2019.03.023

Ropitaux M., Bernard S., Schapman D., Follet-Gueye M.L., Vicré M., Boulogne I., Driouich A. Root border cells and mucilage secretions of soybean, *Glycine max* (Merr) L.: characterization and role in interactions with the oomycete *Phytophthora parasitica*. *Cells*. 2020;9(10):2215. doi 10.3390/cells9102215

Shi C.-L., von Wangenheim D., Herrmann U., Wildhagen M., Kulik I., Kopf A., Ishida T., Olsson V., Anker M.K., Albert M., Butenko M.A., Felix G., Sawa S., Claassen M., Friml J., Aalen R.B. The dynamics of root cap sloughing in *Arabidopsis* is regulated by peptide signalling. *Nat. Plants*. 2018;4(8):596-604. doi 10.1038/s41477-018-0212-z

Shirakawa M., Matsushita N., Fukuda K. Visualization of root extracellular traps in an ectomycorrhizal woody plant (*Pinus densiflora*) and their interactions with root-associated bacteria. *Planta*. 2023;258(6):112. doi 10.1007/s00425-023-04274-1

Tran T.M., MacIntyre A., Hawes M., Allen C. Escaping underground nets: extracellular DNases degrade plant extracellular traps and contribute to virulence of the plant pathogenic bacterium *Ralstonia solanacearum*. *PLoS Pathog*. 2016;12(6):e1005686. doi 10.1371/journal.ppat.1005686

Ulloa-Ogaz A.L., Muñoz-Castellanos L.N., Nevárez-Moorillón G.V. Biocontrol of phytopathogens: Antibiotic production as mechanism of control. In: Méndez-Vilas A. (Ed.). The Battle Against Microbial Pathogens: Basic Science, Technological Advances and Educational Programs. Formatex, 2015;305-309

Vermeer J., McCully M.E. The rhizosphere in *Zea*: New insight into its structure and development. *Planta*. 1982;156:45-61. doi 10.1007/BF00393442

Vicré M., Santaella C., Blanchet S., Gateau A., Driouich A. Root border-like cells of Arabidopsis. Microscopical characterization and role in the interaction with rhizobacteria. *Plant Physiol*. 2005;138:998-1008. doi 10.1104/pp.104.051813

Wang P., Chen X., Goldbeck C., Chung E., Kang B.H. A distinct class of vesicles derived from the *trans*-Golgi mediates secretion of xylogalacturonan in the root border cell. *Plant J*. 2017;92(4):596-610. doi 10.1111/tpj.13704

Watson B.S., Bedair M.F., Urbanczyk-Wochniak E., Huhman D.V., Yang D.S., Allen S.N., Li W., Tang Y., Sumner L.W. Integrated metabolomics and transcriptomics reveal enhanced specialized metabolism in *Medicago truncatula* root border cells. *Plant Physiol*. 2015;167(4):1699-1716. doi 10.1104/pp.114.253054

Weiller F., Moore J.P., Young P., Driouich A., Vivier M.A. The Brassicaceae species *Heliophila coronopifolia* produces root border-like cells that protect the root tip and secrete defensin peptides. *Ann. Bot*. 2017;119(5):803-813. doi 10.1093/aob/mcw141

Wen F., Zhu Y., Hawes M.C. Effect of pectin methylesterase gene expression on pea root development. *Plant Cell*. 1999;11(6):1129-1140. doi 10.1105/tpc.11.6.1129

Wen F., VanEtten H.D., Tsaprailis G., Hawes M.C. Extracellular proteins in pea root tip and border cell exudates. *Plant Physiol*. 2007; 143(2):773-783. doi 10.1104/pp.106.091637

Wen F., White G.J., VanEtten H.D., Xiong Z., Hawes M.C. Extracellular DNA is required for root tip resistance to fungal infection. *Plant Physiol*. 2009;151(2):820-829. doi 10.1104/pp.109.142067

Wen F., Curlango-Rivera G., Huskey D.C., Xiong Z., Hawes M.C. Visualization of extracellular DNA released during border cell separation from the root cap. *Am. J. Bot*. 2017;104(7):970-978. doi 10.3732/ajb.1700142

Wuyts N., Maung Z.T.Z., Swennen R., De Waele D. Banana rhizodeposition: characterization of root border cell production and effects on chemotaxis and motility of the parasitic nematode *Radopholus similis*. *Plant Soil*. 2006;283:217-228. doi 10.1007/s11104-006-0013-4

Zhao X., Misaghi I.J., Hawes M.C. Stimulation of border cell production in response to increased carbon dioxide levels. *Plant Physiol.* 2000;122:181-186. doi 10.1104/pp.122.1.181

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

Received April 8, 2024. Revised November 6, 2024. Accepted November 7, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):927-939

doi 10.18699/vjgb-24-100

Gene networks and metabolomic screening analysis revealed specific pathways of amino acid and acylcarnitine profile alterations in blood plasma of patients with Parkinson's disease and vascular parkinsonism

A.A. Makarova [1, 2] ✉, P.M. Melnikova[2], A.D. Rogachev [2, 3], P.S. Demenkov [1, 2, 4], T.V. Ivanisenko [1, 2, 4], E.V. Predtechenskaya[2], S.Y. Karmanov[1, 2], V.V. Koval [5], A.G. Pokrovsky [2], I.N. Lavrik[1], N.A. Kolchanov [1, 2], V.A. Ivanisenko [1, 2, 4]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] N.N. Vorozhtsov Novosibirsk Institute of Organic Chemistry of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[5] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ makarovaaa@bionet.nsc.ru

Abstract. Parkinson's disease (PD) and vascular parkinsonism (VP) are characterized by similar neurological syndromes but differ in pathogenesis, morphology, and therapeutic approaches. The molecular genetic mechanisms of these pathologies are multifactorial and involve multiple biological processes. To comprehensively analyze the pathophysiology of PD and VP, the methods of systems biology and gene network reconstruction are essential. In the current study, we performed metabolomic screening of amino acids and acylcarnitines in blood plasma of three groups of subjects: PD patients, VP patients and the control group. Comparative statistical analysis of the metabolic profiles identified significantly altered metabolites in the PD and the VP group. To identify potential mechanisms of amino acid and acylcarnitine metabolism disorders in PD and VP, regulatory gene networks were reconstructed using ANDSystem, a cognitive system. Regulatory pathways to the enzymes converting significant metabolites were found from PD-specific genetic markers, VP-specific genetic markers, and the group of genetic markers common to the two diseases. Comparative analysis of molecular genetic pathways in gene networks allowed us to identify both specific and non-specific molecular mechanisms associated with changes in the metabolomic profile in PD and VP. Regulatory pathways with potentially impaired function in these pathologies were discovered. The regulatory pathways to the enzymes ALDH2, BCAT1, AL1B1, and UD11 were found to be specific for PD, while the pathways regulating OCTC, FURIN, and S22A6 were specific for VP. The pathways regulating BCAT2, ODPB and P4HA1 were associated with genetic markers common to both diseases. The results obtained deepen the understanding of pathological processes in PD and VP and can be used for application of diagnostic systems based on the evaluation of the amino acids and acylcarnitines profile in blood plasma of patients with PD and VP.

Key words: metabolomics; amino acids; acylcarnitines; gene networks; genetic markers; Parkinson's disease; vascular parkinsonism; neurodegeneration; dry plasma stains; biomarker.

For citation: Makarova A.A., Melnikova P.M., Rogachev A.D., Demenkov P.S., Ivanisenko T.V., Predtechenskaya E.V., Karmanov S.Y., Koval V.V., Pokrovsky A.G., Lavrik I.N., Kolchanov N.A., Ivanisenko V.A. Gene networks and metabolomic screening analysis revealed specific pathways of amino acid and acylcarnitine profile alterations in blood plasma of patients with Parkinson's disease and vascular parkinsonism *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):927-939. doi 10.18699/vjgb-24-100

Funding. The authors express their gratitude to "Novartis Pharma" for financial support, which facilitated the acquisition of reagents for amino acid and acylcarnitine analysis in this study. The experimental research was conducted with the support of a state assignment of ICBFM SB RAS, No. 121031300045-2. The bioinformatics analysis was supported by the budget project FWNR-2022-0020.

Метод генных сетей и метаболомный анализ позволили выявить специфические пути изменения профиля аминокислот и ацилкарнитинов в плазме крови при болезни Паркинсона и сосудистом паркинсонизме

А.А. Макарова [1, 2] ✉, П.М. Мельникова[2], А.Д. Рогачев [2, 3], П.С. Деменков [1, 2, 4], Т.В. Иванисенко [1, 2, 4], Е.В. Предтеченская[2], С.Ю. Карманов[1, 2], В.В. Коваль [5], А.Г. Покровский [2], И.Н. Лаврик[1], Н.А. Колчанов [1, 2], В.А. Иванисенко [1, 2, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Новосибирский институт органической химии им. Н.Н. Ворожцова Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[5] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ makarovaaa@bionet.nsc.ru

Аннотация. Болезнь Паркинсона (БП) и сосудистый паркинсонизм (СП) характеризуются схожими неврологическими синдромами, но различаются патогенезом, морфологией и терапевтическими подходами. Их молекулярно-генетические механизмы многофакторны и задействуют множество биологических процессов. Для комплексного анализа патофизиологии этих заболеваний необходимо применение методов системной биологии и реконструкции генных сетей. В данном исследовании проведен метаболомный скрининг аминокислот и ацилкарнитинов в плазме крови трех групп испытуемых: пациентов с БП, пациентов с СП и контрольной группы. Сравнительный статистический анализ метаболомных профилей групп пациентов по сравнению с контролем определил значимо измененные уровни метаболитов при болезни Паркинсона и при сосудистом паркинсонизме. Для выявления потенциальных механизмов нарушения метаболизма аминокислот и ацилкарнитинов при БП и СП были реконструированы регуляторные генные сети с помощью когнитивной системы ANDSystem. Пути регуляции ферментов метаболизма значимых метаболитов были найдены для трех групп генетических маркеров: специфических для БП, специфических для СП, а также группы общих маркеров двух заболеваний. Сравнительный анализ молекулярно-генетических путей в генных сетях позволил выявить как специфические, так и общие для БП и СП молекулярные механизмы, ассоциированные с изменением метаболомного профиля. Обнаружены регуляторные пути, функция которых потенциально нарушена при этих патологиях. Специфическими для генетических маркеров БП оказались пути регуляции ферментов ALDH2, BCAT1, AL1B1 и UD11, а для генетических маркеров СП – пути регуляции ферментов OCTC, FURIN и S22A6. Регуляторные пути к ферментам BCAT2, ODPB и P4HA1 были связаны с общими для обоих заболеваний генетическими маркерами. Полученные результаты углубляют понимание патологических процессов при БП и СП и могут быть использованы для применения диагностических систем на основе оценки метаболомного профиля аминокислот и ацилкарнитинов в плазме крови пациентов с болезнью Паркинсона и сосудистым паркинсонизмом.

Ключевые слова: метаболомика; аминокислоты; ацилкарнитины; генные сети; генетический маркер; болезнь Паркинсона; сосудистый паркинсонизм; нейродегенерация; сухие пятна плазмы крови; биомаркер.

Introduction

Parkinson's disease (PD) and vascular parkinsonism (VP) are complex disorders characterized by bradykinesia, muscle rigidity, gait disturbances, and balance impairment in patients. PD is classified as a neurodegenerative disease, while VP, also known as "small vessel disease", arises in the context of cerebrovascular diseases.

In the pathogenesis of PD, disruptions in the nigrostriatal dopaminergic pathway play a crucial role, including depletion of dopamine reserves and neuronal loss in the pars compacta of substantia nigra (Alexander, 2004). Neurodegenerative processes in PD exhibit a distinct morphological staging of progression, beginning with the involvement of the olfactory bulb and the dorsal motor nucleus of the vagus nerve, eventually culminating in the critical loss of neurons in the substantia nigra pars compacta (Braak et al., 2003). This stepwise progression aligns with the gradual clinical manifestation of PD symptoms, starting from autonomic disturbances and advancing to core motor symptoms (bradykinesia, tremor, muscle rigidity) and cognitive deficits. The molecular mechanisms underlying PD are actively investigated within the scientific community. Known key factors include proteolytic stress, impaired energy metabolism in substantia nigra neurons, mitochondrial dysfunction (Levin et al., 2022), and the accumulation of alpha-synuclein (Rocha et al., 2018).

The mechanisms underlying vascular parkinsonism (VP) associated with cerebrovascular diseases (CVD) remain poorly understood. VP often arises in the context of CVD and chronic cerebral circulation disorders, leading to dysfunction of the neurogliovascular unit (Che Mohd Nassir et al., 2021). The symptoms of vascular parkinsonism develop more rapidly than in Parkinson's disease and include lower-body-predominant

bilateral parkinsonism, absence of tremor, pyramidal signs, and cognitive impairments (Vale et al., 2012). Unlike PD, dominated by proteolitic stress, mitochondrial dysfunction and the impairment energy metabolism of *substantia nigra* neurons (Levin et al., 2022), the pathogenesis of which involves the death of dopaminergic neurons and accumulation of alpha-synuclein (Rocha et al., 2018), VP is primarily driven by disturbances in microcirculation and hemodynamics. A key factor in VP development is small cerebral vessels lesion, often associated with a long history of arterial hypertension (Che Mohd Nassir et al., 2021) and diabetes mellitus (Thanvi et al., 2005). Chronic ischemia resulting from cerebrovascular disorders is accompanied by oxidative stress, inflammation, and mitochondrial dysfunction. These pathological processes lead to significant structural and functional changes in the neurogliovascular unit, including endothelial dysfunction, impaired blood-brain barrier permeability, and alterations in astrocytes and pericytes (Narasimhan et al., 2022). Ultimately, this results in white matter damage (leukoaraiosis) and the formation of multiple lacunar infarcts in strategically important areas of the basal ganglia (Zijlmans et al., 1995; Chen Y.-F. et al., 2014; Korczyn, 2015).

Among the common characteristics of these pathological processes are disruptions in the metabolism of lipids, amino acids, and energy molecules, highlighting their importance in the molecular mechanisms of these conditions. Amino acids and acylcarnitines are involved in numerous processes, including neurotransmitter biosynthesis and energy metabolism (Jones et al., 2010; Dalangin et al., 2020). Studies analyzing metabolomic profiles in Parkinson's disease (PD) are available in the literature (Wuolikainen et al., 2016; Zhao et al., 2018; Ostrakhovitch et al., 2022), but the role of amino acids and acylcarnitines requires further investigation. It is also worth noting that, to date, we have not identified any metabolomic studies focused on vascular parkinsonism.

To study complex diseases such as Parkinson's disease and parkinsonism, gene networks have been utilized. These networks allow for integration of knowledge and identification of regulatory mechanisms underlying pathologies at the molecular and genetic levels (Mercatelli et al., 2020). To date, studies on gene networks in Parkinson's disease are presented, including protein-protein interaction networks of PD markers (George et al., 2019a; Tomkins, Manzoni, 2021), gene co-expression networks (George et al., 2019b), regulatory pathways (https://www.kegg.jp/entry/hsa05012), and others. In contrast to PD, studies on the molecular and genetic mechanisms of vascular parkinsonism based on gene networks are sparsely represented in the scientific literature (Chen Y. et al., 2022).

At the ICG SB RAS, the cognitive system ANDSystem was developed for reconstructing and analyzing gene networks using artificial intelligence methods (Demenkov et al., 2012; Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022). ANDSystem has been used for interpreting metabolomic (Rogachev et al., 2021; Ivanisenko V.A. et al., 2022, 2023) and proteomic (Pastushkova et al., 2013, 2019; Binder et al., 2014; Larina et al., 2015) data. Bioinformatics studies conducted with ANDSystem have expanded the understanding of molecular and genetic processes associated with the development of various diseases and the formation of comorbid conditions (Bragina et al., 2014, 2016, 2023; Saik et al., 2016, 2018, 2019; Zolotareva et al., 2019).

The aim of this study was a comparative analysis of the molecular and genetic mechanisms of PD and VP using the gene network reconstruction methods based on metabolomic screening of amino acids and acylcarnitines.

Materials and methods

Characteristics of patients groups. The study included two groups of patients with confirmed diagnoses of Parkinson's disease (PD) and vascular parkinsonism (VP), along with a control group. Differential diagnosis was based on MRI and clinical criteria. Blood samples were collected after discontinuation of L-DOPA (L-dihydroxyphenylalanine) treatment.

The PD group consisted of 9 patients (5 women, 4 men) with a mean age of 72.2 years (age range: 64–88 years). Inclusion criteria were: clinically confirmed PD, stage IV by Hoehn and Yahr, disease duration >5 years, and onset age 55–75 years; symptoms including bradykinesia, resting tremor or muscle rigidity, and response to L-DOPA therapy. The VP group included 9 patients (7 women, 2 men) with a mean age of 74.6 years (age range: 60–89 years). Inclusion criteria were: disease duration >3 years, MRI findings of multi-lacunar status and leukoaraiosis; symptoms such as lower-body-predominant parkinsonism, bilateral onset, and postural instability. The control group consisted of 17 conditionally healthy individuals (11 women, 6 men) with a mean age of 68 years (age range: 51–82 years). Background conditions in this group included chronic arterial hypertension without transient ischemic attacks or stroke and the absence of neurological symptoms.

Collection of biological material and HPLC-MS/MS analysis. Blood samples were collected from a peripheral vein during daytime, three hours after food intake, using 6 mL plasma tubes containing lithium heparin (68 IU) (Vacutainer, BD). Plasma preparation and the analysis of amino acids and acylcarnitines (the list of the analyzed metabolites is provided in Supplementary Material 1)[1] were performed using the HPLC-MS/MS method as previously described (Kasakin et al., 2019). The analysis utilized an API 6500 QTRAP mass spectrometer (AB SCIEX, USA) coupled with an HPLC LC-20AD Prominence chromatograph (Shimadzu, Japan) equipped with an SIL-20AC autosampler (Shimadzu, Japan).

Statistical analysis of experimental data. For the statistical analysis of differences in metabolite levels across the metabolomic profiles of the studied groups, the Mann–Whitney and Kolmogorov–Smirnov tests were applied with a subsequent Benjamini–Hochberg procedure (False Discovery Rate, FDR) for multiple comparisons. Statistical calculations were performed using Python 3.11 with functions from the scipy.stats module.

Formation of lists of enzymes and genetic markers of PD and VP. For each of the metabolites, the concentrations of which were statistically significantly altered in patient groups compared to the control group, lists of biosynthesis and degradation enzymes were made. The enzymes convert-

[1] Supplementary Materials 1–9 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Makarova_Engl_28_8.xlsx

Templates of molecular genetic pathways regulating the biosynthesis and degradation enzymes of significant metabolites by genetic markers of PD, VP, or common to both diseases

Title	Template of regulatory pathway
Protein-protein interactions	Genetic markers – protein-protein interactions → Enzymes
Protein function regulation	Genetic markers – regulation of protein activity/degradation/post-translational modifications/ transport/catalytic reactions → Enzymes
Regulation of expression	Genetic markers – regulation of expression → Genes encoding enzymes – expression → Enzymes
Double regulation of expression	Genetic markers – regulation of expression → Human genes – expression → Human proteins – regulation of expression → Genes encoding enzymes – expression → Enzymes

Note. Genetic markers – proteins encoded by genetic markers (of PD, VP or common markers of both diseases); Enzymes – enzymes of conversion of significant metabolites; Enzyme genes – genes encoding enzymes of conversion of metabolomic markers.

ing significant metabolites were extracted from the KEGG (Kanehisa, 2000) and HMDB (Wishart et al., 2022) databases.

The lists of genetic markers for Parkinson's disease and vascular parkinsonism were extracted from the MalaCards database (https://www.malacards.org/, accessed on: 25.01.2024) (Rappaport et al., 2014). The genetic markers of VP included protein-coding genes annotated in disease terms "Vascular Parkinsonism" and "Vascular Dementia". The genetic markers for PD included protein-coding genes associated with the term "Parkinson's Disease".

Gene networks reconstruction. Gene networks reconstruction was performed using ANDVisio, a graphical user interface of the cognitive system ANDSystem (Ivanisenko V.A. et al., 2015). Regulatory pathways of four types were constructed according to the templates described in the Table. These templates allow to identify molecular genetic pathways including protein-protein interactions, regulation of protein activity, degradation, transport, proteolysis, and also gene expression regulation. The reconstruction of molecular genetic pathways regulating the enzymes that convert metabolites was carried out using the same templates for three sets of genetic markers (PD, VP, and common markers for both diseases).

Results

Statistical analysis of metabolomic data

The statistical analysis of metabolomic data (Supplementary Material 1), aimed at identifying differences in metabolite levels between the PD and VP groups compared to the control group, revealed that out of 44 metabolites with measured concentrations, statistically significant differences (FDR < 0.05) were observed for 18 metabolites in PD and 21 metabolites in VP (Supplementary Material 2).

Both the PD and the VP group differed from the control group in the levels of four out of 14 analyzed amino acids: alanine, proline, isoleucine, and valine. Notably, methionine levels were significantly altered in the PD group but did not distinguish the VP group from the control. Among acylcarnitines, significant differences were identified for 13 metabolites shared between PD and VP (Supplementary Material 2). Specific acylcarnitines, the levels of which significantly differed only in the VP group compared to the control, included acylcarnitines C6, C10, C10:1, and Carnitine.

Reconstruction and analysis of gene networks

To investigate the molecular genetic mechanisms potentially contributing to the altered metabolomic profiles in PD and VP, we utilized the gene network approach. Gene networks enabled the integration of knowledge about the molecular interactions of metabolites with known genetic markers of PD and VP. Genetic markers were defined as genes associated with PD and VP according to the MalaCards database (Rappaport et al., 2014). Lists of 84 genetic markers for Parkinson's disease and 60 markers for vascular parkinsonism are provided in Supplementary Material 3. The intersection of the genetic marker lists for PD and VP showed that 22 genetic markers were shared between these diseases.

To study the role of genetic markers in the regulation of enzymes involved in the conversion of significant metabolites, we applied the gene network approach using the ANDVisio software (Ivanisenko V.A. et al., 2019). This method is based on the automated reconstruction of regulatory molecular genetic pathways using templates specified in the queries to ANDVisio (see the Table).

We analyzed the regulatory pathway templates that start with proteins encoded by genetic markers specific to PD, to VP, and also markers common to both diseases. The lists of these proteins were used as input data for the "Pathway Wizard" module of the ANDVisio software. The regulatory pathways end with the enzymes of biosynthesis and degradation of metabolites identified as significant in the statistical analysis. The lists of enzymes used for the analysis are provided in Supplementary Material 4. The pathways also include intermediate participants (human proteins) that link genetic markers to enzymes. These intermediate proteins were not explicitly specified in the input data, as the software automatically identified such mediators. The regulatory pathways accounted for major types of molecular genetic interactions, including gene expression regulation, protein-protein interactions, and regulation of protein activity, degradation, transport, and catalytic reactions. Illustrations of the gene networks are provided in Supplementary Materials 5 and 6. The number of regulatory connections to each enzyme originating from PD, VP, and shared genetic markers is shown in Supplementary Materials 7–9.

Histograms illustrating the distribution of regulatory connections among participants of the reconstructed gene net-

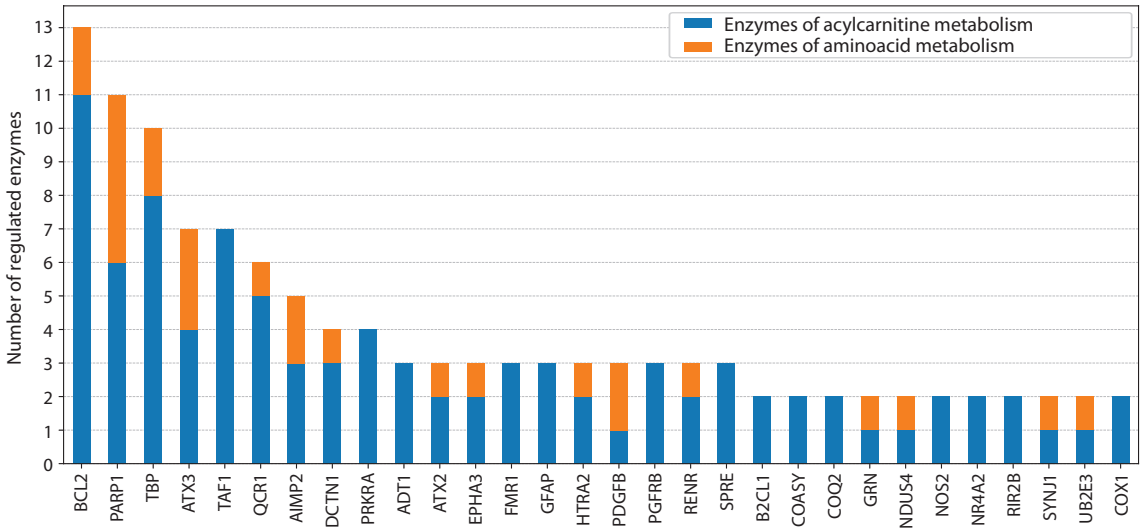


Fig. 1. Distribution of the number of regulatory pathways in the gene network from PD genetic markers to the enzymes of amino acid and acylcarnitine metabolism.

works are shown in Figures 1 and 2. In the histograms, the *X* axis represents the names of genetic markers, while the *Y* axis shows the number of regulatory pathways realized through molecular genetic interactions from the genetic markers (PD, VP, and shared markers) to the enzymes involved in reactions with significant metabolites.

The genetic markers of Parkinson's disease BCL2, TBP, and TAF1 exert greater regulatory influence on acylcarnitine metabolism enzymes, while PARP1 equally affects enzymes involved in both amino acid and acylcarnitine metabolism (Fig. 1). The genetic markers of vascular parkinsonism such as TFAM, CASP3, ALBU, and VEGFA have a stronger regulatory influence on acylcarnitine metabolism enzymes, whereas FBX7, NOTC3, and FA12 predominantly affect amino acid metabolism enzymes (Fig. 2).

Notably, the genetic markers shared between PD and VP participate equally in regulating enzymes involved in amino acid and acylcarnitine metabolism (Fig. 3). The genetic marker LRRK2, according to the gene networks, exerts a greater influence on the regulation of amino acid metabolism.

For some enzymes involved in metabolite conversion, the levels of which significantly differed in PD and VP patients compared to the control group, the regulatory pathways originating from genetic markers of PD, VP, and common genetic markers demonstrated varying quantitative proportions. Histograms depicting the number of regulatory impacts from groups of genetic markers to enzymes of amino acid and acylcarnitine metabolism were based on the gene networks (Supplementary Materials 5 and 6) and are shown in Figures 4 and 5.

The regulatory pathways from PD genetic markers are more prominent for the enzymes ALDH2, BCAT1, AL1B1, and P5CR1, while pathways to BCAT2 and P4HA1 originate more from the genetic markers shared between PD and VP (Fig. 4). Among the enzymes of acylcarnitine metabolism, fatty acid synthase (FAS) is subject to the most significant regulatory influence (Fig. 5). For enzymes FAS, ODPA (PDHA1), and ACACA (ACC1), regulatory pathways are implemented by

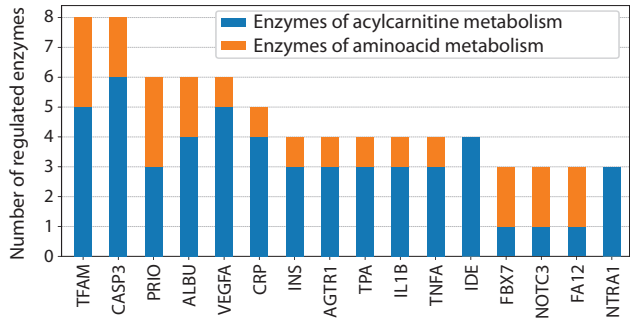


Fig. 2. Distribution of the number of regulatory pathways in the gene network from VP genetic markers to the enzymes of amino acid and acylcarnitine metabolism.

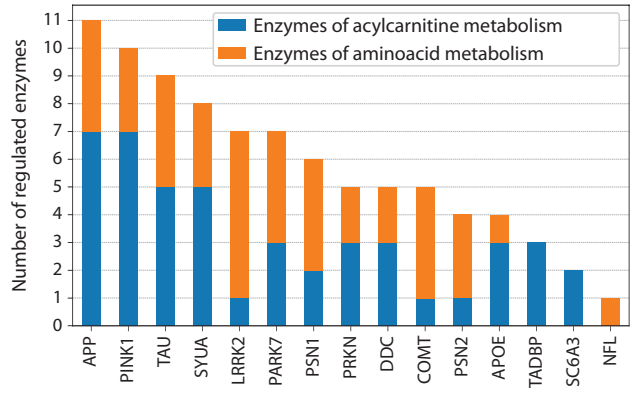


Fig. 3. Distribution of the number of regulatory pathways in the gene network from genetic markers shared between PD and VP to the enzymes of amino acid and acylcarnitine metabolism.

genetic markers specific to both PD and VP. However, in regulation of the enzyme ODPB, shared genetic markers play a more prominent role.

Thus, the metabolomic analysis identified 5 amino acids and 17 acylcarnitines with significantly altered concentrations in

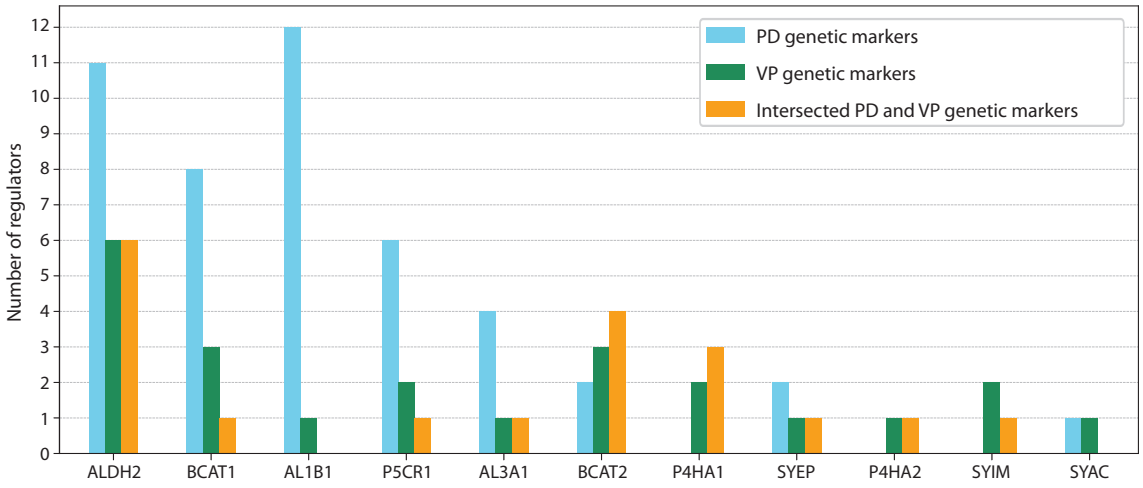


Fig. 4. Distribution of the number of regulatory pathways in the gene networks from genetic markers (PD, VP and shared between PD and VP) to the enzymes of amino acid metabolism.

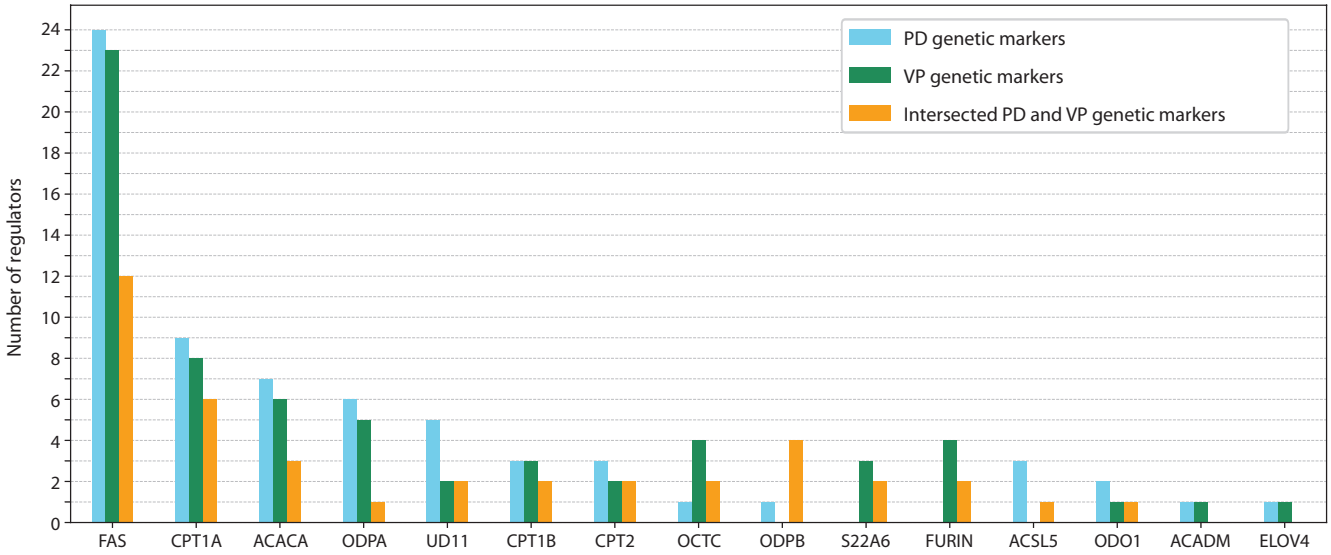


Fig. 5. Distribution of the number of regulatory pathways in the gene networks from genetic markers (PD, VP and shared between PD and VP) to the enzymes of acylcarnitine metabolism.

the PD and VP patient groups compared to the control group (Supplementary Material 2). The gene network approach enabled the reconstruction and analysis of regulatory molecular genetic pathways from PD and VP genetic markers to the enzymes involved in amino acid and acylcarnitine metabolism.

Discussion

Specific and non-specific markers of PD and VP

Our results showed that the lists of 18 and 21 significant metabolites for PD and VP, respectively, had an overlap in 17 common metabolites, which could be considered as non-specific markers for differential diagnosis. The metabolomic analysis thus provided a limited insight into distinguishing between Parkinson's disease and vascular parkinsonism. We hypothesized that while potential metabolomic markers overlap for PD and VP, the molecular mechanisms underlying their metabolic disruptions may differ between the two diseases. It is known that genetic markers play a substantial role in pathological processes. In this regard, the genetic markers may also influence the metabolism of the potential PD and VP markers (amino acids and acylcarnitines) identified in our study.

To test this hypothesis, we reconstructed the gene networks describing regulatory connections from the genetic markers of these diseases to the enzymes involved in the biosynthesis and degradation of significant metabolites. The analysis revealed that disease-specific genetic markers actively regulate enzyme functions and the expression of their encoding genes (Supplementary Materials 5 and 6). Genetic markers were grouped into three categories: specific to PD, specific to VP, and shared between the two diseases. To identify the specific

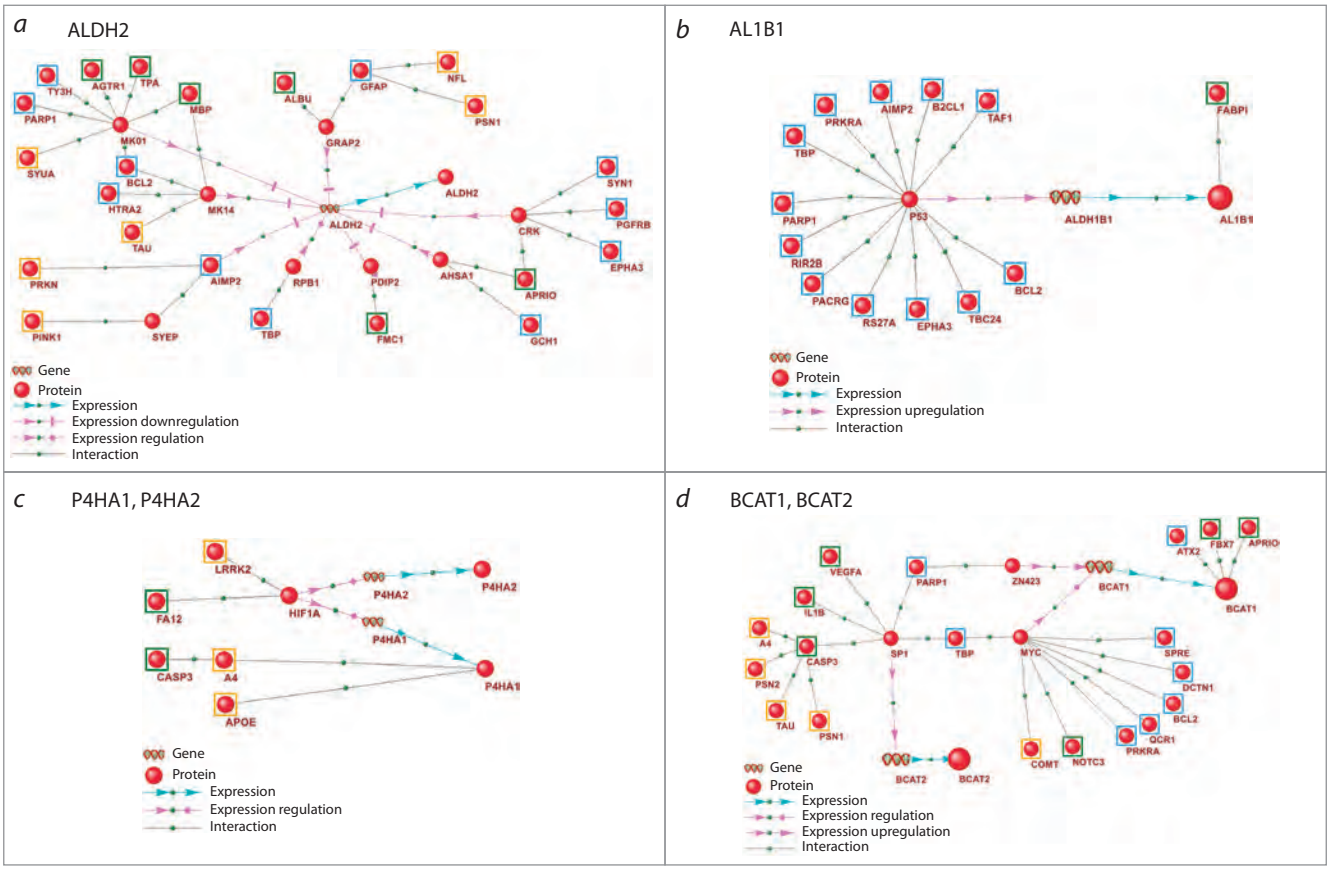


Fig. 6. Gene networks of regulation of the enzymes involved in amino acid metabolism by genetic markers of PD, VP and common markers of PD and VP.
The genetic markers are framed: PD (blue frames), VP (green frames), common markers of PD and VP (orange frames).

molecular mechanisms of disrupted metabolism regulation in PD and VP, we analyzed the regulatory pathways starting from disease-specific genetic markers. Meanwhile, the pathways involving genetic markers shared between PD and VP were hypothesized to define the mechanisms underlying common metabolomic profile disruptions. In the reconstructed gene networks (Supplementary Materials 5 and 6), we highlighted the regulatory pathways involving enzymes previously studied in the context of Parkinson's disease and vascular parkinsonism.

The gene networks of regulation of the enzymes of amino acids metabolism

ALDH2 (aldehyde dehydrogenase 2) was identified among the enzymes with the highest number of regulatory connections from PD and VP genetic markers (Fig. 6*a*). ALDH2 participates in the metabolism of proline, alanine, and fatty acids and is a key enzyme involved in metabolizing aldehydes and cytotoxic metabolites. In the brain, ALDH2 plays a crucial role in preventing "aldehyde load" – the accumulation of aldehydes that, under oxidative stress, can bind to lipids, nucleic acids, and proteins, causing neurotoxic effects (Chen C.-H. et al., 2016). Studies on the association of aldehyde dehydrogenases with PD have shown increased mitochondrial ALDH2 activity in the striatum of PD patients (Michel et al., 2014). ALDH2 may protect neurons from the toxic effects of dopamine metabolites (Chiu et al., 2015), and enhanced ALDH2 activity has been shown to restore neuronal function impaired by hypoxia (Lin et al., 2022).

The enzyme AL1B1 (mitochondrial aldehyde dehydrogenase X) had the highest number of regulatory connections from PD genetic markers in the gene network (Fig. 6*b*). AL1B1 is involved in proline, alanine, and fatty acid metabolism and plays a substantial role in acetaldehyde detoxification and neurotransmitter metabolism (Shortall et al., 2021). It was suggested that AL1B1 deficiency identified in the brain is associated with Parkinson's disease progression (Grünblatt, Riederer, 2016; Odongo et al., 2023). AL1B1 deficiency may lead to the accumulation of aldehydes such as 4-hydroxy-2-nonenal (4-HNE), which can impair mitochondrial function, induce alpha-synuclein aggregation, and trigger neuroinflammation and apoptosis (Wey et al., 2012; Grünblatt, Riederer, 2016).

According to the gene network analysis, the enzymes P4HA1 and P4HA2 were more strongly regulated by the genetic markers shared between PD and VP (Fig. 6*c*). P4HA (prolyl 4-hydroxylase alpha) enzymes catalyze the formation of 4-hydroxyproline, essential for the correct folding of procollagen chains (Song et al., 2023). Additionally, P4HA1 is known to participate in post-ischemic angiogenesis (Xu et al., 2024).

The enzymes BCAT1 and BCAT2 catalyze the reversible transamination of branched-chain amino acids (BCAA) with

alpha-ketoglutarate to form corresponding branched-chain alpha-keto acids and glutamate. Our metabolomic analysis revealed elevated levels of BCAAs, such as valine and isoleucine, in PD and VP patients. Based on our reconstructed gene networks, BCAT1 was one of the enzymes highly regulated by PD genetic markers, while regulatory connections to BCAT2 predominantly originated from shared PD and VP markers (Fig. 6*d*). Defective BCAA metabolism, including BCAT1 disruptions, is associated with key PD features, including motor dysfunction and neurodegeneration (Yao et al., 2018; Sohrabi et al., 2021). In Parkinson's disease *C. elegans* models, knockdown of *bcat1* led to depletion of tricarboxylic acid cycle metabolites and mitochondrial hyperactivity, resulting in oxidative damage to neurons (Mor et al., 2020). Furthermore, a genome-wide association meta-analysis has linked PD to genes encoding BCAA metabolism enzymes (Nalls et al., 2014). Disruptions in BCAA metabolism enzymes have also been observed in vascular dementia. Increased mRNA expression of cytosolic and mitochondrial BCAT was found in cortical samples from patients with vascular dementia, possibly protecting cells from the neurotoxic effects of excess glutamate (Ashby et al., 2017).

The gene networks of regulation of the enzymes of acylcarnitines metabolism

In both patient groups, we identified alterations in the acylcarnitine profile, which plays a critical role in the cellular energy metabolism. As acylcarnitines are the primary carriers of fatty acids to the inner mitochondrial membrane, their metabolism is closely linked to fatty acid metabolism. Fatty acid synthase (FAS) catalyzes the elongation of fatty acids starting from acetyl-CoA and malonyl-CoA. In the gene network regulating acylcarnitine metabolism enzymes, the *FASN* gene had the highest number of "expression regulation" connections from genetic markers of PD and VP (Fig. 7*a*). Notably, the genetic marker PINK1, associated with mitochondrial dysfunction in PD (Narendra et al., 2010), has been implicated in this pathway. Mutations in *PINK1* lead to its deficiency in PD (Valente et al., 2004). It has been shown that FAS repression in *PINK1*-mutant models restores mitochondrial metabolic processes and reduces palmitate levels (Vos et al., 2017). Additionally, FAS is known to play a role in central nervous system myelination and remyelination processes (Dimas et al., 2019).

The gene network analysis revealed numerous regulatory connections to CPT1 (carnitine palmitoyltransferase 1) from the genetic markers specific to both PD and VP (Fig. 7*b*). CPT1 is a transporter protein located on the outer mitochondrial membrane; CPT1 exists in three isoforms in mammalian cells: CPT1A, CPT1B, and CPT1C. CPT1A is more specific to lipogenic tissues (e. g., liver), while CPT1B predominates in tissues with high fatty acid oxidation capacity (e. g., heart and skeletal muscle), and CPT1C is predominantly expressed in neuronal tissue (Wang Muyun et al., 2021). CPT1 enzymes catalyze the transfer of acyl-CoA groups (chain lengths C12–C18) to L-carnitine, forming acylcarnitines (Schlaepfer, Joshi, 2020). Inhibition of lipid metabolism regulated by CPT1A in mouse models of Parkinson's disease has shown promising results, improving motor and sensorimotor functions (Trabjerg et al., 2023). CPT1 has also been implicated in the development of insulin resistance, a condition associated with impaired function of substantia nigra in the brain (Virmani et al., 2015). In early-stage PD patients, reduced levels of long-chain acylcarnitines (C14–C18) were identified, potentially associated with CPT1 deficiency (Saiki et al., 2017).

Regulatory connections to the enzymes ACC1 and ODPA (PDHA1) were characteristic for both PD and VP (Fig. 7*c*, *d*). ACC1 (acetyl-CoA carboxylase 1, ACACA) is the rate-limiting enzyme in *de novo* fatty acid synthesis, converting acetyl-CoA to malonyl-CoA (Wang Y. et al., 2022). In Parkinson's disease models, interaction of phosphorylated alpha-synuclein and ACC1 has been associated with low ATP levels, oxidative stress, and mitochondrial dysfunction (Grassi et al., 2018). PDHA1 (pyruvate dehydrogenase E1 alpha, ODPA) is a key component of the complex that catalyzes the decarboxylation of pyruvate to acetyl-CoA (Børglum et al., 1996). Under stress conditions, PDHA1 suppression enables astrocytes to rely on anaerobic glycolysis, increasing lactate consumption by neurons, conserving glucose, and protecting against oxidative stress (de Holanda Paranhos et al., 2024). Thus, PDHA1 acts as a mediator between cytosolic glycolysis and mitochondrial oxidative phosphorylation (Pavlú-Pereira et al., 2023). Research by Miki Y. et al. (2017) demonstrated that PDHA1 is a component of Lewy bodies in idiopathic PD and PARK14-linked parkinsonism (a familial PD form). Additionally, reduced PDHA1 protein levels have been observed in brain regions such as the striatum and substantia nigra in idiopathic PD patients.

According to the analysis of gene networks, regulatory connections to the enzymes OCTC and FURIN were found to be more specific to VP genetic markers (Fig. 7*e*, *f*). OCTC (peroxisomal carnitine octanoyltransferase), encoded by the *CROT* gene, is involved in the transport of medium- and long-chain acyl-CoA from peroxisomes, which are critical for β-oxidation of fatty acids. OCTC has been associated with calcification of arterial smooth muscle cells, as high OCTC levels were detected near calcified plaque areas (Okui et al., 2021).

Furin (PACE) is a serine convertase involved in atherogenesis. Increased furin activity is associated with cardiovascular disease progression (Wichaiyo et al., 2024), and its inhibition has been shown to slow atherosclerotic lesion progression in mice (Yakala et al., 2019). Furin also affects neuronal tissue by promoting the conversion of brain-derived neurotrophic factor (BDNF) from pro-BDNF to its mature form, potentially influencing neurodegenerative diseases (Wang Mingyue et al., 2021). Furin inhibitors may prevent neuronal damage induced by NMDA signaling (Yamada et al., 2018) and facilitate the conversion of pro-nerve growth factor (pro-NGF) to β-NGF, which influences vascular smooth muscle cells (Urban et al., 2013). Studies on the Parkinson's disease models showed that furin modulates disease progression. For instance, knockdown of *Furin1* in *D. melanogaster* reduced dopaminergic neuron loss caused by mutations in *Lrrk2* (Maksoud et al., 2019). Furthermore, furin is required for cap-dependent LRRK2 translation, impacting postsynaptic signaling (Penney et al., 2016).

Regulatory connections to the enzyme UD11 were predominantly driven by PD genetic markers (Fig. 7*g*). UDP-glu-

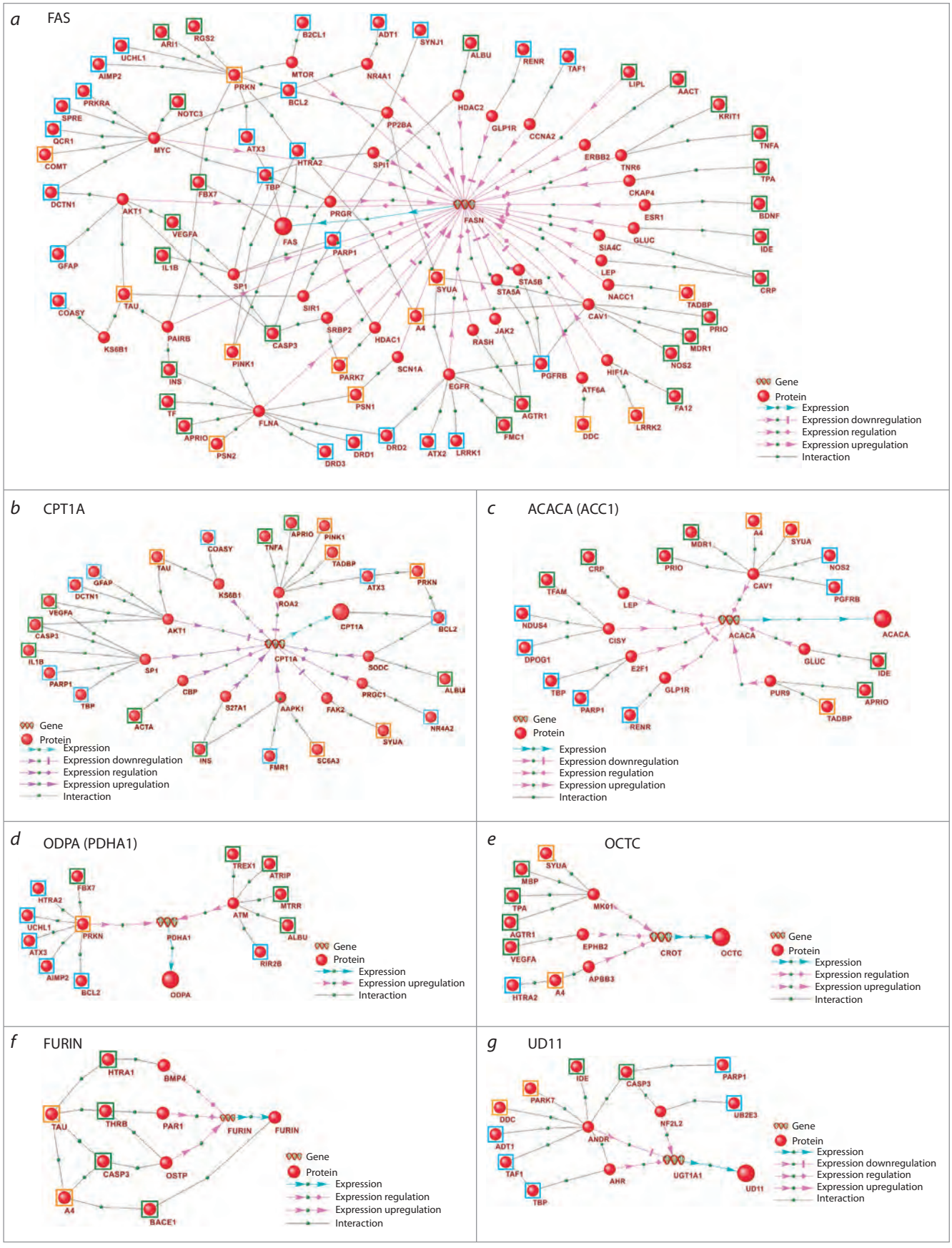


Fig. 7. Gene networks of regulation of the enzymes involved in acylcarnitines metabolism by genetic markers of PD, VP and common markers of PD and VP.

The genetic markers are framed: PD (blue frames), VP (green frames), common markers of PD and VP (orange frames).

curonosyltransferases are enzymes involved in detoxification by glucuronidating substrates, facilitating their excretion (Tukey, Strassburg, 2000). While these enzymes are understudied in the context of PD and VP, the link of UDP-glucuronosyltransferase 1A9 genotype to adverse reactions to catechol-O-methyltransferase inhibitors in PD patients was reported (Ferrari et al., 2012).

The altered metabolomic profiles of amino acids and acylcarnitines in PD and VP may result from distinct molecular genetic mechanisms. In this study, the regulatory pathways specific to PD included the enzymes ALDH2, BCAT1, AL1B1, and UD11. The pathways specific to VP were identified for OCTC, FURIN, and S22A6. For genetic markers shared by PD and VP, regulatory influences were prominent on the enzymes BCAT2, ODPB, and P4HA1. The gene networks analysis for both PD and VP revealed disruptions in lipid metabolism, valine and isoleucine pathways, and mechanisms associated with oxidative stress and mitochondrial dysfunction.

Conclusion

To identify disease-specific molecular genetic mechanisms, we reconstructed gene networks describing the regulation of enzymes involved in the metabolism of potential PD and VP markers identified by HPLC-MS/MS, including several amino acids (alanine, proline, valine, isoleucine, methionine) and 17 acylcarnitines. A comparative analysis of regulatory pathways within these networks revealed both specific and non-specific molecular mechanisms associated with the altered metabolomic profiles of these pathologies. The results obtained highlight the molecular genetic distinctions between PD and VP and may be useful for the development and application of diagnostic systems based on plasma metabolomic profiles of amino acids and acylcarnitines. Notably, this study was the first to apply the gene network analysis to the metabolomic profiles of amino acids and acylcarnitines in patients with vascular parkinsonism and Parkinson's disease, representing a significant step forward in the comparative investigation of these disorders.

References

Alexander G.E. Biology of Parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. *Dialogues Clin. Neurosci*. 2004;6(3):259-280. doi 10.31887/DCNS.2004.6.3/galexander

Ashby E.L., Kierzkowska M., Hull J., Kehoe P.G., Hutson S.M., Conway M.E. Altered expression of human mitochondrial branched chain aminotransferase in dementia with Lewy bodies and vascular dementia. *Neurochem. Res*. 2017;42(1):306-319. doi 10.1007/s11064-016-1855-7

Binder H., Wirth H., Arakelyan A., Lembcke K., Tiys E.S., Ivanisenko V.A., Kolchanov N.A., Kononikhin A., Popov I., Nikolaev E.N., Pastushkova L.K., Larina I.M. Time-course human urine proteomics in space-flight simulation experiments. *BMC Genomics*. 2014;15(S12):S2. doi 10.1186/1471-2164-15-S12-S2

Børglum A.D., Flint T., Hansen L.L., Kruse T.A. Refined localization of the pyruvate dehydrogenase E1α gene (PDHA1) by linkage analysis. *Hum. Genet*. 1996;99(1):80-82. doi 10.1007/s004390050315

Braak H., Tredici K.D., Rüb U., De Vos R.A.I., Jansen Steur E.N.H., Braak E. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging*. 2003;24(2):197-211. doi 10.1016/S0197-4580(02)00065-9

Bragina E.Yu., 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.Yu., 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

Bragina E.Yu., Gomboeva D.E., Saik O.V., Ivanisenko V.A., Freidin M.B., Nazarenko M.S., Puzyrev V.P. Apoptosis genes as a key to identification of inverse comorbidity of Huntington's disease and cancer. *Int. J. Mol. Sci*. 2023;24(11):9385. doi 10.3390/ijms24119385

Che Mohd Nassir C.M.N., Damodaran T., Yusof S.R., Norazit A., Chilla G., Huen I., Kn B.P., Mohamed Ibrahim N., Mustapha M. Aberrant neurogliovascular unit dynamics in cerebral small vessel disease: a rheological clue to vascular Parkinsonism. *Pharmaceutics*. 2021;13(8):1207. doi 10.3390/pharmaceutics13081207

Chen C.-H., Joshi A.U., Mochly-Rosen D. The role of mitochondrial aldehyde dehydrogenase 2 (ALDH2) in neuropathology and neurodegeneration. *Acta Neurol. Taiwan*. 2016;25(4)(4):111-123

Chen Y., Liu Q., Liu J., Wei P., Li B., Wang N., Liu Z., Wang Z. Revealing the modular similarities and differences among Alzheimer's disease, vascular dementia, and Parkinson's disease in genomic networks. *Neuromol. Med*. 2022;24(2):125-138. doi 10.1007/s12017-021-08670-2

Chen Y.-F., Tseng Y.-L., Lan M.-Y., Lai S.-L., Su C.-S., Liu J.-S., Chang Y.-Y. The relationship of leukoaraiosis and the clinical severity of vascular Parkinsonism. *J. Neurol. Sci*. 2014;346(1-2):255-259. doi 10.1016/j.jns.2014.09.002

Chiu C.-C., Yeh T.-H., Lai S.-C., Wu-Chou Y.-H., Chen C.-H., Mochly-Rosen D., Huang Y.-C., Chen Y.-J., Chen C.-L., Chang Y.-M., Wang H.-L., Lu C.-S. Neuroprotective effects of aldehyde dehydrogenase 2 activation in rotenone-induced cellular and animal models of parkinsonism. *Exp. Neurol*. 2015;263:244-253. doi 10.1016/j.expneurol.2014.09.016

Dalangin R., Kim A., Campbell R.E. The role of amino acids in neurotransmission and fluorescent tools for their detection. *Int. J. Mol. Sci*. 2020;21(17):6197. doi 10.3390/ijms21176197

De Holanda Paranhos L., Magalhães R.S.S., De Araújo Brasil A., Neto J.R.M., Ribeiro G.D., Queiroz D.D., Dos Santos V.M., Eleutherio E.C.A. The familial amyotrophic lateral sclerosis-associated A4V SOD1 mutant is not able to regulate aerobic glycolysis. *Biochim. Biophys. Acta Gen. Subjt*. 2024;1868(8):130634. doi 10.1016/j.bbagen.2024.130634

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

Dimas P., Montani L., Pereira J.A., Moreno D., Trötzmüller M., Gerber J., Semenkovich C.F., Köfeler H.C., Suter U. CNS myelination and remyelination depend on fatty acid synthesis by oligodendrocytes. *eLife*. 2019;8:e44702. doi 10.7554/eLife.44702

Ferrari M., Martignoni E., Blandini F., Riboldazzi G., Bono G., Marino F., Cosentino M. Association of UDP-glucuronosyltransferase 1A9 polymorphisms with adverse reactions to catechol-O-methyltransferase inhibitors in Parkinson's disease patients. *Eur. J. Clin. Pharmacol*. 2012;68(11):1493-1499. doi 10.1007/s00228-012-1281-y

George G., Singh S., Lokappa S.B., Varkey J. Gene co-expression network analysis for identifying genetic markers in Parkinson's disease – a three-way comparative approach. *Genomics*. 2019a;111(4):819-830. doi 10.1016/j.ygeno.2018.05.005

George G., Valiya Parambath S., Lokappa S.B., Varkey J. Construction of Parkinson's disease marker-based weighted protein-protein

interaction network for prioritization of co-expressed genes. *Gene*. 2019b;697:67-77. doi 10.1016/j.gene.2019.02.026

Grassi D., Howard S., Zhou M., Diaz-Perez N., Urban N.T., Guerrero-Given D., Kamasawa N., Volpicelli-Daley L.A., LoGrasso P., Lasmézas C.I. Identification of a highly neurotoxic α-synuclein species inducing mitochondrial damage and mitophagy in Parkinson's disease. *Proc. Natl. Acad. Sci. USA*. 2018;115(11):E2634-E2643. doi 10.1073/pnas.1713849115

Grünblatt E., Riederer P. Aldehyde dehydrogenase (ALDH) in Alzheimer's and Parkinson's disease. *J. Neural. Transm*. 2016;123(2): 83-90. doi 10.1007/s00702-014-1320-1

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(S11):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 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(S1):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 Yu.I., Voevoda S.M., Karpenko T.N., Velichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolomics and gene regulatory networks analysis reveal the role of non-structural 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., 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. *Vavilov J. Genet. Breed*. 2023;27(7): 768-775. doi 10.18699/VJGB-23-89

Jones L.L., McDonald D.A., Borum P.R. Acylcarnitines: role in brain. *Prog. Lipid Res*. 2010;49(1):61-75. doi 10.1016/j.plipres.2009.08.004

Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi 10.1093/nar/28.1.27

Kasakin M.F., Rogachev A.D., Predtechenskaya E.V., Zaigraev V.J., Koval V.V., Pokrovsky A.G. Targeted metabolomics approach for identification of relapsing-remitting multiple sclerosis markers and evaluation of diagnostic models. *Med. Chem. Commun*. 2019;10(10): 1803-1809. doi 10.1039/c9md00253g

Korczyn A.D. Vascular Parkinsonism – characteristics, pathogenesis and treatment. *Nat. Rev. Neurol*. 2015;11(6):319-326. doi 10.1038/nrneurol.2015.61

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(01):1540001. doi 10.1142/S0219720015400016

Levin O.S., Bogolepova A.N., Lobzin V.Yu. General mechanisms of the pathogenesis of neurodenerative and cerebrovascular diseases and the possibilities of their correction. *Zhurnal Nevrologii i Psikhiatrii Imeni S.S. Korsakova = S.S. Korsakov Journal of Neurology and Psychiatry*. 2022;122(5):11-16. doi 10.17116/jnevro202212205111 (in Russian)

Lin L., Tao J.-P., Li M., Peng J., Zhou C., Ouyang J., Si Y.-Y. Mechanism of ALDH2 improves the neuronal damage caused by hypoxia/reoxygenation. *Eur. Rev. Med. Pharmacol. Sci.* 2022;26(8):2712-2720. doi 10.26355/eurrev_202204_28601

Maksoud E., Liao E.H., Haghighi A.P. A neuron-glial trans-signaling cascade mediates LRRK2-induced neurodegeneration. *Cell Rep*. 2019;26(7):1774-1786.e4. doi 10.1016/j.celrep.2019.01.077

Mercatelli D., Scalambra L., Triboli L., Ray F., Giorgi F.M. Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta Gene Regul. Mech.* 2020;1863(6):194430. doi 10.1016/j.bbagrm.2019.194430

Michel T.M., Käsbauer L., Gsell W., Jecel J., Sheldrick A.J., Cortese M., Nickl-Jockschat T., Grünblatt E., Riederer P. Aldehyde dehydrogenase 2 in sporadic Parkinson's disease. *Parkinsonism Relat. Disord*. 2014;20:S68-S72. doi 10.1016/S1353-8020(13)70018-X

Miki Y., Tanji K., Mori F., Kakita A., Takahashi H., Wakabayashi K. Alteration of mitochondrial protein PDHA1 in Lewy body disease and PARK14. *Biochem. Biophys. Res. Commun*. 2017;489(4):439-444. doi 10.1016/j.bbrc.2017.05.162

Mor D.E., Sohrabi S., Kaletsky R., Keyes W., Tartici A., Kalia V., Miller G.W., Murphy C.T. Metformin rescues Parkinson's disease phenotypes caused by hyperactive mitochondria. *Proc. Natl. Acad. Sci. USA*. 2020;117(42):26438-26447. doi 10.1073/pnas.2009838117

Nalls M.A., Pankratz N., Lill C.M., Do C.B., Hernandez D.G., Saad M., DeStefano A.L., Kara E., Bras J., Sharma M., … Brice A., Scott W.K., Gasser T., Bertram L., Eriksson N., Foroud T., Singleton A.B. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet*. 2014; 46(9):989-993. doi 10.1038/ng.3043

Narasimhan M., Schwartz R., Halliday G. Parkinsonism and cerebrovascular disease. *J. Neurol. Sci*. 2022;433:120011. doi 10.1016/j.jns.2021.120011

Narendra D.P., Jin S.M., Tanaka A., Suen D.-F., Gautier C.A., Shen J., Cookson M.R., Youle R.J. PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. *PLoS Biol*. 2010;8(1):e1000298. doi 10.1371/journal.pbio.1000298

Odongo R., Bellur O., Abdik E., Çakır T. Brain-wide transcriptome-based metabolic alterations in Parkinson's disease: human inter-region and human-experimental model correlations. *Mol. Omics*. 2023; 19(7):522-537. doi 10.1039/D2MO00343K

Okui T., Iwashita M., Rogers M.A., Halu A., Atkins S.K., Kuraoka S., Abdelhamid I., Higashi H., Ramsaroop A., Aikawa M., Singh S.A., Aikawa E. CROT (Carnitine O-Octanoyltransferase) is a novel contributing factor in vascular calcification via promoting fatty acid metabolism and mitochondrial dysfunction. *Arterioscler. Thromb. Vasc. Biol*. 2021;41(2):755-768. doi 10.1161/ATVBAHA.120.315007

Ostrakhovitch E.A., Song E.-S., Macedo J.K.A., Gentry M.S., Quintero J.E., Van Horne C., Yamasaki T.R. Analysis of circulating metabolites to differentiate Parkinson's disease and essential tremor. *Neurosci. Lett*. 2022;769:136428. doi10.1016/j.neulet.2021.136428

Pastushkova L.Kh., Kireev K.S., Kononikhin A.S., Tiys E.S., Popov I.A., Starodubtseva N.L., Dobrokhotov I.V., Ivanisenko V.A., Larina I.M., Kolchanov N.A., Nikolaev E.N. Detection of renal tissue and urinary tract proteins in the human urine after space flight. *PLoS One*. 2013;8(8):e71652. doi 10.1371/journal.pone.0071652

Pastushkova L.Kh., Kashirina D.N., Brzhozovskiy A.G., Kononikhin A.S., Tiys E.S., Ivanisenko V.A., Koloteva M.I., Nikolaev E.N., Larina I.M. Evaluation of cardiovascular system state by urine proteome after manned space flight. *Acta Astronaut*. 2019;160:594-600. doi 10.1016/j.actaastro.2019.02.015

Pavlú-Pereira H., Florindo C., Carvalho F., Tavares De Almeida I., Vicente J., Morais V., Rivera I. Evaluation of mitochondrial function on

pyruvate dehydrogenase complex deficient patient-derived cell lines. *Endocr. Metab. Immune Disord. Drug Targets*. 2024;24(16):20. doi 10.2174/0118715303280072231004082458

Penney J., Tsurudome K., Liao E.H., Kauwe G., Gray L., Yanagiya A., Calderon M.R., Sonenberg N., Haghighi A.P. LRRK2 regulates retrograde synaptic compensation at the *Drosophila* neuromuscular junction. *Nat. Commun*. 2016;7(1):12188. doi 10.1038/ncomms12188

Rappaport N., Twik M., Nativ N., Stelzer G., Bahir I., Stein T.I., Safran M., Lancet D. MalaCards: a comprehensive automatically-mined database of human diseases. *Curr. Protoc. Bioinformatics*. 2014;47(1):1.24.1-19. doi 10.1002/0471250953.bi0124s47

Rocha E.M., De Miranda B., Sanders L.H. Alpha-synuclein: pathology, mitochondrial dysfunction and neuroinflammation in Parkinson's disease. *Neurobiol. Dis*. 2018;109:249-257. doi 10.1016/j.nbd.2017.04.004

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. doi 10.3390/metabo11030133

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., Demenkov P.S., Ivanisenko T.V., Bragina E.Yu., 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

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(S2):47. doi 10.1186/s12920-019-0492-9

Saiki S., Hatano T., Fujimaki M., Ishikawa K.-I., Mori A., Oji Y., Okuzumi A., Fukuhara T., Koinuma T., Imamichi Y., Nagumo M., Furuya N., Nojiri S., Amo T., Yamashiro K., Hattori N. Decreased long-chain acylcarnitines from insufficient β-oxidation as potential early diagnostic markers for Parkinson's disease. *Sci. Rep*. 2017;7(1):7328. doi 10.1038/s41598-017-06767-y

Schlaepfer I.R., Joshi M. CPT1A-mediated fat oxidation, mechanisms, and therapeutic potential. *Endocrinology*. 2020;161(2):bqz046. doi 10.1210/endocr/bqz046

Shortall K., Djeghader A., Magner E., Soulimane T. Insights into aldehyde dehydrogenase enzymes: a structural perspective. *Front. Mol. Biosci*. 2021;8:659550. doi 10.3389/fmolb.2021.659550

Sohrabi S., Mor D.E., Kaletsky R., Keyes W., Murphy C.T. High-throughput behavioral screen in *C. elegans* reveals Parkinson's disease drug candidates. *Commun. Biol*. 2021;4(1):203. doi 10.1038/s42003-021-01731-z

Song M., Schnettler E., Venkatachalam A., Wang Y., Feldman L., Argenta P., Rodriguez-Rodriguez L., Ramakrishnan S. Increased expression of collagen prolyl hydroxylases in ovarian cancer is associated with cancer growth and metastasis. *Am. J. Cancer Res*. 2023;13(12):6051-6062

Thanvi B., Lo N., Robinson T. Vascular parkinsonism – an important cause of parkinsonism in older people. *Age Ageing*. 2005;34(2):114-119. doi 10.1093/ageing/afi025

Tomkins J.E., Manzoni C. Advances in protein-protein interaction network analysis for Parkinson's disease. *Neurobiol. Dis*. 2021;155:105395. doi 10.1016/j.nbd.2021.105395

Trabjerg M.S., Andersen D.C., Huntjens P., Mørk K., Warming N., Kullab U.B., Skjønnemand M.-L.N., Oklinski M.K., Oklinski K.E., Bolther L., Kroese L.J., Pritchard C.E.J., Huijbers I.J., Corthals A., Søndergaard M.T., Kjeldal H.B., Pedersen C.F.M., Nieland J.D.V. Inhibition of carnitine palmitoyl-transferase 1 is a potential target in a mouse model of Parkinson's disease. *NPJ Parkinsons Dis*. 2023;9(1):6. doi 10.1038/s41531-023-00450-y

Tukey R.H., Strassburg C.P. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu. Rev. Pharmacol. Toxicol*. 2000;40(1):581-616. doi 10.1146/annurev.pharmtox.40.1.581

Urban D., Lorenz J., Meyborg H., Ghosh S., Kintscher U., Kaufmann J., Fleck E., Kappert K., Stawowy P. Proprotein convertase furin enhances survival and migration of vascular smooth muscle cells via processing of pro-nerve growth factor. *J. Biochem*. 2013;153(2):197-207. doi 10.1093/jb/mvs137

Vale T.C., Barbosa M.T., Caramelli P., Cardoso F. Vascular Parkinsonism and cognitive impairment: literature review, Brazilian studies and case vignettes. *Dement. Neuropsychol*. 2012;6(3):137-144. doi 10.1590/S1980-57642012DN06030005

Valente E.M., Abou-Sleiman P.M., Caputo V., Muqit M.M.K., Harvey K., Gispert S., Ali Z., Del Turco D., Bentivoglio A.R., Healy D.G., Albanese A., Nussbaum R., González-Maldonado R., Deller T., Salvi S., Cortelli P., Gilks W.P., Latchman D.S., Harvey R.J., Dallapiccola B., Auburger G., Wood N.W. Hereditary early-onset Parkinson's disease caused by mutations in *PINK1*. *Science*. 2004;304(5674):1158-1160. doi 10.1126/science.1096284

Virmani A., Pinto L., Bauermann O., Zerelli S., Diedenhofen A., Binienda Z.K., Ali S.F., Van Der Leij F.R. The carnitine palmitoyl transferase (CPT) system and possible relevance for neuropsychiatric and neurological conditions. *Mol. Neurobiol*. 2015;52(2):826-836. doi 10.1007/s12035-015-9238-7

Vos M., Geens A., Böhm C., Deaulmerie L., Swerts J., Rossi M., Craessaerts K., Leites E.P., Seibler P., Rakovic A., Lohnau T., De Strooper B., Fendt S.-M., Morais V.A., Klein C., Verstreken P. Cardiolipin promotes electron transport between ubiquinone and complex I to rescue *PINK1* deficiency. *J. Cell Biol*. 2017;216(3):695-708. doi 10.1083/jcb.201511044

Wang Mingyue, Xie Y., Qin D. Proteolytic cleavage of proBDNF to mBDNF in neuropsychiatric and neurodegenerative diseases. *Brain Res. Bull*. 2021;166:172-184. doi 10.1016/j.brainresbull.2020.11.005

Wang Muyun, Wang K., Liao X., Hu H., Chen L., Meng L., Gao W., Li Q. Carnitine palmitoyltransferase system: a new target for anti-inflammatory and anticancer therapy? *Front. Pharmacol*. 2021;12:760581. doi 10.3389/fphar.2021.760581

Wang Yu, Yu W., Li S., Guo D., He J., Wang Yugang. Acetyl-CoA carboxylases and diseases. *Front. Oncol*. 2022;12:836058. doi 10.3389/fonc.2022.836058

Wey M.C.-Y., Fernandez E., Martinez P.A., Sullivan P., Goldstein D.S., Strong R. Neurodegeneration and motor dysfunction in mice lacking cytosolic and mitochondrial aldehyde dehydrogenases: implications for Parkinson's disease. *PLoS One*. 2012;7(2):e31522. doi 10.1371/journal.pone.0031522

Wichaiyo S., Koonyosying P., Morales N.P. Functional roles of furin in cardio-cerebrovascular diseases. *ACS Pharmacol. Transl. Sci*. 2024;7(3):570-585. doi 10.1021/acsptsci.3c00325

Wishart D.S., Guo A., Oler E., Wang F., Anjum A., Peters H., Dizon R., Sayeeda Z., Tian S., Lee B.L., Berjanskii M., Mah R., Yamamoto M., Jovel J., Torres-Calzada C., Hiebert-Giesbrecht M., Lui V.W., Varshavi Dorna, Varshavi Dorsa, Allen D., Arndt D., Khetarpal N., Sivakumaran A., Harford K., Sanford S., Yee K., Cao X., Budinski Z., Liigand J., Zhang L., Zheng J., Mandal R., Karu N., Dambrova M., Schiöth H.B., Greiner R., Gautam V. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res*. 2022;50(D1):D622-D631. doi 10.1093/nar/gkab1062

Wuolikainen A., Jonsson P., Ahnlund M., Antti H., Marklund S.L., Moritz T., Forsgren L., Andersen P.M., Trupp M. Multi-platform mass spectrometry analysis of the CSF and plasma metabolomes of rigorously matched amyotrophic lateral sclerosis, Parkinson's di-

sease and control subjects. *Mol. BioSyst.* 2016;12(4):1287-1298. doi 10.1039/C5MB00711A

Xu Y., Xia D., Huang K., Liang M. Hypoxia-induced P4HA1 overexpression promotes post-ischemic angiogenesis by enhancing endothelial glycolysis through downregulating FBP1. *J. Transl. Med.* 2024;22(1):74. doi 10.1186/s12967-024-04872-x

Yakala G.K., Cabrera-Fuentes H.A., Crespo-Avilan G.E., Rattanasopa C., Burlacu A., George B.L., Anand K., Mayan D.C., Corlianò M., Hernández-Reséndiz S., Wu Z., Schwerk A.M.K., Tan A.L.J., Trigueros-Motos L., Chèvre R., Chua T., Kleemann R., Liehn E.A., Hausenloy D.J., Ghosh S., Singaraja R.R. FURIN inhibition reduces vascular remodeling and atherosclerotic lesion progression in mice. *Arterioscler. Thromb. Vasc. Biol.* 2019;39(3):387-401. doi 10.1161/ATVBAHA.118.311903

Yamada M., Hayashi H., Yuuki M., Matsushima N., Yuan B., Takagi N. Furin inhibitor protects against neuronal cell death induced by activated NMDA receptors. *Sci. Rep.* 2018;8(1):5212. doi 10.1038/s41598-018-23567-0

Yao V., Kaletsky R., Keyes W., Mor D.E., Wong A.K., Sohrabi S., Murphy C.T., Troyanskaya O.G. An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* 2018;36(11):1091-1099. doi 10.1038/nbt.4246

Zhao H., Wang C., Zhao N., Li W., Yang Z., Liu X., Le W., Zhang X. Potential biomarkers of Parkinson's disease revealed by plasma metabolic profiling. *J. Chromatogr. B.* 2018;1081-1082:101-108. doi 10.1016/j.jchromb.2018.01.025

Zijlmans J.C.M., Thijssen H.O.M., Vogels O.J.M., Kremer H.P.H.M.P., Poels P.J.E., Schoonderwaldt H.C., Merx J.L., van't Hof M.A., Thien Th., Horstink M.W.I.M. MRI in patients with suspected vascular parkinsonism. *Neurology*. 1995;45(12):2183-2188. doi 10.1212/WNL.45.12.2183

Zolotareva O., Saik O.V., Königs C., Bragina E.Yu., 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 September 16, 2024. Revised November 7, 2024. Accepted November 8, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):940-949

doi 10.18699/vjgb-24-101

Ontologies in modelling and analysing of big genetic data

N.L. Podkolodnyy [1, 2, 3, 4] ✉, O.A. Podkolodnaya [1], V.A. Ivanisenko [1, 4], M.A. Marchenko [2, 3]

[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] Novosibirsk State University, Novosibirsk, Russia
[4] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

✉ pnl@bionet.nsc.ru

Abstract. To systematize and effectively use the huge volume of experimental data accumulated in the field of bioinformatics and biomedicine, new approaches based on ontologies are needed, including automated methods for semantic integration of heterogeneous experimental data, methods for creating large knowledge bases and self-interpreting methods for analyzing large heterogeneous data based on deep learning. The article briefly presents the features of the subject area (bioinformatics, systems biology, biomedicine), formal definitions of the concept of ontology and knowledge graphs, as well as examples of using ontologies for semantic integration of heterogeneous data and creating large knowledge bases, as well as interpreting the results of deep learning on big data. As an example of a successful project, the Gene Ontology knowledge base is described, which not only includes terminological knowledge and gene ontology annotations (GOA), but also causal influence models (GO-CAM). This makes it useful not only for genomic biology, but also for systems biology, as well as for interpreting large-scale experimental data. An approach to building large ontologies using design patterns is discussed, using the ontology of biological attributes (OBA) as an example. Here, most of the classification is automatically computed based on previously created reference ontologies using automated inference, except for a small number of high-level concepts. One of the main problems of deep learning is the lack of interpretability, since neural networks often function as "black boxes" unable to explain their decisions. This paper describes approaches to creating methods for interpreting deep learning models and presents two examples of self-explanatory ontology-based deep learning models: (1) Deep GONet, which integrates Gene Ontology into a hierarchical neural network architecture, where each neuron represents a biological function. Experiments on cancer diagnostic datasets show that Deep GONet is easily interpretable and has high performance in distinguishing cancerous and non-cancerous samples. (2) ONN4MST, which uses biome ontologies to trace microbial sources of samples whose niches were previously poorly studied or unknown, detecting microbial contaminants. ONN4MST can distinguish samples from ontologically similar biomes, thus offering a quantitative way to characterize the evolution of the human gut microbial community. Both examples demonstrate high performance and interpretability, making them valuable tools for analyzing and interpreting big data in biology.

Key words: ontologies; big data analysis; bioinformatics; systems biology; deep learning; interpretability.

For citation: Podkolodnyy N.L., Podkolodnaya O.A., Ivanisenko V.A., Marchenko M.A. Ontologies in modelling and analysing of big genetic data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8): 940-949. doi 10.18699/vjgb-24-101

Funding. The work was supported by budget projects FWNR-2022-0020 and FWNM-2022-0005.

Онтологии в моделировании и анализе больших генетических данных

Н.Л. Подколодный [1, 2, 3, 4] ✉, О.А. Подколодная [1], В.А. Иванисенко [1, 4], М.А. Марченко [2, 3]

[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] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

✉ pnl@bionet.nsc.ru

Аннотация. Для систематизации и эффективного использования огромного объема экспериментальных данных, накопленных в области биоинформатики и биомедицины, необходимы новые подходы, основанные на онтологиях, включая автоматизированные методы семантической интеграции гетерогенных экспериментальных данных, методы создания больших баз знаний и самоинтерпретируемые методы анализа больших разнородных данных на основе глубокого обучения. В статье кратко представлены особенности предметной области (биоинформатика, системная биология, биомедицина), формальные определения понятия онтологии и графов знаний,

приведены примеры применения онтологий для семантической интеграции гетерогенных данных и создания больших баз знаний, а также интерпретации результатов глубокого обучения на больших данных. В качестве примера успешного проекта описана база знаний Gene Ontology, которая помимо терминологических знаний и аннотаций генов (GOA) включает модели причинных влияний (GO-CAM). Это делает ее полезной не только для геномной биологии, но и для системной биологии, а также для интерпретации крупномасштабных экспериментальных данных. Обсуждается подход к созданию больших онтологий с использованием шаблонов проектирования на примере онтологии биологических атрибутов (OBA). Здесь большая часть классификации автоматически вычисляется на основе ранее созданных эталонных онтологий с помощью автоматизированного логического вывода, за исключением небольшого числа высокоуровневых понятий. Одной из основных проблем глубокого обучения является отсутствие интерпретируемости, поскольку нейронные сети часто функционируют как «черные ящики», не способные объяснить свои решения. В нашей статье описаны подходы к созданию методов интерпретации моделей глубокого обучения и представлены два примера самообъясняемых моделей глубокого обучения на основе онтологий. Модель Deep GONet, которая интегрирует Gene Ontology в иерархическую архитектуру нейронной сети, где каждый нейрон представляет биологическую функцию. Эксперименты с наборами данных диагностики рака показывают, что Deep GONet легко интерпретируется и обладает высокой производительностью для различения раковых и нераковых образцов. Модель ONN4MST, использующая онтологии биома для отслеживания микробных источников образцов, ниши которых ранее были мало изучены или неизвестны, и обнаружения микробных загрязнителей. ONN4MST может отличать образцы от онтологически близких биомов и, таким образом, предлагает количественный способ охарактеризовать развитие микробного сообщества кишечника человека. Оба примера демонстрируют высокую производительность и интерпретируемость, что делает их ценными инструментами для анализа и интерпретации больших данных в биологии.
Ключевые слова: онтологии; биоинформатика; системная биология; анализ больших данных; глубокое обучение; интерпретируемость.

Introduction

The term “Big Data” refers to voluminous datasets that are characterized by significant size, diversity, and complexity, making them difficult to process and analyze using traditional methods. Moreover, such data are often incomplete and uncertain, which complicates the task of controlling their quality and accuracy (Qaiser, Ghulam, 2023).

The emergence of qualitatively new research opportunities based on high-throughput experimental technologies such as massively parallel DNA sequencing, multilocus genotyping, multiparametric gene expression profiling using DNA chips, ChIP-on-chip technology, as well as proteomic and metabolomic technologies, has led to the accumulation of unprecedentedly large volumes of experimental data and knowledge (Stephens et al., 2015). The heterogeneity of molecular biological information and its complexity complicate the analysis, systematization and application of these data to solve specific problems in bioinformatics, biotechnology, pharmacology and personalized medicine.

New approaches to big data processing are required to master, systematize and effectively use huge amounts of data. In particular, this includes automated methods for the semantic integration of heterogeneous data, one of the key stages of which is the harmonization of domain concepts, as well as methods for describing and using them. A coordinated description of a specific domain is called an ontology.

Ontologies allow concepts to be represented in a format suitable for machine processing and act as an intermediary between the user and the information system, as well as between members of the scientific community when exchanging data. Thus, ontologies are becoming an important tool in bioinformatics and systems biology, facilitating the semantic integration of experimental data and knowledge in order to create a “unified picture of the world”. In addition, they help solve problems arising in the analysis of big data, overcoming heterogeneity and deficiencies in data quality, and improving the interpretation of deep learning results. Ontologies increase the scalability and efficiency of processing large amounts of information, which makes them indispensable in modern scientific research.

Earlier, the review (Podkolodnyy et al., 2016) presented examples of ontologies describing biological systems at various levels of organization of living systems. This article will present examples of the application of ontologies for the integration of heterogeneous data and the creation of large knowledge bases, as well as the interpretation of data analysis results.

Formal representation of ontologies

In computer science, the term “ontology” refers to a conceptual model that represents objects, their properties, and the relationships between them (Chandrasekaran et al., 1999). An ontology includes a set of concepts (terms) of a particular subject area and their definitions, as well as all the information associated with these concepts, such as properties, relations, constraints, axioms, and assertions. This information is necessary for describing and solving problems in the chosen subject area (Podkolodnyy et al., 2016).

Thus, a formal model of an ontology is represented as an ordered triple of finite sets O = <T, R, F>, where T is a finite and non-empty set of classes and concepts (concepts, terms) of the subject area considered in a certain context (in our case: bioinformatics, systems biology, biotechnology, and biomedicine); R is a finite set of relations between concepts of a given subject area; F is a finite set of interpretation functions defined by concepts and/or relations of the onto-

logy O, as well as axioms used to model statements that are always true. This constrains the interpretation and ensures the correct use of concepts.

One of the most effective approaches to describing and using domain knowledge is descriptive logics (DL), which define a formal language for describing concepts (concepts, classes, categories, or entities) and relationships between them (called roles), as well as for formulating statements of facts and queries about them, including satisfiability and inclusion checking. In addition, DL includes constructors (operations) for creating conceptual expressions, such as conjunction, disjunction, and relation definition.

From the point of view of descriptive logic, two main categories of knowledge can be distinguished in the domain knowledge base. The first category includes general knowledge about a set of classes of concepts, their properties, and relationships between them, which is referred to as terminological knowledge, or T-Box. The second category covers knowledge about individual objects (instances of classes), their properties, and relationships with other objects, known as assertional knowledge, or A-Box. Thus, the T-Box describes the subject area at the level of abstract concepts, while the A-Box focuses on specific data, representing a database. It is important to note that both components of the knowledge base are interconnected and complement each other.

Knowledge graphs (KGs) are often used to systematically model complex systems, organisms, and diseases, as well as to represent knowledge in bioinformatics and systems biology. According to the definition presented in (Callahan et al. 2024), a knowledge graph is a data structure that represents multiple heterogeneous entities and different types of relationships between them. This structure serves as an abstract framework capable of generating new knowledge and identifying and resolving discrepancies or contradictions, making it useful for a variety of problems and scenarios.

There are three types of knowledge graphs, depending on the complexity of the representation and the functionality of use:

Simple graphs are the most common and basic type of graphs. In such graphs, entities are represented as nodes, and edges are used to model the relationships between them. Simple graphs usually lack formal semantics for edges and nodes, which makes them easy to use, but limits the possibilities for deeper analysis and interpretation of data.

Hybrid graph or property graph. Hybrid graphs are designed to model entities and their relationships using a combination of standard network representations and formal semantics, such as Resource Description Framework (RDF: https://www.w3.org/RDF) and RDF Schema (RDFS: https://www.w3.org/TR/rdf11-mt). Unlike simple graphs, hybrid graphs based on these standards facilitate integration with other resources and provide greater opportunity for automated knowledge inference. This makes them a more powerful tool for representing and processing complex information.

Complex graphs, such as those in the KaBOB system (Livingston et al., 2015; Podkolodnyy et al., 2016), are often built on top of the Web Ontology Language (OWL). Complex graphs are highly expressive, allowing for efficient knowledge generation through deductive inference (Podkolodnyy et al., 2012). Due to its explicit semantics, OWL offers significant advantages over RDF/RDFS in integrating large amounts of biomedical data, making it particularly useful for complex problems in bioinformatics and systems biology.

As an example, Figure 1 provides a high-level network of the core interrelated biomedical concepts needed to model knowledge about pathways, genetic variants, diseases, and pharmaceutical treatments. At the top level are anatomical entities such as tissues, cells, and biological fluids (compartments) containing genomic entities such as DNA, RNA, mRNA, and proteins. DNA encodes genes, which are transcribed into mRNA and translated into proteins, which have molecular functions, can interact with each other, and participate in pathways and biological processes.

Recently, several software systems have been developed, such as KG-HUB (Caufield et al., 2023), Clinical KG (CKG) (Santos et al., 2022), RTX-KG2 (Wood et al., 2022), BioCypher (Lobentanzer et al., 2023), and Knowledge Base Of Biomedicine (KaBOB) (Livingston et al., 2015; Podkolodnyy et al., 2016), which provide broad functionality for creating and using knowledge graphs in bioinformatics and biomedicine, including the integration of large heterogeneous data.

The work (Callahan et al., 2024) describes the semantic ecosystem PheKnowLator (Phenotype Knowledge Translator) for automating the construction of ontological KGs with a fully customizable knowledge representation. The ecosystem includes various components for creating and using KGs to solve various applied problems, as well as pre-built KGs.

Integration of big data and creation of knowledge bases based on ontologies

Currently, in the field of bioinformatics, systems biology, agrobiology, biomedicine, more than a thousand ontologies have been developed that can be used to describe and integrate knowledge, analyze data, and infer new knowledge (https://bioportal.bioontology.org/ontologies).

As an example of one of the most successful projects for creating ontologies and, based on this, creating a knowledge base, we can cite the Gene Ontology (GO) project (http://www.geneontology.org/). GO describes current knowledge about the types of functional characteristics (more than 40 thousand concepts in total) that a gene product may have.

GO consists of 3 sections:

1. Molecular function – an elementary molecular activity or role that a gene or gene product can play in any biological processes. A total of 10,365 terms are described (https://geneontology.org/stats.html. Accessed 2024-09-08).

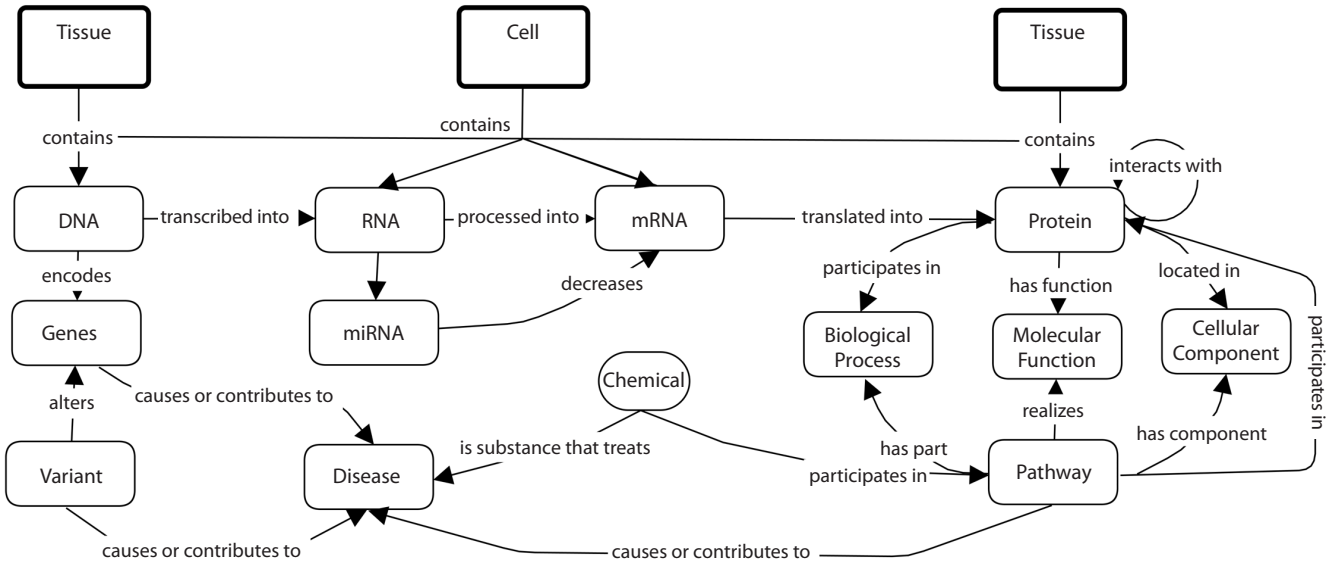


Fig. 1. Representation of knowledge about the levels of biological organization underlying the description of human diseases (Callahan, et al., 2024).

2. Biological process (a total of 26,552 terms are described. Accessed 2024-09-08) – a "biological program" that includes a set of molecular events or activities that act in a coordinated manner to achieve a specific result and relate to the functioning of integrated living units: cells, tissues, organs, and organisms. Unlike a function, a process must have several different stages with a defined beginning and end.
3. Cellular component – a part of the anatomical structure that describes the localization of a gene or its product in an organism, at the levels of cellular structures and macromolecular complexes or groups of gene products. A total of 4,022 terms are described (accessed 2024-09-08).

The main relationships between concepts used in GO include the simple class-subclass relationship (is_a), the part-whole relationship (part_of), the regulates, positively_regulates, and negatively_regulates relationships that describe relationships between biological processes, molecular functions, or biological properties. The transitivity property of the relationships used in GO allows one to construct a lattice of relationships between concepts and perform logical inference about the properties of concepts and their relationships (Podkolodnyy et al., 2016).

A knowledge base has been created based on GO, which in addition to terminological knowledge (GO gene ontology) includes the results of GOA gene annotation (Gene Ontology Annotation – http://www.ebi.ac.uk/GOA), i. e. knowledge about individual objects – genes and their products (Huntley et al., 2015). Currently, GOA includes more than 7.6 million GO annotations for almost 1.54 million proteins and more than 4.4 thousand species of organisms.

Initially, at the early stage of GO development, annotation of a gene or its product (protein or RNA) was carried out independently by molecular functions, biological processes or cellular components. In order to obtain information about the function of a gene or its product (RNA, protein) in a particular biological process and a particular cellular structure, it was necessary to develop another component of the GO knowledge base – the GO-CAM model of causal influences between gene products (Thomas et al., 2019).

GO-CAM links several GO annotations together to create models of biological processes that connect the activities of more than one gene product together into causal networks and allow specification of the biological context (e. g. cell/tissue type) in which the activities occur. As an example, the same biological model describing how the E3 ubiquitin-protein ligase NEDD4 represses RNA transcription in response to UV-induced DNA damage can be represented in two ways: as a set of disparate GO annotations, each capturing a partial description of the overall function (Fig. 2*a*), and as a GO-CAM scheme linking the GO annotations into a structured model of NEDD4 function, including the effect of NEDD4 activity on the activity of the RNA polymerase II macromolecular complex (Fig. 2*b*) (Thomas et al., 2019).

The basic unit of GO-CAM is the gene product activity unit, which combines the GO MF (molecular activity) annotation, together with the GO CC (cellular component) and GO BP (biological process) annotations, which provide the biological context of the activity. The context can be further specified by other ontologies, including Cell Type Ontology (Diehl et al., 2016), tissue/anatomical location (using several different ontologies depending on the species, e. g. the integrated cross-species anatomy ontology covering animals and merging several species-specific ontologies – Uberon (https://obophenotype.github.io/uberon/) (Mungall et al., 2012), or non-animal ontologies such as Plant ontology (https://planteome.org/) (Cooper, Jaiswal, 2016), or a description of a time period (e. g. biological phase GO). Activity units are related to each other by cause-and-effect

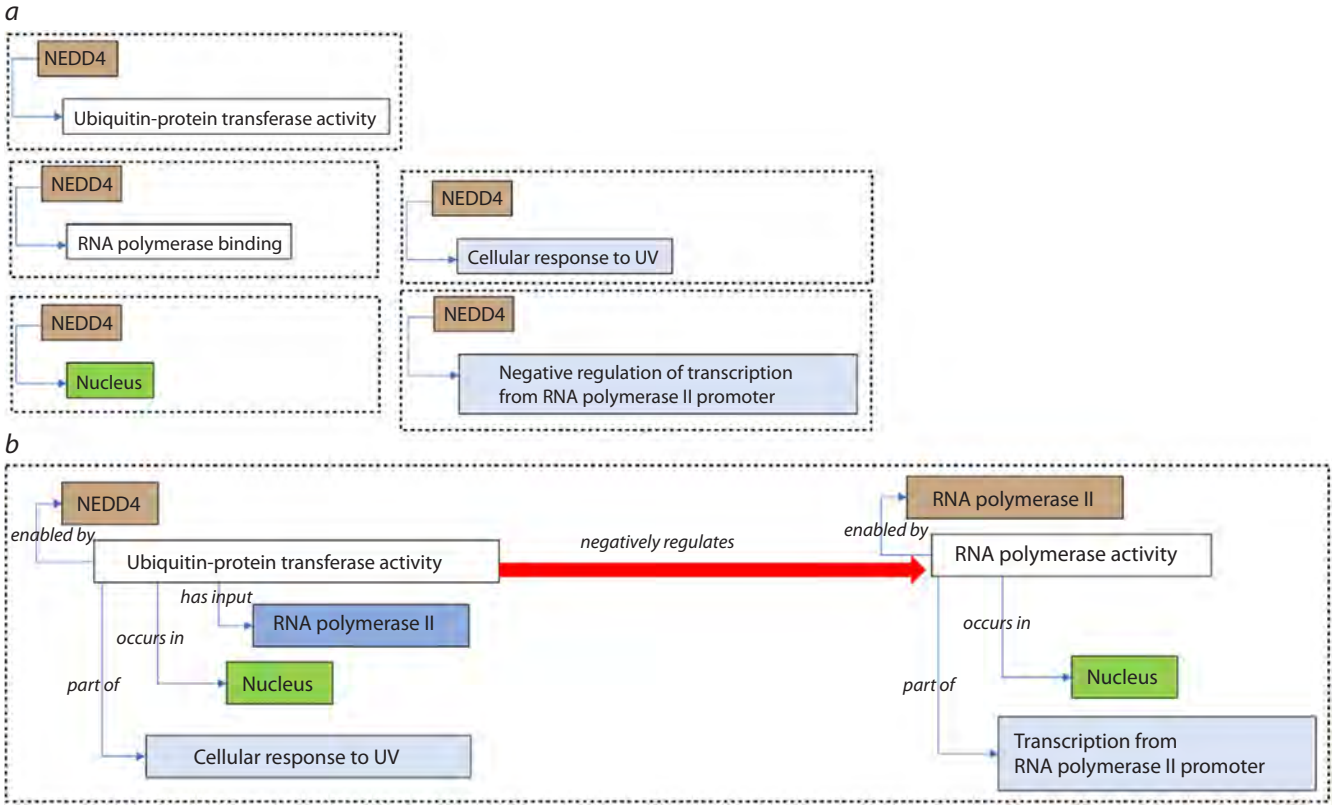


Fig. 2. The same biological model of how NEDD4 represses RNA transcription in response to UV-induced DNA damage described in two ways: *a* – as a set of disparate GO annotations, each capturing a partial description of the overall function; *b* – as a GO-CAM schema linking the GO annotations into a structured model of NEDD4 function, including the effect of NEDD4 activity on the activity of the RNA polymerase II macromolecular complex (Thomas et al., 2019).

relationships from the Relationship Ontology (Smith et al., 2005).

Causal networks in GO-CAM models also enable entirely new applications, such as network analysis of genomic data and logical modeling of biological systems. In addition, the models may also prove useful for pathway visualization. For example, the activity-based GO-CAM representation is compatible with the "activity flow diagrams" of the Systems Biology Graphical Notation (SBGN) standard (Bergmann et al., 2020).

GO-CAM thus provides the opportunity to use the massive GO and GOA knowledge base accumulated over the last 20 years as a basis not only for genomic biology representation of gene function, but also for a broader representation of systems biology and its novel applications to the interpretation of large-scale experimental data.

An example of GO analysis of genes of the associative gene network of rheumatoid arthritis

Earlier, the Institute of Cytology and Genetics SB RAS developed the ANDSystem software and information system for the automated extraction of medical and biological knowledge from scientific publications and a large number of biological and biomedical factual databases (Ivanisenko et al., 2015, 2019). The ANDSystem knowledge base is a unique resource containing formalized information in the form of associative gene networks (knowledge graphs) with almost 44 million interactions of various types between molecular genetic objects.

The original ontology underlying ANDSystem provides a very detailed description of the subject area. The ANDSystem knowledge base describes molecular genetic objects (proteins, genes, metabolites, microRNA), biological processes, phenotypic traits, drugs and their side effects, diseases, etc., as well as more than 25 types of interactions between these objects, including: physical interactions with the formation of molecular complexes (protein/protein, protein/DNA, metabolite/protein); catalytic reactions and proteolytic events involving a substrate/enzyme/product; regulatory interactions, functions/activities, transport and stability of proteins, metabolites and drugs, regulation of protein translation involving miRNA, regulation of biological processes and phenotypic traits involving proteins, metabolites and drugs; associative interactions of genes, proteins, metabolites, biological processes, phenotypic traits with diseases, etc.

An example of a typical task using ANDsystem is the reconstruction of an associative gene network (knowledge graph) of rheumatoid arthritis (RA) containing 1,025 genes/

proteins and more than 20 thousand interactions between them. Analysis of the overrepresentation of biological process terms in Gene Ontology for many rheumatoid arthritis genes, performed using the DAVID system (https://david.ncifcrf.gov/tools.jsp) revealed 376 biological processes statistically significantly associated with rheumatoid arthritis (see the Table). The *p*-values were calculated based on the hypergeometric distribution. The Bonferroni correction was used to account for multiple testing.

Let us consider in more detail the GO:0006955~immune response process, which has the lowest *p*-value, i. e. is most significantly associated with rheumatoid arthritis. Gene Ontology describes 420 genes associated with the "GO:0006955~immune response" term. 158 of them are present in the association network of rheumatoid arthritis (Fig. 3). For random reasons, such a large number of genes can be expected with a very low probability (*p*-value with Bonferroni correction $< 4.69 \cdot 10^{-79}$), which indicates a high significance of the relationship between rheumatoid arthritis and the immune response process and indicates the most important role of the immune system in the pathogenesis of this disease.

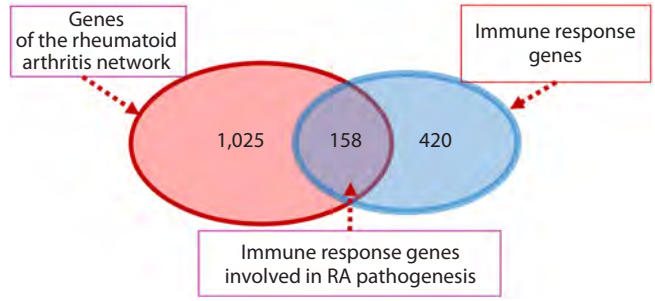


Fig. 3. Venn diagram describing the intersection of genes of the rheumatoid arthritis network and immune response genes (associated with the term GO:0006955~immune response).

The Table presents the list of the first 21 biological processes associated with rheumatoid arthritis and sorted by statistical significance (*p*-value with Bonferroni correction). Most of these terms are somehow related to the immune response and inflammation processes, which play an important role in the pathogenesis of rheumatoid arthritis. These processes are not independent.

List of the first 21 biological processes statistically most significantly associated with rheumatoid arthritis

Biological process (Gene Ontology)	*p*-value with Bonferroni correction
GO:0006955~immune response	$4.69 \cdot 10^{-79}$
GO:0006954~ inflammatory response	$2.13 \cdot 10^{-70}$
GO:0060326~chemotaxis	$2.49 \cdot 10^{-30}$
GO:0007267~cell-cell signaling	$8.59 \cdot 10^{-28}$
GO:0032496~response to lipopolysaccharide	$7.41 \cdot 10^{-27}$
GO:0070098~chemokine-mediated signaling pathway	$3.91 \cdot 10^{-25}$
GO:1990256~signal transduction	$2.91 \cdot 10^{-24}$
GO:0071222~cellular response to lipopolysaccharide	$5.45 \cdot 10^{-24}$
GO:0050729~positive regulation of inflammatory response	$6.31 \cdot 10^{-24}$
GO:2001023~regulation of response to drug	$1.70 \cdot 10^{-23}$
GO:0070374~positive regulation of ERK1 and ERK2 cascade	$8.26 \cdot 10^{-23}$
GO:0001666~response to hypoxia	$9.11 \cdot 10^{-23}$
GO:0071864~positive regulation of cell proliferation	$2.52 \cdot 10^{-22}$
GO:0042102~positive regulation of T cell proliferation	$6.90 \cdot 10^{-22}$
GO:0045087~innate immune response	$2.09 \cdot 10^{-18}$
GO:0032729~positive regulation of interferon-gamma production	$2.38 \cdot 10^{-18}$
GO:0045766~positive regulation of angiogenesis	$2.90 \cdot 10^{-18}$
GO:0043066~negative regulation of apoptotic process	$5.72 \cdot 10^{-18}$
GO:0050731~positive regulation of peptidyl-tyrosine phosphorylation	$8.13 \cdot 10^{-18}$
GO:0007166~cell surface receptor signaling pathway	$8.40 \cdot 10^{-18}$
GO:0007568~aging	$1.28 \cdot 10^{-17}$

Thus, the term "GO:0006955~immune response" is associated with such terms from this table as "GO:0045087~innate immune response", "GO:0032729~positive regulation of interferon-gamma production", "GO:0060326~chemotaxis", "GO:0042102~positive regulation of T cell proliferation", "GO:1990256~signal transduction" and others.

Similarly, the process "GO:0006954~inflammatory response" is associated with the terms "GO:0032496~response to lipopolysaccharide", "GO:0050729~positive regulation of inflammatory response", "GO:1990256~signal transduction", "GO:0001666~response to hypoxia", "GO:0045766~positive regulation of angiogenesis". And even the term "GO:0007568~aging" is related to the term "GO:0006954~inflammatory response", since one of the mechanisms of aging is chronic non-infectious inflammation.

These results on the example of rheumatoid arthritis indicate that the approach to identifying genes associated with a specific disease using ANDsystem and further GO analysis of this group of genes allows us to identify key biological processes involved in the pathogenesis of this disease.

Using ontology design patterns to integrate phenotype and biological attributes ontologies

Ontologies with logically rich axiomatization provide powerful capabilities such as automated reasoning, classification, and logical queries. However, manually creating such ontologies is extremely expensive and requires annotators to be not only domain experts but also have knowledge of logical modeling (Slater et al., 2020).

A popular approach to solving this problem is to use design patterns and template systems for logical axioms (Osumi-Sutherland et al., 2017). This allows separating the curation of reference terms used for logical definitions from their precise axiomatic picture. The central idea is to use a small number of axiom templates that implement design patterns, which can be created and maintained by logic experts, and for content curators to focus on selecting appropriate filler terms (e. g., terms from the Uberon ontology for defining anatomical attributes).

The Biological Attributes Ontology (OBA) is a standardized framework for observable attributes that are characteristics of organisms or parts of organisms (Stefancsik et al., 2023). Unlike most phenotypic ontologies, in OBA, the logical axioms define general attributes without reference to any specific phenotypic changes or states.

OBA was created using the Entity-Quality (EQ) design pattern, in which a phenotypic quality (Q), such as "height", "mass", or "amount" from the Phenotype and Trait Ontology (PATO) (Gkoutos et al., 2005), is combined with an entity (E), such as an anatomical or chemical entity, to form the concept of a "biological attribute" called a "trait". For example, the concept "blood glucose amount" (OBA:VT0000188) includes the class "amount" (PATO:000070), which defines the glucose characteristic – "glucose" (CHEBI:17234) in the blood – "blood" (UBERON:0000178).

Currently, OBA uses ten feature patterns from the Dead Simple OWL Design Patterns (DOS-DP) (Osumi-Sutherland et al., 2017). They were chosen because they cover most of the anatomical, chemical and cellular attributes that are central to genomics data integration.

A rich logical axiomatization based on design patterns is needed to ensure compatibility with existing phenotype ontologies and other data types, such as anatomical, chemical and biological data on metabolic pathways and gene networks.

Most attributes in OBA are inferred using OWL. These inferred definitions use terms from relevant reference ontologies such as Uberon (Mungall et al., 2012) or Chebi (Hastings et al., 2016). Except for a small number of high-level concepts, most of the classification in OBA is automatically computed based on the classifications of various reference ontologies, using automated inference. There are two advantages to this approach: first, no concepts need to be manually classified, which significantly reduces the cost of curating the classification while increasing its completeness. Second, multiple links to reference ontologies can be used for a wide variety of applications, including querying (e. g., retrieving all data where the morphology of a part of the renal system is affected), knowledge graph integration (e. g., automatic linking to phenotypic anomalies from widely used ontologies such as the human phenotype ontology (HPO) or mammalian phenotype ontologies (MP)), and knowledge inference (e. g., inferring missing data) (Dececchi et al., 2015).

Application of ontologies to interpret deep learning

Deep learning (DL) has clearly demonstrated its effectiveness in solving problems in the field of genomics, proteomics, biomedicine, including analysis and automatic functional annotation of DNA, RNA and protein sequences, search for DNA/RNA targets of regulatory RNAs and proteins, prediction of properties and functions of biomolecules, search for 3D protein structure, reconstruction of structures of biomolecules with given properties, prediction of interactions of biomolecules and identification of potential drug candidates on this basis, image processing and analysis, integration of omics data, analysis of complex, heterogeneous and interconnected biological networks (including protein-protein interaction networks, gene regulatory networks and metabolic pathways, semantic networks), modeling of biological systems and processes, etc. (Li et al., 2019; Sapoval et al., 2022).

One of the key problems of deep learning in bioinformatics, systems biology and modern biomedicine is the lack of interpretability of neural network models, which often function as "black box" models.

Interpretability of machine learning algorithms in bioinformatics and biomedicine is important for three main reasons. First, when analyzing complex systems, when there is no theory and a clear decision-making algorithm, it is necessary to understand why the model predicts a given

phenotype. Second, it is important to ensure that the model bases its predictions on a reliable representation of the data and does not focus on irrelevant artifacts. Finally, a model with highly accurate predictions may have revealed interesting patterns that biologists would like to study.

In the formal logical sense, interpretation is the mapping of a formal construct onto the entities and their relationships that it represents. In this sense, one can say that one understands a formal construct if one can relate it to relevant entities and propositions in the real world and reason about the consequences. However, it is important to distinguish the understandability of a model from the understandability of why the model is true or how the model was derived from the data, which raises questions about the validity of the model and the understandability of the learning algorithm.

Two main approaches to interpreting black boxes can be distinguished: a posteriori methods and self-explaining models (Adadi, Berrada, 2018). In the a posteriori method, the black box model is first learned and then an interpretive method is used to explain the predictions. However, explanations often do not match how the deep learning algorithm arrives at a solution. In addition, the explanation procedure is a separate method with its own errors that affect the quality of decisions made. Therefore, such an explanation is not always suitable for biomedicine.

It should be noted that interpretability is a concept specific to a particular domain, so there cannot be a universal definition. Very often, in an interpretable machine learning model, constraints are added to the model form so that it is either useful to someone or obeys structural knowledge of the domain, such as monotonicity (Gupta et al., 2015), causality, structural (generative) constraints, additivity (Lou et al., 2013), or physical constraints that come from knowledge of the subject domain (ontologies).

Currently, several works have been published on building self-explanatory neural networks based on gene expression data using Gene Ontology (GO) knowledge. For example, in the work (Bourgeais et al., 2021), a self-explanatory deep learning model called Deep GONet is proposed, integrating Gene Ontology into a hierarchical neural network architecture. This model is based on a fully connected architecture constrained by Gene Ontology annotations, so that each neuron represents a biological function. Experiments on cancer diagnostic datasets show that Deep GONet is easy to interpret and has high performance in distinguishing cancerous and non-cancerous samples.

Another example of an ontology-based self-explanatory neural network is ONN4MST, a generalization of the Ontology-based Neural Network (ONN) computational model for microbial source tracing (Zha, Ning, 2022). The ONN model uses a novel ontology-based approach that rewards predictions that satisfy the "biome" ontology. In other words, the ONN model can use biome ontology information to model dependencies between biomes and estimate the proportion of different biomes in a community sample.

The knowledge discovery capability of ONN4MST has been demonstrated in various source tracking applications. It enables source tracking of samples, the niches of which were less studied previously or unknown, detection of microbial contaminants, and identification of similar samples from ontologically distant biomes, demonstrating the unique importance of ONN4MST in knowledge discovery from a vast number of microbial community samples from heterogeneous biomes.

ONN4MST can distinguish samples from ontologically similar biomes, thus offering a quantitative way to characterize the evolution of the human gut microbial community. In particular, it is shown that the gut microbiome of centenarians differs from that of normal elderly people and shows a youthful pattern (Zha, Ning, 2022).

Conclusion

The rapid development of experimental technologies in the field of molecular biology has led to the fact that ontological modeling is becoming a basic method in bioinformatics and systems biology for integrating and analyzing heterogeneous experimental data and using them to build mathematical models of molecular genetic systems and processes. The creation of several hundred basic reference ontologies and their verification allows using these ontologies as sources of knowledge for integrating and building complex domain models and knowledge bases aimed at solving specific problems of biomedicine.

Ontologies are of particular importance for interpreting the results of computer predictions obtained using deep learning methods. In order for scientists to trust deep learning, which is often presented as "black box" models, special interpretation methods based on additional knowledge about the subject area or ontologies should be used. Ontologies, patterns of their construction, integration of big data and creation of knowledge graphs play a key role in increasing the interpretability of deep learning models. These tools not only improve the understanding of the results, but also provide higher quality data analysis. With the rapid growth of information volumes and the complexity of deep learning models, the use of ontologies is becoming a necessary step towards creating more transparent and explainable systems.

It can be expected that the new generation of interpretation systems will be able not only to explain the obtained solutions in a way understandable to humans, indicating the quantitative level of uncertainty, but also to suggest additional steps (e. g., additional experiments, clinical studies, etc.) necessary to clarify or reliably confirm their decisions.

References

Adadi A., Berrada M. Peeking inside the Black-Box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-52160. doi 10.1109/ACCESS.2018.2870052

Bergmann F.T., Czauderna T., Dogrusoz U., Rougny A., Drager A., Toure V., Mazein A., Blinov M.L., Luna A. Systems biology graphical notation markup language (SBGNML) version 0.3. *J. Integr. Bioinform*. 2020;17(2-3):20200016. doi 10.1515/jib-2020-0016

Bourgeais V., Zehraoui F., Ben Hamdoune M., Hanczar B. Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data. *BMC Bioinformatics*. 2021;22(S10):455. doi 10.1186/s12859-021-04370-7

Callahan T.J., Tripodi I.J., Stefanski A.L., Cappelletti L., Taneja S.B., Wyrwa J.M., Casiraghi E., Matentzoglu N.A., Reese J., Silverstein J.C., Hoyt C.T., Boyce R.D., Malec S.A., Unni D.R., Joachimiak M.P., Robinson P.N., Mungall C.J., Cavalleri E., Fontana T., Valentini G., Mesiti M., Gillenwater L.A., Santangelo B., Vasilevsky N.A., Hoehndorf R., Bennett T.D., Ryan P.B., Hripcsak G., Kahn M.G., Bada M., Baumgartner W.A., Hunter L.E. An open source knowledge graph ecosystem for the life sciences. *Sci. Data*. 2024;11(1):363. doi 10.1038/s41597-024-03171-w

Caufield J.H., Putman T., Schaper K., Unni D.R., Hegde H., Callahan T.J., Cappelletti L., Moxon S.A.T., Ravanmehr V., Carbon S., Chan L.E., Cortes K., Shefchek K.A., Elsarboukh G., Balhoff J., Fontana T., Matentzoglu N., Bruskiewich R.M., Thessen A.E., Harris N.L., Munoz-Torres M.C., Haendel M.A., Robinson P.N., Joachimiak M.P., Mungall C.J., Reese J.T. KG-Hub – building and exchanging biological knowledge graphs. *Bioinformatics*. 2023;39(7): btad418. doi 10.1093/bioinformatics/btad418

Chandrasekaran B., Josephson J., Benjamins V. What are ontologies, and why do we need them? *IEEE Intell. Syst. Appl*. 1999;14(1):20-26. doi 10.1109/5254.747902

Cooper L., Jaiswal P. The plant ontology: a tool for plant genomics. In Edwards D. (Ed.) Plant Bioinformatics. Methods in Molecular Biology. Vol. 1374. New York: Humana Press, 2016;89-114. doi 10.1007/978-1-4939-3167-5_5

Dececchi T.A., Balhoff J.P., Lapp H., Mabee P.M. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Syst. Biol*. 2015;64(6):936-952. doi 10.1093/sysbio/syv031

Diehl A.D., Meehan T.F., Bradford Y.M., Brush M.H., Dahdul W.M., Dougall D.S., He Y., Osumi-Sutherland D., Ruttenberg A., Sarntivijai S., Van Slyke C.E., Vasilevsky N.A., Haendel M.A., Blake J.A., Mungall C.J. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*. 2016; 7(1):44. doi 10.1186/s13326-016-0088-7

Gkoutos G.V., Schofield P.N., Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. *Brief Bioinform*. 2018;19(5):1008-1021. doi 10.1093/bib/bbx035

Gupta M., Cotter A., Pfeifer J., Voevodski K., Canini K., Mangylov A., Moczydlowski W., van Esbroeck A. Monotonic calibrated interpolated look-up tables. *J. Mach. Learn. Res*. 2016;17:1-47

Hastings J., Owen G., Dekker A., Ennis M., Kale N., Muthukrishnan V., Turner S., Swainston N., Mendes P., Steinbeck C. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res*. 2016;44(D1):D1214-D1219. doi 10.1093/nar/gkv1031

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):D1057-D1063. doi 10.1093/nar/gku1113

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

Li Y., Huang C., Ding L., Li Z., Pan Y., Gao X. Deep learning in bioinformatics: introduction, application, and perspective in big data era. *Methods*. 2019;166:4-21. doi 10.1016/j.ymeth.2019.04.008

Livingston K.M., Bada M., Baumgartner W.A., Hunter L.E. KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*. 2015;16(1):126. doi 10.1186/s12859-015-0559-3

Lobentanzer S., Aloy P., Baumbach J., Bohar B., Carey V.J., Charoentong P., Danhauser K., Doğan T., Dreo J., Dunham I., Farr E., Fernandez-Torras A., Gyori B.M., Hartung M., Hoyt C.T., Klein C., Korcsmaros T., Maier A., Mann M., Ochoa D., Pareja-Lorente E., Popp F., Preusse M., Probul N., Schwikowski B., Sen B., Strauss M.T., Turei D., Ulusoy E., Waltemath D., Wodke J.A.H., Saez-Rodriguez J. Democratizing knowledge representation with BioCypher. *Nat. Biotechnol*. 2023;41(8):1056-1059. doi 10.1038/s41587-023-01848-y

Lou Y., Caruana R., Gehrke J., Hooker G. Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Assoc. for Computing Machinery, 2013;623-631. doi 10.1145/2487575.2487579

Mungall C.J., Torniai C., Gkoutos G.V., Lewis S.E., Haendel M.A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 2012;13(1):R5. doi 10.1186/gb-2012-13-1-r5

Osumi-Sutherland D., Courtot M., Balhoff J., Mungall C. Dead simple OWL design patterns. *J. Biomed. Semant*. 2017;8:18. doi 10.1186/s13326-017-0126-0

Podkolodnyy N.L., Ignatyeva E.V., Podkolodnaya O.A., Kolchanov N.A. Information support of research on transcriptional regulatory mechanisms: an ontological approach. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2012; 16(4/1):742-755 (in Russian)

Podkolodnyy N.L., Podkolodnaya O.A. Ontologies in bioinformatics and systems biology. *Russ. J. Genet. Appl. Res.* 2016;6(7):749-758. doi 10.1134/S2079059716070091

Qaiser A., Ghulam S. Bioinformatics and big data analytics in genomic research. *Med. Pap.* 2023;3(1):165-179. doi 10.31219/osf.io/5grpc

Santos A., Colaço A.R., Nielsen A.B., Niu L., Strauss M., Geyer P.E., Coscia F., Albrechtsen N.J.W., Mundt F., Jensen L.J., Mann M. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol*. 2022;40(5):692-702. doi 10.1038/s41587-021-01145-6

Sapoval N., Aghazadeh A., Nute M.G., Antunes D.A., Balaji A., Baraniuk R., Barberan C.J., Dannenfelser R., Dun C., Edrisi M., Elworth R.A.L., Kille B., Kyrillidis A., Nakhleh L., Wolfe C.R., Yan Z., Yao V., Treangen T.J. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun*. 2022;13(1):1728. doi 10.1038/s41467-022-29268-7

Slater L.T., Gkoutos G.V., Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. *BMC Med. Inform. Decis. Mak*. 2020;20(Suppl.10):311. doi 10.1186/s12911-020-01336-2

Smith B., Ceusters W., Klagges B., Kohler J., Kumar A., Lomax J., Mungall C., Neuhaus F., Rector A.L., Rosse C. Relations in biomedical ontologies. *Genome Biol*. 2005;6(5):R46. doi 10.1186/gb-2005-6-5-r46

Stefancsik R., Balhoff J.P., Balk M.A., Ball R.L., Bello S.M., Caron A.R., Chesler E.J., de Souza V., Gehrke S., Haendel M., Harris L.W., Harris N.L., Ibrahim A., Koehler S., Matentzoglu N., McMurry J.A., Mungall C.J., Munoz-Torres M.C., Putman T., Robinson P., Smedley D., Sollis E., Thessen A.E., Vasilevsky N., Walton D.O., Osumi-Sutherland D. The Ontology of Biological Attributes (OBA)-computational traits for the life sciences. *Mamm. Genome*. 2023;34(3):364-378. doi 10.1007/s00335-023-09992-1

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E. Big Data: astronomical or genomical? *PLoS Biol*. 2015;13(7):e1002195. doi 10.1371/journal.pbio.1002195

Thomas P.D., Hill D.P., Mi H., Osumi-Sutherland D., Van Auken K., Carbon S., Balhoff J.P., Albou L.-P., Good B., Gaudet P., Lewis S.E., Mungall C.J. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet*. 2019;51(10):1429-1433. doi 10.1038/s41588-019-0500-1

Wood E.C., Glen A.K., Kvarfordt L.G., Womack F., Acevedo L., Yoon T.S., Ma C., Flores V., Sinha M., Chodpathumwan Y., Termehchy A., Roach J.C., Mendoza L., Hoffman A.S., Deutsch E.W., Koslicki D., Ramsey S.A. RTX-KG2: a system for building a semantically standardized knowledge graph for translational biomedicine. *BMC Bioinformatics*. 2022;23(1):400. doi 10.1186/s12859-022-04932-3

Zha Y., Ning K. Ontology-aware neural network: a general framework for pattern mining from microbiome data. *Brief. Bioinform*. 2022;23(2):bbac005. doi 10.1093/bib/bbac005

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

Received October 28, 2024. Revised November 8, 2024. Accepted November 11, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):950-959

doi 10.18699/vjgb-24-102

PlantReg: the reconstruction of links between transcription factor regulatory networks and biological processes under their control

V.V. Lavrekha [1, 2#], N.A. Omelyanchuk [1#], A.G. Bogomolov [1], 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 description of the path from a gene to a trait, as the main task of many areas in biology, is currently being equipped with new methods affecting not only experimental techniques, but also analysis of the results. The pleiotropic effect of a gene is due to its participation in numerous biological processes involved in different traits. A widespread use of genome-wide sequencing of transcripts and transcription factor (TF) binding regions has made the following tasks relevant: unveiling pleiotropic effects of TFs based on the functions of their target genes; compiling the lists of TFs that regulate biological processes of interest; and describing the ways of TF functioning (their primary and secondary targets, higher order targets, TF interactions in the process under study). We have previously developed a method for the reconstruction of TF regulatory networks and proposed an approach that allows identifying which biological processes are controlled by these networks and how this control is exerted. In this paper, we have implemented the approach as PlantReg, a program available as a web service. The paper describes how the program works. The input consists of a list of genes and a list of TFs – known or putative transcriptional regulators of these genes. As an output, the program provides a list of biological processes enriched for these genes, as well as information about by which TFs and through which genes these processes are controlled. We illustrated the use of PlantReg deciphering transcriptional regulation of processes initiated at the early salt stress response in *Arabidopsis thaliana* L. With PlantReg, we identified biological processes stimulated by the stress, and specific sets of TFs that activate each process. With one of these processes (response to abscisic acid) as an example, we showed that salt stress mainly affects abscisic acid signaling and identified key TFs in this regulation. Thus, PlantReg is a convenient tool for generating hypotheses about the molecular mechanisms that control plant traits.

Key words: gene ontology; biological processes; gene regulatory networks; *Arabidopsis thaliana*.

For citation: 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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):950-959. doi 10.18699/vjgb-24-102

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

PlantReg: реконструкция связей между регуляторными сетями транскрипционных факторов и контролируемыми ими признаками

В.В. Лавреха [1, 2#], Н.А. Омельянчук [1#], А.Г. Богомолов [1], 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

Аннотация. Описание пути от гена к признаку как основная задача многих отраслей биологии в настоящее время оснащается новыми методами не только в технике экспериментов, но и в системном анализе их результатов. Плейотропный эффект гена осуществляется за счет его участия во многих биологических процессах, вовлеченных в разные признаки. Широкое распространение полногеномного секвенирования транскриптов и районов связывания транскрипционных факторов (ТФ) сделало актуальными задачи установления плейотропных эффектов ТФ за счет знаний о функциях их мишеней, составление списков ТФ, регулирующих интересующие исследователя биологические процессы, описание путей их действия (первичные и вторичные мишени, мишени следующих порядков, взаимодействие между ТФ в исследуемом процессе). Ранее мы разработали метод реконструкции регуляторных сетей ТФ и предложили подход, позволяющий выявлять, какие биологические процессы и каким образом эти сети регулируют. В данной работе мы реализовали этот подход в виде программы

# Equal contribution

PlantReg, доступной пользователям через веб-интерфейс. В статье описан принцип работы программы. На вход подаются список генов и список ТФ – известных или предполагаемых регуляторов транскрипции этих генов. На выходе программа выдает список биологических процессов, которые обогащены в этих генах, а также информацию о том, какими ТФ и через какие гены эти процессы регулируются. Работа PlantReg проиллюстрирована на примере исследования регуляции процессов, инициируемых на начальных этапах ответа на солевой стресс у *Arabidopsis thaliana* L. С помощью программы PlantReg нами выявлены биологические процессы, стимулируемые в раннем ответе на солевой стресс, и специфический набор ТФ, активирующих каждый из этих процессов. На примере одного из таких процессов – ответа на фитогормон абсцизовую кислоту – мы показали, что солевой стресс активирует в основном сигнальный путь этого гормона, и выделили ключевые ТФ в этой регуляции. Таким образом, программа PlantReg – удобный инструмент для создания гипотез о молекулярных механизмах регуляции признаков растений.

Ключевые слова: генная онтология; биологические процессы; регуляторные генные сети; *Arabidopsis thaliana*.

Introduction

The efficient development of transcriptome sequencing methods has opened up wide opportunities not only to study changes in gene expression at the level of transcription, but also to track the regulation of these changes by transcription factors (TFs) and their impact on biological processes (Chen J.W. et al., 2023). In this regard, methods for compilation of TF lists based on the presence of their binding sites in the promoters of differentially expressed genes (DEGs) and methods for gene ontology (GO) terms enrichment analysis of gene lists (i. e., their functional annotation) are now widely used. Nevertheless, identification of the relationship between the outputs of these methods (i. e., determination of TFs that affect specific biological processes, their stages influenced by these TFs, common and specific TFs among the processes) remains a poorly worked out part in the analysis of transcriptomic data. The development of computer programs for this purpose will make this analysis more systematic and build a connection between alterations in gene expression and changes in biological processes.

If TFs regulate each other at the transcription level, their interactions are often represented as graphs – transcription factor regulatory networks (TFRNs), which can be reconstructed using various methods (Hecker et al., 2023). TFRNs allow establishing hierarchy in their architecture and identifying hubs – TFs that are most connected to other TFs. Altering the expression of genes encoding hubs is likely to change the functioning of the entire TFRN, and consequently affects downstream biological processes (He, Zhang, 2006).

We have previously developed a methodology and a software for reconstruction of TFRNs. We have also proposed a bioinformatics approach to identify biological processes under control of TFRNs and regulatory links between TFRN components and the processes (Omelyanchuk et al., 2024). It is based on the following steps. The first step is compilation of a list of TFs enriched in DEG promoters. The TF list is then used for TFRN reconstruction. The second step is functional annotation of the DEG list, after which within every biological process potential regulators of each of its DEGs are extracted from the TF list composed at the first step. After this, the genes are arranged in the order in which they function during a biological process, and the TFs that control the individual stages of this process can be identified. The use of this approach was illustrated in (Omelyanchuk et al., 2024) with the examples of auxin regulation of chlorophyll and lignin biosynthesis, abscisic acid signaling, and ribosome biogenesis.

In this work, we implemented this approach as a PlantReg program, available to users via a web interface (https://plamorph.sysbio.ru/fannotf/). We used PlantReg to investigate the regulation of processes during an early salt stress response in *Arabidopsis thaliana* L., using transcriptomic data from (Wu et al., 2021a). With PlantReg, we found that processes involved in the early reaction to salt stress and coordinated by all TFs within the TFRN include responses to heat, red and far-red light, and salicylic acid. The largest number of processes (programmed cell death, leaf senescence, and responses to blue light, hypoxia, reactive oxygen species, dehydration, abscisic acid, and jasmonic acid) are regulated by at least 70 % of TFs from the TFRN. In the control of the endoplasmic reticulum (ER) unfolded protein response, biosynthesis of indole-containing compounds and S-glucosides, as well as water transport, less than 50 % of the TFRN is involved.

Next, we examined the PlantReg results on the regulation of the abscisic acid (ABA) response during early salt stress in more detail and found that this regulation is primarily mediated through the control of ABA signaling, and its last stage, activation of the master TFs, is modulated most strongly. Both TFRN hubs (WRKY8 and DEAR2) are involved in this activation, and DEAR2 also controls ABA reception. Thus, the PlantReg program is an effective tool for analyzing data on differential gene expression in transcriptomes and creating hypotheses about the molecular mechanisms operating in regulation of biological processes.

Materials and methods

PlantReg implementation. PlantReg workflow is shown in Figure 1. The program takes a list of genes (in this work, we focus on DEG lists) and a list of TFs that are known or putative transcriptional regulators of these genes as input. The FuncAnnot function performs functional annotation of the gene list using the clusterProfiler R package (Yu et al., 2012; Wu et al., 2021b). The result is a file containing information about the GO terms enriched in the DEG list, as well as sublists of genes from the input annotated with the enriched GO terms. The next step is the search for the overlaps between the binding peaks of the input TFs and 5′ regulatory regions of genes from the sublists. For this purpose, the TF-targets function,

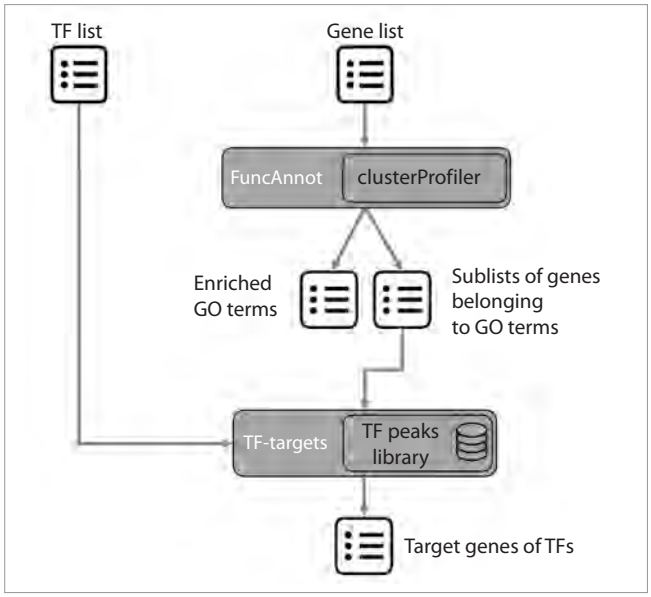


Fig. 1. The PlantReg workflow.

which we developed earlier as part of the CisCross-FindTFnet program (Omelyanchuk et al., 2024), is applied. As output, the user receives a file containing enriched GO terms and their associated DEGs, evidence codes, and TFs, the binding peaks of which are mapped to the 5′ regulatory regions of DEGs associated with the enriched GO terms.

The core of the PlantReg software is implemented in Perl and recruits the clusterProfiler R package. PlantReg is accessible through a web interface (https://plamorph.sysbio.ru/fannotf/). In the web version of PlantReg, two collections of TF binding profiles are available for identifying target genes of TFs. The first collection (GTRD-MACS2) includes 306 sets of ChIP-seq peaks for 131 *A. thaliana* TFs downloaded in BED format from the GTRD database (https://gtrd.biouml.org/#!) (Kolmykov et al., 2021). The second collection (CisCross-MACS2) was obtained by large-scale profiling of *A. thaliana* TF binding sites using DAP-seq (O'Malley et al., 2016) and represents the result of re-processing of raw data from the original study (Lavrekha et al., 2022). This collection contains 608 peak sets for 404 TFs of *A. thaliana*. The ARAPORT11 annotation of *A. thaliana* genome (https://bar.utoronto.ca/thalemine/begin.do) is used to identify 5′ regulatory regions of genes (500, 1,000, 1,500, 2,000, or 2,500 bp upstream of the transcription start) in the PlantReg web version.

Reconstruction of the TFRN for early response to salt stress. To reconstruct the TFRN for early response to salt stress, we used publicly available RNA-seq data for seven-day-old *A. thaliana* seedlings (ecotype Col-0) grown in the light, before and after salt treatment (100 mM NaCl) for 1 h (Wu et al., 2021a). To extract DEGs, we set the FDR threshold at 0.05; among them, we distinguished upregulated and downregulated DEGs (uDEGs and dDEGs, respectively). The TFRN was reconstructed using the CisCross-FindTFnet program (Omelyanchuk et al., 2024) with the following parameters. For mapping of TF binding regions, we used the CisCross-MACS2 collection of peaks, and set the length of the 5′ regulatory regions to 1,000 bp. The positions of transcription start sites were determined according to the ARAPORT11 *A. thaliana* genome annotation. In TF binding peak enrichment analysis of 5′ regulatory regions of uDEGs and dDEG, we controlled FDR at 0.001 using the Benjamini–Hochberg method. To reconstruct "TF-regulator–TF-target" pairs within the TFRN, we used the peak sets corresponding to the binding of TFs to the native leaf genomic DNA possessing methylation marks.

Reconstruction of the links between the TFRN for early response to salt stress and the biological processes it controls. Using PlantReg, we reconstructed the links between the TFRN for early response to salt stress and downstream biological processes. As input, we used a list of TFs from the TFRN, as well as a list of DEGs responding to salt treatment (uDEGs and dDEGs were analyzed separately). The length of the 5′ regulatory regions was set to 1,000 bp, and the CisCross-MACS2 collection was used to map TF binding peaks. For further analysis and interpretation, we only used "TF-regulator–Target gene" pairs reconstructed based on DAP-seq TF binding profiles captured in leaf genomic DNA possessing methylation marks.

Results and discussion

Biological interpretation of PlantReg output data

The PlantReg program is designed to reconstruct molecular mechanisms operating in genetic regulation of traits. To get started, the user needs to have a list of known or putative regulators of differential gene expression. PlantReg performs a functional annotation for the list of DEGs, then searches for potential targets of TFs among DEGs associated with enriched biological processes. The mapping of TF binding peaks in the 5′ regulatory regions of genes is performed using a representative collection of whole-genome TF binding profiles for the species being studied. In the web version, two collections of TF binding profiles for *A. thaliana*, from ChIP-seq or DAP-seq data, are available. The user can choose one of them. The program outputs the relationships between DEGs, the upstream TFs, and the enriched GO terms.

For convenient biological interpretation and subsequent analysis, PlantReg output is organized in five blocks. The first four blocks offer four alternative representations of the same results. So, block (1) characterizes genes. It contains a sublist of DEGs annotated with the enriched GO terms, the list of potential TFs (with an indication of TF family) and the number of TFs for each DEG (Fig. 2*a*). Each DEG is also characterized with the total number and the list of enriched GO terms (with an indication of the evidence code), which facilitates identification of DEGs involved in a wide range of biological processes as well as DEGs specific to particular processes.

Biological processes are the focus of block (2). In this block, for each enriched GO term, a sublist of associated DEGs with the evidence codes is created, as well as a sublist of TFs potentially regulating the expression of these DEGs with an indication of TF family (Fig. 2*b*). This output block allows reconstructing the mechanism of genetic regulation for each biological process.

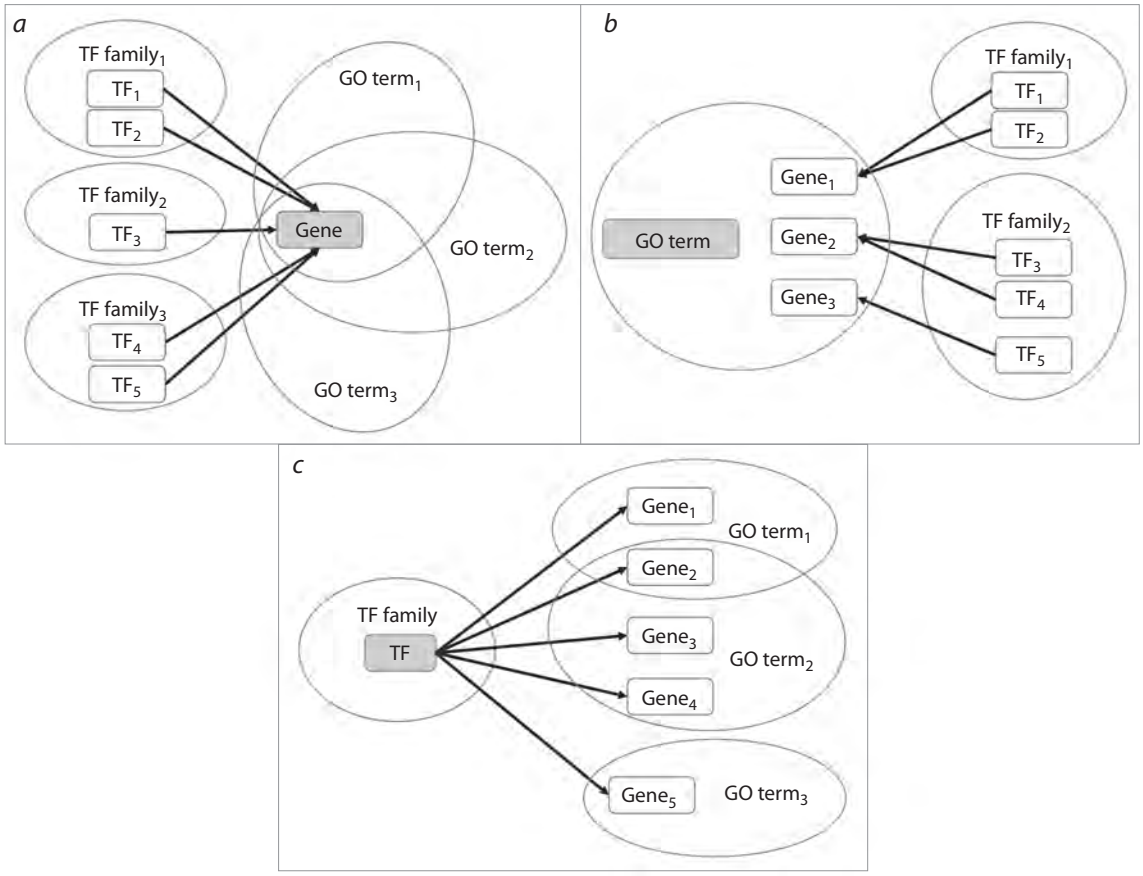


Fig. 2. The PlantReg output representations.

Panels *a*, *b* and *c* correspond to output blocks 1, 2 and 3. The central output element is highlighted in gray.

Block (3) characterizes transcriptional regulators of differential gene expression. It contains a list of TFs, for which the target genes associated with enriched GO terms were found among DEGs (Fig. 2*c*). This output representation is useful for planning the experiments to verify the predicted mechanisms for genetic regulation of biological processes.

Block (4) holds a table where each row contains one DEG, one of the TFs potentially regulating its expression, its family, and one of the GO terms with the evidence codes. This output can be used for further analysis with software tools.

The auxiliary block (5) accommodates the results of functional annotation of DEGs by clusterProfiler with the significance of GO term enrichment.

Functional annotation of the TFRN for early response to salt stress in *A. thaliana*

We used the PlantReg program to investigate the mechanisms that regulate the response to salt stress in the model plant species *A. thaliana*. A list of DEGs that respond to high salt concentration was extracted from publicly available transcriptome data (Wu et al., 2021a). In order to generate a list of potential TF regulators for these genes, we used the previously developed CisCross-FindTFnet tool. Based on the combined analysis of DEGs and TF binding profiles, this tool identifies potential TF regulators of DEGs, classifies them by regulation type, determines the relationships between them and reconstructs a TFRN (Omelyanchuk et al., 2024).

TF regulation types are distinguished based on a set of rules and correspond to the following properties of the regulators. US (upregulated suppressor) is a suppressor induced by the stimulus (in our case, high salt concentration). It suppresses the expression of target genes that were active before the stimulus application. UA (upregulated activator) is induced by the stimulus and activates expression of its target genes. DA (downregulated activator) and DS (downregulated suppressor) are active in the absence of the stimulus. The application of the stimulus inhibits DA expression and, consequently, expression of its target genes. DS suppresses activity of its target genes in the absence of the stimulus; under the stimulus exposure, DS expression is reduced and the activity of its targets is unblocked.

The structure of the early salt stress response TFRN reconstructed with the CisCross-FindTFnet program is shown in Figure 3*a*, and consists only of TFs, the binding sites of which were enriched in uDEGs, i. e., the response to salt stress begins with transcription activation, and TFs in the TFRN are related only to the DS and UA types, i. e., gene activation occurs passively due to stress-induced downregulation of the suppressor (DS) or actively due to stimulation of the activator (UA). Among UA-encoding genes, increased activity under salt stress was previously experimentally shown for *CBF4/DREB1D* (Sakuma et al., 2002), *ERF37/DREB A-4* (Hossain et al., 2016), *RAP2.1/DEAR6* (Ghorbani et al., 2019), *WRKY25* (Jiang, Deyholos, 2009), *ABI5* (Yuan et al., 2011),

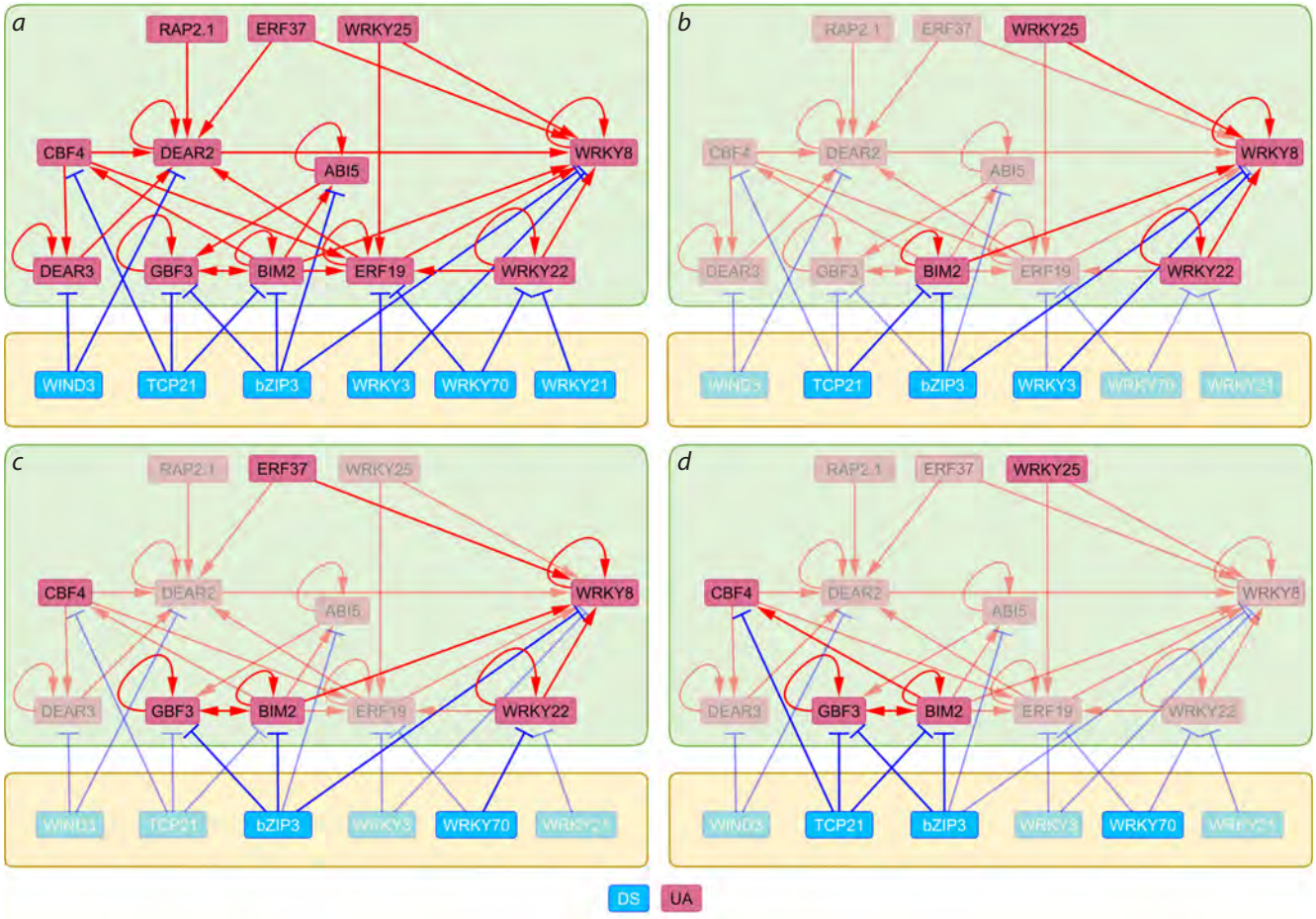


Fig. 3. The reconstructed TFRN for the early salt stress response in *A. thaliana* (*a*) and its participation in the regulation of processes that compose the salt stress response: ER unfolded protein response (*b*), biosynthesis of indole-containing compounds (*c*) and S-glycosides (*d*).

The nodes of the graphs correspond to transcription factors. TF1 and TF2 are connected by an edge directed from TF1 (regulator) to TF2 (target) if the TF1 binding peak is mapped in the 5′ regulatory region of the TF2-encoding gene. The green block highlights the group of TFs (UAs) that are activated and activate their targets in response to salt stress. The yellow block highlights the group of TFs (DSes) that repress genes normally and are themselves repressed by salt stress, which results in passive activation of the DS targets. The nodes and edges involved in the regulation of the process are highlighted in panels *b*–*d*. DSes and UAs denote downregulated suppressors and upregulated activators according to (Omelyanchuk et al., 2024).

GBF3 (Zhang L. et al., 2012, 2017) and *WRKY8* (Hu et al., 2013). Wherein, *ABI5* (Yuan et al., 2011), *GBF3* (Zhang L. et al., 2012, 2017), and *WRKY8* (Hu et al., 2013) play a key role in response to salt stress.

For DSes we identified, it was previously demonstrated that inactivation of *WRKY70* increased plant tolerance to salt stress (Li J. et al., 2013), and *bZIP3* expression was inhibited by salt stress (Liu Y. et al., 2013). Notably, longer salt stress (4 h) activated *WRKY3* (Li P. et al., 2021). Thus, the composition of our reconstructed TFRN is in good accordance with the published data. At the same time, only four TFs out of 18 (22 %) have been previously identified as the key players in salt stress, and only 10 (56 %) have been described to respond to salt stress, i. e., the TFRN contains new potential regulators of this process.

Functional annotation of DEGs showed that the early response to salt stress is accompanied by the ER unfolded protein response, as well as activation of the following processes: programmed cell death, leaf senescence, water transport, biosynthesis of indole-containing compounds and S-glycosides, response to heat, red and far-red light, abscisic, salicylic and jasmonic acids, blue light, hypoxia, reactive oxygen species, and dehydration. A link between the response to salt stress and heat has been shown previously, as heat shock proteins enhance resistance to salt stress and, conversely, overexpression of salt stress proteins provides resistance to heat stress (Azameti et al., 2024; Chaffai et al., 2024; Chang et al., 2024). The relationship of salt stress response to leaf senescence, hypoxia, water transport, responses to blue, red, and far-red light, reactive oxygen species, dehydration, abscisic acid, salicylic acid, and jasmonic acid has also been demonstrated in experiments (Serraj et al., 1994; Szepesi et al., 2009; Khan et al., 2012; Kumar et al., 2014; Joseph, Jini, 2010; Sharma et al., 2022; Kesawat et al., 2023; Lu, Fricke, 2023; Tan et al., 2023; Peng et al., 2024).

Salt stress leads to disruption of protein folding in the endoplasmic reticulum (so-called endoplasmic reticulum stress), and the response to this is optimization of protein folding, resulting in a decrease in unfolded proteins (Liu et al., 2007; Wang et al., 2011). There is evidence for the involvement of

biosynthesis of an indole-containing compound such as melatonin in the response to salt stress (Qi et al., 2020; Shamloo-Dashtpagerdi et al., 2022). Enrichment of salt stress response genes with the gene ontology term "S-glycoside metabolism" has been detected previously (Rodriquez et al., 2021).

We found that all TFs in the TFRN are involved in the regulation of the response to heat, red and far-red light, and salicylic acid. The remaining biological processes fell into two groups: those controlled by at least 70 % of the network TFs and those controlled by less than 50 % of the network TFs. The first group included programmed cell death, leaf senescence, and responses to blue light, hypoxia, reactive oxygen species, dehydration, abscisic acid, and jasmonic acid. The second group comprised the ER unfolded protein response (Fig. 3*b*), biosynthesis of indole-containing compounds (Fig. 3*c*) and S-glycosides (Fig. 3*d*), and water transport (the latter was regulated by only three TFs: BIM2, bZIP3, and WIND3). Thus, using PlantReg, we have shown that the response to salt stress is composed of both processes regulated by the entire TF network and processes controlled by distinct parts of this network.

Among the TFs we have identified as controlling the ER unfolded protein response, only WRKY70 has been shown as a regulator of this process to date (Wang L.Y. et al., 2023), and bZIP3 has been indicated as a possible candidate for this role (Ko et al., 2023).

Glucosinolates, the most diverse and studied group of S-glycosides, are the secondary metabolites of Brassicaceae involved in plant defense (Halkier, Gershenzon, 2006). Currently, they are intensively studied due to their therapeutic and preventive properties against cancer, cardiovascular or neurological diseases. Glucosinolates are categorized into three groups depending on the amino acids from which they are derived: aliphatic glucosinolates (methionine, alanine, leucine, isoleucine, and valine), aromatic glucosinolates (phenylalanine and tyrosine), and indole glucosinolates (tryptophan). For at least three out of seven TFs that we found to control glucosinolate biosynthesis, this function was previously known. CBF4 triggers the synthesis of aliphatic glucosinolates, which also increases salt stress tolerance (Defoort et al., 2018), while WRKY70 suppresses indole-3-ylmethyl glucosinolate biosynthesis (Li J. et al., 2006). *GBF3* expression is significantly reduced in mutants for the *SUR2/CYP83B1* gene that controls the metabolic switch between auxin and indole glucosinolate biosynthesis (Morant et al., 2010).

Regulation of abscisic acid signaling pathway under salt stress in *A. thaliana*

In addition to determining the composition of TFs that control specific processes, PlantReg allows determination of TFs that regulate the activity of individual genes in these processes. The latter provides an opportunity to identify modulators of gene expression consistently at each stage of the process. In this paper, we demonstrate this on the example of reconstructing the mechanism for transcriptional regulation of ABA response under salt stress. According to PlantReg results, all TFs within the salt stress response TFRN except for WRKY21 control ABA response. This regulation starts with the control of ABA level in the cell.

At this stage (stage 1 in Figure 4), the targets of the TFRN include the *ABCG25* and *ATAF1* genes encoding, respectively, the ABA exporter from the cell (Park et al., 2016) and the TF that activates both the ABA biosynthesis gene *NCED3* (Jensen et al., 2013) and the ABA importer gene *ABCG40* (Kang et al., 2015).

In the next stage (stage 2 in Figure 4), ABA binds to and activates the PYRABACTIN RESISTANCE1/PYR1 LIKE/REGULATORY COMPONENTS OF ABA RECEPTORS (PYR/PYL/RCAR) group of receptors (Fidler et al, 2022), among which the salt stress response TFRN controls *PYL7*. It is the most tightly TFRN-controlled gene in ABA signaling, since its expression is managed by half of the TFRN TFs (9 of 18). Under normal conditions, *PYL7* activity is suppressed by bZIP3 and WIND3. Whereas bZIP3 inhibits the activity of 11 ABA signaling genes in addition to *PYL7*, WIND3 is a specific suppressor of *PYL7*. Salt stress activates *PYL7* through seven TFs that form a regulatory loop with *DEAR2* being a hub, directly activated by five TFs (CBF4, DEAR3, ERF19, ERF37, RAP2.1), while the sixth TF (WRKY22) stimulates it through ERF19.

In ABA signaling, PYR/PYL/RCAR receptors inhibit PP2C phosphatase activity, thereby preventing dephosphorylation of SnRK2 kinases (Fidler et al., 2022). Here, the direct TFRN targets are genes encoding the following: PP2C phosphatases PP2C5, ABI2 and HAI/2, as well as the SNRK2.6 activator RPK1 (Shang et al., 2020), PP2C phosphatase regulators EDL3 (Koops et al, 2011), LOG2 (Pan W. et al., 2020), and phospholipase PLDALPHA1, the product of which (phosphatidic acid) inhibits the activity of some PP2C phosphatases (Ndathe, Kato, 2024).

The third stage of ABA signal transduction (stage 3 in Figure 4) begins with the activation of ABA response master TFs by SnRK2 kinases. Notably, one of them, ABI5, is also represented in the TFRN. Except for *ABI5* and *MAPKKK17/18* (initiators of the MAPK cascade) (Zhou M. et al., 2021; Zhao et al., 2023), all other TFRN targets at this stage represent regulators of ABA response master TFs. These include genes encoding kinases CPK4/6, PKS5, EDR1 (Zhu et al., 2007; Wawrzynska et al., 2008; Zhou X. et al., 2015; Zhang H. et al., 2020), transcription factors ABR1 (Sanyal, Pandey, 2024) and HFR1 (Wang Z. et al., 2024), transcriptional regulators VQ18 (Pan J. et al., 2018) and PRN1 (Warpeha et al., 2007), components of the protein degradation complexes PUB9 (Samuel et al., 2008), AFP1 (Lopez-Molina et al., 2003), and RHA2B (Li H. et al., 2011).

Interestingly, within the TFRN, half of DSes and all UAs are involved in the control of the third step of ABA signaling. Both TFRN hubs, DEAR2 and WRKY8, have targets at this stage. Moreover, while DEAR2 has targets at stage 2 as well, WRKY8 is specific for stage 3. WRKY8 and DEAR2 enhance transcription of seven and six activators, respectively. During viral infection, WRKY8 controls ABA signaling as an infection-suppressed activator of *ABI4* (Chen L. et al., 2013). We showed that under salt stress, WRKY8 controls ABA signaling by upregulating *CPK6*. CPK6 kinase stimulates ABF4 and ABI5 through their phosphorylation (Zhang H. et al., 2020). This suggests that the same TF may have different targets in ABA signaling under various stresses.

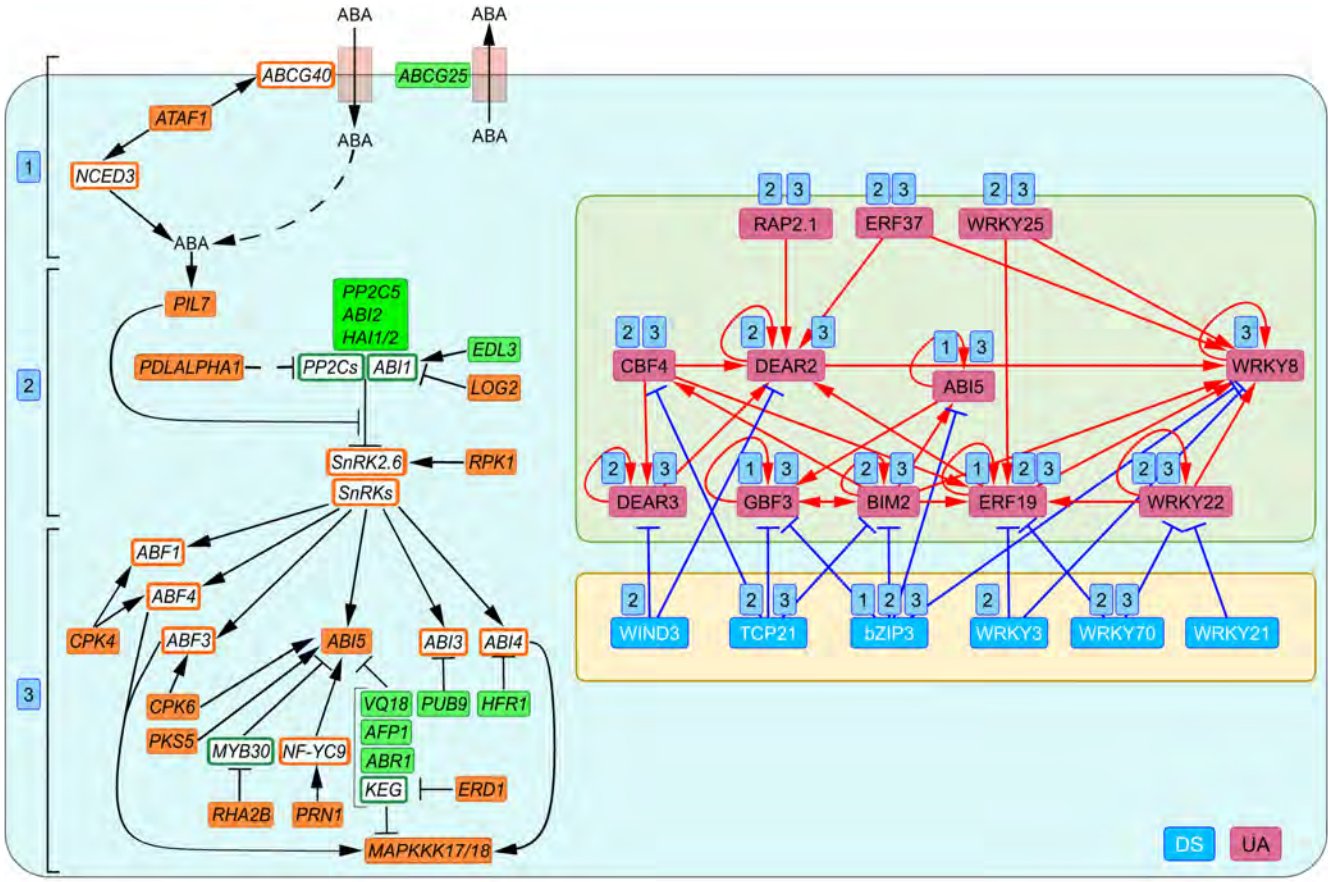


Fig. 4. Transcriptional regulation of ABA level and signaling under early salt stress.
Green and orange rectangles denote uDEGs that encode repressors and activators of the ABA level and signaling pathway, respectively, and are potential TFRN targets. White rectangles in green and orange frames correspond to repressors and activators of ABA level and signaling that are not potential TFRN targets. Numbers in blue rectangles denote the following stages: 1 – control of ABA level; 2 – ABA perception by receptors; 3 – activation of master TFs of ABA response. Abbreviations for the names of ABA transport, biosynthesis and signaling genes: *ATP-BINDING CASETTE G25/40* (*ABCG25/40*), *PYR1 LIKE 7* (*PYL7*), *PROTEIN PHOSPHATASES TYPE 2C* (*PP2Cs*), *ABA INSENSITIVE1/2/3/4/5* (*ABI1/2/3/4/5*), *SNF1-RELATED PROTEIN KINASE* (*SnRKs*), *ABSCISIC ACID RESPONSIVE ELEMENT-BINDING FACTOR1/3/4* (*ABF1/3/4*), *CALCIUM-DEPENDENT PROTEIN KINASE 4/6* (*CPK4/6*), *ABI FIVE BINDING PROTEIN 1* (*AFP1*), *KEEP ON GOING* (*KEG*), *ENHANCED DISEASE RESISTANCE 1* (*EDR1*), *NUCLEAR FACTOR Y9* (*NF-YC9*), *PLANT U-BOX/ARM-REPEAT* (*ATPUB-ARM*) *E3 LIGASE 9* (*PUB9*), *ABA REPRESSOR 1* (*ABR1*), *VQ PROTEIN 18* (*VQ18*), *HIGHLY ABA-INDUCED PP2C GENE 1/2* (*HAI1/2*), *ARABIDOPSIS THALIANA ACTIVATING FACTOR1* (*ATAF1*), *EID1-LIKE 3* (*EDL3*), *LONG HYPOCOTYL IN FAR-RED 1* (*HFR1*), *LOSS OF GDU2* (*LOG2*), *MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 17/18* (*MAPKKK17/18*), *PHOSPHOLIPASE D ALPHA 1* (*PLDALPHA1*), *PIRIN 1* (*PRN1*), *RING-H2 FINGER PROTEIN 2B* (*RHA2B*), *RECEPTOR-LIKE PROTEIN KINASE 1* (*RPK1*), *CALCINEURIN B-LIKE PROTEIN-INTERACTING PROTEIN KINASEs/SOS2-LIKE PROTEIN KINASE* (*PKS5*), *MYB DOMAIN PROTEIN 30* (*MYB30*), *NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 3* (*NCED3*).

Thus, PlantReg demonstrated that within ABA response, the targets of the salt stress response TFRN belong to the genes involved in ABA signaling, in which the most stringent control occurs at the regulation of the master TFs, ABF1/3/4 and ABI3/4/5. Moreover, ABI5, one of the master TFs in ABA signaling, is also one of the TFs within the TFRN of the salt stress response, where its activity is suppressed by bZIP3 before stress and stimulated by BIM2 during stress. ABI5 itself activates *GBF3*, which, like *BIM2*, is repressed by bZIP3 before stress. At the same time, GBF3 and BIM2 mutually activate each other. Thus, BIM2, bZIP3, GBF3, and ABI5 form a clear regulatory circuit in our reconstructed TFRN (Fig. 3*a*, 4).

Interestingly, in the ABA response gene network in (Aerts et al., 2024), the TFs that make up this regulatory loop (BIM2, bZIP3, GBF3, and ABI5) belong to the group of the earliest regulators and share a large number of common targets, i. e. control the same genes. In addition to BIM2, bZIP3, GBF3, and ABI5, our reconstructed TFRN for the salt stress response overlaps with the abscisic acid response gene network from (Aerts et al., 2024) for three other TFs: CBF4, DEAR2, and WRKY3. We identified DEAR2 as a TFRN hub. Moreover, CBF4, DEAR2, and WRKY3 are components of the network connecting its central activating regulatory circuit (BIM2, GBF3, and ABI5) to the second TFRN hub, WRKY8.

WRKY3, along with bZIP3, suppresses WRKY8 before stress (Fig. 3*a*). Under stress conditions, sequential activation of BIM2, CBF4, DEAR2, and WRKY8 occurs. Thus, comparison of the PlantReg results with the abscisic acid response gene network (Aerts et al., 2024) identified TFs that are the key regulators of ABA response. The remaining TFs, RAP2.1, ERF19/37, DEAR3, TCP21, WRKY8/22/25/70, are possibly involved in the control of ABA signaling only under salt stress.

Conclusion

The PlantReg program has shown its efficiency in systematic analysis of the results of whole-genome experiments on differential gene expression. It allows, along with functional

annotation of DEGs, identifying TF targets among them and, based on this, identifying TFs regulating certain biological processes. Combination of PlantReg results with those of programs that reconstruct TFRNs (e. g., CisCross-FindTFnet) allows subdividing a TFRN into subnetworks, which control distinct processes, to identify key TFs in these processes and even at their certain stages. The approaches and methods developed for PlantReg implementation can be successfully used to reconstruct the mechanisms of transcriptional regulation of biological processes in various species.

References

Aerts N., Hickman R., Van Dijken A.J., Kaufmann M., Snoek B.L., Pieterse C.M., Van Wees S.C. Architecture and dynamics of the abscisic acid gene regulatory network. *Plant J*. 2024;119(5):2538-2563. doi 10.1111/tpj.16899

Azameti M.K., Tanuja N., Kumar S., Rathinam M., Imoro A.W.M., Singh P.K., Gaikwad K., Sreevathsa R., Dalal M., Arora A., Rai V., Padaria J.C. Transgenic tobacco plants overexpressing a wheat salt stress root protein (TaSSRP) exhibit enhanced tolerance to heat stress. *Mol. Biol. Rep*. 2024;51(1):791. doi 10.1007/s11033-024-09755-4

Chaffai R., Ganesan M., Cherif A. Mechanisms of plant response to heat stress: recent insights. In: Plant Adaptation to Abiotic Stress: From Signaling Pathways and Microbiomes to Molecular Mechanisms. Singapore: Springer, 2024;83-105. doi 10.1007/978-981-97-0672-3_5

Chang H., Wu T., Shalmani A., Xu L., Li C., Zhang W., Pan R. Heat shock protein HvHSP16.9 from wild barley enhances tolerance to salt stress. *Physiol. Mol. Biol. Plants.* 2024;30(5):687-704. doi 10.1007/s12298-024-01455-4

Chen J.W., Shrestha L., Green G., Leier A., Marquez-Lago T.T. The hitchhikers' guide to RNA sequencing and functional analysis. *Brief. Bioinform.* 2023;24(1):bbac529. doi 10.1093/bib/bbac529

Chen L., Zhang L., Li D., Wang F., Yu D. WRKY8 transcription factor functions in the TMV-cg defense response by mediating both abscisic acid and ethylene signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA.* 2013;110(21):E1963-E1971. doi 10.1073/pnas.1221347110

Defoort J., Van de Peer Y., Vermeirssen V. Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant. *Nucleic Acids Res*. 2018;46(13):6480-6503. doi 10.1093/nar/gky468

Fidler J., Graska J., Gietler M., Nykiel M., Prabucka B., Rybarczyk-Płońska A., Muszyńska E., Morkunas I., Labudda M. PYR/PYL/RCAR receptors play a vital role in the abscisic-acid-dependent responses of plants to external or internal stimuli. *Cells*. 2022;11(8):1352. doi 10.3390/cells11081352

Ghorbani R., Alemzadeh A., Razi H. Microarray analysis of transcriptional responses to salt and drought stress in *Arabidopsis thaliana*. *Heliyon*. 2019;5(11):e02614. doi 10.1016/j.heliyon.2019.e02614

Halkier B.A., Gershenzon J. Biology and biochemistry of glucosinolates. *Annu. Rev. Plant Biol.* 2006;57(1):303-333. doi 10.1146/annurev.arplant.57.032905.105228

He X., Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet*. 2006;2(6):88. doi 10.1371/journal.pgen.0020088

Hecker D., Lauber M., Behjati Ardakani F., Ashrafiyan S., Manz Q., Kersting J., Hoffmann M., Schulz M.H., List M. Computational tools for inferring transcription factor activity. *Proteomics*. 2023;23(23-24):2200462. doi 10.1002/pmic.202200462

Hossain M.A., Henríquez-Valencia C., Gómez-Páez M., Medina J., Orellana A., Vicente-Carbajosa J., Zouhar J. Identification of novel components of the unfolded protein response in Arabidopsis. *Front. Plant Sci*. 2016;7:650. doi 10.3389/fpls.2016.00650

Hu Y., Chen L., Wang H., Zhang L., Wang F., Yu D. Arabidopsis transcription factor WRKY8 functions antagonistically with its interacting partner VQ9 to modulate salinity stress tolerance. *Plant J*. 2013;74(5):730-745. doi 10.1111/tpj.12159

Jensen M.K., Lindemose S., De Masi F., Reimer J.J., Nielsen M., Perera V., Workman C.T., Turck F., Grant M.R., Mundy J., Petersen M., Skriver K. ATAF1 transcription factor directly regulates abscisic acid biosynthetic gene *NCED3* in *Arabidopsis thaliana*. *FEBS Open Bio*. 2013;3:321-327. doi 10.1016/j.fob.2013.07.006

Jiang Y., Deyholos M.K. Functional characterization of Arabidopsis NaCl-inducible *WRKY25* and *WRKY33* transcription factors in abiotic stresses. *Plant Mol. Biol*. 2009;69(1-2):91-105. doi 10.1007/s11103-008-9408-3

Joseph B., Jini D. Salinity induced programmed cell death in plants: challenges and opportunities for salt-tolerant plants. *J. Plant Sci*. 2010;5(4):376-390. doi 10.3923/jps.2010.376.390

Kang J., Yim S., Choi H., Kim A., Lee K.P., Lopez-Molina L., Martinoia E., Lee Y. Abscisic acid transporters cooperate to control seed germination. *Nat. Commun*. 2015;6(1):8113. doi 10.1038/ncomms9113

Kesawat M.S., Satheesh N., Kherawat B.S., Kumar A., Kim H.U., Chung S.M., Kumar M. Regulation of reactive oxygen species during salt stress in plants and their crosstalk with other signaling molecules – current perspectives and future directions. *Plants*. 2023;12(4):864. doi 10.3390/plants12040864

Khan M.I.R., Syeed S., Nazar R., Anjum N.A. An insight into the role of salicylic acid and jasmonic acid in salt stress tolerance. In: Khan N., Nazar R., Iqbal N., Anjum N. (Eds). Phytohormones and Abiotic Stress Tolerance in Plants. Berlin; Heidelberg: Springer, 2012;277-300. doi 10.1007/978-3-642-25829-9_12

Ko D.K., Kim J.Y., Thibault E.A., Brandizzi F. An IRE1-proteasome system signalling cohort controls cell fate determination in unresolved proteotoxic stress of the plant endoplasmic reticulum. *Nat. Plants*. 2023;9(8):1333-1346. doi 10.1038/s41477-023-01480-3

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;49:104-111. doi 10.1093/nar/gkaa1057

Koops P., Pelser S., Ignatz M., Klose C., Marrocco-Selden K., Kretsch T. EDL3 is an F-box protein involved in the regulation of abscisic acid signalling in *Arabidopsis thaliana*. *J. Exp. Bot*. 2011;62(15):5547-5560. doi 10.1093/jxb/err236

Kumar M., Kumari P., Reddy C.R.K., Jha B. Salinity and desiccation induced oxidative stress acclimation in seaweeds. *Adv. Bot. Res.* 2014;71:91-123. doi 10.1016/B978-0-12-408062-1.00004-4

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

Li H., Jiang H., Bu Q., Zhao Q., Sun J., Xie Q., Li C. The Arabidopsis RING finger E3 ligase RHA2b acts additively with RHA2a in regulating abscisic acid signaling and drought response. *Plant Physiol*. 2011;156(2):550-563. doi 10.1104/pp.111.176214

Li J., Brader G., Kariola T., Tapio Palva E. WRKY70 modulates the selection of signaling pathways in plant defense. *Plant J*. 2006;46(3):477-491. doi 10.1111/j.1365-313X.2006.02712.x

Li J., Besseau S., Törönen P., Sipari N., Kollist H., Holm L., Palva E.T. Defense-related transcription factors WRKY70 and WRKY54 modulate osmotic stress tolerance by regulating stomatal aperture in *Arabidopsis*. *New Phytol*. 2013;200(2):457-472. doi 10.1111/nph.12378

Li P., Li X., Jiang M. CRISPR/Cas9-mediated mutagenesis of WRKY3 and WRKY4 function decreases salt and Me-JA stress tolerance in *Arabidopsis thaliana*. *Mol. Biol. Rep*. 2021;48(8):5821-5832. doi 10.1007/s11033-021-06541-4

Liu J.X., Srivastava R., Che P., Howell S.H. Salt stress responses in Arabidopsis utilize a signal transduction pathway related to endoplasmic reticulum stress signaling. *Plant J*. 2007;51(5):897-909. doi 10.1111/j.1365-313X.2007.03195.x

Liu Y., Ji X., Zheng L., Nie X., Wang Y. Microarray analysis of transcriptional responses to abscisic acid and salt stress in *Arabidopsis thaliana*. *Int. J. Mol. Sci*. 2013;14(5):9979-9998. doi 10.3390/ijms14059979

Lopez-Molina L., Mongrand S., Kinoshita N., Chua N.-H. AFP is a novel negative regulator of ABA signaling that promotes ABI5 protein degradation. *Genes Dev*. 2003;17:410-418. doi 10.1101/gad.1055803

Lu Y., Fricke W. Salt stress – regulation of root water uptake in a whole-plant and diurnal context. *Int. J. Mol. Sci*. 2023;24(9):8070. doi 10.3390/ijms24098070

Morant M., Ekstrøm C., Ulvskov P., Kristensen C., Rudemo M., Olsen C.E., Hansen J., Jørgensen K., Jørgensen B., Møller B.L., Bak S. Metabolomic, transcriptional, hormonal, and signaling cross-talk in *superroot2*. *Mol. Plant*. 2010;3(1):192-211. doi 10.1093/mp/ssp098

Ndathe R., Kato N. Phosphatidic acid produced by phospholipase Dα1 and Dδ is incorporated into the internal membranes but not involved in the gene expression of *RD29A* in the abscisic acid signaling network in *Arabidopsis thaliana*. *Front. Plant Sci*. 2024;15:1356699. doi 10.3389/fpls.2024.1356699

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

O'Malley R.C., Huang S.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;165:1280-1292. doi 10.1016/j.cell.2016.04.038

Pan J., Wang H., Hu Y., Yu D. Arabidopsis VQ18 and VQ26 proteins interact with ABI5 transcription factor to negatively modulate ABA response during seed germination. *Plant J*. 2018;95(3):529-544. doi 10.1111/tpj.13969

Pan W., Lin B., Yang X., Liu L., Xia R., Li J., Wu Y., Xie Q. The UBC27-AIRP3 ubiquitination complex modulates ABA signaling by promoting the degradation of ABI1 in Arabidopsis. *Proc. Natl. Acad. Sci. USA*. 2020;117(44):27694-27702. doi 10.1073/pnas.2007366117

Park Y., Xu Z.-Y., Kim S.Y., Lee J., Choi B., Lee J., Kim H., Sim H.-J., Hwang I. Spatial regulation of ABCG25, an ABA exporter, is an important component of the mechanism controlling cellular ABA levels. *Plant Cell*. 2016;28(10):2528-2544. doi 10.1105/tpc.16.00359

Peng Y., Zhu H., Wang Y., Kang J., Hu L., Li L., Zhu K., Yan J., Bu X., Wang X., Zhang Y., Sun X., Ahammed G.J., Jiang C., Meng S., Liu Y., Sun Z., Qi M., Li T., Wang F. Revisiting the role of light signaling in plant responses to salt stress. *Hortic. Res*. 2024;uhae262. doi 10.1093/hr/uhae262

Qi C., Zhang H., Liu Y., Wang X., Dong D., Yuan X., Li X., Zhang X., Li X., Zhang N., Guo Y.-D. CsSNAT positively regulates salt tolerance and growth of cucumber by promoting melatonin biosynthesis. *Environ. Exp. Bot*. 2020;175:104036. doi 10.1016/j.envexpbot.2020.104036

Rodriguez M.C., Mehta D., Tan M., Uhrig R.G. Quantitative proteome and PTMome analysis of *Arabidopsis thaliana* root responses to persistent osmotic and salinity stress. *Plant Cell Physiol*. 2021;62(6):1012-1029. doi 10.1093/pcp/pcab076

Sakuma Y., Liu Q., Dubouzet J.G., Abe H., Shinozaki K., Yamaguchi-Shinozaki K. DNA-binding specificity of the ERF/AP2 domain of *Arabidopsis* DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochem. Biophys. Res. Commun*. 2002;290(3);998-1009. doi 10.1006/bbrc.2001.6299

Samuel M.A., Mudgil Y., Salt J.N., Delmas F., Ramachandran S., Chilelli A., Goring D.R. Interactions between the S-domain receptor kinases and AtPUB-ARM E3 ubiquitin ligases suggest a conserved signaling pathway in Arabidopsis. *Plant Physiol*. 2008;147(4):2084-2095. doi 10.1104/pp.108.123380

Sanyal S.K., Pandey G.K. ERF111/ABR1: An AP2 domain transcription factor serving as a hub for multiple signaling pathways. *J. Plant Growth Regul*. 2024. doi 10.1007/s00344-023-11225-3

Serraj R., Roy G., Drevon J.J. Salt stress induces a decrease in the oxygen uptake of soybean nodules and in their permeability to oxygen diffusion. *Physiol. Plant*. 1994;91(2):161-168. doi 10.1111/j.1399-3054.1994.tb00414.x

Shamloo-Dashtpagerdi R., Aliakbari M., Lindlöf A., Tahmasebi S. A systems biology study unveils the association between a melatonin biosynthesis gene, O-methyl transferase 1 (OMT1) and wheat (*Triticum aestivum* L.) combined drought and salinity stress tolerance. *Planta*. 2022;255(5):99. doi 10.1007/s00425-022-03885-4

Shang Y., Yang D., Ha Y., Shin H.Y., Nam K.H. Receptor-like protein kinases RPK1 and BAK1 sequentially form complexes with the cytoplasmic kinase OST1 to regulate ABA-induced stomatal closure. *J. Exp. Bot*. 2020;71(4):1491-1502. doi 10.1093/jxb/erz489

Sharma M., Irfan M., Kumar A., Kumar P., Datta A. Recent insights into plant circadian clock response against abiotic stress. *J. Plant Growth Regul*. 2022;41(8):3530-3543. doi 10.1007/s00344-021-10531-y

Szepesi Á., Csiszár J., Gémes K., Horváth E., Horváth F., Simon M.L., Tari I. Salicylic acid improves acclimation to salt stress by stimulating abscisic aldehyde oxidase activity and abscisic acid accumulation, and increases Na^+ content in leaves without toxicity symptoms in *Solanum lycopersicum* L. *J. Plant Physiol*. 2009;166(9):914-925. doi 10.1016/j.jplph.2008.11.012

Tan S., Sha Y., Sun L., Li Z. Abiotic stress-induced leaf senescence: regulatory mechanisms and application. *Int. J. Mol. Sci*. 2023;24(15):11996. doi 10.3390/ijms241511996

Wang L.Y., Li J., Gong B., Wang R.H., Chen Y.L., Yin J., Yang C., Lin J.-T., Liu H.-Z., Yang Y., Li J., Li C., Yao N. Orosomucoid proteins limit endoplasmic reticulum stress in plants. *New Phytol*. 2023;240(3):1134-1148. doi 10.1111/nph.19200

Wang M., Xu Q., Yuan M. The unfolded protein response induced by salt stress in *Arabidopsis*. *Methods Enzymol*. 2011;489:319-328. doi 10.1016/B978-0-12-385116-1.00018-2

Wang Z., Mao Y., Liang L., Pedro G.C., Zhi L., Li P., Hu X. HFR1 antagonizes ABI4 to coordinate cytosolic redox status for seed germination under high-temperature stress. *Physiol. Plant*. 2024;176(4):e14490. doi 10.1111/ppl.14490

Warpeha K.M., Upadhyay S., Yeh J., Adamiak J., Hawkins S.I., Lapik Y.R., Anderson M.B., Kaufman L.S. The GCR1, GPA1, PRN1, NF-Y signal chain mediates both blue light and abscisic acid responses in Arabidopsis. *Plant Physiol*. 2007;143(4):1590-1600. doi 10.1104/pp.106.089904

Wawrzynska A., Christiansen K.M., Lan Y., Rodibaugh N.L., Innes R.W. Powdery mildew resistance conferred by loss of the ENHANCED DISEASE RESISTANCE1 protein kinase is suppressed by a missense mutation in *KEEP ON GOING*, a regulator of abscisic acid signaling. *Plant Physiol*. 2008;148(3):1510-1522. doi 10.1104/pp.108.127605

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*. 2021a;7(6):787-799. doi 10.1038/s41477-021-00929-7

Wu T., Hu E., Xu S., Chen M., Guo P., Dai Z., Feng T., Zhou L., Tang W., Zhan L.I., Fu X. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* (*Camb*). 2021b;2(3):100141. doi 10.1016/j.xinn.2021.100141

Yu G., Wang L., Han Y., He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi 10.1089/omi.2011.0118

Yuan K., Rashotte A.M., Wysocka-Diller J.W. ABA and GA signaling pathways interact and regulate seed germination and seedling de-

velopment under salt stress. *Acta Physiol. Plant*. 2011;33:261-271. doi 10.1007/s11738-010-0542-6

Zhang H., Liu D., Yang B., Liu W.Z., Mu B., Song H., Chen B., Li Y., Ren D., Deng H., Jiang Y.Q. Arabidopsis CPK6 positively regulates ABA signaling and drought tolerance through phosphorylating ABA-responsive element-binding factors. *J. Exp. Bot*. 2020;71(1): 188-203. doi 10.1093/jxb/erz432

Zhang L., Li Q., Shen J., Xue J., Han Y. Transcriptional regulatory networks in response to salt and drought stress in *Arabidopsis thaliana*. *J. Med. Plants Res*. 2012;6(6):950-958. doi 10.5897/JMPR11.240

Zhang L., Zhang X., Fan S. Meta-analysis of salt-related gene expression profiles identifies common signatures of salt stress responses in *Arabidopsis*. *Plant Syst. Evol*. 2017;303:757-774. doi 10.1007/s00606-017-1407-x

Zhao G., Cheng Q., Zhao Y., Wu F., Mu B., Gao J., Yang L., Yan J., Zhang H., Cui X., Chen Q., Lu F., Ao Q., Amdouni A., Jiang Y.-Q., Yang B. The abscisic acid-responsive element binding factors MAPKKK18 module regulates abscisic acid-induced leaf senescence in *Arabidopsis*. *J. Biol. Chem*. 2023;299(4):103060. doi 10.1016/j.jbc.2023.103060

Zhou M., Zhang J., Shen J., Zhou H., Zhao D., Gotor C., Romero L.C., Fu L., Li Z., Yang J., Shen W., Yuan X., Xie Y. Hydrogen sulfide-linked persulfidation of ABI4 controls ABA responses through the transactivation of MAPKKK18 in *Arabidopsis*. *Mol. Plant*. 2021; 14(6):921-936. doi 10.1016/j.molp.2021.03.007

Zhou X., Hao H., Zhang Y., Bai Y., Zhu W., Qin Y., Yuan F., Zhao F., Wang M., Hu J., Xu H., Guo A., Zhao H., Zhao Y., Cao C., Yang Y., Schumaker K.S., Guo Y., Xie C.G. SOS2-LIKE PROTEIN KINASE5, an SNF1-RELATED PROTEIN KINASE3-type protein kinase, is important for abscisic acid responses in Arabidopsis through phosphorylation of ABSCISIC ACID-INSENSITIVE5. *Plant Physiol*. 2015;168(2):659-676. doi 10.1104/pp.114.255455

Zhu S.Y., Yu X.C., Wang X.J., Zhao R., Li Y., Fan R.C., Shang Y., Du S.Y., Wang X.F., Wu F.Q., Xu Y.H., Zhang X.Y., Zhang D.P. Two calcium-dependent protein kinases, CPK4 and CPK11, regulate abscisic acid signal transduction in *Arabidopsis*. *Plant Cell*. 2007; 19(10):3019-3036. doi 10.1105/tpc.107.050666

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

Received October 17, 2024. Revised November 20, 2024. Accepted November 20, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):960-973

doi 10.18699/vjgb-24-103

Computational identification of promising genetic markers associated with molecular mechanisms of reduced rice resistance to *Rhizoctonia solani* under excess nitrogen fertilization using gene network reconstruction and analysis methods

E.A. Antropova [1, 2] ✉, A.R. Volyanskaya [1, 2], A.V. Adamovskaya [1, 2], P.S. Demenkov [1, 2, 3, 4], I.V. Yatsyk [1, 2, 4], T.V. Ivanisenko [1, 2, 3, 4], Y.L. Orlov [1, 3, 5, 6], Ch. Haoyu [7], M. Chen [7], V.A. Ivanisenko [1, 2, 3, 4]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Artificial Intelligence Research Center, Novosibirsk State University, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[5] Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow, Russia
[6] Digital Health Center, I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenovskiy University), Moscow, Russia
[7] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China
✉ nzhenia@bionet.nsc.ru

Abstract. Although nitrogen fertilizers increase rice yield, their excess can impair plant resistance to diseases, particularly sheath blight caused by *Rhizoctonia solani*. This pathogen can destroy up to 50 % of the crop, but the mechanisms underlying reduced resistance under excess nitrogen remain poorly understood. This study aims to identify potential marker genes to enhance rice resistance to *R. solani* under excess nitrogen conditions. A comprehensive bioinformatics approach was applied, including differential gene expression analysis, gene network reconstruction, biological process overrepresentation analysis, phylostratigraphic analysis, and non-coding RNA co-expression analysis. The Smart crop cognitive system, ANDSystem, the ncPlantDB database, and other bioinformatics resources were used. Analysis of the molecular genetic interaction network revealed three potential mechanisms explaining reduced resistance of rice to *R. solani* under excess nitrogen: the OsGSK2-mediated pathway, the OsMYB44-OsWRKY6-OsPR1 pathway, and the SOG1-Rad51-PR1/PR2 pathway. Potential markers for breeding were identified: 7 genes controlling rice responses to various stresses and 11 genes modulating the immune system. Special attention was given to key participants in regulatory pathways under excess nitrogen conditions. Non-coding RNA analysis revealed 30 miRNAs targeting genes of the reconstructed gene network. For two miRNAs (Osa-miR396 and Osa-miR7695), about 7,400 unique long non-coding RNAs (lncRNAs) with various co-expression indices were found. The top 50 lncRNAs with the highest co-expression index for each miRNA were highlighted, opening new perspectives for studying regulatory mechanisms of rice resistance to pathogens. The results provide a theoretical basis for experimental work on creating new rice varieties with increased pathogen resistance under excessive nitrogen nutrition. This study opens prospects for developing innovative strategies in rice breeding aimed at optimizing the balance between yield and disease resistance in modern agrotechnical conditions.

Key words: *Oryza sativa*; *Rhizoctonia solani*; plant bioinformatics; differentially expressed genes; genetic regulation; associative gene networks; Smart crop knowledge base; ANDSystem software and information system; nitrogen fertilizer; fungal response.

For citation: Antropova E.A., Volyanskaya A.R., Adamovskaya A.V., Demenkov P.S., Yatsyk I.V., Ivanisenko T.V., Orlov Y.L., Haoyu Ch., Chen M., Ivanisenko V.A. Computational identification of promising genetic markers associated with molecular mechanisms of reduced rice resistance to *Rhizoctonia solani* under excess nitrogen fertilization using gene network reconstruction and analysis methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):960-973. doi 10.18699/vjgb-24-103

Funding. The work of EAA, ARV, AVA, PSD, IVY, TVI, YLO, and VAI was supported by the Russian-Chinese grant from the Russian Science Foundation No. 23-44-00030. The work of ChH and MCh was supported by the National Natural Science Foundation of China (No. 32261133526).

Поиск перспективных генетических маркеров, ассоциированных с молекулярными механизмами снижения устойчивости риса к *Rhizoctonia solani* при избытке азотных удобрений, методом реконструкции и анализа генных сетей

Е.А. Антропова [1, 2] ✉, А.Р. Волянская [1, 2], А.В. Адамовская [1, 2], П.С. Деменков [1, 2, 3, 4], И.В. Яцык [1, 2, 4], Т.В. Иванисенко [1, 2, 3, 4], Ю.Л. Орлов [1, 3, 5, 6], Х. Чао [7], М. Чэнь [7], В.А. Иванисенко [1, 2, 3, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Исследовательский центр в сфере искусственного интеллекта Новосибирского национального исследовательского государственного университета, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[5] Аграрно-технологический институт Российского университета дружбы народов им. Патриса Лумумбы, Москва, Россия
[6] Центр цифровой медицины, Первый Московский государственный медицинский университет им. И.М. Сеченова Минздрава России (Сеченовский Университет), Москва, Россия
[7] Отдел биоинформатики, Колледж естественных наук, Чжэцзянский университет, Ханчжоу, Китай
nzhenia@bionet.nsc.ru

Аннотация. Азотные удобрения, повышающие урожайность риса, при избытке могут снижать устойчивость растений к заболеваниям, в частности к ризоктониозу, вызываемому *Rhizoctonia solani*. Этот патоген способен уничтожить до 50 % урожая, однако механизмы, лежащие в основе снижения устойчивости при избытке азота, остаются малоизученными. Данное исследование направлено на выявление потенциальных генов-маркеров для повышения устойчивости риса к *R. solani* в условиях избытка азота. Применен комплексный биоинформатический подход, включающий анализ дифференциальной экспрессии генов, реконструкцию генных сетей, анализ перепредставленности биологических процессов, филостратиграфический анализ и анализ коэкспрессии некодирующих РНК. Использованы когнитивная система Smart crop, ANDSystem, база данных ncPlantDB и другие биоинформатические ресурсы. Анализ молекулярно-генетической сети взаимодействий выявил три потенциальных механизма, объясняющих снижение устойчивости риса к *R. solani* при избытке азота: OsGSK2-опосредованный путь, путь OsMYB44-OsWRKY6-OsPR1 и путь SOG1-Rad51-PR1/PR2. Идентифицированы потенциальные маркеры для селекции: 7 генов, контролирующих ответы риса на широкий круг стрессов, и 11 генов-модуляторов иммунной системы. Особое внимание уделено ключевым участникам регуляторных путей в условиях избытка азота. Анализ некодирующих РНК выявил 30 микроРНК, мишенями которых являются гены из реконструированной генной сети. Для двух микроРНК (Osa-miR396 и Osa-miR7695) обнаружено около 7400 тыс. уникальных длинных некодирующих РНК (днРНК) с различными индексами коэкспрессии. Выделены топ-50 днРНК с наибольшим индексом коэкспрессии для каждой микроРНК, что открывает новые перспективы в изучении регуляторных механизмов устойчивости риса к патогенам. Полученные результаты создают теоретическую основу для экспериментальных работ по созданию новых сортов риса с повышенной устойчивостью к патогенам в условиях избыточного азотного питания.
Ключевые слова: *Oryza sativa*; *Rhizoctonia solani*; биоинформатика растений; дифференциально экспрессируемые гены; генетическая регуляция; ассоциативные генные сети; база знаний Smart crop; программно-информационная система ANDSystem; азотные удобрения; ответ на грибную инфекцию.

Introduction

Rice (*Oryza sativa* L.) is one of the most economically valuable crops in the world, constituting the main part of the diet for about half of the world's population. Nitrogen fertilizers are widely used in rice production in agricultural enterprises. They account for about 80–90 % of the yield increase obtained from mineral fertilizers (Kumeiko et al., 2013). However, along with the positive effect, nitrogen fertilizers reduce rice resistance to diseases. Excess nitrogen fertilization is one of the main factors contributing to the development of sheath blight disease in rice, caused by the fungus *Rhizoctonia solani* Kühn. Sheath blight causes serious damage to this crop's yield, leading to losses of up to 50 % (Senapati et al., 2022).

Plant susceptibility to pathogenic infections under excess nitrogen fertilization is caused by a complex of factors related to both rapid growth and development, as well as changes in plant defense responses. Excess nitrogen leads to a series of physiological changes that can increase plant susceptibility to pathogens. In particular, accelerated growth can cause weakening of cellular structures, including reduced cell wall strength and decreased cuticle thickness, which facilitates pathogen penetration (Hückelhoven, 2007; Rose et al., 2018). Furthermore, excessive nitrogen nutrition can cause changes in the plant microbiome and stimulate the growth of pathogenic microorganisms in the rhizosphere (Xiong et al., 2021).

At the molecular genetic level, complex regulatory networks including phytohormones, transcription factors, and non-coding RNAs play a key role in forming pathogen resistance. These components participate in complex stress response mechanisms affecting plant immune processes.

Phytohormones, such as salicylic acid, brassinosteroids, jasmonic acid, gibberellins, abscisic acid, auxins, and ethylene, have special significance in response to pathogenic infections (Yang J. et al., 2019). Notably, some of these phytohormones, particularly salicylic and abscisic acids, are also involved in nitrogen compound metabolism, regulating the expression of genes related to nitrogen exchange (Xing et al., 2023). This observation suggests that interference in phytohormone signaling pathways may serve as a mechanism through which excess nitrogen affects plant resistance to pathogens.

Non-coding RNAs (ncRNAs) represent a diverse group of RNA molecules that are not translated into proteins but perform important regulatory functions in the cell. Among them, several main types are distinguished: microRNAs (miRNA), small interfering RNAs (siRNAs), piRNAs (Piwi-interacting RNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and long non-coding RNAs (lncRNAs). Long non-coding RNAs are of particular interest as they play a significant role in gene regulation, affecting mRNA stability and translation, and participating in signaling pathways. In particular, the work of Supriya et al. (2024) shows that lncRNAs are involved in rice response to the fungus *R. solani*.

Despite their importance, lncRNAs remain the least studied among non-coding RNAs (Statello et al., 2021). This is due to their diversity, complexity of functions and mechanisms of action, as well as technical difficulties in their identification and characterization. One approach to studying the functional role of non-coding RNAs is to analyze their co-expression with protein-coding genes, as well as with other types of non-coding RNAs, the function of which has been established. The most comprehensive resource for non-coding RNA co-expression, including rice long non-coding RNAs, is the ncPlantDB database (https://bis.zju.edu.cn/ncPlantDB/).

The study of interactions between these various regulatory elements – phytohormones, transcription factors, and non-coding RNAs – in the context of nitrogen metabolism and pathogen resistance represents a promising research direction. It may lead to a deeper understanding of the mechanisms underlying nitrogen-induced plant disease susceptibility and potentially reveal new ways to enhance crop resistance under intensive nitrogen nutrition.

A widely used approach in computational systems biology for studying complex molecular genetic processes is the gene network method (Kolchanov et al., 2013). For automatic reconstruction of gene networks, the Institute of Cytology and Genetics of SB RAS has developed the ANDSystem cognitive system, which uses artificial intelligence methods to extract knowledge from databases and scientific publication texts (Ivanisenko V.A. et al., 2015, 2019). ANDSystem has been successfully applied to reconstruct associative gene networks and interpret genomic, proteomic, and metabolomic data in various fields of biomedicine and agrobiology. In particular, this software system has been used to reconstruct important molecular genetic mechanisms of various pathological processes and biological phenomena, including asthma (Bragina et al., 2014; Saik et al., 2018; Zolotareva et al., 2019), lymphedema (Saik et al., 2019), tuberculosis (Bragina et al., 2016), hepatitis C (Saik et al., 2016), coronavirus infection (Ivanisenko V.A. et al., 2022), Huntington's disease (Bragina et al., 2023), glioma (Rogachev et al., 2021), post-operative delirium (Ivanisenko V.A. et al., 2023), and others.

In the field of plant biology, ANDSystem has enabled new discoveries about the molecular mechanisms of cell wall functioning in *Arabidopsis thaliana* L. leaves in response to drought (Volyanskaya et al., 2023). Adapting ANDSystem's knowledge extraction methods to potato biology led to the creation of the specialized SOLANUM TUBEROSUM knowledge base, containing information about genetic regulation of potato metabolic pathways (Ivanisenko T.V. et al., 2018), which was used to prioritize potato genes involved in the formation of agronomically valuable traits (Demenkov et al., 2019).

The aim of this study was to conduct a comprehensive bioinformatic analysis of molecular mechanisms of rice response to *R. solani* under excess nitrogen conditions. The study included gene network reconstruction using the Smart Crop knowledge base – a specialized version of ANDSystem configured for rice biology, as well as the application of bioinformatic methods for analyzing the overrepresentation of biological processes, phylostratigraphic analysis of gene evolutionary age, and analysis of non-coding RNA co-expression.

Materials and methods

The study was conducted in several sequential stages (Fig. 1). In the first stage, based on transcriptome data analysis, genes that had been differentially expressed during *R. solani* infection were identified, as well as genes, the differential expression of which had been observed under excess nitrogen conditions. The second stage included the reconstruction of regulatory gene networks involving the identified genes. In the third stage, a structural and functional analysis of the obtained networks was conducted, including assessment of node centrality measures, analysis of biological process enrichment, and determination of gene evolutionary age. Next, analysis of network gene translation regulation by miRNAs was performed, and long non-coding RNA co-expression was investigated. The final stage was aimed at identifying potential markers of resistance to *R. solani* under excess nitrogen conditions.

Publicly available gene expression data. Publicly available transcriptomic data on *O. sativa* response to excess nitrogen fertilization, as well as to the pathogen *R. solani*, were collected from the NCBI GEO (Gene Expression Omnibus) and NCBI SRA (Sequence Read Archive) databases (https://www.ncbi.nlm.nih.gov/sra) (Table 1). For the analysis of *O. sativa* transcriptome under excess nitrogen conditions, one study containing three experiments was found. In this work, plants were treated with excess fertilizer – ammonium nitrate (NH_4NO_3) – at concentrations exceeding the normal level by 4, 16, and 64 times.

The differential expression analysis of *O. sativa* during *R. solani* infection included data from five time-series studies, containing a total of 21 experiments.

Transcriptomic data analysis. SRA Toolkit (v3.1.0) was used to extract FASTQ format files. Read quality control was performed using FastQC (v0.12.0). Filtering and removal of low-quality nucleotides was conducted using Trimmomatic (https://github.com/usadellab/Trimmomatic). A read length of 15 bp and Phred sequence quality score < Q20 were used as thresholds. Reads were mapped to the reference genome of *O. sativa* Japonica Group (IRGSP-1.0), deposited from the EnsemblPlants database (https://plants.ensembl.org/index.html) using the HISAT2 (v2.2.1) tool. SAMtools (v1.20) was used to convert SAM format mapping output files to binary BAM format. HTSeq (v2.0.2) was used for quantification. Read count normalization and differential gene expression analysis were performed using the edgeR (4.0.16) tool implemented in the Bioconductor project (https://www.bioconductor.org/). The TMM (Trimmed Mean of M-values) method was used for normalization. Multiple testing correction was applied using FDR (false discovery rate).

For DNA microarray data analysis, the limma (v3.58.1) package from the Bioconductor project was used. Raw Agilent platform DNA microarray files were read using read.images. Background noise correction and quantile normalization of the data were then performed. The biomaRt (v2.58.2) package (https://bioconductor.org/packages/release/bioc/html/biomaRt.html) was used to map DNA microarray probe identifiers to Ensembl gene identifiers. Differential gene expression analysis was performed using the limma package. An FDR threshold of < 0.05 was used to identify differentially expressed genes.

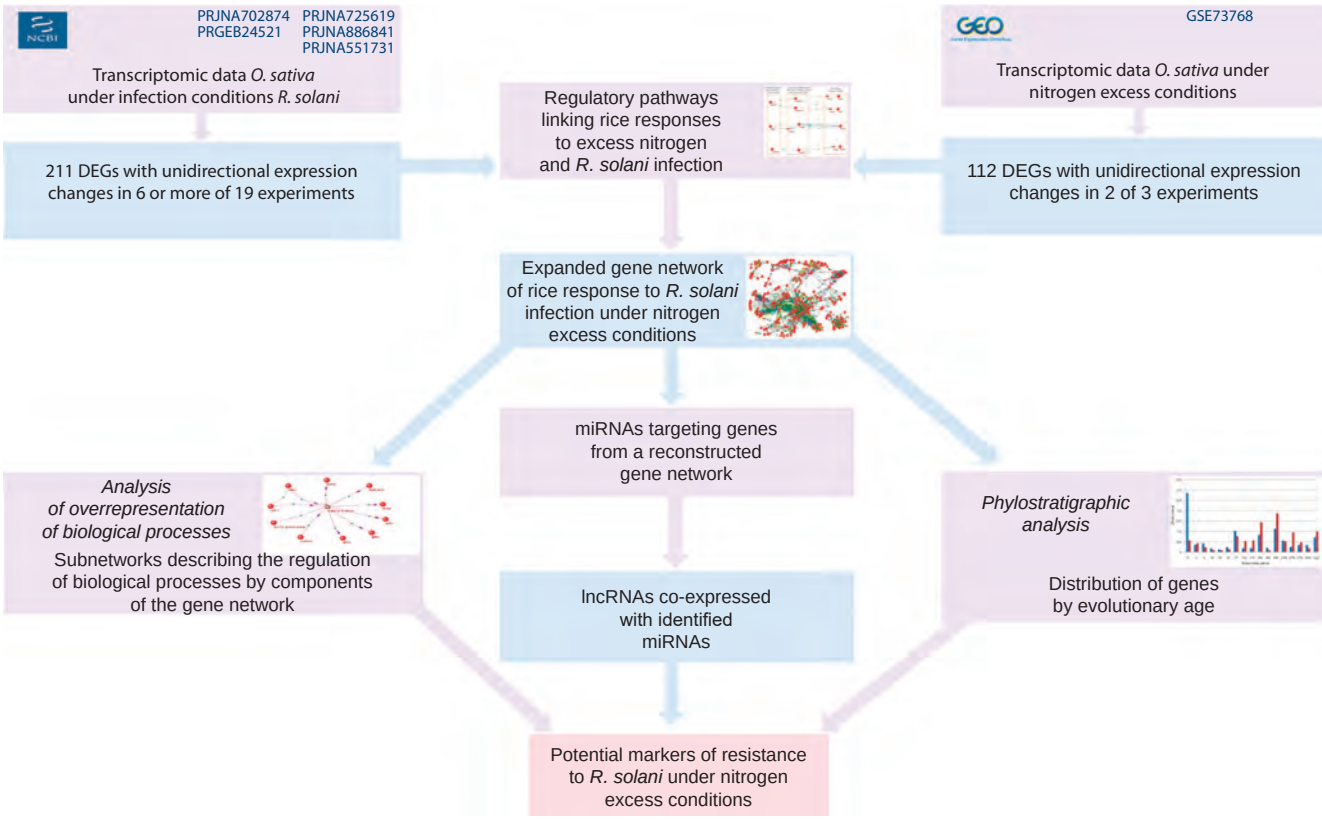


Fig. 1. Research stage diagram.

Table 1. List of publicly available RNA-seq and DNA microarray data used in the study

Stress	Design	Project ID	Subfamily	Genotype	Organ	Data type
Excess nitrogen (NH_4NO_3)	3 experiments: 3 concentrations (×64, ×16, and ×4 normal concentration	GSE73768	ssp. *japonica*	cv. Nipponbare	Shoot	Microarray
R. solani infection	3 experiments: 3 time points (1, 2 and 5 days post-infection)	PRJNA725619	ssp. *indica*	var. BPT- 5204	Leaves	RNA-seq
R. solani infection	4 experiments: 2 time points for 2 varieties (1 and 2 days post-infection)	PRJNA886841	ssp. *japonica*	var. Lemont, var. GD66	Leaves	RNA-seq
R. solani infection	2 experiments: 1 time point for 2 varieties (3 days post-infection)	PRJNA551731	ssp. *japonica*	cv. Yanhui-888, cv. Jingang-30	Leaves	RNA-seq
R. solani infection	6 experiments: 3 time points for 2 varieties (1, 3 and 5 days post-infection)	PRJEB24521	ssp. *japonica*	var. Cocodrie, line MCR	Leaves	RNA-seq
R. solani infection	6 experiments: 3 time points for 2 varieties (1, 2 and 3 days post-infection)	PRJNA702874	ssp. *indica*	Line PAU-ShB8, line PR114	Leaves	RNA-seq

Smart Crop knowledge base. This study used the specialized Smart Crop knowledge base, which is an adapted version of the ANDSystem software and information system, focused on rice and wheat genetics and breeding. System adaptation included configuring three key ANDSystem modules for effective task solving. The first module was the domain-specific ontology module, which was expanded with special dictionaries. These dictionaries covered a wide range of research objects that can be divided into molecular genetic objects (genes, proteins, metabolites, non-coding RNAs, and miRNAs),

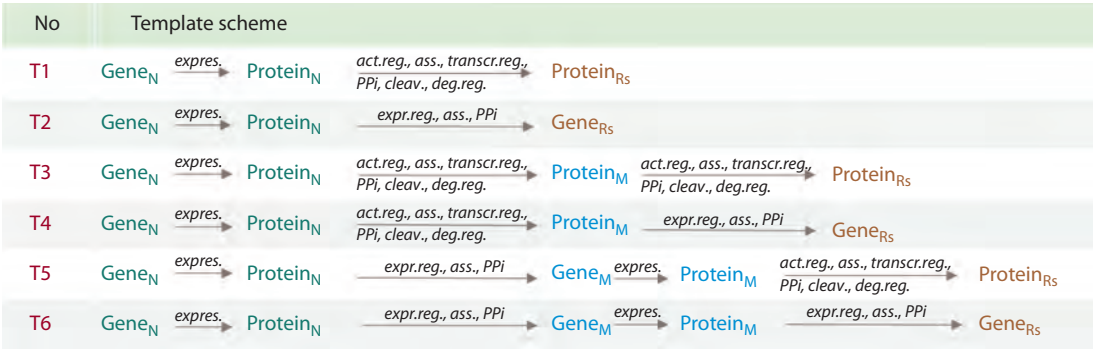


Fig. 2. Template scheme used for searching for molecular genetic pathways in the Smart Crop knowledge base.

Notation: T – template; Gene$_N$ – DEGs of rice under excess nitrogen fertilization; Protein$_N$ – protein products of DEGs under excess nitrogen fertilization; Gene$_M$ – genes encoding mediator proteins; Protein$_M$ – mediator proteins; Gene$_{Rs}$ – rice DEGs in response to *R. solani*; Protein$_{Rs}$ – protein products of rice DEGs in response to *R. solani*; *expres*. – expression; *act.reg.* – regulation of activity; *expr.reg*. – regulation of expression; *ass*. – association; *transcr.reg.* – regulation of transcription; *deg.reg.* – regulation of degradation; *cleav*. – cleavage; *PPi* – protein-protein interaction.

their functional characteristics (biological processes, genetic biomarkers, QTL polymorphisms), phenotypic characteristics (plant varieties, breeding-significant qualities, phenotypic traits, diseases), biotic and abiotic factors (pathogens, pests, and others). Various databases and ontologies were used to form these dictionaries, such as NCBI Gene (https://www.ncbi.nlm.nih.gov/gene/), ChEBI (https://www.ebi.ac.uk/chebi/), MirBase (https://www.mirbase.org/), Gene Ontology (https://cropontology.org/), Wheat Ontology, Rice Ontology, and others (Chao et al., 2023). For example, the gene dictionary from the molecular genetic objects group contains names of approximately 627 thousand genes, including their conventional names and synonyms. Biological processes, belonging to functional characteristics, contain more than 122 thousand names. The pathogen dictionary, included in the biotic factors group, contains about 755 names.

The second important component was the information extraction module from factographic databases, which was configured for automated data extraction from specialized sources in plant biology. These sources included Oryzabase (https://shigen.nig.ac.jp/rice/oryzabase/), GrainGenes (https://wheat.pw.usda.gov/GG3/), ASPNet, and others. The third module was the text-mining module using semantic-linguistic templates and artificial intelligence methods. It was adapted for effective knowledge extraction from text sources, such as scientific articles and patents in plant biology. Based on the analysis of scientific publications performed using this module, more than 4 million interactions between objects represented in the dictionaries were extracted.

Gene network reconstruction and analysis. Gene network reconstruction was performed using the "Query Wizard" and "Pathway Wizard" of the ANDVisio software module (Demenkov et al., 2011), which serves as the user interface in the ANDSystem and Smart Crop systems. The structure of templates used for searching for regulatory pathways in the Smart Crop knowledge base using the "Pathway Wizard" is shown in Figure 2.

Node centrality assessment in the gene network. Node centrality in the gene network was evaluated using the network connectivity measure, defined as the number of connections between a given node and other network nodes.

Biological process enrichment analysis. Gene Ontology biological process enrichment analysis was performed using the PANTHER resource (https://pantherdb.org/).

Long non-coding RNA analysis. Co-expression analysis between miRNAs and lncRNAs was conducted using the ncPlantDB database (https://bis.zju.edu.cn/ncPlantDB/).

Phylostratigraphic analysis. The evolutionary age of genes was determined using the GenOrigin database (http://chenzxlab.hzau.edu.cn/) (Tong et al., 2021), which contains information about the evolutionary age of genes from various organisms, established through phylostratigraphic analysis. To assess the statistical significance of differences in the distribution of genes of different ages between the complete set of rice protein-coding genes and genes in the reconstructed network, a hypergeometric test was applied. The probability of observing m or more genes of a certain age interval among M network genes was calculated using the hypergeom.pmf function from the scipy library. The analysis was conducted for 17 age intervals represented in the GenOrigin database. The following parameters were used in calculations: N – total number of rice protein-coding genes, n – number of rice genes in a given age interval, M – number of genes in the gene network, m – number of network genes in the analyzed age interval. Differences were considered statistically significant at p-value < 0.05.

Results and discussion

Identification of stable differentially expressed genes

To identify differentially expressed genes (DEGs) in rice under excess nitrogen conditions, 3 experiments were analyzed, while under *R. solani* fungus influence, 21 experiments were analyzed using transcriptomic data found in open sources. We considered genes with unidirectional expression changes across different experiments (simultaneous decrease or increase), which we will further refer to as stable DEGs.

In the case of excess nitrogen, only 5 genes were found to be stable DEGs across all three experiments (*Os09g0538000*, *Os05g0162000*, *Os09g0537700*, *Os04g0664900*, *Os06g0113800*). When considering DEGs present in two

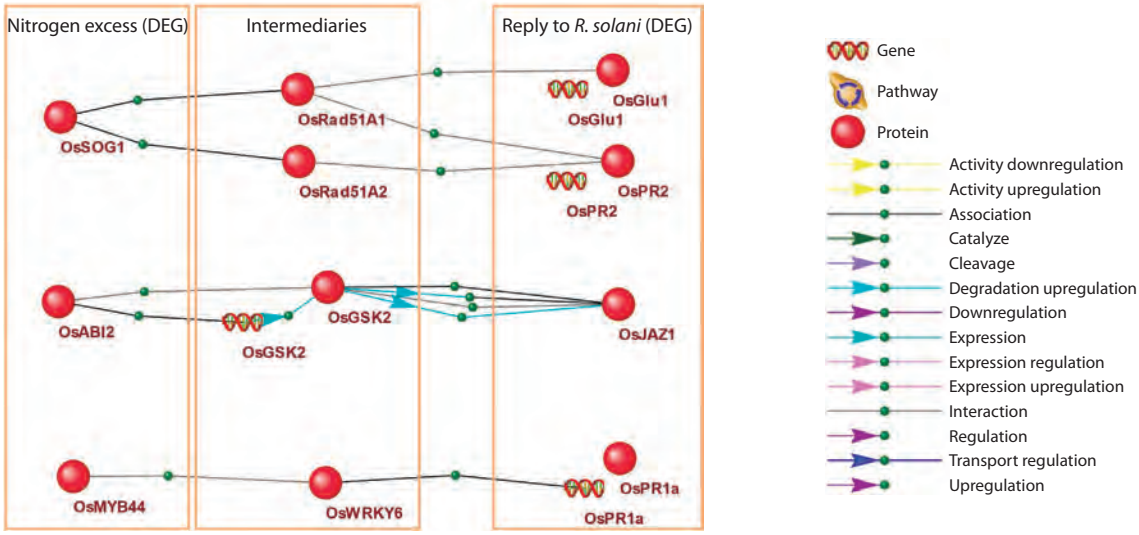


Fig. 3. Regulatory pathways describing the connection between DEGs in rice response to excess nitrogen and *R. solani* infection.

out of three experiments, the number of such genes was 112, which were taken for further analysis.

Analysis of differential gene expression under *R. solani* infection showed that in two out of 21 experiments, no statistically significant DEGs were identified. Analysis of the remaining 19 experiments revealed no genes that were DEGs in every experiment. Only 2 genes were found to be stable DEGs in half or more of the experiments (*Os04g0180500* and *Os09g0255600*). When considering one-third of the experiments (6 or more out of 19), the number of stable DEGs included 211 genes. The number of stable DEGs for a quarter of the experiments (5 or more out of 19) was 463 genes. For further analysis, we chose a threshold value for determining stable DEGs equal to one-third of the experiments (6 or more out of 19), as at this value, the samples of stable DEGs under excess nitrogen and fungal influence were comparable in size.

Reconstruction of molecular genetic pathways describing the relationship between rice responses to excess nitrogen and infection

Using the ANDVisio program, which serves as the user interface for the Smart Crop and ANDSystem knowledge bases, a search was conducted for molecular genetic pathways in the global Smart Crop gene network (Fig. 2), connecting the group of the selected 112 stable DEGs in response to excess nitrogen and 211 stable DEGs in response to *R. solani* fungus. This search resulted in the identification of several regulatory pathways that included 3 proteins encoded by DEGs in response to excess nitrogen, 4 DEGs and their encoded proteins in response to *R. solani* infection, as well as 4 proteins acting as mediators in interactions between the considered DEGs (Fig. 3).

OsABI2-OsGSK2-OsJAZ1 molecular genetic pathway

An important reconstructed pathway (Fig. 3) potentially explaining the mechanism of deteriorated rice resistance to fungus under excess nitrogen is the OsABI2-OsGSK2-OsJAZ1 pathway. The OsABI2 protein (PP2C06, protein phosphatase 2C6) is a product of the *Os01g0583100* gene that is differentially expressed under excess nitrogen: its expression decreases at 16- and 64-fold excess of nitrogen fertilizer concentration (Supplementary Material 1)[1].

It is known that ABI2 is one of the main participants in the ABA (abscisic acid) signaling pathway (Sun et al., 2011), which is an important plant hormone necessary for regulating stomatal closure, leaf senescence, bud dormancy, seed germination inhibition, growth inhibition, and stress responses to drought, salinity, and toxic metals (Chen et al., 2020; Kumar S. et al., 2022). Literature has shown that OsABI2 participates in rice response to excess iron (Junior et al., 2015), in sunflower, its expression increases during drought (Shen et al., 2023), and in rice, during drought, its expression is also noted in roots and stem (Sircar et al., 2022). The presence of this protein in the reconstructed regulatory pathway may indicate its involvement in modulating rice response to the pathogen under excess nitrogen. OsABI2 can exert regulatory influence on OsJAZ1 (jasmonate-Zim-domain protein 1), an important factor in pathogen response, through the mediator OsGSK2.

According to our analysis, OsJAZ1 (*Os10g0392400*) is a DEG with increased expression levels in 7 out of 19 experiments studying *R. solani* influence on rice transcriptome (Supplementary Material 2). In Arabidopsis and cotton, it was shown that the fungus *Verticillium dahliae*, which causes Verticillium wilt, induces JAZ1 phosphorylation through GSK2, and this promotes further JAZ1 degradation (Song Y. et al., 2021). The authors note that in this action, GSK2 is a negative regulator of fungal resistance – its constitutive expression weakened resistance, while *GSK2* gene knockdown increased resistance to *V. dahliae*. Interestingly, *OsGSK2* (*Os05g0207500*) is a DEG in 2 out of 19 analyzed experiments studying *R. solani* influence on transcriptome, where its expression was decreased (Supplementary Material 2). Also, *OsGSK2* is a DEG in response to excess nitrogen in the experiment with the highest nitrogen fertilizer concentration (64 times higher than normal concentration).

[1] Supplementary Materials 1–5 are available at: https://vavilov.elpub.ru/jour/Suppl_Antropova_Engl_28_8.xlsx

In our network, the connection between ABI2 and GSK2 is of the "interaction" type (physical interaction). In Arabidopsis, it was shown that ABI1 and ABI2 interact with the GSK2 protein (Glycogen synthase kinase 2, also known as: brassinosteroid insensitive 2, BIN2) and dephosphorylate it, leading to suppression of its kinase activity and decreased stability. The examined interactions between regulatory pathway participants are consistent with literature data showing that the abscisic acid signaling pathway suppresses the brassinosteroid signaling pathway (Wang H., 2018). In particular, in *O. sativa*, it was demonstrated that ABA acts oppositely to BR (brassinosteroids) in regulating leaf inclination through the BR biosynthesis gene *OsD11* and signaling genes *OsGSK2* and *OsDLT* (Li et al., 2019).

It should be noted that BR represents an important group of plant hormones, in some cases playing an antagonistic role to ABA action. For example, it was shown that BR stimulates seed germination, while ABA promotes their dormancy (Steber, McCourt, 2001).

MYB44-WRKY6-PR1 molecular genetic pathway

Another important regulatory pathway begins with the OsMYB44 protein – a product of the *Os09g0106700* gene that is differentially expressed under excess nitrogen. Notably, it is a DEG in two out of three experiments (gene expression is decreased at 16- and 64-fold excess of nitrogen fertilizer concentration, Supplementary Material 1). The transcription factor MYB44 is known to be an important participant in plant life regulation (root development, somatic embryogenesis, leaf senescence, etc.) and response to biotic and abiotic stresses, such as reactions to drought, cold, phosphate and nitrogen deficiency, and pathogenic organism infection (Wang F. et al., 2023). Interestingly, MYB44 has opposing effects on plant defense reactions. Shim et al. (2013) showed that it enhanced the defensive response to pathogenic bacteria *Pseudomonas syringae* pv. *tomato* induced by salicylic acid but reduced the defensive response against the black spot disease fungus *Alternaria brassicicola*, which is dependent on jasmonic acid. In the pathway under consideration, MYB44 forms a regulatory complex with another TF, WRKY6 (*Os03g0798500*), which regulates inorganic phosphate transport, as shown in potato (Zhou et al., 2017). The transcription factor WRKY6, like MYB44 in *A. thaliana*, acts as a positive regulator of abscisic acid signaling. The WRKY TF family participates in protecting plants from a wide range of stresses, in particular, OsWRKY6 is necessary for rice protection from *Xanthomonas oryzae* pv. *oryzae* (bacterial leaf blight) (Im et al., 2024). It has been shown that OsWRKY6 activates *OsPR1* expression (Im et al., 2022), the final link in the regulatory pathway under consideration.

SOG1-Rad51-PR1/PR2 molecular genetic pathway

This pathway includes three links: SOG1 (suppressor of gamma response1), RAD51 (DNA repair protein RAD51), and the *PR1* (pathogenesis-related protein 1) and *PR2* (pathogenesis-related protein 2) genes (Fig. 3). SOG1 is a plant transcription factor, analogous to the animal p53 protein, playing a crucial role in regulating transcription of genes involved in programmed cell death, DNA damage repair, as well as responses to abiotic stresses and pathogenic infections (Ogita et al., 2018; Yoshiyama, Kimura, 2018). According to our transcriptional data analysis, *SOG1* (*Os06g0267500*) is a DEG under excess nitrogen (expression level increases in two out of three experiments – at 16- and 64-fold excess of nitrogen fertilizer concentration, Supplementary Material 1).

SOG1 is known to be a transcriptional regulator of *OsRad51* (Ogita et al., 2018; Yoshiyama, Kimura, 2018), acting as a mediator in the pathway under consideration. RAD51 is a regulatory protein of plant immune response, and among its direct targets are members of the pathogenesis-related protein family, such as *PR1* and *PR2* (Wang S. et al., 2010). These genes were among the DEGs in response to *R. solani* fungus (Supplementary Material 2).

PR1 (*Os07g0129200*) expression increased in 6 out of 19 experiments studying *R. solani* influence on transcriptome. Seven genes named *PR2* have been found in the rice genome (Yokotani et al., 2014). According to our data, expression of three of them (*Os07g0539900*, *Os01g0940700*, and *Os01g0940800*) increased in 7 out of 19 experiments.

It should be noted that the *PR1* and *PR2* genes were also among the DEGs based on our analysis of transcriptomic data from a series of experiments studying excess nitrogen. Their expression changed significantly in one out of three experiments, where the concentration of nitrogen fertilizers was maximal.

Reconstruction of extended gene network of rice response to *R. solani* infection under excess nitrogen

To identify a broader range of potential participants in the mechanisms of deteriorating rice resistance to *R. solani* fungus under excess nitrogen, we reconstructed an extended gene network based on the regulatory pathways discussed above. Gene network reconstruction was performed automatically using the functional module of the ANDVisio program. This tool allows expanding the initial network by adding new components (genes, proteins, metabolites, etc.) based on data about their interactions contained in the Smart Crop knowledge base. For 15 participants of the initial regulatory pathways (Fig. 3), the knowledge base contained information about their interactions with 358 new proteins and genes. The network reconstructed in this way contained 61 genes, 271 proteins, and 2,359 interactions (Fig. 4). To identify key participants in the reconstructed network, node centrality analysis was conducted using the "Network connectivity" index, indicating the number of nearest neighbors. The highest index value belonged to the OsGSK2 protein, which is a participant in the initial regulatory pathways, mediating interactions between differentially expressed genes. Jaz1 was also among the top three in terms of the "Network connectivity" index. It should be noted that the gene encoding Jaz1 was a stable DEG in response to *R. solani* fungus.

Identification of lncRNAs potentially regulating the identified molecular genetic pathways

To search for lncRNAs potentially involved in regulating the rice gene network response to fungus under excess nitrogen conditions, we analyzed the ncPlantDB database. This database contains information about lncRNA co-expression with miRNAs, obtained from single-cell data analysis.

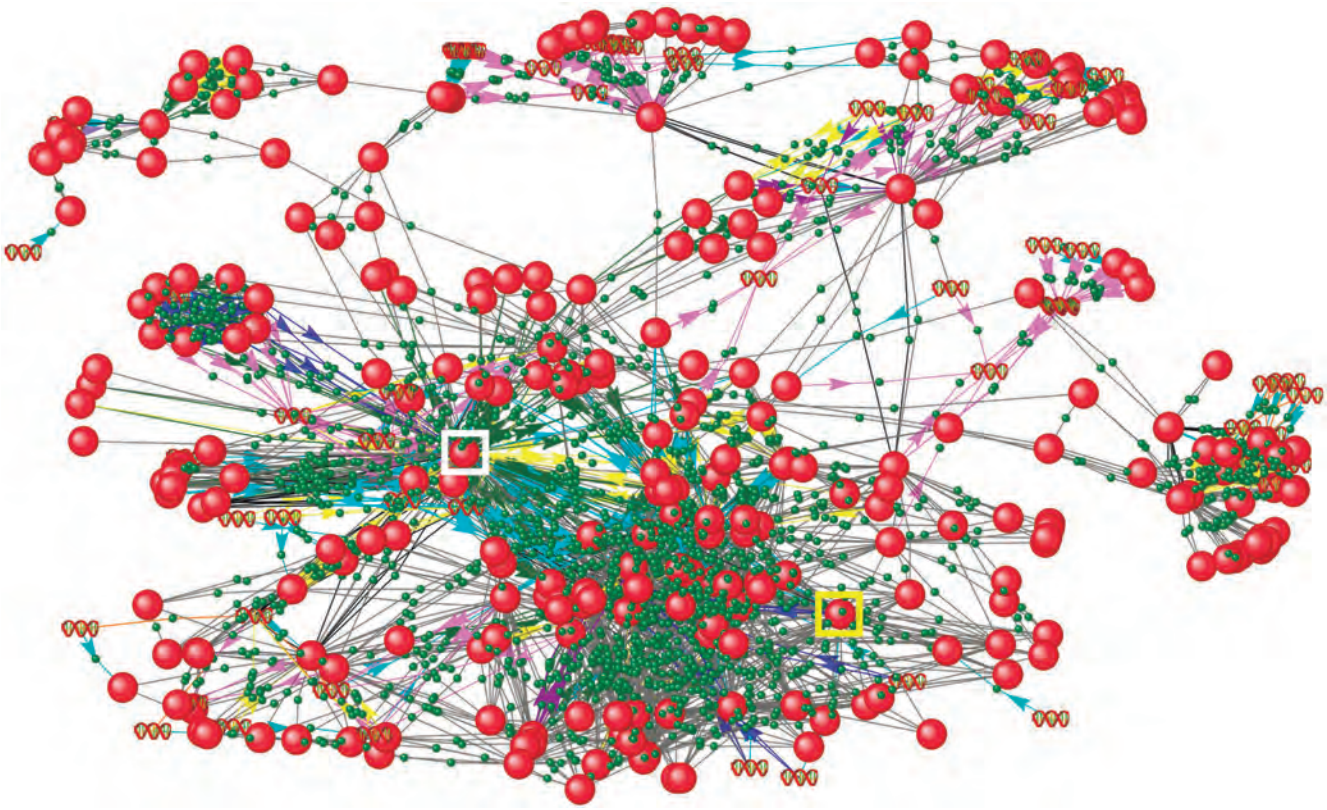


Fig. 4. Extended gene network of rice response to *R. solani* infection under excess nitrogen conditions.

The network includes both initial regulatory pathways and newly identified components (genes and proteins). The JAZ1 and GSK2 proteins are highlighted with yellow and white squares, respectively. Gene and protein designations and their interaction types are similar to those shown in Figure 3.

According to the Smart Crop knowledge base, we found 30 miRNAs that target genes from the reconstructed gene network (Table 2). In the ncPlantDB database, co-expression connections were found for Osa-miR396 and Osa-miR7695 with lncRNAs, with various co-expression degree indices. For two variants of Osa-miR396 (Osa-miR396b and Osa-miR396c), the number of such non-coding RNAs was around 4,000. For Osa-miR7695, about 3,500 co-expression connections with lncRNAs were identified. The total number of unique lncRNAs was approximately 7,400.

Among the identified lncRNAs, special attention should be paid to those with the highest co-expression index. These

Table 2. miRNAs regulating stress response genes in the reconstructed gene network

No.	miRNA	Target gene	Reference
1–3	Osa-miR156	*OsMPKs*, *OsSPL14*	Xie et al., 2006; Kumar K. et al., 2022; Song L. et al., 2021
4	Osa-miR159	*OsGAMYB*	Kumar K. et al., 2022
5	Osa-miR162	*OsDCL1*	Kumar K. et al., 2022
6–8	Osa-miR166	*EIN 2*	Song L. et al., 2021; Kumar K. et al., 2022
9	Osa-miR167	*ARF12*	Kumar K. et al., 2022
10	Osa-miR319	*OsTCP21*	Song L. et al., 2021; Kumar K. et al., 2022
11–12	Osa-miR393	*AFB2/TIR*	Song L. et al., 2021
13–21	Osa-miR396	*OsGRFs*	Song L. et al., 2021
22	Osa-miR398	*SOD*, *CSD1*, *CSD2*	Song L. et al., 2021; Kumar K. et al., 2022
23	Osa-miR408	*OsAAE3*	Charagh et al., 2024
24–29	Osa-miR444	*MADS23/27a/57*	Kumar K. et al., 2022; Song L. et al., 2021
30	Osa-miR7695	*OsNramp6*	Kumar K. et al., 2022; Song L. et al., 2021

Note. miRNAs of the same family are grouped together.

include the top 50 lncRNAs ranked by co-expression index, particularly the group of lncRNAs identified in rice metaxylem that have the same co-expression index with Osa-miR396b, the target genes of which are *GFR1* and *GFR3*: LNC-Os08g15450, LNC-Os04g61735, LNC-Os05g27975, LNC-Os05g62500, and others (Supplementary Material 3).

The search for functions of these lncRNAs in literature data yielded no results. Therefore, the connection of lncRNAs with the gene network may have special significance for further characterization of their functions.

Phylostratigraphic analysis

The application of phylostratigraphic analysis methods to assess the evolutionary age of genes is a promising approach to studying the evolution patterns of gene networks and identifying their key components (Mustafin et al., 2021). In this work, this approach was used to analyze the evolutionary stages at which genes participating in the reconstructed network of response to fungal infection under elevated nitrogen fertilizer concentrations emerged.

Analysis of the evolutionary age distribution of genes showed that the reconstructed network contains genes of different ages, among which several most represented groups can be distinguished (Fig. 5). Age intervals within which the number of genes statistically significantly exceeded the one expected by chance corresponded to the following time points shown in the graph (Fig. 5): (1) 132 million years ($p = 1.85 \cdot 10^{-3}$), (2) 170 million years ($p = 9.16 \cdot 10^{-4}$), and (3) 1,578 million years ($p = 5.41 \cdot 10^{-7}$).

The first group, including 11 genes about 132 million years old, likely emerged at the evolutionary stage of monocot plant appearance (Friis et al., 2004). Representatives of this group include the transcription factor OFP3 (ovate family protein 3). The OFP family is plant-specific, participating in regulation of cellular pluripotency, morphogenesis, and growth in *A. thaliana* (Wang F. et al., 2016). Moreover, it is suggested that changes in transcription factor regulatory networks are an essential feature of monocot plant evolution (Vincentz et al., 2004).

Within the second interval under consideration (170 million years), the age of 12 genes was found. This period is associated with the emergence of flowering plants (van der Kooi, Ollerton, 2020). Members of the WRKY transcription factor family (WRKY6, 40, and 46), involved in molecular mechanisms of flowering regulation (Song H. et al., 2024), fell into this interval. Importantly, WRKY6 is also a participant in the initial regulatory pathways.

The third group included 20 genes, the age of which fell within the third interval (1,578 million years), corresponding to the emergence of red and green algae (Zhang S. et al., 2021). One representative of this group is the *PHT1* (*PHOSPHATE TRANSPORTER1*) gene, the product of which participates in inorganic phosphate uptake and transport (Wang X. et al., 2014). The development of phosphorus assimilation mechanisms could have been significant in plant evolution, as increased phosphate availability in oceans is associated with the growth of larger eukaryotic organisms (Zhang S. et al., 2021).

Another feature of the gene network can be noted: the proportion of "young" genes (less than 1 million years old) was lower than their proportion in the complete genome. The "young" genes falling into this interval include 12 genes, many of which are related to immune responses to varying degrees: *OsPR5* (*OS01G0122000*), *OsNAC6* (*Os01g0672100*), *similar to histone H4* (*OS01G0835900*), *OsMPK3* (*OS02G0148100*), *R2R3-MYB* (*OS02G0641300*), *R2R3-MYB* (*OS06G0205100*), *OsPR1b* (*OS07G0127700*), *histone H4* (*OS07G0549900*), *R2R3MYB-domain protein* (*OS12G0564100*).

The obtained data can contribute to a deeper understanding of the reconstructed gene network functioning mechanisms and serve as a basis for further selection of markers in breeding plants resistant to pathogens under elevated nitrogen fertilizer concentrations.

Search for potential marker-oriented selection targets

To search for potential marker-oriented selection targets, analysis of gene functional significance at the biological process level was conducted. Using the PANTHER resource,

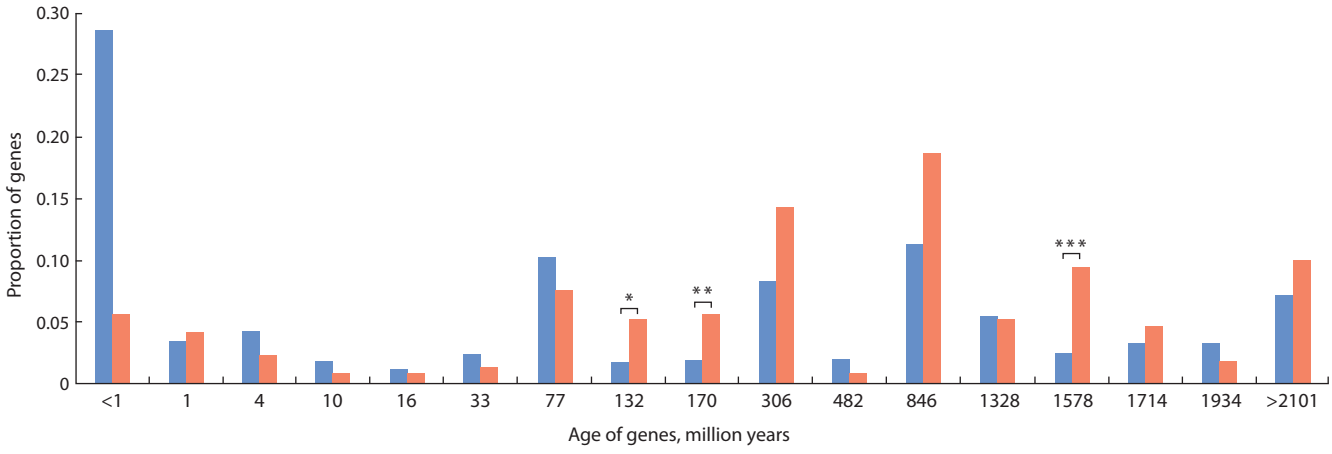


Fig. 5. Distribution of gene evolutionary age in the reconstructed gene network.

The *X*-axis shows the central points of age intervals (million years) according to the GenOrigin database, the *Y*-axis shows the proportion of genes in each age interval. Blue shows the distribution for the complete set of rice protein-coding genes, red shows the distribution for genes in the reconstructed network. Asterisks mark age intervals with statistically significant differences in gene representation: * $p = 1.85 \cdot 10^{-3}$, ** $p = 9.16 \cdot 10^{-4}$, *** $p = 5.41 \cdot 10^{-7}$, hypergeometric test.

Table 3. Results of biological process enrichment analysis for genes in the extended network of rice response to *R. solani* infection under excess nitrogen conditions

Term from Gene Ontology	*p*-value	FDR	Number of genes
Response to hormone	1.09E-38	2.87E-36	44
Hormone-mediated signaling pathway	2.08E-37	4.75E-35	38
Response to chemical	2.15E-30	3.38E-28	47
Response to stress	1.41E-22	1.69E-20	50
Regulation of defense response	4.12E-16	4.41E-14	13
Seed germination	5.17E-10	3.76E-08	6
Response to water deprivation	4.47E-09	2.74E-07	8
Defense response	3.36E-07	1.51E-05	18
Cellular response to abiotic stimulus	3.70E-06	1.46E-04	5
Defense response to fungus	1.81E-03	3.96E-02	4

Note. Analysis was performed using the PANTHER resource. The most significant biological processes related to response to various biotic and abiotic factors are presented.

Gene Ontology term enrichment analysis was performed for the extended gene network. The analysis revealed 239 statistically significant biological processes (Supplementary Material 4), including key signaling pathways and responses to abiotic and biotic stresses, including fungal infections (Table 3).

Although the biological process enrichment analysis provides important information about the functional significance of the gene network, the understanding of specific regulatory mechanisms is necessary for selecting effective markers. The Smart Crop knowledge base contains information about regulatory interactions between genes and biological processes, which allows identifying potential markers not only by their association with key processes but also by their regulatory potential.

To search for potential markers, the gene network was supplemented with regulatory connections to biological processes using ANDVisio (Supplementary Material 5). Regulatory connections between genes and processes were classified as positive (upregulation), negative (downregulation), or without direction (regulation). Figure 6 shows regulatory networks for the processes “response to stress” and “innate immune system”, which play key roles in stress response mechanisms.

It should be noted that “response to stress” was found to be overrepresented among genes in the extended network of rice response to *R. solani* infection under excess nitrogen conditions (Table 3). Three proteins are important regulators of this process (Fig. 6*a*): BZR1 (brassinazole resistant 1), serine-threonine protein kinase SAPK4 (shown in Fig. 6*a* as Ser/Thr protein kinase), and transcription factor SOG1 (shown in Fig. 6*a*, *b* as OsSOG1). BZR1 is known to mediate brassinosteroid signaling by suppressing the transcription of stress response genes (Yang Y.X. et al., 2015; Cao et al., 2024). SAPK4 regulates gene expression in response to salt stress in rice (Diédhiou et al., 2008). SOG1 controls plant response to DNA damage-inducing stresses (Ogita et al., 2018; Yoshiyama et al., 2018). SOG1 is a component of the initial regulatory pathways, which allows it to be classified as a particularly important potential marker. All the considered

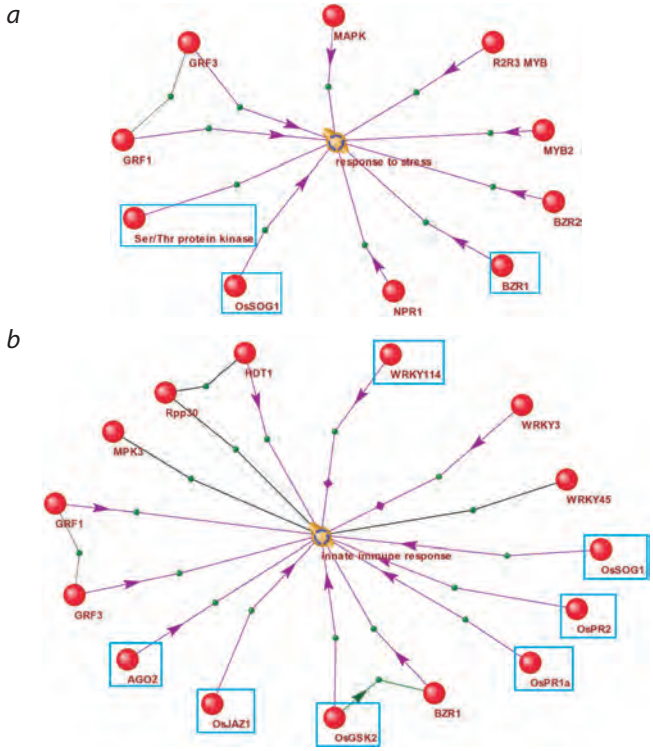


Fig. 6. Regulation of biological processes “response to stress” (*a*) and “innate immune response” (*b*) by proteins that are components of the rice gene network response to pathogenic fungus under excess nitrogen conditions.

Connections between objects marked with black lines indicate association; purple arrows indicate regulatory effects. Blue rectangles highlight proteins discussed in the text.

proteins can be classified as markers controlling responses to a wide spectrum of stress effects. This characteristic makes them especially valuable for further research and potential application in plant biotechnology.

The “innate immune system” process is interesting because it is regulated by thirteen gene network participants that can

be considered as promising markers associated with pathogen resistance (Fig. 6*b*). Key regulators of this process are proteins WRKT114 and AGO2, as well as components of the molecular genetic pathways described above (GSK, PR1, PR2, JAZ1, and SOG1). WRKT114 activates immune response during *Xanthomonas oryzae* pv. *oryzae* infection (Son et al., 2020). AGO2 regulates innate immunity through miRNA-mediated suppression of target genes during *Pseudomonas syringae* pv. *tomato* infection (Zhang X. et al., 2011). The remaining components also make significant contributions to plant immune response regulation (Song Y. et al., 2021; Johnson et al., 2023; Javed et al., 2024).

Characterization of marker genes by evolutionary age

Assessment of gene evolutionary age can provide important information for planning breeding programs, allowing prediction of specificity, functional conservation, and phenotypic effects of candidate genes. The application of gene evolutionary age analysis in experiment planning is illustrated by work on the introgression of the rice *Xa21* gene. This gene provides resistance to rice bacterial blight caused by *X. oryzae* pv. *oryzae*. *Xa21* was isolated from the wild species *Oryza longistaminata* and is an evolutionarily young gene specific to the *Oryza* genus. Introduction of the *Xa21* gene into cultivated rice varieties led to the creation of lines with high disease resistance without negative effects on yield and grain quality (Song W.Y. et al., 1995; Wang G.L. et al., 1996).

Another example is the modification of the *ERF922* gene to increase rice resistance to fungal pathogens using CRISPR/Cas9. *ERF922* is an evolutionarily young gene involved in regulating rice immune response. Its knockout led to increased resistance to rice blast without negative effects on plant growth (Wang F. et al., 2016).

Our phylostratigraphic analysis of the gene network revealed that the average evolutionary age of potential marker genes in the "innate immune response" group is 605 million years, which is significantly less than the corresponding value for the "response to stress" group (1,270 million years). These data confirm the understanding of the evolutionary youth of immune mechanisms (Han, 2019). In the "innate immune response" group, the age range extends from *OsPR1a* (less than 1 million years) to *OsGSK2* (more than 2,101 million years), while in the "response to stress" group, from *OsSOG1* (306 million years) to Ser/Thr protein kinase (1,714 million years).

It is known that genes with a greater evolutionary age participate in the functioning of more fundamental processes (Wolf et al., 2009; Domazet-Lošo, Tautz, 2010). Variations in these genes can affect multiple phenotypic traits, which may complicate selection for target properties. In this regard, evolutionarily young network genes appear most promising for marker-oriented selection: *OsPR5*, *OsNAC6*, *OsMPK3*, *R2R3-MYB*, *OsPR1b*, and *histone H4*.

Conclusion

In this work, a systems approach incorporating a wide range of bioinformatic methods was applied to search for potential marker genes aimed at increasing rice resistance to *R. solani* under excess nitrogen conditions. Methods implemented in the Smart Crop cognitive system, ANDSystem, and other well-known bioinformatic resources were used. The systems analysis, implemented as a data processing pipeline, included: (1) investigation of differential gene expression; (2) reconstruction and analysis of gene networks; (3) analysis of biological process enrichment; (4) analysis of gene network evolution using phylostratigraphic analysis; (5) analysis of omics data on non-coding RNA co-expression.

Analysis of the molecular genetic interaction network connecting rice responses to excess nitrogen and *R. solani* infection allowed us to propose mechanisms explaining the deterioration of rice resistance to fungus under elevated nitrogen fertilizer concentrations. Three potential pathways were identified: (1) the OsGSK2-mediated pathway: OsGSK2 may be a participant in the pathway linking plant responses to excess nitrogen and *R. solani* fungus. At elevated levels, it can worsen plant resistance to fungus, as shown with *Verticillium dahliae* affecting Arabidopsis and cotton. According to our data, active OsGSK2 levels may be elevated under excess nitrogen due to decreased expression of its inhibitor (OsABI2); (2) the OsMYB44-OsWRKY6-OsPR1 pathway: all participants in this pathway are related to plant protection from biotic stresses; (3) the SOG1-Rad51-PR1/PR2 pathway: from transcription factor SOG1 through immune response gene transcription regulator Rad51 to the *PR1* and *PR2* genes, essential participants in pathogen response.

Reconstruction of the extended gene network allowed identification of potential markers for breeding aimed at increasing resistance to pathogens (such as *R. solani*) under excess nitrogen conditions. The found markers are divided into two groups: markers controlling rice responses to a wide range of stresses (7 genes) and markers modulating the immune system (11 genes).

Among the most important markers are genes that are key participants in regulatory pathways underlying the rice gene network response to *R. solani* pathogen under excess nitrogen conditions (*OsGSK2*, *JAZ1*, *PR1*/*PR2*, *SOG1*).

The obtained theoretical results can serve as a foundation for further experimental work on creating new rice varieties with increased pathogen resistance under excess nitrogen fertilizer conditions. The conducted research opens prospects for developing innovative strategies in rice breeding aimed at optimizing the balance between yield and disease resistance in modern agrotechnical conditions.

References

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

Bragina E.Y., Gomboeva D.E., Saik O.V., Ivanisenko V.A., Freidin M.B., Nazarenko M.S., Puzyrev V.P. Apoptosis genes as a key to identification of inverse comorbidity of Huntington's disease and cancer. *Int. J. Mol. Sci*. 2023;24(11):9385. doi 10.3390/ijms24119385

Cao X., Wei Y., Shen B., Liu L., Mao J. Interaction of the transcription factors BES1/BZR1 in plant crowth and stress response. *Int. J. Mol. Sci*. 2024;25(13):6836. doi 10.3390/ijms25136836

Chao H., Zhang S., Hu Y., Ni Q., Xin S., Zhao L., Ivanisenko V.A., Orlov Y.L., Chen M. Integrating omics databases for enhanced crop breeding. *J. Integr. Bioinform*. 2023;20(4):20230012. doi 10.1515/jib-2023-0012

Charagh S., Hui S., Wang J., Raza A., Zhou L., Xu B., Zhang Y., Sheng Z., Tang S., Hu S., Hu P. Unveiling innovative approaches to mitigate metals/metalloids toxicity for sustainable agriculture. *Physiol. Plant*. 2024;176(2):e14226. doi 10.1111/ppl.14226

Chen K., Li G.J., Bressan R.A., Song C.P., Zhu J.K., Zhao Y. Abscisic acid dynamics, signaling, and functions in plants. *J. Integr. Plant Biol*. 2020;62(1):25-54. doi 10.1111/jipb.12899

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*. 2011-2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Demenkov P.S., Saik O.V., Ivanisenko T.V., Kolchanov N.A., Kochetov A.V., Ivanisenko V.A. Prioritization of potato genes involved in the formation of agronomically valuable traits using the SOLANUM TUBEROSUM knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2019;23(3):312-319. doi 10.18699/VJ19.501

Diédhiou C.J., Popova O.V., Dietz K.J., Golldack D. The SNF1-type serine-threonine protein kinase SAPK4 regulates stress-responsive gene expression in rice. *BMC Plant Biol*. 2008;8:49. doi 10.1186/1471-2229-8-49

Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*. 2010;468(7325):815-818. doi 10.1038/nature09632

Friis E.M., Pedersen K.R., Crane P.R. Araceae from the Early Cretaceous of Portugal: evidence on the emergence of monocotyledons. *Proc. Natl. Acad. Sci. USA*. 2004;101(47):16565-16570. doi 10.1073/pnas.0407174101

Han G.Z. Origin and evolution of the plant immune system. *New Phytol*. 2019;222(1):70-83. doi 10.1111/nph.15596

Hückelhoven R. Cell wall-associated mechanisms of disease resistance and susceptibility. *Annu. Rev. Phytopathol*. 2007;45(1):101-127. doi 10.1146/annurev.phyto.45.062806.094325

Im J.H., Choi C., Park S.R., Hwang D.J. The OsWRKY6 transcriptional cascade functions in basal defense and Xa1-mediated defense of rice against *Xanthomonas oryzae* pv. *oryzae*. *Planta*. 2022;255(2):47. doi 10.1007/s00425-022-03830-5

Im J.H., Choi C., Jung M.Y., Park S.R., Hwang D.J. The *OsICS1* is directly regulated by OsWRKY6 and increases resistance against *Xanthomonas oryzae* pv. *oryzae*. *Planta*. 2024;259(6):124. doi 10.1007/s00425-024-04405-2

Ivanisenko T.V., Saik O.V., Demenkov P.S., Khlestkin V.K., Khlestkina E.K., Kolchanov N.A., Ivanisenko V.A. The SOLANUM TUBEROSUM knowledge base: the section on molecular-genetic regulation of metabolic pathways. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018;22(1):8-17. doi 10.18699/VJ18.325 (in Russian)

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., 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., Karpenko T.N., Velichko 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., 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 Journal of Genetics and Breeding*. 2023;27(7):768-775. doi 10.18699/VJGB-23-89

Javed T., Wang W., Yang B., Shen L., Sun T., Gao S.J., Zhang S. Pathogenesis related-1 proteins in plant defense: regulation and functional diversity. *Crit. Rev. Biotechnol*. 2024;1-9. doi 10.1080/07388551.2024.2344583

Johnson L.Y.D., Major I.T., Chen Y., Yang C., Vanegas-Cano L.J., Howe G.A. Diversification of JAZ-MYC signaling function in immune metabolism. *New Phytol*. 2023;239(6):2277-2291. doi 10.1111/nph.19114

Junior A.T.D., Farias D.D., dos Santos R.S., do Amaral M.N., Arge L.W.P., Oliveira D.D.C., Silveira S.F.S., Sousa R.O., Braga E.J.B., Maia L.C., Oliveira A.C. The quest for more tolerant rice: How high concentrations of iron affect alternative splicing? *Transcriptomics*. 2015;3:2. doi 10.4172/2329-8936.1000122

Kolchanov N.A., Ignatyeva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850 (in Russian)

Kumar K., Mandal S.N., Neelam K., de Los Reyes B.G. MicroRNA-mediated host defense mechanisms against pathogens and herbivores in rice: balancing gains from genetic resistance with trade-offs to productivity potential. *BMC Plant Biol*. 2022;22(1):351. doi 10.1186/s12870-022-03723-5

Kumar S., Shah S.H., Vimala Y., Jatav H.S., Ahmad P., Chen Y., Siddique K.H.M. Abscisic acid: metabolism, transport, crosstalk with other plant growth regulators, and its role in heavy metal stress mitigation. *Front. Plant Sci*. 2022;13:972856. doi 10.3389/fpls.2022.972856

Kumeiko Yu.V., Paraschenko V.N., Kremzin N.M. Application of nitrification inhibitor to reduce nitrogen losses and increase the efficiency of nitrogen fertilizers in rice growing. *Sbornik Nauchnykh Trudov Stavropolskogo NII Zhivotnovodstva i Kormoproizvodstva = Proceedings of the Stavropol Research Institute of Animal Husbandry and Forage Production*. 2013;3(6):144-147 (in Russian)

Li Q.F., Lu J., Zhou Y., Wu F., Tong H.N., Wang J.D., Yu J.W., Zhang C.Q., Fan X.L., Liu Q.Q. Abscisic acid represses rice lamina joint inclination by antagonizing brassinosteroid biosynthesis and signaling. *Int. J. Mol. Sci*. 2019;20(19):4908. doi 10.3390/ijms20194908

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Ogita N., Okushima Y., Tokizawa M., Yamamoto Y.Y., Tanaka M., Seki M., Makita Y., Matsui M., Okamoto-Yoshiyama K., Sakamoto T., Kurata T., Hiruma K., Saijo Y., Takahashi N., Umeda M. Identifying the target genes of SUPPRESSOR OF GAMMA RESPONSE 1, a master transcription factor controlling DNA damage response in *Arabidopsis*. *Plant J*. 2018;94(3):439-453. doi 10.1111/tpj.13866

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. doi 10.3390/metabo11030133

Rose J.K.C., Catalá C., Gonzalez-Carranza Z.H., Roberts J.A. Cell wall disassembly. In: Annual Plant Reviews online. Vol. 8. The Plant Cell Wall. Wiley, 2018;264-324. doi 10.1002/9781119312994.apr0075

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., 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

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

Senapati M., Tiwari A., Sharma N., Chandra P., Bashyal B.M., Ellur R.K., Bhowmick P.K., Bollinedi H., Vinod K.K., Singh A.K., Krishnan S.G. *Rhizoctonia solani* Kühn pathophysiology: status and prospects of sheath blight disease management in rice. *Front. Plant Sci*. 2022;13:881116. doi 10.3389/fpls.2022.881116

Shen J., Wang X., Song H., Wang M., Niu T., Lei H., Qin C., Liu A. Physiology and transcriptomics highlight the underlying mechanism of sunflower responses to drought stress and rehydration. *iScience*. 2023;26(11):108112. doi 10.1016/j.isci.2023.108112

Shim J.S., Jung C., Lee S., Min K., Lee Y.W., Choi Y., Lee J.S., Song J.T., Kim J.K., Choi Y.D. AtMYB44 regulates *WRKY70* expression and modulates antagonistic interaction between salicylic acid and jasmonic acid signaling. *Plant J*. 2013;73(3):483-495. doi 10.1111/tpj.12051

Sircar S., Musaddi M., Parekh N. NetREx: Network-based Rice Expression Analysis Server for abiotic stress conditions. *Database* (*Oxford*). 2022;2022:baac060. doi 10.1093/database/baac060

Son S., An H.K., Seol Y.J., Park S.R., Im J.H. Rice transcription factor WRKY114 directly regulates the expression of *OsPR1a* and *Chitinase* to enhance resistance against *Xanthomonas oryzae* pv. *oryzae*. *Biochem. Biophys. Res. Commun*. 2020;533(4):1262-1268. doi 10.1016/j.bbrc.2020.09.141

Song H., Duan Z., Zhang J. WRKY transcription factors modulate flowering time and response to environmental changes. *Plant Physiol. Biochem*. 2024;210:108630. doi 10.1016/j.plaphy.2024.108630

Song L., Fang Y., Chen L., Wang J., Chen X. Role of non-coding RNAs in plant immunity. *Plant Commun*. 2021;2(3):100180. doi 10.1016/j.xplc.2021.100180

Song W.Y., Wang G.L., Chen L.L., Kim H.S., Pi L.Y., Holsten T., Gardner J., Wang B., Zhai W.X., Zhu L.H., Fauquet C., Ronald P. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science*. 1995;270(5243):1804-1806. doi 10.1126/science.270.5243.1804

Song Y., Zhai Y., Li L., Yang Z., Ge X., Yang Z., Zhang C., Li F., Ren M. BIN2 negatively regulates plant defence against *Verticillium dahliae* in *Arabidopsis* and cotton. *Plant Biotechnol. J*. 2021;19(10):2097-2112. doi 10.1111/pbi.13640.

Statello L., Guo C.J., Chen L.L., Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol*. 2021;22(2):96-118. doi 10.1038/s41580-020-00315-9

Steber C.M., McCourt P. A role for brassinosteroids in germination in *Arabidopsis*. *Plant Physiol*. 2001;125(2):763-769. doi 10.1104/pp.125.2.763.

Sun H.L., Wang X.J., Ding W.H., Zhu S.Y., Zhao R., Zhang Y.X., Xin Q., Wang X.F., Zhang D.P. Identification of an important site for function of the type 2C protein phosphatase ABI2 in abscisic acid signalling in *Arabidopsis*. *J. Exp. Bot*. 2011;62(15):5713-5725. doi 10.1093/jxb/err274

Supriya P., Srividya G.K., Solanki M., Manvitha D., Prakasam V., Balakrishnan M., Neeraja C.N., Srinivasa Rao Ch, Sundaram R.M., Mangrauthia S.K. Identification and expression analysis of long non-coding RNAs of rice induced during interaction with *Rhizoctonia solani*. *Physiol. Mol. Plant Pathol*. 2024;134:102389. doi 10.1016/j.pmpp.2024.102389

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

van der Kooi C.J., Ollerton J. The origins of flowering plants and pollinators. *Science*. 2020;368(6497):1306-1308. doi 10.1126/science.aay3662

Vincentz M., Cara F.A., Okura V.K., da Silva F.R., Pedrosa G.L., Hemerly A.S., Capella A.N., Marins M., Ferreira P.C., França S.C., Grivet L., Vettore A.L., Kemper E.L., Burnquist W.L., Targon M.L., Siqueira W.J., Kuramae E.E., Marino C.L., Camargo L.E., Carrer H., Coutinho L.L., Furlan L.R., Lemos M.V., Nunes L.R., Gomes S.L., Santelli R.V., Goldman M.H., Bacci M. Jr, Giglioti E.A., Thiemann O.H., Silva F.H., Van Sluys M.A., Nobrega F.G., Arruda P., Menck C.F. Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol*. 2004;134(3):951-959. doi 10.1104/pp.103.033878

Volyanskaya A.R., Antropova E.A., Zubairova U.S., Demenkov P.S., Venzel A.S., Orlov Y.L., Makarova A.A., Ivanisenko T.V., Gorshkova T.A., Aglyamova A.R., Kolchanov N.A., Chen M., Ivanisenko V.A. Reconstruction and analysis of the gene regulatory network for cell wall function in *Arabidopsis thaliana* L. leaves in response to water deficit. *Vavilov J. Genet. Breed*. 2023;27(8):1031-1041. doi 10.18699/VJGB-23-118

Wang F., Wang C., Liu P., Lei C., Hao W., Gao Y., Liu Y.G., Zhao K. Enhanced rice blast resistance by CRISPR/Cas9-targeted mutagenesis of the ERF transcription factor gene *OsERF922*. *PLoS One*. 2016;11(4):e0154027. doi 10.1371/journal.pone.0154027

Wang F., Yang F., Zhu D., Saniboere B., Zhou B., Peng D. MYB44 plays key roles in regulating plant responses to abiotic and biotic stress, metabolism, and development. *J. Plant Biochem. Biotechnol*. 2024;33(4):462-473. doi 10.1007/s13562-023-00864-y

Wang G.L., Song W.Y., Ruan D.L., Sideris S., Ronald P.C. The cloned gene, *Xa21*, confers resistance to multiple *Xanthomonas oryzae* pv. *oryzae* isolates in transgenic plants. *Mol. Plant-Microbe Interact*. 1996;9(9):850-855. doi 10.1094/mpmi-9-0850

Wang H., Tang J., Liu J., Hu J., Liu J., Chen Y., Cai Z., Wang X. Abscisic acid signaling inhibits brassinosteroid signaling through dampening the dephosphorylation of BIN2 by ABI1 and ABI2. *Mol. Plant*. 2018;11:315-325. doi 10.1016/j.molp.2017.12.013

Wang S., Durrant W.E., Song J., Spivey N.W., Dong X. *Arabidopsis* BRCA2 and RAD51 proteins are specifically involved in defense gene transcription during plant immune responses. *Proc. Natl. Acad. Sci. USA*. 2010;107(52):22716-22721. doi 10.1073/pnas.1005978107

Wang X., Wang Y., Piñeros M.A., Wang Z., Wang W., Li C., Wu Z., Kochian L.V., Wu P. Phosphate transporters OsPHT1;9 and OsPHT1;10 are involved in phosphate uptake in rice. *Plant Cell Environ*. 2014;37(5):1159-1170. doi 10.1111/pce.12224

Wolf Y.I., Novichkov P.S., Karev G.P., Koonin E.V., Lipman D.J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci. USA*. 2009;106(18):7273-7280. doi10.1073/pnas.0901808106

Xie K., Wu C., Xiong L. Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice. *Plant Physiol*. 2006;142(1):280-293. doi 10.1104/pp.106.084475

Xing J., Cao X., Zhang M., Wei X., Zhang J., Wan X. Plant nitrogen availability and crosstalk with phytohormones signallings and their biotechnology breeding application in crops. *Plant Biotechnol. J*. 2023;21(7):1320-1342. doi 10.1111/pbi.13971

Xiong Q., Hu J., Wei H., Zhang H., Zhu J. Relationship between plant roots, rhizosphere microorganisms, and nitrogen and its special focus on rice. *Agriculture*. 2021;11(3):234. doi 10.3390/agriculture 11030234

Yang J., Duan G., Li C., Liu L., Han G., Zhang Y., Wang C. The crosstalks between jasmonic acid and other plant hormone signaling highlight the involvement of jasmonic acid as a core component in plant response to biotic and abiotic stresses. *Front. Plant Sci*. 2019; 10:1349. doi 10.3389/fpls.2019.01349

Yang Y.X., Ahammed G.J., Wu C., Fan S.Y., Zhou Y.H. Crosstalk among jasmonate, salicylate and ethylene signaling pathways in plant disease and immune responses. *Curr. Protein Pept. Sci*. 2015; 16(5):450-461. doi 10.2174/1389203716666150330141638

Yokotani N., Tsuchida-Mayama T., Ichikawa H., Mitsuda N., Ohme-Takagi M., Kaku H., Minami E., Nishizawa Y. OsNAC111, a blast disease-responsive transcription factor in rice, positively regulates the expression of defense-related genes. *Mol. Plant-Microbe Interact*. 2014;27(10):1027-1034. doi 10.1094/MPMI-03-14-0065-R

Yoshiyama K.O., Kimura S. Ser-Gln sites of SOG1 are rapidly hyperphosphorylated in response to DNA double-strand breaks. *Plant Signal. Behav*. 2018;13(6):e1477904. doi 10.1080/15592324.2018. 1477904

Zhang S., Su J., Ma S., Wang H., Wang X., He K., Wang H., Canfield D.E. Eukaryotic red and green algae populated the tropical ocean 1400 million years ago. *Precambrian Res*. 2021;357:106166. doi 10.1016/j.precamres.2021.106166

Zhang X., Zhao H., Gao S., Wang W.C., Katiyar-Agarwal S., Huang H.D., Raikhel N., Jin H. *Arabidopsis* Argonaute 2 regulates innate immunity via miRNA393(*)-mediated silencing of a Golgi-localized SNARE gene *MEMB12*. *Mol. Cell*. 2011;42(3):356-366. doi 10.1016/j.molcel.2011.04.010

Zhou X., Zha M., Huang J., Li L., Imran M., Zhang C. StMYB44 negatively regulates phosphate transport by suppressing expression of *PHOSPHATE1* in potato. *J. Exp. Bot.* 2017;68(5):1265-1281. doi 10.1093/jxb/erx026

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 September 27, 2024. Revised November 14, 2024. Accepted November 14, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):974-981

doi 10.18699/vjgb-24-104

Reconstruction of gene regulatory networks from single cell transcriptomic data

M.A. Rybakov[1, 2], N.A. Omelyanchuk [1], 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. Gene regulatory networks (GRNs) – interpretable graph models of gene expression regulation – are a pivotal tool for understanding and investigating the mechanisms utilized by cells during development and in response to various internal and external stimuli. Historically, the first approach for the GRN reconstruction was based on the analysis of published data (including those summarized in databases). Currently, the primary GRN inference approach is the analysis of omics (mainly transcriptomic) data; a number of mathematical methods have been adapted for that. Obtaining omics data for individual cells has made it possible to conduct large-scale molecular genetic studies with an extremely high resolution. In particular, it has become possible to reconstruct GRNs for individual cell types and for various cell states. However, technical and biological features of single-cell omics data require specific approaches for GRN inference. This review describes the approaches and programs that are used to reconstruct GRNs from single-cell RNA sequencing (scRNA-seq) data. We consider the advantages of using scRNA-seq data compared to bulk RNA-seq, as well as challenges in GRN inference. We pay specific attention to state-of-the-art methods for GRN reconstruction from single-cell transcriptomes recruiting other omics data, primarily transcription factor binding sites and open chromatin profiles (scATAC-seq), in order to increase inference accuracy. The review also considers the applicability of GRNs reconstructed from single-cell omics data to recover and characterize various biological processes. Future perspectives in this area are discussed.

Key words: gene regulatory network; single-cell data; RNA sequencing; scRNA-seq; scATAC-seq.

For citation: Rybakov M.A., Omelyanchuk N.A., Zemlyanskaya E.V. Reconstruction of gene regulatory networks from single cell transcriptomic data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024; 28(8):974-981. doi 10.18699/vjgb-24-104

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

Методы реконструкции генных регуляторных сетей на основе транскриптомных данных отдельных клеток

М.А. Рыбаков[1, 2], Н.А. Омельянчук [1], 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

Аннотация. Генные регуляторные сети – интерпретируемые графовые модели регуляции экспрессии генов – являются важным инструментом для понимания и исследования механизмов, которые клетки реализуют в процессе развития и при ответе на различные внутренние и внешние стимулы. Исторически первый подход для реконструкции генных регуляторных сетей основывался на анализе литературных сведений, в том числе обобщенных в базах данных. В настоящее время основной способ системной реконструкции генных регуляторных сетей – анализ омиксных (в первую очередь транскриптомных) данных; разработан ряд математических подходов для решения этой задачи. Развитие технологий получения омиксных данных для отдельных клеток сделало возможным проведение широкомасштабных молекулярно-генетических исследований с беспрецедентно высоким уровнем разрешения. В частности, появилась возможность реконструировать генные регуляторные сети для отдельных клеточных типов и для различных стадий развития клеток. Однако технические и биологические особенности омиксных данных отдельных клеток требуют специальных программ для решения этой задачи. В обзоре описаны подходы и программы, которые разработаны и используются для построения генных регуляторных сетей по транскриптомным данным отдельных клеток (scRNA-seq). Разбираются преимущества применения транскриптомных данных для отдельных клеток по сравнению с транскриптомами многоклеточных образцов,

а также их недостатки в рамках решения задачи реконструкции регуляторных генных сетей. Существенное внимание уделяется повышению точности генных регуляторных сетей, построенных по транскриптомным данным отдельных клеток с помощью привлечения других омиксных данных, в первую очередь данных по сайтам связывания транскрипционных факторов и профилирования районов открытого хроматина (scATAC-seq). Рассматриваются вопросы применимости получаемых сетей в молекулярно-генетических исследованиях, приводятся примеры успешного использования генных регуляторных сетей, реконструированных различными методами с применением омиксных данных отдельных клеток для решения конкретных биологических задач. Обсуждаются перспективные направления развития этой области.
Ключевые слова: регуляторная генная сеть; данные для отдельных клеток; секвенирование РНК; scRNA-seq; scATAC-seq.

Introduction

A gene network is a group of coordinately expressed genes that interact with each other through the RNAs and proteins they encode, as well as the products of protein activity (Kolchanov et al., 2013). Gene networks are a central object of systems biology. To explore specific aspects more deeply, specialized types of gene networks are distinguished. Among them, gene regulatory networks (GRNs) hold a special place, as they describe the regulation of gene expression by transcription factors (TFs) – a key mechanism for a flexible implementation of genetic information (Huynh-Thu, Sanguinetti, 2019). GRNs are visualized as graphs of interactions between TFs and the genes they regulate (Fig. 1*a*) (Badia-i-Mompel et al., 2023). Each node in a GRN represents a gene (some of which encode TFs), while edges correspond to regulatory relationships between TF-encoding genes and other genes (these relationships may reflect true molecular interactions between TFs and promoters of their target genes or merely their statistical correlation). An edge may have a sign indicating whether it describes activation or inhibition of transcription, and a weight reflecting the strength of the regulator's influence. Thus, GRNs represent models of the logic of regulatory events between genes during execution of cellular programs (Tieri, Castiglione, 2021). They provide a viable alternative to classical modeling with differential equations when kinetic information is unavailable.

GRNs can be constructed based on information about TFs and their target genes from publications or inferred *de novo* from transcriptomic data (Badia-i-Mompel et al., 2023). Bulk RNA-seq results in expression levels for each gene aggregated across all cells in a tissue or organ sample. Bulk RNA-seq data can be presented as a so-called expression matrix, which provides the expression values for each gene (depicted in lines) across different samples (depicted in columns) (Fig. 1*b*). Given that gene expression levels in these matrices result from regulation mediated by TF binding to gene promoters, a mathematical model can be constructed to explain the observed gene expression levels (Mercatelli et al., 2020; Nguyen et al., 2021). Most GRN inference methods designed for transcriptomic data are based on this premise (Mercatelli et al., 2020). Currently, GRN reconstruction from RNA-seq data is one of the topics in systems biology, within which a large number of methods and software programs have been developed (Nguyen et al., 2019; Mercatelli et al., 2020).

At the same time, the approach described above has drawbacks. First, transcriptomic data do not contain explicit information about specific regulatory events (e. g., TF binding to the promoters of the genes they regulate); all TF-target links are mathematically inferred from gene expression levels. As a result, non-existent (erroneous) connections may be reconstructed. Incorporating data that directly describe transcriptional regulation (e. g., genome-wide open chromatin profiles or TF binding sites) can significantly improve GRN accuracy (Sönmezer et al., 2020; Isbel et al., 2022). Second, RNA-seq data do not account for the heterogeneity of cell populations, whereas gene expression can vary dramatically among different cell types. This issue is addressed by scRNA-seq (Tang et al., 2009).

Single cell transcriptomic data represent an expression matrix where lines correspond to genes and columns correspond to cells (Fig. 1*c*), which can be grouped by cell types using special approaches (Luecken, Theis, 2019). scRNA-seq opens up opportunities to investigate biological processes at the level of individual cell types and provides new perspectives for GRN reconstruction and analysis (Nguyen et al., 2021). GRNs for individual cell types will allow the discovery of regulatory circuits specific to cell states or degrees of differentiation.

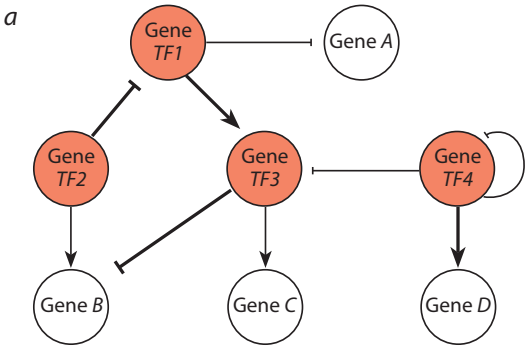
In this review, we discuss methods for GRN inference from scRNA-seq data, with a detailed focus on the incorporation of other omics data, primarily TF binding sites and open chromatin profiles. Special attention is given to biological results that have been achieved through GRN analysis.

Single-cell transcriptomes as a data source for GRN inference

Besides enabling inference of cell-type specific GRNs, scRNA-seq data offer other advantages over bulk RNA-seq. Since the number of interactions within a GRN is typically quite large, a substantial number of transcriptomic profiles (columns in the expression matrix, Fig. 1) is required for their accurate reconstruction. This is not always achievable with bulk RNA-seq data (Fig. 1*b*) (Altay, 2012), whereas scRNA-seq data contain a representative set of transcriptomes (ranging from several hundred to several thousand) (Fig. 1*c*) (Luecken, Theis, 2019).

The ultimate purpose of GRNs is to outline the dynamics of gene expression regulation in biological processes, including cell differentiation and responses to various internal and external stimuli. For the most accurate GRN inference from bulk RNA-seq data, time series experiments are required. In contrast, scRNA-seq data from one sample can contain information about gene expression changes over time if cells within the sample participate in the same biological process (e. g., differentiation) and are undergoing different stages (Saelens et al., 2019; Hou et al., 2023). In such cases, computational

a



b

	Samples			
	s1	s2	s3	s4
Gene *TF1*	3	2	2	3
Gene *TF2*	6	4	5	6
Gene *TF3*	6	3	4	5
Gene *TF4*	3	2	3	4
Gene *A*	2	1	2	2
Gene *B*	2	1	1	3
Gene *C*	1	1	2	1
Gene *D*	1	1	1	0

c

	Cells									
	a	b	c	d	e	f	g	h	i	j
Gene *TF1*	1	0	0	0	0	0	1	0	0	1
Gene *TF2*	0	1	0	1	1	0	0	2	1	0
Gene *TF3*	1	0	0	0	0	1	2	0	1	1
Gene *TF4*	1	1	0	1	0	0	0	0	0	0
Gene *A*	0	0	0	1	0	0	0	1	0	0
Gene *B*	0	0	0	1	1	0	0	0	0	0
Gene *C*	0	0	0	0	0	1	0	0	0	0
Gene *D*	0	1	0	0	0	0	0	0	0	0

Fig. 1. Gene regulatory network and transcriptomic data behind its construction.

a – visualization of a GRN graph model; *b* – gene expression matrix constructed from bulk RNA-seq data for several samples (s1–s4); *c* – gene expression matrix constructed from scRNA-seq data for a single sample. The graph nodes denote genes, edges reflect regulatory links, including their direction, type (activation or inhibition of transcription), and magnitude (the larger the weight of the edge, the stronger the regulator's influence on transcription). Red nodes correspond to TF-coding genes, white nodes correspond to other genes. In the GRN, edges originate only from TF-coding genes. In panel (*c*), different colors denote different cell types.

positioning of cells along a pseudotime trajectory (with the order of cells defined by the distance between their transcriptomes) allows for a good approximation of gene expression dynamics throughout the process.

However, it is important to remember that, in some samples, cells may be in the same state or they can participate in numerous independent processes, making reconstruction of biologically meaningful pseudotime trajectories impossible (Pratapa et al., 2020). Therefore, when selecting a method for GRN inference, it is crucial to determine whether pseudotime information is present in the single-cell transcriptome dataset, as some methods are designed specifically for data with cellular dynamics, while others are only suitable for static data. There are also methods that can be applied to both types of data.

At the same time, scRNA-seq data have some features that complicate their analysis, in particular, GRN reconstruction (Wagner et al., 2016; Nguyen et al., 2021). These concern transient activation or low expression of certain genes, gene expression changes during cell cycle, and other factors. The widespread use of scRNA-seq technology in biology has led to development of multiple algorithms for analyzing the data it generates, each addressing these challenges in different ways.

Reconstruction of GRNs from scRNA-seq data

In this section, we describe the main categories of popular algorithms used for GRN inference from scRNA-seq data (correlation- and mutual information-based methods, regression, Bayesian and logical networks, mathematical modeling with differential equations) (Fig. 2). It is worth noting that in benchmarking of GRN inference tools on both simulated and real scRNA-seq data, no single method has proven to be universally superior (Chen, Mar, 2018; Blencowe et al., 2019; Pratapa et al., 2020). Such variability may be attributed to the fact that each method is suitable for specific types and sources of data for which it was developed.

Correlation-based algorithms

Pearson correlation, a widely recognized statistical index for calculating the association between two variables, has been applied to measure the co-expression of TF-coding genes and their potential targets in RNA-seq and scRNA-seq datasets (Hong et al., 2013; Nguyen et al., 2021). Being symmetric in its arguments, correlation does not predict the directionality of regulatory interactions. It can identify associations between pairs of genes that do not necessarily have a direct regulatory relationship. Methods such as PPCOR (Kim, 2015) account for the influence of other genes by calculating semi-partial correlation coefficients. LEAP (Specht, Li, 2017), an algorithm specifically designed for the analysis of single-cell data, computes the maximum Pearson correlation between each pair of genes over varying lag-windows, given that the cells were arranged in a pseudotime order. Since this type of correlation is not symmetric, LEAP is capable of reconstructing directed gene regulatory networks. As a result of testing this program on transcriptomes from 564 individual mouse dendritic cells, LEAP identified several thousand previously unknown links between genes (Shalek et al., 2014).

Mutual information-based algorithms

Information-theoretic approaches utilize mutual information, which measures the reduction in entropy for one variable (e. g., the expression level of one gene) given the value of another variable (e. g., the expression level of another gene) (Chan et al., 2017; Qiu et al., 2020; Chang et al., 2024). To reduce false positives arising from indirect interactions between two genes, methods such as PIDC (Chan et al., 2017) use partial information decomposition (PID) to compute the proportional unique contribution (PUC) for a pair of genes that cannot be explained by the expression of a third gene. Since this relationship is symmetric, the reconstructed edges are undirected.

PIDC has been successfully applied to reconstruct GRNs from single-cell transcriptomes for three processes in mice: differentiation of megakaryocytes and erythrocytes from a common precursor, early embryogenesis, and embryonic hematopoiesis. In all three examples, PIDC identified previously unknown links, effectively highlighted gene modules at different stages of differentiation, and suggested gene interactions that facilitate transitions between stages. In a systematic

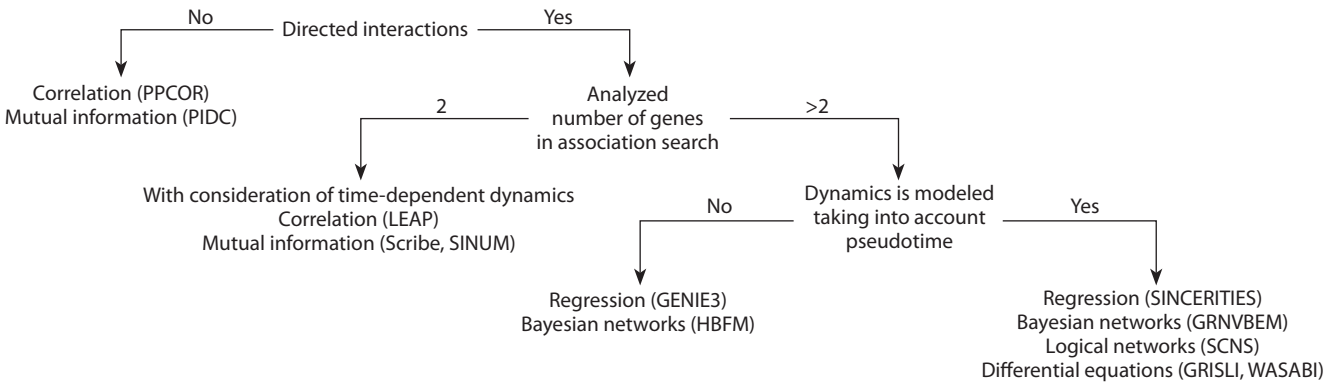


Fig. 2. The main categories of popular algorithms used for GRN inference from scRNA-seq data.

evaluation of 12 different GRN inference tools, PIDC was identified as one of the most effective (Pratapa et al., 2020).

Scribe (Qiu et al., 2020) uses pseudotime to compute restricted directed information (RDI). This measure assesses the mutual information between the preceding expression level of a TF-coding gene and the current expression level of a target gene, which is conditioned by the regulator expression earlier in the pseudotime series. Since the mutual information between preceding and current expression is asymmetric, Scribe can infer directed edges. Scribe has been applied both for verifying the existence of individual connections in various gene networks and for inferring the GRN of early embryogenesis in *Caenorhabditis elegans*, where the known hierarchy of transcriptional regulation of genes was reproduced.

The third program, SINUM, which also evaluates mutual information between any two genes and determines whether they are dependent or independent in a specific cell, has been tested on various types of data and has shown high effectiveness in identifying cell types, their marker genes, and gene connections, as well as in studying changes in gene associations during the differentiation of human embryonic stem cells into endoderm (Chang et al., 2024).

Regression-based algorithms

GRNs can be reconstructed by modeling the expression of each gene as a function of the expression levels of other genes and solving the resulting system of equations by using regression-based methods (Huynh-Thu et al., 2010; Gao et al., 2017; Moerman et al., 2018). GENIE3 employs a random forest method, which is based on an ensemble of regression trees (Huynh-Thu et al., 2010). The weight of the edge from a TF to a target gene arises from the significance of the TF in predicting the expression of the target gene, averaged across all regression trees in the random forest. GENIE3 was developed and has been widely used for bulk RNA-seq data analysis. The GRNBoost2 software enhances the scalability of GENIE3, particularly in terms of efficiently processing large datasets from single cells (Moerman et al., 2018). Both GENIE3 and GRNBoost2 have demonstrated their effectiveness in reconstructing GRNs from single-cell transcriptomes, showing good overlap with known biological interactions (Kang et al., 2021).

The SINCERITIES algorithm was specifically designed for single-cell transcriptomes and solves a regression model, which is based on temporal or pseudo-temporal changes in the distributions of gene expression levels (Gao et al., 2017). GRNBoost2 and SINCERITIES have been identified among the most effective algorithms for GRN inference in benchmarking of 12 programs based on different types of modeling (Pratapa et al., 2020). However, a recent comparative analysis of performance across different datasets and metrics revealed that GRNBoost2 generally outperforms SINCERITIES and more accurately identifies hubs in GRNs (Stock et al., 2024).

Bayesian networks

Another GRN inference approach models regulatory interactions within a Bayesian network. The GRNVBEM algorithm works with time samples, i.e. it requires that cells be sorted according to pseudotime beforehand (Sanchez-Castillo et al., 2017). Then it models the fold changes in gene expression between successive time points as a linear combination from the expression of gene regulators at the immediate previous time sample within the Bayesian network. The reconstruction of GRNs for early embryogenesis in mice and kidney cells of *Danio rerio* using this method allowed for the identification of hubs and the formation of hypotheses about differentiation regulators.

The HBFM method is based on gene co-expression analysis that employs a sparse hierarchical Bayesian factor model to reduce the impact of high intercellular variability and noise in single-cell datasets on the predicted network (Sekula et al., 2020). When analyzing single-cell transcriptomes from mouse brains, the program identified a significant number of known and putative protein-protein interactions from the STRING database.

Logical networks

While the previously presented methods infer networks that describe the regulatory effects of individual TFs, they do not account for the logical rules governing the combinatorial effect of multiple TFs on the expression of a target gene (Nguyen et al., 2021). For example, regulatory mechanisms may involve the activation of a gene only in the presence of

several specific TFs or, alternatively, its inhibition by another TF regardless of additional factors. Boolean networks are capable of characterizing these combinations of interactions by representing the active or inactive state of a gene as a binary variable, discretized using a gene expression threshold, and combining these states using AND, OR, and NOT operations to explain the expression of all genes in the system.

The SCNS program computes logical rules that explain the progression of gene expression from one pseudotime point to another (Woodhouse et al., 2018). Application of this program to transcriptomes from early-stage human embryo cells resulted in reconstruction of a core GRN for preimplantation embryonic development. The LogicNet algorithm employs probabilistic continuous logic to build a Boolean network, in which gene expression is modeled as a continuous rather than binary variable between 0 and 1, allowing for the construction of GRNs with directed and signed edges (Malekpour et al., 2020). Using LogicNet, GRNs for early embryogenesis in mice were constructed.

Differential equations

The presence of pseudotime information in scRNA-seq data allows for modeling gene expression using ordinary differential equations (ODEs) (Nguyen et al., 2021). Here, the rate of expression changes for a target gene is a function of expression of the gene encoding its TF regulator. By solving this system of equations, regulatory relationships can be determined based on the weight of each TF in the function, which describes changes in gene expression. The SCODE algorithm makes a simplifying assumption that changes in gene expression can be defined as a linear combination of reduced dimensional spaces to effectively solve a less complex system of equations using linear regression (Matsumoto et al., 2017). Alternatively, GRISLI estimates the rate at which the expression of each gene changes according to the dynamic process in each cell (Aubin-Frankowski, Vert, 2020). It subsequently simplifies the system of equations based on the assumption that the inferred GRN has few regulatory edges compared to the number of genes in the network, thereby reducing the problem to sparse regression.

A valuable feature of GRISLI is that it allows cells to follow multiple differentiation trajectories, whereas most methods permit only a linear, non-branching trajectory. The DynGENIE3 algorithm applies the random forest approach of GENIE3 to solve a system of ODEs, where the change in the expression of one gene is defined as a potentially nonlinear combination of the expression of other genes (Huynh-Thu, Geurts, 2018).

Another class of approaches is based on the observation that variations in gene expression from cell to cell may arise from the stochastic nature of molecular regulatory interactions (Nguyen et al., 2021). The piecewise-deterministic Markov process (PDMP) defines ODEs for gene expression as a function of a stochastic two-state Markov process indicating whether the transcription of the gene is activated, rather than directly as a function of the expression of regulating TFs (Herbach et al., 2017).

For each gene, the probability function representing transitions between active and inactive states includes a weight for each potential regulator. PDMP uses maximum likelihood estimation to determine these weights and thus infers the edges of the GRN. The WASABI algorithm implements an alternative maximum likelihood estimation based on the concept that observed increases or decreases in gene expression should precede transitions between active and inactive states in an earlier time window (Bonnaffoux et al., 2019). The application of WASABI for reconstructing the GRN of erythrocyte differentiation in birds revealed its unusual properties of this GRN – absence of hubs, a distributed network structure, and control of the expression of most genes directly by the factor inducing differentiation.

Refinement of GRNs reconstructed from scRNA-seq data through the recognition of TF binding sites

Despite the widespread use of scRNA-seq data for inferring GRNs, the accuracy of reconstructing the actual regulatory mechanisms based on these data remains unsatisfactory (Chen, Mar, 2018; Pratapa et al., 2020). This issue arises because programs for GRN inference from transcriptomic data are based on the assumption that the identified associations between the expression levels of TF-coding genes and their potential target genes imply direct transcriptional regulation. However, the observed associations may be caused by other biological phenomena or even random factors. Transcriptomic data do not contain direct information about regulatory events (e. g., TF binding to gene regulatory regions). Thus, it is challenging to distinguish between direct and indirect regulation based solely on scRNA-seq data.

To address these issues and enhance the effectiveness of GRN inference, it is necessary to incorporate additional data that directly characterize the factors involved in transcriptional regulation. For example, genome sequences bearing regulatory codes can be used to identify potential TF binding sites. In this case, the presence of a TF binding motif in the regulatory region of the target gene testifies in favor of direct TF-target gene regulation.

Accordingly, SCENIC utilizes a database of TF binding motifs to refine GRNs inferred with GENIE3 (Aibar et al., 2017). It keeps the links in the network only if the motifs, which correspond to the TF binding sites, are enriched in the promoter regions of the target genes. A later version, pySCENIC, employs parallelization to improve SCENIC efficiency (Van de Sande et al., 2020). In both studies, SCENIC successfully identified cell types in mouse and human brains (including those represented by as few as two to six cells), as well as stages of tumor development that are more difficult to distinguish than cell types (Aibar et al., 2017; Van de Sande et al., 2020). It also found a specific set of TFs for each cell type and tumor stage, including previously unknown oncological markers. The role of some of these markers in tumor progression was experimentally validated in the same studies.

Integration of scRNA-seq and scATAC-seq data for GRN reconstruction

In the genome, DNA is packaged into nucleosomes – the basic structural units of chromatin, which hinder TF bind-

ing to DNA, thereby preventing gene transcription (Parmar, Padinhateeri, 2020). Activation of genes is only possible when their regulatory regions are free from nucleosomes. The nucleosomal packaging of DNA is a regulated process and varies depending on conditions and cell types. The scATAC-seq (single-cell Assay for Transposase-Accessible Chromatin using sequencing) technology allows for identification of open chromatin areas, i. e., DNA regulatory regions that are accessible for TF binding, in individual cells (Buenrostro et al., 2015). Thus, scATAC-seq data can contribute to a more accurate reconstruction of direct regulatory relationships between TFs and their targets in GRNs.

It has been shown that integrating bulk RNA-seq and ATAC-seq (or other epigenomic data) significantly enhances the accuracy of GRN inference (Qin et al., 2014; Wang et al., 2015; Ackermann et al., 2016). This methodology is also applicable to single-cell sequencing data. However, due to the specificity of transcriptomic and epigenomic profiles by cell type and conditions, combining RNA-seq data with ATAC-seq or ChIP-seq data typically requires that both datasets be obtained from cells of the same type under identical conditions.

Current technologies allow for simultaneous sequencing of the transcriptome and epigenome in the same cell (Angermueller et al., 2016; Hu et al., 2016; Chen et al., 2019). An alternative is the integration of scRNA-seq and scATAC-seq data obtained from different biological samples of the same nature. In this case, an additional challenge for GRN reconstruction is establishing the correspondence between cell clusters representing the same type, condition, or state across two types of sequencing data. So-called diagonal integration methods are being developed to address this challenge (Argelaguet et al., 2021).

Since scATAC-seq is most frequently used for epigenome profiling in individual cells, several bioinformatics tools have been developed to integrate scRNA-seq and scATAC-seq data for GRN inference (Loers, Vermeirssen, 2024). GRNs reconstructed based on these data are specifically referred to as enhancer GRNs (eGRNs). STREAM reconstructs eGRNs based on jointly profiled scRNA-seq and scATAC-seq data, using a Steiner tree problem model, a hybrid biclustering pipeline, and submodular optimization to infer gene networks (Li et al., 2024). STREAM has been tested on single-cell data from human organs with pathologies (Alzheimer's disease and lymphocytic lymphoma) and has demonstrated its effectiveness in reconstructing TF–open binding site–gene connections along a pseudotime trajectory and in identifying transcriptional regulations specific to these diseases.

There are also programs that utilize the results of preliminary separate analyses of scRNA-seq and scATAC-seq data. For example, scMTNI takes as input a cell differentiation scheme, scRNA-seq results, and prior networks based on scATAC-seq for each cell type (Zhang et al., 2023). The application of scMTNI to scRNA-seq and scATAC-seq data on cell reprogramming in mice and differentiation of human hematopoietic cells allowed for the construction of eGRNs for both linear and branching lineages and the identification of regulators and other components of eGRNs specific to their fate transitions.

Conclusion

The identification of gene relationships in regulation of their expression is a key to understanding the mechanisms that ensure the realization of genetic information into specific phenotypic traits. The reconstruction of GRNs based on omics data from individual cells provides a unique opportunity to systematically investigate the mechanisms of cellular differentiation, as it theoretically allows for the reconstruction of regulatory gene networks for specific cell types and even at distinct stages of their development. To date, a number of methods have been worked out for reconstructing such GRNs, many of which are available to users as a software. However, despite the promising nature of this approach, its potential has not yet been fully realized. Not all available methods are user-friendly or easy to interpret.

The shortage of methods for verifying the reconstructed GRNs is also an ongoing challenge. Perhaps for this reason, the use of these models in specific biological studies remains limited, and there are only a handful of successful applications of single cell GRNs to address biological questions. Further advancements in molecular genetic technologies for studying individual cells and computational methods for analyzing the data they generate (particularly for the purpose of reconstructing and analyzing GRNs) will significantly narrow the gap between our knowledge of the molecular determinants of traits (including at the cellular level) and the transcriptional cascades triggered by external or internal stimuli. Breakthrough discoveries made with GRNs reconstructed from single cell omics data are likely awaiting us in the future.

References

Ackermann A.M., Wang Z., Schug J., Naji A., Kaestner K.H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab*. 2016;5(3):233-244. doi 10.1016/j.molmet.2016.01.002

Aibar S., González-Blas C.B., Moerman T., Huynh-Thu V.A., Imrichova H., Hulselmans G., Rambow F., Marine J., Geurts P., Aerts J., Van Den Oord J., Atak Z.K., Wouters J., Aerts S. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods.* 2017;14(11):1083-1086. doi 10.1038/nmeth.4463

Altay G. Empirically determining the sample size for large-scale gene network inference algorithms. *IET Syst. Biol*. 2012;6(2):35-43. doi 10.1049/iet-syb.2010.0091

Angermueller C., Clark S.J., Lee H.J., Macaulay I.C., Teng M.J., Hu T.X., Krueger F., Smallwood S.A., Ponting C.P., Voet T., Kelsey G., Stegle O., Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*. 2016;13(3):229-232. doi 10.1038/nmeth.3728

Argelaguet R., Cuomo A.S.E., Stegle O., Marioni J.C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol*. 2021;39(10):1202-1215. doi 10.1038/s41587-021-00895-7

Aubin-Frankowski P., Vert J. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics*. 2020;36(18):4774-4780. doi 10.1093/bioinformatics/btaa576

Badia-i-Mompel P., Wessels L., Müller-Dott S., Trimbour R., Flores R.O.R., Argelaguet R., Saez-Rodriguez J. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet*. 2023;24(11):739-754. doi 10.1038/s41576-023-00618-5

Blencowe M., Arneson D., Ding J., Chen Y.W., Saleem Z., Yang X. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerg. Top. Life Sci*. 2019;3(4):379-398. doi 10.1042/ETLS20180176

Bonnaffoux A., Herbach U., Richard A., Guillemin A., Gonin-Giraud S., Gros P., Gandrillon O. WASABI: a dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics*. 2019;20(1):220. doi 10.1186/s12859-019-2798-1

Buenrostro J.D., Wu B., Litzenburger U.M., Ruff D., Gonzales M.L., Snyder M.P., Chang H.Y., Greenleaf W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015; 523(7561):486-490. doi 10.1038/nature14590

Chan T.E., Stumpf M.P., Babtie A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst*. 2017;5(3):251-267.e3. doi 10.1016/j.cels.2017.08.014

Chang L., Hao T., Wang W., Lin C. Inference of single-cell network using mutual information for scRNA-seq data analysis. *BMC Bioinformatics*. 2024;25(S2):292. doi 10.1186/s12859-024-05895-3

Chen S., Mar J.C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*. 2018;19(1):232. doi 10.1186/s12859-018-2217-z

Chen S., Lake B.B., Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol*. 2019;37(12):1452-1457. doi 10.1038/s41587-019-0290-0

Gao N.P., Ud-Dean S.M.M., Gandrillon O., Gunawan R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*. 2017;34(2):258-266. doi 10.1093/bioinformatics/btx575

Herbach U., Bonnaffoux A., Espinasse T., Gandrillon O. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol*. 2017;11(1):105. doi 10.1186/s12918-017-0487-0

Hong S., Chen X., Jin L., Xiong M. Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res*. 2013;41(8): e95. doi 10.1093/nar/gkt145

Hou W., Ji Z., Chen Z., Wherry E.J., Hicks S.C., Ji H. A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat. Commun*. 2023;14(1):7286. doi 10.1038/s41467-023-42841-y

Hu Y., Huang K., An Q., Du G., Hu G., Xue J., Zhu X., Wang C., Xue Z., Fan G. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol*. 2016;17(1):88. doi 10.1186/s13059-016-0950-z

Huynh-Thu V.A., Geurts P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Sci. Rep*. 2018;8(1):3384. doi 10.1038/s41598-018-21715-0

Huynh-Thu V.A., Sanguinetti G. Gene regulatory network inference: An introductory survey. *Methods Mol. Biol*. 2019;1883:1-23. doi 10.1007/978-1-4939-8882-2_1

Huynh-Thu V.A., Irrthum A., Wehenkel L., Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*. 2010;5(9):e12776. doi 10.1371/journal.pone.0012776

Isbel L., Grand R.S., Schübeler D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat. Rev. Genet*. 2022;23(12):728-740. doi 10.1038/s41576-022-00512-6

Kang Y., Thieffry D., Cantini L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Front. Genet*. 2021;12:617282. doi 10.3389/fgene.2021.617282

Kim S. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods*. 2015;22(6):665-674. doi 10.5351/CSAM.2015.22.6.665

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2): 833-850 (in Russian)]

Li Y., Ma A., Wang Y., Guo Q., Wang C., Fu H., Liu B., Ma Q. Enhancer-driven gene regulatory networks inference from single-cell RNA-seq and ATAC-seq data. *Brief. Bioinform*. 2024;25(5):bbae369. doi 10.1093/bib/bbae369

Loers J.U., Vermeirssen V. A single-cell multimodal view on gene regulatory network inference from transcriptomics and chromatin accessibility data. *Brief. Bioinform*. 2024;25(5):bbae382. doi 10.1093/bib/bbae382

Luecken M.D., Theis F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol*. 2019;15(6):e8746. doi 10.15252/msb.20188746

Malekpour S.A., Alizad-Rahvar A.R., Sadeghi M. LogicNet: probabilistic continuous logics in reconstructing gene regulatory networks. *BMC Bioinformatics*. 2020;21(1):318. doi 10.1186/s12859-020-03651-x

Matsumoto H., Kiryu H., Furusawa C., Ko M.S.H., Ko S.B.H., Gouda N., Hayashi T., Nikaido I. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*. 2017;33(15):2314-2321. doi 10.1093/bioinformatics/btx194

Mercatelli D., Scalambra L., Triboli L., Ray F., Giorgi F.M. Gene regulatory network inference resources: A practical overview. *Biochim. Biophys. Acta Gene Regul. Mech*. 2020;1863(6):194430. doi 10.1016/j.bbagrm.2019.194430

Moerman T., Santos S.A., González-Blas C.B., Simm J., Moreau Y., Aerts J., Aerts S. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*. 2018; 35(12):2159-2161. doi 10.1093/bioinformatics/bty916

Nguyen H., Shrestha S., Tran D., Shafi A., Draghici S., Nguyen T. A comprehensive survey of tools and software for active subnetwork identification. *Front. Genet*. 2019;10:155. doi 10.3389/fgene.2019.00155

Nguyen H., Tran D., Tran B., Pehlivan B., Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief. Bioinform*. 2021;22(3):bbaa190. doi 10.1093/bib/bbaa190

Parmar J.J., Padinhateeri R. Nucleosome positioning and chromatin organization. *Curr. Opin. Struct. Biol*. 2020;64:111-118. doi 10.1016/j.sbi.2020.06.021

Pratapa A., Jalihal A.P., Law J.N., Bharadwaj A., Murali T.M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods*. 2020;17(2):147-154. doi 10.1038/s41592-019-0690-6

Qin J., Hu Y., Xu F., Yalamanchili H.K., Wang J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*. 2014;67(3):294-303. doi 10.1016/j.ymeth.2014.03.006

Qiu X., Rahimzamani A., Wang L., Ren B., Mao Q., Durham T., McFaline-Figueroa J.L., Saunders L., Trapnell C., Kannan S. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Syst*. 2020;10(3):265-274. doi 10.1016/j.cels.2020.02.003

Saelens W., Cannoodt R., Todorov H., Saeys Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol*. 2019;37(5):547-554. doi 10.1038/s41587-019-0071-9

Sanchez-Castillo M., Blanco D., Tienda-Luna I.M., Carrion M.C., Huang Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*. 2017;34(6):964-970. doi 10.1093/bioinformatics/btx605

Sekula M., Gaskins J., Datta S. A sparse Bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data. *BMC Bioinformatics*. 2020;21(1):361. doi 10.1186/s12859-020-03707-y

Shalek A.K., Satija R., Shuga J., Trombetta J.J., Gennert D., Lu D., Chen P., Gertner R.S., Gaublomme J.T., Yosef N., Schwartz S., Fowler B., Weaver S., Wang J., Wang X., Ding R., Raychowdhury R., Friedman N., Hacohen N., Park H., May A.P., Regev A. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510(7505):363-369. doi 10.1038/nature13437

Sönmezer C., Kleinendorst R., Imanci D., Barzaghi G., Villacorta L., Schübeler D., Benes V., Molina N., Krebs A.R. Molecular co-occupancy identifies transcription factor binding cooperativity *in vivo*. *Mol. Cell*. 2020;81(2):255-267. doi 10.1016/j.molcel.2020.11.015

Specht A.T., Li J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*. 2017;33(5):764-766. doi 10.1093/bioinformatics/btw729

Stock M., Popp N., Fiorentino J., Scialdone A. Topological benchmarking of algorithms to infer gene regulatory networks from single-cell RNA-seq data. *Bioinformatics*. 2024;40(5):btae267. doi 10.1093/bioinformatics/btae267

Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B.B., Siddiqui A., Lao K., Surani M.A. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. 2009;6(5):377-382. doi 10.1038/nmeth.1315

Tieri P., Castiglione F. Modeling macrophage differentiation and cellular dynamics. In: Wolkenhauer O. (Ed.). Systems Medicine. Integrative, Qualitative and Computational Approaches. Academic Press, 2021;511-520. doi 10.1016/B978-0-12-801238-3.11644-7

Van de Sande B., Flerin C., Davie K., De Waegeneer M., Hulselmans G., Aibar S., Seurinck R., Saelens W., Cannoodt R., Rouchon Q., Verbeiren T., De Maeyer D., Reumers J., Saeys Y., Aerts S. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc*. 2020;15(7):2247-2276. doi 10.1038/s41596-020-0336-2

Wagner A., Regev A., Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol*. 2016;34(11):1145-1160. doi 10.1038/nbt.3711

Wang P., Qin J., Qin Y., Zhu Y., Wang L.Y., Li M.J., Zhang M.Q., Wang J. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Res*. 2015;43(W1):264-269. doi 10.1093/nar/gkv398

Woodhouse S., Piterman N., Wintersteiger C.M., Göttgens B., Fisher J. SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol*. 2018;12(1):59. doi 10.1186/s12918-018-0581-y

Zhang S., Pyne S., Pietrzak S., Halberg S., McCalla S.G., Siahpirani A.F., Sridharan R., Roy S. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat. Commun*. 2023;14(1):3064. doi 10.1038/s41467-023-38637-9

Conflict of interest. The authors declare no conflict of interest.
Received October 28, 2024. Revised November 21, 2024. Accepted November 22, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):982-992

doi 10.18699/vjgb-24-105

Comparison of brain activity metrics in Chinese and Russian students while perceiving information referencing self or others

Q. Si[1], J. Tian[1], V.A. Savostyanov [1, 3], D.A. Lebedkin [1, 4], A.V. Bocharov [1, 3], 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] Tomsk State University, Tomsk, Russia

a.savostianov@g.nsu.ru

Abstract. Neurocomputing technology is a field of interdisciplinary research and development widely applied in modern digital medicine. One of the problems of neuroimaging technology is the creation of methods for studying human brain activity in socially oriented conditions by using modern information approaches. The aim of this study is to develop a methodology for collecting and processing psychophysiological data, which makes it possible to estimate the functional states of the human brain associated with the attribution of external information to oneself or other people. Self-reference is a person's subjective assessment of information coming from the external environment as related to himself/herself. Assigning information to other people or inanimate objects is evaluating information as a message about someone else or about things. In modern neurophysiology, two approaches to the study of self-referential processing have been developed: (1) recording brain activity at rest, then questioning the participant for self-reported thoughts; (2) recording brain activity induced by self-assigned stimuli. In the presented paper, a technology was tested that combines registration and analysis of EEG with viewing facial video recordings. The novelty of our approach is the use of video recordings obtained in the first stage of the survey to induce resting states associated with recognition of information about different subjects in later stages of the survey. We have developed a software and hardware module, i. e. a set of related programs and procedures for their application consisting of blocks that allow for a full cycle of registration and processing of psychological and neurophysiological data. Using this module, brain electrical activity (EEG) indicators reflecting individual characteristics of recognition of information related to oneself and other people were compared between groups of 30 Chinese (14 men and 16 women, average age 23.2 ± 0.4 years) and 32 Russian (15 men, 17 women, average age 22.1 ± 0.4 years) participants. We tested the hypothesis that differences in brain activity in functional rest intervals between Chinese and Russian participants depend on their psychological differences in collectivism scores. It was revealed that brain functional activity depends on the subject relevance of the facial video that the participants viewed between resting-state intervals. Interethnic differences were observed in the activity of the anterior and parietal hubs of the default-mode network and depended on the subject attribution of information. In Chinese, but not Russian, participants significant positive correlations were revealed between the level of collectivism and spectral density in the anterior hub of the default-mode network in all experimental conditions for a wide range of frequencies. The developed software and hardware module is included in an integrated digital platform for conducting research in the field of systems biology and digital medicine.

Key words: neurocomputing technologies; hardware-software module; data processing methods; self-referential processes; resting-state EEG; default-mode network; interethnic differences; collectivism.

For citation: Si Q., Tian J., Savostyanov V.A., Lebedkin D.A., Bocharov A.V., Savostyanov A.N. Comparison of brain activity metrics in Chinese and Russian students while perceiving information referencing self or others. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):982-992. doi 10.18699/vjgb-24-105

Funding. The development of the hardware-software module was carried out within the framework of the budgetary project ICG SB RAS No. FWNR-2022-0020. The part of the study concerning the preparation of psychological tests and selection of experimental groups was carried out with the financial support of the Development Program of Tomsk State University (Priority-2030).

Сопоставление показателей мозговой активности у китайских и российских студентов в условиях распознавания информации, отнесенной к себе и другим людям

Ц. Сы[1], Ц. Тянь[1], В.А. Савостьянов [1, 3], Д.А. Лебедкин [1, 4], А.В. Бочаров [1, 3], 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] Национальный исследовательский Томский государственный университет, Томск, Россия
a.savostianov@g.nsu.ru

Аннотация. Нейровычислительные технологии – область междисциплинарных исследований и разработок, которая находит широкое применение в современной цифровой медицине. Одна из задач нейровычислительных технологий состоит в создании методик изучения мозговой активности человека в условиях социально-ориентированной деятельности при помощи современных информационных подходов. Цель предлагаемого исследования – разработать методику сбора и обработки психофизиологических данных, позволяющую изучать функциональные состояния головного мозга человека, ассоциированные с отнесением внешней информации к самому субъекту или другим людям. Под самоотнесением (самореференцией) понимается субъективная оценка человеком поступающей из внешней среды информации как имеющей отношение к нему самому. Отнесение информации к другим людям или неодушевленным объектам – это оценка информации как сообщения о ком-то другом или о вещах. В современной нейрофизиологии сложились два подхода к исследованию самореференции: 1 – регистрация мозговой активности в условиях покоя с последующим опросом участника на предмет выявления самоотнесенных мыслей; 2 – регистрация мозговой активности, вызванной самоотнесенными стимулами. В представленной работе была апробирована технология, сочетающая регистрацию и анализ ЭЭГ с просмотром видеозаписей изображений лица самого испытуемого или незнакомого ему человека. Новизна нашего подхода состоит в использовании видеозаписей человеческого лица, полученных на первом этапе обследования, для индукции состояний покоя, ассоциированных с распознаванием информации о разных субъектах, на более поздних этапах обследования. Нами был разработан программно-аппаратный модуль, т. е. комплект связанных друг с другом программ и процедур их применения, состоящий из блоков, позволяющих проводить полный цикл регистрации и обработки психологических и нейрофизиологических данных. При помощи этого модуля показатели электрической активности головного мозга (ЭЭГ), отражающие индивидуальные особенности распознавания информации, отнесенной к самому себе и другим людям, были сопоставлены между группами из 30 китайских (14 мужчин и 16 женщин, средний возраст 23.2 ± 0.4 года) и 32 российских (15 мужчин, 17 женщин, средний возраст 22.1 ± 0.4 года) участников. Мы проверили гипотезу, что различия в мозговой активности в интервалах функционального покоя между китайскими и российскими участниками зависят от их психологических различий в показателях коллективизма. Было выявлено, что функциональная активность мозга зависит от субъектной отнесенности лицевого видео, которое участники просматривали между интервалами покоя. Межнациональные различия наблюдались в активности переднего и заднего хабов дефолт-системы и зависели от субъектной отнесенности информации. У китайских, но не у российских участников выявлены достоверные положительные корреляции между уровнем коллективизма и спектральной плотностью в переднем хабе дефолт-системы во всех экспериментальных условиях для широкого ряда частотных диапазонов. Разработанный программно-аппаратный модуль включен в интегрированную цифровую платформу для проведения исследований в области системной биологии и цифровой медицины.

Ключевые слова: нейровычислительные технологии; программно-аппаратный модуль; методы обработки данных; самоотнесенные процессы; ЭЭГ покоя; дефолт-система мозга; межнациональные различия; коллективизм.

Introduction

Neurocomputing technology is a technical field aimed at the development of methods for collection and computer analysis of neurophysiological data, which is widely used in digital medicine to create new approaches to diagnosis and therapy of diseases. The purpose of neurocomputing technologies is to develop programs and devices for obtaining information about the anatomo-functional organization of the nervous system in the norm and in pathologies.

The theory of reference was proposed in the works of logicians and linguists of the first half of the 20th century (for an overview, see Yakovleva, 2011). Information referencing is the evaluation of incoming information as being related to a particular object or subject. The term "self-reference" refers to the evaluation of an event as being related to the very subject perceiving information about that event (Northoff et al., 2005; Neff, McGehee, 2010). The term "self-reference" is fundamentally different in its content from the terms "reflection" (thinking about oneself) and "self-control" (controlling one's actions), as it does not refer to behavior management or self-assessment, but to the domain of analyzing the incoming information from the external environment as relevant or irrelevant to oneself. To date, two fundamentally different approaches to the study of neurophysiological markers of subjective attribution of information have emerged. In the first approach, brain activity (recorded via EEG, MEG, or fMRI) is recorded in conditions of functional rest, i. e., without performing experimental tasks (Knyazev et al., 2012, 2016). After completing the recording of brain activity, participants are surveyed about their focus on self-referential events. Another approach is to present participants with several sets of stimuli with unambiguous

attribution to self, familiar or unfamiliar people, or inanimate objects (Quevedo et al., 2018; Knyazev et al., 2020, 2024).

The goal of our study is to develop a new experimental model that combines both approaches described above to study the self-referential activity of the human brain, i. e., those neurophysiological processes that underlie the self-reference of information. In this model, the participant is presented with external information (viewing video images) about him/herself or another person versus observing an inanimate object. In the intervals between viewing the video images, the participant closes their eyes and does not receive external stimulation for some time. The proposed technology includes a technique for organizing data collection based on combining EEG recording with video recording of human faces (Savostyanov et al., 2022), a technique for preprocessing EEG data to clean the target signal from irrelevant noise (Delorme, Makeig, 2004), a technique for localizing the sources of brain signals on the cortical surface and searching for statistical relationships between neurophysiological activity and psychological characteristics of the survey participants (Pascual-Margui, 2002). In addition, our approach includes psychological testing to identify participants' personality traits and severity of depression symptoms. Within the framework of the proposed article, we will test the created technology to search for neurophysiological differences caused by different attitudes toward the self in groups of Russian and Chinese students. We hypothesize that Russians are more inclined to individualistic definition of their own personality, whereas the Chinese are more characterized by collectivistic ways of self-definition. The developed methods and computer programs for data collection and processing, as well as the actual data collected in this study, are included as one of the modules of the integrated digital platform "Bioinformatics and Systems Computational Biology", which is being developed at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences.

Materials and methods

Software module for data collection and processing. We have created a software module for data collection and processing, which is included in the integrated digital platform "Bioinformatics and Systems Computational Biology" that is being developed at ICG SB RAS. The module consists of both software products developed by the staff of ICG SB RAS and software tools from open sources. In total, all the blocks of the module allow us to carry out a complete cycle of collection and processing of psychological and neurophysiological data, starting from preliminary interviewing of participants to obtain their consent to be examined, and ending with statistical processing of the obtained results. The list of programs included in the module is given in Table 1.

Subjects. 30 undergraduate and PhD students from China (14 males and 16 females, mean age 23.2 ± 0.4 years) and 32 Russian undergraduate and PhD students (15 males, 17 females, mean age 22.1 ± 0.4 years), all studying at Novosibirsk State University, were invited. Before beginning the experiment, all participants completed a questionnaire that included questions about the presence of neurological or psychiatric diseases and alcohol or other psychoactive substance use. In addition, all participants gave informed consent to undergo the experimental examination in accordance with the Helsinki Declaration on Biomedical Ethics. The experimental protocol was approved by the ethical committee of the Scientific Research Institute of Neurosciences and Medicine.

Psychological evaluation. Participants filled out psychological questionnaires for trait-dependent and state-dependent anxiety (STAI: State-Trait Anxiety Inventory, Spielberger et al., 1970; Russian-language adaptation by Khanin, 1976), a questionnaire to assess the severity of depression symptoms (BDI: Beck's Depression Inventory, Beck et al., 1996), the Collective and Individual Self-Concept Test (SCS: Self Construal Scale, Singelis, 1994), and the Relationally-Interdependent Self-Construal (RISC: Relational-interdependent self-construal, Cross et al., 2000). The survey was conducted using a special Internet application developed on the Yandex platform. Russian participants filled out questionnaires in Russian; Chinese participants, in Chinese.

Experiment design, stages of data acquisition and processing. The experiment method and data processing steps are presented in the form of a flowchart in Figure 1. EEG was recorded in a sound- and light-isolated room. During the course of the experiment, three conditions were fulfilled. In the first experimental condition, EEG was recorded for 12 minutes without functional load (3 intervals of 2 minutes each with eyes closed and 3 intervals of 2 minutes each with eyes open). During the intervals when the subject opened their eyes, they saw a black screen of a computer monitor. During this period, the subject had a video image of their face recorded together with the EEG for all 12 minutes. The second and third conditions differed from the first in that in the second condition, with eyes open, the subject saw the video of their face obtained during the first condition, and in the third condition, they were presented with a video of an unfamiliar person's face (always a male for a male subject and a female for a female subject). The order of the second and third task was randomly switched. For about half of the subjects, the second task came first, followed by the third task; for the other half, *vice versa*.

EEG recording. EEG was recorded using an NVX-132 amplifier, Russia. 128 EEG channels were arranged according to the international 5–5 % system with reference electrode Cz, ground electrode AFz, and additional channels for EOG and ECG. Bandwidth was set at 0.1–100 Hz, signal sampling frequency, at 1,000 Hz. The EEG recording was done using the NeoRec recorder software.

EEG preprocessing. Re-reference to the average was performed to remove artifacts of tonic scalp muscle tension. Oculomotor and other artifacts were removed from the EEG using Independent Component Analysis (ICA) from the EEGLAB software package version 14.1.2b for the

Table 1. List of hardware and software blocks included in the module for registration and processing of neurophysiological data

Module block number	Name of the software product and its developer	Hardware required for the software to run	Block purpose
1	A special online form implemented on the Yandex platform by ICG SB RAS employees	Digital mobile device with Internet connection	Conducting remote psychological testing of participants to assess their personality traits and severity of depression symptoms
2	NeoRec Program, Medical Computer Systems, https://mks.ru/	Bioelectric signal amplifier NVX-132	Recording of brain bioelectrical activity under conditions of functional rest
3	Open Broadcaster Software, OBS Studio, https://obsproject.com/	Video camera connected to a recording computer	Registration of human face video recordings
4	A program for markup of EEG recordings based on facial video. Implemented by ICG SB RAS staff on the Inquisit platform, https://www.millisecond.com/	Bioelectric signal amplifier NVX-132, Steam Tracker for synchronization of event marks	Presentation of facial video recordings to the subject with simultaneous recording and annotation of EEG into resting and stimulation segments
5	EEGlab_toolbox, Swartz Center for Computational Neuroscience, https://sccn.ucsd.edu/eeglab/index.php	Personal computer	Pre-processing of EEG recordings, including frequency filtering, signal re-reference, Independent Component Analysis, and removal of extracerebral noise from the EEG signal
6	eLoreta: low resolution brain electromagnetic tomography, The KEY Institute for Brain-Mind Research, https://www.uzh.ch/keyinst/loreta.htm	Personal computer	Computation of spectral density in different frequency ranges
7			Localization of brain activity sources on the surface of the cerebral cortex
8			Conducting regression and correlation analyses to look for associations between participants' psychological traits and their neurophysiological characteristics
9	IBM SPSS software, IBM, https://www.ibm.com/spss	Personal computer	Statistical analysis of the obtained results

MATLAB environment (Delorme, Makeig, 2004). ICA is a widely used data analysis technique that allows, among other things, to separate signal from noise. The EEG recordings were then divided into periods when the participant had their eyes closed and periods when their eyes were open. Further analysis was performed only for those EEG intervals that were recorded with closed eyes but were enclosed by the periods of the corresponding stimulus observation. Once these EEG segments were extracted, they were divided into two-second time intervals.

Brain activity sources localization on the cortex surface. Further analysis was performed using the eLoreta software package (Pascual-Margui, 2002). eLoreta is a mathematical model and a software product based on this model, aimed at solving the inverse problem of EEG, i. e. at reconstructing the sources of functional processes in the brain based on computer analysis of the distribution of electrical signals on the surface of the head. eLoreta allows localization of brain activity sources based on interpolation of data from numerous EEG electrodes.

For each two-second interval, spectral density values were calculated in the frequency bands of delta (2–4 Hz), theta (4–8 Hz), alpha-1 (8–10 Hz), alpha-2 (10–12 Hz), beta-1 (12–16 Hz), beta-2 (16–20 Hz), beta-3 (20–25 Hz), and gamma (25–35 Hz) rhythms. Then, for each participant, the total spectrum over the entire EEG trial interval was calculated separately for each of the three experimental conditions (150 to 170 two-second intervals were used for each participant). Spectra were computed independently for each of the 128 EEG channels included in the data processing. Source-level analysis of spectral density comparisons between different conditions ("blank screen", "own face", and "other's face") was carried out in the eLoreta software. A 3,000 ms segment of the EEG recording with a sampling rate of 1,000 Hz after the onset of the block was used to calculate the spectral density

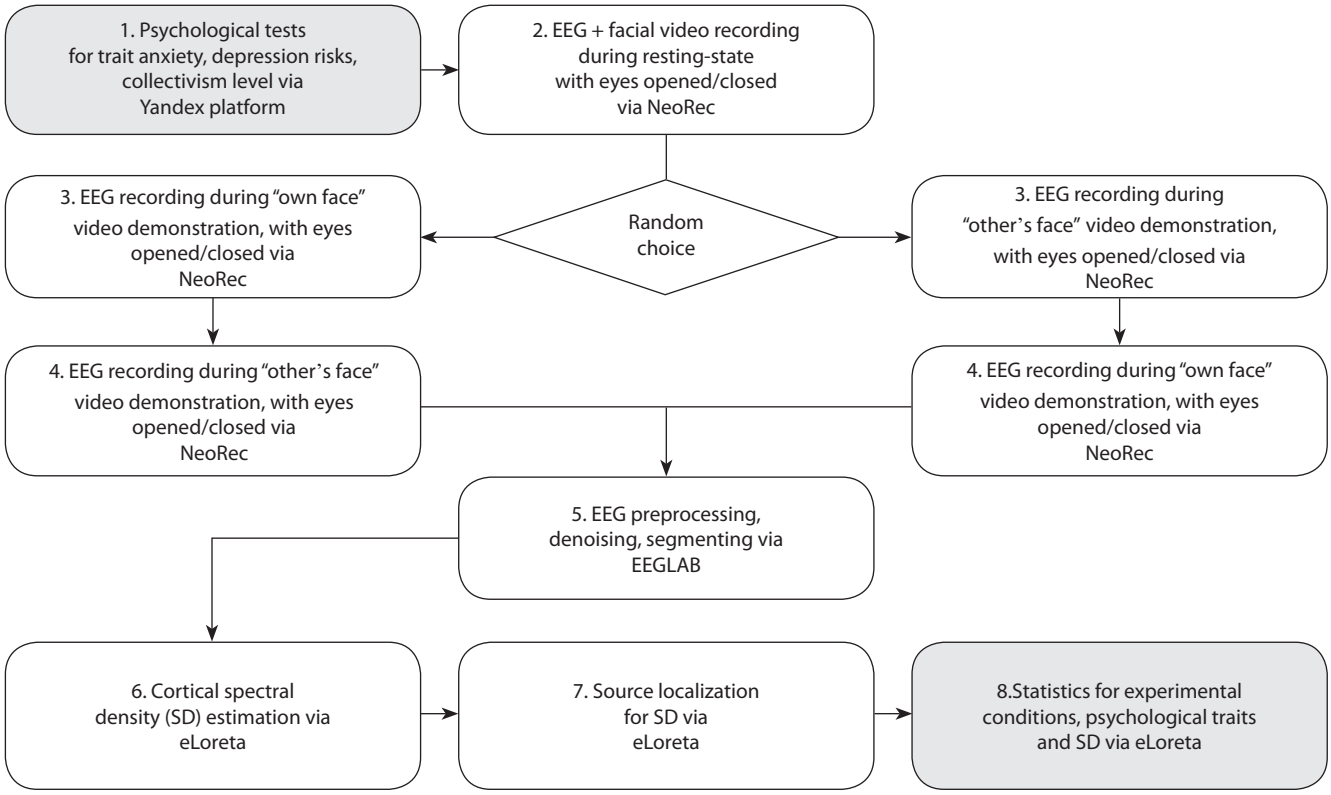


Fig. 1. Flowchart of data collection and processing stages with references to the computer programs used in our study.

of the sources in the eLoreta program (Pascual-Margui, 2002).

Statistical analysis of the results. Statistical analysis of the psychological assessment results was performed in the IBM SPSS software program. Comparisons were performed using one-way ANOVA with psychological traits as an independent variable, and intergroup factors "group" (Russian or Chinese), "gender" (male or female) and age as segregating variables.

Dependencies between experimental conditions and EEG metrics, and between psychological and ethnic characteristics and EEG metrics were assessed in the eLoreta package. The statistical significance of comparison results between different conditions was assessed using *t*-statistics for paired groups, with a randomization method of statistical nonparametric mapping (SnPM) that includes correction for multiple comparisons. The SnPM randomization method in eLoreta is based on a bootstrapping approach and is performed by multiple nonparametric permutation comparisons. A total of 5,000 randomizations were used to correct for multiple comparisons. Correlation analysis was performed to find the dependency of the spectral density on measures of personality traits and depression symptoms severity.

Results

Statistically significant results of the study and methods of their acquisition are presented in Table 2.

Results of psychological assessment

For the index of the anxiety trait according to the STAI test, the main effect of the "ethnicity" factor was not reliable ($p > 0.3$). A significant effect of the "gender" factor was found, $F(1; 62) = 6.47$, $p = 0.014$, $\eta^2 = 0.100$, mean anxiety in women (30.6 ± 1.6) was higher than in men (24.8 ± 1.7). The BDI test revealed a statistically significant value of the "ethnicity" factor, $F(1; 62) = 18.62$, $p < 0.0001$, $\eta^2 = 0.243$. The mean depression symptoms severity index was higher in the Chinese group (9.2 ± 1.1) than in the Russian group (2.8 ± 1.0).

The RISC questionnaire revealed statistically significant differences between the ethnic groups, $F(1; 62) = 7.27$, $p = 0.009$, $\eta^2 = 0.111$ in the importance of family values. The value of family was higher for Chinese participants (5.1 ± 0.2) than for Russian participants (4.3 ± 0.2). The SCS questionnaire also revealed a highly significant value for the "ethnicity" factor, $F(1; 62) = 23.41$, $p < 0.0001$, $\eta^2 = 0.288$ for the collectivism indicator. For participants from the Chinese group, the collectivism index was higher (5.0 ± 0.1) than for participants from the Russian group (4.5 ± 0.1) (Fig. 2). There was a significant interaction between the factors "gender" and "nationality" for this indicator, $F(1; 62) = 5.87$, $p = 0.019$, $\eta^2 = 0.092$. Russian (4.6 ± 0.1) and Chinese women (4.9 ± 0.1) did not differ significantly in this respect, whereas for Russian (4.3 ± 0.1) and Chinese (5.2 ± 0.1) men, the differences were more substantial.

Table 2. The main statistical results of the study, methods and software products used for obtaining them

Result	Significance level	Statistics method	Statistics software
Psychological differences in the level of collectivism between Russian and Chinese subjects	$p < 0.0001$	One-way analysis of variance (ANOVA)	IBM SPSS
Psychological differences in the severity of depression symptoms between Russian and Chinese subjects	$p < 0.0001$	One-way analysis of variance (ANOVA)	IBM SPSS
Differences in spectral density for different experimental conditions in both groups	$p < 0.01$	*t*-statistics for dependent samples	eLoreta
Differences in spectral density between the Russian and Chinese groups	$p < 0.05$	*t*-statistics for independent samples	eLoreta
Correlations between spectral density and measures of personality traits including the collectivism level	$p < 0.05$	Regression analysis with an independent variable	eLoreta

Results of eLoreta when comparing different experimental conditions for a generalized group (62 subjects, both Chinese and Russian participants)

Using the eLoreta software package, spectral density metrics were compared for EEG intervals with eyes closed, which followed intervals of observing one's own face, another person's face, or a blank screen. It was found that spectral density in the frequency ranges of delta (2–4 Hz), alpha-2 (10–12 Hz), and gamma (25–35 Hz) rhythms was higher with eyes closed after observing one's own face than with eyes closed after observing a blank screen. It should be specifically noted that muscle artifacts were removed from the EEG recordings using independent component analysis. According to Delorme and Makeig (2004), this method gives the ability to remove more than 80 % of all muscle noise. This suggests that the amplitude of electrical potentials in the delta and gamma bands was not simply due to surface tonic EMG. The statistically most reliable differences ($p = 0.0036$) were recorded for areas of the prefrontal cortex of both hemispheres (medial frontal area, 11 Brodmann's area, and orbitofrontal cortex, 47 Brodmann's area) in the range of the alpha-2 rhythm (Fig. 3*a*). Similar results were found when comparing the "other's face" and "blank screen" conditions (Fig. 3*b*). Also, as in the first comparison, spectral density in the prefrontal cortex in the alpha-2 rhythm band is shown to be higher for the "other's face" condition compared to the "blank screen" condition ($p = 0.002$). When comparing EEG intervals recorded after observing a stranger's face, it was found that the spectral densities in the frequency bands of alpha-1 (8–10 Hz) and alpha-2 (10–12 Hz) rhythms in EEG intervals with eyes closed after observing a videotape of one's own face were higher than in intervals with eyes closed after observing a stranger's face. Significant differences in spectral density for these conditions ($p = 0.0104$) were found for the parietal cortex (superior parietal lobe, 7 Brodmann's area, Fig. 3*c*).

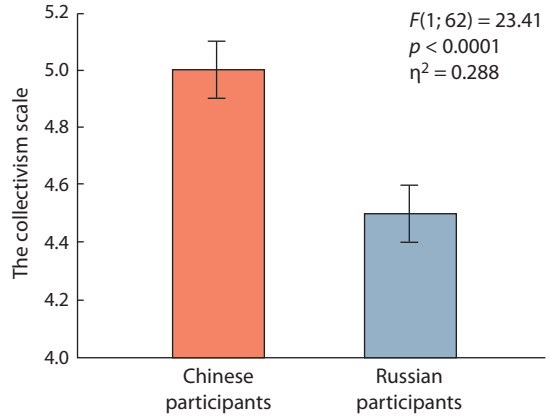


Fig. 2. Differences between Chinese and Russian participants in terms of the collectivism score from the SCS questionnaire.

Results of eLoreta when comparing different experimental conditions for Chinese and Russian participants

Comparison of spectral density indices between the groups of Chinese and Russian subjects in intervals with eyes closed following the observation of a blank screen did not reveal any statistically significant intergroup differences. In this condition, both groups showed similar spectral density distributions in all cortical areas and all frequency bands. Cross-ethnic comparisons in the eyes-closed condition following observation of a videorecording of one's face revealed significant differences in the alpha-2 and gamma rhythms ($p = 0.044$) (Fig. 4). Chinese participants in comparison with Russian participants showed increased spectral density in the

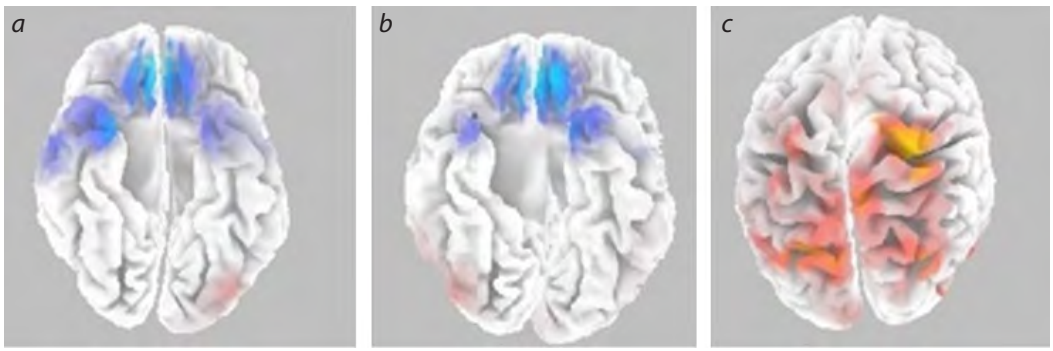


Fig. 3. Comparison of spectral density in the alpha-2 (10–12 Hz) rhythm when comparing intervals with eyes closed between conditions (*a*) blank screen *vs* own face; (*b*) blank screen *vs* other's face; (*c*) own face *vs* other's face.

The cortical regions in which spectral density is significantly higher for the "face" conditions than for the blank screen condition are marked in blue. Red color indicates cortical regions in which spectral density is significantly higher for the "own face" condition compared to the "other's face" condition.

alpha-2 band in the parietal and temporal cortex (38 Brodmann's area), whereas Russian participants in comparison with Chinese participants showed increased spectral density in the medial frontal cortex in the gamma rhythm band (3, 4, and 6 Brodmann's areas).

Cross-ethnic comparisons in the eyes-closed condition between the observation of a stranger's face video also revealed reliable differences in the ranges of alpha-2 and gamma rhythms ($p = 0.0002$), but they differed significantly from the results obtained for the own-face condition both in the topography of the effect and in the directionality of the cross-ethnic differences. Chinese participants in comparison to Russian participants showed significantly higher spectral density in the right inferior temporal cortical area (38 Brodmann's area) in the gamma band, whereas Russian participants in comparison to Chinese participants showed higher spectral density values in both bands (alpha-2 and gamma rhythms) in the prefrontal cortical areas (medial frontal area, 11 Brodmann's area and orbitofrontal cortex, 47 Brodmann's area) (Fig. 5).

Results of eLoreta in identifying the effects of psychological measures dependent on participants' ethnicity and gender

The correlations between the SCS collectivism score for the combined group of Russian and Chinese subjects were statistically insignificant. There was no significance for the "blank screen" condition ($p = 0.1954$). For the "own face" ($p = 0.0968$) and "other's face" ($p = 0.0664$) conditions for both groups, the *p*-levels were

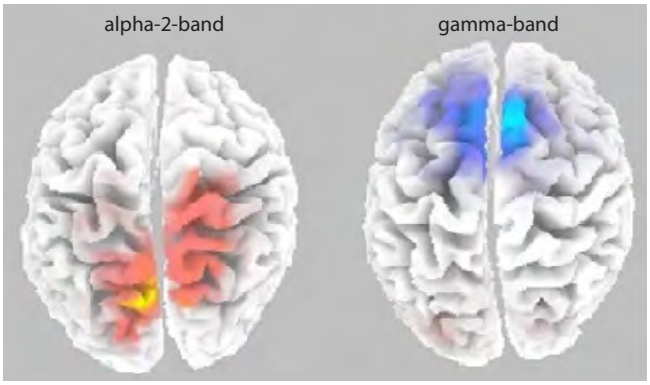


Fig. 4. Comparison of spectral density in the alpha-2 (10–12 Hz) and gamma (25–35 Hz) bands when comparing Chinese and Russian groups for EEG intervals with eyes closed between the participants' observation of their own face. This figure shows the superior surface of the cerebral cortex.

Chinese participants are characterized by a greater, when compared to Russian participants, spectral density of the alpha-2 rhythm in the posterior (parietal and temporal) cortical regions (areas marked in red), whereas Russian participants were found to have significantly greater values of gamma rhythm spectral density in the medial frontal cortical regions (marked in blue).

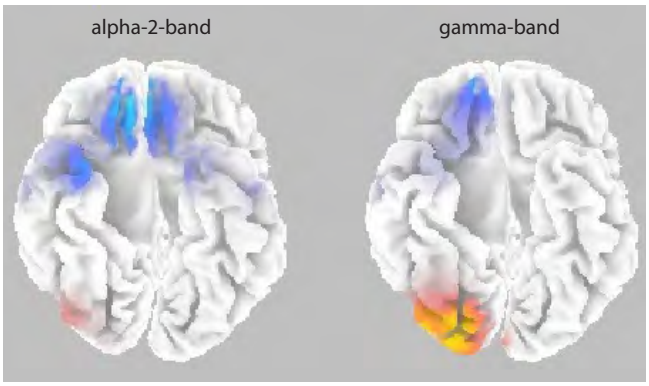


Fig. 5. Comparison of spectral density in the alpha-2 (10–12 Hz) and gamma (25–35 Hz) bands when comparing the Chinese and Russian groups for EEG of "eyes closed" intervals between the intervals of the participants' observation of a stranger's face. This figure shows the inferior surface of the cerebral cortex.

The Chinese group is characterized by a greater, when compared to the Russian group, spectral density of gamma rhythms in the right inferior temporal cortex (areas marked in red), whereas the Russian group showed significantly greater values of spectral density of both alpha-2 and gamma rhythms in the prefrontal cortex (marked in blue).

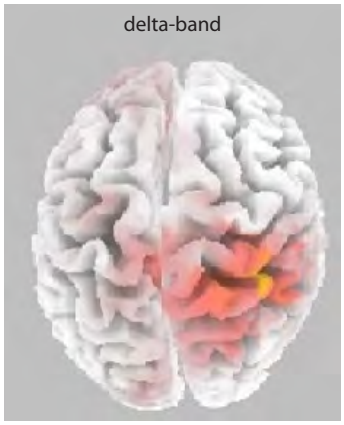


Fig. 6. Correlations between the collectivism level and delta rhythm spectral density in the group of Russian participants in the "eyes closed" intervals following the observation of a stranger's face.

The cortical areas in which reliable positive correlations of the collectivism level with EEG spectral density measures were found are marked in red. The figure shows the convexital surface of the brain.

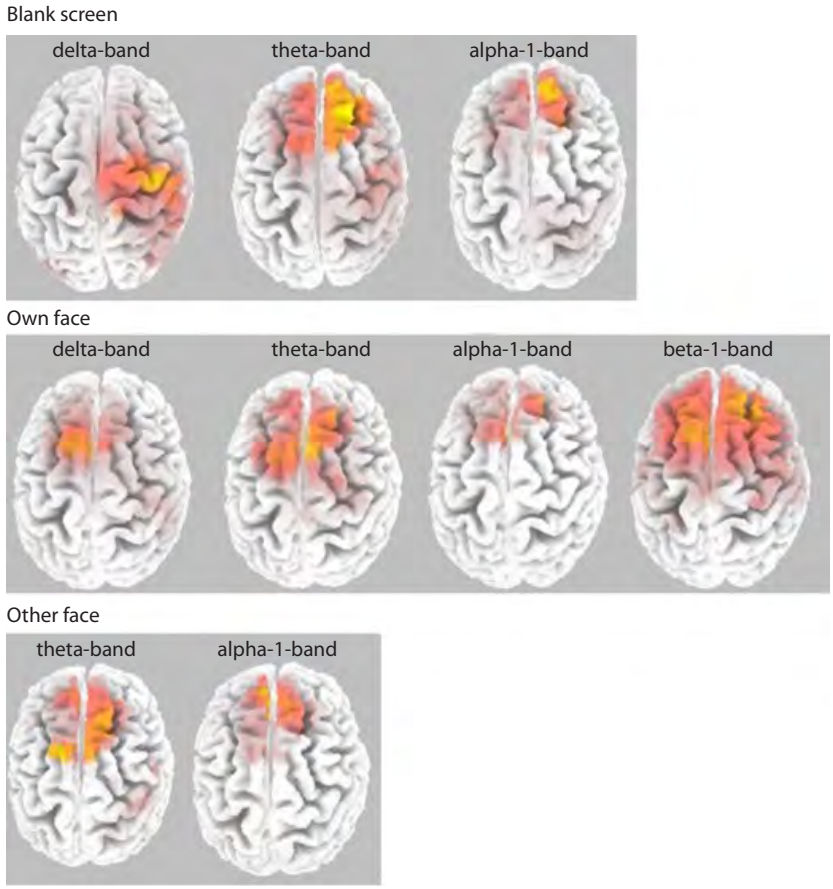


Fig. 7. Correlations between the collectivism index and spectral density for the blank screen (first row), own face (second row), and other person' face (third row) conditions in different frequency bands for Chinese participants.

The cortical areas that showed positive correlations between the level of collectivism and spectral density on EEG are marked in red. The figure shows the convexital surface of the brain.

close to, but did not reach, a significant value.

In the Russian sample for the collectivism index, no significant correlations were found for the "blank screen" or "own face" conditions. Significant correlations were found only for the spectral density in the delta band for the "other's face" condition (p = 0.043) in the right temporal cortex (Brodmann's area 22) (Fig. 6).

In contrast to the Russian sample, for the Chinese participants, statistically significant correlations with the collectivism score were found for all three conditions (for "blank screen" p = 0.001, for "own face" p = 0.0032, for "other's face" p = 0.0334). One can also notice that positive correlations with the collectivism score in the Chinese group were found for a wide range of delta, theta, alpha, and beta rhythms. These correlations are mainly found within the anterior cluster of the default-mode network (medial sections of the frontal and prefrontal cortex), and partially in the right parieto-temporal cortex (Fig. 7).

Discussion

Development of a hardware-software module for data collection and analysis

The aim of this work was to create a neurocomputing technology and develop a hardware and software module for collecting and analyzing data to study brain processes underlying personal self-reference. We had previously proposed an approach that combines the analysis of resting EEG with the analysis of facial mimetic muscle activity recorded under the same conditions (Savostyanov et al., 2022). The main result of the new work is the demonstration of the possibility of using facial video recordings obtained at the initial stage of the experiment to initiate the participants' processes of referencing information to themselves or others. Such data collection model is combined with well-known approaches for cleaning the EEG signal from noise (Delorme, Makeig, 2004) and localizing sources of brain activity on the surface of the cortex (Pascual-Margui, 2002).

One of the results of the study is the development of a hardware-software module that includes several sequentially connected blocks for experiment planning, data collection, preprocessing and analysis, as well as for intergroup statistical comparisons. In the future, this module can be used to conduct a wide range of neurophysiological studies, including the identification of markers of affective diseases such as depression, anxiety disorder, or autism spectrum disorders.

Neurophysiological correlates of self-referential information processing

Researchers' interest in studying the neurophysiological mechanisms of self-referential information processing is driven, firstly, by the fundamental role of self-reference in the formation of human personality, and secondly, by the presence of a wide range of psychiatric diseases, the symptoms of which

are various disorders in personal self-assessment (Bradley et al., 2016; Quevedo et al., 2018). In modern neurophysiology, there is a debate about the presence or absence of a specific anatomical substrate for self-referential processes in the brain (Northoff, Bermpohl, 2004; Northoff et al., 2005; Hu et al., 2016). The default-mode network, i. e., several interconnected cortical areas that show a decrease in the level of physiological activity when a person transitions from a resting state to performing cognitive tasks, is often considered as the main self-referential structure of the brain (Raichle, 2015; Knyazev et al., 2020, 2024).

The construction of a model of one's own personality is significantly determined by the subject's sociocultural specificity. In a classic study by Markus and Kitayama (1991), it was shown that representatives of Western (American) and Oriental (Japanese) cultures differ fundamentally in the criteria of the so-called "self-concept", i. e. the way of self-identification. Most Americans demonstrated individualistic personal attitudes, whereas collectivism was characteristic of the Japanese. In a cross-cultural study by G.G. Knyazev et al. (2012), a comparison of EEG correlates reflecting default-mode network activity at rest in representatives of Russian and Chinese (Taiwan) cultures was conducted. It was shown that most participants from Taiwan were characterized by dominance of the anterior (medial prefrontal cortex) hub of the default-mode network of the brain, whereas Russian participants showed dominance of the posterior (precuneus) part of this system (Knyazev et al., 2012). A hypothesis was proposed that interethnic differences in electrophysiological processes in the default-mode network may be caused by differences in self-concept according to the individualism-collectivism criterion, characteristic of representatives of Russian (predominantly individualistic) and Chinese (collectivistic) cultures. In our case, we experimentally tested the hypothesis of Knyazev et al. (2012) using data from the psychological questionnaires SCS and RISC.

Results of interethnic comparisons

The present study compared two samples of non-clinical subjects living in Russia at the time of the survey – Russian and Chinese. The examination included filling out psychological tests to identify the personality traits of the participants and the severity of their depression symptoms. The neurophysiological part of the examination consisted of EEG recording in three experimental conditions: (1) in the intervals between observation of a blank screen, (2) in the intervals between viewing a video of the participant's own face, and (3) in the intervals between viewing a video of the face of a person unfamiliar to the participant.

Psychological comparisons showed that Russian and Chinese subjects did not differ in the anxiety trait (STAI test). As for the severity of depression (BDI test), it was found that Chinese subjects expressed depression symptoms more strongly than Russian participants. This difference can be explained by the fact that Chinese participants had been away from their home for a long time, whereas Russian participants were in more familiar conditions. In the measures of collectivism for both tests we used (RISC and SCS), highly reliable differences were found between Chinese and Russian participants. As expected, significantly higher collectivism scores were found for Chinese participants than for Russian participants.

Spectral density comparisons between the condition pairs "blank screen" *vs* "own face", "blank screen" *vs* "other's face", "own face" *vs* "other's face" for a generalized group of all participants regardless of their ethnicity and gender revealed statistically significant differences, predominantly in the alpha-2 rhythm range. Differences between neutral (blank screen) and both social (both own and other's face) conditions were localized within the anterior hub of the default-mode network (medial prefrontal cortex). In both cases, the spectral density of the alpha rhythm was higher for the social than for the neutral condition. Differences between own and strangers' faces were localized within the posterior hub of the default-mode network (medial parietal cortex) and were expressed in higher spectral density for own than for strangers' faces.

Interethnic differences, without accounting for sex and psychological differences, were not detected in the EEG recorded in the intervals between blank screen observations, but were detected for the intervals between observations of both own and strangers' faces. For the "own-face" condition, differences were found in the range of the alpha-2 rhythm in the posterior hub of the default-mode network (Chinese participants had higher spectral density than Russian participants), and in the range of the gamma rhythm in the anterior hub of the default-mode network (Russian participants had higher spectral density than Chinese participants). For the "foreign face" condition, a higher density of both alpha and gamma rhythm sources was detected in the anterior hub of the default-mode network in Russian participants, whereas for Chinese participants, a higher spectral density was detected in the temporal cortex. Thus, our result generally confirms the conclusion of G.G. Knyazev et al. (2012) about the presence of interethnic differences in the operation of the anterior and posterior hubs of the default-mode network.

In the group of Russian subjects, assessments of collectivism correlated with brain activity indices only for the "stranger's face" condition. These correlations involved the posterior hub of the default-mode network. In contrast, in Chinese subjects, collectivism appeared to be a psychological metric for which multiple valid correlations were found for all three experimental conditions and several frequency ranges simultaneously. Most of the significant correlations in the Chinese group were found for brain structures from the anterior (medial frontal, medial prefrontal cortex) hub of the default-mode network. Thus, we confirm the hypothesis that the differences in default-mode network activity between Russian and Chinese subjects are mainly due to their differences in the collectivism index.

In general, thanks to the new experimental model proposed in this study, we were able to confirm G.G. Knyazev's hypothesis that cross-cultural differences in default-mode net-

work activity between Chinese and Russian participants are associated with their differences in collectivism indicators.

As a result of the study, we carried out the initial stage of development of a complex neurocomputing technology for collecting and analyzing psychological and physiological data, which allows to investigate the dynamics of processing self-referential information depending on the cultural features of the survey participants. The hardware-software module that we have developed is included in the integrated digital platform "Bioinformatics and Systems Computational Biology" being developed at ICG SB RAS under the budget project No. FWNR-2022-0020. It can be expected that the obtained approach will be further combined with the results of neurocomputer studies based on fMRI processing (Haxby et al., 2001) or with the data from psychogenetic studies. For example, for a portion of our subjects, data concerning their single-nucleotide polymorphisms in loci of brain neurotransmitter systems have been collected (Ivanov et al., 2022). Therefore, the results of psychological and neurophysiological studies can be compared with the genetic characteristics of the participants. In addition, convolutional neural networks using EEG metrics as input parameters can be used to classify participants into subgroups associated with different levels of stress (Fu et al., 2023).

Conclusion

1. Brain electrical activity recorded during the intervals of functional rest following stimulation differs for conditions after presentation of neutral, self-referential, or other-referential information to participants. This dependence is evident in measures of the spectral density of the alpha-2 rhythm in cortical regions that are part of the brain's default-mode network.
2. Functional activity of the default-mode network in Chinese and Russian subjects differs in resting intervals following the observation of subject-referencing stimuli, but does not differ for intervals following the observation of a blank screen. Functional activity in the anterior and posterior hubs of the default-mode network depends significantly on the ethnicity of the participants.
3. Functional activity in the anterior hub of the default-mode network is associated with collectivism in Chinese participants but not in Russian participants.

Limitations

1. During EEG recording, scalp EMG, which measures psychoemotional load, was not recorded. Although we performed the procedure of computing and applying the average reference, we can assume that the effects of personality traits and ethnicity in the gamma and beta bands are related not only to cerebral but also to muscular activity.
2. We chose standard rather than personalized frequency range boundaries, which may reduce the accuracy of identifying personalized EEG correlates of cognitive processes, especially for the alpha rhythm. Unfortunately, the software package we chose does not allow us to analyze spectral density in personalized ranges.
3. Although all female participants were interviewed before the experiment to establish the week of their menstrual cycle, we did not consider the psychoendocrinological factor of hormonal fluctuation in women when analyzing the EEG results, which may have reduced the accuracy of the findings.

We acknowledge all the limitations listed above and will strive to address them in future studies.

References

Beck A.T., Steer R.A., Brown G.K. Manual for the Beck Depression Inventory II. San Antonio, TX: Psychological Corporation, 1996

Bradley K.A., Colcombe S., Henderson S.E., Alonso C.M., Milham M.P., Gabbay V. Neural correlates of self-perceptions in adolescents with major depressive disorder. *Dev. Cogn. Neurosci*. 2016; 19:87-97. doi 10.1016/j.dcn.2016.02.007

Cross S.E., Bacon P.L., Morris M.L. The relational-interdependent self-construal and relationships. *J. Pers. Soc. Psychol*. 2000;78(4):791-808

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*. 2004;134(1):9-21. doi 10.1016/j.jneumeth.2003.10.009

Fu X., Tamozhnikov S.S., Saprygin A.E., Istomina N.A., Klemeshova D.I., Savostyanov A.N. Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):851-858. doi 10.18699/VJGB-23-98

Haxby J.V., Gobbini M.I., Furey M.L. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*. 2001;293(5539):2425-2430. doi 10.1126/science.1063736

Hu C., Di X., Eickhoff S.B., Zhang M., Peng K., Guo H., Sui J. Distinct and common aspects of physical and psychological self-representation in the brain: a meta-analysis of self-bias in facial and self-referential judgements. *Neurosci. Biobehav. Rev*. 2016;61:197-207. doi 10.1016/j.neubiorev.2015.12.003

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

Khanin Y.L. A Brief Guide to the C.D. Spielberger State and Trait Anxiety Scale. Leningrad, 1976 (in Russian)

Knyazev G.G., Savostyanov A.N., Volf N.V., Liou M., Bocharov A.V. EEG correlates of spontaneous self-referential thoughts: a cross-cultural study. *Int. J. Psychophysiol*. 2012;86(2):173-181. doi 10.1016/j.ijpsycho.2012.09.002

Knyazev G.G., Savostyanov A.N., Bocharov A.V., Tamozhnikov S.S., Saprigyn A.E. Task-positive and task-negative networks and their relation to depression: EEG beamformer analysis. *Behav. Brain Res*. 2016;306:160-169. doi 10.1016/j.bbr.2016.03.033

Knyazev G.G., Savostyanov A.N., Bocharov A.V., Levin E.A., Rudych P.D. Intrinsic connectivity networks in the self- and other-referential processing. *Front. Hum. Neurosci*. 2020;14:579703. doi 10.3389/fnhum.2020.579703

Knyazev G.G., Savostyanov A.N., Bocharov A.V., Saprigyn A.E. Representational similarity analysis of self-versus other-processing: effect of trait aggressiveness. *Aggress. Behav*. 2024;50(1):e22125. doi 10.1002/ab.22125

Markus H.R., Kitayama S. Culture and the self: implications for cognition, emotion, and motivation. *Psychol. Rev*. 1991;98(2):224-253. doi 10.1037/0033-295X.98.2.224

Neff K.D., McGehee P. Self-compassion and psychological resilience among adolescents and young adults. *Self Identity*. 2010;9(3):225-240. doi 10.1080/15298860902979307

Northoff G., Bermpohl F. Cortical midline structures and the self. *Trends Cogn. Sci*. 2004;8(3):102-107. doi 10.1016/j.tics.2004.01.004

Northoff G., Heinzel A., De Greck M., Bermpohl F., Dobrowolny H., Panksepp J. Self-referential processing in our brain – a meta-analysis of imaging studies on the self. *NeuroImage*. 2005;31(1):440-457. doi 10.1016/j.neuroimage.2005.12.002

Pascual-Margui R.D. Standardized low-resolution brain electromagnetic tomography (sLORETA). Technical details. *Methods Find. Exp. Clin. Pharmacol*. 2002;24(Suppl.D):5-12

Quevedo K., Harms M., Sauder M., Scott H., Mohamed S., Thomas K.M., Schallmo M.-P., Smyda G. The neurobiology of self face recognition among depressed adolescents. *J. Affect. Disord*. 2018; 229:22-31. doi 10.1016/j.jad.2017.12.023

Raichle M.E. The brain's default mode network. *Annu. Rev. Neurosci*. 2015;38:433-447. doi 10.1146/annurev-neuro-071013-014030

Savostyanov A.N., Vergunov E.G., Saprygin A.E., Lebedkin D.A. Validation of a face image assessment technology to study the dynamics of human functional states in the EEG resting-state paradigm. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):765-772. doi 10.18699/VJGB-22-92

Singelis T.M. The measurement of independent and interdependent self-construals. *Personality Social Psychol. Bull*. 1994;20(5):580-591. doi 10.1177/0146167294205014

Spielberger C.D., Gorsuch R.L., Lushene R.E. Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press, 1970

Yakovleva E.V. Theory of reference and theory of psyhosys tematics. *Izvestiia Rossiiskogo Gosudarstvennogo Pedagogicheskogo Universiteta im. A.I. Gertsena = Izvestia: Herzen University Journal of Humanities and Sciences*. 2011;131:226-233 (in Russian)

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

Received October 5, 2024. Revised November 12, 2024. Accepted November 13, 2024.

BIOMEDICINE
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):993-1007

doi 10.18699/vjgb-24-106

A concept of natural genome reconstruction.
Part 2. Effect of extracellular double-stranded DNA fragments on hematopoietic stem cells

V.S. Ruzanova [1&], S.G. Oshikhmina [1, 2&], A.S. Proskurina [1], G.S. Ritter [1], S.S. Kirikovich [1], E.V. Levites [1], Y.R. Efremov [1], T.V. Karamysheva[1], M.I. Meschaninova [3], A.L. Mamaev[4], O.S. Taranov [5], A.S. Bogachev[2], S.V. Sidorov[6], S.D. Nikonov[7], O.Y. Leplina [8], A.A. Ostanin [8], E.R. Chernykh [8], N.A. Kolchanov [1], E.V. Dolgova [1#], S.S. Bogachev [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
[3] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Laboratory Angiopharm LLC, Novosibirsk, Russia
[5] State Scientific Center of Virology and Biotechnology "Vector" of Rospotrebnadzor, Koltsovo, Novosibirsk region, Russia
[6] City Clinical Hospital No. 1, Novosibirsk, Russia
[7] Novosibirsk Tuberculosis Research Institute, Novosibirsk, Russia
[8] Research Institute of Fundamental and Clinical Immunology, Novosibirsk, Russia
labmolbiol@mail.ru

Abstract. In this part of the study, the first component of the concept of "natural genome reconstruction" is being proven. It was shown with mouse and human model organisms that CD34+ hematopoietic bone marrow progenitors take up fragments of extracellular double-stranded DNA through a natural mechanism. It is known that the process of internalization of extracellular DNA fragments involves glycocalyx structures, which include glycoproteins/protein glycans, glycosylphosphatidylinositol-anchored proteins and scavenger receptors. The bioinformatic analysis conducted indicates that the main surface marker proteins of hematopoietic stem cells belong to the indicated groups of factors and contain specific DNA binding sites, including a heparin-binding domain and clusters of positively charged amino acid residues. A direct interaction of CD34 and CD84 (SLAMF5) glycoproteins, markers of hematopoietic stem cells, with double-stranded DNA fragments was demonstrated using an electrophoretic mobility shift assay system. In cells negative for CD34, which also internalize fragments, concatemerization of the fragments delivered into the cell occurs. In this case, up to five oligonucleotide monomers containing 9 telomeric TTAGGG repeats are stitched together into one structure. Extracellular fragments delivered to hematopoietic stem cells initiate division of the original hematopoietic stem cell in such a way that one of the daughter cells becomes committed to terminal differentiation, and the second retains its low-differentiated status. After treatment of bone marrow cells with hDNA^{gr}, the number of CD34+ cells in the colonies increases to 3 % (humans as the model organism). At the same time, treatment with hDNA^{gr} induces proliferation of blood stem cells and their immediate descendants and stimulates colony formation (mouse, rat and humans as the model organisms). Most often, the granulocyte-macrophage lineage of hematopoiesis is activated as a result of processing extracellular double-stranded DNA. The commitment process is manifested by the appearance and repair of pangenomic single-strand breaks. The transition time in the direction of differentiation (the time it takes for pangenomic single-strand breaks to appear and to be repaired) is about 7 days. It is assumed that at the moment of initiation of pangenomic single-strand breaks, a "recombinogenic situation" ensues in the cell and molecular repair and recombination mechanisms are activated. In all experiments with individual molecules, recombinant human angiogenin was used as a comparison factor. In all other experiments, one of the experimental groups consisted of hematopoietic stem cells treated with angiogenin.

Key words: hematopoietic stem cells; extracellular DNA; internalization; terminal differentiation; single-strand breaks.

For citation: Ruzanova V.S., Oshikhmina S.G., Proskurina A.S., Ritter G.S., Kirikovich S.S., Levites E.V., Efremov Y.R., Karamysheva T.V., Meschaninova M.I., Mamaev A.L., Taranov O.S., Bogachev A.S., Sidorov S.V., Nikonov S.D., Leplina O.Y., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Dolgova E.V., Bogachev S.S. A concept of natural genome reconstruction. Part 2. Effect of extracellular double-stranded DNA fragments on hematopoietic stem cells. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):993-1007. doi 10.18699/vjgb-24-106

Funding. This work was supported by the Ministry of Science and Higher Education of the Russian Federation for the Institute of Cytology and Genetics (state budget-funded project No. FWNR-2022-0016) and by LLC "ES.LAB DIAGNOSTIC", I.N. Zaitseva and A.A. Purtov.

&, # Equal contribution of authors.

Концепция природной реконструкции генома.
Часть 2. Влияние фрагментов экстраклеточной двуцепочечной ДНК на гемопоэтические стволовые клетки

В.С. Рузанова [1&], С.Г. Ошихмина [1, 2&], А.С. Проскурина [1], Г.С. Риттер [1], С.С. Кирикович [1], Е.В. Левитес [1], Я.Р. Ефремов [1], Т.В. Карамышева[1], М.И. Мещанинова [3], А.Л. Мамаев[4], О.С. Таранов [5], А.С. Богачев[2], С.В. Сидоров[6], С.Д. Никонов[7], О.Ю. Леплина [8], А.А. Останин [8], Е.Р. Черных [8], Н.А. Колчанов [1], Е.В. Долгова [1#], С.С. Богачев [1#] ✉

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] ООО «Лаборатория Ангиофарм», Новосибирск, Россия
[5] Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р. п. Кольцово, Новосибирская область, Россия
[6] Городская клиническая больница № 1, Новосибирск, Россия
[7] Новосибирский научно-исследовательский институт туберкулеза, Новосибирск, Россия
[8] Научно-исследовательский институт фундаментальной и клинической иммунологии, Новосибирск, Россия
✉ labmolbiol@mail.ru

Аннотация. В настоящей части исследования доказывается первая составляющая концепции «природной реконструкции генома». На модельных организмах мышь и человек показано, что CD34+ гемопоэтические предшественники костного мозга захватывают фрагменты экстраклеточной двуцепочечной ДНК естественным природным механизмом. Известно, что в процессе интернализации экстраклеточных фрагментов ДНК принимают участие структуры гликокаликса, в состав которых входят гликопротеины/протеогликаны, гликозилфосфатидилинозитол-заякоренные белки и скавенджер-рецепторы. Проведенный биоинформационный анализ свидетельствует, что основные поверхностные маркерные белки гемопоэтических стволовых клеток относятся к указанным группам факторов и содержат специфические сайты связывания ДНК, включающие гепарин-связывающий домен и кластеры положительно заряженных аминокислотных остатков. С использованием системы Electrophoretic mobility shift assay показано прямое взаимодействие CD34 и CD84 (SLAMF5) гликопротеинов, маркеров гемопоэтических стволовых клеток, с фрагментами двуцепочечной ДНК. В клетках, негативных по CD34, также интернализующих фрагменты, происходит конкатемеризация доставленных внутрь клетки фрагментов. При этом в одну структуру сшивается до пяти мономеров олигонуклеотидов, содержащих девять теломерных повторов TTAGGG. Доставленные в гемопоэтические стволовые клетки экстраклеточные фрагменты инициируют деление исходной гемопоэтической стволовой клетки таким образом, что одна из дочерних клеток уходит в терминальную дифференцировку, а вторая сохраняет свой низкодифференцированный статус. В составе колоний после обработки клеток костного мозга препаратом hDNA^{gr} количество CD34+ клеток возрастает до 3 % (модельный организм – человек). Одновременно обработка препаратом hDNA^{gr} индуцирует пролиферацию стволовых клеток крови и их ближайших потомков и стимулирует колониеобразование (модельные организмы – мышь, крыса, человек). Наиболее часто в результате обработки экстраклеточной двуцепочечной ДНК активируется гранулоцитарно-макрофагальный росток кроветворения. Процесс коммитирования манифестируется появлением и репарацией пангеномных одноцепочечных разрывов. Время перехода в направлении дифференцировки (время появления и репарации пангеномных одноцепочечных разрывов) составляет около 7 суток. Предполагается, что в момент инициации пангеномных одноцепочечных разрывов в клетке создается «рекомбиногенная ситуация» и активируются молекулярные репаративно-рекомбинационные механизмы. Во всех проведенных экспериментах по анализу индивидуальных молекул в качестве фактора сравнения использовался ангиогенин рекомбинантный человеческий. Во всех других экспериментах одной из сравниваемых групп являлись гемопоэтические стволовые клетки, обработанные ангиогенином.

Ключевые слова: гемопоэтические стволовые клетки; экстраклеточная ДНК; интернализация; терминальная дифференцировка; одноцепочечные разрывы.

Introduction

Hematopoietic stem cell (HSC) and its bone marrow (BM) niche constitute a unique cell system, which maintains the balance of blood cell elements and repairs tissue and organs throughout life. The HSC concept is complex; it characterizes a number of cellular states and various cell types of different anatomical localization, developing into different cell lineages. Three HSC classes are distinguished: myeloid-biased, lymphoid-biased, and balanced cells; all of them vary in their differentiation capacity, which is fixed epigenetically. Clonal analysis indicates that these cell classes are comprised of two populations: short-lived HSCs and long-lived progenitors. The first cell population enters the differentiation and proliferation phase within a few weeks, while long-lived progenitors remain in the quiescent G0 phase for a long time (Muller-Sieburg, Sieburg, 2008).

It is generally believed that long-lived quiescent mouse HSCs have the following phenotype: Lin– Kit+ Sca-1+

CD150+ CD34– Flk2– CD48–. There are 30,000 BM mononuclear cells per one HSC, and about 80 % of HSCs remain quiescent throughout life (in humans), preserving their stemness (Morita et al., 2010; Zhang, Sadek, 2014; Wilkinson et al., 2020).

The HSC is surrounded by different cell types; these cells create a niche for the implementation of HSC functions. The stem cell niche is composed of endothelial cells, multiple mesenchymal cells (adipocytes, CXCL12+, adventitial reticular [CAR] cells, osteoclast-like cells [OLCs], leptinR+ and nestin+ cells, and NG2+ arteriolar wall cells), non-myelin-forming Schwann cells, and hematopoietic cells (macrophages and megakaryocytes) (Lévesque et al., 2010; Mendelson, Frenette, 2014; Kumar, Geiger, 2017; Szade et al., 2018; Lucas, 2019).

Two types of HSC niches are currently distinguished in the adult human BM. The osteoblastic niche is responsible for the quiescent state of early primitive progenitors that retain stemness for a long time. Once activated, HSCs differentiate into blood precursors located within a vascular niche, adjacent to sinusoid endothelial cells (Redondo et al., 2017).

The fundamental characteristic of the primitive HSC is its immanent choice: to either maintain the quiescent state and divide symmetrically into two identical HSCs or divide asymmetrically and give rise to a committed cell with further development of a certain cell lineage.

The HSC function is directly associated with the balance between quiescence and activation. A decreased ability of the HSC to exit quiescence results in insufficient blood cell reproduction. At the same time, if an unreasonably high number of cells exit quiescence and do not return to this state after activation, the HSC pool is depleted, resulting in BM function failure (Scharf et al., 2020). HSCs of a young organism are known to divide symmetrically and proliferate more often, while progenitors in adult and aging organisms are mainly quiescent (Desterke et al., 2021).

The establishment of the HSC state involves numerous factors. First of all, these are the anatomical localization of HSCs and the stem niche preserving them, and the local hypoxia level. Hypoxia is one of the key factors determining the HSC state, and the majority of quiescent and primitive HSCs are located in hypoxic BM areas with reduced blood perfusion (Forristal, Levesque, 2014; Zhang, Sadek, 2014). Factors secreted by the stem niche and HSCs, so-called membrane-associated factors (Winkler et al., 2012; Forristal, Levesque, 2014; Goncalves et al., 2016; Silberstein et al., 2016; Redondo et al., 2017; Chen T.L. et al., 2018; Scharf et al., 2020; Desterke et al., 2021), are important participants of the processes determining the HSC biological state. Furthermore, the same factor can induce quiescence in one HSC type and transition to the cycle and commitment in another type, as it was shown for angiogenin (Goncalves et al., 2016). Migrating peripheral leukocytes, histamine and TNF-α secreted by them, and other BM and peripheral blood cells induce activation of quiescent progenitors (Lucas, 2019; Pinho, Frenette, 2019). Different pharmacological agents, inflammation, starvation, environmental xenobiotics, and radiation also determine the HSC's fate (Chen T.L. et al., 2018; Scharf et al., 2020; Kiang et al., 2021; Wang et al., 2021).

Unsymmetrical division with subsequent commitment and proliferation is the basic mechanism of replenishment of blood cell populations. This process presents a finely regulated sequence of events, involving a diverse and abundant set of inducers. As previously mentioned, terminal differentiation, proliferation, and mobilization of HSCs can be activated by such environmental factors and body physiological systems as integral stimuli forming the common response vector of the HSC and its environment (the stem niche). This process results in activation of molecular signaling cascades and gene platforms determining the fate of the HSC and its committed progenitor (Kulkarni, Kale, 2020). Inflammation is one of the initiating factors in this process. As a result of the inflammatory response, a huge variety of active molecules are released into the bloodstream and lymphatic system, including a palette of pro-inflammatory cytokines, glucocorticoids (Pierce et al., 2017), granulocyte-macrophage colony-stimulating factor (GM-CSF), etc., which are the trigger releasing the resting HSC into the cycle. In addition, a large amount of apoptotic cell DNA (self-DNA) and pathogen-associated double-stranded DNA (dsDNA) and RNA appears in the bloodstream during both sterile and pathogen-induced inflammation (Jiang, Pisetsky, 2005; Saitoh et al., 2010; Lauková et al., 2019; Korabecna et al., 2020; Kananen et al., 2023). The involvement of the inflammatory process in HSC terminal differentiation indicates that all factors released into the blood during inflammation, including fragments of extracellular self/pathogen-associated DNA, affect the decision-making of primitive progenitors in a transient, competitive or restricted manner (Seita, Weissman, 2010). The inflammation is considered to shift differentiation of hematopoietic progenitors in the myeloid direction (Kovtonyuk et al., 2016).

Our recent studies have shown that stem cells of different genesis, cancer stem cells (Ritter et al., 2022), and HSCs (Potter et al., 2024) internalize extracellular dsDNA fragments through a natural mechanism. We propose that this newly discovered feature of poorly differentiated cells, including HSCs, is a transitional intermediary element in understanding the processes that take place in hematopoietic precursors, including the exit to terminal differentiation and proliferation upon their interaction with extracellular dsDNA fragments circulating in the blood.

There is another phenomenon that is the cornerstone of the concept proposed in the first part of the study. It is the presence of single-strand breaks (nicks) in the stem cell genome and their association with terminal differentiation of progenitors.

This phenomenon was first reported in studies conducted on a series of eukaryotic models at the end of the previous

century. To analyze the events occurring in the nuclear chromatin during commitment, the following inducers were used: DMSO, sodium butyrate, butyrylcholine, and retinoic acid. Single-strand breaks were detected using sedimentation assay (Jacobson et al., 1975; Scher, Friend, 1978), hydroxyapatite chromatography (Pulito et al., 1983), alkaline filter elution (McMahon et al., 1984; Boerrigter et al., 1989; Kaminskas, Li, 1989), *in situ* nick translation (Iseki, 1986; Patkin et al., 1995), and alkaline electrophoresis (McMahon et al., 1984; Vatolin et al., 1997). It turned out that formation and repair of single-strand breaks is a dose- and time-dependent process that does not correlate with the direction of differentiation (Scher, Friend, 1978; Farzaneh et al., 1982).

Chromatin nicking was shown to be associated with the activity of calcium/magnesium-dependent DNases, i. e. it is an enzymatic process, and single-strand breaks occur randomly (McMahon et al., 1984; Kaminskas, Li, 1989). Repair of single-strand breaks involves ADP-ribosyl transferase, which, in turn, is also believed to regulate differentiation through stimulation of ligase activity (Farzaneh et al., 1982; Johnstone, Williams, 1982). Quite peculiar and complex results were obtained in the study (Patkin et al., 1995). In this work, using *in situ* nick translation, the authors established that metaphase chromosomes in stem cells contain numerous nicks in the phase of transition to a committed state.

Thus, the presence of single-strand breaks was shown to closely correlate with terminal differentiation of stem cells. This event is considered the earliest manifestation of initiated commitment. These breaks are not associated with apoptosis, they do not result in cell death, and chromatin integrity is restored after a certain time. A possible explanation for this phenomenon is activation of genes necessary for commitment at this point in time (Jacobson et al., 1975; Farzaneh et al., 1982). We believe that this phenomenon is the cornerstone of the entire differentiation process: a biological, supramolecular, and large-scale manifestation of a change in the cell biological status. It is pangenomic single-strand breaks that allow the cell, apparently with minimal energy costs, to reorganize the chromatin topology of the undifferentiated state into a new architecture required for cell specialization (which, naturally, is associated with a fundamental change in the platform of expressed genes, as follows from the reasoning in the work (Jacobson et al., 1975; Farzaneh et al., 1982)). This is the phenomenon we attempted to characterize in the current part of the study within the new experimental framework, where extracellular dsDNA fragments act as the inducer.

Unfortunately, we did not manage to find studies on the presence and repair of single-strand breaks in hematopoietic stem cells in the available literature for the past 20 years. It is absolutely unclear why this area characterizing terminal transition of poorly differentiated stem cells of various origin has not received further development.

Therefore, in the second part of the work cycle, we analyzed internalization of dsDNA fragments in cells and their induction of terminal differentiation of progenitors, which manifests itself in the formation and repair of pangenomic single-strand breaks.

Materials and methods

Experimental animals. The following animals were used in the study: male CBA/Lac mice aged 2–5 months, 9–12 months old male CBA/Lac mice, male Wistar rats aged 2–6 months, and 18–22 months old male Wistar rats. All animals were bred at the Conventional Vivarium of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia). Animals were kept in groups of 6–10 mice and 3–4 rats per cage with free access to food and water. All animal experiments were approved by the Animal Care and Use Committee of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Mice were withdrawn from the experiment by cervical dislocation, and rats were either euthanized using CO_2 or decapitated.

Human bone marrow cells. Cryopreserved bone marrow cells from patients with Hodgkin lymphoma were used in the study. Cells were provided by the Cryobank of the Research Institute of Fundamental and Clinical Immunology.

hDNA^{gr}. The hDNA^{gr} preparation (DNA genome reconstructor) was isolated from placentas of healthy women. Total genome DNA was fragmented to 1–20 nucleosome monomers (200–2,000 bp) by ultrasonic disintegration, deproteinized using proteinase K, and extracted with phenol-chloroform.

Angiogenin. Recombinant human angiogenin was provided by Angiopharm Laboratory LLC (Novosibirsk, Russia). Angiogenin was labeled with Cy5 according to the manufacturer's instructions (Lumiprobe, Germany).

TAMRA-labeled DNA probe. Human *AluI* repeat DNA was labeled with the fluorescent dye TAMRA by PCR using TAMRA-5′-dUTP (deoxyuridine triphosphate) as described in (Dolgova et al., 2014).

Assessment of change in gel mobility of the complex of CD34 and SLAMF5 proteins and DNA probes. To analyze the interaction of the CD34 and SLAMF5 proteins with TAMRA-labeled DNA probe and P^{32}-labeled double-stranded $(TTAGGG)_9$ telomeric repeat, protein and DNA samples were incubated at different ratios and for different time periods in 10 mM PBS buffer at 37 °C (see Figure 1 caption). Incorporation of γP^{32}-ATP and native polyacrylamide gel electrophoresis were performed according to standard procedures (Maniatis et al., 1984; DNA Cloning…, 1985).

Isolation of bone marrow cells. To isolate the BM, animals were withdrawn from the experiment, femurs and tibias were isolated, epiphyses were removed, and BM cavity was washed with IMDM + 2 % FBS. The resulting cell suspension was passed through a 21-gauge needle several times to eliminate BM rosettes and then through a 40-μm filter. Cells were pelleted for 10 minutes at 400 *g*

and resuspended in red blood cell lysis buffer containing 130 mM ammonium chloride for 3–5 min. The buffer was then diluted 10-fold with PBS, cells were re-pelleted, resuspended in IMDM medium, and counted in a Goryaev chamber.

Internalization of DNA and angiogenin by human and mouse HSCs. To stain HSC colonies, mouse anti-Sca-1 and anti-c-Kit antibodies and 0.1 μg of TAMRA-labeled DNA were added to cells in 100 μl of IMDM medium using the manufacturer's protocol. The resulting mixture was carefully plated in 35-mm Petri dishes with HSC colonies by avoiding the contact with methylcellulose and colonies and then spread over a small surface area. A laser scanning confocal microscope LSM 780 NLO (Zeiss) and ZenLight software were used for data collection and imaging.

To quantify TAMRA-positive (TAMRA+) cells in BM cells and colony cell suspension, 1×10^6 cells were incubated in 400 μl of IMDM supplemented with 0.1 μg of TAMRA-labeled DNA for 30 min at room temperature in the dark. Cells were pelleted for 5 min at 400 *g* and 25 °C, washed in a small medium volume, and resuspended in the final medium volume. The same protocol was used for staining and analysis of c-Kit+/Sca-1+/TAMRA+ cells.

For fluorescence confocal microscopy analysis, 5 μg of Cy5-labeled angiogenin with and without antibodies was added to 3×10^6 BM cells and colonies resuspended in 1 ml of cell culture medium in a 12-well plate. After 30–60-min incubation, cells were analyzed on a laser scanning confocal microscope LSM 780 NLO (Zeiss) using ZenLight software. FACS analysis of cells was performed on a BD FACSAria III flow cytometer at the Flow Cytometry Center for Collective Use of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences.

DNA quantification in HSCs. For incubation of HSC colony cells with the human Alu repeat, colonies obtained after BM cell induction with $hDNA^{gr}$ were collected from two 35-mm Petri dishes on day 10 by adding 8 ml of IMDM. Cells were pelleted by centrifugation at 400 *g* for 8 min, washed with 2 ml of the medium, and re-pelleted. A fragment of the human Alu repeat was added to cells to a concentration of 0.23 μg per 1×10^6 cells; the mixture was incubated for 30 min. Cells were washed, pelleted by centrifugation at 400 *g* for 5 min, and resuspended in 1 ml of PBS.

Real-time PCR was conducted using the BioMaster RT-qPCR kit (SYBR Green dye) (#RM03-200, Biolabmix, Russia). Standard M13 primers (M13 forward: 5′-GTAAA-ACGAC-GGCCA-G-3′, M13 reverse: 5′-CAGGA-AAC AG-CTATG-AC-3′) and different amounts of Alu repeat DNA (0–5,000 pg) were used to obtain the calibration curve. Each concentration was used in triplicate. The linear dependence of Ct on Alu DNA load was constructed using Bio-Rad CFX Manager v3.1 software.

Treatment of BM cells with inducers. BM cells isolated from old animals and BM sections from patients with Hodgkin lymphoma were incubated with inducers ($hDNA^{gr}$ or angiogenin or two inducers simultaneously) for one hour in the 5 % CO_2 atmosphere with 95 % humidity at 37 °C at the following ratio: 500 μg of $hDNA^{gr}$ or 500 ng of angiogenin or 500 μg of hDNAgr and 500 ng of angiogenin in 1 ml in serum-free MDM medium per 3×10^6 cells. Control (untreated) BM cells were incubated in serum-free IMDM complimented with the PBS volume equal to that of the inducer added to activate BM cells. We use the term "inducer", which designates both DNA and angiogenin in the current study, to characterize any intended and expected HSC response induced by exposure to them.

Cultivation of BM cells in methylcellulose medium. BM cells with/without inducer activation were pelleted for 10 min at 400 *g* and resuspended in IMDM + 2 % FBS. To quantify and analyze myeloid precursors, we placed mouse BM cells in the MethoCult M3434 methylcellulose medium, and rat and human bone marrow cells, in the MethoCult H4034 methylcellulose medium (Stem Cell Technologies). Methylcellulose analysis, colony counting, and cell isolation from methylcellulose after cultivation were carried out according to the manufacturer's instructions. The analysis was performed in 35-mm Petri dishes, which were stored in a Petri dish of a larger diameter with additional humidification of the internal atmosphere during colony formation.

Comet tail assay for analysis of single- and double-strand breaks. BM cells isolated from old mice and BM sections from patients with Hodgkin lymphoma after incubation in the presence/absence of inducers ($hDNA^{gr}$, angiogenin, and $hDNA^{gr}$+angiogenin) were cultured for 10–12 days in methylcellulose medium. Colonies isolated from methylcellulose were pooled and washed from the medium according to the manufacturer's instructions. The resulting colony cells were counted in a Goryaev chamber and incubated with inducers. Cells were re-pelleted for 10 min at 400 *g*, resuspended in IMDM + 2 % FBS, placed in methylcellulose, and seeded into 24-well plates. A cell sample was collected every day at the same time (24, 48, 72, 96, 120, and 144 hrs after the start of treatment with inducers) and washed from methylcellulose. Colony cells were embedded into slow-melting 1 % agarose blocks in the amount of 5×10^3 cells per 1 block. Blocks were stored in 0.5 M EDTA at 4 °C prior to analysis. The zero point presents colony cells prior to repeated treatment with inducers.

Prior to electrophoresis, blocks were rinsed in TE buffer, incubated with a lysis buffer (50 mM EDTA, 1 % sarcosyl (Serva, Heidelberg, Germany), and 1 mg/ml proteinase K (Thermo Fisher Scientific, Waltham, USA)) for 20 min at 50 °C.

Prior to native electrophoresis, blocks were stained for 10 min in TAE buffer containing 0.5 μg/ml ethidium bromide (Medigen, Novosibirsk, Russia). Blocks were fixed on an agarose support, native electrophoresis was performed in 1×TAE buffer at 36 V and 299 mA (Model H4 Horizontal Gel Electrophoresis System (BRL, USA)) for 30 min.

Alkaline electrophoresis was carried out in a buffer containing 300 mM NaOH and 1 mM EDTA (pH > 13). Prior to alkaline electrophoresis, blocks were rinsed in the electrophoretic buffer and fixed on an agarose support. The support with blocks was placed in the electrophoretic buffer for 30 minutes. Alkaline electrophoresis was performed at 36 V and 299 mA (Model H4 Horizontal Gel Electrophoresis System chamber (BRL, USA)) for 30 min. After electrophoresis, the support with blocks was transferred to a neutral buffer containing 0.4 M Tris (pH 7.5) for 15 min. The neutral buffer was then replaced with a new one, 1 μg/ml of ethidium bromide was added, and the support with blocks was stained for 30 min.

The support with blocks was rinsed with distilled water. Preparations were obtained and dried at 37 °C for 24 hrs. After drying, preparations were washed in distilled water for 0.5–1 h. Microscopic analysis was performed on a Zeiss Axio Imager M2 (Carl Zeiss Microscopy, Oberkochen, Germany) at the Center for Collective Use for Microscopic Analysis of Biological Objects of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Comet tail values were assessed using CASP (CASP, Wrocław, Poland) and ImageJ software.

Statistical analysis. Statistical analysis was performed using Statistica 8 software (StatSoft, USA). The reliability of differences was assessed using the Mann–Whitney U-test. Statistical significance is indicated in figure legends ($p < 0.05$ or $p < 0.01$).

Results

HSC capability to internalize dsDNA fragments

Our recent studies (Dolgova et al., 2014; Petrova et al., 2022; Ritter et al., 2022) report a new general biological property of stem cells of various genesis. We confirmed experimentally that mouse HSCs, as well as all poorly differentiated cells of higher eukaryotes analyzed by us, including cancer stem cells, can capture dsDNA fragments from the environment through a natural mechanism. The interaction of extracellular DNA molecules with the cell and their internalization are mediated by the glycocalyx elements of glycoproteins/proteoglycans, glycosylphosphatidylinositol-anchored proteins, and the scavenger receptor system through the caveolae/clathrin-dependent mechanism. The most important and characteristic feature is the uniqueness of the pattern of glycoproteins/proteoglycans, glycosylphosphatidylinositol-anchored proteins, and scavenger receptors located on the surface of an individual cell type. This uniqueness is determined and limited by three functional domains composed of their different representatives, namely, molecules of glycoproteins/proteoglycans, glycosylphosphatidylinositol-anchored proteins, and scavenger receptors. In other words, each stem cell can have at least three functional domains that determine its interaction with extracellular double-stranded nucleic acids and internalization of the latter. For dsDNA molecules, the heparin-binding domain, which is presented in various cell surface proteins either by the C1q domain, heparin-binding domain or the domain of positively charged amino acids, is the main binding site (Petrova et al., 2022; Ritter et al., 2022).

Is this work, we also carried out FACS and immunofluorescence analysis of the capability of human HSCs to internalize extracellular dsDNA fragments in comparison with mouse HSCs. Recombinant human angiogenin was used as a reference factor, since its effect on the cell is well-studied. We also quantified extracellular dsDNA internalized in human CD34+ HSCs.

As mentioned above, glycocalyx factors (glycoproteins/proteoglycans, glycosylphosphatidylinositol-anchored proteins, and scavenger receptors) play a major role in DNA internalization into stem cells. We analyzed the recent literature, presenting an atlas of human HSC surface markers, for the presence of these types of proteins (Rix et al., 2022). We found that specific domains determining internalization of extracellular dsDNA fragments (clusters of positively charged amino acid residues and the heparin-binding domain) are located in the sequences of the selected proteins. The analysis results are presented in the Table. We found that several surface glycoproteins characteristic of HSCs, mainly CD34, contain domains required for internalization.

Characterization of direct molecular interaction between dsDNA *Alu*-TAMRA/telomeric repeat ($n = 9$) and HSC marker proteins CD34 and CD84 (SLAMF5). In our studies (Petrova et al., 2022; Ritter et al., 2022), we propose and confirm the hypothesis that dsDNA internalization in various stem cells is mediated by the developed glycocalyx structure on these cell membranes. The glycocalyx is composed of proteinglycans-glycoproteins, glycosylphosphatidylinositol-anchored proteins, and scavenger receptors. The interaction with these proteins is considered to have a complex physical and molecular hierarchy, and the physical contact between dsDNA and the above factors is believed to be the basis for "dragging" dsDNA into the cell.

In the current series of experiments, we attempted to assess the possibility of a direct physical interaction between the two types of molecules: dsDNA and HSC marker proteins. The following repeats were used as the dsDNA substrate: TAMRA-labeled *AluI* probe, which is commonly used in the laboratory, and a telometic repeat ($n = 9$) in the form of P^{32}-labeled 54-bp double-stranded oligonucleotide. CD34 and CD84 (SLAMF5) were selected as response factors. The main characteristics of the interaction between these proteins and dsDNA are presented in the Table. Experimental results are shown in Figure 1 and described in detail in the figure caption. In this part of the study, in a direct experiment, we first demonstrated the possibility of the chemical/molecular/physical interaction between dsDNA and specific HSC surface markers CD34 and SLAMF5.

Specific human HSC surface proteins containing domains of positively charged amino acids and the heparin-binding domain

Surface HSC markers	Name	Positively charged amino acids	Heparin/DNA-binding sites
CD90	Thy-1 membrane glycoprotein	-FSLTRETKKHVLFGTVG-	–
CD34	CD34 molecule	-LVRRGARAGPRMPRGW- -ISSKLQLMKKHQSD-	-EVRPQCLLLVLANRTE-
KIT	KIT proto-oncogene, receptor tyrosine kinase	-FLRRKRDS- -ADKRRSVRIG-	–
VNN2 (GPI-80)	Vanin 2	-EGKLVARYHKVC-	–
SPN (CD43)	Sialophorin	-LLLWRRRQKRRTGA- -FGRRKSRQGS-	-RQKRRTGALVLSRGGKRN-
CD44	CD44 molecule	-ILAVCIAVNSRRRCGQKKKLV-	–
CD9	CD9 molecule	-AIRRNREM-	–
CD48	CD48 molecule	-FESKFKGRVRLD- -GDKRPLPKEL-	–
CD84	CD84 molecule	-TTKRYNLQIYRRLGKPKITQ-	-LFKRRQGRIF- (a-helix)
ITGA6 (CD49f)	Integrin subunit alpha 6	-ESHNSRKKREI-	-TLKRQKQK- -FFKRSRYD-
GPRC5C	G-protein coupled receptor class C group 5 member C	-CGRYKRWRKHGV-	–
PROCR (EPCR)	Protein C receptor	–	–
RET	Ret proto-oncogene	-VSRRARRFA-	-ALRRPKCA-
PROM1 (CD133)	Prominin 1	-QVRTRIKRSRKLA-	-DCKKNRGT
CD59	CD59 molecule	–	–
PTPRC	Protein tyrosine phosphatase receptor type C	-DLHKKRSC- -ELRHSKRKDS-	-LRRQRCL- (a-helix)

Note. Clusters of positively charged amino acids are highlighted in green, DNA-binding sites are indicated in red, and heparin-binding sites are denoted in blue.

Demonstration of internalization of extracellular dsDNA fragments in HSCs (Sca1+ for mouse and CD34+ for human). Using fluorescence microscopy and FACS, we demonstrated the presence of labeled dsDNA probe in human CD34+ BM cells and mouse Sca1 BM cells. Mouse primitive Sca1 hematopoietic cells and human CD34+ stem cells also internalize the reference factor human recombinant angiogenin (Supplementary Material 1)[1]. Analysis of the amount of dsDNA probe delivered into human CD34+ HSCs indicates that ~0.02 % of extracellular fragments (in terms of the haploid genome) are found in the internal space of this cell type. The calculations obtained are in agreement with our numerous estimates, indicating that stem cells of various genesis, depending on their origin and state, capture ~0.01–3.0 % of extracellular dsDNA fragments (in terms of the haploid genome) (Dolgova et al., 2013, 2016, 2019; Potter et al., 2018, 2024).

[1] Supplementary Materials 1–6 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Ruzanova_Engl_28_8.pdf

We carried out a series of experiments that directly demonstrated internalization of extracellular DNA fragments in HSCs (Sca1+ for mouse and CD34+ for human) derived from BM cells (Fig. 2*A*, *B*). Molecule internalization in the cell includes the following phases: mobilization on the cytoplasmic membrane, internalization, and the presence and processing stage. In this regard, in order to avoid speculations on whether DNA molecules mobilized on the cytoplasmic membrane are detected in the experiment, we developed and applied a protocol of cell sample preparation, which is described in Supplementary Material 2.

It can be seen that original dsDNA probe molecules developed into forms containing up to 6–7 repeats (300–350 bp) of the original fragment (54 bp) (indicated with black arrows) in cells negative for both mouse and human HSC markers (Fig. 2*C*). This fact is in good agreement with our previous results (Dolgova et al., 2013; Potter et al., 2018, 2024). In addition, the presence of labeled

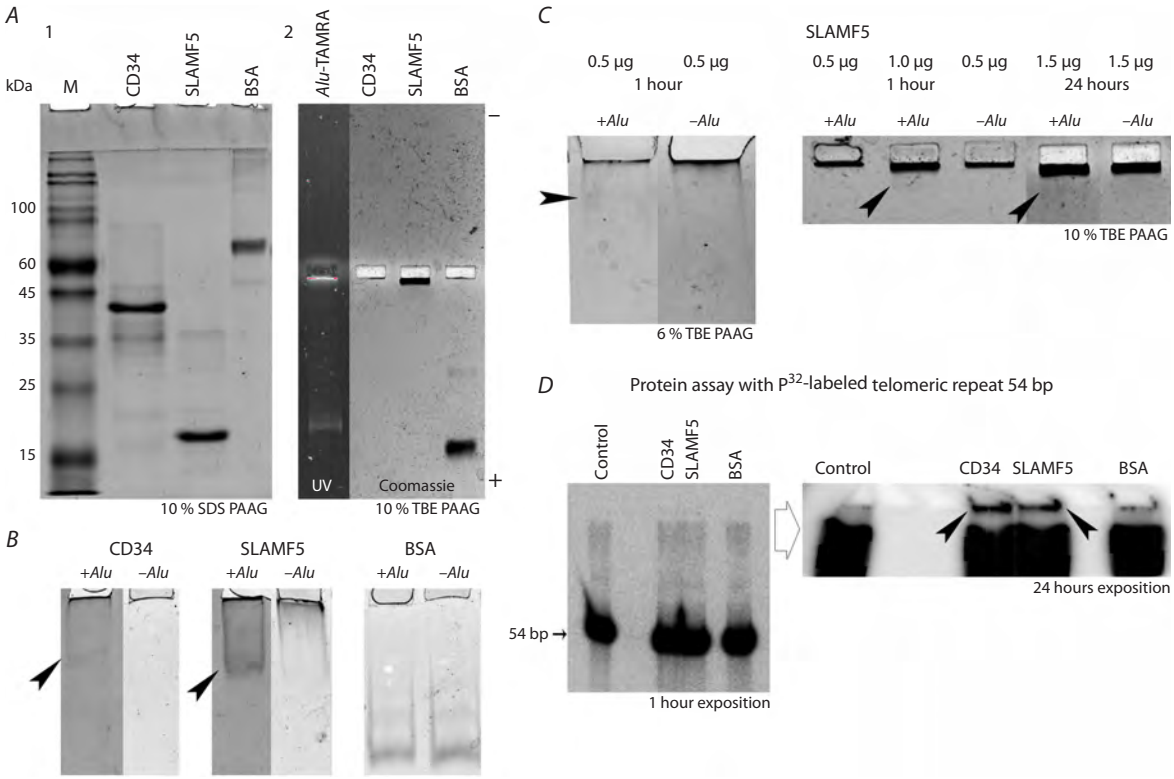


Fig. 1. Analysis of direct molecular interaction between *Alu*-TAMRA/telomeric repeat dsDNA (n = 9) and HSC marker proteins CD34 and CD84 (SLAMF5).

A – electrophoresis of analyzed factors in 10 % SDS (1) and 10 % native tris-borate horizontal (2) polyacrylamide gel. HSC markers do not have electrophoretic mobility in native conditions and thus do not enter the gel. The part of the gel with a dark field on the right panel demonstrates migration of the *AluI* dsDNA probe. *B* – change in electrophoretic mobility of factor samples after formation of complexes with the TAMRA-labeled *AluI* DNA probe. The migrating fraction of proteins (CD34 and SLAMF5) is clearly seen, which indicates that protein molecules are charged; the charge is apparently due to the DNA molecule the protein has formed a physical bond with (indicated with arrows). No changes in protein migration are detected in BSA. *C* – evaluation of some parameters of TAMRA *AluI* DNA probe-SLAMF5 complex formation. The left panel presents an electropherogram of the DNA probe-SLAMF5 complex in a native 6 % polyacrylamide gel. The amount of protein loaded on the gel is the same in control and experimental samples. The formation of a migrating protein fraction and a simultaneous decrease in its amount at the start are clearly visible. The right panel (10 % native tris-borate gel) shows the results for several modes of the DNA-SLAMF5 complex formation (indicated with arrows). It was found that the protein and DNA binding is not determined by time and the factor molar ratio. This fact indicates the absence of a stoichiometry between the TAMRA *AluI* DNA probe and SLAMF5. *D* – DNA-protein interactions between CD34, SLAMF5, and BSA using P^{32}-labeled double-stranded oligonucleotide containing 9 telomeric repeats (54 bp). Specific interactions between DNA and proteins are clearly detected in the CD34 and SLAMF5 samples (indicated with arrows).

material in the genomic DNA fraction is clearly noted in the mouse model.

The present study was not intended to provide a deep analysis of cell populations capable of capturing extracellular DNA. This study is focused exceptionally on internalization. Similar to our previous works, the study results show that CD34+ cells capture extracellular DNA. In addition, we also showed that a population of CD34– cells, which is also present in the BM, is capable of internalizing extracellular dsDNA fragments; this population may include any variants of both multipotent progenitors and committed progeny.

Terminal differentiation, HSC proliferation, and formation of colonies induced by angiogenin, hDNA^{gr}, and (angiogenin+hDNA^{gr})

Deproteinized human genomic dsDNA fragmented to 1–10 nucleosome monomers, namely hDNA^{gr}, or genome reconstructor, was used in the study. The length of 1–10 nucleosome monomers is the physiological size of DNA molecules (self-DNA) in apoptotic cells, which are always present in the peripheral blood. The inducer human recombinant angiogenin was used as a comparison factor.

We performed a series of experiments on analysis of the stimulation of colony formation and proliferative activity of BM HSCs after treatment with the selected inducers in three models: mouse BM cells, rat BM cells, and cryopreserved human BM cells. We found that cell treatment with angiogenin, hDNA^{gr}, and angiogenin+hDNA^{gr} stimulates colony formation (an increase in the total number) in the studied models. The number of new colonies in mouse and human models in some cases increased by 20–30 % when using hDNA^{gr} (Supplementary Material 3, Fig. 1*A*, *C*). A significant increase in the number of colonies was noted in the mouse model after treatment with both angiogenin and angiogenin+hDNA^{gr}.

Angiogenin reliably stimulates cell proliferation in growing colonies in the mouse model. CFU–GM is the main

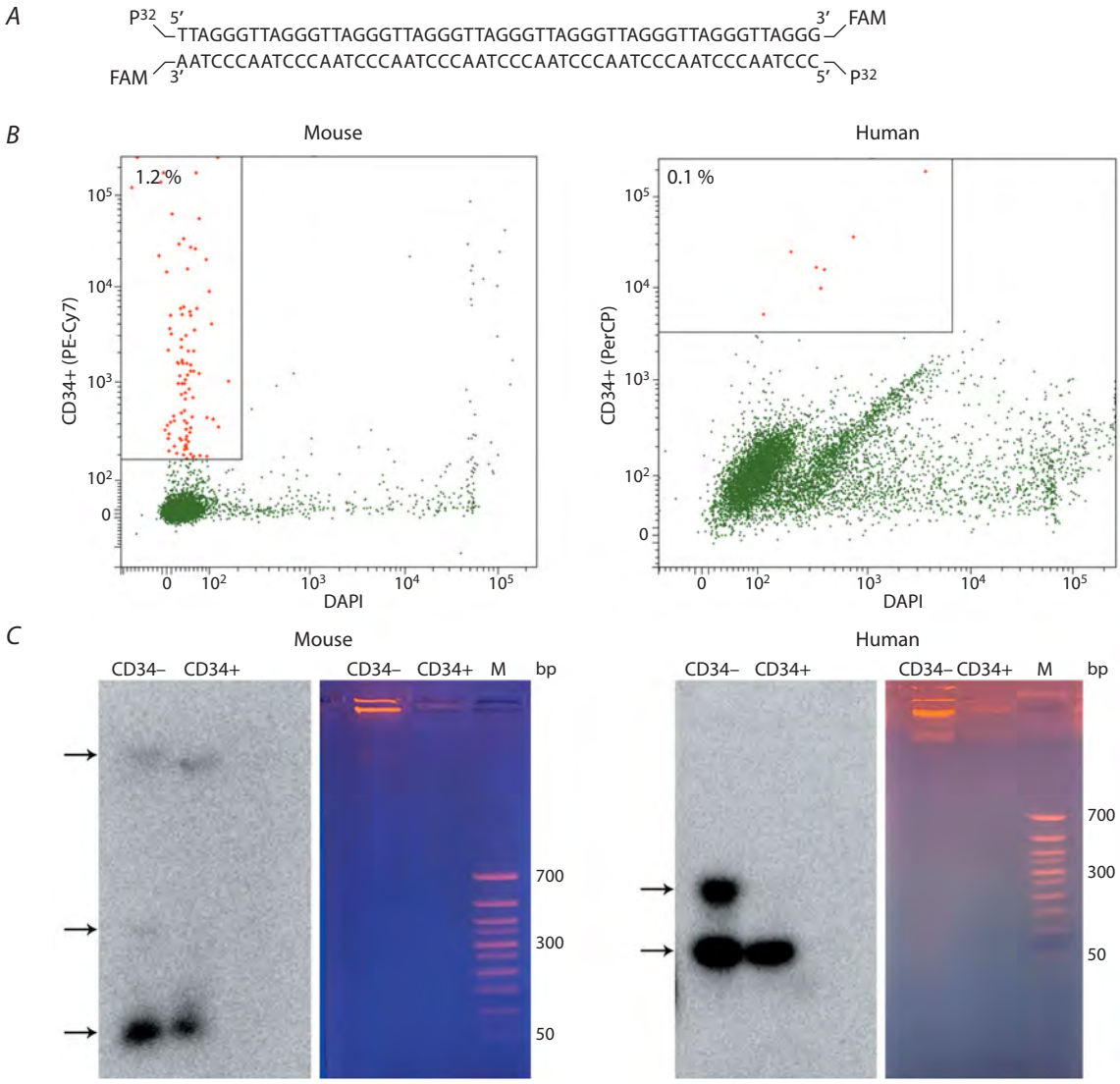


Fig. 2. Direct experiment on internalization of extracellular dsDNA fragments.

A – DNA probe structure; *B* – FACS analysis of mouse and human BM cell samples using the CD34 marker; *C* – Gel electrophoresis and autoradiography of DNA found in the internal compartments of sorted mouse and human HSCs. Arrows indicate bands corresponding to the DNA probe, concatemeric (circle?) form, and genomic DNA label.

responsive lineage, which is reliably confirmed in the human model. Treatment of BM cells with activators neither induces apoptosis nor stimulates CD34+ cell survival. Addition of hDNA^{gr} and angiogenin+hDNA^{gr} to freshly thawed human samples enhances CD34+ cell proliferation. At the same time, angiogenin neither shows any stimulatory effect nor affects the ability of hDNA^{gr} to enhance CD34+ cell proliferation (Supplementary Material 3, Fig. 1).

A comparison was also made of the proliferative activity of CD34+ HSCs for the synthesis of the proliferative factor Ki-67 after treatment with inducers before seeding on methylcellulose and the proliferative activity of these cells, expressed in the number of cells per colony after incubation on methylcellulose for 11–15 days. No correlation was found between the two parameters (Supplementary Material 3, Fig. 2).

Assessment of the ability of colony cells selected on days 7 and 15 of culturing in methylcellulose to internalize a TAMRA-labeled 500 bp PCR fragment

The main keynote of all our studies is the confirmed statement that extracellular DNA fragments are captured by primitive stem progenitors. In humans, these cells are CD34+ progenitors. In the study performed in a mouse model (Potter et al., 2024), we showed that the number of primitive hematopoietic progenitors increases significantly in colonies formed after induction of terminal differentiation by extracellular dsDNA fragments. This makes it possible to use these progenitors to analyze various events occurring in HSCs, which is impossible in case of BM HSCs.

A similar study was conducted in a human cryopreserved BM cell model. We estimated the percentage of CD34+

stem cells in colonies formed by HSCs after their induction in BM by angiogenin, hDNA^{gr}, and angiogenin+hDNA^{gr}. Treatment of HSCs in BM by hDNA^{gr} on day 15 of culturing resulted in an increase in the number of cells in the colony to 2.7 % versus 1.56 % in an individual experiment (GM-CSF-stimulated BM cells). This indicates that colonies contain a sufficient number of cells able to internalize extracellular genetic material in an amount required for reliable detection of extracellular DNA in the cell. At the same time, neither angiogenin nor angiogenin+hDNA^{gr} increased the number of hematopoietic precursors in colonies (Supplementary Material 4).

Analysis of formation of pangenomic single-strand breaks in the cells of colonies of primitive progenitor descendants treated by hDNA^{gr} as part of BM cells

Early studies analyzed in the Introduction section showed that the genome of embryonic stem cells is exposed by pangenomic single-strand breaks during commitment upon induction of terminal differentiation. These single-strand breaks are repaired without causing cell death. We believe that this process is important for the change in chromatin architecture characterizing undifferentiated blood stem cells to the spatial organization of expressing genes characteristic of committed progeny (Jacobson et al., 1975; Scher, Friend, 1978; Farzaneh et al., 1982; McMahon et al., 1984; Boerrigter et al., 1989; Kaminskas, Li, 1989; Vatolin et al., 1997).

We hypothesized that this process is common for all types of primitive progenitors, including HSCs. The analysis performed in the first part of our study and in the work (Potter et al., 2024) demonstrated that the selected inducers cause colony formation and terminal differentiation of activated BM HSCs in mice, rats, and humans. This means that formation of pangenomic single-strand breaks may also be an integral part of HSC biology. The content of HSC colonies in mice was 12–15 % (Potter et al., 2024). In human, the cell content is ~3 % (Supplementary Material 4). This indicates that there will be a sufficient number of cells retaining the undifferentiated state and undergoing terminal differentiation in the colony formed by BM HSCs after a single induction of BM cells and repeated induction of colony cells on day 15 for identification of single-strand breaks.

The work was performed in mouse and human models using the following inducers: hDNA^{gr}, angiogenin, and angiogenin+hDNA^{gr}. We also quantified single-strand breaks in the DNA of colony cells on day 15 after all the procedures described above.

The analysis revealed significant and reliable differences in the studied parameters between different sample and control points (Fig. 3). An increase in the number of cells with the maximum level of tail DNA after 72–96 hrs and 96 hrs of incubation of hDNA^{gr}-treated cells was noted in the human and mouse models, respectively. The use of angiogenin alone has virtually no effect on the induction of single-strand breaks and increase in the tail DNA content. Apparently, complete repair of single-strand breaks takes place on days 7–9 of incubation in the human model (Supplementary Material 5).

The obtained results on changes in comet tail lengths at specific time points made it possible to estimate the approximate number of induced pangenomic single-strand breaks (Fig. 4).

Several assumptions were made to estimate the number of single-strand breaks. One DNA strand of a chromosome was considered to break as a nick by forming two equal parts. Any other scenario required the use of a powerful mathematical framework, which did not correspond to the study goals. The smallest chromosome size is $\sim 50 \times 10^6$ bp. In this regard, we calculated the number of breaks based on this length. This simplest scenario suggested that, if the DNA strand breaks into two equal parts forming a nick (alkaline conditions), then the length of the tail formed by one strand decreases by half. In case there are two nicks, each of the previous parts decreases by another half, etc. That is, if the tail length is considered 10 in scale units at the first point, it corresponds to either the absence of breaks or their native number. In that case, the tail length twice as long (20) corresponds to the formation of one break per the initial molecule length (chromosome). Thus, transfer to the next interval requires all DNA fragments formed at the previous stage to have another break. Hence, the number of breaks is estimated using the formula $2n + 1$, where n is the number of breaks for the previous interval. The box thickness on the graph shows the number of cells in the specific interval. The number of breaks calculated for the interval was multiplied by the number of cells in the same interval. The average number of breaks per cell was calculated for the specified time point. Based on these data, a graph of the change in the number of breaks depending on time was constructed.

The conducted analysis demonstrated that, using the above calculation protocol, the maximum number of single-strand breaks is ~2.5–3.5 nicks per 5×10^6 chromatin bp and takes place at the time point of 72–96 hrs (for two independent experiments). The number of nicks in the control sample is in the range of 1.0–1.5 nicks per 5×10^6 chromatin bp (Fig. 4).

In a sample treated with angiogenin, a slightly higher number of nicks compared to the control sample can be detected in cells at the time point of the maximum chromatin perturbation. This does not contradict the results on colony stimulation, which demonstrate a positive effect of angiogenin on the formation of several types of colonies.

Discussion

The discovered fact of dsDNA fragment internalization in HSCs with subsequent induction of terminal differentiation and colony formation suggested that, similar to embryonic stem cells (Vatolin et al., 1997), single-strand breaks are also induced in hematopoietic stem cells at the

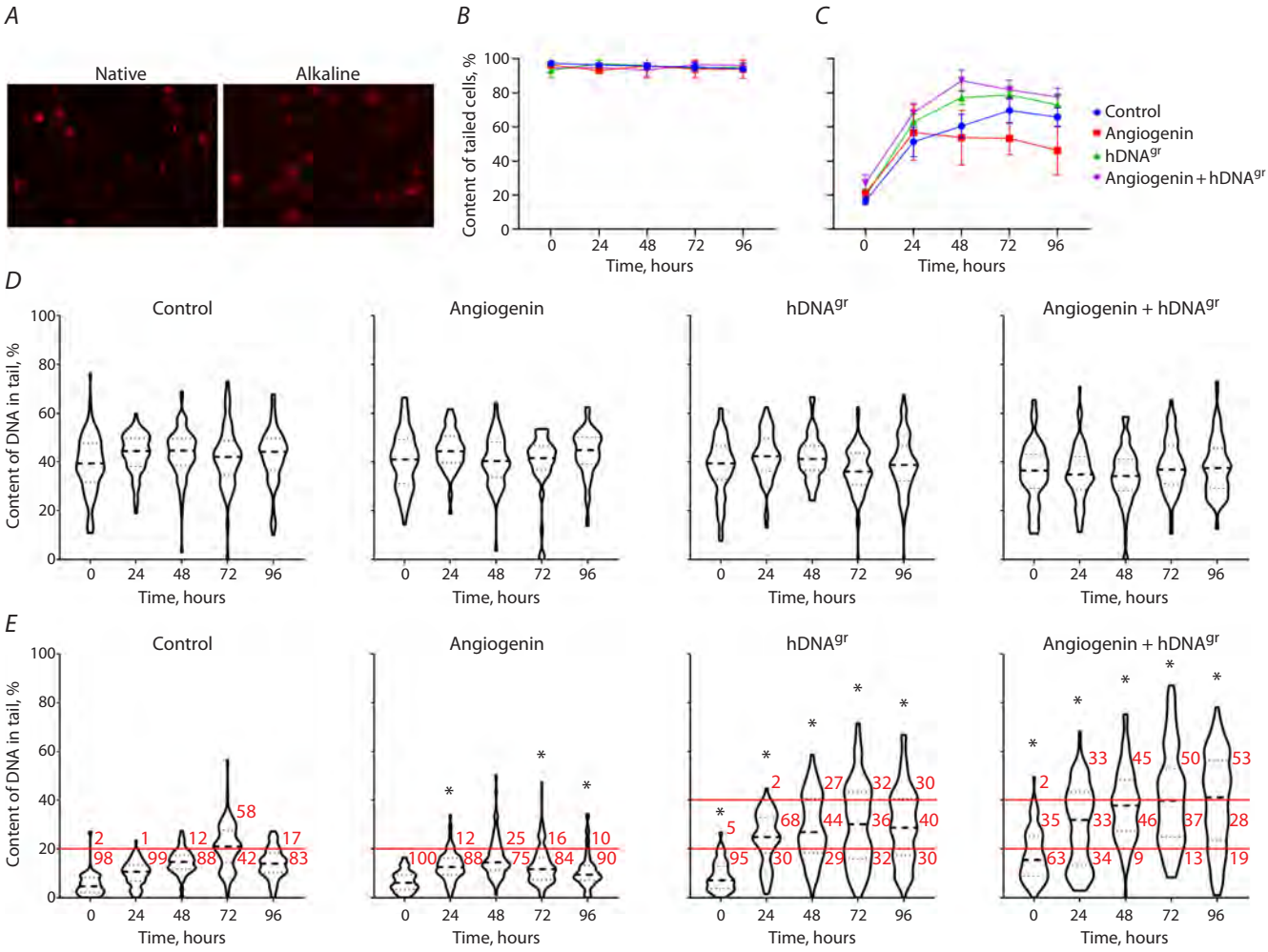


Fig. 3. Human model. *A* – cells and comet tails in native and alkaline electrophoresis. *B*, *C* – content of cells with a tail in native (*B*) and alkaline (*C*) electrophoresis. *D*, *E* – diagrams showing the number of cells with different tail DNA levels in native (*D*) and alkaline (*E*) electrophoresis.

The bold dashed line indicates the median value, the thin dashed line shows the interquartile range. The percentage of cells with the tail DNA level of 0–20 %, 20–40 % and >40 % is indicated in red (the corresponding ranges are highlighted with red lines). * Significant differences compared to the control group, $p < 0.01$, Mann–Whitney test.

state of terminal differentiation. The analysis performed in the two selected models indicated a similar biological phenomenon in HSCs. Pangenomic single-strand breaks are formed, developed, and repaired in HSCs at the phase of terminal differentiation. Together with the experimental data presented in the literature, the obtained result indicates that this is a general biological process. Pangenomic single-strand breaks are a necessary condition for reorientation of the activity of gene platforms determining the undifferentiated state to gene platforms characteristic of the committed cell state.

For the past two decades, the main attention of researchers was focused on double-strand breaks and the variety of processes associated with their formation, as well as repair and recombination events mediated by these breaks in cells (So et al., 2017). Nevertheless, the scientific community has renewed its interest in nicks, or single-strand chromosome breaks, in the past years, as shown in some reviews (Xu, 2015; Vriend, Krawczyk, 2017; Maizels, Davis, 2018; Zilio, Ulrich, 2021). The keynote of the new surge of interest in nicked chromatin DNA is the forgotten concept of nick-initiated homologous recombination. The performed analysis indicates that nicks are no less important as intermediates of chromatin DNA metabolism, inducing repair and recombination processes in the cell, than double-strand breaks. However, unlike double-strand breaks, repair of single-strand breaks (nicks) much less frequently leads to fatal changes in the genome structure. Homologous recombination is the main mechanism of single-strand break repair.

The above indicates that single-strand breaks are inducers of recombinogenic state of the cell. The idea of the recombinogenic state is most fully described in our pioneering review (Likhacheva et al., 2008). The term “recombinogenic state” characterizes the activity of the cell molecular machine launched by a change in the higher-order chromatin architecture. Single-strand breaks are one of the inducers of such a change.

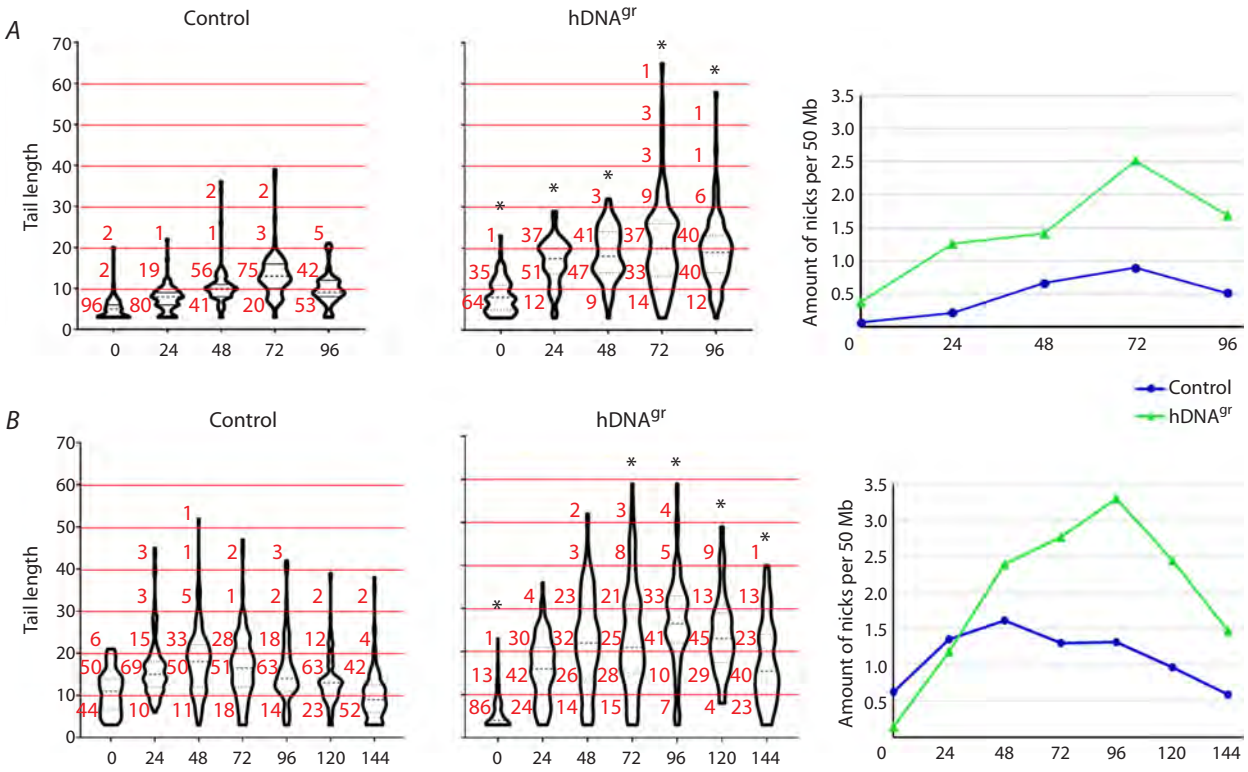


Fig. 4. Results of two independent experiments in the human model (*A*, *B*).

Diagrams for control and hDNA^{gr}-treated cells are presented; they show the comet tail length in arbitrary units (*Y* axis) and time intervals with a 24-h step (*X* axis). The percentage of cells with the comet tail length within the corresponding interval is shown in red. * Reliable differences compared to the control group, $p < 0.01$, Mann–Whitney test. Graphs on the right show dependence of the calculated number of nicks per 50×10^6 bp (*Y* axis) on the time interval (*X* axis).

The main thesis in the review is that, if there are internalized extracellular dsDNA fragments in the cell in the activated recombinogenic state, these fragments become natural participants in the repair-recombination process activated by molecular mechanisms. This means that these fragments can participate in the recombination process as a natural recombination substrate. Hence, a general biological mechanism explaining the presence of extrachromosomal genetic information in the recipient genome as a result of either direct homologous integration of extracellular dsDNA fragments or formation of stable, genetically active extrachromosomal complexes has been found.

We characterized two phenomena with the involvement of dsDNA fragments in the repair-recombination process in our studies. In the work (Likhacheva et al., 2007), we demonstrated the participation of exogenous human DNA in the rescue of mouse BM progenitors from a lethal dose of gamma radiation, resulting in the survival of experimental animals. The mechanism of HSC rescue is associated with internalization of dsDNA fragments into the blood stem cell and repair-recombination correction of double-strand chromatin breaks induced by severe irradiation. In a series of other studies, we showed the involvement of extracellular dsDNA in suppressing the repair of interstrand cross-links in tumor stem cells. The outcome of this participation is inability of the tumor stem cell to complete the repair of cytostatic-induced chromatin damage resulting in its further apoptotic death (Ruzanova et al., 2022). Numerous other studies indicate that single-strand breaks induce homologous recombination of the genetic material in the cell nucleus (Vriend, Krawczyk, 2017; Maizels, Davis, 2018).

Conclusion

Thus, extracellular dsDNA fragments are internalized in HSCs through a natural mechanism, induce terminal differentiation of blood stem cells, and stimulate colony formation. Pangenomic single-strand breaks are the molecular manifestation of these processes. The formation of pangenomic single-strand breaks induces the recombinogenic state of the blood stem cell. During this process, extracellular dsDNA fragments can integrate into the recipient HSC genome. From a theoretical standpoint, a series of integration scenarios are possible: the ends-in/ends-out mechanism, reciprocal homologous recombination, gene conversion or single-strand annealing, and non-homologous integration (Rubnitz, Subramani, 1984; Hastings et al., 1993; Li et al., 2001; Langston, Symington, 2004; Chen J.M. et al., 2007; Rass et al., 2012).

In the following parts of our research series, we present experimental evidence of both integration of extracellular dsDNA fragments into the HSC genome and formation of

circular structures complexing with chromosomal DNA preserved under sever fractionation conditions. Comments on events associated with HSC terminal differentiation after extracellular dsDNA internalization are presented in Supplementary Material 6. In addition, an apparent discrepancy with the flow cytometry data, indicating that CD34+ HSCs do not disappear but, on the contrary, increase their number in colonies compared to the original BM cell sample, is discussed (Supplementary Material 4).

References

Boerrigter M.E.T.I., Mullaart E., Van Der Schans G.P., Vijg J. Quiescent human peripheral blood lymphocytes do not contain a sizable amount of preexistent DNA single-strand breaks. *Exp. Cell Res*. 1989;180(2):569-573. doi 10.1016/0014-4827(89)90085-2

Chen J.M., Cooper D.N., Chuzhanova N., Férec C., Patrinos G.P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet*. 2007;8(10):762-775. doi 10.1038/NRG2193

Chen T.L., Chiang Y.W., Lin G.L., Chang H.H., Lien T.S., Sheh M.H., Sun D.S. Different effects of granulocyte colony-stimulating factor and erythropoietin on erythropoiesis. *Stem. Cell Res. Ther*. 2018; 9(1):119. doi 10.1186/S13287-018-0877-2

Desterke C., Bennaceur-Griscelli A., Turhan A.G. EGR1 dysregulation defines an inflammatory and leukemic program in cell trajectory of human-aged hematopoietic stem cells (HSC). *Stem. Cell Res. Ther.* 2021;12(1):419. doi 10.1186/S13287-021-02498-0

DNA Cloning. A practical approach. Ed. D.M. Glover. IRL Press, 1985

Dolgova E.V., Efremov Y.R., Orishchenko K.E., Andrushkevich O.M., Alyamkina E.A., Proskurina A.S., Bayborodin S.I., Nikolin V.P., Popova N.A., Chernykh E.R., Ostanin A.A., Taranov O.S., Omigov V.V., Minkevich A.M., Rogachev V.A., Bogachev S.S., Shurdov M.A. Delivery and processing of exogenous double-stranded DNA in mouse CD34+ hematopoietic progenitor cells and their cell cycle changes upon combined treatment with cyclophosphamide and double-stranded DNA. *Gene*. 2013;528(2):74-83. doi 10.1016/j.gene.2013.06.058

Dolgova E.V., Alyamkina E.A., Efremov Y.R., Nikolin V.P., Popova N.A., Tyrinova T.V., Kozel A.V., Minkevich A.M., Andrushkevich O.M., Zavyalov E.L., Romaschenko A.V., Bayborodin S.I., Taranov O.S., Omigov V.V., Shevela E.Y., Stupak V.V., Mishinov S.V., Rogachev V.A., Proskurina A.S., Mayorov V.I., Shurdov M.A., Ostanin A.A., Chernykh E.R., Bogachev S.S. Identification of cancer stem cells and a strategy for their elimination. *Cancer Biol. Ther*. 2014;15(10):1378-1394. doi 10.4161/cbt.29854

Dolgova E.V., Potter E.A., Proskurina A.S., Minkevich A.M., Chernych E.R., Ostanin A.A., Efremov Y.R., Bayborodin S.I., Nikolin V.P., Popova N.A., Kolchanov N.A., Bogachev S.S. Properties of internalization factors contributing to the uptake of extracellular DNA into tumor-initiating stem cells of mouse Krebs-2 cell line. *Stem. Cell Res. Ther*. 2016;7(1):76. doi 10.1186/s13287-016-0338-8

Dolgova E.V., Petrova D.D., Proskurina A.S., Ritter G.S., Kisaretova P.E., Potter E.A., Efremov Y.R., Bayborodin S.I., Karamysheva T.V., Romanenko M.V., Netesov S.V., Taranov O.S., Ostanin A.A., Chernykh E.R., Bogachev S.S. Identification of the xenograft and its ascendant sphere-forming cell line as belonging to EBV-induced lymphoma, and characterization of the status of sphere-forming cells. *Cancer Cell Int*. 2019;19:120. doi 10.1186/S12935-019-0842-X

Farzaneh F., Zalin R., Brill D., Shall S. DNA strand breaks and ADP-ribosyl transferase activation during cell differentiation. *Nature*. 1982;300(5890):362-366. doi 10.1038/300362A0

Forristal C.E., Levesque J.-P. Targeting the hypoxia-sensing pathway in clinical hematology. *Stem Cells Transl. Med*. 2014;3(2):135-140. doi 10.5966/SCTM.2013-0134

Goncalves K.A., Silberstein L., Li S., Severe N., Hu M.G., Yang H., Scadden D.T., Hu G.F. Angiogenin promotes hematopoietic regeneration by dichotomously regulating quiescence of stem and progenitor cells. *Cell*. 2016;166(4):894-906. doi 10.1016/J.CELL.2016.06.042

Hastings P.J., McGill C., Shafer B., Strathern J.N. Ends-in *vs*. ends-out recombination in yeast. *Genetics*. 1993;135(4):973-980. doi 10.1093/GENETICS/135.4.973

Iseki S. DNA strand breaks in rat tissues as detected by in situ nick translation. *Exp. Cell Res*. 1986;167(2):311-326. doi 10.1016/0014-4827(86)90172-2

Jacobson G.K., Pinon R., Esposito R.E., Esposito M.S. Single-strand scissions of chromosomal DNA during commitment to recombination at meiosis. *Proc. Natl. Acad. Sci. USA*. 1975;72(5):1887-1891. doi 10.1073/PNAS.72.5.1887

Jiang N., Pisetsky D.S. The effect of inflammation on the generation of plasma DNA from dead and dying cells in the peritoneum. *J. Leukoc. Biol*. 2005;77(3):296-302. doi 10.1189/JLB.0704411

Johnstone A.P., Williams G.T. Role of DNA breaks and ADP-ribosyl transferase activity in eukaryotic differentiation demonstrated in human lymphocytes. *Nature*. 1982;300(5890):368-370. doi 10.1038/300368A0

Kaminskas E., Li J.C. DNA fragmentation in permeabilized cells and nuclei. The role of (Ca^{2+} + Mg^{2+})-dependent endodeoxyribonuclease. *Biochem. J*. 1989;261(1):17-21. doi 10.1042/BJ2610017

Kananen L., Hurme M., Bürkle A., Moreno-Villanueva M., Bernhardt J., Debacq-Chainiaux F., Grubeck-Loebenstein B., Malavolta M., Basso A., Piacenza F., Collino S., Gonos E.S., Sikora E., Gradinaru D., Jansen E.H.J.M., Dollé M.E.T., Salmon M., Stuetz W., Weber D., Grune T., Breusing N., Simm A., Capri M., Franceschi C., Slagboom E., Talbot D., Libert C., Raitanen J., Koskinen S., Härkänen T., Stenholm S., Ala-Korpela M., Lehtimäki T., Raitakari O.T., Ukkola O., Kähönen M., Jylhä M., Jylhävä J. Circulating cell-free DNA in health and disease – the relationship to health behaviours, ageing phenotypes and metabolomics. *GeroScience*. 2023;45(1):85-103. doi 10.1007/S11357-022-00590-8

Kiang J.G., Zhai M., Lin B., Smith J.T., Anderson M.N., Jiang S. Co-therapy of pegylated G-CSF and ghrelin for enhancing survival after exposure to lethal radiation. *Front. Pharmacol*. 2021;12:628018. doi 10.3389/FPHAR.2021.628018

Korabecna M., Zinkova A., Brynychova I., Chylikova B., Prikryl P., Sedova L., Neuzil P., Seda O. Cell-free DNA in plasma as an essential immune system regulator. *Sci. Rep*. 2020;10(1):17478. doi 10.1038/S41598-020-74288-2

Kovtonyuk L.V., Fritsch K., Feng X., Manz M.G., Takizawa H. Inflamm-aging of hematopoiesis, hematopoietic stem cells, and the bone marrow microenvironment. *Front. Immunol*. 2016;7:502. doi 10.3389/FIMMU.2016.00502

Kulkarni R., Kale V. Physiological cues involved in the regulation of adhesion mechanisms in hematopoietic stem cell fate decision. *Front. Cell Dev. Biol*. 2020;8:611. doi 10.3389/FCELL.2020.00611

Kumar S., Geiger H. HSC niche biology and HSC expansion *ex vivo*. *Trends Mol. Med*. 2017;23(9):799. doi 10.1016/J.MOLMED.2017.07.003

Langston L.D., Symington L.S. Gene targeting in yeast is initiated by two independent strand invasions. *Proc. Natl. Acad. Sci. USA*. 2004;101(43):15392-15397. doi 10.1073/PNAS.0403748101

Lauková L., Bertolo E.M.J., Zelinková M., Borbélyová V., Čonka J., Gaál Kovalčíková A., Domonkos E., Vlková B., Celec P. Early dynamics of plasma DNA in a mouse model of sepsis. *Shock*. 2019;52(2):257-263. doi 10.1097/SHK.0000000000001215

Lévesque J.P., Helwani F.M., Winkler I.G. The endosteal 'osteoblastic' niche and its role in hematopoietic stem cell homing and mobilization. *Leukemia.* 2010;24(12):1979-1992. doi 10.1038/leu.2010.214

Li J., Read L.R., Baker M.D. The mechanism of mammalian gene replacement is consistent with the formation of long regions of heteroduplex DNA associated with two crossing-over events. *Mol. Cell. Biol*. 2001;21(2):501-510. doi 10.1128/MCB.21.2.501-510.2001

Likhacheva A.S., Nikolin V.P., Popova N.A., Rogachev V.A., Prokhorovich M.A., Sebeleva T.E., Bogachev S.S., Shurdov M.A. Exogenous DNA can be captured by stem cells and be involved in their rescue from death after lethal-dose γ-radiation. *Gene Therapy Mol. Biol*. 2007;11:305-314

Likhacheva A.S., Rogachev V.A., Nikolin V.P., Popova N.A., Shilov A.G., Sebeleva T.E., Strunkin D.N., Chernykh E.R., Gel'fgat E.L., Bogachev S.S., Shurdov M.A. Involvement of exogenous DNA in the molecular processes in somatic cell. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2008;12(3):426-473 (in Russian)

Lucas D. Leukocyte trafficking and regulation of murine hematopoietic stem cells and their niches. *Front. Immunol*. 2019;10:387. doi 10.3389/FIMMU.2019.00387/BIBTEX

Maizels N., Davis L. Initiation of homologous recombination at DNA nicks. *Nucleic Acids Res*. 2018;46:6962-6973. doi 10.1093/NAR/GKY588

Maniatis T., Fritch E., Sambrook D. Methods of Genetic Engineering. Molecular Cloning. Moscow: Mir Publ., 1984 (in Russian)

McMahon G., Alsina J.L., Levy S.B. Induction of a Ca^{2+}, Mg^{2+}-dependent endonuclease activity during the early stages of murine erythroleukemic cell differentiation. *Proc. Natl. Acad. Sci. USA.* 1984;81(23):7461-7465. doi 10.1073/PNAS.81.23.7461

Mendelson A., Frenette P.S. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nat. Med*. 2014;20(8):833-846. doi 10.1038/NM.3647

Morita Y., Ema H., Nakauchi H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med*. 2010;207(6):1173-1182. doi 10.1084/JEM.20091318

Muller-Sieburg C., Sieburg H.B. Stem cell aging: survival of the laziest? *Cell Cycle*. 2008;7(24):3798-3804. doi 10.4161/CC.7.24.7214

Patkin E.L., Kustova M.E., Noniashvili E.M. DNA-strand breaks in chromosomes of early mouse embryos as detected by *in situ* nick translation and gap filling. *Genome*. 1995;38:381-384. doi 10.1139/G95-049

Petrova D.D., Dolgova E.V., Proskurina A.S., Ritter G.S., Ruzanova V.S., Efremov Y.R., Potter E.A., Kirikovich S.S., Levites E.V., Taranov O.S., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. The new general biological property of stem-like tumor cells (Part II: Surface molecules, which belongs to distinctive groups with particular functions, form a unique pattern characteristic of a certain type of tumor stem-like cells). *Int. J. Mol. Sci.* 2022;23(24):15800. doi 10.3390/ijms232415800

Pierce H., Zhang D., Magnon C., Lucas D., Christin J.R., Huggins M., Schwartz G.J., Frenette P.S. Cholinergic signals from the CNS regulate G-CSF-mediated HSC mobilization from bone marrow via a glucocorticoid signaling relay. *Cell Stem. Cell*. 2017;20:648-658.e4. doi 10.1016/J.STEM.2017.01.002

Pinho S., Frenette P.S. Haematopoietic stem cell activity and interactions with the niche. *Nat. Rev. Mol. Cell Biol*. 2019;20(5):303-320. doi 10.1038/S41580-019-0103-9

Potter E.A., Proskurina A.S., Ritter G.S., Dolgova E.V., Nikolin V.P., Popova N.A., Taranov O.S., Efremov Y.R., Bayborodin S.I., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. Efficacy of a new cancer treatment strategy based on eradication of tumor-initiating stem cells in a mouse model of Krebs-2 solid adenocarcinoma. *Oncotarget*. 2018;9(47):28486-28499. doi 10.18632/oncotarget.25503

Potter E.A., Dolgova E.V., Proskurina A.S., Ruzanova V.S., Efremov Y.R., Kirikovich S.S., Oshikhmina S.G., Mamaev A.L., Taranov O.S., Bryukhovetskiy A.S., Grivtsova L.U., Kolchanov N.A., Ostanin A.A., Chernykh E.R., Bogachev S.S. Stimulation of mouse hematopoietic stem cells by angiogenin and DNA preparations. *Braz. J. Med. Biol. Res*. 2024;57:e13072. doi 10.1590/1414-431X2024E13072

Pulito V.L., Miller D.L., Sassa S., Yamane T. DNA fragments in Friend erythroleukemia cells induced by dimethyl sulfoxide. *Proc. Natl. Acad. Sci. USA*. 1983;80(19):5912-5915. doi 10.1073/PNAS.80.19.5912

Rass E., Grabarz A., Bertrand P., Lopez B.S. Double strand break repair, one mechanism can hide another: alternative non-homologous end joining. *Cancer Radiother.* 2012;16:1-10. doi 10.1016/J.CANRAD.2011.05.004

Redondo P.A., Pavlou M., Loizidou M., Cheema U. Elements of the niche for adult stem cell expansion. *J. Tissue Eng*. 2017;8:2041731417725464. doi 10.1177/2041731417725464

Ritter G.S., Dolgova E.V., Petrova D.D., Efremov Y.R., Proskurina A.S., Potter E.A., Ruzanova V.S., Kirikovich S.S., Levites E.V., Taranov O.S., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. The new general biological property of stem-like tumor cells. Part I. Peculiarities of the process of the double-stranded DNA fragments internalization into stem-like tumor cells. *Front. Genetics*. 2022;13:954395. doi 10.3389/fgene.2022.954395

Rix B., Maduro A.H., Bridge K.S., Grey W. Markers for human haematopoietic stem cells: the disconnect between an identification marker and its function. *Front. Physiol*. 2022;13. doi 10.3389/FPHYS.2022.1009160

Rubnitz J., Subramani S. The minimum amount of homology required for homologous recombination in mammalian cells. *Mol. Cell. Biol*. 1984;4(11):2253-2258. doi 10.1128/MCB.4.11.2253-2258.1984

Ruzanova V., Proskurina A., Efremov Y., Kirikovich S., Ritter G., Levites E., Dolgova E., Potter E., Babaeva O., Sidorov S., Taranov O., Ostanin A., Chernykh E., Bogachev S. Chronometric administration of cyclophosphamide and a double-stranded DNA-Mix at interstrand crosslinks repair timing, called "Karanahan" therapy, is highly efficient in a weakly immunogenic Lewis carcinoma model. *Pathol. Oncol. Res*. 2022;28. doi 10.3389/PORE.2022.1610180

Saitoh T., Fujita N., Yoshimori T., Akira S. Regulation of dsDNA-induced innate immune responses by membrane trafficking. *Autophagy*. 2010;6:430-432. doi 10.4161/AUTO.6.3.11611

Scharf P., Broering M.F., da Rocha G.H.O., Farsky S.H.P. Cellular and molecular mechanisms of environmental pollutants on hematopoiesis. *Int. J. Mol. Sci*. 2020;21(19):6996. doi 10.3390/IJMS21196996

Scher W., Friend C. Breakage of DNA and alterations in folded genomes by inducers of differentiation in Friend erythroleukemic cells. *Cancer Res*. 1978;38:841-849

Seita J., Weissman I.L. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med*. 2010;2(6):640-653. doi 10.1002/WSBM.86

Silberstein L., Goncalves K.A., Kharchenko P.V., Turcotte R., Kfoury Y., Mercier F., Baryawno N., Severe N., Bachand J., Spencer J.A., Papazian A., Lee D., Chitteti B.R., Srour E.F., Hoggatt J., Tate T., Lo Celso C., Ono N., Nutt S., Heino J., Sipilä K., Shioda T., Osawa M., Lin C.P., Hu G.-fu, Scadden D.T. Proximity-based differential single-cell analysis of the niche to identify stem/progenitor cell regulators. *Cell Stem Cell*. 2016;19(4):530-543. doi 10.1016/J.STEM.2016.07.004

So A., Le Guen T., Lopez B.S., Guirouilh-Barbat J. Genomic rearrangements induced by unscheduled DNA double strand breaks in somatic

mammalian cells. *FEBS J*. 2017;284(15):2324-2344. doi 10.1111/FEBS.14053

Szade K., Gulati G.S., Chan C.K.F., Kao K.S., Miyanishi M., Marjon K.D., Sinha R., George B.M., Chen J.Y., Weissman I.L. Where hematopoietic stem cells live: the bone marrow niche. *Antioxid. Redox Signal*. 2018;29:191. doi 10.1089/ARS.2017.7419

Vatolin S.Y., Okhapkina E.V., Matveeva N.M., Shilov A.G., Baiborodin S.I., Philimonenko V.V., Zhdanova N.S., Serov O.L. Scheduled perturbation in DNA during in vitro differentiation of mouse embryo-derived cells. *Mol. Reprod. Dev.* 1997;47(1):1-10. doi 10.1002/(SICI)1098-2795(199705)47:1<1::AID-MRD1>3.0.CO;2-R

Vriend L.E.M., Krawczyk P.M. Nick-initiated homologous recombination: protecting the genome, one strand at a time. *DNA Repair*. 2017;50:1-13. doi 10.1016/J.DNAREP.2016.12.005

Wang S., Zhang Y., Meng W., Dong Y., Zhang S., Teng L., Liu Y., Li L., Wang D. The involvement of macrophage colony stimulating factor on protein hydrolysate injection mediated hematopoietic function improvement. *Cells*. 2021;10(10):2776. doi 10.3390/CELLS10102776

Wilkinson A.C., Igarashi K.J., Nakauchi H. Haematopoietic stem cell self-renewal in vivo and ex vivo. *Nat. Rev. Genet.* 2020;21(9):541-554. doi 10.1038/s41576-020-0241-0

Winkler I.G., Barbier V., Nowlan B., Jacobsen R.N., Forristal C.E., Patton J.T., Magnani J.L., Lévesque J.P. Vascular niche E-selectin regulates hematopoietic stem cell dormancy, self renewal and chemoresistance. *Nat. Med*. 2012;18(11):1651-1657. doi 10.1038/NM.2969

Xu S.Y. Sequence-specific DNA nicking endonucleases. *Biomol. Concepts*. 2015;6(4):253-267. doi 10.1515/BMC-2015-0016

Zhang C.C., Sadek H.A. Hypoxia and metabolic properties of hematopoietic stem cells. *Antioxid. Redox. Signal*. 2014;20(12):1891-1901. doi 10.1089/ARS.2012.5019

Zilio N., Ulrich H.D. Exploring the SSBreakome: genome-wide mapping of DNA single-strand breaks by next-generation sequencing. *FEBS J*. 2021;288(13):3948-3961. doi 10.1111/FEBS.15568

Conflict of interest. The authors declare no conflict of interest.
Received June 26, 2024. Revised September 19, 2024. Accepted September 26, 2024.

BIOMEDICINE
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):1008-1017

doi 10.18699/vjgb-24-107

GlucoGenes®, a database of genes and proteins associated with glucose metabolism disorders, its description and applications in bioinformatics research

V.V. Klimontov [1] ✉, K.S. Shishin [1], R.A. Ivanov [2], M.P. Ponomarenko [2], K.A. Zolotareva [2], S.A. Lashin [1, 2]

[1] Research Institute of Clinical and Experimental Lymphology – Branch of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ klimontov@mail.ru

Abstract. Data on the genetics and molecular biology of diabetes are accumulating rapidly. This poses the challenge of creating research tools for a rapid search for, structuring and analysis of information in this field. We have developed a web resource, GlucoGenes®, which includes a database and an Internet portal of genes and proteins associated with high glucose (hyperglycemia), low glucose (hypoglycemia), and both metabolic disorders. The data were collected using text mining of the publications indexed in PubMed and PubMed Central and analysis of gene networks associated with hyperglycemia, hypoglycemia and glucose variability performed with ANDSystems, a bioinformatics tool. GlucoGenes® is freely available at: https://glucogenes.sysbio.ru/genes/main. GlucoGenes® enables users to access and download information about genes and proteins associated with the risk of hyperglycemia and hypoglycemia, molecular regulators with hyperglycemic and antihyperglycemic activity, genes up-regulated by high glucose and/or low glucose, genes down-regulated by high glucose and/or low glucose, and molecules otherwise associated with the glucose metabolism disorders. With GlucoGenes®, an evolutionary analysis of genes associated with glucose metabolism disorders was performed. The results of the analysis revealed a significant increase (up to 40 %) in the proportion of genes with phylostratigraphic age index (PAI) values corresponding to the time of origin of multicellular organisms. Analysis of sequence conservation using the divergence index (DI) showed that most of the corresponding genes are highly conserved (DI < 0.6) or conservative (DI < 1). When analyzing single nucleotide polymorphism (SNP) in the proximal regions of promoters affecting the affinity of the TATA-binding protein, 181 SNP markers were found in the GlucoGenes® database, which can reduce (45 SNP markers) or increase (136 SNP markers) the expression of 52 genes. We believe that this resource will be a useful tool for further research in the field of molecular biology of diabetes.

Key words: gene; protein; diabetes mellitus; hyperglycemia; hypoglycemia; glucose variability; database; phylostratigraphic index; single nucleotide polymorphism.

For citation: Klimontov V.V., Shishin K.S., Ivanov R.A., Ponomarenko M.P., Zolotareva K.A., Lashin S.A. GlucoGenes®, a database of genes and proteins associated with glucose metabolism disorders, its description and applications in bioinformatics research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8): 1008-1017. doi 10.18699/vjgb-24-107

Funding. The GlucoGenes® database and web resource was created with the support of the Russian Science Foundation (grant No. 20-15-00057-П).

Acknowledgements. The sections on evolutionary analysis of genes and SNP analysis were performed using the Bioinformatics Sharing Centre supported by Budget Project No. FWNR-2022-0006. The authors express their sincere gratitude to O.V. Saik (RICEL – branch of ICG SB RAS) for her significant contribution to data collection and valuable advice on the development of the database. We also thank A.M. Mukhin (ICG SB RAS) for technical assistance in creating the web resource.

База данных о генах и белках, ассоциированных с нарушениями метаболизма глюкозы (GlucoGenes®): описание и возможности применения в биоинформатических исследованиях

В.В. Климонтов [1] ✉, К.С. Шишин [1], Р.А. Иванов [2], М.П. Пономаренко [2], К.А. Золотарева [2], С.А. Лашин [1, 2]

[1] Научно-исследовательский институт клинической и экспериментальной лимфологии – филиал Федерального исследовательского центра Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ klimontov@mail.ru

Аннотация. Данные в области генетики и молекулярной биологии сахарного диабета стремительно накапливаются. Это ставит задачу создания исследовательских инструментов для быстрого поиска, структурирования и анализа информации в этой области. Мы разработали базу данных о генах и белках человека, ассоциированных с высоким уровнем глюкозы (гипергликемией), низким уровнем глюкозы (гипогликемией) и обоими нарушениями. Сведения были собраны с помощью текст-майнинга научных публикаций, проиндексированных в PubMed и PubMed Central, и анализа генных сетей гипергликемии, гипогликемии и вариабельности гликемии, выполненного с помощью биоинформатической системы ANDSystems. Созданный ресурс (GlucoGenes®) доступен по адресу: https://glucogenes.sysbio.ru/genes/main. Ресурс предоставляет информацию о генах и белках, связанных с риском развития гипергликемии и гипогликемии; регуляторных молекулах с гипергликемической и антигипергликемической активностью; генах, экспрессия которых повышается при высоком и/или низком уровне глюкозы; генах, экспрессия которых снижается при высоком и/или низком уровне глюкозы, а также о молекулах, связанных с нарушениями метаболизма глюкозы иным образом. На основе ресурса проведен эволюционный анализ генов, ассоциированных с нарушениями метаболизма глюкозы. Результаты анализа выявили значительное увеличение (до 40 %) доли генов, имеющих филостратиграфический индекс (phylostratigraphy age index, PAI), соответствующий времени происхождения многоклеточных организмов. Анализ консервативности последовательностей белков по индексу дивергенции (divergency index, DI) показал, что большинство соответствующих генов высококонсервативны (DI < 0.6) или консервативны (DI < 1). При анализе однонуклеотидного полиморфизма (SNP) в проксимальных районах промоторов, влияющих на сродство TATA-связывающего белка, в базе данных GlucoGenes® найден 181 SNP-маркер, который может снижать (45 SNP-маркеров) или повышать (136 SNP-маркеров) экспрессию 52 генов. Мы полагаем, что разработанный ресурс станет полезным инструментом для дальнейших исследований в области молекулярной биологии диабета.

Ключевые слова: ген; белок; сахарный диабет; гипергликемия; гипогликемия; вариабельность глюкозы; база данных; филостратиграфический индекс; однонуклеотидный полиморфизм.

Introduction

Diabetes is one of the most common and socially significant human diseases. According to experts from the International Diabetes Federation, the number of people living with diabetes worldwide reached 537 million in 2021 and is expected to rise to 783 million by 2045. In addition, more than 540 million people have impaired glucose tolerance (International Diabetes Federation, 2021).

In recent years, significant progress has been made in understanding the molecular mechanisms underlying the development of diabetes and its complications. Genome-wide association studies have identified a number of novel genetic loci that modulate the risk of diabetes and diabetic complications in European and Asian populations. Proteomics, metabolomics and multiomics studies have shed light on the molecular basis of disease pathogenesis (Langenberg, Lotta, 2018; Lyssenko, Vaag, 2023; Shojima et al., 2023).

At the same time, the effects of exposure to high glucose concentrations on the regulation of gene expression in different tissues have been identified (Vaulont et al., 2000; Hall et al., 2018; Vega et al., 2020; Zhang S. et al., 2021). It has been shown that the effects of high glucose levels on gene expression can be prolonged and exacerbated by epigenetic modifications. This mechanism is considered to be important for the phenomenon of metabolic memory and the development of diabetic complications (Dhawan et al., 2022). Abnormally low glucose levels are also associated with a number of biochemical shifts. These shifts are primarily related to the response of the cardiovascular and nervous systems to hypoglycemia (Hanefeld et al., 2016; Rehni, Dave, 2018).

The molecular effects of repeated episodes of high and low glucose levels, which characterize the phenomenon of high glycemic variability (GV), attract increasing attention. Elevated GV has been found to increase the risk of microvascular and macrovascular diabetic complications and is associated with increased all-cause and cardiovascular mortality (Ceriello et al., 2019; Wilmot et al., 2019). At the molecular level, the pathophysiological changes associated with high GV are realized through increased or decreased expression of a large number of genes and altered activity of signaling pathways such as PI3K/Akt, NF-κB, MAPK (ERK), JNK and TGF-β/Smad (Klimontov et al., 2021b).

Given the vast number and diversity of molecular changes in diabetes, advanced analytical tools are necessary to form a comprehensive and holistic understanding of the disease's pathogenesis. Artificial intelligence, bioinformatics, and integrative systems biology provide new opportunities for studying complex diseases such as diabetes (Nielsen, 2017; Klimontov et al., 2021a; Orlov et al., 2021; Putra et al., 2024).

A promising approach in this field is the analysis of gene networks, i. e. groups of genes that function in a coordinated manner, interact with each other, and determine specific phenotypic traits of an organism (Kolchanov et al., 2013). Previously, using text mining of scientific publications indexed in the PubMed and PubMed Central (PMC) databases, along with bioinformatic analysis, we reconstructed gene networks associated with glucose metabolism disorders (GMDs): hyperglycemia, hypoglycemia, and GV, gathering a large amount of data on molecules and proteins related to these metabolic disorders in some way (Saik, Klimontov, 2020–2022).

Based on the obtained data, a database was created containing information about genes and proteins associated with GMDs: hyperglycemia, hypoglycemia, and both conditions.

In this article, we present a description of the database's capabilities and provide the results of two bioinformatics studies conducted using it. The first study involved an evolutionary analysis of GMD genes, and the second involved an analysis of single nucleotide polymorphisms (SNPs) in 90 base-pair proximal regions of human gene promoters associated with GMDs.

Materials and methods

Development of the web resource. The material for creating the GlucoGenes® web resource was previously accumulated during the reconstruction and analysis of gene networks for hyperglycemia, hypoglycemia, and GV, conducted using ANDSystem, a bioinformatics tool (developed at Institute of Cytology and Genetics SB RAS; access: https://anddigest.sysbio.ru/index.php). ANDSystem constructs associative gene networks based on text mining of scientific publications indexed in the PubMed and PMC databases (Ivanisenko et al., 2015, 2019). The details of the analysis of GMD gene networks generated using ANDSystem were described previously (Saik, Klimontov, 2020–2022).

A relational data model and the PostgreSQL database management system (https://www.postgresql.org/) were chosen for the software implementation of the database.

For the design of the web resource, a client-server architecture was chosen, consisting of three main components: client, server and database. The Vue.js and Flask frameworks were used for development, and access management to the database is implemented through programmatic access based on REST technology.

Phylostratigraphic analysis and divergence analysis of genes associated with GMDs. Phylostratigraphic analysis is a method aimed at determining the evolutionary origin of genes by analyzing the presence of their orthologs, which are genes encoding homologous proteins that have diverged through speciation in the genomes of different species. This approach identifies key points in genome evolution, where a sharp increase occurred in the number of new genes, and helps to identify genes unique to specific taxa (Domazet-Lošo, Tautz, 2010).

We performed evolutionary analysis of genes included in the GlucoGenes® database using the phylostratigraphy age index (PAI) and divergency index (DI). The PAI value indicates how far the taxon reflecting the gene's age is from the root of the phylogenetic tree (Mustafin et al., 2021). The taxon reflecting the gene's age is considered as the taxon where the studied species diverged from the most distantly related taxon in which an ortholog of the gene has been found. The higher the PAI value of a gene, the younger it is. The Orthoweb software package (https://orthoweb.sysbio.cytogen.ru/run.html) was used for phylostratigraphic analysis. For PAI calculation, the method based on KEGG orthology groups was used (Kanehisa et al., 2016).

DI is an indicator of a gene's evolutionary variability. DI is calculated based on the *dN*/*dS* ratio, where *dN* is the proportion of nonsynonymous substitutions in the DNA sequences of the studied gene and its ortholog; *dS* is the proportion of synonymous substitutions. This index was calculated by comparing human genes with genes from closely related organisms in the Hominidae family, specifically orthologs found in the western lowland gorilla *Gorilla gorilla gorilla*, Sumatran orangutan *Pongo abelii*, and common chimpanzee *Pan troglodytes*. The LPB93 model (Yang, 2007) was used to calculate *dN*/*dS* values. A DI value ranging from 0 to 1 indicates that a gene is undergoing stabilizing selection, a value of 1 indicates neutral evolution, and a value greater than 1 indicates positive selection.

Analysis of SNPs in 90-bp proximal regions of human gene promoters associated with GMD. In the Human_SNP_TATAdb knowledge base (Filonov et al., 2023), we searched for SNP variants in 90-bp proximal regions of human gene promoters associated with GMD that could statistically significantly decrease or increase the affinity of TATA-binding protein (TBP) to these promoters and consequently affect gene expression levels. Among all these SNPs, only those with clinical manifestations described in the publicly available ClinVar database (Landrum et al., 2014) were selected for further work. Finally, using the PubMed database, we performed curated annotations of how polymorphic changes in gene expression affected glucose levels in patients carrying these SNPs for all clinically relevant SNP markers located in promoters of genes associated with glucose metabolism disorders.

Results

GlucoGenes® web resource

The GlucoGenes® web resource is freely available at: https://glucogenes.sysbio.ru/genes. The interface of the resource is shown in Fig. 1.

The GlucoGenes® database consists of six components (tables). The Genes table contains gene names, descriptions and NCBI identifiers. The Proteins table includes protein names, descriptions, UniProt database identifiers, and links to the corresponding genes in the GlucoGenes® database. The Glycemia_related_conditions table provides information on glycemic disorders (hyperglycemia and hypoglycemia). The Types_of_glycemia_gene_association table contains information on the types of associations between molecules and glycemic disorders. The References table contains article identifiers in PubMed or PubMed Central with brief data extracts. The Glycemia_gene_association table is a summary table that aggregates information from all of the above-mentioned tables. The structure of the database is shown in Fig. 2. The database currently includes 561 genes associated with GMDs and 2,115 references to literature sources.

The GlucoGenes® web portal consists of four functional sections.

1. Homepage: The homepage provides general information about the resource and the terms used. We define hyperglycemia or high glucose levels in the culture medium as High Glucose (HG) and hypoglycemia or low glucose levels in the culture medium as Low Glucose (LG).

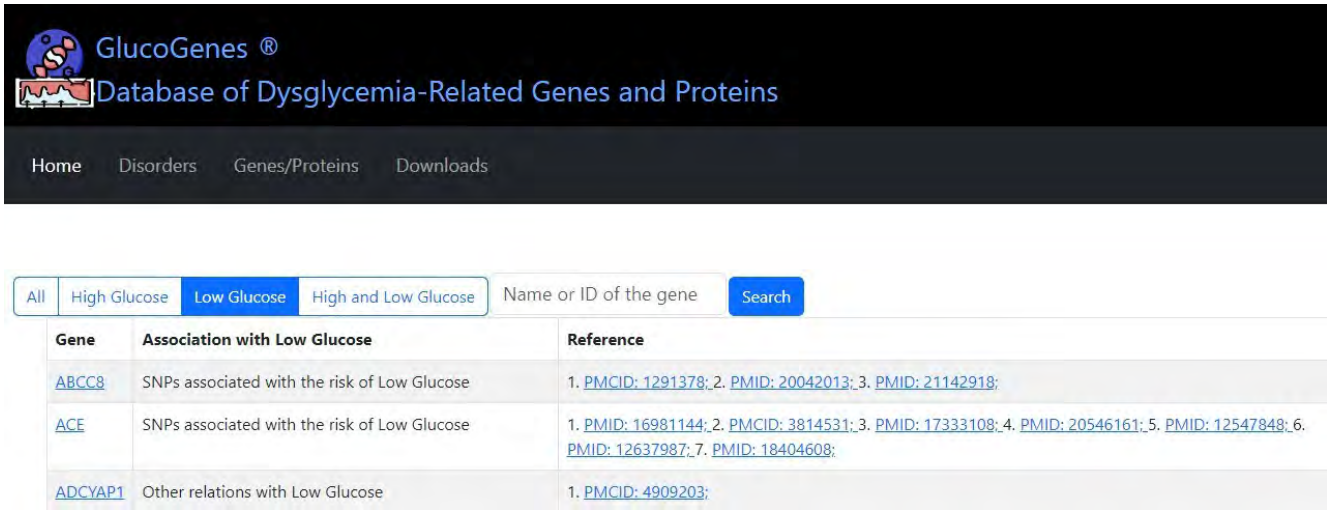


Fig. 1. GlucoGenes® website interface.

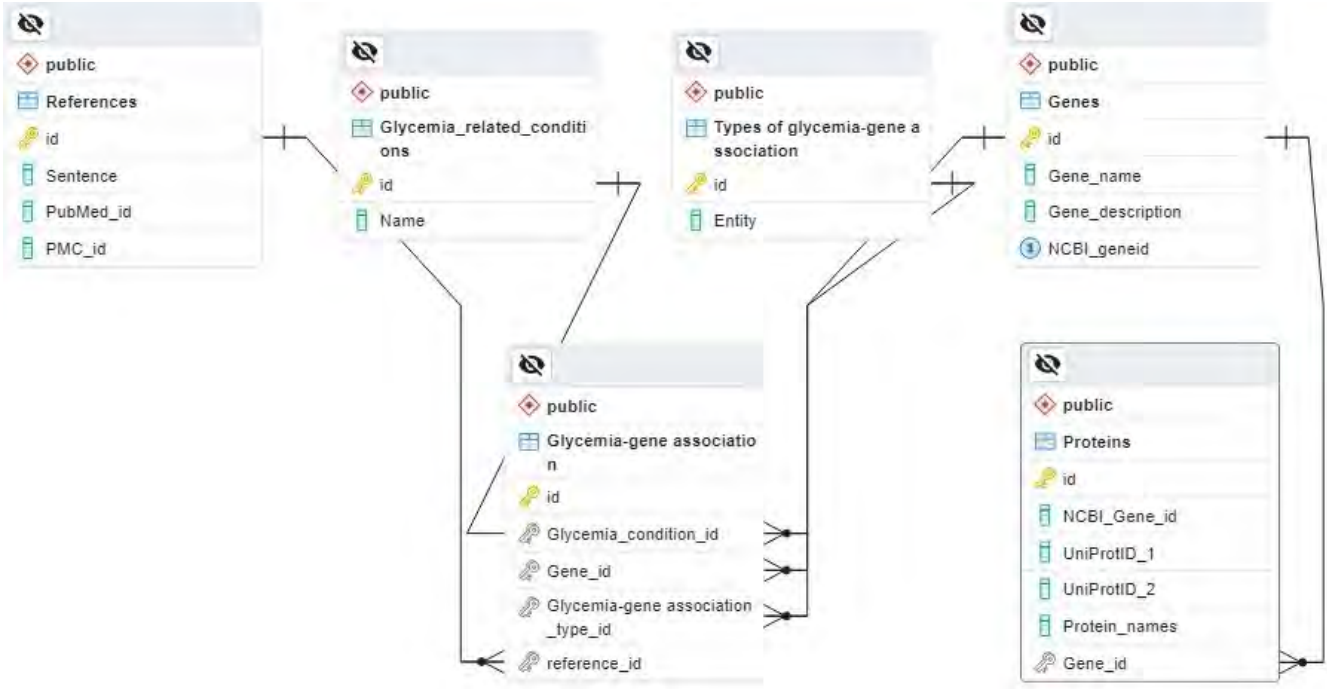


Fig. 2. Entity-relationship diagram of the GlucoGenes® database.

2. Disorders page: This page presents lists of genes associated with HG, LG, as well as with high and low glucose levels (HLG). For each gene, the type of association with glycemic disorders is indicated. The following categories of associations with glycemic disorders are highlighted: SNPs associated with HG, LG, or HLG; proteins with hyperglycemic activity; proteins with antihyperglycemic effects; genes up-regulated by HG; genes up-regulated by LG; genes down-regulated by HG; genes down-regulated by LG; and other associations with HG, LG, or HLG. For each gene and its association, references to relevant publications in PubMed are provided.
3. Genes/Proteins catalog: This section allows users to find gene names and NCBI gene identifiers, names of protein(s) encoded by the gene, and types of associations with GMD.
4. Downloads page: From this page, users can download lists of genes associated with HG, LG, and HLG, as well as all associated genes in Excel format. Search within the system is available by gene name, NCBI gene identifier, or type of GMD.

Data from the portal can also be accessed without using the graphical user interface via a REST application programming interface (API). This interface allows users to retrieve

Table 1. Lists of human protein-coding genes analyzed through phylostratigraphic analysis

Gene group	Description	Number of genes
All genes in the *Homo sapiens* genome	All human protein-coding genes for which PAI and DI values were calculated	19,566
Genes associated only with hyperglycemia	List of genes from GlucoGenes associated with high glucose levels	430
Genes associated only with hypoglycemia	List of genes from GlucoGenes associated with low glucose levels	140
Genes associated with both high and low glucose levels	List of genes from GlucoGenes associated with both high and low glucose levels	151

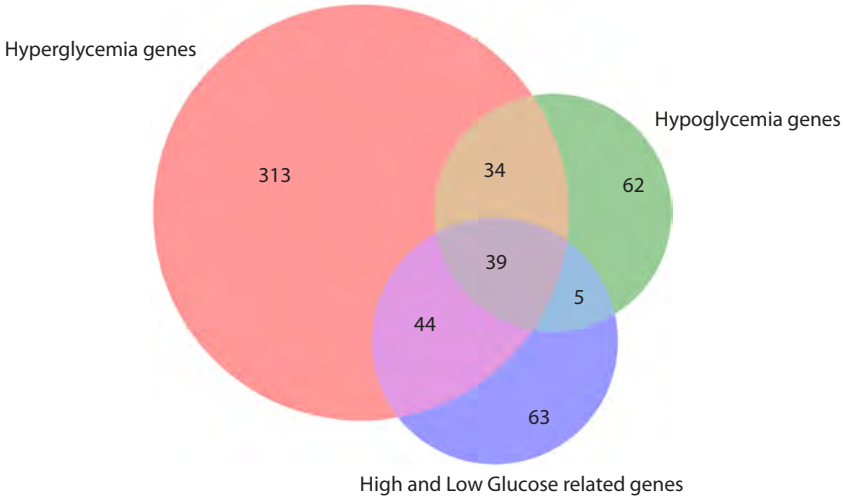


Fig. 3. Venn diagram showing intersections of gene groups.

required information by sending a request to the web server in the form of a URL string. In response to such a request, the server returns results as a text page or file, where the information is structured according to the JavaScript Object Notation (JSON) format (http://json.org/). The resulting text file can be opened with any text editor. It can also be processed using various software tools, including user-written programs in general-purpose modeling environments (e. g., Matlab, Scilab) or high-level programming languages (e. g., Python, R, C++, Java).

An example of a REST request is given below (the result is a text file in structured JSON format): https://glucogenes.sysbio.ru/api/genes/<geneid> – returns a card with a description of the gene <geneid>.

Evolutionary characteristics of genes associated with GMDs. We calculated PAI indices and plotted their distribution both for the list of protein-coding genes in the *Homo sapiens* genome and for genes represented in the GlucoGenes® database including those associated with hyperglycemia, hypoglycemia and glucose fluctuations (Table 1). Please, note that some genes were associated with more than one GMD (Fig. 3).

The distribution of PAI values for all genes in the human genome is multimodal, with two pronounced peaks at the levels of Cellular Organisms, Metazoa and Vertebrata-Euteleostomi (Fig. 4). The first peak is the largest; almost 55 % of genes in the *H. sapiens* genome have a PAI between 1 and 3. The second peak covers 32 % of the genes.

The results of the analysis showed a significant increase (up to 40 %) in the proportion of genes involved in glucose regulation with a PAI index = 3 in all three categories (Fig. 4). In particular, this group includes the *TCF7L2*, *PPARG*, *GCGR*, *IRS1* and *MTNR1B* genes, the products of which are important regulators of glucose metabolism.

Sequence conservation analysis for the same gene lists (Fig. 5) showed that most of the genes studied are highly conserved (DI < 0.6) or conserved (DI < 1). This indicates the conservation of their functions during evolution and highlights their critical role in biological processes related to glucose regulation. However, several genes with a DI greater than 1 were identified, indicating recent exposure to positive selection. These genes include *SPP1*, *CALCA*, *CD33*, *SULT2A1*, *TNF*, *ECM1*, *CYP3A4* and *EDN1*.

Analysis of SNPs in 90-bp proximal regions of human gene promoters associated with GMD. A total of 181 SNP markers were identified in the GlucoGenes® database, which may either decrease (45 SNP markers) or increase (136 SNP markers) the expression of 52 human genes, thereby altering glucose levels in patients carrying minor alleles of these SNPs. Table 2 provides an example of 10 SNPs located in the promoters of the human *ABCC8*, *INSR*, and *PGM1* genes, available in the ClinVar database (Landrum et al., 2014).

PAI	Taxon
1	Cellular Organisms
2	Eukaryota
3	Metazoa
4	Chordata
5	Craniata
6	Vertebrata
7	Euteleostomi
8	Mammalia
9	Eutheria
10	Euarchontoglires
11	Primates
12	Haplorrhini
13	Catarrhini
14	Hominidae
15	Homo
16	*Homo sapiens*

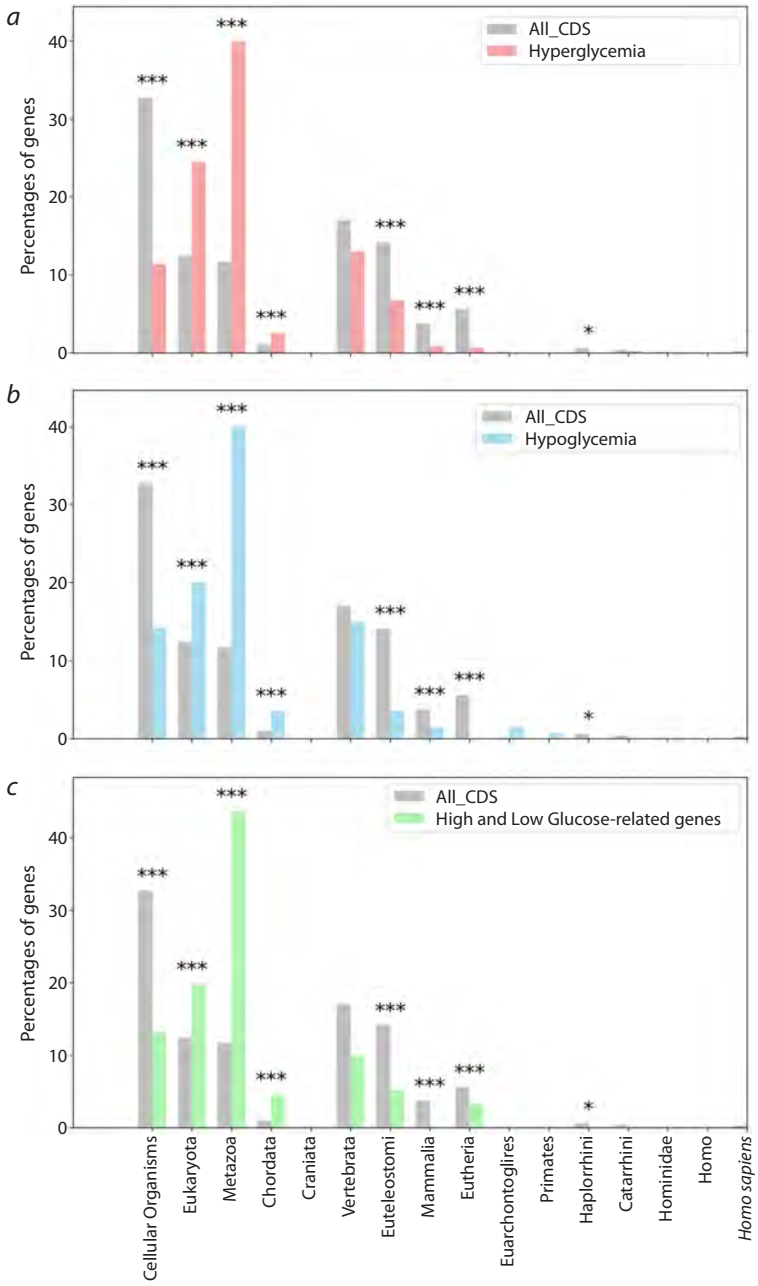


Fig. 4. Distribution of protein-coding genes associated with GMDs by PAI values.

Here and in Fig. 5: *a* – all human protein-coding genes (All_CDS) as a control group compared to genes associated with high glucose levels (Hyperglycemia); *b* – all human protein-coding genes (All_CDS) as a control group compared to genes associated with low glucose levels (Hypoglycemia); *c* – all human protein-coding genes (All_CDS) as a control group compared to genes associated with both high and low glucose levels (High and Low Glucose-related genes). Columns marked with asterisks indicate statistically significant differences between gene samples from the database and the sample of all human protein-coding genes: **p*-value < 0.05, ****p*-value < 0.001. Statistical testing was performed using the chi-square test.

According to the data presented in Table 2, minor alleles of the *ABCC8*, *INSR*, and *PGM1* gene promoters exhibit altered affinity for TBP, which may affect the expression levels of these genes and explain their association with GMDs. More detailed information on the identified SNP markers can be found in Supplementary Material[1].

[1] Supplementary Material is available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Klimontov_Engl_28_8.pdf

Discussion

Advances in the study of the molecular biology of diabetes open up broad opportunities for the implementation of precision medicine technologies in the treatment of this disease. In particular, the identification of disease-specific biomarkers offers new prospects for diagnosis, monitoring, prognosis of the disease and its outcomes, pharmacogenetics of modern glucose-lowering drugs, as well as the search for new thera-

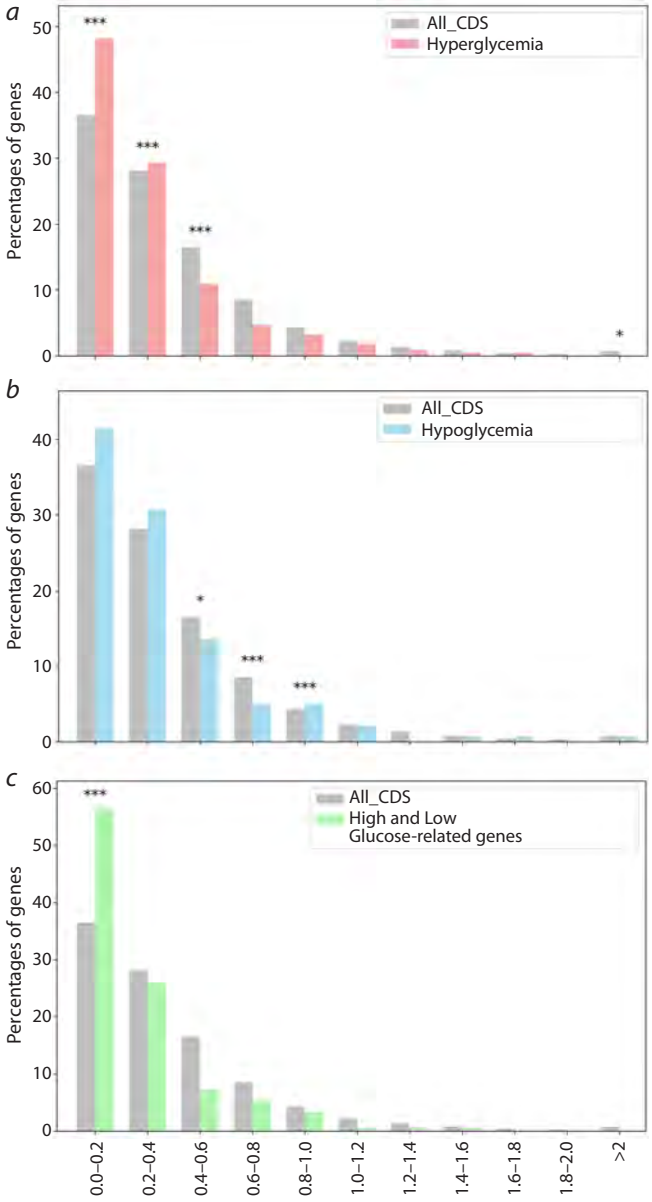


Fig. 5. Distribution of protein-coding genes associated with GMD by DI values.

peutic agents (Chung et al., 2020). The rapid accumulation of data on the molecular basis of genetic predisposition to diabetes and the molecular mechanisms of its complications underscores the need for research tools to facilitate structured information retrieval in this field.

We have developed a database of genes and proteins that have demonstrated associations with GMDs, including hyperglycemia, hypoglycemia, or both. The web-based resource, named GlucoGenes® (https://glucogenes.sysbio.ru/genes/main), can be utilized to collect, search, and visualize information on genes and proteins associated with GMDs. Access to the database integrated into GlucoGenes® is provided via a REST-based API for record browsing. A graphical user interface allows users to view records and export their content in Excel format. The database contains catalogs of genes and proteins associated with GMDs, including information on the types of associations and links to abstracts of relevant publications in PubMed or full-text articles in PMC. Gene and protein lists are available for download. A limitation of this resource is that it accumulates data only from articles indexed in the PubMed and PMC. Regular information updates are evidently required.

The developed resource may prove useful for addressing research challenges in bioinformatics and the molecular biology of diabetes. Specifically, it can be applied to select genes and proteins for studying genetic predisposition to diabetes in various populations, investigating the molecular aspects of pathogenesis, searching for potential biomarkers of diabetic complications, identifying potential therapeutic targets, and other tasks. In this study, we present examples of using the developed resource to solve research tasks in bioinformatics studies.

The first task focuses on the evolutionary origin of genes associated with GMDs. Evolutionary analysis of genes using phylostratigraphy is a key tool in biology, enabling an understanding of the fundamental mechanisms underlying the diversity of life on Earth. The evolutionary history of genes provides insights into how various functions and structures have evolved and adapted to environmental changes. This knowledge not only aids in reconstructing phylogenetic trees but also helps to identify genes responsible for adaptive changes and specific physiological processes, such as glucose metabolism. The conducted phylostratigraphic analysis revealed that among genes associated with glucose metabolism, a significant proportion (up to 40 %) are genes with PAI = 3, corresponding to the origin of multicellular organisms (Maloof et al., 2010). Most of the studied genes were found to be highly conserved (DI < 0.6) or conserved (DI < 1). The obtained results emphasize the importance of GMD-associated genes in regulating specialized metabolic processes characteristic of complex organisms.

During the second task, data from the web resource were used to analyze SNPs in proximal regions of human gene promoters that affect the affinity for TBP. The integration of GlucoGenes® data with information on SNP associations with various human diseases from other databases, on the one hand, and bioinformatic assessments of changes in glucose levels in patients carrying these SNPs, on the other hand, reflects the molecular mechanisms through which GMDs may influence the progression of these diseases.

Conclusion

GlucoGenes® is a resource that combines a graphical user interface with a database of genes and proteins associated with hyperglycemia, hypoglycemia, and both metabolic disorders. The resource has been utilized for bioinformatic analysis of the evolutionary characteristics of genes associated with these disorders, as well as for the analysis of SNPs in proximal promoter regions of genes that affect the

Table 2. SNPs affecting TBP affinity in the promoters of the *ABCC8*, *INSR* and *PGM1* genes associated with GMDs

GlucoGenes® web resource		ClinVar database (Landrum et al., 2014)					Human_SNP_TATAdb database (Filonov et al., 2023)						PubMed database
		Association with SNP		dbSNP database (Day, 2010)			K_D, nM/l, *in silico*		Significance				Effect on disease course (Ref)
Gene symbol	Glycaemia status	Disease	dbSNP ID:min	5′ flank, 10 bp	WT → min	3′ flank, 10 bp	WT	min	z	*p*	ρ	Δ	
							$M_0 \pm SEM$	$M_0 \pm SEM$					
PGM1	Low Glucose	PGM1-congenital disorder of glycosylation, inborn genetic diseases	rs199928376:T	tgccgccaac	G→T	gggtaaggga	46.52 ± 3.87	36.68 ± 3.28	3.89	10^{-3}	B	↑	Excess PGM1: A response to glucose deficiency in lung cancer (Li et al., 2020)
INSR	High and Low Glucose & High Glucose	Insulin-resistant diabetes mellitus & acanthosis nigricans; Rabson-Mendenhall syndrome, leprechaunism syndrome	rs532052290:T	cgcctcggag	C→T	tcttgggtgg	62.14 ± 4.83	17.87 ± 1.92	18.81	10^{-6}	A	↑	Excess *Insr*: Improved health outcomes in db/db mice (Sasaki et al., 2015)
ABCC8	High and Low Glucose & Low Glucose & High Glucose	ABCC8-related disorder, hypoglycemia, hyperinsulinism, diabetes mellitus, monogenic diabetes	rs1409900082:A	agcactttcc	G→A	tggtcctcag	74.04 ± 5.71	30.75 ± 2.85	14.58	10^{-6}	A	↑	Maternal diet effects on offspring in mice (Zhang Q. et al., 2024): Excess Abcc8 protects offspring
			rs1409900082:T	agcactttcc	G→T	tggtcctcag	74.04 ± 5.71	52.37 ± 3.26	6.99	10^{-6}	A		
			rs1955706366:A	cctgtccagt	G→A	ctaagcgagt	53.2 ± 4.54	12.08 ± 1.4	20.58	10^{-6}	A		
			rs1957150758:A	cttggccgac	C→A	actttccgca	74.04 ± 5.71	47.43 ± 3.75	8.06	10^{-6}	A		
			rs541244107:A	agctgcaagg	G→A	ccgggcggcg	59.19 ± 4.91	50.3 ± 3.4	3.04	10^{-2}	C		
			rs541244107:C	agctgcaagg	G→C	ccgggcggcg	59.19 ± 4.91	47.44 ± 3.82	3.83	10^{-3}	B		
			rs565662182:A	tgccccccat	G→A	gaggagcagt	53.2 ± 4.54	31.21 ± 2.78	8.65	10^{-6}	A		
			rs58241708:T	tgcagagatc	C→T	ttcctgtcca	53.2 ± 4.54	45.33 ± 3.8	2.68	10^{-2}	C		

Note. WT and min – ancestral (wild-type) and minor (pathological) alleles of SNPs; K_D – equilibrium dissociation constant of the TBP-promoter complex, expressed in nanomoles per liter, nM/l; M_0 and SEM – context-dependent *in silico* assessment and its standardized error; z, *p* and ρ – Fisher's z-statistic, its significance level, and heuristic prioritization of *in silico* evaluations from best (A) to worst (D) in alphabetical order; Δ – expression increased (↑) or decreased (↓).

affinity for TBP. It has been demonstrated that a significant proportion of genes associated with GMDs are evolutionarily ancient and conserved. SNP markers that can decrease (45 SNP markers) or increase (136 SNP markers) the expression of 52 genes have been identified.

References

Ceriello A., Monnier L., Owens D. Glycaemic variability in diabetes: clinical and therapeutic implications. *Lancet Diabetes Endocrinol.* 2019;7(3):221-230. doi 10.1016/S2213-8587(18)30136-0

Chung W.K., Erion K., Florez J.C., Hattersley A.T., Hivert M.F., Lee C.G., McCarthy M.I., Nolan J.J., Norris J.M., Pearson E.R., Philipson L., McElvaine A.T., Cefalu W.T., Rich S.S., Franks P.W. Precision medicine in diabetes: a Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia.* 2020;63(9):1671-1693. doi 10.1007/s00125-020-05181-w

Day I.N. dbSNP in the detail and copy number complexities. *Hum. Mutat.* 2010;31(1):2-4. doi 10.1002/humu.21149

Dhawan P., Vasishta S., Balakrishnan A., Joshi M.B. Mechanistic insights into glucose induced vascular epigenetic reprogramming in type 2 diabetes. *Life Sci.* 2022;298:120490. doi 10.1016/j.lfs.2022.120490

Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*. 2010;468(7325):815-819. doi 10.1038/nature09632

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 Journal of Genetics and Breeding*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85

Hall E., Dekker Nitert M., Volkov P., Malmgren S., Mulder H., Bacos K., Ling C. The effects of high glucose exposure on global gene expression and DNA methylation in human pancreatic islets. *Mol. Cell. Endocrinol.* 2018;472:57-67. doi 10.1016/j.mce.2017.11.019

Hanefeld M., Frier B.M., Pistrosch F. Hypoglycemia and cardiovascular risk: is there a major link? *Diabetes Care.* 2016;39(S.2):S205-S209. doi 10.2337/dcS15-3014

International Diabetes Federation. IDF Diabetes Atlas, 10th ed. Brussels, 2021

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(S2):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

Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):D457-D462. doi 10.1093/nar/gkv1070

Klimontov V.V., Berikov V.B., Saik O.V. Artificial intelligence in diabetology. *Sakharnyi Diabet = Diabetes Mellitus.* 2021a;24(2):156-166. doi 10.14341/DM12665 (in Russian)

Klimontov V.V., Saik O.V., Korbut A.I. Glucose variability: How does it work? *Int. J. Mol. Sci.* 2021b;22(15):7783. doi 10.3390/ijms22157783

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2013;17(4/2):833-850 (in Russian)

Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., Maglott D.R. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42:D980-D985. doi 10.1093/nar/gkt1113

Langenberg C., Lotta L.A. Genomic insights into the causes of type 2 diabetes. *Lancet.* 2018;391(10138):2463-2474. doi 10.1016/S0140-6736(18)31132-2

Li Y., Liang R., Sun M., Li Z., Sheng H., Wang J., Xu P., Liu S., Yang W., Lu B., Zhang S., Shan C. AMPK-dependent phosphorylation of HDAC8 triggers PGM1 expression to promote lung cancer cell survival under glucose starvation. *Cancer Lett*. 2020;478:82-92. doi 10.1016/j.canlet.2020.03.007

Lyssenko V., Vaag A. Genetics of diabetes-associated microvascular complications. *Diabetologia.* 2023;66(9):1601-1613. doi 10.1007/s00125-023-05964-x

Maloof A.C., Porter S.M., Moore J.L., Dudás F.Ö., Bowring S.A., Higgins J.A., Fike D.A., Eddy M.P. The earliest Cambrian record of animals and ocean geochemical change. *Geol. Soc. Am. Bull.* 2010;122(11-12):1731-1774. doi 10.1130/B30346.1

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nielsen J. Systems biology of metabolism: A driver for developing personalized and precision medicine. *Cell Metab*. 2017:25(3):572-579. doi 10.1016/j.cmet.2017.02.002

Orlov Y.L., Anashkina A.A., Klimontov V.V., Baranova A.V. Medical genetics, genomics and bioinformatics aid in understanding molecular mechanisms of human diseases. *Int. J. Mol. Sci.* 2021;22(18):9962. doi 10.3390/ijms22189962

Putra S.E.D., Martriano Humardani F., Antonius Y., Jonathan J., Thalia Mulyanata L. Epigenetics of Diabetes: A bioinformatic approach. *Clin. Chim. Acta.* 2024;557:117856. doi 10.1016/j.cca.2024.117856

Rehni A.K., Dave K.R. Impact of hypoglycemia on brain metabolism during diabetes. *Mol. Neurobiol.* 2018;55(12):9075-9088. doi 10.1007/s12035-018-1044-6

Saik O.V., Klimontov V.V. Bioinformatic reconstruction and analysis of gene networks related to glucose variability in diabetes and its complications. *Int. J. Mol. Sci.* 2020;21(22):8691. doi 10.3390/ijms21228691

Saik O.V., Klimontov V.V. Hypoglycemia, vascular disease and cognitive dysfunction in diabetes: insights from text mining-based reconstruction and bioinformatics analysis of the gene networks. *Int. J. Mol. Sci.* 2021;22(22):12419. doi 10.3390/ijms22222212419

Saik O.V., Klimontov V.V. Gene networks of hyperglycemia, diabetic complications, and human proteins targeted by SARS-CoV-2: what is the molecular basis for comorbidity? *Int. J. Mol. Sci.* 2022;23:7247. doi 10.3390/ijms23137247

Sasaki T., Kuroko M., Sekine S., Matsui S., Kikuchi O., Susanti V.Y., Kobayashi M., Tanaka Y., Yuasa T., Kitamura T. Overexpression of insulin receptor partially improves obese and diabetic phenotypes in db/db mice. *Endocr. J.* 2015;62(9):787-796. doi 10.1507/endocrj.ej15-0255

Shojima N., Yamauchi T. Progress in genetics of type 2 diabetes and diabetic complications. *J. Diabetes Investig.* 2023;14(4):503-515. doi 10.1111/jdi.13970

Vaulont S., Vasseur-Cognet M., Kahn A. Glucose regulation of gene transcription. *J. Biol. Chem.* 2000;275(41):31555-31558. doi 10.1074/jbc.R000016200

Vega M.E., Finlay J., Vasishtha M., Schwarzbauer J.E. Elevated glucose alters global gene expression and tenascin-C alternative splicing in mesangial cells. *Matrix Biol. Plus.* 2020;8:100048. doi 10.1016/j.mbplus.2020.100048

Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591. doi 10.1093/molbev/msm088

Wilmot E.G., Choudhary P., Leelarathna L., Baxter M. Glycaemic variability: The under-recognized therapeutic target in type 1 diabetes care. *Diabetes Obes. Metab.* 2019;21(12):2599-2608. doi 10.1111/dom.13842

Zhang S., Ke Z., Yang C., Zhou P., Jiang H., Chen L., Li Y., Li Q. High glucose causes distinct expression patterns of primary human skin cells by RNA sequencing. *Front. Endocrinol.* 2021;12:603645. doi 10.3389/fendo.2021.603645

Zhang Q., Xiao X., Zheng J., Li M., Yu M., Ping F., Wang T., Wang X. DNA methylation regulates pancreatic gene expression and links maternal high-fat diet to the offspring glucose metabolism. *J. Nutr. Biochem*. 2024;123:109490. doi 10.1016/j.jnutbio.2023.109490

Conflict of interest. The authors declare no conflict of interest.
Received October 2, 2024. Revised November 8, 2024. Accepted November 12, 2024.

BIOMEDICINE
Original article

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(8):1018-1024

doi 10.18699/vjgb-24-108

Association of autistic personality traits with the EEG scores in non-clinical subjects during the facial video viewing

A.N. Savostyanov [1, 2, 3], D.A. Kuleshov [1, 4, 5], D.I. Klemeshova[1, 2], M.S. Vlasov [6], A.E. Saprygin [1, 2]

[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
[4] Trofimuk Institute of Petroleum Geology and Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[5] Siberian State University of Telecommunications and Informatics, Novosibirsk, Russia
[6] Altai State Pedagogical University, Biysk Branch named after V.M. Shukshin, Biysk, Russia
a-sav@mail.ru

Abstract. A software information module of the experimental computer platform "EEG_Self-Construct" was developed and tested in the framework of this study. This module can be applied for identification of neurophysiological markers of self-referential processes based on the joint use of EEG and facial video recording to induce the brain's functional states associated with participants' personality traits. This module was tested on a group of non-clinical participants with varying degrees of severity of autistic personality traits (APT) according to the Broad Autism Phenotype Questionnaire. The degree of individual severity of APT is a quantitative characteristic of difficulties that a person has when communicating with other people. Each person has some individual degree of severity of such traits. Patients with autism are found to have high rates of autistic traits. However, some individuals with high rates of autistic traits are not accompanied by clinical symptoms. Our module allows inducing the brain's functional states, in which the EEG indicators of people with different levels of APT significantly differ. In addition, the module includes a set of software tools for recording and analyzing brain activity indices. We have found that relationships between brain activity and the individual level of severity of APT in non-clinical subjects can be identified in resting-state conditions following recognition of self-referential information, while recognition of socially neutral information does not induce processes associated with APT. It has been shown that people with high scores of APT have increased spectral density in the delta and theta ranges of rhythms in the frontal cortical areas of both hemispheres compared to people with lower scores of APT. This could hypothetically be interpreted as an index of reduced brain activity associated with recognition of self-referential information in people with higher scores of autistic traits. The software module we are developing can be integrated with modules that allow identifying molecular genetic markers of personality traits, including traits that determine the predisposition to mental pathologies.

Key words: information-digital platforms in medicine; neurocomputation technologies; resting-state EEG; autistic personality traits; Broad Autism Phenotype; self-referential processing; default-mode network.

For citation: Savostyanov A.N., Kuleshov D.A., Klemeshova D.I., Vlasov M.S., Saprygin A.E. Association of autistic personality traits with the EEG scores in non-clinical subjects during the facial video viewing. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):1018-1024. doi 10.18699/vjgb-24-108

Funding. The part of the study concerning the preparation of psychological tests, selection of experimental groups and EEG registration was carried out with the financial support of the Russian Science Foundation (RSF) within the framework of research project No. 22-15-00142.

Acknowledgements. The development of the hardware and software module was carried out within the framework of the budget project of Institute of Cytology and Genetics SB RAS No. FWNR-2022-0020.

Ассоциация аутистических личностных черт у неклинических испытуемых с показателями ЭЭГ в условиях просмотра видеозаписей лица

А.Н. Савостьянов [1, 2, 3], Д.А. Кулешов [1, 4, 5], Д.И. Клемешова[1, 2], М.С. Власов [6], А.Е. Сапрыгин [1, 2]

[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
[4] Институт нефтегазовой геологии и геофизики им. А.А. Трофимука Сибирского отделения Российской академии наук, Новосибирск, Россия
[5] Сибирский государственный университет телекоммуникаций и информатики, Новосибирск, Россия
[6] Бийский филиал им. В.М. Шукшина Алтайского государственного педагогического университета, Бийск, Россия
a-sav@mail.ru

Аннотация. В рамках проводимого исследования разработан и апробирован программно-информационный модуль экспериментально-компьютерной платформы "EEG_Self-Construct", позволяющий выявлять нейрофизиологические маркеры самореферентных процессов на основе совместного использования ЭЭГ и регистрации видеозаписей лица для индукции функциональных состояний головного мозга, ассоциированных с личностными особенностями участников. Этот модуль был апробирован на группе неклинических участников с разной степенью выраженности аутистических личностных черт (АЛЧ), измеренных с помощью опросника расширенного фенотипа аутизма. Степень индивидуальной выраженности АЛЧ – это количественный показатель, который характеризует затруднения, возникающие у человека при коммуникации с другими людьми. У каждого человека имеется некоторая индивидуальная степень выраженности таких черт. Высокие значения аутистических черт определяются у пациентов с аутизмом. Однако существуют также люди, у которых высокие значения АЛЧ не сопровождаются клинической симптоматикой. Разработанный нами модуль дает возможность индуцировать функциональные состояния головного мозга, в которых ЭЭГ-показатели людей с разным уровнем АЛЧ достоверно различаются. Кроме того, модуль включает комплект программного обеспечения для регистрации и анализа индексов мозговой активности. Нами установлено, что зависимости между мозговой активностью и индивидуальным уровнем выраженности АЛЧ у неклинических испытуемых могут быть выявлены в условиях функционального покоя, следующих за распознаванием самоотнесенной информации, тогда как распознавание социально нейтральной информации не индуцирует процессы, связанные с аутистичностью. Показано, что у людей с высокими значениями АЛЧ наблюдаются повышенные показатели спектральной плотности в диапазонах дельта- и тета-ритмов в лобных отделах обоих полушарий в сравнении с людьми с низкой степенью аутистичности. Это может быть гипотетически интерпретировано как индекс сниженной мозговой активности, ассоциированной с распознаванием самоотнесенной информации у людей с высокой аутистичностью. Разрабатываемый нами программный модуль может быть интегрирован с модулями, позволяющими выявлять молекулярно-генетические маркеры личностных черт, включая черты, определяющие предрасположенность к психиатрическим патологиям.
Ключевые слова: информационно-цифровые платформы в медицине; нейровычислительные технологии; ЭЭГ покоя; аутистические черты; расширенный аутистический фенотип; самореференция; дефолт-система мозга.

Introduction

The development of new approaches to identifying predisposition to certain types of behavior, including an increased risk of developing mental disorders, is based on testing individuals using genetic, neurophysiological and behavioral methods, accumulating experimental information in databases and analyzing it using a wide range of information technologies (Ivanov et al., 2022; Lin et al., 2022).

According to modern concepts, autism is a disease that is associated with disturbances in the brain and manifests itself in the social sphere (Baron-Cohen, 2002; Lavenne-Collot et al., 2023). This disease manifests itself in three domains of behavior: social interaction, communication (use of verbal and non-verbal stimuli), as well as limited and repetitive patterns in behavior, interests and activities (Baron-Cohen, 2009; Murray et al., 2017). In the 1980s, autism was recognized as a spectrum of conditions (disorders), which can be individual for each patient (Lovaas, 1987).

There is no strict boundary between a "healthy person" and an "autistic person", since each person can be assigned a certain rate of some autistic personality traits (APT) measured by the Broad Autism Phenotype Questionnaire, BAPQ (Piven et al., 1997). The higher the rate of APT, the more the subject's behavior resembles that of an autistic person. It is believed that the manifestation of APT is clinical in nature if its rate exceeds a certain threshold. However, there is a phenomenon of "non-clinical autism", when a person with an expressed APT does not consider it necessary to seek medical help. At the same time, a significant part of such "non-clinical autistic persons" turn out to be adapted people who, during their lives, demonstrate a level of social success that is no different from individuals with low rates of autistic traits. It is assumed that there are some compensatory mechanisms that may be formed depending on the influence of the environment and can both weaken and strengthen the manifestation of APT in subjects (Frith, 1991; Georgiades et al., 2017).

Since autism and APT are associated with behavioral difficulties in social communication, most neurophysiological (Tsai et al., 2013; Tseng et al., 2015) and genetic (Genovese, Butler, 2023) studies compare the brain responses of individuals with different degrees of autistic traits to the presentation of external stimuli, the recognition of which is essential for the regulation of interpersonal communication. For experimental research of the phenomenon of autism, approaches such as psychological testing using questionnaires, recording and analysis of EEG under stimulation conditions are used. Facial photographs (Harms et al., 2010; Tseng et al., 2015) or speech tasks (Tsai et al., 2013) are usually used as stimuli. However, some studies demonstrate the association of the severity of autism with brain activity under resting-state conditions without recognition of external stimuli (Harikumar et al., 2021).

An effective method is the registration of a facial video to induce psychological states that differ in participants with different degrees of expression of personality traits (Si et al., 2024).

Another approach used is to record the EEG without any additional stimulation. It is based on the hypothesis about the functional role of the default mode network of the brain in organizing self-reference processes. The default mode network is a set of cortical areas that demonstrate increased activation under resting-state conditions, but decrease the level of activation when performing tasks associated with attention to external stimulation. The default mode network is

considered as a brain structure involved in the assessment of socially significant stimuli that the subject attributes to oneself (Northoff et al., 2005). It is assumed that clinical forms of autism are accompanied by a decrease in the activity of the default mode network (Ronde et al., 2024). The functioning of the default mode network can be associated not only with the characteristics of individuals' social behavior, but also with the characteristics of their genome (Fanelli et al., 2024).

Previously, we proposed an approach for joint registration and processing of EEG and facial video that allows combining brain activity analysis with assessment of facial muscle dynamics (Savostyanov et al., 2022). In this study, we propose a methodology based on the use of video fragments obtained at the first stage of the study to stimulate participants at later stages of the study. As shown below, this approach provides useful information for identifying markers of autistic traits in non-clinical subjects.

To provide information support for the conducted research, we are developing the "EEG_AutisticTrait" software information module, which is an important component of the "EEG_Self-Construct" experimental computer module. It provides a full cycle of information support for research, including: (a) accumulation and storage of the results of examining people using psychological, neurophysiological and genetic methods that make it possible to identify individual characteristics of social communication associated with autism; (b) computer processing of experimental data using regression, correlation and factor analysis methods that compare behavioral and neurophysiological indicators (Si et al., 2024); (c) visualization of primary experimental data and results of data analysis.

The fundamental novelty of the proposed approach is that time intervals of EEG recordings under resting-state conditions in the intervals between recognition of self-referential or non-self-referential stimuli are used to identify neurophysiological markers of APT. This approach allows inducing mental states associated with self-reference in the intervals of functional rest.

Materials and methods

The sequence of stages of the experimental computer module "EEG_Self-Construct" and the list of software tools required for the implementation of these stages are presented in Table 1. The module contains both software products developed by ICG SB RAS staff and programs taken from open sources. All modules allow for a full cycle of data collection and processing required to establish markers of autistic personality traits.

Study participants. The study involved volunteers, among which students of Novosibirsk State University prevailed. The sample included 43 participants aged from 18 to 48 years (19 males and 24 females). All participants had no neurological or mental disorders at the time of the study and did not use any psychoactive substances or pharmacological drugs. Participants gave informed consent to undergo the experimental study in accordance with the Helsinki Declaration on Biomedical Ethics. The experimental protocol was approved by the Ethics Committee of the Research Institute of Neuroscience and Medicine.

Table 1. List of stages of module operation and software tools required to perform each stage

Name of the module operation stage	Software packages required to implement the stage	The result of passing the stage of the module's work
Stage 1. Extracting lists of candidate genes and brain structures associated with personality traits from natural language texts	ANDSystem Software	List of candidate genes for behavioral genetics studies, lists of brain structures for neurophysiological studies
Stage 2. Planning the experimental design and data processing	EventIDE	Protocols of behavioral and neurophysiological experiments, protocols of data processing
Stage 3. Development of experimental paradigms for psychophysiological studies	EventIDE, Millisecond Software	Software scripts for conducting experiments
Stage 4. Registration of EEG/ECG signals	NeoRec System	EEG and ECG signal recordings with event tagging
Stage 5. Development of an experimental database	ICBrainDB	Network database of psychological, neurophysiological and genetic data
Stage 6. Preprocessing of EEG/ECG signals	EEGlab_toolbox	Neurophysiological signals cleared of irrelevant noise
Stage 7. Localization of signal sources on the surface of the brain cortex	EEGLab_toolbox, eLoreta	Maps of localization of brain activity sources
Stage 8. Statistical processing of behavioral, physiological and genetic data	IBM SPSS Statistics, GNU PSPP	Results of statistical comparisons of experimental samples, list of neurophysiological and genetic markers of personality traits

Psychological testing was performed using a special Internet form implemented on the Yandex platform by ICG SB RAS staff. All participants filled out the Russian-language version of the BAPQ to assess the severity of autistic traits according to the Broad Autism Phenotype Questionnaire (Hurley et al., 2007, translated by M.S. Vlasov). This test includes 36 questions concerning a person's ability to control one's behavior in social situations. In addition, the participants filled out psychological questionnaires on personal and situational anxiety by C. Spielberger (Spielberger, 1970; Russian adaptation by (Khanin, 1976)), a questionnaire for assessing personality traits by L. Goldberg "Markers of the Big Five Factors" (translated and validated by G.G. Knyazev et al. (2010), a questionnaire on affiliation with one's family (Cross et al., 2000), and a questionnaire on emotional intelligence (Knyazev et al., 2012).

Experiment. The program for conducting the experiment is implemented on the Inquisit platform (https://www.millisecond.com/). In the experiment, the participants fulfilled three conditions. In the first condition, the EEG was recorded for 12 minutes without a functional load. The subject had three 2-minute intervals with closed eyes and three 2-minute intervals with open eyes. During the intervals when the subject opened one's eyes, a black computer screen was presented to the subject. During this period, the subject's face was recorded along with the EEG for all 12 minutes. The second and third conditions differed from the first in that in the second condition, with open eyes, the subject watched a video recording of his or her own face, obtained from the first condition, and in the third condition, he/she was shown a video recording of a stranger's face (always a man for a male subject, and a woman for a female subject). The order of the second and third tasks was changed randomly.

EEG registration and processing. The NeoRec software (by "Medical Computer Systems", https://mks.ru/) was used to register neurophysiological data. EEG was registered using a 130-channel amplifier NVX-132, Russia, 128 EEG channels located according to the international 5-5 % scheme with a reference electrode Cz, ground electrode AFz, bandwidth 0.1–100 Hz, signal sampling frequency 1000 Hz. In addition to EEG, EOG and ECG were additionally registered.

Muscle and other artifacts were removed from the EEG using independent component analysis with the EEGlab_toolbox software package (Delorme, Makeig, 2004; https://sccn.ucsd.edu/eeglab/index.php). Then, fragments corresponding to periods when the participant sat with eyes closed were extracted from the EEG recordings. Further analysis was performed only for those intervals of the EEG recordings in which the participant did not see either video recordings or a blank screen, but which were recorded immediately after observing the corresponding stimuli. After extracting these EEG fragments, they were divided into two-second time intervals. Further analysis was performed using the eLoreta software package (Pascual-Margui, 2002; https://www.uzh.ch/keyinst/loreta.htm).

In our case, the neurophysiological states detected using eLoreta were compared with the psychological characteristics of the subjects to determine the markers of APT. For each two-second interval, the spectral density values were calculated in the frequency of delta (2–4 Hz), theta (4–8 Hz), alpha-1 (8–10 Hz), alpha-2 (10–12 Hz), beta-1 (12–16 Hz), beta-2 (16–20 Hz), beta-3 (20–25 Hz) and gamma (25–35 Hz) bands. Then, for each participant, the total spectrum indicator was calculated for the entire EEG testing interval separately for each of the three experimental conditions (for each participant, from 150 to 170 two-second intervals were used for this). The spectra were calculated independently for each of the 128 EEG channels included in data processing. A 3000 ms EEG recording segment with a sampling frequency of 1000 Hz after the onset of the block was used to calculate the spectral density of the sources in eLoreta (Pascual-Margui, 2002).

Statistical analysis. The validity of psychological tests was assessed using the IBM SPSS software package, IBM, https://www.ibm.com/spss. Regression analysis was performed in the eLoreta package to find the dependence of spectral density on the indicators of individual BAPQ score independently for each of the three experimental conditions. Additional correction for multiple comparisons was not performed.

Results

Results of psychological testing

To assess the reliability of the Russian version of the BAPQ test, we determined the internal consistency of responses to 36 items of this questionnaire using Cronbach's alpha. The Cronbach's alpha value was 0.838, which indicates a fairly high internal consistency. In addition, we assessed the correlation of individual BAPQ scores with scores on various scales of well-validated psychological measures. Table 2 shows the correlation between autistic traits (BAPQ scores) and other personality traits assessed in this study. The BAPQ score correlates reliably positively with anxiety and negatively with extroversion, the ability to express positive emotions and affiliation with the family.

eLoreta results for detecting effects of autistic traits

Correlations between BAPQ autistic traits scores were statistically significant only for the "own face" condition ($p = 0.0340$) in the delta (2–4 Hz) and theta (4–8 Hz) bands (see the Figure). For both bands, eLoreta revealed a positive association between the spectral density scores and individual severity of autistic traits in the frontal cortex of both hemispheres, i. e. higher BAPQ autistic traits scores corresponded to higher spectral density scores. There was no significance for the "blank screen" condition ($p = 0.28640$). For the "another person's face" condition ($p = 0.0932$), the p-value was close to, but did not reach, significance.

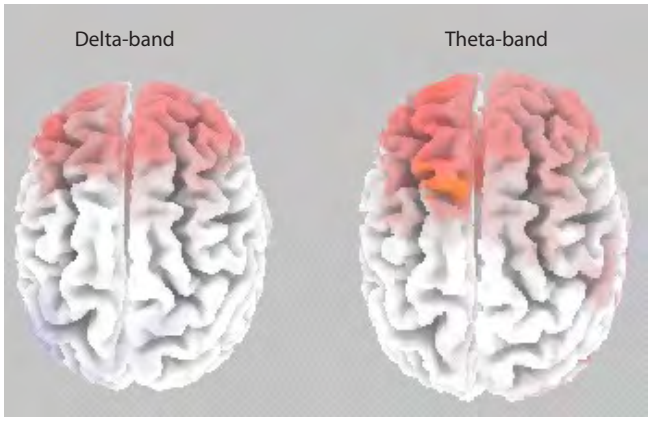
Discussion

Identification of neurophysiological markers of personality traits, including traits associated with predisposition to diseases, involves the use of complex multicomponent tools

Table 2. Correlation between autistic traits (BAPQ score) and other personality traits

BAPQ	Anxiety	Extroversi	Affiliation with the family	Ability to express positive emotions
Person correlation	0.407**	–0.524**	–0.351**	–0.278*
2-tailed *p*-value	0.002	0.003	0.007	0.036
N	43	43	43	43

* Significant correlation, *p*-value < 0.05 (two-tailed).
** Significant correlation, *p*-value < 0.01 (two-tailed).



Correlation of the spectral density in the delta (2–4 Hz) and theta (4–8 Hz) bands with the severity of autistic traits (measured by BAPQ) in a group of 43 participants for EEG intervals with eyes closed between viewing one's own face.

The cortical areas showing positive correlations of autistic traits with spectral density ($p < 0.04$) are marked in red. A significantly positive association is observed between autistic traits and spectral density in the frontal areas of both hemispheres.

for planning experiments, collecting, storing and analyzing data, comparing the results of different studies and organizing access to different programs and the data obtained with their help. An important component of such tools is the opportunity to develop and implement new paradigms for conducting neurophysiological research. For example, in (Si et al., 2024), a software module was developed to identify cross-national characteristics in the processes of self-attribution of information to the subject oneself or to other people, which is crucial for the search for markers of depression.

In the search for markers of predisposition to psychiatric disorders, an important task is the reconstruction and analysis of gene networks underlying the regulation of psycho-emotional states in humans and animals (Savostyanov, Makarova, 2024). An example of a module aimed at reconstructing and comparing gene networks of anxiety in mice and humans is described in (Savostyanov, Makarova, 2024). Using this module, it is possible to identify brain structures in which differential gene expression is detected in animals that differ in their level of anxiety. In the future, such structures can be considered as areas of interest for identifying neurophysiological markers of anxiety disorder in humans.

The software-information module "EEG_AutisticTrait" was tested to identify neurophysiological markers of autistic personality traits. Using a special Yandex platform, comprehensive testing of participants was conducted using several questionnaires, including a test for individual expression of autistic personality traits (the Russian version of BAPQ). The Cronbach's alpha for the Russian version of BAPQ was 0.83, which indicates a fairly high internal consistency of this questionnaire. Negative correlations of autistic personality traits with extroversion, emotional intelligence and affiliation with the family, and positive correlations between autistic personality traits and anxiety were also found, which is in good agreement with the general understanding of psychologists about autistic traits.

At the neurophysiological level, positive correlations were found between BAPQ scores and the spectral density in the delta and theta bands for the experimental condition associated with self-referential visual information, but no reliable relationships were found for the conditions following viewing a socially neutral stimulus (blank screen) or information related to other individuals. According to the literature (Knyazev, 2007), high values of the spectral density of the delta and theta rhythm under resting-state conditions are most often interpreted as an indicator of reduced functional brain activity. With this approach, our results can be hypothetically explained as a correlate of reduced brain activity in conditions following the presentation of self-referential information in individuals with more vivid autistic traits compared to individuals with lower levels of autistic traits.

Significantly, we identified neurophysiological correlates of autistic traits only for the self-referential condition. In the socially neutral condition, there was no tendency for BAPQ scores to be related to brain activity, whereas for the "another person's face" condition, there was a marginal statistical tendency for the result to be significant. It can be assumed that resting-state EEG activity in non-clinical subjects is weakly associated with their level of autism, which explains the failure of previous attempts to identify any relationships between autistic traits and resting-state EEG in such participants. However, viewing video recordings related to the

participant oneself (and to a lesser extent, to other people) activates processes in the brain associated with the recognition of socially significant information, which makes EEG indices more dependent on autistic traits than in the case of viewing socially neutral stimuli.

Conclusion

The approach we propose is based on the integration of psychological and neurophysiological methods of data collection and analysis. In the future, it is planned to evaluate the dependence of autistic traits on the genetic characteristics of the subjects. It is also desirable to evaluate the effect of the expression level of various genes in the brain on the severity of personality traits. The assessment of the level of gene expression in the brain cannot be performed on a living person, which suggests the need to combine data obtained on people and on experimental animals (Savostyanov, Makarova, 2024). Such a study requires the development of special tools for the accumulation, storage and analysis of data, which will be created on the basis of the Bioinformatics and Systems Computational Biology platform. In the future, this tool can be used to assess the neurophysiological correlates of various personality traits in healthy controls and subjects with different pathologies, which will make it possible to conduct new comprehensive studies within the framework of system neurobiology.

References

Baron-Cohen S. The extreme male brain theory of autism. *Trends Cogn. Sci*. 2002;6(6):248-254. doi 10.1016/s1364-6613(02)01904-6

Baron-Cohen S. Autism: the empathizing-systemizing (E-S) theory. *Ann. N.Y. Acad. Sci.* 2009;1156:68-80. doi 10.1111/j.1749-6632.2009.04467.x

Cross S.E., Bacon P.L., Morris M.L. The relational-interdependent self-construal and relationships. *J. Pers. Soc. Psychol*. 2000;78(4):791-808

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*. 2004;134(1):9-21. doi 10.1016/j.jneumeth.2003.10.009

Fanelli G., Robinson J., Fabbri C., Bralten J., Mota N.R., Arenella M., Sprooten E., Franke B., Kas M., Andlauer T.F., Serretti A. Shared genetics linking sociability with the brain's default mode network. *medRxiv*. [Preprint]. 2024. May 25:2024.05.24.24307883. doi 10.1101/2024.05.24.24307883

Frith U. Asperger and his syndrome. In: Frith U. (Ed.). Autism and Asperger Syndrome. Cambridge University Press, 1991;1-36

Genovese A., Butler M.G. The autism spectrum: behavioral, psychiatric and genetic associations. *Genes* (*Basel*). 2023;14(3):677. doi 10.3390/genes14030677

Georgiades S., Bishop S.L., Frazier T. Editorial perspective: longitudinal research in autism − introducing the concept of 'chronogeneity'. *J. Child Psychol. Psychiatry*. 2017;58:634-636. doi 10.1111/jcpp.12690

Harikumar A., Evans D.W., Dougherty C.C., Carpenter K.L.H., Michael A.M. A review of the default mode network in autism spectrum disorders and attention deficit hyperactivity disorder. *Brain Connect*. 2021;11(4):253-263. doi 10.1089/brain.2020.0865

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

Hurley L., Parlier M., Reznick J., Piven J. The broad autism phenotype questionnaire. *J. Autism Dev. Disord.* 2007;37(9):1679-1690. doi 10.1007/s10803-006-0299-3

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

Khanin Yu.L. Quick Guide to C.D. Spielberger's Scale of State and Trait Anxiety. Leningrad, 1976 (in Russian)

Knyazev G.G. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neurosci. Biobehav. Rev.* 2007;31(3):377-395. doi 10.1016/j.neubiorev.2006.10.004

Knyazev G.G., Mitrofanova L.G., Bocharov A.V. Validization of Russian version of Goldberg's ''Big-five factor markers»" inventory. *Psikhologicheskii Zhurnal*. 2010;31(5):100-110 (in Russian)

Knyazev G.G., Mitrofanova L.G., Razumnikova O.M., Barchard K. Adaptation of Russian language version of K. Barchard's Emotional Intelligence Questionnaire. *Psikhologicheskii Zhurnal*. 2012;33(4):112-120 (in Russian)

Lavenne-Collot N., Tersiguel M., Dissaux N., Degrez C., Bronsard G., Botbol M., Berthoz A. Self/other distinction in adolescents with autism spectrum disorder (ASD) assessed with a double mirror paradigm. *PLoS One*. 2023;18(3):e0275018. doi 10.1371/journal.pone.0275018

Lin M., Wang Y., Lopez-Naranjo C., Hu S., Reyes R.C.G., Paz-Linares D., Areces-Gonzalez A., Hamid A.I.A., Evans A.C., Savostyanov A.N., Calzada-Reyes A., Villringer A., Tobon-Quinero C.A., Garcia-Agustin D., Yao D., Dong L., Aubet-Vazquez E., Reza F., Razzaq F.A., Omar H., Abdullah J.M., Galler J.R., Ochoa-Gomez J.F., Prichep L.S., Galan-Garcia L., Morales-Chacon L., Valdes-Sosa M.J., Trondle M., Zulkifly M.F.M., Rahman M.R.B.A., Milakhina N.S., Langer N., Rudych P., Koenig T., Virues-Alba T.A., Lei X., Bringas-Vega M.L., Bosch-Bayard J.F., Valdes-Sosa P.A. Harmonized-Multinational qEEG norms (HarMNqEEG). *NeuroImage*. 2022;256:119190. doi 10.1016/j.neuroimage.2022.119190

Lovaas O.I. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *J. Consult. Clin. Psychol.* 1987;55(1):3-9. doi 10.1037/0022-006x.55.1.3

Murray K., Johnston K., Cunnane H., Kerr Ch., Spain D., Gillan N., Hammond N., Murphy D., Happe F. A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. *Autism Res*. 2017;10(6):1120-1132. doi 10.1002/aur.1744

Northoff G., Heinzel A., De Greck M., Bermpohl F., Dobrowolny H., Panksepp J. Self-referential processing in our brain – a meta-analysis of imaging studies on the self. *NeuroImage*. 2005;31(1):440-457. doi 10.1016/j.neuroimage.2005.12.002

Pascual-Margui R.D. Standardized low-resolution brain electromagnetic tomography (sLORETA). Technical details. *Methods Find. Exp. Clin. Pharmacol*. 2002;24(Suppl. D):5-12

Piven J., Palmer P., Jacobi D., Childress D., Arndt S. Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. *Am. J. Psychiatry*.1997;154(2):185-190. doi 10.1176/ajp.154.2.185

Ronde M., van der Zee E.A., Kas M.J.H. Default mode network dynamics: An integrated neurocircuitry perspective on social dysfunction in human brain disorders. *Neurosci. Biobehav. Rev*. 2024;164:105839. doi 10.1016/j.neubiorev.2024.105839

Savostyanov A.N., Vergunov E.G., Saprygin A.E., Lebedkin D.A. Validation of a face image assessment technology to study the dynamics of human functional states in the EEG resting-state paradigm.

Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2022;26(8):765-772. doi 10.18699/VJGB-22-92

Savostyanov V.A., Makarova A.A. Reconstruction and analysis of the gene network for regulation of trait anxiety level in mice by means of ANDSystem software. In: IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, 2024;2340-2343. doi 10.1109/EDM61683.2024.10615053

Si Q., Tian J., Savostyanov V.A., Lebedkin D.A., Bocharov A.V., Savostyanov A.N. Comparison of brain activity indexes in the Chinese and Russian students under recognition of self- and other-related information. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):982-992. doi 10.18699/vjgb-24-105

Spielberger C.D., Gorsuch R.L., Lushene R.E. Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press, 1970

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., 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

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

Received October 28, 2024. Revised November 12, 2024. Accepted November 13, 2024.

Алфавитный указатель авторов статей, опубликованных в журнале в 2024 г.

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 публикации 26.12.2024. Формат 60 × 84 $^{1}/_{8}$. Уч.-изд. л. 30.5.