2021 • 25 • 1

Print ISSN 2500-0462
Online ISSN 2500-3259

Научный рецензируемый журнал

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.
Периодичность 8 выпусков в год
DOI 10.18699/VJ21.001

Учредители

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Межрегиональная общественная организация Вавиловское общество генетиков и селекционеров

Сибирское отделение Российской академии наук

2021 • 25 • 1

Print ISSN 2500-0462
Online ISSN 2500-3259

Scientific Peer Reviewed Journal

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

Founded in 1997
Published 8 times annually
DOI 10.18699/VJ21.001

Founders
Federal State Budget Scientific Institution "The Federal Research Center Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences"
The Vavilov Society of Geneticists and Breeders
Siberian Branch of the Russian Academy of Sciences

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

СОДЕРЖАНИЕ • 2021 • 25 • 1

5 ОТ РЕДАКТОРА

Биоинформатика и системная компьютерная биология

7 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2. *А.В. Цуканов, В.Г. Левицкий, Т.И. Меркулова*

18 ОБЗОР
Геномная изменчивость в регуляторных районах генов, ассоциированная с заболеваниями человека: механизмы влияния на транскрипцию генов и полногеномные информационные ресурсы, обеспечивающие исследование этих механизмов. *Е.В. Игнатьева, Е.А. Матросова*

30 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Улучшение качества сборки *de novo* транскриптомов ячменя на основе гибридного подхода для линий с изменениями окраски колоса и стебля. *Н.А. Шмаков*

39 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Поиск участников сигнального пути ауксина к его транспортерам PIN на основе метаанализа транскриптомов, индуцированных ауксином. *В.В. Коврижных, З.С. Мустафин, З.З. Багаутдинова*

46 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Филостратиграфический анализ генных сетей заболеваний человека. *З.С. Мустафин, С.А. Лашин, Ю.Г. Матушкин*

57 ОБЗОР
Пангеномы сельскохозяйственных растений. *А.Ю. Пронозин, М.К. Брагина, Е.А. Салина*

64 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса. *Е.А. Урбанович, Д.А. Афонников, С.В. Николаев*

71 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Автоматическое фенотипирование морфологии колоса тетра- и гекса-плоидных видов пшеницы методами компьютерного зрения. *А.Ю. Пронозин, А.А. Паулиш, Е.А. Заварзин, А.Ю. Приходько, Н.М. Прохошин, Ю.В. Кручинина, Н.П. Гончаров, Е.Г. Комышев, М.А. Генаев*

82 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19. *О.И. Криворотько, С.И. Кабанихин, М.И. Сосновская, Д.В. Андорная*

92 ОБЗОР
Механический стресс клеток мозга, локальная трансляция и нейродегенеративные заболевания: молекулярно-генетические аспекты. *Т.М. Хлебодарова*

Биотехнология

101 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Биоинформационный анализ сплайс-лидерного транс-сплайсинга у регенерирующего плоского червя *Macrostomum lignano* показал его преобладание среди консервативных генов и генов стволовых клеток. *К.В. Устьянцев, Е.В. Березиков (на англ. языке)*

108 ОБЗОР
Macrostomum lignano как модельный объект для исследования генетики и геномики паразитических плоских червей. *К.В. Устьянцев, В.Ю. Вавилова, А.Г. Блинов, Е.В. Березиков*

117 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Трансгенная клеточная линия с индуцируемой транскрипцией для исследования механизмов экспансии (CGG)n повторов. *И.В. Грищенко, А.А. Тулупов, Ю.М. Рымарева, Е.Д. Петровский, А.А. Савелов, А.М. Коростышевская, Ю.В. Максимова, А.Р. Шорина, Е.М. Шитик, Д.В. Юдкин*

125 ОБЗОР
Продукция субтилизиновых протеаз в бактериях и дрожжах. *А.С. Розанов, С.В. Шеховцов, Н.В. Богачева, Е.Г. Першина, А.В. Ряполова, Д.С. Бытяк, С.Е. Пельтек*

VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2021 • 25 • 1

5 FROM THE EDITOR

Bioinformatics and computational systems biology

7 ORIGINAL ARTICLE
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. *A.V. Tsukanov, V.G. Levitsky, T.I. Merkulova*

18 REVIEW
Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms. *E.V. Ignatieva, E.A. Matrosova*

30 ORIGINAL ARTICLE
Improving the quality of barley transcriptome *de novo* assembling by using a hybrid approach for lines with varying spike and stem coloration. *N.A. Shmakov*

39 ORIGINAL ARTICLE
The auxin signaling pathway to its PIN transporters: insights based on a meta-analysis of auxin-induced transcriptomes. *V.V. Kovrizhnykh, Z.S. Mustafin, Z.Z. Bagautdinova*

46 ORIGINAL ARTICLE
Phylostratigraphic analysis of gene networks of human diseases. *Z.S. Mustafin, S.A. Lashin, Yu.G. Matushkin*

57 REVIEW
Crop pangenomes. *A.Yu. Pronozin, M.K. Bragina, E.A. Salina*

64 ORIGINAL ARTICLE
Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm. *E.A. Urbanovich, D.A. Afonnikov, S.V. Nikolaev*

71 ORIGINAL ARTICLE
Automatic morphology phenotyping of tetra- and hexaploid wheat spike using computer vision methods. *A.Yu. Pronozin, A.A. Paulish, E.A. Zavarzin, A.Yu. Prikhodko, N.M. Prokhoshin, Yu.V. Kruchinina, N.P. Goncharov, E.G. Komyshev, M.A. Genaev*

82 ORIGINAL ARTICLE
Sensitivity and identifiability analysis of COVID-19 pandemic models. *O.I. Krivorotko, S.I. Kabanikhin, M.I. Sosnovskaya, D.V. Andornaya*

92 REVIEW
The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases. *T.M. Khlebodarova*

Biotechnology

101 ORIGINAL ARTICLE
Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes. *K.V. Ustyantsev, E.V. Berezikov*

108 REVIEW
Macrostomum lignano as a model to study the genetics and genomics of parasitic flatworms. *K.V. Ustyantsev, V.Yu. Vavilova, A.G. Blinov, E.V. Berezikov*

117 ORIGINAL ARTICLE
A transgenic cell line with inducible transcription for studying (CGG)n repeat expansion mechanisms. *I.V. Grishchenko, A.A. Tulupov, Y.M. Rymareva, E.D. Petrovskiy, A.A. Savelov, A.M. Korostyshevskaya, Y.V. Maksimova, A.R. Shorina, E.M. Shitik, D.V. Yudkin*

125 REVIEW
Production of subtilisin proteases in bacteria and yeast. *A.S. Rozanov, S.V. Shekhovtsov, N.V. Bogacheva, E.G. Pershina, A.V. Ryapolova, D.S. Bytyak, S.E. Peltek*

Dear colleagues, dear readers,

this issue of the journal focuses on bioinformatics. In the last decade, a rapid improvement of methods for decoding genomes resulted in an information explosion of such a power that genetics has become the largest source of data not only in world science, but also in all other aspects of human activity, including social networks. Studies looking into the human genome become more and more intensive common with the advent of large international projects. As of August 14, 2020, the 1000 Genomes Project (https://www.internationalgenome.org/) had sequenced 3202 genomes. The 100,000 Genomes Project (https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/) has sequenced the genomes of 85,000 patients with rare diseases/cancer.

As of July 31, 2020, the 1000 Bull Genomes Project (http://www.1000bullgenomes.com) had sequenced the genomes of more than 5000 animals in 200 cattle breeds and species. This resulted in the identification of more than 155 million genetic variants (SNPs and small deletions/insertions). As of August 1, 2019, the sheep genome sequence project SheepGenomesDB (https://sheepgenomesdb.org) had sequenced the genomes of 935 animals in 69 breeds and found more than 50 million genetic variants. As of November 9, 2020, the 1000 Goat Genome Project (http://www.goatgenome.org/vargoats_data_access.html) had collected data on 127,852,473 genetic variants identified in 1159 animals in 101 goat breeds.

Marker-oriented and genomic breeding as well as genomic editing called for the deciphering of the genomes of the main agricultural plants: wheat, maize, barley, rice, soybean, common bean, potato, a wide range of vegetables and fruits and others (http://plants.ensembl.org/species.html; http://www.plantgdb.org/prj/GenomeBrowser/). A large project seeking to study genetic variation in rice based on the sequencing of a 3000-strong collection of accessions from 89 countries (the 3,000 Rice Genomes Project. *Gigascience*. 2014;3:7. DOI 10.1186/2047-217X-3-7) has achieved completion. One of the ongoing projects is 10KP, a whole genome plant sequencing plan started in 2018 and aiming at the complete sequencing of 10,000 plants in the main clades of embryophytes, green algae and protists (with the exception of fungi) (Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.M., Li F.W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K. 10KP: A phylodiverse genome sequencing plan. *Gigascience*. 2018; 7(3):1-9. DOI 10.1093/gigascience/giy013). The genomes sequenced to date include more than 36,000 whole viral genomes (https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/), 163,645 whole bacterial genomes and 1886 whole Archaeal genomes (https://gold.jgi.doe.gov/distribution). More than 2590 fungal genomes have been sequenced under the 1000 Fungal Genomes Project (https://mycocosm.jgi.doe.gov/pages/fungi-1000-projects.jsf).

A tremendous wealth of protein sequence and annotation data been collected. The Universal Protein Resource (UniProt, https://www.uniprot.org/) is a database containing the descriptions of 563,082 experimentally confirmed primary protein structures; the TrEMBL database (https://www.uniprot.org/statistics/TrEMBL) contains more than 190 million amino acid sequences obtained by automated computer-aided genome annotation. Progress in methods for the physical and chemical studies of proteins has resulted in a fast-paced accumulation of data on their spatial structure (174,507 entries in the Protein Data Bank (PDB), https://www.rcsb.org/). Of invaluable importance is the information on protein structure contained in the mass spectrometry database Chemdata.nist.gov (https://chemdata.nist.gov/) including the descriptions of more than 100 million mass spectra of chemical peptides and metabolites from a range of tissues, biological liquids and cells.

At present, KEGG Pathway (manual annotation), STRING (https://string-db.org/), GeneMANIA (https://genemania.org/), Pathway Commons (https://www.pathwaycommons.org/) and other resources contain more than 70,000 gene regulatory networks, signal transduction pathways and metabolic pathways reconstructed to date.

As far as medicine is concerned, it will benefit from the gigantic bodies of data on human genetic variation: dbSNP (https://www.ncbi.nlm.nih.gov/) contains more than 72 million entries with human SNPs (of which ~24,000 are associated with various diseases) and Ensembl (http://www.ensembl.org/Homo_sapiens) contains more than 667 million entries with human SNPs.

The information explosion in genetics has become quite a challenge, for the rates at which genomic data are accumulated are well ahead of the rates at which these data can be analyzed in a computer-aided manner, and so most genome projects end up in formal assemblies with very rough annotations (or without them) (https://gold.jgi.doe.gov/). This suggests the fundamental importance of information technologies and bioinformatics for the storage, processing and analysis

of genomic data for the benefit of addressing fundamental and applied problems in genetics, medicine, pharmacology, agriculture, biotechnology and biosafety.

The understanding and practical use of tremendously large bodies of exceptionally complex genetic experimental data asked for modern information technologies, efficient methods for computer-aided analysis of big data and mathematical modeling of biological systems and processes at different hierarchical levels of organization of living systems – from genomes, genes, proteins, metabolic pathways and gene regulatory networks, including cells and tissues, to whole organisms, populations and ecosystems. In this issue, the reader will find papers on such aspects of bioinformatics as computer-aided genomics and transcriptomics, computer-aided systems biology, computer-aided evolutionary biology and automated analysis of plant phenotypes.

Nikolay A. Kolchanov,

Scientific Editor of the issue,
Full Member of the Russian Academy of Sciences,
Academic Advisor, the Institute of Cytology and Genetics
of the Siberian Branch of the Russian Academy of Sciences

Вавиловский журнал генетики и селекции. 2021;25(1):7-17
DOI 10.18699/VJ21.002

Original Russian text www.bionet.nsc.ru/vogis/

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

A.V. Tsukanov[1] ✉, V.G. Levitsky[1, 2], T.I. Merkulova[1, 2]

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

Abstract. The most popular model for the search of ChIP-seq data for transcription factor binding sites (TFBS) is the positional weight matrix (PWM). However, this model does not take into account dependencies between nucleotide occurrences in different site positions. Currently, two recently proposed models, BaMM and InMoDe, can do as much. However, application of these models was usually limited only to comparing their recognition accuracies with that of PWMs, while none of the analyses of the co-prediction and relative positioning of hits of different models in peaks has yet been performed. To close this gap, we propose the pipeline called MultiDeNA. This pipeline includes stages of model training, assessing their recognition accuracy, scanning ChIP-seq peaks and their classification based on scan results. We applied our pipeline to 22 ChIP-seq datasets of TF FOXA2 and considered PWM, dinucleotide PWM (diPWM), BaMM and InMoDe models. The combination of these four models allowed a significant increase in the fraction of recognized peaks compared to that for the sole PWM model: the increase was 26.3 %. The BaMM model provided the main contribution to the recognition of sites. Although the major fraction of predicted peaks contained TFBS of different models with coincided positions, the medians of the fraction of peaks containing the predictions of sole models were 1.08, 0.49, 4.15 and 1.73 % for PWM, diPWM, BaMM and InMoDe, respectively. Thus, FOXA2 BSs were not fully described by only a sole model, which indicates theirs heterogeneity. We assume that the BaMM model is the most successful in describing the structure of the FOXA2 BS in ChIP-seq datasets under study.
Key words: transcription factor binding sites (TFBS); TFBS *de novo* searching; ChIP-seq; heterogeneity of TFBS.

For citation: 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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):7-17. DOI 10.18699/VJ21.002

Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2

А.В. Цуканов[1] ✉, В.Г. Левицкий[1, 2], 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
✉ tsukanov@bionet.nsc.ru

Аннотация. В настоящее время самой распространенной моделью поиска сайтов связывания транскрипционных факторов (ССТФ) в пиках ChIP-seq является позиционная весовая матрица (position weight matrix, PWM). Но эта модель не учитывает взаимосвязи между частотами встреч нуклеотидов в разных позициях ССТФ, поэтому не способна гарантировать определение всех возможных структурных вариантов ССТФ. На сегодняшний день уже предложены альтернативные модели, например BaMM и InMoDe, которые учитывают такие взаимосвязи. Однако применение этих моделей обычно сводилось к сравнению их точности с точностью традиционной модели PWM, тогда как анализ совместной встречаемости и относительного расположения ССТФ разных моделей в пиках не производился. В нашей работе мы предлагаем конвейер программ MultiDeNA, позволяющий сочетать разные модели *de novo* поиска ССТФ для выявления структурной гетерогенности ССТФ в данных ChIP-seq. Разработанный конвейер включает этапы построения моделей на основе заданного набора пиков, оценки точности распознавания моделей с помощью перекрестных тестов, выбора порогов, сканирования пиков ChIP-seq и классификацию пиков по результатам сканирования. С применением конвейера нами проведен анализ 22 экспериментов ChIP-seq для ТФ FOXA2 с помощью четырех моделей: PWM, diPWM, BaMM и InMoDe. Показано, что сочетание моделей позволяет существенно увеличить

общее количество распознанных пиков (на 26.3 %) по сравнению с применением только PWM; при этом основной вклад в распознавание внесла модель BaMM. В значительной доле пиков разные модели распознают совпадающие ССТФ; однако для моделей PWM, diPWM, BaMM и InMoDe медианы доли пиков, которые содержали ССТФ только одной модели, составили 1.08, 0.49, 4.15 и 1.73 % соответственно. Таким образом, совокупность ССТФ FOXA2 не описывается полностью только одной моделью, что свидетельствует о наличии структурной гетерогенности в ССТФ у FOXA2.
Ключевые слова: сайты связывания транскрипционных факторов (ССТФ); *de novo* поиск ССТФ; ChIP-seq; гетерогенность ССТФ.

Introduction

Transcription factors (TFs) are proteins that can recognize certain regions of genomic DNA (TF binding sites, TFBS) (Lambert et al., 2018). The main function of TFs is to increase or decrease a level of gene transcription (Latchman, 2001). The key stage of the regulation of gene expression is TF binding to DNA. This binding initiates a chain of molecular events that ensure the assembly and regulate the activity of the pre-initiation complex of RNA polymerase II, both through direct or indirect contacts with the components of this complex, and through the involvement of various modifying chromatin and remodeling proteins. As a consequence, local changes in the structure of chromatin allow the transcription initiation (Iwafuchi-Doi, 2019; Srivastava, Mahony, 2020). Therefore, one of the most important tasks of modern molecular biology is to identify genomic TFBSs.

Currently, the ChIP-seq technique is widely used to solve this problem (Farnham, 2009; Park, 2009). This technique is based on the chromatin immunoprecipitation with antibodies to an investigated TF with consequent high-throughput sequencing of precipitated DNA. Primary ChIP-seq data processing identifies DNA regions, or peaks, in which a target TF was directly or through intermediate proteins binds DNA (Furey, 2012). However, lengths of peaks are usually equal to hundreds of bp, but a length of TFBS does not exceed 20–25 bp (Levitsky et al., 2007; Kulakovskiy et al., 2018). Thus, the next stage of the bioinformatics processing of ChIP-seq data is to search exact positions of TFBS in peaks. To date, many tools have been developed to solve this issue, the overwhelming majority of them are based on the model of position weight matrix (PWM) (Stormo, 2000), including such popular ones as ChIPMunk (Kulakovskiy, Makeev, 2009) and Homer (Heinz et al., 2010). It is no exaggeration to say that the use of different implementations of the PWM model are included in almost every pipeline of ChIP-seq data processing (Lloyd, Bao, 2019).

The application of the standard PWM-based approach to the processing of ChIP-seq data showed that for most TFs about a half of peaks did not contain detected PWM hits (Worsley Hunt, Wasserman, 2014; Gheorghe et al., 2019). Traditionally, this was associated with the main disadvantage of PWM, the hypothesis of independence of nucleotides frequencies in different positions of TFBS, which is not always true. This may negatively affect the recognition accuracy (Benos et al., 2002; Keilwagen, Grau, 2015). Therefore, alternative models of TFBS recognition have being developed. They took into account dependencies between nucleotides occurrences in a site model (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). Thus, the simplest alternative model was the dinucleotide position weight matrix (diPWM), it took into account dependences between adjacent nucleotides (Zhang M., Marr, 1993; Kulakovskiy et al., 2013). On the other hand, models such as BaMM (Siebert, Söding, 2016) and InMoDe (Eggeling et al., 2017) have been proposed. They were constructed using Markov chains, which took into account dependences of positions using the concept of Markov chain order, i. e. a length for which nucleotide frequencies can be mutually dependent (an order usually does not exceed 5 nt).

Authors of these alternative models proved that their models might outperform in recognition accuracy the standard PWM. However, these models were not applied to solve the problem of incomplete recognition of TFTS in ChIP-seq peaks. We assume that this problem is partially related to the structural heterogeneity of binding sites of TFs, and the number of recognized peaks can be significantly increased with the combination of different models together. In this case, the ChIP-seq peaks contain both predicted TFBS with application of a sole model, or with two models, etc. (Ignatieva et al., 2004; Levitsky et al., 2014, 2016). Earlier, we used the training sample of 53 known TF sites of the FOXA subfamily and analyzed ChIP-seq data of FOXA2 (Wederell et al., 2008; Wallerman et al., 2009) with alternative models ChIPMunk (PWM) (Kulakovskiy, Makeev, 2009) and SiteGA (Levitsky et al., 2007) with experimentally fitted model's thresholds (EMSA experiment, electrophoretic mobility shift assay, shift in electrophoretic mobility analysis). We showed that both models together found FOXA2 sites in more than 95 % of peaks (Levitsky et al., 2014). This conclusion was consistent with the absence in literature any data about indirect interaction of this well-studied TF with genomic DNA.

The given example indicates that combination of alternative models with PWM model for analyzing ChIP-seq data is promising. However, until now there has been no systematic research on this topic. Alternative models of TFBS search are not widely used, despite that about 20 years ago it was proved that there is a dependence of the nucleotide frequencies in different positions in TFBS (Bulyk et al., 2002). As the indicator of the popularity of different models, we use the number of citations of papers devoted to specific *de novo* TFBS searching programs for ChIP-seq data analysis. Thus, at the end of 2020, papers

devoted to the implementation of the PWM model MEME, HOMER and ChIPMunk (Bailey, Elkan, 1994; Heinz et al., 2010; Kulakovskiy et al., 2010; Machanick, Bailey, 2011) have the total number of citations over 6000. However, papers devoted to alternative models BaMM, InMoDe and diChIPMunk (Kulakovskiy et al., 2013; Siebert, Söding, 2016; Eggeling et al., 2017; Kiesel et al., 2018) have about 50 citations. Moreover, specific studies of individual ChIP-seq experiments were usually analyzed only with the standard PWM model. This situation we explain as follow. First, the PWM model is understandable and anyone can simply interpret it. Second, advantages of alternative models are insufficiently understandable. E. g., hardly anyone thought that alternative models were able systematically to find out TFBS of a different structure.

In this paper, we propose the pipeline that combines four *de novo* models of TFBS search, namely ChIPMunk/diChIPMunk implementations of PWM/diPWM (Kulakovskiy et al., 2010, 2013), and the Markov models InMoDe (Eggeling et al., 2017) and BaMM (Siebert, Söding, 2016). The pipeline evaluates the recognition accuracy of these models, selects their thresholds and classifies ChIP-seq peaks by comparing respective scan results. This approach expands the understanding of TFBS structural diversity, especially in cases when the PWM model is unable to find TFBS in a peak. We applied the pipeline for 22 ChIP-seq datasets for TF FOXA2.

Materials and methods

For the analysis we used the set of preprocessed 22 ChIP-seq datasets for TF FOXA2 in the bed format from the ReMap database http://remap.univ-amu.fr/ (Chèneby et al., 2020), see the Table. Only the best 4000 peaks we used for analysis in each sample (see below).

The input of our pipeline includes a dataset of ChIP-seq peaks with notation of genome version (mm10 or hg38) and the list of available TFBS search programs (PWM, diPWM, BaMM, InMoDe). The notation of genome version allows selection of the list of promoters in the fasta format (5′-regions of protein-coding genes, 2000 bp upstream transcription start sites). This promoter dataset is required for concordant threshold selection for all models. The total sizes of these samples were 19795/19991 genes for the human/mouse genomes (GRCh38.p13/GRCm38.p6 versions). We used the reference genome to extract nucleotide sequences of the peaks.

Pipeline for searching heterogeneity of TFBS. We have developed the MultiDeNA pipeline (multiple de novo analysis, https://github.com/ubercomrade/MultiDeNA) to search TFBS in ChIP-seq data with several *de novo* models. This pipeline allows obtaining the classification of ChIP-seq peaks, which is used to estimate the structural diversity of TFBS. The pipeline currently uses ChIPMunk (PWM), diChIPMunk (diPWM), BaMM, and InMoDe models, as well as the bedtools (Quinlan, Hall, 2010) and TomTom (Gupta et al., 2007) support programs. The schematic diagram of the program pipeline is shown in Fig. 1. The

The list of ChIP-seq experiments used in our study

No.	GEO/ ENCODE ID	Cell line/tissue	Treatment	TomTom
1	ENCSR066EBK	Hep-G2	–	+
2	GSE90454	BJ1-hTERT	Mimosine	+
3	GSE90454	A-549	–	+
4	ENCSR000BRE	A-549	–	+
5	GSE92491	BJ1-hTERT	Mimosine	+
6	GSE90454	BJ1-hTERT	–	+
7	ENCSR080XEY	Liver	–	+
8	ENCSR310NYI	Liver	–	+
9	ENCSR000BNI	Hep-G2	–	+
10	GSE90454	BJ1-hTERT	–	+
11	ERP004206	H9	–	+
12	GSE92491	BJ1-hTERT	Mimosine	–
13	GSE90454	KerCT	–	+
14	GSE90454	BJ1-hTERT	Mimosine	–
15	GSE90454	BJ1-hTERT	Mimosine	+
16	GSE90454	BJ1-hTERT	Mimosine	+
17	GSE90454	BJ1-hTERT	GATA4	–
18	ERP008682	Pancreas	CARN1618	+
19	GSE90454	BJ1-hTERT	Mimosine	–
20	GSE92491	BJ1-hTERT	CDT1	+
21	GSE90454	Hep-G2	–	–
22	GSE92491	BJ1-hTERT	FOXA2 and GATA4 coexpression	–

Note: GEO/ENCODE – unique identifier of databases (GSE*/ENC*). TomTom – result of filtering data using TomTom software (see "Comparison of found TFBS with known ones using TomTom tool"). (+)/(–) – the frequency matrix built on the basis of the TFBS found by ChIPMunk (PWM) is significantly similar (p-value < 0.001)/not similar (p-value > 0.001) to the frequency matrix of the FOXA2 TFBS from HOCOMOCO FOXA2_HUMAN.H11MO.0.A.

pipeline includes the following steps: (1) data preparation, (2) building of a model, (3) model accuracy assessment, (4) threshold selection and search of TFBS in ChIP-seq peaks with the fixed thresholds and (5) classification of ChIP-seq peaks according to results of TFBS recognition. Each stage of the program pipeline is described below.

Preparing initial data for analysis. The preparation of the data included the sorting of peaks according the value $-10 \cdot \log_{10}$ (p-value) that characterized the peak quality. This value was previously calculated for each peak by the MACS program (Zhang Y. et al., 2008). The pipeline of ReMap database (Chèneby et al., 2020) used this program to process raw ChIP-seq data. For each ChIP-seq dataset,

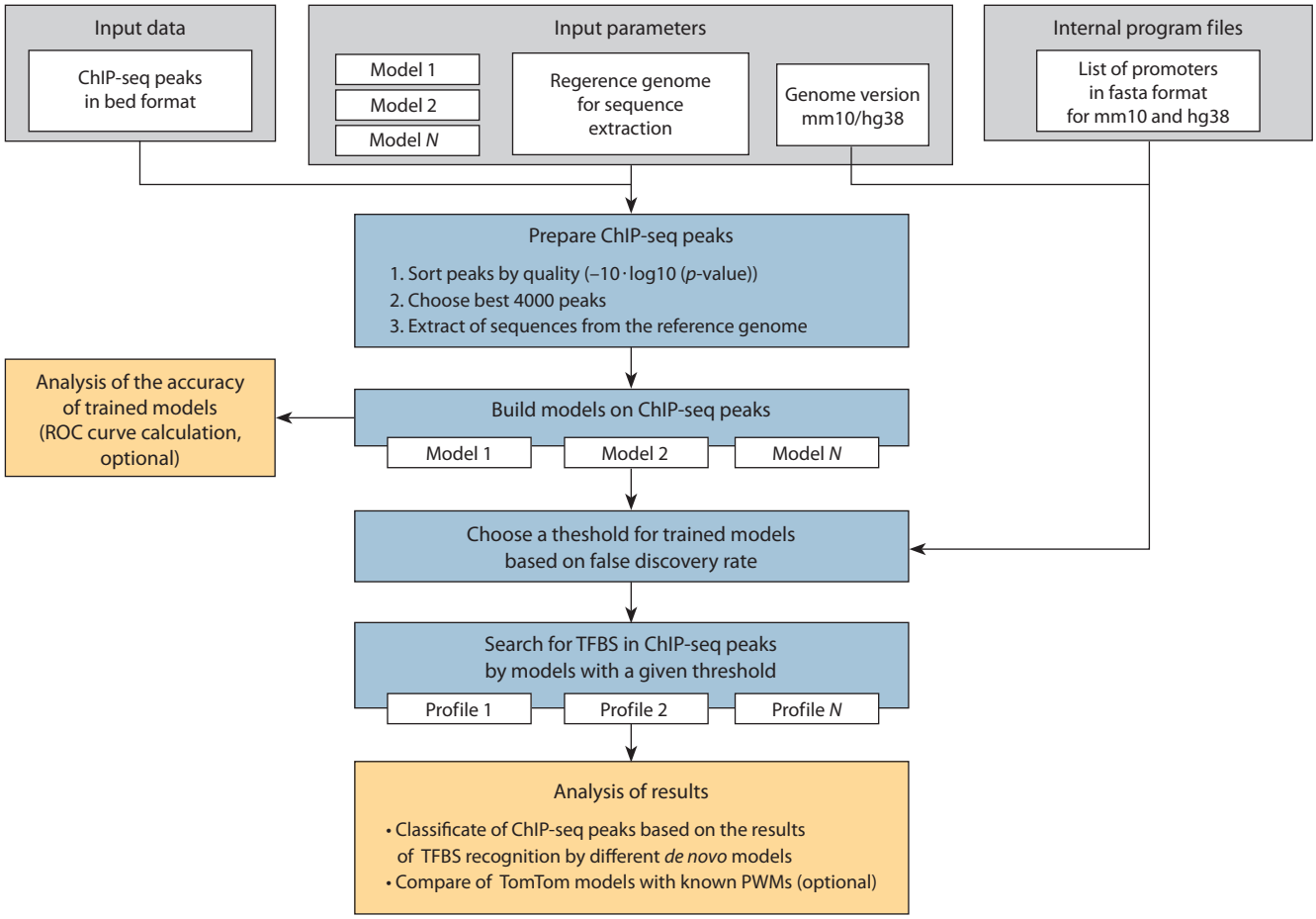


Fig. 1. The scheme of MultiDeNA workflow.

we took in analysis top-scoring 4000 peaks. Next, nucleotide sequences of the peaks we extracted from the genome using bedtools (Quinlan, Hall, 2010).

Training *de novo* models and assessing the TFBS recognition accuracy. In order to recognize TFBS in peaks, it is necessary to build *de novo* models. The PWM and diPWM models we build with ChIPMunk and diChIPMunk, respectively (Kulakovskiy et al., 2010, 2013).The construction of alternative models we carry out with BaMM (Siebert, Söding, 2016) and InMoDe (Eggeling et al., 2017).

To improve the recognition accuracy of PWM model, we selected it optimal length by the cross-validation procedure. We used the same length for the construction of other models. This procedure included the following steps: (1) to divide the ChIP-seq dataset randomly into the training (90 % of the peaks) and the control (remaining 10 % of the peaks) samples; (2) to build a model with the training sample; (3) to get recognition scores of a model with the control sample to calculate true positive rate (TPR); (4) to generate the sample of random sequences by shuffling of nucleotides in the control sample; (5) to get scores of a model with the sample of random sequences to calculate the false positives rate (FPR); (6) repetition of steps 1–5 several times; (7) to calculate the ROC-curve (receiver operating characteristic). We compared different models with the pAUC value (partial area under curve), we calculated it as the part of the area under ROC curve for all FPR values less than 0.001 (McClish, 1989; Siebert, Söding, 2016). The method described above for choosing the optimal PWM length was developed earlier (Levitsky et al., 2007; Kulakovskiy et al., 2013). The accuracy of all models we assessed with the same approach.

Next, a model can be applied to a nucleotide sequence with the same length as a model site. The result of applying this model is the recognition score. The larger score points to the higher probability of estimated nucleotide sequence to be a functional TFBS.

Threshold selection for models according to false positive rate estimates. To compare the results of TFBS search of different models correctly, it is necessary to set thresholds for all models uniformly. We set these thresholds for all models according the fixed FPR. To calculate this FPR, we use the negative sample, which included 5′-regions of protein-coding genes (2000 base pairs from transcription start sites).

We calculate FPR as follows. The scores of a model we determine for each site in the negative sample at each position and DNA strand. Then, the FPR for each unique

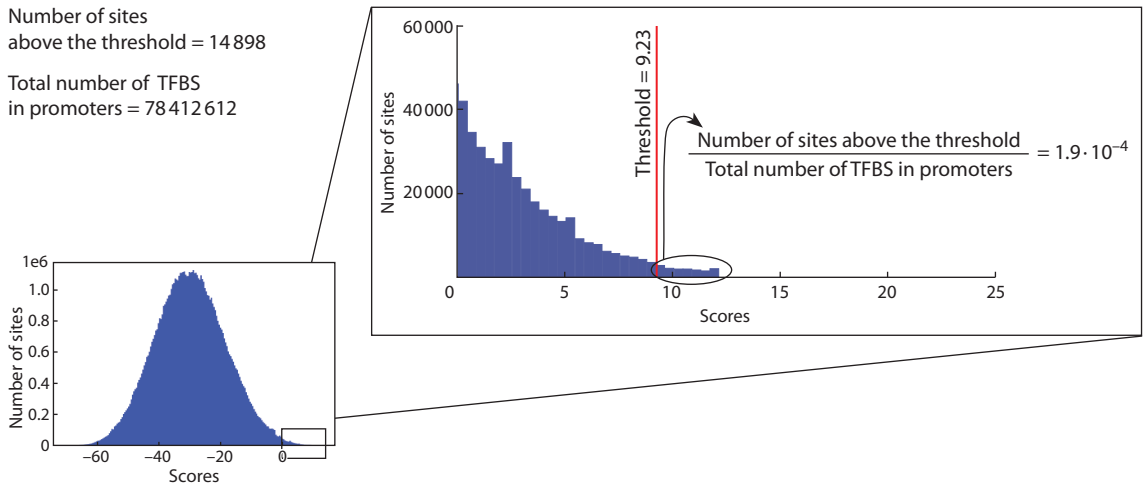


Fig. 2. The approach of threshold selection for a model through estimation of false positive rate with the whole-genome promoter dataset.

score threshold we calculate as the ratio of the total count of predicted TFBS, for which the score is the same or higher than this threshold, to the total number of positions in the sequence sample available for such TFBS. We choose for recognition of TFBS in peaks thresholds for all models respecting the FPR $1.9 \cdot 10^{-4}$. An example of choosing a threshold for PWM for the GSE92491 dataset is shown in Fig. 2.

Classification of ChIP-seq peaks based on the results of TFBS recognition by different models. After threshold selection for all models, we search TFBS in ChIP-seq peaks. Further, these peaks we classify into fractions depending on the presence/absence of sites of different models (PWM, diPWM, BaMM, InMoDe). We use two types of classification. One of them take into account the location of TFBS of different models in a peak, and another did not (see previously developed method, Levitsky et al., 2014, 2016). In particular, we carry out for each pair of models the classification of peaks with taking into account positions of TFBS of different models. Totally, there are six pairs of models: PWM and diPWM, PWM and BaMM, PWM and InMoDe, BaMM and diPWM, BaMM and InMoDe, InMoDe and diPWM. If a peak includes TFBS of a single model only, then this peak we classify as the peak of the corresponding model. If there are only two different models with hits in a peak, then two outcomes are possible (Fig. 3).

In the first case, if there is at least one pair of sites from two models that has at least one common position, then such peak we classify as the "intersection". Otherwise, if a peak contains sites of different models, but these sites are not intersected, then a peak is classified as "no intersection". If sites are absent in a peak, then we classify it as "no sites". Such classification of ChIP-seq peaks for the two models can be represented as the pie chart (Fig. 4).

The classification of peaks, without taking into account positions of sites of different models we carry out as follows. We identify following groups of peaks: peaks with

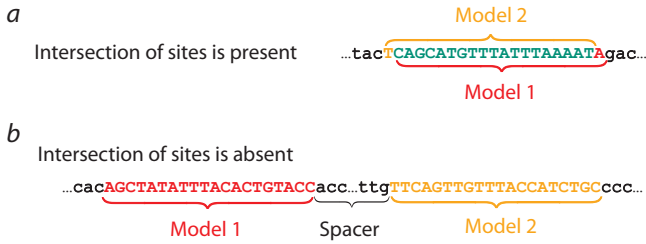


Fig. 3. The example of classification for two ChIP-seq peaks containing sites of two various models (Model 1, Model 2). Colors mark options of intersected (*a*) or not intersected sites (*b*).

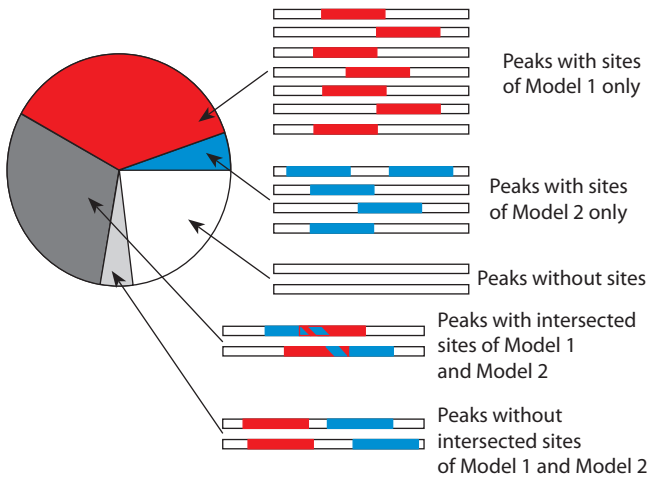


Fig. 4. Peak classification for two models (Model 1, Model 2) with taking into account the intersection of TFBS.

sites of one model only, peaks containing sites of all models, and also several groups of peaks respecting combination of various models.

Comparison of found TFBS with known ones using TomTom tool. To assess whether a predicted site matches to known FOXA2 sites, we use the TomTom motif

comparison program (Gupta et al., 2007). This program is designed to assess the similarity between nucleotide frequency matrices. For each PWM model, we construct a nucleotide frequency matrix based on the sites it find. Next, using TomTom, we evaluate the similarity of this matrix to the frequency matrix of the FOXA2 from the HOCOMOCO database (ID HOCOMOCO FOXA2_HUMAN.H11MO.0.A, Kulakovskiy et al., 2018). If the *p*-value of the matrix comparison is below 0.001, then a ChIP-seq dataset we consider as enriched with FOXA2 BS (see the Table).

Statistical data analysis. Data analysis and visualization we perform in the Python 3.8 programming language in the Jupyter environment using the numpy, matplotlib, seaborn, and statannot packages. The distributions of peak fractions respecting to various models we compare with the Mann–Whitney U-test, corrected for multiple comparisons (Bonferroni approach).

Results

Filtering data based on TomTom's motif comparison

To ensure that the trained models find sites corresponding to known FOXA2 sites we apply the filter based on the TomTom program. For this, we build the frequency matrices respecting a trained model and we compare it with the known matrix of FOXA2 from the HOCOMOCO database. This procedure left only 16 ChIP-seq datasets out of total 22 (see the Table), therefore, these 16 sets we use in further analysis.

Classification of ChIP-seq peaks without taking into account the intersection of TFBS positions found by different *de novo* models

The main result of MultiDeNA pipeline is the classification of peaks. It allows establishing how the models are related to each other in terms of their ability to identify TFBS in peaks. We used two types of peak classification. The first one takes into account an intersection of positions of predicted TFBS of different models, the second one did not take it into account (see "Classification of ChIP-seq peaks based on the results of TFBS recognition by different models"). The example of results classification for GSE90454.FOXA2.KerCT dataset is given in Fig. 5.

Let us consider in more detail the classification of ChIP-seq peaks based on the results of the TFBS search with four models without taking into account site positions. It can be seen that all models jointly recognized 88.35 % of the peaks (3534 out of 4000, the sum of all areas within the Venn diagram, see Fig. 5, *a*, *b*). The overlap fraction of all models amounts 34.25 % (1370 out of 4000 peaks). Two non-PWM models BaMM and InMoDe make the significant contributions to peak recognition. They totally add 34.55 % of all peaks (696 + 647 + 39 = 1382 out of 4000). This fraction is almost the same as the overlap fraction of all models (1370). The BaMM model makes the largest independent contribution to recognition of sites, it adds 17.4 % of the peaks (696), in contrast to other models that add 0.525 % (21), 0.975 % (39) and 0.2 % (8) (PWM, InMoDe and diPWM respectively).

To assess the structural diversity of the TFBS, we build logos for peak fractions "only PWM", "only diPWM", "only BaMM", "only InMoDe" and "all models" (see Fig. 5, *c*). All logos contain the GTAAACA consensus. However, the "only PWM", "only diPWM" and "only InMoDe" fractions have the higher occurrence of GT than AT at the first two nucleotides of the consensus. It can also be noted that the 5′-ends of all logos are diverse in nucleotide content.

Classification of ChIP-seq peaks with taking into account the intersection of TFBS positions found by different models

The classification of peaks described above (without taking into account the positions of the TFBS) does not take into account positions of sites in peaks. To consider this circumstance we classify peaks with taking into account positions. We perform this for each pair of models (PWM–diPWM, PWM–BaMM, PWM–InMoDe, diPWM–BaMM, diPWM–InMoDe, InMoDe–BaMM). The results of the classification of peaks for GSE90454.FOXA2.KerCT are shown as the pie charts in Fig. 6.

All pairs of model combinations have very small fraction of "without intersection" peaks, ranging from 0.3 to 6.9 %. On the other hand, all cases were characterized by the large fraction of peaks "with intersection" (BaMM–InMoDe 53.6 %, PWM–diPWM 44.4 %, diPWM–BaMM 41.0 %, PWM–BaMM 37.3 %, diPWM–InMoDe 35.4 %, PWM–InMoDe 31.6 %). This fraction is larger for methodologically close pairs of models BaMM–InMoDe and PWM–diPWM (see Fig. 6). The fraction of the peaks with TFBS found with only a single model is the highest for BaMM model. In pairs of models PWM–BaMM, diPWM–BaMM, and InMoDe–BaMM, BaMM contributes greatly (39.2, 36.4 and 26.8 %, respectively).

Evaluation of the recognition TFBS accuracy by different models for FOXA2

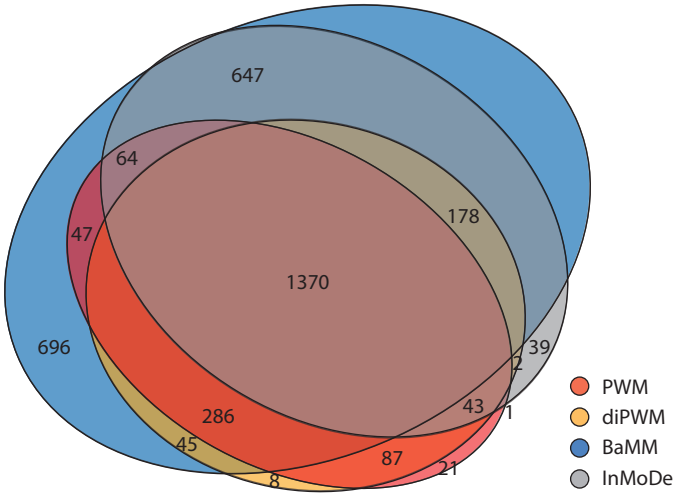
To compare the recognition accuracy of different models we calculate pAUC values from ROC curves obtained with the cross-validation procedure (see "Training *de novo* models and assessing the TFBS recognition accuracy") (Fig. 7, *a*). According to the results obtained, the values of the pAUC medians for the PWM, diPWM, BaMM and InMoDe models are 8.0E–4, 8.1E–4, 7.3E–4, and 5.6E–4, respectively. The differences between pAUC values were not significant ($p > 0.05$) for paired comparisons of PWM, diPWM, and BaMM, but the InMoDe model has significantly less values than any other model ($p < 0.05$).

Comparison of fractions of peaks with TFBS found by each model with that for all models. To investigate contributions of different models to the efficiency of TFBS search and to evaluate the overall result of several models, we determine fractions of peaks containing at least

a

Peaks with sites found by														
all models	all models				only two models						only one model			
	except PWM	except diPWM	except BaMM	except InMoDe	PWM diPWM	PWM BaMM	PWM InMoDe	diPWM BaMM	diPWM InMoDe	BaMM InMoDe	PWM	diPWM	BaMM	InMoDe
1370	178	64	43	286	87	47	1	45	2	647	21	8	696	39

b



c

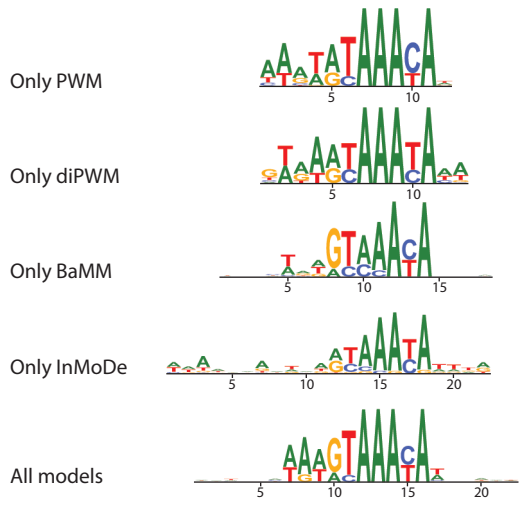


Fig. 5. The classification of peaks of the GSE90454.FOXA2.KerCT ChIP-seq dataset according to prediction results of all four models.
(*a*) Table, (*b*) Venn diagram, (*c*) Logo for fraction of the peaks respecting to predictions of sole models and that for the overlapping fraction of all models.

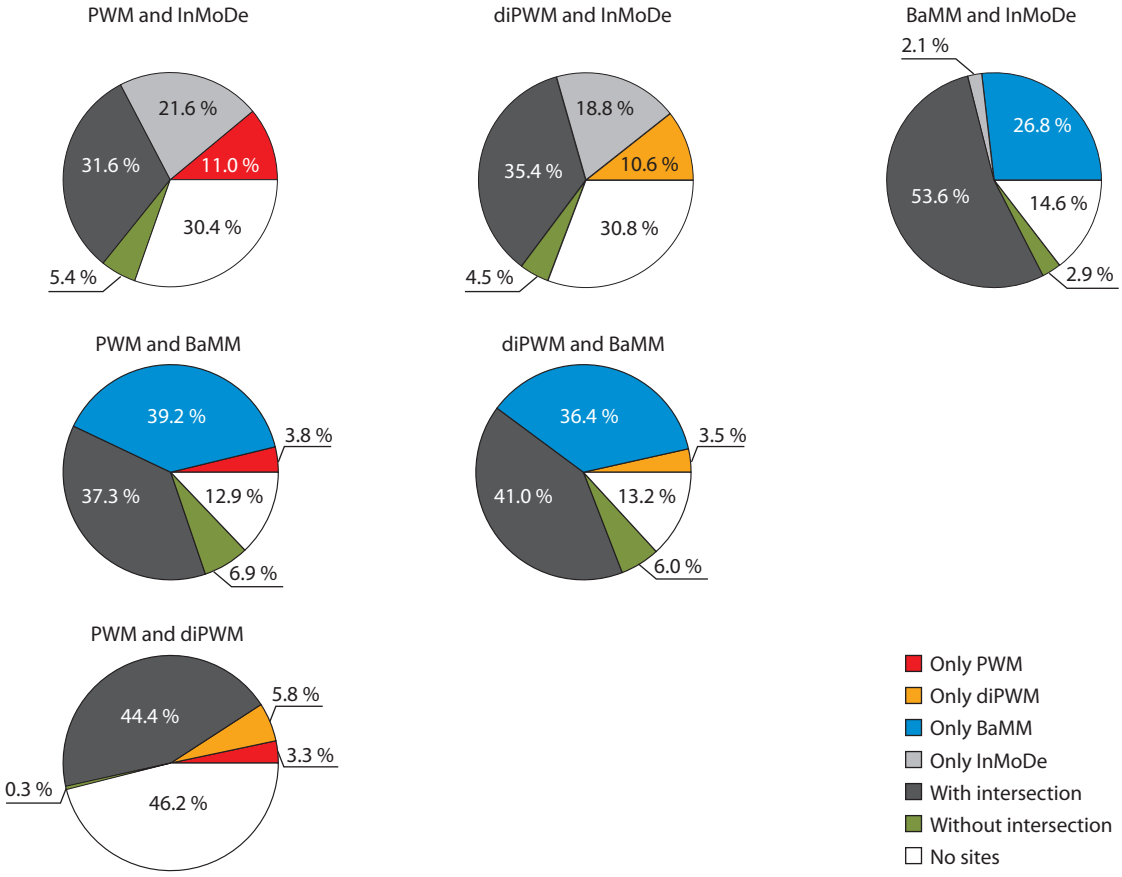


Fig. 6. Classification of the GSE90454.FOXA2.KerCT ChIP-seq dataset with taking into account intersection of TFBS positions respecting to different models.

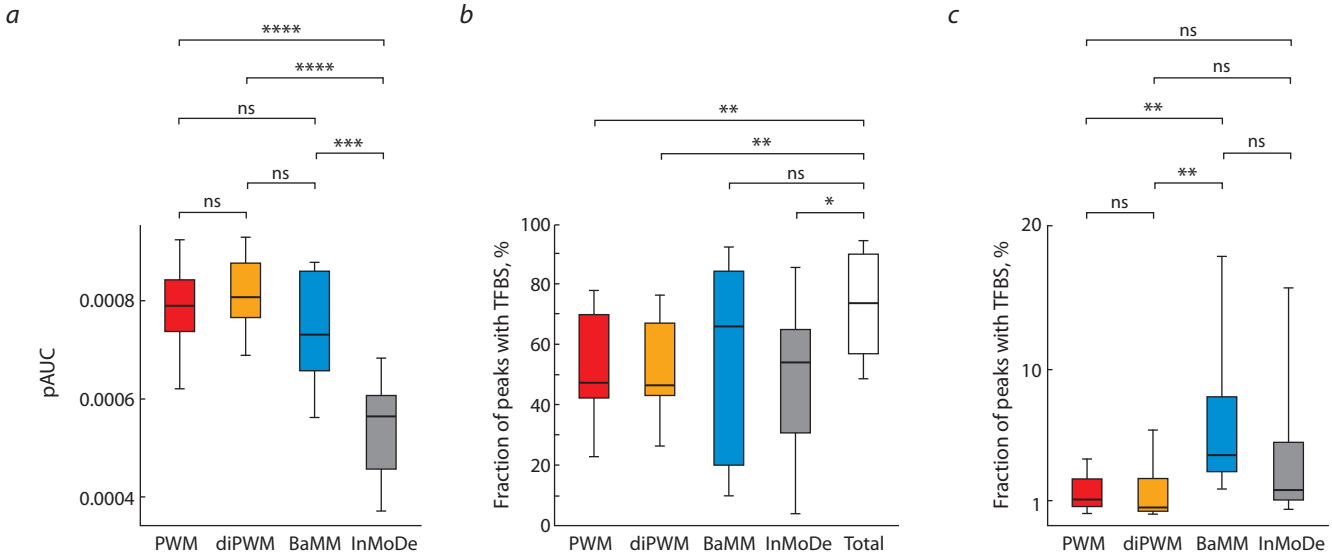


Fig. 7. The distribution of quartiles for recognition measures. The bottom part of the box denotes the minimum value of parameter; the top part denotes the maximum value of parameter. (*a*) Values of pAUC for all models. (*b*) Fractions of peaks recognized with a single models (PWM, diPWM, BaMM, InMoDe) and with all models together (Total). (*c*) Fractions of peaks contained only TFBS recognized with a single model.

ns – $p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

one TFBS for each sole model and those for all models together (see Fig. 7, *b*). The medians of recognized peaks fractions are 47.3, 46.4, 65.8, and 54 % for sole PWM, diPWM, BaMM and InMoDe, respectively. The median of recognized peaks fraction of joined results of all four models' case is 73.6 %. Consequently, together, all models add 26.3 % peaks containing TFBS to the fraction of sole PWM model, which is consistent with the earlier obtained result of using two fundamentally different PWM and SiteGA models (Levitsky et al., 2014). At the same time, the median values respecting fractions for the PWM, diPWM, and InMoDe models significantly lower ($p < 0.05$) than that obtained by combining all models. Thus, the approach using the combination of different models allows better identification of peaks with TFBS for FOXA2 than that using only one model. However, for BaMM, the fraction of recognized peaks did not statistically differ ($p > 0.05$) from the result obtained by combining the four models. Hence, the BaMM model makes the main contribution to the recognition of FOXA2 peaks and, among the other models this model better describes the structure of FOXA2 sites. However, the rest three models add 7.8 % of the peaks to the BaMM result, which proves the effectiveness of using different models together.

Comparison of fractions of peaks containing TFBS found by single models. As it is shown above, the combination of different models increases the number of peaks with TFBS. Hence, each model recognizes TFBS that others do not. To assess the independent contributions of all models to the search for TFBS, we determine the fractions of peaks containing TFBS of only one model (see Fig. 7, *c*). As one can see, each model (PWM, diPWM, BaMM, InMoDe) is able to find TFBS that other models do not find. The medians of peaks containing TFBS respecting a single model are 1.08, 0.49, 4.15, and 1.73 %, respectively for PWM, diPWM, BaMM, and InMoDe. At the same time, the results for BaMM are significantly different ($p < 0.05$) from those for both PWM and diPWM. It also confirms the assumption that the BaMM model better recognizes FOXA2 sites. However, each model contributes to site recognition. Consequently, each model reveals certain structural variant of TFBS.

Cross-validation test for PWM models with participation of their own training dataset and other ChIP-seq datasets

To estimate the dependence of specificity of various models for different ChIP-seq datasets as a function of a selection of particular dataset as the training sample, we performed the cross-validation test as follow. The accuracy of each PWM model we assessed not only within the same ChIP-seq training dataset, but also for the rest 15 datasets (control datasets). For the case of training dataset, we performed several iterations to divide the total training sample into 90 % of the peaks that we used to build a model, and the remaining 10 % of the peaks we used to estimate the performance. For each case we calculated the accuracy estimate pAUC (see "Training *de novo* models and assessing the TFBS recognition accuracy"), the results we presented in the form of the heatmap (Fig. 8). The heatmap shows that only in three cases ENCSR000BRE.A-549, ENCSR000BNI.Hep-G2 and ERP008682.pancreas other models have very low pAUC scores, and for five cases GSE90454.A-549, ENCSR066EBK.Hep-G2, GSE90454.KerCT, ENCSR080XEY.liver and ENCSR310NYI.liver, all models have high pAUC values.

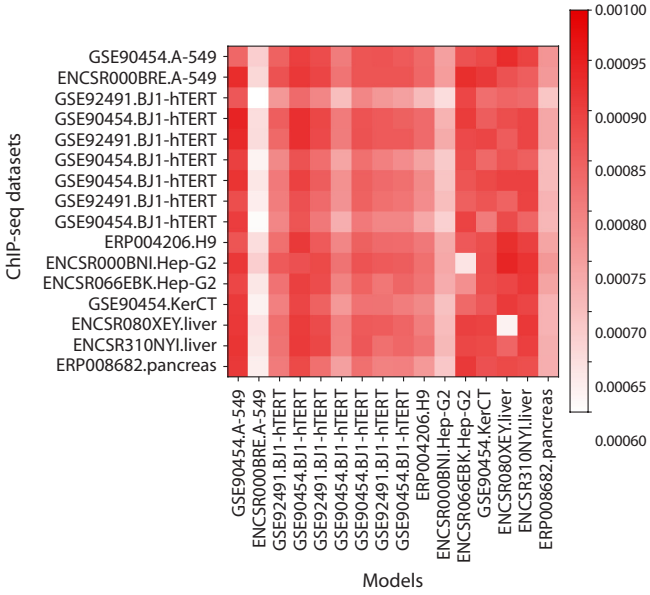


Fig. 8. The heatmap of cross-validation test results for PWM models.

Colors mark pAUC values. Each diagonal cell implies that control and training datasets are the same. Remaining cells refer to distinct training and control datasets. Rows mean models and columns denote ChIP-seq datasets.

Discussion

Based on all data obtained, we conclude that the joint use of alternative models allows us to expand the number of detected peaks containing TFBS relative to application of sole PWM.

This result can be explained by the presence of different structural types of TFBS of FOXA2. This is in agreement with experimental data obtained for a number of other TFs, including members of the FOX family. Thus, it was shown that HOXB13 and FOXC2 are able to bind with the same affinity to completely different sequences CAATAAA/TCGTAAA (Morgunova et al., 2018) and GTAAACA/ACAAATA (Chen et al., 2019), respectively. It was recently found that TF FOXN3 is able to bind to two fundamentally different types of TFBS, which had different lengths (Rogers et al., 2019). In addition, small changes in the structure of the TFBS depend on the cooperative interaction between TFs (Morgunova, Taipale, 2017). Hence, we propose that FOXA2 also can bind to different structural types of BS.

To take into account all the TFBS structural types, the only PWM model for site recognition may not be enough, this problem partially was solved using several PWMs (Bi et al., 2011; Mitra et al., 2018) or using alternative models (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). However, previously alternative models were usually compared with PWM only in terms of the recognition accuracy (Siebert, Söding, 2016), or according the number of recognized TFBSs (Samee et al., 2019). In the current study, we took in analysis FOXA2 ChIP-seq data. We compared not only the accuracy and the number of peaks recognized, but also we estimated independent contributions of each model and assessed the joint contribution for all pairs of models, and also we tested positions of hits in peaks for each pair of models. The results for the accuracy assessment (see Fig. 7, *a*) showed that the InMoDe model had the lowest accuracy relative to other models, and the BaMM, diPWM and PWM models were comparable in accuracy. In terms of expanding the total fraction of peaks with TFBS, the BaMM model performed the best, since this model found the largest fraction of peaks with TFBS that other models do not find. Nevertheless, all alternative diPWM, BaMM and InMoDe models allow expanding the pool of recognized TFBS relative to sole PWM, but PWM also makes an independent contribution to the total number of peaks with recognized TFBS.

Conclusion

We have developed the pipeline MultiDeNA, which allows uniform processing of ChIP-seq data using different TFBS models. Currently, it can be used to build PWM, diPWM, InMoDe, BaMM models. MultiDeNA includes the steps of preparing data, building models, evaluating recognition accuracy, scanning peaks, combining results, and analyzing them. The developed pipeline of programs processed datasets from the ReMap database, including 22 ChIP-seq experiments for TF FOXA2. We have shown that combined use of different models allows increasing the total fraction of recognized peaks up to 73.6 % (relative to sole PWM model, the fraction of recognized peaks increased by 26.3 %). We have shown that different models tend to recognize the same sites of FOXA2 in the large fraction of peaks, thereby revealing the most common structural type of TFBS in these peaks. Also, each model found TFBS that other models did not predict. The BaMM model performed the best with 4.15 % of peaks containing only its sites, versus 1.08, 0.49, 1.73 % for PWM, diPWM and InMoDe, respectively. We proposed that the heterogeneity of sites for FOXA2 is revealed only if two or more models are applied. The diPWM model showed worst result in sole application in comparison with other models (diPWM recognized TFBS in 46.4 % of the peaks). The best model for the FOXA2 sites was BaMM; it found TFBS in 65.8 % of the peaks. Hence, we assumed that the BaMM model could better describe BS for FOXA2.

References

Bailey T.L., Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proc. Int. Conf. Intell. Syst. Mol. Biol. 1994;2:28-36. DOI citeulike-article-id:878292. PMID 7584402.

Benos P.V., Bulyk M.L., Stormo G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* 2002;30(20):4442-4451. DOI 10.1093/nar/gkf578.

Bi Y., Kim H., Gupta R., Davuluri R.V. Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One.* 2011;6(9):e24210. DOI 10.1371/journal.pone.0024210.

Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002;30(5):1255-1261. DOI 10.1093/nar/30.5.1255.

Chen X., Wei H., Li J., Liang X., Dai S., Jiang L., Guo M., Qu L., Chen Z., Chen L., Chen Y. Structural basis for DNA recognition by FOXC2. *Nucleic Acids Res.* 2019;47(7):3752-3764. DOI 10.1093/nar/gkz077.

Chèneby J., Ménétrier Z., Mestdagh M., Rosnet T., Douida A., Rhalloussi W., Bergon A., Lopez F., Ballester B. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 2020;48(D1):D180-D188. DOI 10.1093/nar/gkz945.

Eggeling R., Grosse I., Grau J. InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics.* 2017;33(4):580-582. DOI 10.1093/bioinformatics/btw689.

Farnham P.J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 2009;10(9):605-616. DOI 10.1038/nrg2636.

Furey T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* 2012;13(12):840-852. DOI 10.1038/nrg3306.

Gheorghe M., Sandve G.K., Khan A., Chèneby J., Ballester B., Mathelier A. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* 2019;47(4):e21. DOI 10.1093/nar/gky1210.

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

Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 2010;38(4):576-589. DOI 10.1016/j.molcel.2010.05.004.

Ignatieva E.V., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina T.V., Merkulova T.I. Comparison of the results of search for the SF-1 binding sites in the promoter regions of the steroidogenic genes, using the SiteGA and SITECON methods. In: Proc. Fourth Int. Conf. Bioinform. Genome Regul. Struct. (BGRS). 2004;1:69-72.

Iwafuchi-Doi M. The mechanistic basis for chromatin regulation by pioneer transcription factors. *WIREs Syst. Biol. Med.* 2019;11(1):e1427. DOI 10.1002/wsbm.1427.

Keilwagen J., Grau J. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* 2015;43(18):e119. DOI 10.1093/nar/gkv577.

Kiesel A., Roth C., Ge W., Wess M., Meier M., Söding J. The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.* 2018;46(W1):W215-W220. DOI 10.1093/nar/gky431.

Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics.* 2010;26(20):2622-2623. DOI 10.1093/bioinformatics/btq488.

Kulakovskiy I., Levitsky V., Oshchepkov D., Bryzgalov L., Vorontsov I., Makeev V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.* 2013;11(01):1340004. DOI 10.1142/S0219720013400040.

Kulakovskiy I.V., Makeev V.J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics* (*Oxf.*). 2009;54(6):667-674. DOI 10.1134/S0006350909060013.

Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.

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.

Latchman D.S. Transcription factors: bound to activate or repress. *Trends Biochem. Sci.* 2001;26(4):211-213. DOI 10.1016/S0968-0004(01)01812-6.

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C.T. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinform.* 2007;8(1):1-20. DOI 10.1186/1471-2105-8-481.

Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.

Levitsky V.G., Oshchepkov D.Y., Klimova N.V., Ignatieva E.V., Vasiliev G.V., Merkulov V.M., Merkulova T.I. Hidden heterogeneity of transcription factor binding sites: a case study of SF-1. *Comput. Biol. Chem.* 2016;64:19-32. DOI 10.1016/j.compbiolchem.2016.04.008.

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

Machanick P., Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics.* 2011;27(12):1696-1697. DOI 10.1093/bioinformatics/btr189.

Mathelier A., Wasserman W.W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 2013;9(9):e1003214. DOI 10.1371/journal.pcbi.1003214.

McClish D.K. Analyzing a portion of the ROC curve. *Med. Decis. Mak.* 1989;9(3):190-195. DOI 10.1177/0272989X8900900307.

Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol.* 2018;14(4):1-20. 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.

Morgunova E., Yin Y., Das P.K., Jolma A., Zhu F., Popov A., Xu Y., Nilsson L., Taipale J. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife.* 2018;7:1-21. DOI 10.7554/eLife.32963.

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

Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842. DOI 10.1093/bioinformatics/btq033.

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.DOI 10.1016/j.molcel. 2019.01.019.

Samee M.A.H., Bruneau B.G., Pollard K.S. A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.* 2019;8(1):27-42. DOI 10.1016/j.cels.2018.12.001.

Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.* 2016;44(13):6055-6069. DOI 10.1093/nar/gkw521.

Srivastava D., Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta – Gene Regul. Mech.* 2020;1863(6):e194443. DOI 10.1016/j.bbagrm.2019.194443.

Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics.* 2000;16(1):16-23. DOI 10.1093/bioinformatics/16.1.16.

Wallerman O., Motallebipour M., Enroth S., Patra K., Bysani M.S.R., Komorowski J., Wadelius C. Molecular interactions between

HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res.* 2009;37(22):7498-7508. DOI 10.1093/nar/gkp823.

Wederell E.D., Bilenky M., Cullum R., Thiessen N., Dagpinar M., Delaney A., Varhol R., Zhao Y., Zeng T., Bernier B., Ingham M., Hirst M., Robertson G., Marra M.A., Jones S., Hoodless P.A. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* 2008;36(14):4549-4564. DOI 10.1093/nar/gkn382.

Worsley Hunt R., Wasserman W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* 2014;15(7):412. DOI 10.1186/s13059-014-0412-4.

Yang L., Zhou T., Dror I., Mathelier A., Wasserman W.W., Gordân R., Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* 2014;42(D1):D148-D155. DOI 10.1093/nar/gkt1087.

Zhang M.O., Marr T.G. A weight array method for splicing signal analysis. *Bioinformatics.* 1993;9(5):499-509. DOI 10.1093/bioinformatics/9.5.499.

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

ORCID ID

A.V. Tsukanov orcid.org/0000-0002-5174-6609
V.G. Levitsky orcid.org/0000-0002-4905-3088

Acknowledgements. This work was supported by the Russian Foundation for Basic Research No. 18-29-13040 and the state budget project No. 0259-2019-0008.

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

Received October 10, 2020. Revised January 10, 2021. Accepted January 12, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):18-29
DOI 10.18699/VJ21.003

Original Russian text www.bionet.nsc.ru/vogis/

Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms

E.V. Ignatieva[1, 2] ✉, E.A. Matrosova[1, 2]

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

Abstract. Whole genome and whole exome sequencing technologies play a very important role in the studies of the genetic aspects of the pathogenesis of various diseases. The ample use of genome-wide and exome-wide association study methodology (GWAS and EWAS) made it possible to identify a large number of genetic variants associated with diseases. This information is accumulated in the databases like GWAS central, GWAS catalog, OMIM, ClinVar, etc. Most of the variants identified by the GWAS technique are located in the noncoding regions of the human genome. According to the ENCODE project, the fraction of regions in the human genome potentially involved in transcriptional control is many times greater than the fraction of coding regions. Thus, genetic variation in noncoding regions of the genome can increase the susceptibility to diseases by disrupting various regulatory elements (promoters, enhancers, silencers, insulator regions, etc.). However, identification of the mechanisms of influence of pathogenic genetic variants on the diseases risk is difficult due to a wide variety of regulatory elements. The present review focuses on the molecular genetic mechanisms by which pathogenic genetic variants affect gene expression. At the same time, attention is concentrated on the transcriptional level of regulation as an initial step in the expression of any gene. A triggering event mediating the effect of a pathogenic genetic variant on the level of gene expression can be, for example, a change in the functional activity of transcription factor binding sites (TFBSs) or DNA methylation change, which, in turn, affects the functional activity of promoters or enhancers. Dissecting the regulatory roles of polymorphic loci have been impossible without close integration of modern experimental approaches with computer analysis of a growing wealth of genetic and biological data obtained using omics technologies. The review provides a brief description of a number of the most well-known public genomic information resources containing data obtained using omics technologies, including (1) resources that accumulate data on the chromatin states and the regions of transcription factor binding derived from ChIP-seq experiments; (2) resources containing data on genomic loci, for which allele-specific transcription factor binding was revealed based on ChIP-seq technology; (3) resources containing *in silico* predicted data on the potential impact of genetic variants on the transcription factor binding sites.
Key words: transcription regulation; genetic variability; pathogenic genetic variants; transcription regulatory regions; transcription factor binding sites (TFBSs); genomic databases.

For citation: Ignatieva E.V., Matrosova E.A. Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):18-29. DOI 10.18699/VJ21.003

Геномная изменчивость в регуляторных районах генов, ассоциированная с заболеваниями человека: механизмы влияния на транскрипцию генов и полногеномные информационные ресурсы, обеспечивающие исследование этих механизмов

Е.В. Игнатьева[1, 2] ✉, Е.А. Матросова[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
✉ eignat@bionet.nsc.ru

Аннотация. Полногеномные и полноэкзомные технологии секвенирования играют важную роль в исследованиях генетических аспектов патогенеза различных заболеваний. Широкое применение методов полногеномного и полноэкзомного анализа ассоциаций позволяет идентифицировать множество вариантов геномной изменчивости (ГИ), ассоциированных с заболеваниями. Эта информация накапливается в базах данных GWAS central, GWAS catalog, OMIM, ClinVar и др. Большинство вариантов, идентифицированных методикой полногеномного анализа ассоциаций, располагается в некодирующих областях генома человека. По данным проекта ENCODE, доля участков в геноме человека, потенциально задействованных в регуляции транскрипции, во много раз превышает долю кодирующих областей. Таким образом, геномная изменчивость в некодирующих областях генома может повышать предрасположенность к заболеваниям, нарушая функционирование различных регуляторных элементов (промоторов, эн-

хансеров, участков, определяющих 3D структуру хроматина и т. д.). Однако идентификация механизмов влияния патогенных вариантов ГИ на риск развития заболеваний затруднена ввиду большого разнообразия регуляторных элементов. В обзоре рассмотрены молекулярно-генетические механизмы влияния патогенных вариантов ГИ на экспрессию генов. При этом внимание сосредоточено на транскрипционном уровне регуляции как ключевой стадии, запускающей последовательность этапов экспрессии любого гена. Пусковым событием, опосредующим влияние патогенного варианта ГИ на уровень экспрессии гена, может быть, например, изменение функциональной активности сайтов связывания транскрипционных факторов или уровня метилирования ДНК, что, в свою очередь, отражается на функциональной активности промоторов или энхансеров. Выявление регуляторных эффектов полиморфных локусов невозможно без тесной интеграции современных экспериментальных подходов с компьютерным анализом больших массивов генетических данных, получаемых на основе омиксных технологий. В обзоре кратко описаны наиболее известные открытые полногеномные информационные ресурсы, содержащие данные, полученные на основе омиксных технологий, в том числе: ресурсы, накапливающие сведения о состоянии хроматина и участках его связывания с транскрипционными факторами, выявленными с помощью технологии ChIP-seq; ресурсы по геномным локусам, для которых на основе данных ChIP-seq выявлено аллель-специфичное связывание с транскрипционными факторами; а также ресурсы, содержащие предсказанные *in silico* данные о потенциальном влиянии геномной изменчивости на сайты связывания транскрипционных факторов.
Ключевые слова: регуляция транскрипции; геномная изменчивость; патогенные геномные варианты; районы, регулирующие транскрипцию; сайты связывания транскрипционных факторов; геномные базы данных.

Introduction

At present, largely due to the widespread use of the technology of genome-wide and exome-wide association study (GWAS and EWAS), a large number of polymorphisms associated with diseases have been identified. For example, GWAS central (https://www.gwascentral.org/) contains information on more than 70 million associations between ~3.2 million SNPs and 1451 diseases or phenotypic characteristics (Beck et al., 2020). Experimental datasets of comparable volume have been accumulated in a number of other databases on genotype-phenotype associations (GWAS catalog, OMIM, ClinVar, HGMD, PheGenI, EGA, GAD, dbGaP).

Currently, a large amount of experimental data has been obtained about the disease-associated genetic variants (GVs), but the molecular mechanisms underlying these associations are extremely poorly understood. This is due to the fact that only a relatively small proportion of pathogenic GVs is located in the coding regions of the human genome, changes in the nucleotide sequence of which disrupt the structure and function of proteins. A huge mass of polymorphic loci associated with diseases is located in non-coding regions of the genome (introns, 5′- and 3′-flanking regions of genes, intergenic regions). For example, according to GWAS data, ~90 % of the total number of variants associated with diseases are located in noncoding regions of the human genome (Maurano et al., 2012; Farh et al., 2015).

It is known that non-coding regions of the genome contain regions that perform a wide range of regulatory functions: promoter regions, enhancers, negative regulatory elements, nuclear matrix attachment regions, regions that determine the structure of topologically associating domains (TADs), and other features of 3D organization of genome (Mathelier et al., 2015; Meddens et al., 2019; Ibrahim, Mundlos, 2020). The proportion of regions in the human genome potentially involved in the transcriptional regulation is extremely high. According to the ENCODE project, the chromatin regions corresponding to the peaks of transcription factor (TF) binding identified by the ChIP-seq occupy ~8.1 % of the total genomic DNA (ENCODE Project Consortium, 2012), which is significantly higher than the proportion of coding regions of the human genome (~1.2 %). Considering that not all known TFs and not all cell lines were studied in the ENCODE project, an obviously larger fraction of genomic DNA is involved in the interaction with TFs. The total length of human genome regions with enhancer-associated chromatin features also significantly exceeds the total size of the coding regions: for example, in only one cell type studied (H1-ES), enhancer regions occupy ~3.2 % (Roadmap Epigenomics Consortium et al., 2015).

Studies aimed at identifying the mechanisms of the influence of pathogenic GVs on the predisposition to diseases are carried out very actively, which is reflected in a number of review publications (Mathelier et al., 2015; Merkulov et al., 2018; Smith et al., 2018; Wang et al., 2019; Vohra et al., 2020). The most discussed effect of pathogenic GVs is a change in the binding activity of TFBSs (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Mathelier et al., 2015; Gorbacheva et al., 2018). It has also been shown that polymorphic loci can be associated with alteration of DNA methylation patterns (Howard et al., 2014; Kumar D. et al., 2017; Rahbar et al., 2018; Schmitz et al., 2019) and modifications of histone proteins (Kilpinen et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019), with structural change in chromatin loops (Visser et al., 2015; Zhang et al., 2018) and, as one of the manifestations of this process, with changes in the TADs structure (Cong et al., 2019; Mei et al., 2019). Examples of such effects will be discussed below (Table 1).

The effects of genetic variants on the functional activity of transcription factor binding sites

The key role in the transcriptional regulation is played by transcription factors – proteins that can specifically bind to DNA of the regulatory regions of genes and to initiate the transcription complexes formation. The human genome contains more than 1500 genes encoding TFs (Wingender et al., 2013). TF binding sites, as a rule, have a length of 10–25 nucleotides (Levitsky et al., 2014; Kulakovskiy et al., 2018).

Nucleotide substitutions, as well as short insertions/deletions at polymorphic loci, can disrupt TFBSs or create them *de novo* (see Table 1), and this, in turn, can have both negative and positive effects on the level of gene transcription (Chen L. et al., 2013; Gorbacheva et al., 2018). Such GVs (and the cor-

Table 1. Examples of polymorphic loci associated with pathologies and mechanisms of their action on the gene expression level

Disease or pathology	Polymorphic locus	Localization	Mechanism	Reference
Atopic asthma	rs928413 A→G	*IL33* promoter region	rs928413(G) allele creates a binding site for the transcription factor CREB1 leading to increased expression level of *IL33*	Gorbacheva et al., 2018
Obesity	rs1421085 T→C	Intron of the *FTO* gene which contains the regulatory region of the *IRX3* and *IRX5* genes (the distance between rs1421085 and TSSs of *IRX3* and *IRX5* is ~517,000 and ~1,164,000 bases)	rs1421085(C) variant disrupts a conserved motif for the ARID5B repressor, which leads to derepression of a potent preadipocyte enhancer and a doubling of *IRX3* and *IRX5* expression	Claussnitzer et al., 2015
Pancreatic cancer	rs2001389 A→G	The boundary of TAD located on chromosome 10	The allele G of rs2001389 weakens the CTCF binding activity of DNA, eliminating TAD boundary and altering 3D chromatin structure, and it is related to the lower expression of a putative antioncogene *MFSD13A*	Mei et al., 2019
Disturbances of lipid metabolism	rs174537 G→T	An enhancer region of the FADS cluster	Individuals that have rs174537(T) allele exhibited a higher level of DNA methylation at CpG sites located within regulatory region of FADS cluster, which led to a decrease in transcriptional activity of *FADS1* and *FADS2*	Howard et al., 2014
Atopic dermatitis	rs612529 T→C	*VSTM1* promoter region	The rs612529(T) allele facilitates binding of the transcription factor PU.1, that acts as docking site for DNA demethylases. In carriers of pathogenic variant C, the interaction of PU.1 with DNA is disrupted, as a result, the methylation level of the *VSTM1* promoter is elevated, and this is accompanied by a downregulation of *VSTM1* expression	Kumar D. et al., 2017
Fragile X syndrome	CGG repeat expansion. Healthy individuals harbor between 5 and 55 copies of the CGG repeats, affected patients harbor more than 100 copies	The 5′-untranslated region of *FMR1* gene	CGG repeat expansion disrupts the structure of TAD, that includes *FMR1*. In individuals with mutation-length CGG triplet repeats, the 5′-boundary region of TAD is ablated (this region is hypermethylated and its CTCF occupancy is lost). As a result, one subTAD dissolves. *FMR1*, which is normally associated with the downstream TAD, shifts to the upstream TAD. In this case, *FMR1* promoter is hypermethylated, and *FMR1* expression is downregulated	Sun et al., 2018
Rheumatoid arthritis and type-2 diabetes mellitus	rs7873784 G→C	The 3′-untranslated region of *TLR4* gene	rs7873784(C) allele creates a binding site for transcription factor PU.1, a known regulator of *TLR4* expression. Functional PU.1 binding site augments the enhancer activity of *TLR4* 3′-UTR that leads to increased *TLR4* expression	Korneev et al., 2020
Breast cancer	rs4321755 C→T	Enhancer region of *MRPS30* and *RP11-53O19.1* genes	The risk allele rs4321755(T) creates a GATA3 binding motif within an enhancer, resulting in stronger binding of GATA3 and chromatin accessibility, thereby activating interaction between the enhancer and *MRPS30/RP11-53O19.1* divergent promoter and increasing the expression of *MRPS30* and *RP11-53O19.1* genes	Zhang et al., 2018

responding polymorphic loci) that affect the transcriptional activity of genes are usually called regulatory variants (Kumar S. et al., 2017; Guo, Wang, 2018; Merkulov et al., 2018).

Pathological (that is, associated with a disease) can be both an allelic variant of the DNA sequence containing a disrupted TFBS (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Kumar D. et al., 2017; Mei et al., 2019) and an allelic variant, leading to creation of TFBS *de novo* (Gorbacheva et al., 2018; Zhang et al., 2018; Korneev et al., 2020) (see Table 1).

Pathological GVs, affecting the binding activity of TFBSs, can be located not only in promoter regions, but also in regulatory regions located at considerable distance from transcription start sites (TSSs) of genes: enhancers (Lewinsky et al., 2005; Zhang et al., 2018; Meddens et al., 2019), regulatory regions with repressive function (Claussnitzer et al., 2015),

and TAD boundary regions (Mei et al., 2019) (see Table 1). For example, the rs1421085 T→C substitution associated with obesity impairs the functioning of the negative regulatory region controlling expression of the *IRX3* and *IRX5* genes (Claussnitzer et al., 2015). The rs1421085 locus is located in the intron of the *FTO* gene (Fig. 1) at a considerable distance from the transcription start sites of *IRX3* and *IRX5* (~520,000 and ~1,164,000 bases). Normally, the DNA region containing allele T interacts with a repressor factor ARID5B, leading to a decrease in transcriptional activity of *IRX3* and *IRX5* genes. In carriers of the mutant variant of the DNA sequence (allele C), the binding site of the ARID5B repressor factor is disrupted, which causes an excessively high expression of the *IRX3* and *IRX5* genes and activates adipogenesis (Claussnitzer et al., 2015).

Occasionally a nucleotide substitution at a polymorphic locus disrupts the TFBS and this, in turn, affects the functional activity of the TAD (see Table 1). This effect was found in the case of A→G (rs2001389), associated with the risk of pancreatic cancer (Fig. 2). The rs2001389 locus is located in the region that determines the structure of chromatin loops within the TAD. This TAD contains 91 genes and is formed by spatially adjacent chromatin regions (Mei et al., 2019). The DNA region containing the risk allele G is characterized by a reduced ability to interact with CTCF, which in this case acts as a structural protein of chromatin. Normally, CTCF binding ensures the functioning of one of the regions that determines the structure of chromatin loops within the considered TAD. The pathogenic allele G alters the activity of CTCF binding motif within TAD boundary disrupting the stability of corresponding 3D structure of chromatin. As a result, the expression of the genes within this TAD is impaired. In this case, the greatest decrease in *MFSD13A* expression is observed.

The effects of genetic variability on DNA methylation and gene transcriptional activity

DNA methylation doesn't change the nucleotide sequence and is the addition of a methyl group to the fifth carbon atom of cytosine (Angeloni, Bogdanovic, 2019). An increase in the level of DNA methylation, as a rule, leads to a long-term inactivation of the expression of genes lying in the methylated region, since, according to the generally accepted concept, methylation of a DNA region facilitates recruiting protein complexes, containing histone deacetylase (HDAC) (Jones et al., 1998; Nan et al., 1998). DNA methylation can also decrease the ability of some TFs to interact with DNA: it is known that CTCF factors and factors from the ETS family have such sensitivity to methylation (Wang et al., 2019). In contrast, another transcription factor, ZFP57, binds only to methylated DNA (Quenneville et al., 2011). Thus, cytosine methylation can activate different mechanisms of gene transcription regulation, and not always an increase in the methylation level of the regulatory DNA region is associated with a decrease in the expression of the corresponding gene (Izzi et al., 2016; Wang et al., 2019).

Genetic variability affects significantly the methylation of DNA regions that have regulatory potential. Thus, a genome-wide analysis of the methylation patterns of DNA collected from 24 subjects from Norfolk Island genetic isolate (Benton

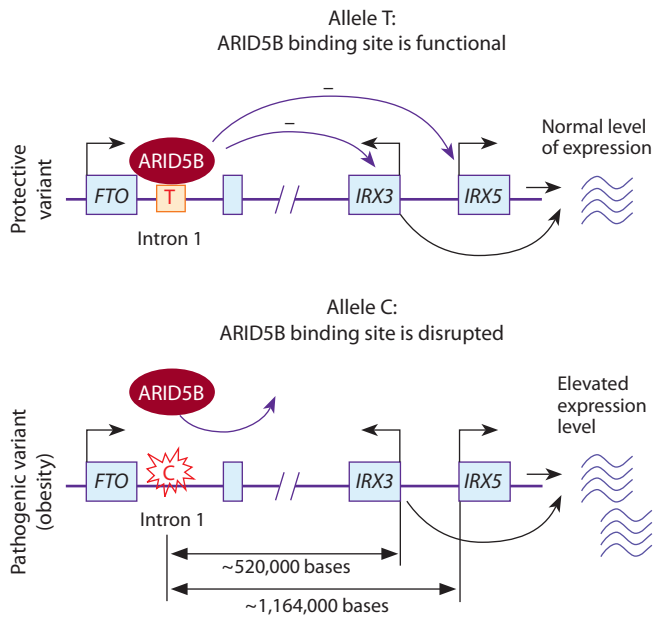


Fig. 1. Disruption of the binding site caused by T→C substitution (rs1421085) weakens ARID5B repressor binding to the regulatory region of the *IRX3* and *IRX5* genes. As a result, the level of expression of *IRX3* and *IRX5* is increased.

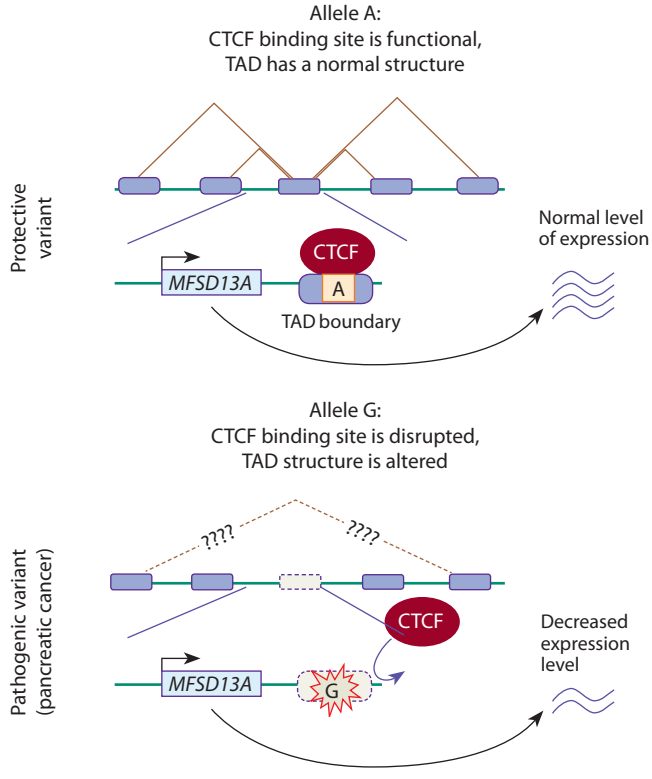


Fig. 2. Disruption of the CTCF binding site caused by the nucleotide substitution (rs2001389) eliminates one of the boundary regions that determine the TAD structure. As a result, the tumor suppressor gene *MFSD13A* expression is downregulated.

The contacts between chromatin regions within the TAD are shown with brown lines. Interrogation points in the bottom figure indicate the lack of data on the new structure of TAD.

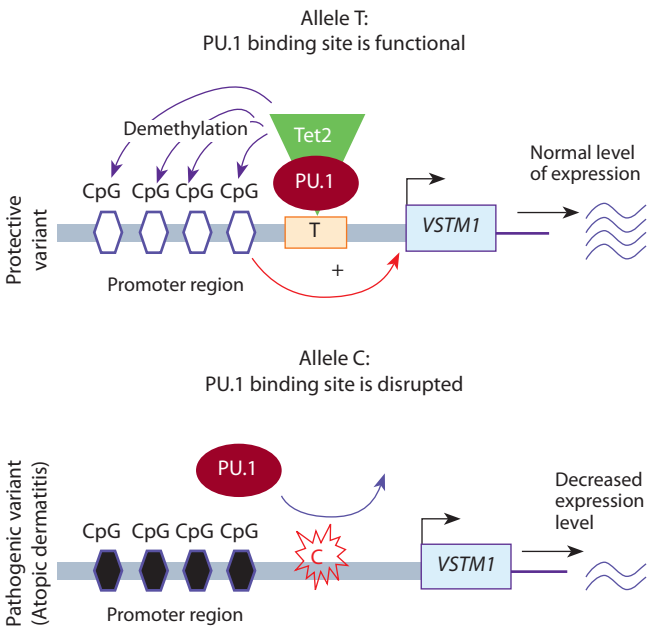


Fig. 3. Disruption of the PU.1 binding site caused by the T→C (rs612529) nucleotide substitution reduces the activity of demethylases (for example, Tet2) that maintain the *VSTM1* promoter region in an active state, and therefore *VSTM1* expression is downregulated.

et al., 2019), identified 12,761 regions containing at least two CpG dinucleotides and having an allele-specific methylation level. In most cases (98 %), regions with allele-specific methylation level are co-localized with single nucleotide variants presented in dbSNP (Benton et al., 2019).

This study (Benton et al., 2019) also analyzed the location of allele-specific methylation regions relative to the set of polymorphic loci associated with human diseases extracted from the GWAS catalog database. It turned out that polymorphic loci associated with diseases overlap with regions of allele-specific methylation twice more often than it would be expected by chance. This means that the change in methylation levels due to genetic variability is one of the factors that increase the risk of disease.

As an example, consider the rs174537 (G→T) polymorphic locus located in the enhancer of the *FADS1* and *FADS2* genes encoding fatty acid desaturases 1 and 2. The T variant of the rs174537 locus is associated with an increased risk of pathological disturbances of lipid metabolism (see Table 1). It was shown that individuals that have rs174537(T) allele had a higher methylation level of the regulatory region of the *FADS1* and *FADS2* genes in human liver (Howard et al., 2014), which led to the suppression of the transcriptional activity of *FADS1* and *FADS2*.

Occasionally, in one of the allelic variants, DNA demethylation occurs, initiated by TF binding to DNA (see Table 1). For example, such a mechanism was revealed for rs612529 T→C. This locus is located in the promoter region of the *VSTM1* (Fig. 3). The low expression of *VSTM1* in monocytes provokes the development of atopic dermatitis. In this cell type, the promoter region containing the protective variant T interacts with the transcription factor PU.1 more actively than the other one containing variant C. PU.1 initiates DNA demethylation by recruiting DNA demethylases (for example, Tet2). As a result, carriers of the T allele have completely demethylated *VSTM1* promoter, and *VSTM1* expression is activated. In carriers of pathogenic variant C, the interaction of PU.1 with DNA is disrupted, as a result, methylation level of the *VSTM1* promoter is elevated, and this is accompanied by a decrease in *VSTM1* expression (Kumar D. et al., 2017).

The effects of the genetic variability on the chromatin states and chromatin spatial organization

Pathogenic GVs may impair the chromatin state (Kilpinen et al., 2013). There are cases when the presence of a pathogenic GV was accompanied by a change in the patterns of histone modification and the appearance (or disappearance) of DNase I hypersensitive sites (McVicker et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019). In these cases, allele-specific contacts between promoters and enhancers were identified, the number of which correlated with the activity of the enhancer regions.

There are also known cases when structural variations of the genome (insertions, deletions, duplications, inversions, translocations longer than 50 nucleotides) lead to a change in the spatial organization of chromatin, thereby disrupting the expression of genes associated with pathological processes (Sun et al., 2018; Ibrahim, Mundlos, 2020). For example, the expansion of CGG trinucleotide repeats in the 5′-untranslated region (5′-UTR) of the *FMR1* gene, associated with the fragile X syndrome, disrupts the structure of TAD, that includes *FMR1* (Fig. 4, see Table 1). Normally, *FMR1* is very close to the 5′-boundary region of TAD (in Fig. 4, this is TAD1). The DNA region corresponding to this 5′-boundary is hypomethylated and is occupied by CTCF. In individuals with mutation-length CGG triplet repeats (more than 100), this boundary is ablated (this region is hypermethylated and its CTCF occupancy is lost). As a result, TAD1 dissolves and the boundary of the other TAD (in Fig. 4 it is designated as TAD2) shifts to the 3′-region of *FMR1*. Therefore, *FMR1* is within the TAD2, which normally does not contain this gene. In this case, *FMR1* promoter is hypermethylated, and *FMR1* expression is inactivated (Park et al., 2015; Sun et al., 2018).

To study molecular-genetic mechanisms of the effect of genome variability on the 3D chromatin structure, it is necessary to reconstruct the spatial genome organization. The following basic levels of the 3D genome organization have been identified: (1) regulatory DNA loops that bring together promoters and enhancers; (2) topologically associating domains (TADs), within which DNA regions have more contacts with each other than with neighboring domains; (3) A and B compartments corresponding to transcriptionally active and condensed chromatin; and finally (4) chromosome territories (Fishman et al., 2018; Hansen et al., 2018). Disruption of 3D contacts between promoters and enhancers within the TAD, caused, for example, by chromosomal rearrangements, can significantly affect the transcriptional activity of a gene, increasing risk of diseases (Lupiáñez et al., 2015).

The Institute of Cytology and Genetics SB RAS has developed an experimental computer approach for prediction

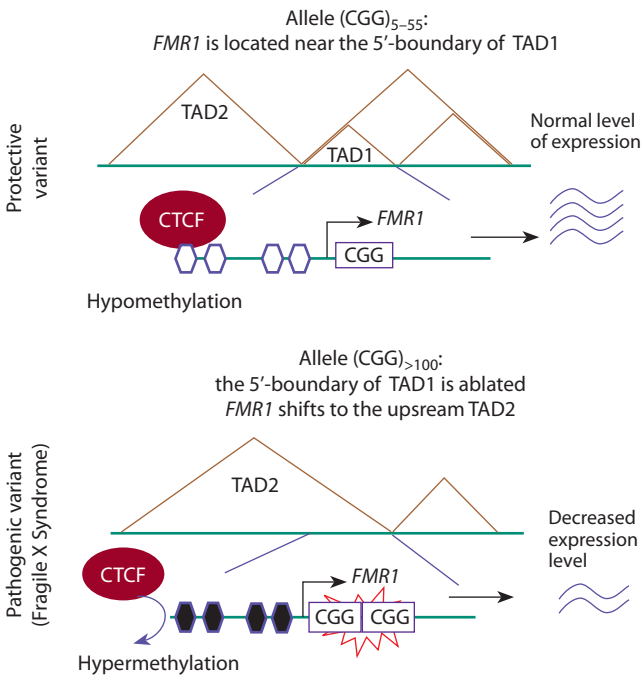


Fig. 4. With an increase in the number of CGG triplet repeats in the 5′-untranslated region of the *FMR1* gene, the DNA region corresponding to the TAD1 boundary region is hypermethylated. This leads to impaired binding of CTCF factors and disrupts a barrier function of the boundary region.

The brown lines show the contacts between chromatin loops within TADs.

physical contacts between promoters and enhancers within the 3D chromatin structure (Fishman et al., 2018; Belokopytova et al., 2020; Belokopytova, Fishman, 2021). The approach is based on the following information: (1) cell type; (2) cell-specific localization of enhancers in the linear genome (from the ENCODE database); (3) transcriptional activity of promoters (from RNA-seq experiments); (4) boundaries of chromatin loop extrusion (based on ChIP-seq mapping of CTCF occupancy in a definite cell type); (5) orientation of CTCF binding motifs (based on motif prediction pipeline); (6) A or B chromatin compartment (according to Hi-C experiments). Analysis of these data using the original 3DPredictor program (Belokopytova et al., 2020), developed on the basis of machine learning algorithms, allows to predict the frequencies of physical contacts between promoters and enhancers in the 3D genome structure with an accuracy that exceeds the accuracy of other known prediction methods.

The 3DPredictor was used to analyze the 3D genome structure in homozygous *DelB/DelB* mice that have a deletion of the 1.5 Mb genomic region containing *Epha4*. This deletion is accompanied by the appearance of additional contacts between *Pax3* gene and *Epha4* enhancer region, altering *Pax3* expression and leading to brachydactyly. Mice with the *DelB/DelB* genotype are a genetic model of human pathology accompanied by limb malformations (Lupiáñez et al., 2015). Testing 3DPredictor on this model has demonstrated the high efficiency of the program: in homozygous *DelB/DelB* mice, ectopic contacts between the *Pax3* gene and *Epha4* enhancers cluster were predicted (Belokopytova et al., 2020), and these predictions were in good agreement with the experimental data.

Genetic variability: combined analysis of heterogeneous big biological and genetic data

As noted above, many polymorphic loci associated with diseases are located at a considerable distance from the coding regions of genes (ENCODE Project Consortium, 2012; Maurano et al., 2012). Additional studies are needed to identify the molecular-genetic mechanisms of the influence of such GVs on the predisposition to diseases. The purpose of such studies is to clarify the regulatory role of GVs. A typical example is the work (Zhang et al., 2018), which made it possible to find a functionally active regulatory variant rs4321755 associated with the risk of breast cancer. The rs4321755 locus is located in a distant enhancer that regulates the expression of the *MRPS30* and *RP11-53O19.1* genes (see Table 1). It turned out that in the presence of the pathogenic variant rs4321755(T), a new GATA3 binding site is created. The transcription factor GATA3 increases the functional activity of the enhancer, this leads to the formation of more contacts between the enhancer and the divergent promoter of the *MRPS30* and *RP11-53O19.1* genes, and increased expression level of these genes. To identify this functionally significant regulatory variant, the authors developed an integrated experimental computer method based on a combined analysis of heterogeneous big biological and genetic data, including: (1) data on allele-specific expression obtained from RNA-seq in combination with data on haplotypes; (2) expression quantitative trait loci (eQTL); (3) genomic distribution of DNAse I hypersensitive sites; (4) localization of ChIP-seq peaks from ENCODE and GEO databases; (5) localization of regulatory motives predicted by computer programs. Similar scenarios for integrated experimental computer research have been implemented in the other studies (Chen C.-Y. et al., 2014; Claussnitzer et al., 2015; Zhao et al., 2019; Li et al., 2020).

This kind of research became possible due to (1) the development of modern high-throughput experimental approaches that allow producing data of different types on a genome-wide scale (parallel high-throughput sequencing, ChIP-seq, 3C, Hi-C, ChIA-PET techniques, DNAse I footprinting, bisulfite sequencing, etc.); (2) development of public information resources accumulating such experimental data. Table 2 provides a brief description of information resources containing genomic data obtained on the basis of omics technologies and used to study the mechanisms by which GVs alter the level of transcription. These resources present (1) the human genome annotation (GENCODE); (2) genome variability in human populations (HapMap, 1000 Genomes Project, IGSR, dbSNP); (3) GVs associated with diseases (GWAS central, GWAS catalog, ClinVar, HGMD, OMIM, etc.); (4) modifications of the chromatin (ENCODE, NIH Roadmap Epigenomics Mapping Consortium); (5) expression quantitative trait loci (GTEx project, eQTL databases, exSNP, etc.); (6) profiling of transcription factor binding events by ChIP-seq (Cistrome Data Browser, GTRD, ReMap); (7) allele-specific binding of TFs, identified using ChIP-seq data in combination with the data on the genotypes of the studied cells (AlleleDB,

Table 2. Information resources on genomic data obtained on the basis of the modern high-performance experimental methods

Information resource	URL	Description
	The human genome annotation	
GENCODE*	https://www.gencodegenes.org/	Reference quality human gene annotations created by merging the results of manual and computational gene annotation methods
	Genetic diversity in human populations	
HapMap (Haplotype Map)	https://www.genome.gov/10001688/international-hapmap-project ftp://ftp.ncbi.nlm.nih.gov/hapmap/	A map of haplotype blocks of the human genome and the specific SNPs that identify the haplotypes (tag SNPs)
1000 Genomes Project (1KGP)	https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/	Genetic variants (single nucleotide polymorphisms, insertions/deletions, structural variants) and genotypes identified in individuals from 26 populations
International Genome Sample Resource (IGSR)	https://www.internationalgenome.org	Combining 1000 Genomes Project data with the other large datasets generated on 1000 Genomes samples by projects such as GEUVADIS, who generated RNA-Seq data on the 1000 Genomes European samples and the YRI population, and ENCODE, who have carried out extensive assays on the NA12878 cell line
dbSNP	https://www.ncbi.nlm.nih.gov/snp/	Human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with population frequency, publication, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations. The human data in dbSNP include submissions from the SNP Consortium, variations mined from genome sequence as part of the human genome project, and individual lab contributions of variations in specific genes, mRNAs, ESTs, or genomic regions
	Disease-associated genetic variants	
GWAS central (Genome-wide association studies central)	https://www.gwascentral.org/	Allele and genotype frequency data, genetic association significance findings. GWAS central gathers datasets from public domain projects, and also encourage direct data submission from the community
GWAS catalog (Genome-wide association studies catalog)	https://www.ebi.ac.uk/gwas/home	Data on associations between polymorphic loci and phenotypic traits extracted from the published GWA studies
OMIM (Online Mendelian Inheritance in Man)	https://www.ncbi.nlm.nih.gov/omim	A compendium of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. OMIM is based on the peer-reviewed biomedical literature
ClinVar (Clinical Variations)	https://www.ncbi.nlm.nih.gov/clinvar/	A public archive of reports of the relationships among human variations and phenotypes
HGMD (The Human Gene Mutation Database)	http://www.hgmd.cf.ac.uk/ac/index.php	All published gene lesions responsible for human inherited disease
PheGenI (The Phenotype-Genotype Integrator)	https://www.ncbi.nlm.nih.gov/gap/phegeni	Phenotype-oriented resource that merges GWAS catalog data with several other databases (Gene, dbGaP, OMIM, eQTL and dbSNP)
EGA (The European Genome-phenome Archive)	https://ega-archive.org/	Data on the relationship between genotypes and phenotypes obtained by various experimental methods (GWAS, exome sequencing, whole-genome sequencing, single-cell sequencing, genotyping)
dbGaP (The database of Genotypes and Phenotypes)	https://www.ncbi.nlm.nih.gov/gap/	Data and results from studies that have investigated the interaction of genotype and phenotype in humans. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits
	Chromatin modifications and chromatin states	
ENCODE (The Encyclopedia of DNA Elements)	http://genome.ucsc.edu/ENCODE/ https://www.encodeproject.org/	Genome-wide profiles of histone modifications, genome-wide DNA methylation profiles, regions of TF binding derived from ChIP-seq experiments, interactions between genomic loci, genomic distribution of DNAse I hypersensitive sites, expression data for more than 300 cell types
NIH Roadmap Epigenomics Mapping Consortium	http://www.roadmapepigenomics.org/	Human epigenomic data (DNA methylation profiles, histone modifications, chromatin accessibility, etc.). Annotation of the human genome in accordance with the classifications of chromatin states (15, 18, 25-state models)

End of Table 2

Information resource	URL	Description
	Expression quantitative trait loci (eQTL)	
Genotype-Tissue Expression (GTEx) project	https://www.gtexportal.org/home/	Expression and eQTL data in 54 human cell types with a healthy phenotype
eQTL databases	https://www.hsph.harvard.edu/liming-liang/software/eqtl/	Expression quantitative trait loci derived from lymphoblastoid cell lines
exSNP	http://www.exsnp.org/	eQTL data from six cell types (LCLs, B cells, monocytes, brain, liver, and skin) integrated with SNPs in disease risk loci from GWA studies of seven common human diseases
eQTL Catalogue	https://www.ebi.ac.uk/eqtl/	Cis-eQTLs and splicing QTLs from all available public studies on human (including GTEx project data)
eQTL Browser	http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/	eQTLs identified in recent studies in multiple tissues
	Profiling of transcription factor binding events by ChIP-seq	
Cistrome Data Browser	http://cistrome.org/db/#/	The ChIP-seq, DNase-seq and ATAC-seq data: (1) genomic regions interacting with TFs, (2) DNase I hypersensitive sites, (3) the binding locations of modified histone proteins. The data has been assigned statuses according to six quality control criteria
Gene Transcription Regulation Database (GTRD)	https://gtrd.biouml.org/#!	A collection of ChIP-seq experiments aimed at finding TF binding sites in the human and mouse genomes
ReMap (Global map of regulatory elements)	http://remap.univ-amu.fr/	A collection of ChIP-seq, ChIP-exo, DAP-seq experiments from public resources (GEO, ENCODE, ENA). Chromatin regions in contact with TFs, transcriptional coactivators, and chromatin remodeling factors
	Allele-specific binding of TFs, identified using ChIP-seq data in combination with the data on the genotypes of the studied cells	
AlleleDB	http://alleledb.gersteinlab.org/	Genomic annotation of cis-regulatory SNVs associated with allele-specific binding and expression derived from RNA-seq and ChIP-seq data of 383 individuals from the 1000 Genomes Project
AlleleSeq	http://alleleseq.gersteinlab.org/	Allele-specific binding of six TFs (cFos, cMyc, JunD, Max, NfκB, CTCF) identified using variation data for NA12878 from the 1000 Genomes Project as well as matched, deeply sequenced RNA-Seq and ChIP-Seq data sets generated for this purpose
	The effects of genetic variants on TFBSs predicted *in silico* by computer programs	
HaploReg	https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php	Annotation of polymorphic loci within haplotype blocks that were defined using LD information from the 1000 Genomes Project. Annotation includes: (1) chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, (2) sequence conservation across mammals, (3) the effect of GVs on regulatory motifs, (4) the effect of GVs on expression from eQTL studies
SNP2TFBS	http://ccg.vital-it.ch/snp2tfbs/	Genetic variants from 1000 Genomes Project, which, according to *in silico* predictions, affect the similarity of TFBSs with weight matrices
rSNPBase	http://rsnp3.psych.ac.cn/index.do	SNP-related regulatory elements (TF binding regions, TADs, mature miRNA regions, predicted miRNA target sites, etc.), SNP-related regulatory element-target gene pairs, SNP-based regulatory networks
rVarBase	http://rv.psych.ac.cn/	Annotation of polymorphic loci (including copy number variations). Annotation includes (1) chromatin state, (2) related regulatory element (CpG islands, matched TF binding sites, miRNA target sites, etc.), (3) target genes
	Information resources integrating or accumulating diverse types of data	
UCSC Genome Browser	https://genome.ucsc.edu/	Data is integrated based on a graphical interface that allows visualizing genome sequences along with a large number of annotations and features (positions of transcripts, GC percent, chromatin states, histone marks, contacts between chromatin regions, expression, genetic variability, etc.). Data can be retrieved in text format via special Table Browser program
Ensembl Genome Browser	https://www.ensembl.org/index.html	Data is integrated based on a graphical interface that allows visualizing genome sequences along with a large number of annotations and features (positions of transcripts, GC percent, chromatin states, genetic variability, etc.). Tables of Ensembl data can be downloaded via the highly customizable BioMart data mining tool
GEO (Gene Expression Omnibus)	https://www.ncbi.nlm.nih.gov/gds	The largest public repository that archives and freely distributes comprehensive sets of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community

* GENCODE reference gene annotations for the human and mouse genomes are also available through the UCSC Genome Browser (https://genome.ucsc.edu/) and the Ensembl genome browser (https://www.ensembl.org/index.html).

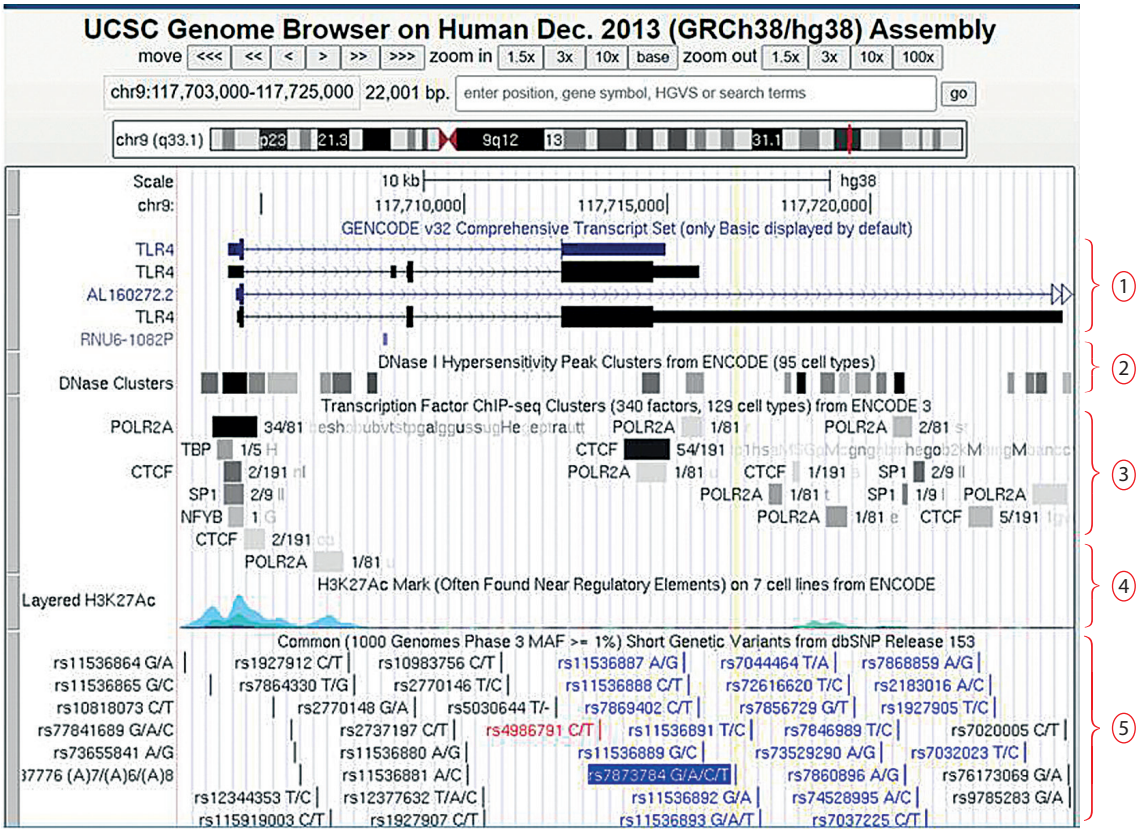


Fig. 5. The view of the human genomic region (chromosomal coordinates chr9: 117,703,000–117,725,000) displayed by the Genome Browser of the University of California, Santa Cruz, USA (UCSC Genome Browser, https://genome.ucsc.edu/).

(1) transcripts of the *TLR4* gene, displayed according to the GENCODE v32 release; (2) DNase I hypersensitivity peak clusters derived from assays in 95 cell types (as a part of the ENCODE project); (3) transcription factor binding derived from a large collection of ChIP-seq experiments performed by the ENCODE project; (4) levels of enrichment of the H3K27Ac histone mark across the genome as determined by a ChIP-seq assay on 7 cell lines from ENCODE (H3K27Ac is the acetylation of lysine 27 of the H3 histone protein, and it is often found near regulatory elements); (5) short genetic variants from dbSNP release 153. The yellow vertical line marks the position the SNP rs7873784 located in the 3′-UTR of *TLR4* gene and associated with development of rheumatoid arthritis and type 2 diabetes (see Table 1). According to (Korneev et al., 2020), the G⟶C substitution at the rs7873784 locus creates PU.1 binding site, that increases the activity of the enhancer located in the 3′-UTR of the *TLR4* gene.

AlleleSeq); (8) the effects of genetic variability on TFBSs predicted *in silico* by computer programs (HaploReg, SNP2TFBS, rSNPBase, rVarBase).

A separate category of information resources includes: (1) the genome browser of the University of California, Santa Cruz, USA (UCSC Genome Browser, https://genome.ucsc.edu/) and (2) the genome browser of the Ensembl database which is a joint research project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute (Ensembl Genome Browser, https://www.ensembl.org/index.html). These genome browsers integrate data on genome sequences and its features obtained by different research groups using a wide range of experimental methods (Lee et al., 2020; Yates et al., 2020). The websites of these browsers provide access to the primary DNA sequences and genome annotations for many organisms (including vertebrates and several other model species). Browser's graphical interfaces allow to obtain scalable maps of genomic regions and to visualize interactively a large number of annotations and features (for example, positions of transcripts, positions of GVs, chromatin regions interacting with TFs detected by ChIP-seq experiments, data on genome-wide mapping of DNase I hypersensitive sites, etc.) (Fig. 5).

The websites of the UCSC Genome Browser and Ensembl Genome Browser provide access to software tools for extraction data as text files: UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) and BioMart data mining tool (https://www.ensembl.org/info/data/biomart/index.html).

Information resources on allele-specific binding of transcription factors and on the effects of genetic variants on TFBSs predicted *in silico*

As noted above, the influence of pathogenic GVs on gene expression is often mediated through a change in the functional activity of TFBSs. In this regard, information resources that include whole genome data on allele-specific binding of TFs, identified based on the ChIP-seq method, can be extremely useful. A range of approaches have been developed to identify allele-specific binding of TFs (Rozowsky et al., 2011; Reddy et al., 2012; Waszak et al., 2014; Younesy et

al., 2014). These approaches are based on the analysis of the ChIP-seq data in combination with the sequencing data, which allow to find heterozygous loci within a single genome and to phase genotypes of the studied cells. Thus, for each type of cells examined, its own set of genomic loci interacting with a specific transcription factor in an allele-specific manner can be identified. For example, in (Cavalli et al., 2016a), the ChIP-seq data for 55 TFs in the HepG2 cells and 57 TFs in the HeLa-S3 cells were analyzed. In HepG2 cells, 3001 genomic loci with allele-specific signals were found, and 712 loci were found in HeLa-S3 cells. The authors note the pronounced tissue-specific nature of allele-specific TF binding: of the entire set of identified loci, only 34 were found in both cell lines (Cavalli et al., 2016a).

The data on allele-specific binding of TFs are collected in the following information resources: AlleleDB (http://alleledb.gersteinlab.org/) (Chen J. et al., 2016), AlleleSeq (http://alleleseq.gersteinlab.org/) (Rozowsky et al., 2011) (see Table 2), as well as in the supplemental files to publications (Cavalli et al., 2016a, b, 2019; Shi et al., 2016).

Studies aimed at identifying allele-specific TF binding made it possible to estimate the number of genetic variants that affect the binding of a particular transcription factor to DNA in a particular cell type. The average number of such events registered for a single transcription factor can range from 19 to 37 for cells with a normal karyotype (GM12878, H1-hESC) and from 12 to 55 for cancer cell lines (SK-N-SH, K562) (Cavalli et al., 2016a, b).

When generating hypotheses on the mechanisms that mediate the effect of GVs on disease risk, one can also use the data on the effects of genetic variants on the functional activity of TFBSs predicted *in silico.* Such information is accumulated in specialized databases: HaploReg (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward, Kellis, 2012), SNP2TFBS (http://ccg.vital-it.ch/snp2tfbs/) (Kumar S. et al., 2017), rSNPBase (http://rsnp3.psych.ac.cn/index.do) (Guo, Wang, 2018), rVarBase (http://rv.psych.ac.cn) (see Table 2).

Conclusion

A significant proportion of pathogenic genetic variants associated with diseases are located in non-coding regions of the human genome. Such genetic variants can with a high degree of probability disrupt functional activity of regulatory regions that control the transcriptional activity of genes. The examples of the mechanisms of influence of pathogenic genetic variants on gene expression considered in this review confirm this possibility. The studies that have made it possible to identify these mechanisms are complex and are based on the analysis of big heterogeneous genetic data. The online omics data resources provide ample opportunities for such research. Further development of experimental techniques and bioinformatics methods for analyzing the data obtained with the help of this techniques, as well as an increase in the set of investigated cell types, will significantly expand these possibilities.

References

Angeloni A., Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem*. 2019;63(6):707-715. DOI 10.1042/EBC20190030.

Beck T., Shorter T., Brookes A.J. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res*. 2020;48(D1):D933-D940. DOI 10.1093/nar/gkz895.

Belokopytova P., Fishman V. Predicting genome architecture: challenges and solutions. *Front. Genet*. 2021. DOI 10.3389/fgene.2020.617202.

Belokopytova P.S., Nuriddinov M.A., Mozheiko E.A., Fishman D., Fishman V. Quantitative prediction of enhancer-promoter interactions. *Genome Res*. 2020;30(1):72-84. DOI 10.1101/gr.249367.119.

Benton M.C., Lea R.A., Macartney-Coxson D., Sutherland H.G., White N., Kennedy D., Mengersen K., Haupt L.M., Griffiths L.R. Genome-wide allele-specific methylation is enriched at gene regulatory regions in a multi-generation pedigree from the Norfolk Island isolate. *Epigenetics Chromatin*. 2019;12(1):60. DOI 10.1186/s13072-019-0304-7.

Cavalli M., Baltzer N., Umer H.M., Grau J., Lemnian I., Pan G., Wallerman O., Spalinskas R., Sahlén P., Grosse I., Komorowski J., Wadelius C. Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci. Rep.* 2019;9(1):2695. DOI 10.1038/s41598-019-39633-0.

Cavalli M., Pan G., Nord H., Wallén Arzt E., Wallerman O., Wadelius C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics.* 2016a;107(6):248-254. DOI 10.1016/j.ygeno.2016.04.006.

Cavalli M., Pan G., Nord H., Wallerman O., Wallén Arzt E., Berggren O., Elvers I., Eloranta M.L., Rönnblom L., Lindblad Toh K., Wadelius C. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* 2016b;135(5):485-497. DOI 10.1007/s00439-016-1654-x.

Chen C.-Y., Chang I.-S., Hsiung C.A., Wasserman W.W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics*. 2014;7:34. DOI 10.1186/1755-8794-7-34.

Chen J., Rozowsky J., Galeev T.R., Harmanci A., Kitchen R., Bedford J., Abyzov A., Kong Y., Regan L., Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.* 2016;18(7):11101. DOI 10.1038/ncomms11101.

Chen L., Liang Y., Qiu J., Zhang L., Chen X., Luo X., Jiang J. Significance of rs1271572 in the estrogen receptor beta gene promoter and its correlation with breast cancer in a southwestern Chinese population. *J. Biomed. Sci*. 2013;20:32. DOI 10.1186/1423-0127-20-32.

Claussnitzer M., Dankel S.N., Kim K.-H., Quon G., Meuleman W., Haugen C., Glunk V., Sousa I.S., Beaudry J.L., Puviindran V., Abdennur N.A., Liu J., Svensson P.-A., Hsu Y.-H., Drucker D.J., Mellgren G., Hui C.-Ch., Hauner H., Kellis M. *FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med*. 2015;373:895-907. DOI 10.1056/NEJMoa1502214.

Cong Z., Li Q., Yang Y., Guo X., Cui L., You T. The SNP of rs6854845 suppresses transcription via the DNA looping structure alteration of super-enhancer in colon cells. *Biochem. Biophys. Res*. 2019;514:734-741. DOI 10.1016/j.bbrc.2019.04.190.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. DOI 10.1038/nature11247.

Farh K.K.-H., Marson A., Zhu J., Kleinewietfeld M., Housley W.J., Beik S., Shoresh N., Whitton H., Ryan R.J.H., Shishkin A.A., Hatan M., Carrasco-Alfonso M.J., Mayer D., Luckey C.J., Patsopoulos N.A., De Jager P.L., Kuchroo V.K., Epstein C.B., Daly M.J., Hafler D.A., Bernstein B.E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343. DOI 10.1038/nature13835.

Fishman V.S., Salnikov P.A., Battulin N.R. Interpreting chromosomal rearrangements in the context of 3-dimentional genome organization: a practical guide for medical genetics. *Biochemistry*. 2018;83(4):393-401. DOI 10.1134/S0006297918040107.

Gorbacheva A.M., Korneev K.V., Kuprash D.V., Mitkin N.A. The risk G allele of the single-nucleotide polymorphism rs928413 creates a

CREB1-binding site that activates *IL33* promoter in lung epithelial cells. *Int. J. Mol. Sci*. 2018;19(10):2911. DOI 10.3390/ijms1910 2911.

Guo L., Wang J. rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res*. 2018;46(D1):D1111-D1116. DOI 10.1093/nar/gkx1101.

Hansen A.S., Cattoglio C., Darzacq X., Tjian R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus.* 2018; 9(1):20-32. DOI 10.1080/19491034.2017.1389365.

Howard T.D., Mathias R.A., Seeds M.C., Herrington D.M., Hixson J.E., Shimmin L.C., Hawkins G.A., Sellers M., Ainsworth H.C., Sergeant S., Miller L.R., Chilton F.H. DNA methylation in an enhancer region of the FADS cluster is associated with FADS activity in human liver. *PLoS One*. 2014;9(5):e97510. DOI 10.1371/journal.pone.0097510.

Ibrahim D.M., Mundlos S. Three-dimensional chromatin in disease: what holds us together and what drives us apart? *Curr. Opin. Cell Biol*. 2020;64:1-9. DOI 10.1016/j.ceb.2020.01.003.

Izzi B., Pistoni M., Cludts K., Akkor P., Lambrechts D., Verfaillie C., Verhamme P., Freson K., Hoylaerts M.F. Allele-specific DNA methylation reinforces *PEAR1* enhancer activity. *Blood.* 2016;128: 1003-1012. DOI 10.1182/blood-2015-11-682153.

Jones P.L., Veenstra G.J., Wade P.A., Vermaak D., Kass S.U., Landsberger N., Strouboulis J., Wolffe A.P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet*. 1998;19:187-191. DOI 10.1038/561.

Kilpinen H., Waszak S.M., Gschwind A.R., Raghav S.K., Witwicki R.M., Orioli A., Migliavacca E., Wiederkehr M., Gutierrez-Arcelus M., Panousis N., Yurovsky A., Lappalainen T., Romano-Palumbo L., Planchon A., Bielser D., Bryois J., Padioleau I., Udin G., Thurnheer S., Hacker D., Core L.J., Lis J.T., Hernandez N., Reymond A., Deplancke B., Dermitzakis E.T. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342:744-747. DOI 10.1126/science.1242463.

Korneev K.V., Sviriaeva E.N., Mitkin N.A., Gorbacheva A.M., Uvarova A.N., Ustiugova A.S., Polanovsky O.L., Kulakovskiy I.V., Afanasyeva M.A., Schwartz A.M., Kuprash D.V. Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression. *Biochim. Biophys. Acta Mol. Basis Dis.* 2020;1866(3):165626. DOI 10.1016/j.bbadis.2019.165626.

Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.

Kumar D., Puan K.J., Andiappan A.K., Lee B., Westerlaken G.H., Haase D., Melchiotti R., Li Z., Yusof N., Lum J., Koh G., Foo S., Yeong J., Alves A.C., Pekkanen J., Sun L.D., Irwanto A., Fairfax B.P., Naranbhai V., Common J.E., Tang M., Chuang C.K., Jarvelin M.R., Knight J.C., Zhang X., Chew F.T., Prabhakar S., Jianjun L., Wang Y., Zolezzi F., Poidinger M., Lane E.B., Meyaard L., Rötzschke O. A functional SNP associated with atopic dermatitis controls cell type-specific methylation of the *VSTM1* gene locus. *Genome Med*. 2017;9(1):18. DOI 10.1186/s13073-017-0404-6.

Kumar S., Ambrosini G., Bucher P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res*. 2017;45(D1):D139-D144. DOI 10.1093/nar/gkw1064.

Lee C.M., Barber G.P., Casper J., Clawson H., Diekhans M., Gonzalez J.N., Hinrichs A.S., Lee B.T., Nassar L.R., Powell C.C., Raney B.J., Rosenbloom K.R., Schmelter D., Speir M.L., Zweig A.S., Haussler D., Haeussler M., Kuhn R.M., Kent W.J. UCSC Genome Browser enters 20th year. *Nucleic Acids Res*. 2020;48(D1):D756-D761. DOI 10.1093/nar/gkz1012.

Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.

Lewinsky R.H., Jensen T.G.K., Møller J., Stensballe A., Olsen J., Troelsen J.T. T_{-13910} DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity *in vitro*. *Hum. Mol. Genet*. 2005;14(24):3945-3953. DOI 10.1093/hmg/ddi418.

Li S., Li Y., Li X., Liu J., Huo Y., Wang J., Liu Z., Li M., Luo X.-J. Regulatory mechanisms of major depressive disorder risk variants. *Mol. Psychiatry*. 2020;25(9):1926-1945. DOI 10.1038/s41380-020-0715-7.

Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Klopocki E., Horn D., Kayserili H., Opitz J.M., Laxova R., Santos-Simarro F., Gilbert-Dussardier B., Wittler L., Borschiwer M., Haas S.A., Osterwalder M., Franke M., Timmermann B., Hecht J., Spielmann M., Visel A., Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell.* 2015;161(5):1012-1025. DOI 10.1016/j.cell.2015.04.004.

Mathelier A., Shi W., Wasserman W.W. Identification of altered cis-regulatory elements in human disease. *Trends Genet*. 2015;31(2): 67-76. DOI 10.1016/j.tig.2014.12.003.

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., Sandstrom R., Qu H., Brody J., Shafer A., Neri F., Lee K., Kutyavin T., Stehling-Sun S., Johnson A.K., Canfield T.K., Giste E., Diegel M., Bates D., Hansen R.S., Neph S., Sabo P.J., Heimfeld S., Raubitschek A., Ziegler S., Cotsapas C., Sotoodehnia N., Glass I., Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. DOI 10.1126/science.1222794.

McVicker G., van de Geijn B., Degner J.F., Cain C.E., Banovich N.E., Raj A., Lewellen N., Myrthil M., Gilad Y., Pritchard J.K. Identification of genetic variants that affect histone modifications in human cells. *Science.* 2013;342:747-749. DOI 10.1126/science.1242429.

Meddens C., van der List A.C.J., Nieuwenhuis E.E.S., Mokry M. Non-coding DNA in IBD: from sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut*. 2019;68(5):928-941. DOI 10.1136/gutjnl-2018-317516.

Mei S., Ke J., Tian J., Ying P., Yang N., Wang X., Zou D., Peng X., Yang Y., Zhu Y., Gong Y., Zhong R., Chang J., Miao X. A functional variant in the boundary of a topological association domain is associated with pancreatic cancer risk. *Mol. Carcinog*. 2019;58(10): 1855-1862. DOI 10.1002/mc.23077.

Merkulov V.M., Leberfarb E.Y., Merkulova T.I. Regulatory SNPs and their widespread effects on the transcriptome. *J. Biosci*. 2018;43(5): 1069-1075. DOI 10.1007/s12038-018-9817-7.

Nan X., Ng H.H., Johnson C.A., Laherty C.D., Turner B.M., Eisenman R.N., Bird A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*. 1998;393:386-389. DOI 10.1038/30764.

Park C.-Y., Halevy T., Lee D.R., Sung J.J., Lee J.S., Yanuka O., Benvenisty N., Kim D.-W. Reversion of *FMR1* methylation and silencing by editing the triplet repeats in fragile X iPSC-derived neurons. *Cell. Rep*. 2015;13(2):234-241. DOI 10.1016/j.celrep.2015.08.084.

Quenneville S., Verde G., Corsinotti A., Kapopoulou A., Jakobsson J., Offner S., Baglivo I., Pedone P.V., Grimaldi G., Riccio A., Trono D. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell*. 2011;44(3):361-372. DOI 10.1016/j.molcel.2011.08.032.

Rahbar E., Waits C.M.K., Kirby E.H., Jr., Miller L.R., Ainsworth H.C., Cui T., Sergeant S., Howard T.D., Langefeld C.D., Chilton F.H. Allele-specific methylation in the *FADS* genomic region in DNA from human saliva, CD4+ cells, and total leukocytes. *Clin. Epigenetics*. 2018;10:46. DOI 10.1186/s13148-018-0480-5.

Reddy T.E., Gertz J., Pauli F., Kucera K.S., Varley K.E., Newberry K.M., Marinov G.K., Mortazavi A., Williams B.A., Song L., Crawford G.E., Wold B., Willard H.F., Myers R.M. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res*. 2012;22(5):860-869. DOI 10.1101/gr.131201.111.

Roadmap Epigenomics Consortium, Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M.J., … Hirst M., Meissner A., Milosavljevic A., Ren B., Stamatoyannopoulos J.A., Wang T., Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015; 518(7539):317-330. DOI 10.1038/nature14248.

Rozowsky J., Abyzov A., Wang J., Alves P., Raha D., Harmanci A., Leng J., Bjornson R., Kong Y., Kitabayashi N., Bhardwaj N., Rubin M., Snyder M., Gerstein M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol*. 2011;7:522. DOI 10.1038/msb.2011.54.

Schmitz R.J., Lewis Z.A., Goll M.G. DNA methylation: shared and divergent features across eukaryotes. *Trends Genet*. 2019;35(11): 818-827. DOI 10.1016/j.tig.2019.07.007.

Shi W., Fornes O., Mathelier A., Wasserman W.W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* 2016;44(21):10106-10116. DOI 10.1093/nar/gkw691.

Smith A.J.P., Deloukas P., Munroe P.B. Emerging applications of genome-editing technology to examine functionality of GWAS-associated variants for complex traits. *Physiol. Genomics*. 2018;50(7): 510-522. DOI 10.1152/physiolgenomics.00028.2018.

Sun J.H., Zhou L., Emerson D.J., Phyo S.A., Titus K.R., Gong W., Gilgenast T.G., Beagan J.A., Davidson B.L., Tassone F., Phillips-Cremins J.E. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell*. 2018;175(1):224-238. DOI 10.1016/j.cell.2018.08.005.

Visser M., Palstra R.J., Kayser M. Allele-specific transcriptional regulation of *IRF4* in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the *IRF4* promoter. *Hum. Mol. Genet*. 2015;24(9):2649-2661. DOI 10.1093/hmg/ddv029.

Vohra M., Sharma A.R., Prabhu B.N., Rai P.S. SNPs in sites for DNA methylation, transcription factor binding, and miRNA targets leading to allele-specific gene expression and contributing to complex disease risk: a systematic review. *Public Health Genomics*. 2020;23: 1-16. DOI 10.1159/000510253.

Wang H., Lou D., Wang Z. Crosstalk of genetic variants, allele-specific DNA methylation, and environmental factors for complex disease risk. *Front. Genet*. 2019;9:695. DOI 10.3389/fgene.2018.00695.

Ward L.D., Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930-D934. DOI 10.1093/nar/gkr917.

Waszak S.M., Kilpinen H., Gschwind A.R., Orioli A., Raghav S.K., Witwicki R.M., Migliavacca E., Yurovsky A., Lappalainen T., Hernandez N., Reymond A., Dermitzakis E.T., Deplancke B. Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics*. 2014;30(2):165-171. DOI 10.1093/bioinformatics/btt667.

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.

Yates A.D., Achuthan P., Akanni W., Allen J., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Azov A.G., Bennett R., Bhai J., … Perry E., Ruffier M., Trevanion S.J., Cunningham F., Howe K.L., Zerbino D.R., Flicek P. Ensembl 2020. *Nucleic Acids Res*. 2020; 48(D1):D682-D688. DOI 10.1093/nar/gkz966.

Younesy H., Möller T., Heravi-Moussavi A., Cheng J.B., Costello J.F., Lorincz M.C., Karimi M.M., Jones S.J.M. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics*. 2014;30(8): 1172-1174. DOI 10.1093/bioinformatics/btt744.

Zhang Y., Manjunath M., Zhang S., Chasman D., Roy S., Song J.S. Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. *Cancer Res*. 2018;78(7):1579-1591. DOI 10.1158/0008-5472.CAN-17-3486.

Zhao T., Hu Y., Zang T., Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet*. 2019;10:1021. DOI 10.3389/fgene.2019.01021.

ORCID ID
E.V. Ignatieva orcid.org/0000-0002-8588-6511

Acknowledgements. The study was supported from the funds of the budget project No. 0259-2021-0009.
Conflict of interest. The authors declare no conflict of interest.
Received December 28, 2020. Revised January 18, 2021. Accepted January 18, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):30-38
DOI 10.18699/VJ21.004

Original Russian text www.bionet.nsc.ru/vogis/

Improving the quality of barley transcriptome *de novo* assembling by using a hybrid approach for lines with varying spike and stem coloration

N.A. Shmakov[1, 2]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomics Center, Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
shmakov@bionet.nsc.ru

Abstract. *De novo* transcriptome assembly is an important stage of RNA-seq data computational analysis. It allows the researchers to obtain the sequences of transcripts presented in the biological sample of interest. The availability of accurate and complete transcriptome sequence of the organism of interest is, in turn, an indispensable condition for further analysis of RNA-seq data. Through years of transcriptomic research, the bioinformatics community has developed a number of assembler programs for transcriptome reconstruction from short reads of RNA-seq libraries. Different assemblers makes it possible to conduct a *de novo* transcriptome reconstruction and a genome-guided reconstruction. The majority of the assemblers working with RNA-seq data are based on the De Bruijn graph method of sequence reconstruction. However, specifics of their procedures can vary drastically, as do their results. A number of authors recommend a hybrid approach to transcriptome reconstruction based on combining the results of several assemblers in order to achieve a better transcriptome assembly. The advantage of this approach has been demonstrated in a number of studies, with RNA-seq experiments conducted on the Illumina platform. In this paper, we propose a hybrid approach for creating a transcriptome assembly of the barley *Hordeum vulgare* isogenic line Bowman and two nearly isogenic lines contrasting in spike pigmentation, based on the results of sequencing on the IonTorrent platform. This approach implements several *de novo* assemblers: Trinity, Trans-ABySS and rnaSPAdes. Several assembly metrics were examined: the percentage of reference transcripts observed in the assemblies, the percentage of RNA-seq reads involved, and BUSCO scores. It was shown that, based on the summation of these metrics, transcriptome meta-assembly surpasses individual transcriptome assemblies it consists of.
Key words: RNA-seq; transcriptomics; *de novo* transcriptome reconstruction; IonTorrent.

For citation: Shmakov N.A. Improving the quality of barley transcriptome *de novo* assembling by using a hybrid approach for lines with varying spike and stem coloration. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):30-38. DOI 10.18699/VJ21.004

Улучшение качества сборки *de novo* транскриптомов ячменя на основе гибридного подхода для линий с изменениями окраски колоса и стебля

Н.А. Шмаков[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр, Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
shmakov@bionet.nsc.ru

Аннотация. Реконструкция транскриптома *de novo* – важная стадия биоинформатического анализа данных RNA-seq, которая позволяет получить последовательности транскриптов, присутствующих в изучаемом биологическом образце. Наличие точной и полной последовательности транскриптома организма, в свою очередь, является необходимым условием для дальнейшей работы с данными RNA-seq. Биоинформатическим сообществом было создано множество программ-сборщиков для реконструкции транскриптома из коротких прочтений RNA-seq. Сборщики позволяют проводить как *de novo* реконструкцию транскриптома, так и реконструкцию, основанную на картировании коротких прочтений RNA-seq на последовательность референсного генома организма. Большинство *de novo* сборщиков, работающих с данными RNA-seq, применяют технологию реконструкции последовательностей методом графов де Брёйна. Однако детали их работы могут существенно различаться, поэтому различия могут встречаться и в результатах. Некоторые авторы рекомендуют для получения более полной и качественной сборки использовать гибридную сборку транскриптома – подход, основанный на комбинации результатов работы нескольких сборщиков. Преимущество такого подхода было продемонстрировано

в ряде исследований по анализу транскриптомов на платформе Illumina. Нами предложен гибридный подход по созданию сборок транскриптома ячменя *Hordeum vulgare* изогенной линии Bowman и двух почти изогенных линий, полученных на основе Bowman и контрастных по окраске колоса, используя данные, полученные при секвенировании матричной РНК на платформе IonTorrent. В данном подходе применяются несколько индивидуальных сборщиков: Trans-ABySS, rnaSPAdes и Trinity. Были оценены некоторые показатели, характеризующие полноту и точность сборки: доля обнаруженных в сборке известных транскриптов ячменя, доля задействованных в сборке прочтений из библиотек RNA-seq, значение критерия BUSCO. По совокупности этих показателей метасборки демонстрируют более высокое качество полученного транскриптома по сравнению с индивидуальными сборщиками.
Ключевые слова: RNA-seq; транскриптомика; *de novo* реконструкция транскриптома; IonTorrent.

Introduction

Next generation massively parallel sequencing technology applied to RNA (RNA-seq) is a method of choice in modern transcriptomics researches. It involves several steps: extraction of total mRNA of a biological sample, fragmentation of mRNA and simultaneous sequencing a large number of obtained short fragments (Engström et al., 2013; Hrdlickova et al., 2017).

De novo transcriptome assembly from sequenced fragments is one of the most important stages of transcriptome profiling experiment (Chang et al., 2014). It allows researcher to obtain sequences of mRNA molecules from the studied sample. Presently there are two main approaches to transcriptome sequences reconstruction from short read libraries – so-called OLC method (Overlap–Layout–Consensus) and de Bruijn graph method (Li et al., 2012; Schliesky et al., 2012). OLC method is based on pairwise alignment of reads and construction of oriented graphs where each node is one read. Overlaps between reads represent edges of the graph. This method is more suitable for contig assembly from a relatively smaller number of long reads with large overlapping regions, and thus is more frequently used to assemble sequences obtained with Saenger sequencing method or third generation sequencing methods (Cui et al., 2020).

The other method is based on construction of de Bruijn graph in which nodes are represented by *k*-mers – nucleotide sequences of given length *k*. Next, all paths on the graph that comprise sequences of short reads in RNA-seq libraries are marked. Then, all paths on the graph that contain continuous sequences of overlapping reads are marked. Thus, sequences of contigs consisting of short reads from the libraries are obtained. This method is implemented in several assemblers, namely Trinity (Grabherr et al., 2013), Trans-ABySS (Robertson et al., 2010), SOAPdenovo-Trans (Xie et al., 2014), Oases (Schulz et al., 2012).

An important parameter for de Bruijn graphs-based assemblers is *k* – length of *k*-mers used in de Bruijn graph construction. *k*-mers are words located in the nodes of de Bruijn graph. This parameter can be set by user prior to starting assembler program. Increasing *k* results in higher precision of assembly, but at the same time it makes it more computationally difficult (Fu et al., 2018). At larger *k*, the assembler might fail to detect a limited intersection between the reads, if its size is smaller than *k*. Often the following strategy is used: several preliminary assemblies are conducted at different values of *k*, then assemblies are combined, and redundancy reduction is performed, which results in a final *de novo* transcriptome assembly (Wang, Gribskov, 2017).

Since a large number of transcriptome *de novo* assemblers have been developed to date, researches were dedicated to answering the question of performance and precision of these programs. Reviews comparing different transcriptome assemblers usually mention Trinity, SOAPdenovo-Trans, Velvet-Oases among the best and most popular tools (Jain et al., 2013; Honaas et al., 2016; Wang, Gribskov, 2017). Trinity distributive, aside from the assembler itself, includes a large number of utilities for assembly quality assessment, removal of poorly represented contigs and other manipulations with the *de novo* assembly. SOAPdenovo-Trans is mentioned as the program fitting for large plant transcriptomes *de novo* assembly (Payá-Milans et al., 2018).

Given the diversity of modern assemblers, none of them are perfect and capable to satisfy all the requirements for completeness and quality of the assembly. Thus, it was proposed that implementing several *de novo* assemblers followed by creating of a single ‘meta-assembly’ may further increase sensitivity and precision of transcriptome sequences reconstruction (Cerveau, Jackson, 2016). Meta-assembly is then defined as a junction of all the *de novo* assemblies obtained with different tools after redundancy reduction. Redundancy reduction is a procedure of removal every contig that is a substring of at least one other contig in a given set. This approach was earlier tested for transcriptome assembly of non-model species using three assemblers – Trinity, Trans-ABySS and rnaSPAdes (Evangelistella et al., 2017). Furthermore, attempts were undertaken to obtain meta-assembly of transcriptome based on genome-guided assemblies (Venturini et al., 2018).

However, to the best of our knowledge, no attempts were made to evaluate performance of this approach on data obtained with IonTorrent sequencing platform. Meanwhile, IonTorrent platform, although being less popular than Illumina, is still in demand in biological researches, including studies of microbial metagenomes (Lee et al., 2019), interspecific diversity of earthworms (Shekhovtsov et al., 2019), transgenic rat lines (Bürckert et al., 2017), sequencing plant genomes (Salina et al., 2018). Furthermore, studies on comparison of Illumina and IonTorrent platforms have been performed that show IonTorrent reads having somewhat lower quality and precision that Illumina reads, and have greater discrepancy of read lengths (Lahens et al., 2017).

This research aims to create a computational pipeline based on transcriptome meta-assembly creation using *de novo* assemblers Trinity, Trans-ABySS and rnaSPAdes, as well as genome-guided version of Trinity based on reference genome. Computational pipeline was tested on transcriptome assembly

of *Hordeum vulgare* L. barley isogenic line Bowman and nearly-isogenic lines iBw*Alm* with partial albinism of the spike and BLP with partial melanism of the spike. It was observed that quality of the transcriptome assemblies performed with different tools vary; however, in general they complement each other. Highest quality is observed for the transcriptome meta-assembly, which outstrips individual assemblies based on a number of metrics that characterize overall assembly quality.

Materials and methods

Short read libraries. Libraries of *H. vulgare* isogenic line Bowman and two nearly isogenic lines: iBw*Alm* (characterized by spike partial albinism) and BLP (characterized by partial melanism of the spike) transcriptome were used. The data was obtained from NCBI SRA database BioProjects PRJNA342150 (libraries of NIL i:Bw*Alm* and isogenic line Bowman) and PRJNA399215 (libraries of NIL BLP and isogenic line Bowman).

In PRJNA342150 experiment, transcriptomes of NIL i:Bw*Alm*, based on isogenic line Bowman, plants lemma and line Bowman, taken as a control, plants lemma were compared (Shmakov et al., 2016). For each of the lines three biological replicates were taken. Thus, in this experiment six short read libraries were sequenced. We will refer to this experiment as 'alm experiment' in further text.

In PRJNA399215, transcriptomes of NIL BLP, based on Bowman isogenic line, plants lemma and isogenic line Bowman, taken as a control group, plants lemma were compared (Glagoleva et al., 2017). We will refer to this experiment as 'blp experiment' in further text.

All libraries were obtained by sequencing using IonTorrent platform. The libraries then underwent filtration procedure, during which adapter sequences were removed using CutAdapt software version 1.9.1 (Martin, 2011), reads with mean quality score below 20 and lengths below 50 or above 270 were removed using Prinseq-lite software version 0.20.4 (Schmieder, Edwards, 2011). Table 1 lists metrics of the libraries used in this research.

Transcriptome reconstruction. In this work, three transcriptome assemblers were used: Trinity (Grabherr et al., 2013) version 2.2.0, Trans-ABySS (Robertson et al., 2010) version 2.0.1 and rnaSPAdes (Bushmanova et al., 2018) version 3.12.0. All three tools were listed among the best in performance and quality in a number of researches dedicated to comparison of transcriptome *de novo* assemblers (Honaas et al., 2016; Lafond-Lapalme et al., 2017; Fu et al., 2018; Hölzer, Marz, 2019).

Libraries from the two experiments were processed independently. Individual transcriptome assemblies obtained with each of the software tools were reconstructed as follows.

Trinity assembler was run with default parameters; all six libraries belonging to the respective experiment were given as input files. While running SPAdes assembler, likewise, all six libraries belonging to the respective experiment were given as input files. When launching SPAdes assembler, options '–iontorrent' and '–only-assembler' were specified.

Trans-ABySS assembly was conducted for each of the libraries separately, with resulting assemblies combined using transabyss-merge tool from Trans-ABySS software package. This assembly was performed with default parameters, with *k*-mer size equal to 32. In the same way, assemblies were conducted with *k*-mer sizes of 48 and 64. Thus, three *de novo* assemblies were obtained with Trans-ABySS, differing by *k*-mer lengths. Next, the three assemblies were combined with transabyss-merge. Resulting assembly was further used as an individual *de novo* transcriptome assembly obtained with Trans-ABySS software.

Additionally, genome-guided transcriptome assembly was performed using Trinity software. First, short read libraries were mapped to barley genome. Mapping files in the SAM (sequence alignment/mapping) format were then concatenated using merge tool from samtools software package version 1.6 into a single alignment file combining mapping of all six libraries belonging to respective experiment. This file, together with the six libraries from the respective experiment, were processed with Trinity tool in genome-guided transcriptome

Table 1. Metrics of the libraries implemented in the work

Experiment	Line	Library	Raw size	Clean size	Read mean length
PRJNA342150	i:Bw*Alm*	Alm_1	4596395	3874912	166.94
		Alm_2	3056413	2372255	199.52
		Alm_3	5794644	5332600	181.47
	Bowman	A_bow_1	4122599	2450068	175.49
		A_bow_2	4023501	2356572	126.56
		A_bow_3	6887599	6523266	201.68
PRJNA399215	BLP	Blp_1	3583148	1311442	185.39
		Blp_2	4710862	1687289	156.96
		Blp_3	4070591	1864073	146.02
	Bowman	B_bow_1	1769261	438702	164.66
		B_bow_2	3740926	1092191	199.48
		B_bow_3	5253524	2364034	209.00

assembly mode, with a specified parameter of maximal intron length of 500 000 nucleotides.

In order to remove redundancy of assemblies, tr2aacds.pl tool from software package Evidential Gene (Gilbert, 2019) version 20.05.2020 was implemented. Each of the individual assemblies was processed with this software. Thus, three non-redundant transcriptome *de novo* assemblies and one non-redundant genome-guided transcriptome assembly were obtained. We will further refer to the *de novo* assemblies as short versions of respective software names: abyss, spades and trinity assemblies constructed using Trans-ABySS, rnaSPAdes and Trinity, respectively. We will further refer to genome-guided transcriptome assembly as GG (short of genome-guided).

In order to create an optimal meta-assembly of the transcriptome, individual assemblies were concatenated into one file, which was then processed with tr2aacds.pl tool for redundancy removal. It should be noted that only contigs containing open reading frames are considered, as tr2aacds.pl only uses contigs with predicted open reading frames with length above threshold value for further analysis. Figure 1 illustrates main stages of non-redundant meta-assembly construction.

Thus, for each of the two experiments, four individual assemblies were created: spades and trinity assemblies, consisting of six short libraries belonging to the respective experiment; abyss assembly performed for each of the libraries separately with three different *k*-mer length values, which were later combined into a single abyss transcriptome assembly using transabyss-merge script; genome-guided GG transcriptome assembly performed on six libraries belonging to the respective experiment and alignment file combined from six libraries alignments to the barley genome. Finally, four individual assemblies for each of the experiments were combined into the barley transcriptome meta-assembly.

Transcriptome assemblies quality assessment. In order to analyze qualities of assemblies, each one was processed with the following tools: BUSCO (Simão et al., 2015) version 3.0.2 for completeness assessment based on presence of characteristic sequences for plants; TransRate (Smith-Unna et al., 2016) version 1.0.3 for contigs annotation and completeness of known barley genes presence in the assembly. Then, comparison of CDS lists detected by TransRate in each individual assembly was performed. Based on overlapping of the lists of CDS detected in each assembly, Venn diagrams illustrating the part of each individual assembly in the structure of meta-assembly were drawn.

Next, contigs of two meta-assemblies of barley transcriptome belonging to two experiments were aligned to the *H. vulgare* genome using rnaQUAST software (Bushmanova et al., 2016). rnaQUAST counts several characteristics of assembly mapping to genome, and allows the user to evaluate the assembly's quality based on these characteristics. Specifically, this tool divides the contigs into three groups: contigs mapped to the reference and interlocking with known annotated genes; contigs mapped to the genome but lacking significant overlaps with the known annotated genes; and contigs with no homology to the known genome. We will further refer to this last group of contigs as 'new contigs'.

Transcriptome assemblies' quality comparison. In order to compare the assemblies' quality numerically, an approach

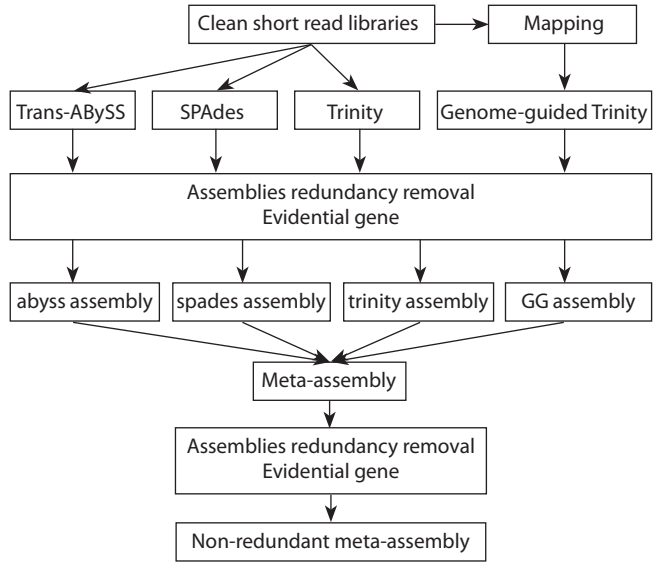


Fig. 1. Pipeline of individual *de novo* barley transcriptome assemblies and barley transcriptome meta-assembly acquisition.

suggested in the Hölzner and Marz publication (Hölzner, Marz, 2019) was implemented. This method is to normalize a selected number of parameters that reflect *de novo* transcriptome assembly quality using the following formula:

$$N_j^i = \frac{R_j^i - \min(V^i)}{\max(V^i) - \min(V^i)},$$

where R_j^i is a value of parameter i for the transcriptome assembly j before normalization, N_j^i is this parameter's value after normalization, V^i is a vector of all values of the parameter i for all k *de novo* transcriptome assemblies before normalization: $V^i = (V_1^i, \ldots, V_k^i)$. Thus, after normalization each of the parameters takes a value from 0 to 1 for each of the *de novo* assemblies. Next, for each of the assemblies all the normalized parameters are summed, and assemblies are sorted based on the summed values of normalized parameters. The assembly with the highest value of summed normalized parameters is considered to have the highest quality.

To compare individual assemblies and meta-assemblies of barley transcriptome obtained while working with the short read libraries belonging to two experiments, seven parameters characterizing different aspects of transcriptome assemblies were used: (1) N50; (2) median of contig lengths distribution; (3) number of BUSCO genes detected in the assembly (both complete and fragmentary genes); (4) percentage of contigs with homology to known barley CDS detected using TransRate; (5) number of barley CDS that contigs from *de novo* assembly are homologous with; (6) amount of barley CDS with at least 95 % of the lengths covered with aligned contigs; (7) percentage of short reads from the library that was used in construction of the *de novo* assembly that were mapped back to the assembly using kallisto software.

Parameters 1 and 2 reflect distribution of contig lengths. Parameters 3, 4, 5 and 6 show completeness of the transcriptome assembly. Parameter 7 shows completeness of the transcriptome assembly and how fully were the libraries used in the process of assembly construction.

Results

alm experiment

For barley line i:Bw*Alm* and control isogenic line Bowman four *de novo* assemblies of lemma transcriptome, and one meta-assembly consisting of the four individual assemblies were obtained. Table 2 lists results of *de novo* transcriptome assembly of barley lines i:Bw*Alm* and Bowman, including common for the two lines meta-assembly.

Transcriptome meta-assembly of lines i:Bw*Alm* and Bowman obtained from *de novo* assemblies created with rnaSPAdes, Trans-ABySS and Trinity and genome-guided Trinity assemblies, consists of 169232 contigs before redundancy removal. Non-redundant meta-assembly consists of 68414 contigs with total length of 46440750 bases. Longest contig consists of 9920 nucleotides, mean contig length is 678.8 nucleotides, N50 is 936 nucleotides. Redundancy removal reduced meta-assembly size to 40.4 % of initial.

Coverage of contigs with short reads from the libraries was estimated for individual assemblies and meta-assembly of transcriptome using pseudo-alignment technique. It was observed that the highest percentage of reads was mapped to the abyss transcriptome assembly, while the lowest – to the spades assembly. 61.47 % of all the short reads were mapped to the meta-assembly of the transcriptome (see Table 2).

Search of known annotated barley CDS in transcriptome assemblies was carried out using TransRate software tool. Results of CDS identification for the assemblies are listed in the Table 3.

The highest amount of known CDS (29790) was detected in meta-assembly of transcriptome. Moreover, the highest amount of CDS with coverage no less than 95 % was detected in meta-assembly. However, the highest percentage of contigs that show homology to known barley CDS was detected for the spades assembly – 90.3 %. In meta-assembly this metric is only 62.7 %, which is lower than in any of individual assemblies.

Furthermore, in order to estimate contribution of each of the assemblers into the transcriptome meta-assembly structure, overlapping of CDS lists detected in individual assemblies was counted. Resulting overlaps are illustrated in Figure 2. As seen from Figure 2, 7191 barley CDS were detected in all four individual assemblies; 9305 CDS were detected in three out of four assemblies. 14615 CDS were detected in only single individual assembly, out of which the largest amount (5173) were detected only in trinity assembly, the lowest amount (2086) – only in spades assembly. The biggest intersection of CDS lists were observed between trinity assembly and GG assembly – 18258 CDS.

In contigs of each of the assemblies open reading frames (ORF) were predicted. ORF detected in the contigs of meta-assembly encode 58636 protein products with lengths equal to or greater than 30 amino acid residues. These protein products were used then to evaluate integrity of the assemblies using BUSCO software, which is shown in Figure 3. Transcriptome meta-assembly contains more complete BUSCO sequences than any individual transcriptome assembly, and less fragmented and absent BUSCO sequences. This suggests that meta-assembly has higher quality and integrity.

blp experiment

For RNA-seq libraries from blp experiment, individual transcriptome assemblies and transcriptome meta-assembly were obtained, and quality comparison of the assemblies was performed. Table 4 lists main parameters of the assemblies.

Resulting transcriptome meta-assembly of barley lines Bowman and BLP consists of 133070 contigs. After redundancy removal meta-assembly contains 32466 contigs with total length of 25184753 nucleotides. Thus, redundancy re-

Table 2. Characteristics of barley *de novo* transcriptome assemblies in alm experiment

Assembly	Assembly size, contigs		N50	Mean length	Reads mapped, %
	Redundant	Non-redundant			
abyss	705015	40806	1076	723.6	67.08
spades	22649	19181	1130	1072.65	39.13
trinity	267201	52005	976	741.19	64.97
GG	451309	57240	766	594.82	61.37
Meta-assembly	169232	68414	936	678.82	61.47

Table 3. Numbers of barley CDS detected in *de novo* transcriptome assemblies in alm experiment

Assembly	Contigs		CDS detected	p_95
	detected	%		
abyss	30530	0.748	22420	2542
spades	17323	0.903	14989	644
trinity	35547	0.684	27173	1779
GG	38686	0.676	26978	2240
Meta-assembly	42887	0.627	29790	3073

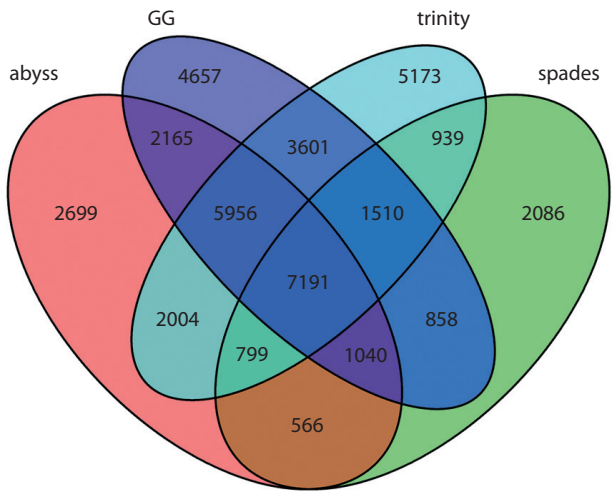


Fig. 2. Venn diagram illustrating overlaps of CDS lists detected in individual transcriptome assemblies in alm experiment.

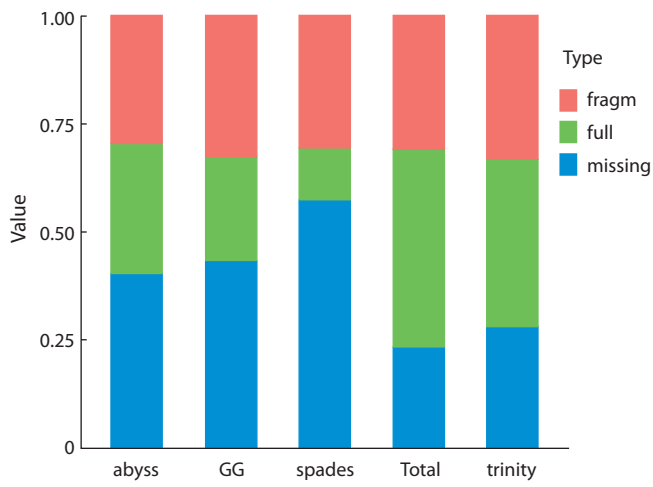


Fig. 3. BUSCO criterion of completeness of transcriptome assembly in alm experiment.

moval reduced assembly size to 24.4 % of initial size. Also, it is worth noting that meta-assembly in blp experiment has a higher N50 value than any of the individual assemblies it consists of. 72.1 % of short reads from blp experiment libraries were mapped back to the transcriptome meta-assembly. For this indicator, meta-assembly is behind GG assembly (77.6 %), but ahead of three other individual assemblies.

Search of known barley CDS was carried out in transcriptome *de novo* assemblies of barley lines under investigation using TransRate software. Results of the search are shown in Table 5. As can be seen from Table 5, from as low as 19848 contigs in spades assembly to as much as 29412 contigs in GG assembly show homology to known barley CDS. Meanwhile, the highest amount of barley CDS were detected in trinity assembly, however, the highest amount of barley CDS with no less than 95 % length covered with contigs is detected in transcriptome meta-assembly – 1825 CDS. Percentage of contigs from the assembly for which homology to known CDS was detected is 74.5 % in meta-assembly which is lower that in any individual assembly except for trinity assembly.

Search of overlaps between lists of CDS detected in individual assemblies was performed, and contribution of individual assemblies into meta-assembly structure was evaluated (Fig. 4). 9742 CDS were detected in all four individual transcriptome *de novo* assemblies. 8656 CDS were detected in only one of individual assemblies, of which the largest amount – 3554 were unique for abyss assembly, lowest

Table 4. Characteristics of barley *de novo* transcriptome assemblies in blp experiment

Assembly	Assembly size, contigs		N50	Mean length	Reads mapped, %
	Redundant	Non-redundant			
abyss	214465	34987	606	490.32	68.75
spades	31453	24401	1046	824.6	58.25
trinity	116897	34363	891	661.59	66.55
GG	122304	39319	976	707.83	77.55
Meta-assembly	133070	32466	1056	775.73	72.07

Table 5. Numbers of barley CDS detected in *de novo* transcriptome assemblies in blp experiment

Assembly	Contigs		CDS detected	p_95
	detected	%		
abyss	25804	0.738	18981	1224
spades	19848	0.813	16818	1017
trinity	22793	0.663	21885	1478
GG	29412	0.748	19947	1597
Meta-assembly	24194	0.745	19665	1825

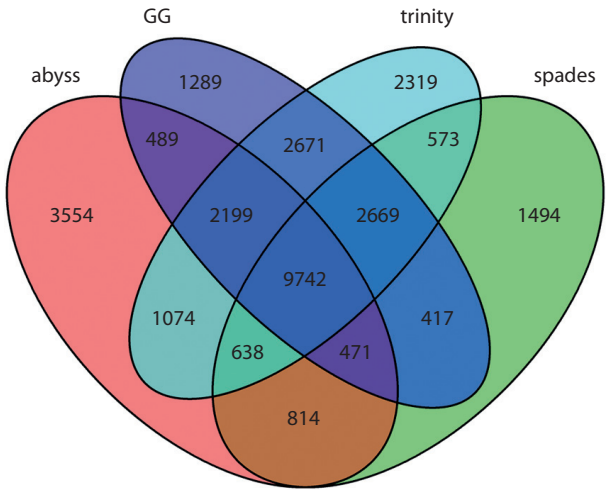


Fig. 4. Overlaps between CDS lists detected in individual transcriptome assemblies in blp experiment.

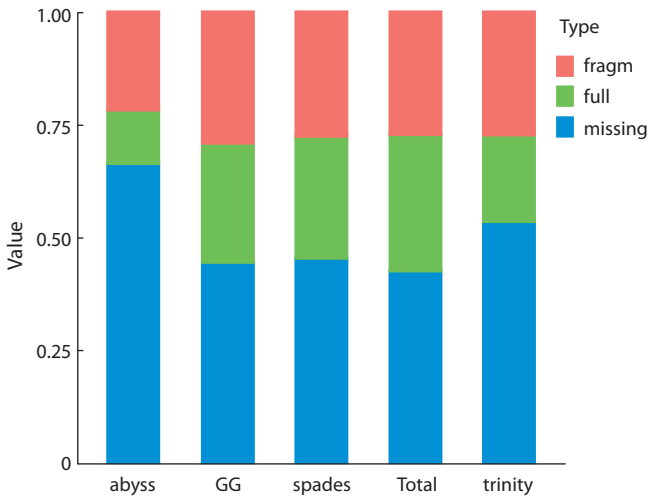


Fig. 5. BUSCO criterion of transcriptome assembly completeness in blp experiment.

amount – 1289 were unique for GG assembly. The highest amount of common CDS is between GG and trinity assemblies – 17281 CDS were detected in both of these assemblies.

Transcriptome assemblies' integrity estimation were carried out using BUSCO tool (Fig. 5). Meta-assembly was shown to have higher completeness than any of the individual assemblies, as it has the highest amount of complete BUSCO sequences detected and lowest amount of BUSCO sequences non-detected. In total 57.6 % of all BUSCO sequences from embryophyte set were detected in non-redundant meta-assembly as completely or partially.

Comparison of *de novo* assemblies' quality

Seven metrics of individual *de novo* assemblies and meta-assembly were evaluated in order to assess quality of the assemblies. These metrics indicate lengths of contigs in *de novo* assemblies (N50 and median of lengths distribution), presence of known barley CDS in the *de novo* assembly (percentage of contigs with homology to known barley CDS, amount of detected CDS and amount of CDS with at least 95 % of length covered) and genes characteristic to vascular plants (BUSCO completeness criterion), and fullness of libraries short reads implementation in the assembly creation (percentage of pseudo-aligned reads). Values of these metrics were normalized and brought into the range of values from 0 to 1 (Hölzner, Marz, 2019), then sums of normalized metrics were taken for each of the individual assemblies and for the meta-assembly. The largest values of the sums show the most optimal transcriptome assembly (Table 6).

As can be seen from the Table 6, highest values of normalized metrics are attributed to the transcriptome meta-assemblies in both experiments. This, together with highest amount of detected genes characteristic to vascular plants detected with BUSCO software, and highest amounts of fully reconstructed barley CDS indicates that meta-assemblies created by combining of individual *de novo* transcriptome assemblies and redundancy removal outstrip individual assemblies by quality.

Table 6. Summarized values of normalized quality metrics for *de novo* transcriptome assemblies in experiments alm and blp

Assembly	alm experiment (lines i:Bw*Alm* and Bowman)	blp experiment (lines BLP and Bowman)
abyss	4.16	1.72
spades	3.00	3.86
trinity	4.07	3.61
GG	2.85	5.22
Meta-assembly	4.32	5.56

Discussion

In this work, an approach to *de novo* transcriptome reconstruction based on creation of meta-assembly from several individual assemblies was tested. It was observed that transcriptome meta-assemblies have higher integrity judging by a number of criteria such as amount of detected BUSCO fragments, amount of barley CDS to which contigs in transcriptome assembly show homology, and percentage of pseudo-aligned to the assembly reads from RNA-seq libraries. Thus, it could be concluded that aforementioned approach to transcriptome *de novo* reconstruction based on creation of several individual assemblies followed by their combining into meta-assembly increases quality of *de novo* reconstructed transcriptome.

Comparison of several aligners showed that rnaSPAdes tool reconstructs fewer contigs, while Trans-ABySS reconstructs the highest amount of contigs. Trinity assembler reconstructs comparable quantities of contigs when run in two modes – *de novo* and genome-guided. At the same time, redundancy removal reduces sizes of Trans-ABySS assemblies most severely – in alm experiment 94.3 % of all contigs reconstructed by Trans-ABySS were removed, in blp experiment – 83.7 %. In the case of spades assembly, 15.3 and 22.4 % of all the contigs were removed, respectively. In trinity assemblies on average 80.5 and 70.6 % of contigs were removed, in genome-guided assemblies – 87.3 and 67.8 % of contigs,

respectively. Genome-guided assemblies have the highest sizes after redundancy removal in both experiments, spades assemblies – the lowest.

Spades reconstructs the largest contigs of all individual assemblers, which is indicated by highest N50 values and medians of contig lengths distribution. Lowest N50 value in alm experiment was observed in GG assembly, in blp experiment – in abyss assembly.

The highest completeness of all individual assemblies according to BUSCO criterion in alm experiment is attributed to trinity assembly. In blp experiment it is attributed to GG assembly. The lowest completeness according to BUSCO criterion is attributed to spades assembly in alm experiment and abyss assembly in blp experiment.

Conclusion

To conclude, in the two experiments difference in performance of the *de novo* transcriptome assemblers is observed, despite IonTorrent short read libraries being used in both experiments, and reconstructed transcriptome belonging to the same organism – *H. vulgare* barley. This suggests that implemented assemblers are sensitive to the input data, and their performance can vary depending on the data used.

However, on both accounts transcriptome meta-assemblies created from combined individual assemblies have higher quality than all individual assemblies, which indicates the effectiveness of the approach to *de novo* transcriptome reconstruction as building of meta-assemblies combining results of several individual *de novo* transcriptome assemblers.

References

Bürckert J.P., Dubois A.R.S.X., Faison W.J., Farinelle S., Charpentier E., Sinner R., Wienecke-Baldacchino A., Muller C.P. Functionally convergent B cell receptor sequences in transgenic rats expressing a human B cell repertoire in response to tetanus toxoid and measles antigens. *Front. Immunol.* 2017. DOI 10.3389/fimmu.2017.01834.

Bushmanova E., Antipov D., Lapidus A., Przhibelskiy A.D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *BioRxiv.* 2018. DOI 10.1101/420208.

Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A.D. rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies. *Bioinformatics*. 2016;32(14):2210-2212. DOI 10.1093/bioinformatics/btw218.

Cerveau N., Jackson D.J. Combining independent *de novo* assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC Bioinform.* 2016;17:525. PMid: 27938328. DOI 10.1186/s12859-016-1406-x.

Chang Z., Wang Z., Li G. The impacts of read length and transcriptome complexity for *de novo* assembly: a simulation study. *PLoS One.* 2014;9(4):e94825. PMid: 24736633. DOI 10.1371/journal.pone.0094825.

Cui J., Shen N., Lu Z., Xu G., Wang Y., Jin B. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the *Arabidopsis* transcriptome. *Plant Methods.* 2020;16:85. DOI 10.1186/s13007-020-00629-x.

Engström P.G., Steijger T., Sipos B., Grant G.R., Kahles A., Rätsch G., Goldman N., Hubbard T.J., Harrow J., Guigó R., Bertone P., Alioto T., Behr J., Bohnert R., Campagna D., Davis C.A., Dobin A., Gingeras T.R., Jean G., Kosarev P., Li S., Liu J., Mason C.E., Molodtsov V., Ning Z., Ponstingl H., Prins J.F., Ribeca P., Seledtsov I., Solovyev V., Valle G., Vitulo N., Wang K., Wu T.D., Zeller G. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods.* 2013;10:1185-1191. PMid: 24185836. DOI 10.1038/nmeth.2722.

Evangelistella C., Valentini A., Ludovisi R., Firrincieli A., Fabbrini F., Scalabrin S., Cattonaro F., Morgante M., Mugnozza G.S., Keurentjes J.J.B., Harfouche A. De novo assembly, functional annotation, and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock. *Biotechnol. Biofuels.* 2017;10:138. DOI 10.1186/s13068-017-0828-7.

Fu S., Ma Y., Yao H., Xu Z., Chen S., Song J., Au K.F. IDP-denovo: *de novo* transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics.* 2018;34(13):2168-2176. PMid: 28407034. DOI 10.1093/bioinformatics/bty098.

Gilbert D.G. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. *PeerJ.* 2019;7:e6374. DOI 10.7717/peerj.6374.

Glagoleva A.Y., Shmakov N.A., Shoeva O.Y., Vasiliev G.V., Shatskaya N.V., Börner A., Afonnikov D.A., Khlestkina E.K. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the *Black lemma and pericarp* (*Blp*) gene. *BMC Plant Biol.* 2017;17:182. DOI 10.1186/s12870-017-1124-1.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 2013;29:644-652. PMid: 21572440. DOI 10.1038/nbt.1883.Trinity.

Hölzer M., Marz M. *De novo* transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience.* 2019;8(5):giz039. PMid: 31077315. DOI 10.1093/gigascience/giz039.

Honaas L.A., Wafula E.K., Wickett N.J., Der J.P., Zhang Y., Edger P.P., Altman N.S., Chris Pires J., Leebens-Mack J.H., DePamphilis C.W. Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS One.* 2016;11(1):e0146062. PMid: 26731733. DOI 10.1371/journal.pone.0146062.

Hrdlickova R., Toloue M., Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA.* 2017;8:e1364. PMid: 27198714. DOI 10.1002/wrna.1364.

Jain P., Krishnan N.M., Panda B. Augmenting transcriptome assembly by combining *de novo* and genome-guided tools. *PeerJ.* 2013;1:e133. PMid: 24024083. DOI 10.7717/peerj.133.

Lafond-Lapalme J., Duceppe M.O., Wang S., Moffett P., Mimee B. A new method for decontamination of *de novo* transcriptomes using a hierarchical clustering algorithm. *Bioinformatics.* 2017;33(9):1293-1300. PMid: 28011783. DOI 10.1093/bioinformatics/btw793.

Lahens N.F., Ricciotti E., Smirnova O., Toorens E., Kim E.J., Baruzzo G., Hayer K.E., Ganguly T., Schug J., Grant G.R. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genom.* 2017;18:602. PMid: 28797240. DOI 10.1186/s12864-017-4011-0.

Lee S., La T.M., Lee H.J., Choi I.S., Song C.S., Park S.Y., Lee J.B., Lee S.W. Characterization of microbial communities in the chicken oviduct and the origin of chicken embryo gut microbiota. *Sci. Rep.* 2019;9:6838. PMid: 31048728. DOI 10.1038/s41598-019-43280-w.

Li Z., Chen Y., Mu D., Yuan J., Shi Y., Zhang H., Gan J., Li N., Hu X., Liu B., Yang B., Fan W. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief Funct. Genomics.* 2012;11(1):25-37. PMid: 22184334. DOI 10.1093/bfgp/elr035.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal.* 2011;17(1):10-12. PMid: 1000006697. DOI 10.14806/ej.17.1.200.

Payá-Milans M., Olmstead J.W., Nunez G., Rinehart T.A., Staton M. Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species. *GigaScience.* 2018;7(12):giy132. PMid: 30418578. DOI 10.1093/gigascience/giy132.

Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S.D., Mungall K., Lee S., Okada H.M., Qian J.Q., Griffith M., Ray-

mond A., Thiessen N., Cezard T., Butterfield Y.S., Newsome R., Chan S.K., She R., Varhol R., Kamoh B., Prabhu A.L., Tam A., Zhao Y., Moore R.A., Hirst M., Marra M.A., Jones S.J.M., Hoodless P.A., Birol I. *De novo* assembly and analysis of RNA-seq data. *Nat. Methods.* 2010;7(11):909-912. DOI 10.1038/nmeth.1517.

Salina E.A., Nesterov M.A., Frenkel Z., Kiseleva A.A., Timonova E.M., Magni F., Vrána J., Šafár J., Šimková H., Doležel J., Korol A., Sergeeva E.M. Features of the organization of bread wheat chromosome 5BS based on physical mapping. *BMC Genom.* 2018; 19:80. PMid: 29504906. DOI 10.1186/s12864-018-4470-y.

Schliesky S., Gowik U., Weber A.P.M., Bräutigam A. RNA-seq assembly – are we there yet? *Front. Plant Sci.* 2012;3:220. DOI 10.3389/fpls.2012.00220.

Schmieder R., Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics.* 2011;27:863-864. PMid: 21278185. DOI 10.1093/bioinformatics/btr026.

Schulz M.H., Zerbino D.R., Vingron M., Birney E. *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics.* 2012;28(8):1086-1092. PMid: 22368243. DOI 10.1093/bioinformatics/bts094.

Shekhovtsov S.V., Ershov N.I., Vasiliev G.V., Peltek S.E. Transcriptomic analysis confirms differences among nuclear genomes of cryptic earthworm lineages living in sympatry. *BMC Evol. Biol.* 2019; 19:50. PMid: 30813890. DOI 10.1186/s12862-019-1370-y.

Shmakov N.A., Vasiliev G.V., Shatskaya N.V., Doroshkov A.V., Gordeeva E.I., Afonnikov D.A., Khlestkina E.K. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq. *BMC Plant Biol.* 2016;16. DOI 10.1186/s12870-016-0926-x.

Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31: 3210-3212. PMid: 26059717. DOI 10.1093/bioinformatics/btv351.

Smith-Unna R., Boursnell C., Patro R., Hibberd J.M., Kelly S. TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res.* 2016;26:1134-1144. PMid: 27252236. DOI 10.1101/gr.196469.115.

Venturini L., Caim S., Kaithakottil G.G., Mapleson D.L., Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience.* 2018;7(8):giy093. PMid: 30052957. DOI 10.1093/gigascience/giy093.

Wang S., Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics.* 2017;33(3):327-333. PMid: 27694201. DOI 10.1093/bioinformatics/btw625.

Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Huang W., He G., Gu S., Li S., Zhou X., Lam T.W., Li Y., Xu X., Wong G.K.S., Wang J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics.* 2014;30(12):1660-1666. DOI 10.1093/bioinformatics/btu077.

Acknowledgements. The work was supported by Russian Science Foundation project No. 18-14-00293 (problem statement, algorithm creation, data analysis). Computational resources of Core Facility 'Bioinformatics' supported by budget project No. 0259-2021-0009 were implemented in this work.

Conflict of interest. The author declares no conflict of interest.

Received November 24, 2020. Revised January 15, 2021. Accepted January 15, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):39-45
DOI 10.18699/VJ21.005

Original Russian text www.bionet.nsc.ru/vogis/

The auxin signaling pathway to its PIN transporters: insights based on a meta-analysis of auxin-induced transcriptomes

V.V. Kovrizhnykh[1, 2] ✉, Z.S. Mustafin[1], Z.Z. Bagautdinova[1]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ vasilinakovr@gmail.com

Abstract. Active polar transport of the plant hormone auxin carried out by its PIN transporters is a key link in the formation and maintenance of auxin distribution, which, in turn, determines plant morphogenesis. The plasticity of auxin distribution is largely realized through the molecular genetic regulation of the expression of its transporters belonging to the PIN-FORMED (PIN) protein family. Regulation of auxin-response genes occurs through the ARF-Aux/IAA signaling pathway. However, it is not known which ARF-Aux/IAA proteins are involved in the regulation of *PIN* gene expression by auxin. In *Arabidopsis thaliana*, the PIN, ARF, and Aux/IAA families contain a larger number of members; their various combinations are possible in realization of the signaling pathway, and this is a challenge for understanding the mechanisms of this process. The use of high-throughput sequencing data on auxin-induced transcriptomes makes it possible to identify candidate genes involved in the regulation of PIN expression. To address this problem, we created an approach for the meta-analysis of auxin-induced transcriptomes, which helped us select genes that change their expression during the auxin response together with *PIN1*, *PIN3*, *PIN4* and *PIN7*. Possible regulators of ARF-Aux/IAA signaling pathway for each of the PINs under study were identified, and so were the aspects of their regulatory circuits both common for groups of *PIN* genes and specific for each *PIN* gene. Reconstruction of gene networks and their analysis predicted possible interactions between genes and served as an additional confirmation of the pathways obtained in the meta-analysis. The approach developed can be used in the search for gene expression regulators in other genome-wide data.
Key words: *Arabidopsis thaliana*; auxin; PIN-FORMED; auxin-response genes; meta-analysis; gene network.

For citation: Kovrizhnykh V.V., Mustafin Z.S., Bagautdinova Z.Z. The auxin signaling pathway to its PIN transporters: insights based on a meta-analysis of auxin-induced transcriptomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):39-45. DOI 10.18699/VJ21.005

Поиск участников сигнального пути ауксина к его транспортерам PIN на основе метаанализа транскриптомов, индуцированных ауксином

В.В. Коврижных[1, 2] ✉, З.С. Мустафин[1], З.З. Багаутдинова[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ vasilinakovr@gmail.com

Аннотация. Активный полярный транспорт гормона растений ауксина, осуществляемый его транспортерами, – ключевое звено в формировании и поддержании распределения ауксина, которое, в свою очередь, определяет морфогенез растения. Пластичность распределения ауксина в большой степени реализуется через молекулярно-генетическую регуляцию им экспрессии транспортеров семейства PIN-FORMED (PIN) белков. Регуляция ауксином экспрессии чувствительных к нему генов происходит через ARF-Aux/IAA-зависимый сигнальный путь. Однако неизвестно, какие ARF-Aux/IAA белки участвуют в регуляции ауксином экспрессии генов *PIN*. У *Arabidopsis thaliana* семейства белков PIN, ARF и Aux/IAA многочисленны, возможны различные комбинации представителей этих семейств в реализации сигнального пути, что создает сложность для понимания механизмов этого процесса. Использование данных высокопроизводительного секвенирования транскриптомов, индуцированных ауксином (RNA-Seq), делает возможным обнаружение генов-кандидатов, участвующих в регуляции экспрессии белков PIN. Мы разработали алгоритм метаанализа ауксин-индуцированных транскриптомов, с помощью которого отобрали гены, изменяющие свою экспрессию в ответе на ауксин вместе с *PIN1*, *PIN3*, *PIN4*, *PIN7*, и предсказали возможные регуляторы ARF-Aux/IAA сигнального пути для каждого из дифференциально экспрессирующихся *PIN*. Применяя сравнительный анализ, мы определили общие и специфичные аспекты в регуляторных контурах, исследуемых *PIN*. Реконструкция генных сетей и их оценка показали возможные взаи-

модействия между генами и послужили дополнительным подтверждением большинства сигнальных путей, полученных в метаанализе. С помощью комплексного подхода мы предсказали, что регуляция ауксином экспрессии *PIN* происходит через несколько ARF-Aux/IAA регуляторных контуров, опосредованных комбинацией *ARF4*, *ARF10* и *IAA4*, *IAA12*, *IAA17*, *IAA18* и *IAA32*. Часть из них являются специфичными при формировании ауксинового ответа с участием отдельных белков PIN, тогда как другие – общими для нескольких белков PIN. Разработанный алгоритм метаанализа можно применять для решения других задач поиска регуляторов экспрессии генов с привлечением полногеномных данных.
Ключевые слова: *Arabidopsis thaliana*; ауксин; PIN-FORMED; ауксин-регулируемые гены; метаанализ полногеномных данных; генные сети.

Introduction

The key role of auxin in regulation of plant growth and development is a well known fact (Mroue et al., 2018). A significant part of auxin is synthesized in the shoot apical meristems and then transferred to the root, providing there the development of lateral and adventitious roots, as well as the maintenance of the stem cell niche in the root apical meristem. At the cellular level, auxin role in physiological process is carried out by its concentration-dependent effect on cell division and elongation rate (Campanoni, Nick, 2005). Therefore, the formation and maintenance of auxin concentration gradients plays a vital role in morphogenesis. For example, in experiments on root decapitation, it was shown that auxin distribution with a concentration maximum located at a certain distance from the new root tip can be formed again in a few hours (Grieneisen et al., 2007; Mironova et al., 2010). In this case, the regeneration of meristem and normal root functioning occurs only after recovery of auxin distribution pattern (Xu et al., 2006).

The *PIN-formed* (*PIN*) family genes, which encode eight transmembrane transporter proteins in *Arabidopsis thaliana*, carry out auxin efflux from the cell (Weijers et al., 2001; Petrasek, 2006). PIN1-4, PIN7 transporters are polar localized on the cell plasma membrane, thereby the directed auxin flows are formed in the tissue. For example, at the individual cells level in *A. thaliana* root tip auxin fluxes forms hormone distribution with maximum in quiescent center (QC), which maintains the stem cell niche in the root (Feraru, Friml, 2008). In most cases, the PIN function is fundamental in formation and maintenance of auxin distribution. It was shown experimentally that there is a complex network of auxin-dependent regulation for PIN expression, which includes positive and negative feedbacks (Gelder et al., 2001; Friml, 2004; Sauer et al., 2006; Vieten et al., 2007). In the article of A. Vieten et al. (2005) it was experimentally shown that treatment with exogenous auxin leads to an increase in *PINs* transcription in the root, and the optimal auxin concentration for maximum increase differs for each of these genes. Later we showed that transcriptional and posttranscriptional regulation of *PIN1* expression by auxin have distinctive features (Omelyanchuk et al., 2016). At the transcriptional level, an increase in *PIN1* expression occurs in a wide range of exogenous auxin concentrations, while the PIN1 protein level changes nonlinearly, increasing with raising from low auxin concentration to medium, and then further increase in auxin concentration leads to PIN1 level decreasing.

The major mechanism of auxin-dependent genes regulation occurs through the ARF-Aux/IAA signaling pathway (Ulmasov et al., 1997). When auxin is absent, ARF transcription factors are bound by Aux/IAA co-repressors. Upon entering the cell, auxin interacts with TIR1 receptor, which forms SCF^{TIR1} ubiquitin ligase complex together with other proteins (Dharmasiri et al., 2005; Kepinski, Leyser, 2005). Further, this complex binds to Aux/IAA proteins, regulating their degradation in 26S proteasome (Calderon-Villalobos et al., 2010; Hayashi, 2012). Thus, ARF transcription factors activate or suppress transcription of auxin response genes. In *A. thaliana* genome, 29 *Aux/IAA* and 23 *ARF* genes were found; their expression in different cell types is various, creating sufficient molecular complexity to provide a variety of auxin responses (Remington et al., 2004; Teale et al., 2006). However, it is not known which ARF-Aux/IAA proteins are involved in auxin regulation of *PIN* expression. It is only known that ARF binding sites were found in promoters of all PINs with bioinformatics methods (Habets, Offringa, 2014).

Reconstruction of the auxin signaling pathway to its PIN transporters is challenging for direct solution by experimental methods. Here, we carried out a meta-analysis of auxin-induced transcriptomes in order to obtain a list of genes that significantly change expression together with *PINs* in response to auxin. A complex approach, including a comparative analysis of these lists and gene networks reconstructed based on those lists, predicted the participants in the ARF-Aux/IAA signaling pathway involved in *PIN* expression regulation by auxin. Thus, the common signaling pathways for *PIN1*, *PIN3*, *PIN7* are mediated by combination of *ARF4* with *IAA12* and *IAA18.* At the same time, the specific auxin regulation for individual *PINs* is probably carried out by other proteins of ARF-Aux/IAA signaling pathway. For example, our results showed that *ARF10* and *IAA32* were present only in the list of genes, which significantly change expression along with *PIN4*. In addition, we noted the genes that are associated with post-transcriptional regulation of PINs activity in the candidate genes list.

Materials and methods

Information used in the meta-analysis. In this study, publicly available data on *A. thaliana* auxin-induced transcriptomes (microarrays and RNA sequencing) were used. Most of the data were previously presented in (Cherenkov et al., 2018). The summary table of the data has been expanded by the information from (Omelyanchuk et al., 2017). As a result, we took the results of 22 experiments for the meta-analysis. Genes were considered differentially expressed (DEG) if the *p*-value (according to Benjamini–Hochberg) was less than 0.05. The sets of experiments (Supplementary 1)[1] for each PIN were al-

[1] Supplementary materials 1–3 are available in the online version of the paper: http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx1.pdf

located according to the algorithm we developed (see section "Results. Meta-analysis algorithm"). Work with the summary table and lists of data was carried out using standard methods of Excel (filters, conditional formatting).

Gene networks reconstruction. Based on lists of DEGs, gene networks were reconstructed using the String resource (https://string-db.org/) (Szklarczyk et al., 2019). String creates gene networks using user-specified criteria, combining the genes according to the following types of links: experimentally determined (e.g. affinity chromatography), databases (an edge retrieved from the data in databases), textmining (genes found together in publications), co-expression (the same expression patterns of mRNA), neighborhood (calculated based on the proximity of the distance between genes in different genomes), gene fusion (hybrid genes formed in the course of evolution from previously independent genes as a result of chromosomal rearrangements), co-occurrence (presence or absence of linked proteins across species), protein homology. Each link has its own score, calculated through the String algorithms.

Results

Meta-analysis algorithm

Stage 1: data collection. We form a summary table of all publicly available microchip experiments and RNA sequencing data on the topic of interest. In our case, this is information about differentially expressed genes in response to auxin treatment for *A. thaliana*. The collected data can be geterogenous, for example, our meta-analysis contains data from 22 experiments, containing two samples types (root, whole seedling), three development stages (3-, 5–7-, 10–12 dag seedlings), five time intervals of treatment (0.5–1 h, 2–4 h, 6–8 h, 12–24 h), six types of auxin and its concentrations (0.1; 1; 5; 10 μM IAA; 10 μM NAA; 10 μM IBA).

Stage 2: selection of the experiments appropriate to the task. In the summary table obtained at Stage 1, we find the experiments, in which there was a change in gene expression, for which we are looking for regulators. In accordance with our issue, it is known that *A. thaliana* has eight PIN transporters. We found *PIN1* (in five experiments), *PIN3* (in eight experiments), *PIN4* (in one experiment) and *PIN7* (in six experiments) differentially expressed in these auxin-induced public transcriptomes.

Stage 3: identification of genes that change their expression under auxin influence along with *PIN* genes. Separately, for each *PIN* we selected only those DEGs that changed exclusively in experiments where this *PIN* changes expression, and in other experiments DEG was absent. Thus, we identify genes potentially involved in *PIN* regulation by auxin. There also may be genes that are direct targets of auxin gradient changes due to PIN proteins activity. For each studied *PIN*, a table is formed that contains information about activation of suppression of each DEG under auxin treatment. The DEG is marked in the table only if it is differentially expressed along with *PIN* in at least one experiment.

Stage 4: the formation of DEGs lists that significantly change expression together with *PIN*. We used the binomial distribution to determine the number of experiments, in which the gene is a DEG along with *PIN*, to consider this event non-random ($p > 95$ %). For each gene list, the significance threshold differs according to amount of experiments, in which a certain *PIN* is differentially expressed (see Stage 2). In our case, for *PIN3* DEG is considered significant if its expression changes occur in three or more experiments, for *PIN1* and *PIN7* – in two or more experiments. Since *PIN4* is differentially expressed only in one experiment, the list of DEGs that change expression along with *PIN4* will not vary from Stage 3.

Stage 5: identification of common and specific gene groups. Comparing DEG lists from previous stage with each other we highlight genes found in several lists, i. e. common for *PINs*, and also mark genes found only in one list, thereby identifying genes that specifically change expression together with a certain *PIN*.

Stage 6: gene networks reconstruction. Using prepared lists of DEGs from Stage 4, we create gene networks for each *PIN* and reconstruct interactions between all genes of each list. The connectivity of this network reflects the gene set, for which one of interaction types available in the String database has been found (textminig, co-expression, co-occurrence, etc.).

Stage 7: analysis of gene networks composition. First of all, we pay attention to genes for which links to the genes under study are found in String, paying attention to the type of the interaction. Then from the ontologies list we select biological process that are related to the studied issue. In our study, we chose the auxin-activated signaling pathway.

Using the meta-analysis algorithm described above, we obtained several candidate genes, which regulate *PIN* expression with a high probability. Next, we describe the results of the reconstruction of auxin signaling pathway to its PIN transporters.

Meta-analysis of auxin-induced transcriptomes

Initially, the collected auxin-induced transcriptomes contained more than 20 thousand DEGs that change expression in response to auxin treatment. Among these DEGs, there were four members of *PIN* family: *PIN1*, *PIN3*, *PIN4*, *PIN7*. After performing the meta-analysis algorithm described above, we selected four lists of DEGs, jointly changing the expression with *PIN1*, *PIN3*, *PIN4*, *PIN7*, respectively (Supplementary 2). In total, expression of 531 genes significantly increased and 236 genes decreased their expression jointly with *PINs* (Fig. 1). Together with *PIN1*, the expression of 378 genes was significantly altered, of which 375 genes increased the expression level in auxin response similar to *PIN1*. For the rest of *PIN* genes, the difference in number of suppressed and activated potential regulators was not so great.

Then, we compared the lists with each other and determined common DEGs for several *PINs* and specific DEGs to each *PIN* gene. Twelve groups of genes were obtained: specific auxin-activated genes and specific suppressed genes were found for each *PIN*, as well as two groups of auxin-activated genes common for (*PIN1*, *PIN3*, *PIN7*) and (*PIN1*, *PIN7*); two groups of suppressed genes by auxin, common to (*PIN3*, *PIN7*) and (*PIN1*, *PIN3*). Activated and suppressed *PIN4* potential regulators don't overlap with those for other *PINs*. Since among potential regulators of *PIN* activity there were

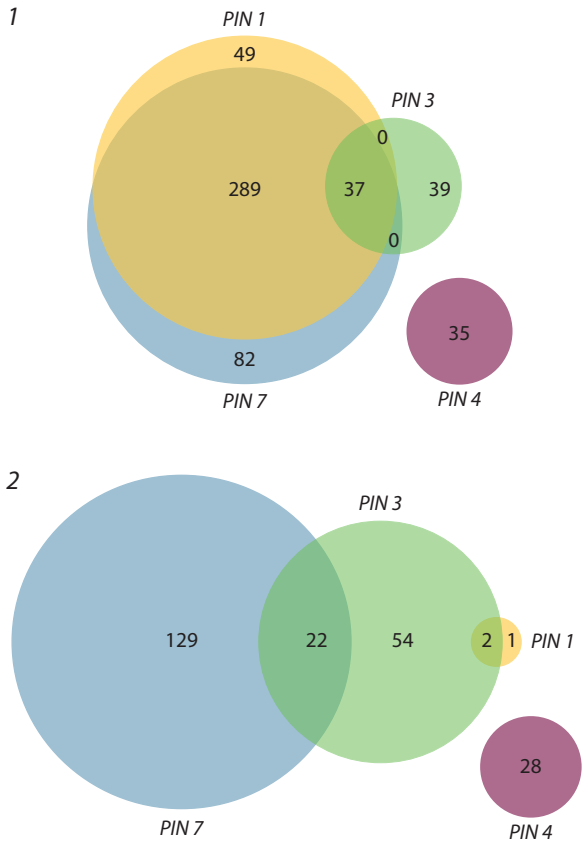


Fig. 1. Twelve groups of genes identified in meta-analysis that significantly change their expression together with *PIN1*, *PIN3*, *PIN4* and *PIN7*.
1 – auxin activated genes; *2* – auxin inhibited genes.

participants of auxin signaling pathway, we searched for them in the lists (see Supplementary 2) and described to which DEG groups they belong.

Prediction of auxin-dependent regulators of *PIN* gene expression

Since the meta-analysis predicted auxin-dependent regulators of *PIN* gene expression, we isolated genes for transcriptional and post-transcriptional regulation in DEG lists. We searched for possible transcriptional regulators only among ARF transcription factors and IAA proteins. Possible post-transcriptional regulators have been identified among members of known protein families that affect the PIN protein localization on cell membrane.

Possible regulators of PIN expression at the transcriptional level

As a result of meta-analysis, we found that *ARF4* and *IAA12*, *IAA18* are the common potential regulators for (*PIN1*, *PIN3*, *PIN7*). *IAA4* has been identified as a specific regulator for *PIN1*, while *ARF10* and *IAA32* presumably mediated auxin response for *PIN4*. In addition, *IAA17* was found in a group of genes that change their expression with *PIN1* and *PIN7*. Interestingly, we didn't find transcription factors of Aux/IAA family among specific regulators of *PIN3* and *PIN7*, but we did find regulators belonging to other transcription factors families. Therefore, there are obvious differences in ARF-Aux/IAA sets for studied *PIN* genes, which may also cause differences in dose-dependent regulation of these transporters by auxin.

Possible regulators of PIN polar localization

According to the published data, PIN proteins circulate between plasma membrane and cytoplasm in vesicles. This process is regulated by BIG, GN, ARF1 proteins and AGC, PID kinases families, and their functioning is controlled by auxin (Dhonukshe, 2011). Moreover, the polar localization of PIN proteins is also influenced by ABCB1, ABCB19 and ROPGEF protein family (Pan et al., 2015). In the course of data meta-analysis, among DEGs in response to auxin treatment we found a downregulation of *BIG4* and *ROPGEF11* in gene lists that change expression jointly with *PIN7* and *PIN4*, respectively. An upregulation was noted for *WAG2* (member of AGC kinase family) in the group of genes that change their expression along *PIN1* and *PIN7*.

In addition, in our opinion, it is interesting that *RGF6/GLV1/CLEL6* RNA of signal peptide was upregulated in response to auxin in experiments where activity of *PIN1* and *PIN7* is increased. Another peptide from RGF/GLV/CLEL family, *RGF8/GLV6/CLEL2*, was increased in experiments where only *PIN7* changed expression.

Thus, the formation of auxin response for (*PIN1*, *PIN3*, *PIN7*) group is due to common signaling pathways mediated by *ARF4* and *IAA12*, *IAA18*. Additionally, there are ARF-Aux/IAA specific paths for *PIN1* and *PIN4*. Also among the known auxin-sensitive genes affecting PIN polar localization, we found downregulation of *BIG4* and *ROPGEF11*, which probably contributes to specific responses of *PIN7* and *PIN4*, respectively.

Reconstruction of gene networks

We used the lists of DEGs for each *PIN* and reconstructed gene networks, which made it possible to evaluate described DEG interaction and, most importantly, how all these DEGs can affect PIN expression activity. As a result, we obtained the connected networks, in which interactions with *PIN* genes were found, only for *PIN1*, *PIN3* and *PIN7*. The meta-analysis, from which gene lists for network reconstruction were made, provides significance in itself, so we used a linkage threshold of 0.4. Since we are interested in reconstruction of auxin signaling pathway, we noted only this biological process in String. Notably, most links are formed based on automatic analysis of the articles texts. In the gene network reconstructed based on DEGs that change expression along with PIN1, 12 genes related to the activation of auxin signaling pathway were found (Supplementary 3). At the same time, IAA12, IAA17 (AXR3), WAG2, AUX1 were directly associated with PIN1, the other genes of auxin response were associated with PIN1 indirectly (Fig. 2). It can also be noted that AIL6/PLT3 and AVP1, which are related to the auxin-regulated organ development in *Arabidopsis*, were directly associated with PIN1 (Krizek, 2011). These genes can be attributed to genes that are direct targets of auxin gradient changes under PIN action. Among these genes, the links between PIN1 and AIL6 and

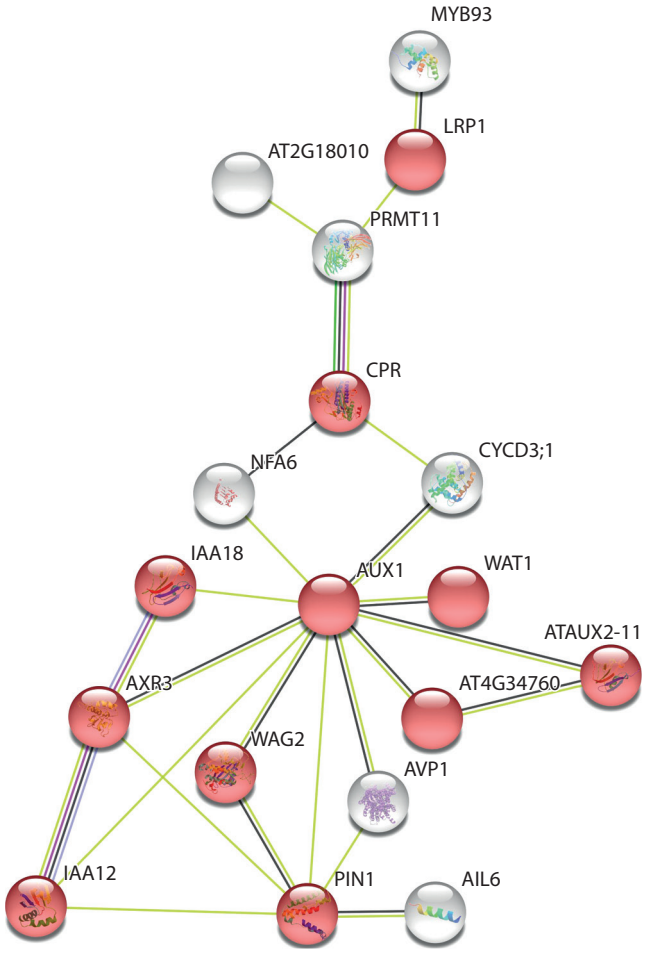


Fig. 2. A fragment of the gene network, containing genes associated with PIN1 and genes related to the auxin signaling pathway.

Red circles denote genes traditionally related to the auxin signaling pathway. Grey – genes identified in meta-analysis, for which direct or indirect links to PIN1 were found in String. The color of the link reflects what kind of data String used for the creation of interaction. Yellow links are based on textmining; black – on co-expression data, blue – on protein homology, pink – on protein-protein interactions.

WAG2 were constructed based on co-expression data of RNA sequencing experiments.

Reconstructed gene network for DEGs that alter expression jointly with PIN3 contained eight genes traditionally related to auxin signaling pathway (see Supplementary 3). Direct interactions to PIN3 have been found for AUX1, IAA12 and SAUR9. In the gene network for PIN7, fourteen genes belonged to traditional auxin signaling pathway. At the same time PIN7 directly interacts with IAA12, IAA17 (AXR3), AUX1, LPR1 and WAG2 (see Supplementary 3). In addition, PIN7 had direct links with ABCG33, NFA6, PHOT1, YUC2, YUC6, related to other biological processes controlled by auxin. Reconstruction of gene networks is an additional verification of the fact that regulation of PIN expression by auxin likely occurs with participation of IAA12 and IAA17. It should be noted that the absence of direct connections with PINs for the rest of predicted by meta-analysis ARF-Aux/IAA regulators does not exclude them from the list of candidates for experimental verification in the future.

Discussion

Phytohormones are actively involved in the processes of plant growth and morphogenesis. The action of auxin in these processes is well studied and it is based on the changes in auxin distribution in tissues (Mroue et al., 2018). Consequently, auxin concentration is a limiting factor in determining cell fate. Proteins-transporters of the PIN family play an important role in the realization of the morphogenetic action of auxin, since they create directed fluxes of this hormone in tissues and, thus, mediate the formation of auxin concentration gradients (Vanneste, Friml, 2009).

An important aspect in the process described above is the presence of positive and negative feedback loops in the mutual regulation of auxin efflux from the cell through PIN functioning and the number of these transporters controlled by auxin. The regulation of auxin-sensitive genes expression is mediated by two proteins families. The first family is ARF transcription factors, which bind to AuxRe site in the promoter of the auxin-sensitive gene and act as an activator or repressor of gene expression (Ulmasov et al., 1997). In some sources, only ARF5-ARF8, ARF19 are supposed to be activators of expression, but there is no experimental confirmation of this (Guilfoyle, Hagen, 2007). The second is the Aux/IAA corepressors, which in the absence of auxin are associated with ARF.

Previously, it was reported, that PIN1–4, PIN7 expression was downregulated in axr3/iaa17 and solitary-root-1(slr-1)/iaa14 mutants (Vieten et al., 2005) and PIN1 expression is regulated by ARF5 transcription factor (Wenzel et al., 2007), which interacts with IAA12 (Hamann, 2002). In the present work, using computer methods of meta-analysis for genome-wide data and gene networks reconstruction, we predicted the details of the auxin signaling pathway to its PIN transporters. The results indicate that there are common mechanisms for *PIN1*, *PIN3*, *PIN7* and *PIN1*, *PIN7* transcription regulation by auxin, as well as specific mechanisms for PIN expression regulation by auxin. By the common mechanism for *PIN1*, *PIN3*, *PIN7*, we predict the activation of their expression through ARF4-IAA12, ARF4-IAA18, and for *PIN1* and *PIN7* – additionally through ARF4-IAA17. Specific mechanisms are implemented via ARF4-IAA4 and ARF10-IAA32 for *PIN1* and *PIN4*, respectively. The interactions between these ARFs and IAAs have been experimentally confirmed (Paponov et al., 2008). Recently, it was shown that salinity downregulates *PIN* expression and leads to stabilization of IAA17 (Liu et al., 2015). Moreover, this type of stress causes a decrease in the size of root apical meristem due to a decline in auxin accumulation, mediated by PIN1, PIN3, PIN7 downregulation. In our data, in auxin-induced transcriptomes, an increase in the expression of *PIN1* and *PIN7* is accompanied by an increase in *IAA17* expression.

For signal peptides of the RGF/GLV/CLEL family, it was previously noted that during gravitropism they change the auxin gradient in the hypocotyl and root (Whitford et al., 2012). At the root, this is due to regulation of PIN2 protein localization by peptides of this family. It was shown that peptides GLV3 and, possibly, GLV6 and GLV9, are secreted from the cortex and endodermis and pass into the outer layers

to regulate PIN2 localization. The GLV1 peptide is not expressed in the root, but is present in the hypocotyl, where it also changes the auxin gradient during gravitropism, both during overexpression and loss of function upon mutation (Whitford et al., 2012). According to our data, RGF/GLV/CLEL peptides are involved in the signaling pathway that regulates PIN1 and PIN7 protein localization, and possibly indirectly affect the increase in the expression of these *PIN* genes. Overexpression or treatment of GLV1 leads to lengthening of the root and its apical meristem due to the fact that the zone of cell division in the root increases, i. e., cells later proceed to differentiation (Fernandez et al., 2013). This transition is also associated with a change in auxin distribution, which is formed by its transporters.

Conclusion

Thus, created algorithm for the meta-analysis of genome-wide data was applied to finding participants and reconstructing the auxin signaling pathway to its transporters. We were able to reveal that auxin controls *PIN1*, *PIN3*, *PIN7* expression both through common regulators and specifically, while for *PIN4* only specific regulators have been identified. We found published experimental data that partially support our assumptions. As a result of computer research, we have nominated new candidates for experimental verification.

References

Calderon-Villalobos L.I., Tan X., Zheng N., Estelle M. Auxin perception – structural insights. *Cold Spring Harb. Perspect. Biol.* 2010;2: a005546-a005546. DOI 10.1101/cshperspect.a005546.

Campanoni P., Nick P. Auxin-dependent cell division and cell elongation. 1-Naphthaleneacetic acid and 2,4-dichlorophenoxyacetic acid activate different pathways. *Plant Physiol.* 2005;137:939-948. DOI 10.1104/pp.104.053843.

Cherenkov P., Novikova D., Omelyanchuk N., Levitsky V., Grosse I., Weijers D., Mironova V. Diversity of cis-regulatory elements associated with auxin response in *Arabidopsis thaliana*. *J. Exp. Bot.* 2018; 69:329-339. DOI 10.1093/jxb/erx254.

Dharmasiri N., Dharmasiri S., Estelle M. The F-box protein TIR1 is an auxin receptor. *Nature*. 2005;435:441-445. DOI 10.1038/nature 03543.

Dhonukshe P. PIN polarity regulation by AGC-3 kinases and ARF-GEF. *Plant Signal*. *Behav.* 2011;6:1333-1337. DOI 10.4161/psb.6.9. 16611.

Feraru E., Friml J. PIN polar targeting. *Plant Physiol.* 2008;147:1553-1559. DOI 10.1104/pp.108.121756.

Fernandez A., Hilson P., Beeckman T. GOLVEN peptides as important regulatory signalling molecules of plant development. *J. Exp. Bot.* 2013;64:5263-5268. DOI 10.1093/jxb/ert248.

Friml J. A PINOID-dependent binary switch in apical-basal PIN polar targeting directs auxin efflux. *Science.* 2004;306:862-865. DOI 10.1126/science.1100618.

Geldner N., Friml J., Stierhof Y.-D., Jürgens G., Palme K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature.* 2001;413:425-428. DOI 10.1038/35096571.

Grieneisen V.A., Xu J., Marée A.F.M., Hogeweg P., Scheres B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature.* 2007;449:1008-1013. DOI 10.1038/nature 06215.

Guilfoyle T.J., Hagen G. Auxin response factors. *Curr. Opin. Plant Biol.* 2007;10:453-460. DOI 10.1016/j.pbi.2007.08.014.

Habets M.E.J., Offringa R. PIN-driven polar auxin transport in plant developmental plasticity: a key target for environmental and endogenous signals. *New Phytol.* 2014;203:362-377. DOI 10.1111/nph. 12831.

Hamann T. The *Arabidopsis BODENLOS* gene encodes an auxin response protein inhibiting MONOPTEROS-mediated embryo patterning. *Genes Dev.* 2002;16:1610-1615. DOI 10.1101/gad.229402.

Hayashi K. The interaction and integration of auxin signaling components. *Plant Cell Physiol.* 2012;53:965-975. DOI 10.1093/pcp/ pcs035.

Kepinski S., Leyser O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature.* 2005;435:446-451. DOI 10.1038/nature 03542.

Krizek B.A. Auxin regulation of *Arabidopsis* flower development involves members of the AINTEGUMENTA-LIKE/PLETHORA (AIL/PLT) family. *J. Exp. Bot.* 2011;62:3311-3319. DOI 10.1093/ jxb/err127.

Liu W., Li R.-J., Han T.-T., Cai W., Fu Z.-W., Lu Y.-T. Salt stress reduces root meristem size by nitric oxide-mediated modulation of auxin accumulation and signaling in *Arabidopsis*. *Plant Physiol.* 2015;168:343-356. DOI 10.1104/pp.15.00030.

Mironova V.V., Omelyanchuk N.A., Yosiphon G., Fadeev S.I., Kolchanov N.A., Mjolsness E., Likhoshvai V.A. A plausible mechanism for auxin patterning along the developing root. *BMC Syst. Biol.* 2010; 4:98. DOI 10.1186/1752-0509-4-98.

Mroue S., Simeunovic A., Robert H.S. Auxin production as an integrator of environmental cues for developmental growth regulation. *J. Exp. Bot.* 2018;69:201-212. DOI 10.1093/jxb/erx259.

Omelyanchuk N.A., Kovrizhnykh V.V., Oshchepkova E.A., Pasternak T., Palme K., Mironova V.V. A detailed expression map of the PIN1 auxin transporter in *Arabidopsis thaliana* root. *BMC Plant Biol.* 2016;16:5. DOI 10.1186/s12870-015-0685-0.

Omelyanchuk N.A., Wiebe D.S., Novikova D.D., Levitsky V.G., Klimova N., Gorelova V., Weinholdt C., Vasiliev G.V., Zemlyanskaya E.V., Kolchanov N.A., Kochetov A.V., Grosse I., Mironova V.V. Auxin regulates functional gene groups in a fold-change-specific manner in *Arabidopsis thaliana* roots. *Sci. Rep.* 2017;7:2489. DOI 10.1038/ s41598-017-02476-8.

Pan X., Chen J., Yang Z. Auxin regulation of cell polarity in plants. *Curr. Opin. Plant Biol*. 2015;28:144-153. DOI 10.1016/j.pbi.2015. 10.009.

Paponov I.A., Paponov M., Teale W., Menges M., Chakrabortee S., Murray J.A.H., Palme K. Comprehensive transcriptome analysis of auxin responses in *Arabidopsis. Mol. Plant.* 2008;1:321-337. DOI 10.1093/mp/ssm021.

Petrasek J. PIN proteins perform a rate-limiting function in cellular auxin efflux. *Science.* 2006;312:914-918. DOI 10.1126/science. 1123542.

Remington D.L., Vision T.J., Guilfoyle T.J., Reed J.W. Contrasting modes of diversification in the *Aux/IAA* and *ARF* gene families. *Plant Physiol.* 2004;135:1738-1752. DOI 10.1104/pp.104. 039669.

Sauer M., Balla J., Luschnig C., Wisniewska J., Reinohl V., Friml J., Benkova E. Canalization of auxin flow by Aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes Dev.* 2006;20:2902-2911. DOI 10.1101/gad.390806.

Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J., von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47: D607-D613. DOI 10.1093/nar/gky1131.

Teale W.D., Paponov I.A., Palme K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* 2006;7:847-859. DOI 10.1038/nrm2020.

Ulmasov T., Murfett J., Hagen G., Guilfoyle T.J. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell.* 1997;9:1963-1971. DOI 10.1105/tpc.9.11.1963.

Vanneste S., Friml J. Auxin: a trigger for change in plant development. *Cell.* 2009;136:1005-1016. DOI 10.1016/j.cell.2009.03.001.

Vieten A., Sauer M., Brewer P.B., Friml J. Molecular and cellular aspects of auxin-transport-mediated development. *Trends Plant Sci.* 2007;12:160-168. DOI 10.1016/j.tplants.2007.03.006.

Vieten A., Vanneste S., Wisniewska J., Benkova E., Benjamins R., Beeckman T., Luschnig C., Friml J. Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression. *Development.* 2005;132:4521-4531. DOI 10.1242/dev.02027.

Weijers D., Franke-van Dijk M., Vencken R.J., Quint A., Hooykaas P., Offringa R. An Arabidopsis Minute-like phenotype caused by a semi-dominant mutation in a RIBOSOMAL PROTEIN S5 gene. *Development.* 2001;128:4289-4299.

Wenzel C.L., Schuetz M., Yu Q., Mattsson J. Dynamics of MONOPTEROS and PIN-FORMED1 expression during leaf vein pattern formation in *Arabidopsis thaliana*. *Plant J.* 2007;49:387-398. DOI 10.1111/j.1365-313X.2006.02977.x.

Whitford R., Fernandez A., Tejos R., Pérez A.C., Kleine-Vehn J., Vanneste S., Drozdzecki A., Leitner J., Abas L., Aerts M., Hoogewijs K., Baster P., De Groodt R., Lin Y.-C., Storme V., Van de Peer Y., Beeckman T., Madder A., Devreese B., Luschnig C., Friml J., Hilson P. GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. *Dev. Cell.* 2012;22:678-685. DOI 10.1016/j.devcel.2012.02.002.

Xu J., Hofhuis H., Heidstra R., Sauer M., Friml J., Scheres B. A molecular framework for plant regeneration. *Science.* 2006;311:385-388. DOI 10.1126/science.1121790.

ORCID ID

Z.S. Mustafin orcid.org/0000-0003-2724-4497

Acknowledgements. This work was supported by budget project No. 0259-2021-0009 and the Presidents grant RF MK-3470.2021.1.4.

Funding transparency. The authors do not hold financial interests in the presented materials or methods.

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

Received October 24, 2020. Revised January 12, 2021. Accepted January 14, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):46-56
DOI 10.18699/VJ21.006

Original Russian text www.bionet.nsc.ru/vogis/

Phylostratigraphic analysis of gene networks of human diseases

Z.S. Mustafin[1] ✉, S.A. Lashin[1, 2], Yu.G. Matushkin[1]

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

Abstract. Phylostratigraphic analysis is an approach to the study of gene evolution that makes it possible to determine the time of the origin of genes by analyzing their orthologous groups. The age of a gene belonging to an orthologous group is defined as the age of the most recent ancestor of all species represented in that group. Such an analysis can reveal important stages in the evolution of both the organism as a whole and groups of functionally related genes, in particular gene networks. In addition to investigating the time of origin of a gene, the level of its genetic variability and what type of selection the gene is subject to in relation to the most closely related organisms is studied. Using the Orthoscape application, gene networks from the KEGG Pathway, Human Diseases database describing various human diseases were analyzed. It was shown that the majority of genes described in gene networks are under stabilizing selection and a high reliable correlation was found between the time of gene origin and the level of genetic variability: the younger the gene, the higher the level of its variability is. It was also shown that among the gene networks analyzed, the highest proportion of evolutionarily young genes was found in the networks associated with diseases of the immune system (65 %), and the highest proportion of evolutionarily ancient genes was found in the networks responsible for the formation of human dependence on substances that cause addiction to chemical compounds (88 %); gene networks responsible for the development of infectious diseases caused by parasites are significantly enriched for evolutionarily young genes, and gene networks responsible for the development of specific types of cancer are significantly enriched for evolutionarily ancient genes.
Key words: evolution; phylostratigraphic analysis; ortholog; gene network; gene age.

For citation: 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

Филостратиграфический анализ генных сетей заболеваний человека

З.С. Мустафин[1] ✉, С.А. Лашин[1, 2], Ю.Г. Матушкин[1]

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

Аннотация. Филостратиграфический анализ – это подход к исследованию эволюции генов, позволяющий определить время возникновения генов за счет анализа филогенетических деревьев организмов, обладающих ортологичными к исследуемому генами. Такой анализ может открыть важные этапы в эволюции как организма в целом, так и групп функционально связанных генов, в частности генных сетей. В дополнение к исследованию времени возникновения гена изучается уровень его генетической изменчивости и то, какому типу отбора подвержен ген по отношению к наиболее близкородственным организмам. С помощью приложения Orthoscape были проанализированы генные сети из базы данных KEGG Pathway, Human Diseases, ассоциированные с заболеваниями человека. Выявлено, что большинство генов, описанных в генных сетях, подвержены стабилизирующему отбору, обнаружена высокая достоверная корреляция между временем возникновения гена и уровнем генетической изменчивости, которой он подвержен, – чем моложе ген, тем выше уровень генетической изменчивости. Было также показано, что среди проанализированных генных сетей наибольшая доля эволюционно молодых генов обнаружена в сетях, связанных с заболеваниями иммунной системы (65 %), а эволюционно древних генов – в сетях, ответственных за формирование зависимостей человека от веществ, вызывающих привыкание к химическим соединениям (88 %); генные сети, связанные с развитием инфекционных заболеваний, вызванных паразитами, достоверно обогащены эволюционно молодыми генами, а генные сети, ответственные за развитие специфических типов рака, – эволюционно древними генами.
Ключевые слова: эволюция; филостратиграфия; ортолог; генная сеть; возраст гена.

Introduction

The study of key factors that influence to the development of diseases is one of the most important research areas in both medicine and biology (Stepanov, 2016). It is known that the formation of phenotypic traits that provide the adaptation of organisms to environmental conditions is controlled not by individual genes, but by gene networks – groups of coordinately functioning genes and their products (RNA, proteins, metabolites, etc.) (Kolchanov et al., 2013). The task of highlighting the key structural features of networks, network elements, and their numerical description arises. One of the important characteristics in evolutionary biology is the age of a gene. The age of a gene belonging to an orthologous group is defined as the age of the most recent ancestor of all species represented in this group (Liebeskind et al., 2016).

Modern methods of analysis make it possible to evaluate the evolutionary characteristics of genes, in particular, phylostratigraphic analysis, a methodology proposed in 2007 by T. Domazet-Lošo, which allows to determine the age of a gene using a special index. The index is derived from analysis of orthologous genes and comparison of the position of organisms whose genes are considered in the analysis on a phylogenetic tree (Domazet-Lošo et al., 2007).

There are many software tools to work with gene networks. Some of them focus on reconstructing networks based on data from biological databases: String (Szklarczyk et al., 2019), GeneMANIA (Montojo et al., 2010). The others have extensive functionality for visualizing network elements and identifying its structural features: Cytoscape (Shannon et al., 2003), yEd (https://www.yworks.com/products/yed). Cytoscape has an advantage that in addition to its extensive capabilities of constructing networks, layouting and painting their elements and analyzing structural features, it allows users to write their own applications in Java. It makes possible for community to implement any functionality and plug in to Cytoscape. For example, well-known tools String and GeneMANIA for networks reconstruction from the list of genes based on extracting interactions from biological databases have their own plugins in Cytoscape and allow to use their functionality by combining it with the capabilities of Cytoscape and its other plugins. Also, the plugins allow to import networks from databases, for example, Pathway Commons (Cerami et al., 2011) or KEGG Pathway (Kanehisa et al., 2017), without hard parsing the formats of network representation.

The results of gene network analysis by one of such applications – Orthoscape (Mustafin et al., 2017), are presented in this paper. Orthoscape can analyze the evolutionary features of genes in gene network. It has been shown that most of the genes described in gene networks are under influence of stabilizing selection. A high reliable correlation has been found between the time of occurrence of a gene and the level of its genetic variability – the younger the gene, the higher the level of genetic variability is. Among the gene networks analyzed, the highest proportion of evolutionary young genes was detected in the networks associated with immune diseases (65 %), and the highest proportion of evolutionary ancient genes was detected in the networks responsible for the substance dependencies (88 %); gene networks responsible for the development of infectious diseases caused by parasites are significantly enriched in evolutionary young genes, and gene networks responsible for the development of specific types of cancer are significantly enriched in evolutionary ancient genes.

Materials and methods

Input data. Gene networks from KEGG Pathway, Human Diseases are used in this work. These networks are divided into 11 categories (with total number of 80 networks): neurodegenerative diseases (5 networks), cardiovascular diseases (5 networks), immune diseases (8 networks), endocrine and metabolic diseases (6 networks), infectious diseases: bacterial (10 networks), infectious diseases: viral (9 networks), infectious diseases: parasitic (6 networks), drug resistance: antineoplastic (4 networks), cancers: overview (7 networks), cancers: specific types (15 networks), substance dependence (5 networks).

The data required for the analysis, such as lists of orthologous genes, nucleotide sequences of genes and amino acid sequences of the proteins they encode, protein domains, taxonomic information about organisms whose genes were considered in the analysis were also taken from the KEGG database.

Software used. The analysis was performed using the Cytoscape software package (Shannon et al., 2003). CyKEGG Parser plugin was used to import networks from the KEGG Pathway (Nersisyan et al., 2014). Orthoscape plugin was used to perform phylostratigraphic analysis and analysis of so called divergence index – the index of evolutionary variability (Mustafin et al., 2017).

Methods for estimation the evolutionary characteristics of genes. Orthoscape allows to estimate two evolutionary characteristics of genes. The first one is *phylostratigraphic age index* (PAI). This index shows how far from the root of the phylogenetic tree is the taxon reflecting the age of the gene, i. e., the taxon where the studied species diverged from the most distant related taxon in which the ortholog of the studied gene was found. Thus, the more PAI of the gene, the younger it is (Fig. 1). KEGG Orthology service is used in Orthoscape to calculate PAI, which makes it possible to consider orthologous genes among all homologs.

Figure 1 shows examples of determining the age of a gene and the phylostratigraphic index, using human genes. On the left (*a*) the case when the most distant organism in which the ortholog of the studied gene was found is the bonobo is shown. The node most distant from the root of the phylogenetic tree that is common to *H. sapiens* and *P. paniscus* (bonobos) is Hominidae. It corresponds to the phylostratigraphic index is equal to 13. On the right (*b*) is the gene whose ortholog was found in *M. domestica* (gray short-tailed opossum), the most distant node is Mammalia, and the phylostratigraphic index of the gene is equal to 7. Since the PAI in example (*a*) is larger than the PAI in example (*b*), we

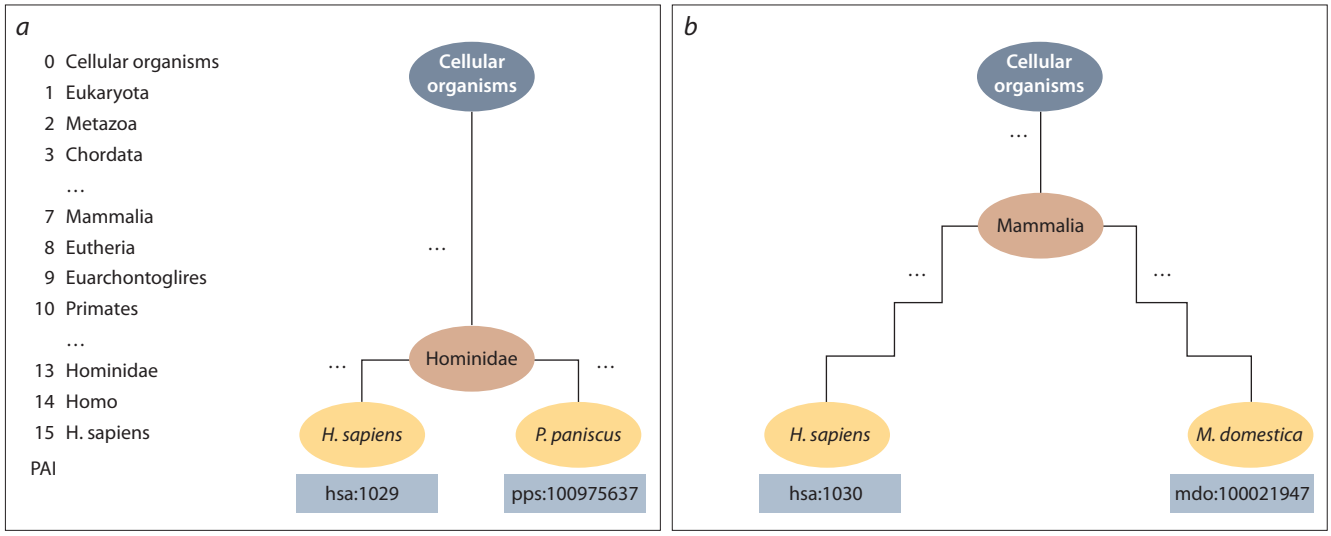


Fig. 1. The example of PAI determination for two *Homo sapiens* (human) genes.

a – the example of evolutionary young gene is hsa:1029, most distant from the studied organism in which the orthologous gene was found *Pan paniscus* (bonobo); *b* – the example of evolutionary older gene is hsa:1030, most distant from the studied organism in which the orthologous gene was found is *Monodelphis domestica* (gray short-tailed opossum). We can conclude that the gene on example (*a*) is evolutionary younger than the gene on example (*b*). The scale on the left shows the PAI index, which corresponds to the depth of a node in the phylogenetic tree (see Table 1 for details).

can conclude that the gene in example (*a*) is evolutionary younger than the gene in example (*b*).

An important characteristic for the phylostratigraphic analysis is the list of taxonomic units describing the stages of divergence on the phylogenetic tree. Table 1 shows the complete list of taxa used in the analysis to determine the phylostratigraphic age index of *H. sapiens* genes, as well as the approximate evolutionary age of these taxa in millions of years from our time. It should be noted that the discussions on this topic are ongoing and there are different data of the age; the values in the table reflect approximate estimates.

Table 1. The list of taxons used in phylostratigraphic analysis of *H. sapiens* genes

PAI	Taxon	Age (Mya)
0	Cellular organism (tree root)	4100 (Bell et al., 2015)
1	Eukaryota	1850 (Leander, 2020)
2	Metazoa	665 (Maloof et al., 2010a)
3	Chordata	541 (Maloof et al., 2010b)
4	Craniata	535 (Maloof et al., 2010b)
5	Vertebrata	525 (Shu et al., 1999)
6	Euteleostomi	420 (Diogo, 2007)
7	Mammalia	225 (Datta, 2005)
8	Eutheria	160 (Luo et al., 2011)
9	Euarchontoglires	65 (Kumar et al., 2013)
10	Primates	55 (Chatterjee et al., 2009)
11	Haplorrhini	50 (Dunn et al., 2016)
12	Catarrhini	44 (Harrison, 2013)
13	Hominidae	17 (Hey, 2005)
14	Homo	2.8 (Schrenk et al., 2014)
15	Homo sapiens	0.35 (Scerri et al., 2018)

Orthoscape also allows to estimate *divergence index* (DI). DI shows the type of selection to which the gene is influenced. This index is calculating based on the *dN/dS* ratio, where *dN* reflects the rate of nonsynonymous substitutions between the sequences of analyzed gene and its orthologous gene (the substitutions changing the amino acid encoded) and *dS* – reflects the rate of synonymous substitutions (the substitutions without changing the amino acid encoded). The index value from 0 to 1 indicates that the gene is under stabilizing selection, value is equal to 1 indicates neutral evolution, and greater than 1 indicates a driving selection. The analysis of this index makes sense only when comparing closely related organisms, because it can't take into account multiple substitutions in the same position, which will be inevitably accumulated when comparing the evolutionary distant organisms. Calculation of *dN/dS* takes place in two phases:

1. Alignment of original sequences. To align the sequences, the Needleman–Wunsch algorithm is used. The task is to align aminoacid sequences and nucleotide triplets correspond to them and remove the gaps from the result.
2. Aligned sequences are given as input to the PAML (phylogenetic analysis by maximum likelihood) (Yang, 2007) software. Various methods are used to calculate *dN/dS*. They take into account different positions of triplets, their frequency of occurrence, and other factors. There are Nei–Gojobori (Nei, Gojobori, 1986), Yang & Nielsen (Yang, Nielsen, 2000), LWL85 (Li, 1985), LWLm (Li, 1993), LPB93 (Pamilo, Bianchi, 1993) methods implemented in PAML. To count *DI*, Orthoscape uses LPB93

method. The formula to count *DI* is based on *dN/dS* for every gene-ortholog pair

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

where $dnds_i$ – *dN/dS* value for gene and *i*-th ortholog; *n* – number or orthologous genes.

Results and discussion

The analysis of evolutionary characteristics of gene networks

80 networks from KEGG Pathway, Human Diseases were analyzed using Orthoscape software. First of all, PAI and DI values for genes in network have been calculated. Based on these data, PAI values for every network have been calculated (Table 2) as an average PAI value of genes involved in network. Finally, PAI of the category of diseases has been calculated as an average of PAI value of networks from this category. The same metrics have been calculated for DI.

There are big PAI variations are observed among the analyzed 80 networks: from 0.44 (i. e., in "Nicotine addiction" gene network, the most of the genes are evolutionary ancient) to 6.38 (i. e., in "Asthma" gene network, the most of the genes are evolutionary young). The DI variation is usually within $0 < DI < 1$ interval, i. e., within stabilizing selection interval; however, the level of variability of genes involved in different networks also varies greatly, from 0.16 to 0.64. The diseases "Asthma" and "Nicotine addiction" are the most exuding according to the PAI and DI indices. In the "Asthma" network, evolutionary young and variable genes prevail, and in the "Nicotine addiction" network, evolutionary ancient and conservative genes prevail. Fig. 2 shows the result of PAI analysis for the "Asthma" and "Nicotine addiction" networks, and Fig. 3 shows the DI results of the same networks.

The most part of genes in the "Asthma" network (Fig. 2, *a*) are evolutionary young (colored in green and yellow), with origin on Vertebrata level. On the contrary, in the "Nicotine addiction" network (Fig. 2, *b*) all genes have been identified as evolutionary ancient, with origin from the cellular life form (Cellular organisms) to multicellular (Metazoa) stages.

Analysis of the DI indicates that almost all the genes involved in the "Asthma" network (see Fig. 3, *a*) are more evolutionary variable than the genes involved in the "Nicotine addiction" network (see Fig. 3, *b*), whose genes are very conservative.

Let's take a look at the estimations of PAI values for 11 disease categories (see Table 2). The most segregated networks are from 4 categories. High PAI and DI values is characteristic of Immune diseases (8 networks) and Infectious diseases: Parasitic (6 networks). Low PAI and DI value is characteristic of cancers: specific types (15 networks) and substance dependence (5 networks).

Genes from the categories above, as well as the complete set of 1436 genes, were divided into two groups: 1) a group of evolutionary ancient genes with PAI < 5 (the age of the genes corresponds to the period of evolution from Cellular organisms to Chordata); 2) a group of evolutionary young genes with PAI ≥ 5 (the age of the genes corresponds to the period of evolution from the Craniata to modern humans).

Contingency tables were created and Fisher's exact test was used to assess whether the difference in the partitioning of genes into groups in the category from the partitioning in the full list of genes was significant (Table 3).

The average PAI of all 1436 genes studied was equal to 2.49. The results from Table 3 show that gene networks from category Immune diseases have not only the highest value of the PAI (5.21), but also a significantly different distribution of the proportion of evolutionary young and ancient genes in comparison with such proportion among all genes analyzed (the last row of the Table 3).

The part of evolutionary young genes in Immune diseases category is 65 %. The most part of genes origin was at vertebrata stages (Vertebrata and Euteleostomi taxa), that corresponds to modern data about the development of specific immunity: it exists in cartilaginous fish (sharks and rays) and, therefore, appeared at least 400–500 million years ago. These fishes have genes related to the genes of the *Ig* variable region (*IgV*), or T-cell receptor (*TcR*) genes. At the same time, even more primitive vertebrates, the roundworms (hagfish and lampreys), do not have an acquired immunity system; they have neither *IgV* nor *TkR* genes (Galaktionov, 2015). The analysis also revealed a small fraction of evolutionary ancient genes in the Immune diseases category. This is consistent with the knowledge that some functions of the immune system originated as early as in unicellular organisms, such as the ability to phagocytose; cells with the T-lymphocyte marker first discovered in ringworms and the histocompatibility system – in sponges (Khaitov, 2009). On the contrary, the highest proportion of evolutionary ancient genes is characteristic of the "Substance dependence" diseases category, which includes genes responsible for addiction to chemicals (88 %). Most of the genes considered are involved in nervous system function, including neurotransmitter function.

The infectious diseases parasitic category, which includes genes associated with infectious diseases caused by parasites (53 % of the evolutionary young genes), has a significant difference in the proportion of evolutionary ancient and evolutionary young genes from that among all the genes analyzed. In the case of the infectious diseases parasitic category, the high proportion of evolutionary young genes can be directly related to the high proportion of evolutionary young genes and the high evolutionary variability of genes found in the Immune diseases category. It is infectious diseases that are one of the most important drivers of immune system evolution. At the same time, infectious diseases of different nature and the immune system co-evolve in the process of forming mechanisms to fight each other (Sasaki et al., 2000; Khakoo, 2004; Zheleznikova, 2014).

It should be noted, that there is a significant excess of the proportion of ancient genes over young genes compared to their distribution (ancient/young) in the full sample of

Table 2. Average values of PAI and DI indices for genes involved in gene networks from the KEGG Pathway, Human Diseases database

No.	Network*	PAI	DI	No.	Network*	PAI	DI
1	Asthma[1]	6.38	0.64	41	Epithelial cell signaling in Helicobacter pylori infection[3]	2.27	0.20
2	Graft-versus-host disease[1]	6.29	0.54	42	Dilated cardiomyopathy (DCM)[8]	2.19	0.26
3	Autoimmune thyroid disease[1]	5.61	0.49	43	Pathogenic Escherichia coli infection[3]	2.19	0.27
4	Allograft rejection[1]	5.53	0.46	44	Human papillomavirus infection[5]	2.18	0.29
5	Malaria[2]	5.49	0.46	45	Human T-cell leukemia virus 1 infection[5]	2.16	0.29
6	African trypanosomiasis[2]	5.12	0.47	46	Hypertrophic cardiomyopathy (HCM)[8]	2.14	0.30
7	Inflammatory bowel disease (IBD)[1]	4.95	0.35	47	Bladder cancer[7]	2.13	0.26
8	Rheumatoid arthritis[1]	4.70	0.40	48	Pancreatic cancer[7]	2.10	0.20
9	Staphylococcus aureus infection[3]	4.41	0.53	49	Proteoglycans in cancer[4]	2.06	0.25
10	Type I diabetes mellitus[9]	4.40	0.42	50	Prion diseases[10]	2.05	0.29
11	Primary immunodeficiency[1]	4.24	0.39	51	Viral carcinogenesis[4]	1.94	0.24
12	Systemic lupus erythematosus[1]	3.97	0.42	52	Non-small cell lung cancer[7]	1.93	0.25
13	Tuberculosis[3]	3.96	0.34	53	Pathways in cancer[4]	1.86	0.24
14	Pertussis[3]	3.87	0.37	54	Small cell lung cancer[7]	1.84	0.26
15	Legionellosis[3]	3.84	0.34	55	Chronic myeloid leukemia[7]	1.82	0.21
16	Salmonella infection[3]	3.77	0.26	56	Shigellosis[3]	1.81	0.27
17	Viral myocarditis[8]	3.66	0.35	57	Parkinson disease[10]	1.76	0.20
18	Leishmaniasis[2]	3.60	0.33	58	Glioma[7]	1.74	0.25
19	Chagas disease (American trypanosomiasis)[2]	3.58	0.29	59	Endometrial cancer[7]	1.72	0.24
20	Chemical carcinogenesis[4]	3.56	0.56	60	Melanoma[7]	1.71	0.24
21	Measles[5]	3.53	0.30	61	Colorectal cancer[7]	1.65	0.21
22	Toxoplasmosis[2]	3.42	0.28	62	Insulin resistance[9]	1.64	0.25
23	Influenza A[5]	3.35	0.35	63	Endocrine resistance[6]	1.62	0.22
24	Amoebiasis[2]	3.26	0.36	64	Central carbon metabolism in cancer[4]	1.61	0.26
25	Herpes simplex virus 1 infection[5]	3.26	0.34	65	Thyroid cancer[7]	1.57	0.24
26	Kaposi sarcoma-associated herpesvirus infection[5]	3.13	0.29	66	Breast cancer[7]	1.55	0.30
27	Antifolate resistance[6]	3.00	0.40	67	Alcoholism[11]	1.48	0.17
28	Hepatitis C[5]	2.92	0.30	68	Cocaine addiction[11]	1.42	0.14
29	Platinum drug resistance[6]	2.80	0.29	69	Bacterial invasion of epithelial cells[3]	1.42	0.15
30	Acute myeloid leukemia[7]	2.80	0.30	70	Huntington disease[10]	1.42	0.20
31	Arrhythmogenic right ventricular cardiomyopathy[8]	2.79	0.25	71	Renal cell carcinoma[7]	1.41	0.16
32	Amyotrophic lateral sclerosis (ALS)[10]	2.75	0.27	72	Vibrio cholerae infection[3]	1.35	0.18
33	Epstein-Barr virus infection[5]	2.54	0.35	73	Prostate cancer[7]	1.33	0.29
34	Transcriptional misregulation in cancer[4]	2.53	0.29	74	Type II diabetes mellitus[9]	1.30	0.29
35	AGE-RAGE signaling pathway in diabetic complications[9]	2.52	0.28	75	Basal cell carcinoma[7]	1.20	0.23
36	Hepatitis B[5]	2.50	0.27	76	Morphine addiction[11]	1.06	0.16
37	Non-alcoholic fatty liver disease[9]	2.44	0.27	77	Maturity onset diabetes of the young[9]	1.04	0.19
38	EGFR tyrosine kinase inhibitor resistance[6]	2.43	0.20	78	Choline metabolism in cancer[4]	1.03	0.19
39	Alzheimer disease[10]	2.42	0.26	79	Amphetamine addiction[11]	0.75	0.18
40	Fluid shear stress and atherosclerosis[8]	2.40	0.26	80	Nicotine addiction[11]	0.44	0.16

* Category 1 – immune diseases; 2 – infectious diseases parasitic; 3 – infectious diseases bacterial; 4 – cancers overview; 5 – infectious diseases viral; 6 – drug resistance antineoplastic; 7 – cancers specific types; 8 – cardiovascular diseases; 9 – endocrine and metabolic diseases; 10 – neurodegenerative diseases; 11 – substance dependence.

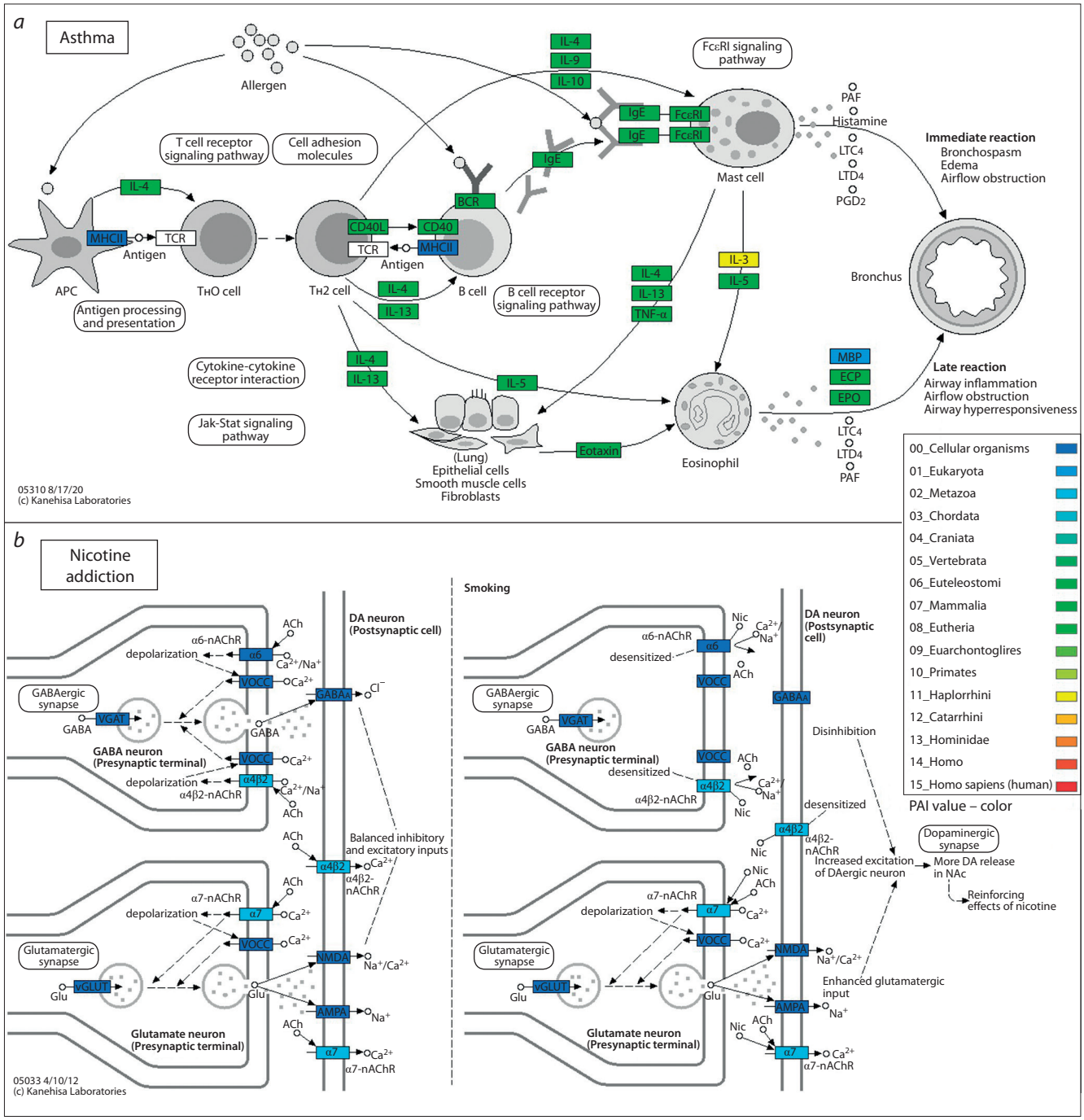


Fig. 2. Gene networks schemes of diseases "Asthma" (*a*) and "Nicotine addiction" (*b*) taken from the KEGG Pathway, Human Disease database with PAI values.

Gene coding the proteins in these networks are shown as rectangles with gene name, the color reflects the gene age. The genes colored in blue and cyan correspond to the most evolutionary ancient taxa, green and yellow correspond to evolutionary younger in compare with taxa colored in blue.

genes analyzed in cancers specific types category, which includes genes associated with carcinogenesis. This result is consistent with the current ideas that gene networks involved in cancer development processes were formed during the stages of multicellular organisms origin (Domazet-Lošo, Tautz, 2010).

Let us consider two categories of diseases in more details: (1) immune diseases with the highest proportion of evolutionary young genes and (2) substance dependence with the highest proportion of evolutionary ancient genes (Fig. 4).

Figure 4 shows the PAI distributions for 13 networks (8 immune diseases networks and 5 substance dependence networks) as "violin plot" graphs. The lower and upper points of each graph show the minimum and maximum PAI values, the orange star shows the median PAI values, and the width of the graph for each position on the ordinate

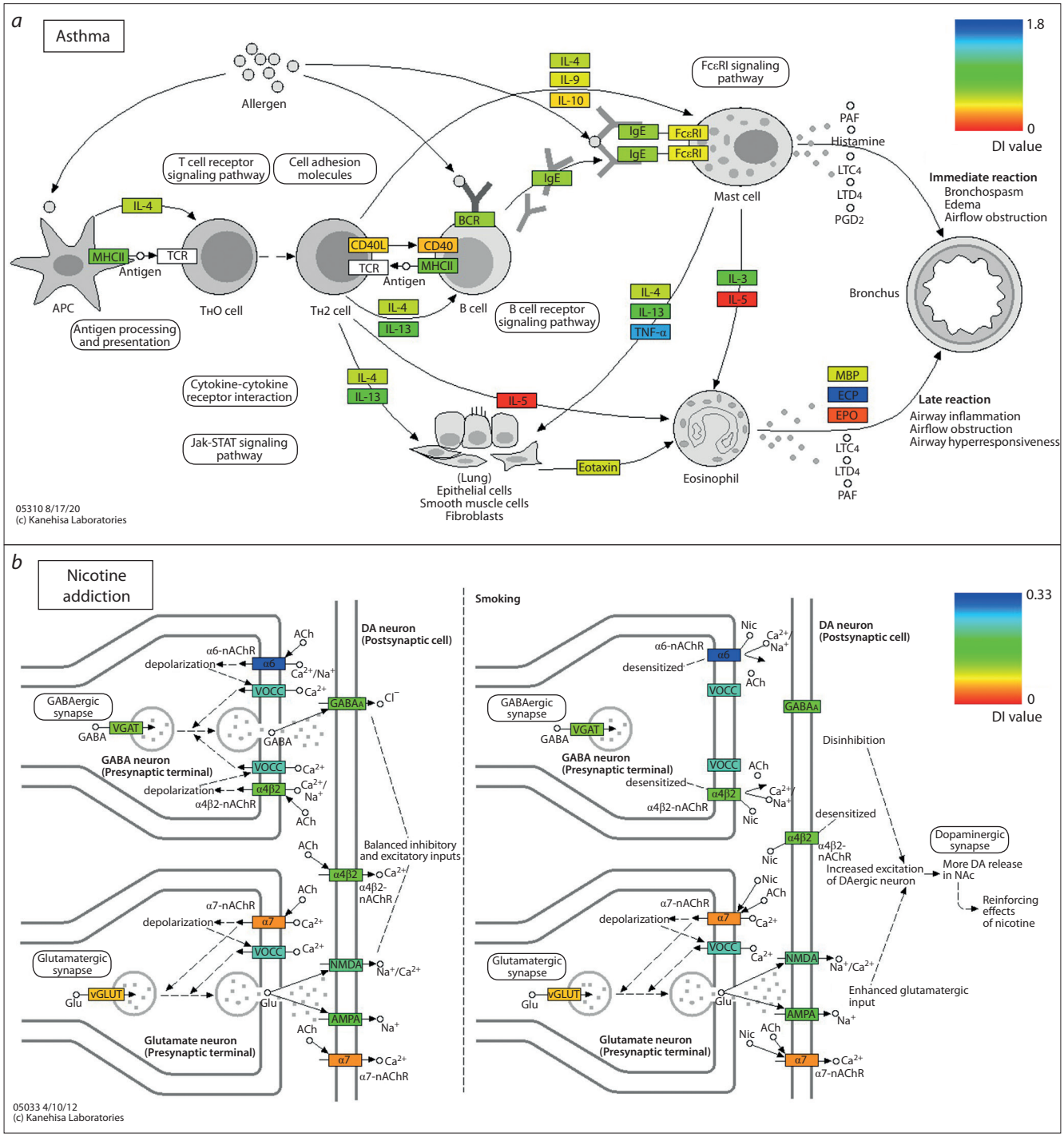


Fig. 3. Gene networks schemes of diseases "Asthma" (*a*) and "Nicotine addiction" (*b*) taken from KEGG Pathway, Human Disease database with DI values.

Gene coding the proteins in these networks are shown as rectangles with gene name, the color reflects the gene variability level. In the upper right part of the graph of each network placed the color scheme for DI. The scale for each network is individual, and even the most variable genes involved in the "Nicotine addiction" network have minimal variability compared to genes involved in the "Asthma" network.

axis (i. e. for each PAI) shows the proportion of genes with that particular PAI. The median for distributions of immune diseases varies in the range (5, 7) (from Vertebrata to Mammalia), and the distributions themselves have a character expressed in decreasing the number of genes with a corresponding PAI value as PAI decreases. The median for distributions of substance dependence varies in the range (0, 1) (Cellular organisms and Eukaryota), and the distributions themselves have a character expressed in increasing the number of genes with a corresponding PAI value as PAI decreases. These distributions are fundamentally different when comparing the proportion of evolutionary ancient and

Table 3. The results of the Fisher's exact test comparing the distribution of evolutionary ancient and evolutionary young genes among all genes described in the human disease gene networks from KEGG Pathway, Human Diseases, and among genes within the same category

KEGG Pathway, Human Diseases category	Genes		PAI	*p*-value
	evolutionary ancient	evolutionary young		
Immune diseases	56	106	5.21	8.84×10^{-15}
Infectious diseases parasitic	74	84	4.08	2.79×10^{-6}
Cancers specific types	187	54	1.77	4.41×10^{-4}
Substance dependence	69	9	1.03	1.75×10^{-5}
Total of 1436 genes	952	484	2.49	–

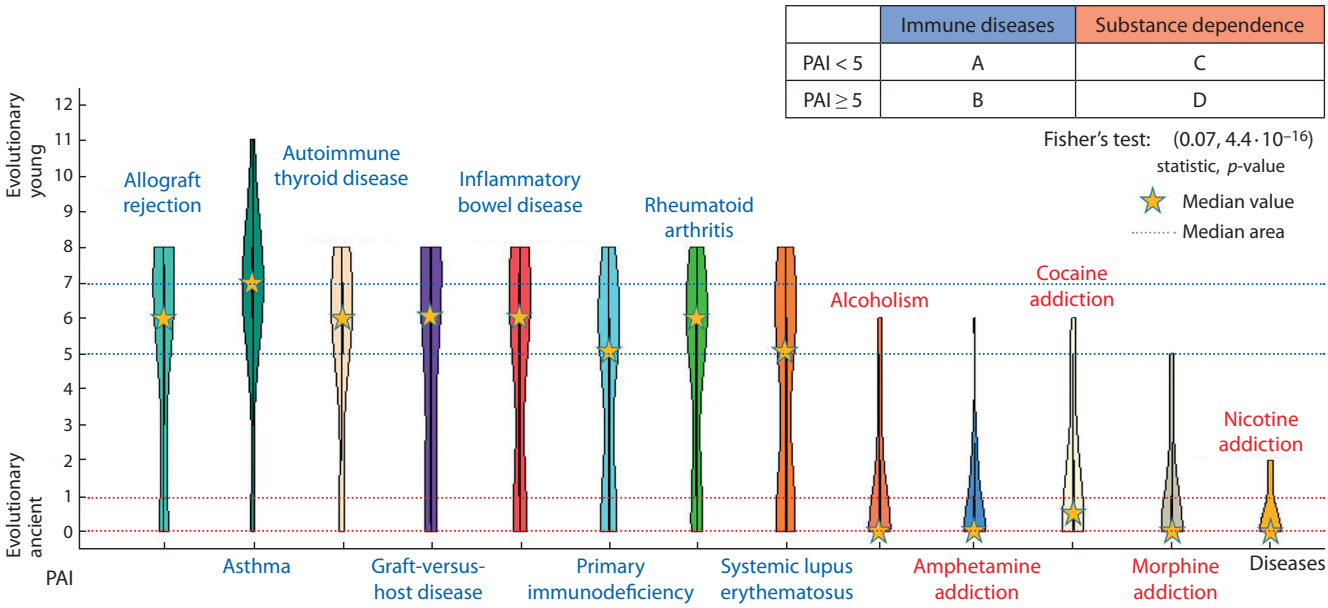


Fig. 4. Distribution of PAI among eight networks from category immune diseases (blue) and five networks from category substance dependence (marked in red).

Plots are visualized with R package vioplot, script is created by Orthoscape.

evolutionary young genes, as shown also by Fisher's exact test with p-value $= 4.4 \times 10^{-16}$.

Figure 5 shows the distribution of PAI among all the genes involved in the 80 gene networks considered from the KEGG Pathway, Human Diseases. This distribution has two peaks. The left peak includes genes formed early in evolution (from the emergence of cellular organization of life to chordates), and the right one includes genes formed at subsequent stages of evolution (vertebrates to placentals). There were more evolutionary ancient genes than evolutionary young ones.

Figure 6 shows the distribution of DI among all the genes involved in considered gene networks from the KEGG Pathway, Human Diseases. The DI analysis makes it possible to estimate what type of selection the genes are influenced. However, it only makes sense when the sequences of the analyzed genes are compared with the orthologous genes of evolutionary close organisms. To calculate *dN/dS*, human gene sequences were compared with the sequences of

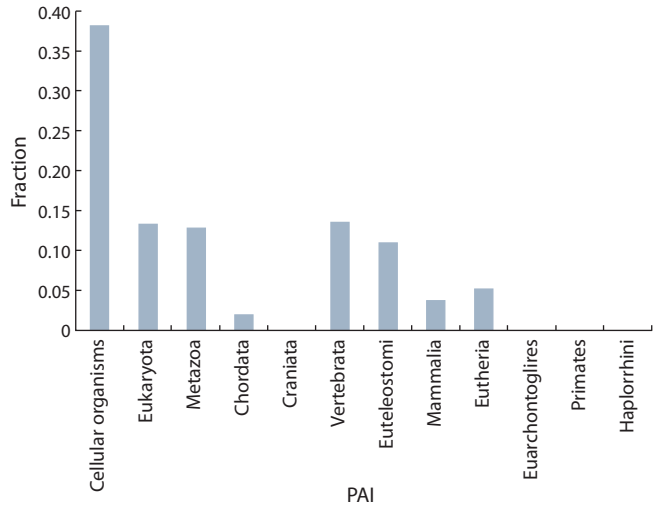


Fig. 5. Distribution of PAI among all genes involved in the considered gene networks from the KEGG Pathway, Human Diseases.

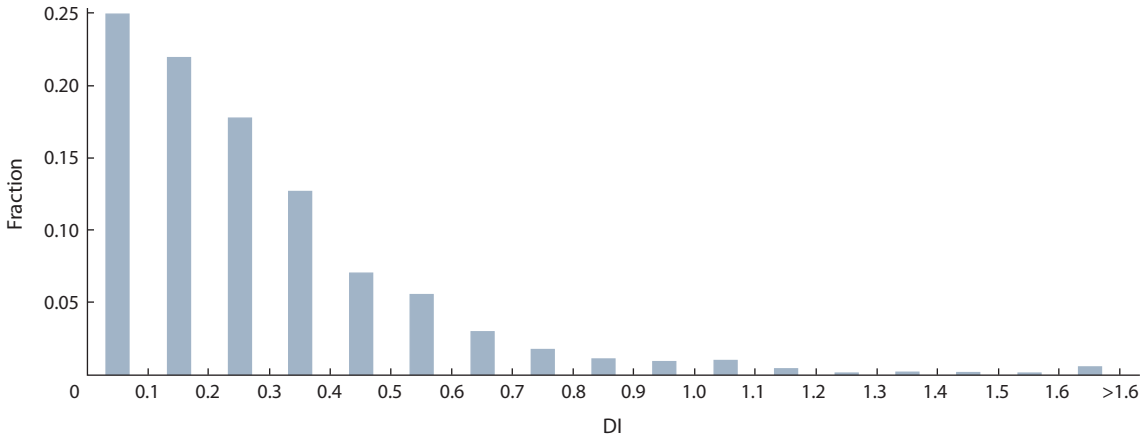


Fig. 6. Distribution of DI among all genes involved in the considered gene networks from the KEGG Pathway, Human Diseases.

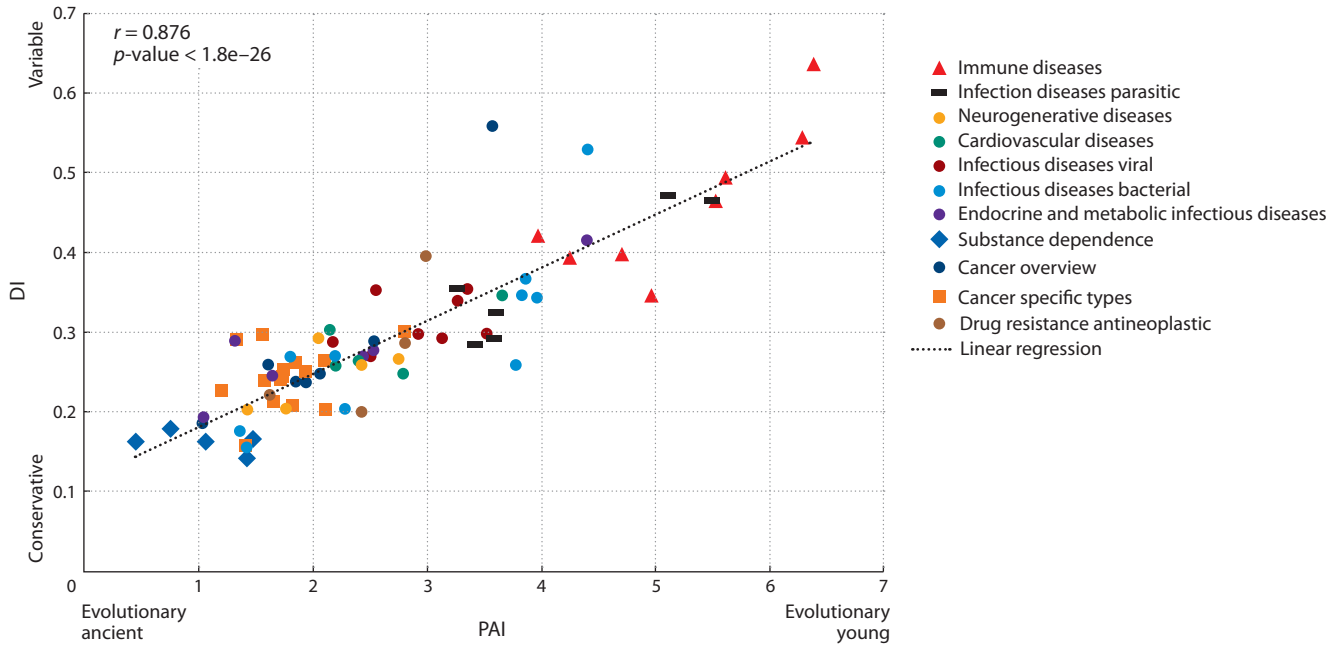


Fig. 7. Scatter plot for mean values of PAI and DI indices for 80 gene networks from KEGG Pathway, Human Diseases database.
The figures of different forms and sizes show the different diseases categories.

orthologous genes of other hominids; if there were several orthologs, the average value of *dN/dS* was used as the DI. Only 38 of the 1,436 genes had DI values > 1 (nine of them fall into one category, immune diseases). The vast majority of genes included in studied gene networks evolved in the mode of stabilizing selection (DI < 1).

It was interesting to study the relationship between PAI and DI for the 80 gene networks we studied. Figure 7 presents the results of this analysis in a single graph, taking into account the categorization of diseases. Figures of different colors and sizes indicate different disease categories.

The analysis showed that there is a large significant correlation between PAI and DI with the value of the correlation coefficient ($r = 0.876$, p-value $< 1.8 \times 10^{-26}$). It means there is a relationship between the average evolutionary age of genes in gene networks and the level of their genetic variability: the less the evolutionary age of genes, the greater the level of their genetic variability. This agrees well with the fact that evolutionary ancient genes are involved in key processes for organism functionality; they are a subject to many restrictions by other genes and molecular-genetic systems organization peculiarities, so they are not characterized by high variability. On the contrary, evolutionary young genes provide adaptation to modern life conditions and are characterized by higher variability.

Conclusion

Phylostratigraphic analysis is a modern methodology that allows genome-wide estimation of gene ages based on data on the similarity of genetic sequences and the origin of or-

ganisms. Together with information on what type of selection a gene is subject to as a unit of heredity, the results of the analysis allow us to estimate the role of certain genes in the evolution of the gene networks of an organism.

Analysis of gene networks from the KEGG Pathway, Human Diseases database shows several trends. The vast majority of the genes involved in the gene networks studied evolved in the mode of stabilizing selection (DI < 1). There is significant ($r = 0.876$, p-value $< 1.8 \times 10^{-26}$) correlation between the average evolutionary age of genes in gene networks and their level of genetic variability: the lower the evolutionary age of genes, the greater the genetic variability is. Some categories of gene networks are especially distinguished by the proportion of evolutionary young and evolutionary ancient genes. The highest proportion of evolutionary young genes (65 %) was found in gene networks from immune diseases category. The highest proportion of evolutionary ancient genes (88 %) was found in gene networks from substance dependence category.

It is also shown that gene networks responsible for the functioning of infectious diseases caused by parasites are significantly enriched with evolutionary young genes, and gene networks responsible for the development of specific cancer types are significantly enriched with evolutionary ancient genes. Such results indicate an active process of adaptation of the human immune system to emerging threats. In addition, the genes involved in chemical addictive diseases have a minimum number of substitutions, i. e., such genes are as conservative as possible. Separate work can be carried out in this direction, with expansion of the original networks thanks to the classifiers and databases currently available.

References

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;112:14518-14521. DOI 10.1073/pnas.1517557112.

Cerami E.G., Gross B.E., Demir E., Rodchenkov I., Babur Ö., Anwar N., Schultz N., Bader G.D., Sander C. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39: 685-690. DOI 10.1093/nar/gkq1039.

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.

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;25:200-207. DOI 10.1671/0272-4634(2005)025(0200:EMWTEU)2.0.CO;2.

Diogo R. The Origin of Higher Clades: Osteology, Myology, Phylogeny and Evolution of Bony Fishes and the Rise of Tetrapods. New York: CRC Press, 2007.

Domazet-Lošo T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 2007;23:533-539. DOI 10.1016/j.tig.2007.08.014.

Domazet-Lošo T., Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* 2010;8:66.

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.

Galaktionov V.G. Immunology: a Guide for University Students Studying in Track 510600 “Biology” and Biological Specialties. Moscow: Academia Publ., 2004. (in Russian)

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

Hey J. The ancestor’s tale A pilgrimage to the dawn of evolution. *J. Clin. Invest.* 2005;115:1680-1680.

Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45:D353-D361.

Khaitov R.M. Immunology: a Guide for Students of Medical Universities. Moscow, 2016. (in Russian)

Khakoo S.I. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*. 2004;305(5685):872-874.

Kolchanov N.A., Ignat’eva E.V., Podkolodnaya O.A., Likhoshvay 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 V., Hallström B.M., Janke A. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PLoS One.* 2013;8(4):e60019.

Leander B.S. Predatory protists. *Curr. Biol.* 2020;30:R510-R516.

Li W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 1993;36(1):96-99.

Li W.H., Wu C.I., Luo C.C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 1985;2(2):150-174.

Liebeskind B.J., McWhite C.D., Marcotte E.M. Towards consensus gene ages. *Genome Biol. Evol.* 2016;8(6):1812-1823.

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

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; 122:1731-1774.

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.

Montojo J., Zuberi K., Rodriguez H., Kazi F., Wrig G., Donaldson S.L., Morris Q., Bader G.D. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop. *Bioinformatics.* 2010;26:2927-2928.

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(S1):1-9.

Nei M., Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 1986;3:418-426.

Nersisyan L., Samsonyan R., Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. *F1000Res.* 2014;3:145.

Pamilo P., Bianchi N.O. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 1993;10(2): 271-281.

Sasaki K., Tsutsumi A., Wakamiya N. Mannose-binding lectin polymorphisms in patients with hepatitis C virus infection. *Scand. J. Gastroenterol.* 2000;35(9):960-965.

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 sub-

divided populations across Africa, and why does it matter? *Trends Ecol. Evol.* 2018;33(8):582-594.

Schrenk F., Kullmer O., Bromage T. The earliest putative homo fossils. In: Handbook of Paleoanthropology. Berlin; Heidelberg: Springer, 2014;1-19.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.

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.

Stepanov V.A. Evolution of genetic diversity and human diseases. *Russ. J. Genet*. 2016;52(7):746-756.

Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J., von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019; 47(D1):D607-D613.

Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591.

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

Zheleznikova G.F. Infection and immunity: strategies from both sides. *Med. Immunol.* 2014;8(5-6):597-614. DOI 10.15789/1563-0625-2006-5-6-597-614. (in Russian)

ORCID ID

Z.S. Mustafin orcid.org/0000-0003-2724-4497
S.A. Lashin orcid.org/0000-0003-3138-381X
Yu.G. Matushkin orcid.org/0000-0001-7754-8611

Acknowledgements. The work was supported by the Russian Foundation for Basic Research No. 20-04-00885 A and the budget project No. 0259-2021-0009.

Funding transparency. The authors do not hold financial interests in the presented materials or methods.

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

Received January 14, 2021. Revised January 20, 2021. Accepted January 20, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):57-63
DOI 10.18699/VJ21.007

Original Russian text www.bionet.nsc.ru/vogis/

Crop pangenomes

A.Yu. Pronozin[1] ✉, M.K. Bragina[1, 2], E.A. Salina[1, 2]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pronozinartem95@gmail.com

Abstract. Progress in genome sequencing, assembly and analysis allows for a deeper study of agricultural plants' chromosome structures, gene identification and annotation. The published genomes of agricultural plants proved to be a valuable tool for studing gene functions and for marker-assisted and genomic selection. However, large structural genome changes, including gene copy number variations (CNVs) and gene presence/absence variations (PAVs), prevail in crops. These genomic variations play an important role in the functional set of genes and the gene composition in individuals of the same species and provide the genetic determination of the agronomically important crops properties. A high degree of genomic variation observed indicates that single reference genomes do not represent the diversity within a species, leading to the pangenome concept. The pangenome represents information about all genes in a taxon: those that are common to all taxon members and those that are variable and are partially or completely specific for particular individuals. Pangenome sequencing and analysis technologies provide a large-scale study of genomic variation and resources for an evolutionary research, functional genomics and crop breeding. This review provides an analysis of agricultural plants' pangenome studies. Pangenome structural features, methods and programs for bioinformatic analysis of pangenomic data are described.
Key words: agricultural plants; genomes; pangenomes; genes; evolution; bioinformatics analysis; computational pipelines.

For citation: Pronozin A.Yu., Bragina M.K., Salina E.A. Crop pangenomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):57-63. DOI 10.18699/VJ21.007

Пангеномы сельскохозяйственных растений

А.Ю. Пронозин[1] ✉, М.К. Брагина[1, 2], Е.А. Салина[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ pronozinartem95@gmail.com

Аннотация. Секвенирование генома организма – важный этап в его генетических исследованиях. Расшифровка геномной последовательности открывает широкие возможности для изучения строения структуры хромосом, распределения повторенных и кодирующих последовательностей, идентификации и аннотации генов. При исследовании сельскохозяйственных растений это позволяет анализировать функции генов, разрабатывать маркеры для поиска ассоциаций с фенотипическими признаками. При решении этих задач геном вида часто представлен последовательностью одного организма (так называемым референсным геномом). В последнее время, однако, появляется много свидетельств в пользу того, что большие структурные изменения генома, включая вариации числа копий генов и вариации наличия/отсутствия генов, преобладают в сельскохозяйственных культурах, играют ключевую роль в генетическом определении агрономически важных признаков и приводят к значительным вариациям функционального набора генов и генного состава у представителей одного вида. Такие структурные вариации не могут быть представлены на основе одной лишь референсной последовательности и описываются исходя из концепции пангенома. Пангеном – это информация о полном наборе генов таксона, среди которых можно выделить набор универсальных генов, общих для всех представителей таксона, и вариабельных генов, которые являются частично или полностью специфичными для его представителей. Анализ пангеномов дает более точное понимание генетического разнообразия генофонда. Технологии секвенирования и анализа пангеномов позволяют обеспечить возможность масштабного изучения геномных вариаций, доступ к более широкому спектру геномных данных в селекционных программах и помогут ускорить селекцию культурных растений для создания сортов со стабильно высокой урожайностью и устойчивостью к стрессам. В работе представлен краткий обзор исследования пангеномов сельскохозяйственных растений, описаны их структурные особенности, методы и программы биоинформатического анализа пангеномных данных.
Ключевые слова: сельскохозяйственные растения; геномы; пангеномы; гены; эволюция; биоинформатический анализ; вычислительные конвейеры.

Introduction

The genome sequence is the basis for a chromosome structure studying, a distribution of repetitive and coding sequences, and genes identification and annotation (Bragina et al., 2019). The different species genomes information allows a comparative phylogenetic analysis to study relationships among species, their origins, and evolutionary features (Marchant et al., 2016; Wendel et al., 2016). In agricultural plants, all these allows to assess the impact of a genetic variability on a gene function, to identify the genes responsible for the most valuable traits in crops (Schnable et al., 2009; Wing et al., 2018).

A single organism chromosome sequences serve as the basis ("reference" genome) for studying other genomes of the same species. The number of sequenced, assembled and annotated plant reference genomes increases every year (Bragina et al., 2019). The Ensembl Plants database version 48 (September 2020) contains 93 assembled and annotated plant genomes (Howe et al., 2020). Based on the reference genome sequencing and the sequencing of the same species representatives genomes (usually based on short-reading technology), genetic variability analysis, the study of the genome single-nucleotide polymorphisms (SNPs) and large structural variants (SVs) are performed. The large structural variants are the most difficult to identify using a short-read sequencing, but due to the third-generation sequencing technologies (Li et al., 2018), the SVs identification is becoming more accessible and reliable. There is a growing evidence that structural variations, including copy number variations (CNVs) and presence/absence variations (PAVs), are prevalent in crops and lead to significant variations in gene content between individuals of the same species (Springer et al., 2009; Hirsch et al., 2014; Li et al., 2014; Lu et al., 2015; Zhao Q. et al., 2018).

Genomes and pangenome

For a more efficient analysis and description of the genetic diversity, the concept of "pangenome" was proposed (Tettelin et al., 2005). The pangenome represents the information about the complete set of genes in a biological cluster (taxon), such as species, among which one can distinguish a set of universal (core) genes that are common to all organisms, and a set of unique (variable) genes that are partially shared or individually specific (Tettelin et al., 2005). Until recently pangenome studies have been focused on finding genes presence or absence in organisms to determine the universal or unique set of genes.

The concept of the "pangenome" was proposed in (Tettelin et al., 2005) for the *Streptococcus agalactiae* bacterial species. To date, there are several definitions of this term, wich are based on two main concepts: a function based and a structure based (Tranchant-Dubreuil et al., 2018). The structural concept considers the pangenome as complete set of taxon genomic sequences. Within this concept, taxon members genomic sequences (of the same species or genus) are compared with each other and on this basis their common unique (non-redundant) set of DNA fragments of the same length (100 bp or more, depending on the species) is determined. These sequences describe the structure of the pangenome (Snipen et al., 2009; Alcaraz et al., 2010).

The second pangenome concept is based on its functional representation. In this case, the pangenome can be described as a set of all genes for particular taxon representatives (Plissonneau et al., 2018). However, for a large number of related organisms, such a set is degenerate, because they contain a large number of genes with a high level of similarity in primary structure, and, consequently, in function. Pangenome redundancy can be eliminated by combining similar gene sequences into functional families (Sun et al., 2016). In this case, the representative genes of the same functional family in different organisms are considered as one sequence in terms of function.

The set of organisms in pangenome analysis usually limited to a single species. However, some authors use a broader interpretation of the pangenome. For example, V.V. Tetz (2003) considers the pangenome as a complete genes set of all living organisms, viruses and mobile elements.

Pangenome structural features

Pangenome genes can be divided into two groups according to their occurence in different organisms (Golicz et al., 2016). The first group includes genes that are found in all members of the taxon. This group of genes is called the universal set or core gene set. The second group of genes includes genes that occur in a part of the taxon. This genes group is called indispensable, accessory or variable genes. Among the second genes group, the unique genes that are present only in the single individual are of particular interest. Universal and variable genes represent the functional core and the diversity of species members, respectively.

From an evolutionary perspective, universal genes are mostly responsible for vital functions and they tend to be conserved within a species. In contrast, variable genes and their specific part, unique genes, contribute to the diversity of the species, enabling them to adapt to different environmental conditions. The proportion of unique genes in the studied crops pangenomes ranges from 8 to 61 % (Tao et al., 2019). However, the resulting size of the unique genome is likely to be underestimated due to the inability of current strategies and technologies to detect all functional changes in genes.

Based on the sequence of one genome it is impossible to determine, which genes are common to all species members. However, each new sequence can be assigned to a universal or variable part of the pangenome. The more taxon genomes are sequenced, the more unique genes are found. This results to a pangenome size increasing with an increase in the genomes number. However, for a universal genes set, increasing genomes number leads to the opposite result: some universal genes may be absent in other species members. As a result, the pangenome size – the set of all the different species genes – increases, while the estimated size of the universal genes set tends to decrease (Golicz et al., 2016; Wang et al., 2018). This relation is shown schematically in Fig. 1. Each point on the graph corresponds to an estimate of the genes number in the pangenome for a set of k genomes (taken randomly from the full sample of N genomes under study). With k increasing, the estimate of the total pangenome genes number increases (red

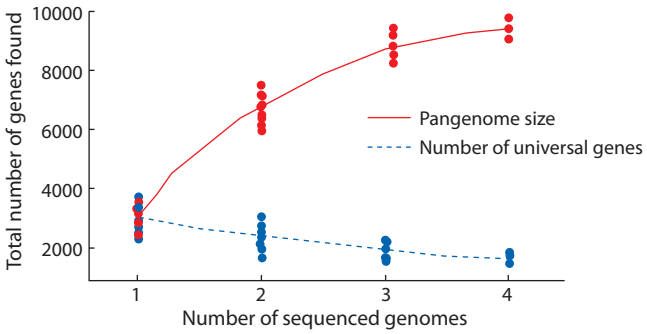


Fig. 1. The pangenome size and the universal gene number dependence on the number of sequenced genomes.

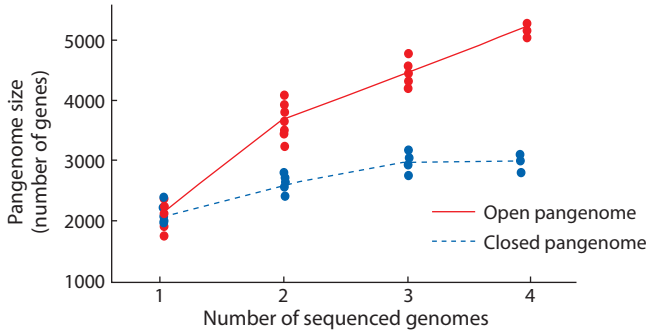


Fig. 2. The dependence of the genes number in the pangenome (*Y*-axis) from the number of sequenced taxon representatives (*X*-axis) for two pangenome types: open and closed.

For open genomes, number of genes raise monotonically, for closed – reaches a plateau.

line), and the unique genes number decreases (blue dashed line). Examples of dependencies for real pangenomes can be found at https://pangp.zhaopage.com. Thereby, the organisms sample sizes significantly affects the pangenome size estimation and the universal gene proportion in it.

In addition to the sequenced genomes number, the pangenome unique gene size and proportion is also influenced by many factors. The choice of a sample for analysis is one of them: (1) wild and cultivated species together will give a larger pangenome with a higher percentage of unique genes than only cultivated plants (Montenegro et al., 2017; Zhao Q. et al., 2018); (2) the ploidy level, mode of reproduction, bottlenecks during domestications, etc. A plant species with higher levels of ploidy and outbreeding and reduced diversity because of domestication tend to have a higher percentage of unique genes (Tao et al., 2019).

It can be assumed that the addition of an unlimited number of new genomes to the pangenome could lead to its unlimited growth. However, the gene diversity studies in crop species have shown the number of unique genes decrease as the number of sequenced samples increases. This suggests that, given a certain number of taxon representatives, the inclusion of additional genomes in the pangenome will no lead to a further increase in its genes number. Such pangenomes are called "closed". The "closed" pangenome was found in tomato (Gao et al., 2019), corn (Hirsch et al., 2014), rice (Wang et al., 2018), soybeans (Li et al., 2014), sunflower (Hübner et al., 2019), *Brachypodium distachyon* (Gordon et al., 2017), *Brassica napus* (Hurgobin et al., 2018) and *Brassica oleracea* (Golicz et al., 2016).

However, there are also "open" pangenomes, in which the total genes number grows with each new sample added. Open pangenomes are specific for microorganisms, for example for the wheat leaves septoria fungal pathogen *Zymoseptoria tritici* (Plissonneau et al., 2018). The bacterium *Paenibacillus polymyxa* pangenome also belongs to the open type (Zhou et al., 2020).

If organisms from the population are randomly selected, the pangenom type can be estimated by plotting the number of found genes in each new genomic sequence (Fig. 2). The pangenome genes number reaching a plateau after analysis of certain genomic sequences number characterizes "closed" pangenomes (see Fig. 2, blue dashed line). The "open" pangenomes are characterized by a constant increase in size when new genomes are added (see Fig. 2, red line).

The comparison of the pangenome size and the universal and variable pangenome parts for some plant species is shown in (Supplimentary 1)[1]. The data obtained demonstrates the number of samples for pangenome analysis varies from three (*B. rapa*) to three thousand (*Oryza sativa*). The genes number in pangenomes varies from 35 thousand in diploid rice to 128 thousand in hexaploid bread wheat. The proportion of universal genes ranged from 41 % in *Medicago truncatula* to 84 % in *B. rapa*.

Pangenome functional features

Researches show that universal genes are responsible for fundamental cellular processes, while variable genes are associated primarily with functions that can give an advantage in different environmental conditions. Thus, *Brachypodium distachyon* pangenome analysis demonstrated universal gene set annotations are enriched with terms such as "glycolysis", "steroid", "glycosylation", "co-enzyme" (Gordon et al., 2017). Variable genes sets annotations were most of all enriched with terms "protective function", "development". In the same work, it was shown the nonsynonymous/synonymous substitution rate ratio in variable genes are higher than in universal genes. In addition, the universal genes orthologs in rice and sorghum were found to be more conservative than orthologs of the variable genes set. Universal genes expression level is generally higher than variable genes (Gordon et al., 2017). Similar results were obtained in the soybeans (Li et al., 2014; Liu et al., 2020), cabbage (Golicz et al., 2016), and wheat (Montenegro et al., 2017) pangenomes analysis.

The analysis of several agricultural plant pangenomes showed (Tao et al., 2019) that (1) the variable genes sequences are more mutable than universal genes; (2) the nonsynonymous substitution rate ratio is higher in variable genes; (3) variable genes are characterized by a wide function diversity; (4) the variable and universal genes functional characteristics are different, the variable genes are more related to the response to environmental factors, receptor activity and

[1] Supplementary materials 1–3 are available in the online version of the paper: http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx2.pdf

signal transduction, the universal genes are more related to basic cellular functions. Thus, the universal genes represent the conservative core of the pangenome (and species, respectively), while the variable genes represent its mutable part (both in terms of function and in terms of primary structure and expression patterns).

Pangenomes and pantranscriptomes

The transcriptome analysis is another gene set analysing method in several members of a taxon. The transcript nucleotide sequences (mainly mRNA), their expression levels estimation and the isoforms presence can be obtained by high-throughput sequencing (RNA-seq), which is significantly cheaper than the genome sequencing. Transcriptomic data allows estimating genes presence in the genome only if they are expressed in a plant tissue or organ. Thus, a set of transcripts cannot represent the full genome gene composition, but it is possible to obtain an approximate estimation (especially if a transcripts set from different tissues at different stages of development is analyzed). In this case, the transcriptome assembly requires significantly less computational resources, and the current methods allow obtaining it with high quality.

A study of the 503 inbred maize lines pantranscriptome revealed genetic diversity in protein-coding genes: more than 1.5 million single-nucleotide variations were found, and mutations associated with plant development traits (timing of several growth phases) were identified (Hirsch et al., 2014).

M. Jin et al. (2016) also analysed the 368 inbred maize lines pantranscriptome. The analysis identified more than two thousand sequences that were not represented in the maize reference genome, including genes responsible for the biotic stress response. Variations that are associated with the gene expression level (eQTL) were analysed. The analysis' results were projected to metabolic networks, which allowed to specify their functioning mechanisms.

Y. Ma et al. (2019) analysed 288 barley transcriptome sequencing experiments. Among the collected transcripts, about 30 % showed no similarity to the reference genome. The results of the pantranscriptome analysis revealed that pathogen resistance genes are more numerous in wild-grown barley, and such genes were subjected to greater selection pressure during domestication compared to genes in other species.

Pangenome construction methods

The pangenome bioinformatic analysis can be divided into the following main steps:

1. The pangenome sequence assembling.
2. The conserved and variable genomic sequences regions identification.
3. Genes identification/prediction and functional annotation.
4. Polymorphisms identification.
5. Storage, rapid access and visualization of the pangenomic data.

The following pangenome assembly strategies exist: assembly-alignment; metagenome approach; mapping-assembly (Golicz et al., 2016; Hurgobin, Edwards, 2017; Tranchant-Dubreuil et al., 2018).

Assembly-then-map. This strategy consists of each taxon separately *de novo* assembly, followed by sequences alignment with each other as well as with the reference genome to decrease redundancy and identify a set of common and variable sequence regions. Several software packages have been developed for the genome assembly: Velvet (Zerbino et al., 2008), SOAPdenovo (Xie et al., 2014), ALLPATHS (Butler et al., 2008) and MaSuRCA (Zimin et al., 2013). This approach is time-consuming and computationally intensive. The *de novo* assembly strategy has been used for the pangenome analysis of cultivated soybean (Li et al., 2010), wild soybean (Li et al., 2014), rice (Wang et al., 2018), *B. oleracea* (Golicz et al., 2016) and *Medicago truncatula* (Zhou et al., 2020).

Metagenomic-like approach. This strategy consists to all sequenced fragments from different taxon representatives combining into one pool and the *de novo* assembling pangenome sequences from these fragments. Each assembled contig is then assigned to a particular genome by the sample original reads alignment to the metagenomic assembly and then contig coverage is evaluated. This method allows low-coverage sequencing results to be handled. The metagenomic approach has been used to analyse the genome of rice (Yao et al., 2015) and tomato (Gao et al., 2019).

Map-then-assembly. This strategy uses one complete genome assembly (reference sequence) as the basis for the genome assembly of the other taxon members (guide assembly). The reads from a single species are mapped to the reference genome, and not mapped reads are discarded and assembled separately. The reference genome sequence is complemented with new sequences, and the samples are compared with the reference genome. This method reduces the time required to construct a pangenome. If a genomic segment is found in more than one sample, the segment will be integrated from the first sample while the *de novo* method creates two complete genomes. This strategy has been used in the sunflower pangenome analysis (Hübner et al., 2019).

It should also be noted, that in a number of studies, the researchers did not use the genomic sequences assembly, but aligned short reads to a reference genome. This approach allows assessing the SNP and phenotypic plants characteristics relations. Methods based on the short reads alignment are also described, which allows the identification of structural rearrangements, duplications and gene losses (Zhao et al., 2013). The alignment method was used in the maize pantranscriptome analysis (Hirsch et al., 2014), in the assessment of CNV's changes in the potato pangenome analysis (Żmieńko et al., 2014).

Pangenome analysis and annotation methods

Based on a comparison of sequences, genome annotation allows identifying gene sequences in taxon representatives' genomes, to determine orthologous genes and universal and variable genes families. Several software packages are designed for pangenomes automatic annotation. They perform the main steps of the pangenome sequence analysis and annotation. The capabilities of a number of these programs are briefly described below.

PGAP (Zhao Y. et al., 2012) performs large-scale gene search, functional annotation, orthologous gene clusters ontology term enrichment, species evolution analysis, pangenome structural analysis, and the universal and variable pangenome parts identification. In the updated version of this program, PGAP-X (Zhao Y. et al., 2018), methods for presentation and visualization of pangenome analysis results are further developed.

PpsPCP (Tahir ul Qamar et al., 2019) was developed for a pangenome PAV identification. The analysis is based on a full-genome taxon and a reference genome sequences comparison in several rounds with sequential correction of both gene set and gene alignment sites in the reference genome. As a result, a pangenome gene set is created by combining the individual genome sequences with the reference genome and their annotation.

BPGA (Chaudhari et al., 2019) provides a wide range of pangenome analysis opportunities: gene clustering based on sequence similarity, orthologs presence/absence analysis, the pangenome and its universal part sizes plotting, phylogenetic tree reconstruction, metabolic pathway and functional annotation analysis, GC composition deviation assessment, various statistical pangenome characteristics calculation, and several other features.

panX (Ding et al., 2018) aims to identify orthologous genes clusters. The sequence comparison clustering, verification and refinement of cluster composition based on evolutionary distance analysis and phylogenetic reconstruction, and assesses the association between the gene composition of individual taxon members and their phenotypes are used.

Pan4Draft (Veras et al., 2018) is designed to improve pangenome annotation by adding sequence information on unfinished genomes. An annotation and assembly to the chromosome level in these genomes is incomplete, but their sequences contain genomic DNA fragments and provide valuable information about the species genome diversity. Information about plant pangenome analysis methods and software for processing and analysis of plant pangenome are provided in Supplementary 2 and 3.

Pangenomic data use perspectives

Currently, the research field of the crop pangenomes sequencing and analysis is developed rapidly and provides more and more information about genetic variations and new genes.

One of the fundamental problems in the crop pangenomes study is to evaluate the genetic diversity of their cultivated representatives as well as wild relatives. This analysis allows us to establish the origin and evolution of cultivated plants, to estimate the breeding process impact on the genetic structure of varieties. Thus, the pangenome analysis helps to answer a number of important questions about patterns of the genome evolution at species level, about mechanisms of the genes *de novo* origination, the gene functions diversity and their associations with phenotypic traits of plants.

One of the important directions of the crop pangenome research is the wild relatives' genome sequencing and analysis. It is supposed that wild relatives of cultivated plants may contain a pool of genes related to adaptation of organisms to environmental conditions, response to biotic stresses; i. e. those genes that may have been lost by cultivated plants as a result of artificial selection (bottleneck effect) (Goncharov, Kondratenko, 2008; Goncharov, 2013; Purugganan, 2019). The discovered genes can be further used to create new genotypes that are more resistant to pathogens, pests and abiotic stress. Thus, the study of agricultural plant pangenomes has not only a fundamental aspect, but is also important in terms of practical breeding.

Conclusion

A better understanding of genetic diversity, combined with advanced sequencing technologies and high-throughput phenotyping can facilitate trait analysis to identify useful genetic mutations. In addition, it allows to access a wider range of genetic resources helps to select the best strategies in breeding programmes and ultimately accelerates crop breeding to develop varieties with consistently high yields under stressful conditions.

Pangenomic studies offer a wider understanding of the crop gene pools genetic diversity than genome resequencing studies and thus can be extremely useful for the crop improvement. Nevertheless, the knowledge obtained through pangenomic researches requires integration with QTL/GWAS and genome resequencing studies to identify important genes and alleles to be used in an effective breeding strategy.

References

Alcaraz L.D., Moreno-Hagelsieb G., Eguiarte L.E., Souza V., Herrera-Estrella L., Olmedo G. Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics*. 2010;11(1):332.

Bragina M.K., Afonnikov D.A., Salina E.A. Progress in plant genome sequencing: research directions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2019;23(1):38-48. DOI 10.18699/VJ19.459. (in Russian)

Butler J., MacCallum I., Kleber M., Shlyakhter I.A., Belmonte M.K., Lander E.S., Nusbaum C., Jaffe D.B. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* 2008;18(5): 810-820. DOI 10.1101/gr.7337908.

Chaudhari N.M., Gupta V.K., Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline. *Sci. Rep*. 2019;6(1):1-10. DOI 10.1038/srep24373.

Ding W., Baumdicker F., Neher R.A. panX: pan-genome analysis and exploration. *Nucleic Acids Res*. 2018;46(1):e5-e5. DOI 10.1093/nar/gkx977.

Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D.M., Thannhauser T.W., Burzynski-Chang E.A., Fish T.L., Stromberg K.A., Sacks G.L., Foolad M.R., Diez M.J., Blanca J., Canizares J., Xu Y., Knaap E., Huang S., Klee H.J., Giovannoni J.J., Fei Z. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet*. 2019;51(6). DOI 1044.10.1038/s41588-019-0410-2.

Golicz A.A., Batley J., Edwards D. Towards plant pangenomics. *Plant Biotechnol. J*. 2016;14(4):1099-1105. DOI 10.1111/pbi.12499.

Goncharov N.P. Plants domestication. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2): 884-899. 2013;17(4/2):884-899. (in Russian)

Goncharov N.P., Kondratenko E.Ja. Wheat origin, domestication and evolution. *Informatcionniy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders.* 2008;12(1-2):159-179. (in Russian)

Gordon S.P., Contreras-Moreira B., Woods D.P., Des Marais D.L., Burgess D., Shu S., Stritt C., Roulin A.C., Schackwitz W., Tyler L., Martin J., Lipzen A., Dochy N., Phillips J., Barry K., Geuten K., Budak H., Juenger T.E., Amasino R., Caicedo A.L., Goodstein D., Davidson P., Mur L.A.J., Figueroa M., Freeling M., Catalan P., Vogel J.P. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun*. 2017;8(1):2184. DOI 10.1038/s41467-017-02292-8.

Hirsch C.N., Foerster J.M., Johnson J.M., Sekhon R.S., Muttoni G., Vaillancourt B., Peñagaricano F., Lindquist E., Pedraza M., Barry K., Leon N., Kaeppler Sh.M., Buell R.C. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 2014;26(1):121-135. https://doi.org/10.1105/tpc.113.119982.

Howe K.L., Contreras-Moreira B., De Silva N., Maslen G., Akanni W., Allen J., Carbajo M. Ensembl Genomes 2020 – enabling non-vertebrate genomic research. *Nucleic Acids Res*. 2020;48(D1):D689-D695. DOI 10.1093/nar/gkz890.

Hübner S., Bercovich N., Todesco M., Mandel J.R., Odenheimer J., Ziegler E., Lee J.S., Baute G.J., Owens G.L., Grassa C.J., Ebert D.P., Ostevik K.L., Moyers B.T., Yakimowski S., Masalia R.R., Gao L., Ćalić I., Bowers J.E., Kane N.C., Swanevelder D.Z.H., Kubach T., Muños S., Langlade N.B., Burke J.M., Rieseberg L.H. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*. 2019;5(1):54-69. DOI 10.1038/s41477-018-0329-0.

Hurgobin B., Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete. *Biology*. 2017;6(1):21. DOI 10.3390/biology6010021.

Hurgobin B., Golicz A.A., Bayer P.E., Chan C.K., Tirnaz S., Dolatabadian A., Schiessl S.V., Samans B., Montenegro J.D., Parkin I.A., Pires J.C. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus. Plant Biotechnol. J*. 2018;16(7):1265-1274. DOI 10.1111/pbi.12867.

Jin M., Liu H., He C., Fu J., Xiao Y., Wang Y., Xie W., Wang G., Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci. Rep*. 2016;6:18936. DOI 10.1038/srep18936.

Li C., Lin F., An D., Wang W., Huang R. Genome sequencing and assembly by long reads in plants. *Genes*. 2018;9(1):6. DOI 10.3390/genes9010006.

Li R., Zhu H., Ruan J., Qian W., Fang W., Shi Z., Li Y., Li Sh., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wang J. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20(2):265-272. DOI 10.1101/gr.097261.109.

Li Y.H., Zhou G., Ma J., Jiang W., Jin L.G., Zhang Z., Guo Y., Zhang J., Sui Y., Zheng L., Zhang S.S. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol*. 2014;32(10):1045. DOI 10.1038/nbt.2979.

Liu Y., Du H., Li P., Shen Y., Peng H., Liu S., Zhou G., Zhang H., Liu Z., Shi M., Huang X., Li Y., Zhang M., Wang Z., Zhu B., Han B., Liang C., Tian Z. Pan-genome of wild and cultivated soybeans. *Cell*. 2020;182(1):162-176. DOI 10.1016/j.cell.2020.05.023.

Lu F., Romay M.C., Glaubitz J.C., Bradbury P.J., Elshire R.J., Wang T., Li Y., Li Y., Semagn K., Zhang X., Hernandez A.G. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun*. 2015;6:6914. DOI 10.1038/ncomms7914.

Ma Y., Liu M., Stiller J., Liu Ch. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. *BMC Genomics*. 2019;20(1):12. https://doi.org/10.1186/s12864-018-5357-7.

Marchant D.B., Soltis D.E., Soltis P.S. Genome evolution in plants. *eLS*. 2016;1-8. DOI 10.1002/9780470015902.a0026814.

Montenegro J.D., Golicz A.A., Bayer P.E., Hurgobin B., Lee H., Chan C.K., Visendi P., Lai K., Doležel J., Batley J., Edwards D. The pangenome of hexaploid bread wheat. *Plant J*. 2017;90(5):1007-1013. DOI 10.1111/tpj.13515.

Plissonneau C., Hartmann F.E., Croll D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol*. 2018;16(1):5. DOI 10.1186/s12915-017-0457-4.

Purugganan M.D. Evolutionary insights into the nature of plant domestication. *Curr. Biol*. 2019;29(14):R705-R714. DOI 10.1016/j.cub.2019.05.053.

Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Minx P. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326(5956):1112-1115. DOI 10.1126/science.1178534.

Snipen L., Almøy T., Ussery D.W. Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*. 2009;10(1):385. DOI 10.1186/1471-2164-10-385.

Springer N.M., Ying K., Fu Y., Ji T., Yeh C.T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A.L., Barbazuk W.B., Jeddeloh J.A., Nettleton D., Schnable P.S. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*. 2009;5(11):e1000734. DOI d10.1371/journal.pgen.1000734.

Sun C., Hu Z., Zheng T., Lu K., Zhao Y., Wang W., Shi J., Wang C., Lu J., Zhang D., Li Z., Wei C. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res*. 2016;45(2):597-605. DOI 10.1093/nar/gkw958.

Tahir Ul Qamar M., Zhu X., Xing F., Chen L.L. ppsPCP: a plant presence/absence variants scanner and pan-genome construction pipeline. *Bioinformatics*. 2019;35(20):4156-4158. DOI 10.1093/bioinformatics/btz168.

Tao Y., Zhao X., Mace E., Henry R., Jordan D. Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant*. 2019;12(2):156-169. DOI 10.1016/j.molp.2018.12.016.

Tets V.V. Pangenome. *Citoligiya = Cytology.* 2003;45(5):526-531. (in Russian)

Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., DeBoy R.T. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA*. 2005;102(39):13950-13955. DOI 10.1073/pnas.0506758102.

Tranchant-Dubreuil C., Rouard M., Sabot F. Plant pangenome: impacts on phenotypes and evolution. *Ann. Plant Rev. Online*. 2018;453-478. DOI 10.1002/9781119312994.apr0664.

Veras A., Araujo F., Pinheiro K., Guimarães L., Azevedo V., Soares S., Costa da Silva A., Ramos R. Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes. *Sci. Rep*. 2018;8(1):1-8. DOI 10.1038/s41598-018-27800-8.

Wang W., Mauleon R., Hu Z., Chebotarov D., Tai S., Wu Z., Li M., Zheng T., Fuentes R.R., Zhang F., Mansueto L. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018;557(7703):43. DOI 10.1038/s41586-018-0063-9.

Wendel J.F., Jackson S.A., Meyers B.C., Wing R.A. Evolution of plant genome architecture. *Genome Biol*. 2016;17:37. DOI 10.1186/s13059-016-0908-1.

Wing R.A., Purugganan M.D., Zhang Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* 2018;19:505-517. DOI 10.1038/s41576-018-0024-z.

Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Zhou X., Lam T., Li Y., Xu X., Wong G.K., Wang J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2014;30(12):1660-1666. DOI 10.1093/bioinformatics/btu077.

Yao W., Li G., Zhao H., Wang G., Lian X., Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* 2015;16:187. DOI 10.1186/s13059-015-0757-3.

Zerbino D.R., Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. DOI 10.1101/gr.074492.107.

Zhao M., Wang Q., Wang Q., Jia P., Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics.* 2013;14(1). DOI 10.1186/1471-2105-14-S11-S1.

Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet*. 2018;50(2):278-284. DOI 10.1038/s41588-018-0041-z.

Zhao Y., Sun C., Zhao D., Zhang Y., You Y., Jia X., Yang J., Wang L., Wang J., Fu H., Kang Y., Chen F., Yu J., Wu J., Xiao J. PGAP-X: extension on pan-genome analysis pipeline. *BMC Genomics.* 2018; 19(1):115-124. DOI 10.1186/s12864-017-4337-7.

Zhao Y., Wu J., Yang J., Sun S., Xiao J., Yu J. PGAP: pan-genomes analysis pipeline. *Bioinformatics*. 2012;28(3):416-418. DOI 10.1093/bioinformatics/btr655.

Zhou L., Zhang T., Tang S., Fu X., Yu Sh. Pan-genome analysis of *Paenibacillus polymyxa* strains reveals the mechanism of plant growth promotion and biocontrol. *Antonie van Leeuwenhoek.* 2020;113: 1539-1558. DOI 10.1007/s10482-020-01461-y.

Zimin A.V., Marçais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. The MaSuRCA genome assembler. *Bioinformatics*. 2013; 29(21): 2669-2677. DOI 10.1093/bioinformatics/btt476.

Żmieńko A., Samelak A., Kozłowski P., Figlerowicz M. Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 2014;127: 1-18. DOI 10.1007/s00122-013-2177-7.

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
E.A. Salina orcid.org/0000-0001-8590-847X

Acknowledgements. This work was carried out with funding from the Russian Science Foundation, grant No. 18-14-00293. The authors are grateful to N.A. Shmakov and D.A. Afonnikov for their assistance with the text. We consider it our pleasant duty to thank the anonymous reviewers for their valuable comments.

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

Received November 4, 2020. Revised December 27, 2020. Accepted January 3, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):64-70
DOI 10.18699/VJ21.008

Original Russian text www.bionet.nsc.ru/vogis/

Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm

E.A. Urbanovich[1] ✉, D.A. Afonnikov[2, 3], S.V. Nikolaev[2, 4]

[1] Novosibirsk State Technical University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K.I. Skryabin, Moscow, Russia
✉ e.urbanovich98@gmail.com

Abstract. Determining the quantitative content of chlorophylls in plant leaves by their reflection spectra is an important task both in monitoring the state of natural and industrial phytocenoses, and in laboratory studies of normal and pathological processes during plant growth. The use of machine learning methods for these purposes is promising, since these methods allow inferring the relationships between input and output variables (prediction model), and in order to improve the quality of the prediction, a researcher may modify predictors and selects a set of method parameters. Here, we present the results of the implementation and evaluation of the random forest algorithm for predicting the total concentration of chlorophylls *a* and *b* from the reflection spectra of plant leaves in the visible and infrared wavelengths. We used the reflection spectra for 276 leaf samples from 39 plant species obtained from open sources. 181 samples were from the sycamore maple (*Acer pseudoplatanus* L.). The reflection spectrum represented wavelengths from 400 to 2500 nm with a step of 1 nm. The training set consisted of the 85 % of *A. pseudoplatanus* L. samples, and the performance was evaluated on the remaining 15 % samples of this species (validation sample). Six models based on the random forest algorithm with different predictors were evaluated. The selection of control parameters was performed by cross-checking on five partitions. For the first model, the intensity of the reflection spectra without any transformation was used. Based on the analysis of this model, the optimal ranges of wavelengths for the remaining five models were selected. The best results were obtained by models that used a two-point estimation of the derivative of the reflection spectrum in the visible wavelength range as input data. We compared one of these models (the two-point estimation of the derivative of the reflection spectrum in the range of 400–800 nm with a step of 1 nm) with the model by other authors (which is based on the functional dependence between two unknown parameters selected by the least squares method and two reflection coefficients, the choice of which is described in the article). The comparison of the results of predictions of the model based on the random forest algorithm with the model of other authors was carried out both on the validation sample of maple and on the sample from other plant species. In the first case, the predictions of the method based on a random forest had a lower estimate of the standard deviation. In the second case, the predictions of this method had a large error for small values of chlorophyll, while the third-party method had acceptable predictions. The article provides the analysis of the results, as well as recommendations for using this machine learning method to assess the quantitative content of chlorophylls in leaves.

Key words: random forest; remote methods; leaf optics; pigments.

For citation: Urbanovich E.A., Afonnikov D.A., Nikolaev S.V. Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):64-70. DOI 10.18699/VJ21.008

Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса

Е.А. Урбанович[1] ✉, Д.А. Афонников[2, 3], С.В. Николаев[2, 4]

[1] Новосибирский государственный технический университет, 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
[4] Московская государственная академия ветеринарной медицины и биотехнологии – МВА им. К.И. Скрябина, Москва, Россия
✉ e.urbanovich98@gmail.com

Аннотация. Определение количественного содержания хлорофиллов в листьях растений по их спектрам отражения – важная задача как при мониторинге состояния естественных и промышленных фитоценозов, так и в лабораторных исследованиях нормальных и патологических процессов в ходе роста растения. Применение для этих целей методов машинного обучения является перспективным, поскольку они позволяют «автомати-

чески» строить решающие правила для получения результата (модель предсказания), а исследователю (для повышения качества предсказания) остаются модификация предикторов и выбор множества параметров метода. В статье приведены результаты построения решающих правил алгоритмом случайного леса (random forest) для предсказания суммарной концентрации хлорофиллов *a* и *b* по спектрам отражения листьев растений в видимом и инфракрасном (ИК) диапазонах длин волн. Набор данных взят из открытых источников. Они включали 276 образцов листьев 39 видов растений. При этом 181 образец получен при анализе листьев белого клена (*Acer pseudoplatanus* L.). Спектр отражения представлен в диапазоне 400–2500 нм с шагом 1 нм. Обучение происходило на 85 % образцов *A. pseudoplatanus* L., оценка качества предсказания – на оставшихся 15 % образцов этого вида (валидационная выборка). Построено шесть моделей на основе алгоритма случайного леса с разными предикторами. Подбор управляющих параметров осуществляли при помощи перекрестной проверки на пяти разбиениях. Предикторами первой модели выступали имеющиеся значения по спектру отражения без какой-либо обработки с нашей стороны. После проведения анализа этой модели были выбраны диапазоны длин волн предикторов для оставшихся пяти моделей. Лучшие предсказания имеют модели с разностной производной спектра отражения в видимом диапазоне длин волн. Модель с первой производной спектра отражения в диапазоне 400–800 нм с шагом 1 нм брали для сравнения с моделью других авторов. Этой моделью выступает функциональная зависимость с двумя неизвестными параметрами, подбираемыми методом наименьших квадратов, и двумя коэффициентами отражения, выбор которых описывается в настоящей статье. Сравнение результатов предсказаний модели с применением алгоритма случайного леса проводили как на валидационной выборке клена, так и на выборке из других видов растений. В первом случае предсказания метода на основе случайного леса имели меньшую оценку среднеквадратического отклонения. Во втором случае предсказания этого метода были с большой ошибкой при малых значениях хлорофилла, в то время как сторонний метод имел приемлемые предсказания. В статье приводятся анализ результатов и рекомендации по применению этого метода машинного обучения для оценки количественного содержания хлорофиллов в листьях.

Ключевые слова: случайный лес; дистанционные методы; оптика листа растения; пигменты.

Introduction

Pigments are low-molecular-weight compounds that give color to plant organs and play an important role in their life, performing photosynthetic, protective and metabolic functions. In terrestrial plants, the most well-known pigments are chlorophylls (which provide the green color of plant organs and play a crucial role in photosynthesis), carotenoids (which give red and yellow color, also participate in photosynthesis), anthocyanins (which give a purple color, perform protective functions), as well as a number of other compounds (Croft, Chen, 2018). Photosynthetic pigments, chlorophylls and carotenoids, attract the most attention from researchers; they have different absorption spectra and perform different functions in the process of photosynthesis, which is due to structural differences between the molecules of these substances.

Chlorophyll in plants is represented by two types of molecules, *a* and *b*, which have structural differences and differ in their light-absorbing properties (Du et al., 1998). It allows photosynthetic organisms to collect sunlight at different wavelengths to maximize the light energy available for photosynthesis. Changes in the concentrations of photosynthetic pigments are closely related to the physiological state of plants. For example, when the leaves of plants wither, there is a rapid decrease in the concentration of chlorophylls compared to carotenoids, thereby increasing the ratio of carotenoids to chlorophylls causes the leaves to turn red and yellow (Croft, Chen, 2018). The content of pigments, in particular chlorophylls *a* and *b*, can thus serve as an indicator of the state of plants during normal growth and during the development of infections, as well as stress, photosynthetic activity, metabolic disorders, etc. (Młodzińska, 2009). The need to determine the physiological state of plants often arises in the course of solving many scientific and practical problems, so methods for assessing the content of pigments in plant organs and tissues are constantly being developed and improved.

Quantitative and qualitative information about pigments can be obtained using chemical methods (Lichtenthaler, 1987; Porra et al., 1989; Wellburn, 1994). However, for many tasks, a more convenient approach is to use remote methods based on the light reflection spectra from the plant leaf (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003). The reflectivity of the leaf in the optical and infrared (IR) wavelengths (400–2500 nm) depends on various biochemical and physical factors, including the content of chlorophyll and other leaf pigments, nitrogen, water, as well as on the internal structure of the leaves and the characteristics of their surface (Croft, Chen, 2018). Plant pigments are characterized by the absorption of electromagnetic radiation in the visible (400–700 nm) and near-IR (1300–2500 nm) wavelength ranges. The absorption of the leaf components in the near-infrared region in the range of 750–1300 nm is low, since in this wavelength range there is an intense reflection from the components of the internal structure of the leaves. Thus, the reflection coefficient in the near-IR range depends on both the concentration of enzymes and the structure of the leaf. All these facts make it possible to use remote observation methods in both the visible and near-infrared wavelength ranges to monitor the physiological state of plants (Merzlyak et al., 2003; Alt et al., 2020).

One of the approaches to estimating the content of chlorophylls from the reflection spectrum is to select empirical dependencies (indices) between the reflection coefficients at certain wavelengths, the choice of which is also an important part of the method, and the content of chlorophylls (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003; Suo et al., 2010; Nikolaev et al., 2018). The success of such a "classical" approach directly depends on the depth of our understanding of the physics of the process.

Currently, machine learning methods are often used to predict the characteristics of biological objects (Doktor et

al., 2014; Feng et al., 2020). Their advantage is that usually a complex nonlinear dependence on many variables can be approximated with the necessary accuracy by machine learning methods. In simple cases, the data is fed to the program input without any processing, however, the accuracy of the predicted parameter will be quite high. Each machine learning method has its own ways to improve the accuracy of the prediction, for example, by varying the control actions. There are also ways to transform the input data to improve the result. Thus, in the analysis of spectra, the calculation of the derivative makes it possible to eliminate additive components and highlight such characteristic features of the spectrum as the positions of maxima, minima, and points.

The aim of our research was to develop a machine learning method using a random forest algorithm to predict the total concentration of chlorophylls *a* and *b* in plant leaves from the values of the reflection spectra in the visible and infrared wavelength ranges. The accuracy of the prediction is evaluated in comparison with the results obtained from the analytical functional dependence, and the advantages and disadvantages of both approaches are determined.

Materials and methods

Experimental data. The characteristics of the leaf reflection spectra at different concentrations of chlorophylls *a* and *b* were downloaded from the EcoSIS database (ecosis.org), set angers2003 (Jacquemound et al., 2003; Féret et al., 2008). 276 leaf samples of 39 plant species were examined. 181 samples are the leaves of sycamore maple (*Acer pseudoplatanus* L.). The data on the reflection spectrum are presented in the range of 400–2500 nm with a step of 1 nm. The ASD FieldSpec spectrum radiometer is used for this purpose; the pigment concentrations were determined by the Lichtenheler method and are presented in units of measurement of μg/cm^2 (see details in (Jacquemound et al., 2003; Féret et al., 2008)).

Mathematical statement of the problem. Let there be a general set of R_λ^{gen} of all possible reflection coefficients of plant leaves for given wavelengths λ and Chl^{gen} – the values of the sum of the concentration of chlorophylls *a* and *b* corresponding to R_λ^{gen}. We have an R_λ – subsample of R_λ^{gen} and *Chl* – values of the sum of the concentration of chlorophylls *a* and *b* corresponding to R_λ. It is required to construct the functional f: $R_\lambda^{\text{gen}} \rightarrow Chl^{\text{gen}}$ from the set (R_λ, *Chl*). Moreover, since this idealized functional cannot be implemented, we get an approximating functional: $\tilde{f}$: $R_\lambda \rightarrow \widehat{Chl}$.

Building a prediction model using the random forest method. The random forest (RF) method was chosen for constructing the functional (Breiman, 2001; Hastie et al., 2009). It allows you to get the accuracy of the prediction of the target function, as a rule, higher than in the case of linear regression methods. The idea of the algorithm is to apply an ensemble of decision trees. Each decision tree in this ensemble sets a piecewise constant function, which is obtained by minimizing the loss function (for example, the mean square of the deviation). The algorithm combines two main ideas: the Breiman bagging method (Breiman, 1996) and the random subspace method proposed by T.K. Ho (1998). In our work, we used the implementation of the random forest method from the sklearn library (scikit-learn.org) of the Python language.

To predict the chlorophyll concentrations by the random forest method, several models that differed in the input data sets were taken. First, each set was characterized by an interval of wavelengths, the intensity of reflection at which was taken into account. In total, several sets of intervals were considered: 400–2450, 400–800 nm, and a combined set of two intervals of 500–600 and 680–740 nm. Second, the models differed in the type of input data. These included the values of the intensity of the reflection spectra at certain wavelengths (base data type), the values of the first derivatives of the spectral curves for the same wavelengths (der data type), and the values of the second derivatives (der2 data type). Some models were based on only one data type, while others shared multiple data types. Such combinations were marked with a summation sign (for example, base+der).

In this paper, six models have been considered. They are designated as RF-(X–Y)-Z, where (X–Y) – intervals of wavelengths, Z – type data model: RF-(400–2450)-base (the intensity of the spectrum in intervals of wavelengths 400–2450 nm); RF(400–800)-base (the intensity of the spectrum in intervals of wavelengths 400–800 nm); RF(400–800)-base+der (intensity spectrum and the first derivative in the intervals of wavelengths 400–800 nm); RF(400–800)-der (first derivative in the intervals of wavelengths 400–800 nm); RF(400–800)-der+der2 (first and second derivatives in the interval of wavelengths 400–800 nm); RF(500–600; 680–740)-base+der+der2 (intensities, first and second derivatives in the wavelength ranges 500–600 and 680–740 nm).

As an approximation of the derivative of the spectral curves, the first-order finite difference with a change equal to 1 was used, which was calculated by the formula $D_i = R_i - R_{i-1}$. In this calculation, there is no derivative for the first value. For simplicity, the finite difference is referred to the derivative throughout the text. The second derivative was calculated as the derivative of the derivative of the spectral curve.

When configuring the random forest algorithm, the following control parameters were selected:

- max_depth: [2, 3, 4, 5, 6] – the maximum depth of the tree;
- max_features: [2, 7, sqrt, log2, auto] – the number of features that the partition is searched for (auto – all features);
- n_estimators: [5, 10, 15, 30, 40] – the number of trees in the random forest ensemble;
- random_state: 20200605.

The specified parameters of the algorithm were selected by cross-checking on five samples of the same size obtained from a randomly mixed initial training sample. Four subsamples were used for training the model, and the fifth one was used for testing it. To determine the best control parameters, the test results (mean square deviation of the target indicator – *mse*) were averaged between models with the same control parameters (i. e., obtained during cross-validation) and sorted. The control parameters for which the average *mse* is the minimum are the best. As the final model, one of the five models with the best control parameters is selected, which has the minimum *mse* when tested among the models obtained by the cross-validation method.

The maximum depth of the trees is chosen to be 6, which gives $2^6 = 64$ intervals for partitioning the parameter space, despite the fact that the sample length taken to build the model

is 123. The depth increasing could lead to overfitting. The number of trees in the forest (up to 40) may seem redundant for 123 sample values, but the parameters of each of the decision trees were selected on different subspaces (since the random subspace method is used), and the dimension of the features was always greater than the number of elements in the sample.

It should be noted that the algorithm implemented in the sklearn library allows us to obtain the informativeness of each of the model features and select the most informative ones for the obtained decision rules (Breiman, 2001; Hastie et al., 2009; Louppe et al., 2013).

Construction of empirical functional dependencies. As a functional of $\tilde{f}: R_\lambda \to \widehat{Chl}$ we additionally chose an empirical dependence from the work (Gitelson et al., 2003) (the GGM method, which we named after the authors' surnames), represented by the expression

$$\widehat{Chl} = \alpha \cdot \left(\frac{1}{R_\lambda} - \frac{1}{R_{\text{NIR}}} \right) \cdot R_{\text{NIR}} + \beta, \qquad (1)$$

where $\widehat{Chl}$ is the total concentration of chlorophylls *a* and *b*; R_λ is the reflection coefficient at the wavelength λ; R_{NIR} is the reflection coefficient in the near-infrared range (for example, at a wavelength of 800 nm); α and β are selected in such a way as to minimize the selected loss function. A.A. Gitelson and co-authors (2003) recommend choosing wavelengths from the range $\lambda \in [525; 555] \cup [695; 725]$. According to the authors, the advantage of this algorithm is that the R_{NIR} coefficient "corrects" the influence of the plant tissue structure on the reflection spectrum and allows us to extend the found function to plants with different leaf structure.

The comparison of methods for predicting the concentration of chlorophyll. The sycamore maple sample from the angers2003 data set was randomly divided into a training and a validation sample in the ratio of 85 : 15. For the methods used in this work for predicting the random forest algorithm (RF) and functional dependence (GGM), the optimal parameters are selected on the training sample. The quality control of the algorithms is carried out on a validation sample represented by a sycamore maple and on a sample of non-maple samples. The following metrics were used to evaluate the accuracy of predicting chlorophyll concentrations: *mse*, mean absolute error (*mae*), and determination coefficient R^2. The formulas for calculating metrics are as follows:

$$mse = \frac{1}{n}\sum_1^n (x_i - \hat{x}_i)^2,$$

$$mae = \frac{1}{n}\sum_1^n |x_i - \hat{x}_i|,$$

$$R^2 = 1 - \frac{\sum_1^n (x_i - \hat{x}_i)^2}{\sum_1^n (x_i - \bar{x}_i)^2},$$

where x is the true values; $\hat{x}$ is the predicted values; n is the number of samples, and $\bar{x}$ is the mathematical expectation for the true values. In terms of optimization, *mae* and R^2 are equivalent. The coefficient of determination R^2 is convenient because it is a dimensionless value usually in the range [0; 1], the value of $R^2 < 0$ shows that the arithmetic mean $\bar{x}$ has a better result than the predictions of the constructed model).

Results

Selection of parameters for the functional dependence method. For the GGM prediction on the training sample, we selected the coefficients α and β of equations (1), as well as the values λ so as to maximize the value of R^2. The value λ_{NIR} = 800 nm is selected as the wavelength in the near-infrared range. To get the coefficients α and β, we took a linear model based on the least squares method (the LinearRegression class from the sklearn.linear_model package). For each $\lambda \in [400; 800]$ with a step of 1 nm, a specific type of GGM curve was found. The coefficients of determination R^2 for the predictions of the obtained models are shown in Fig. 1. The highest coefficient of determination was achieved at the wavelength λ = 705 nm. The result is consistent with the recommended range $\lambda \in [525; 555] \cup [695; 725]$ (Gitelson et al., 2003). The RF method is compared with the GGM model obtained at this wavelength λ = 705 nm.

Results of constructing an algorithm based on the random forest method. The characteristics of the accuracy of the prediction of chlorophyll concentrations (the values of the *mse*, *mae*, R^2 parameters) for all six models in the test sample are shown in the table. The RF-(400–800)-der and RF-(400–800)-der+der2 methods demonstrated high prediction accuracy. As the best of them, the RF-(400–800)-der method was selected as having a smaller number of input parameters.

The selection of wavelengths, the reflection coefficients for which were taken as input features for predicting chlorophyll concentrations by the random forest method, was carried out on the basis of the first model (RF-(400–2450)-base). This is due to the fact that at first it was not known whether the entire spectrum was needed, or only a part of it was necessary, and which one. As mentioned earlier, the RF algorithm allows you to evaluate the information content of the features the training took place on. After configuring the control parameters of the RF-(400–2450)-base model, we took the obtained parameters to re-train the models on five training samples (from cross-validation). For these five models, we identified 10 features with the greatest contribution to the prediction. The results are shown in Fig. 2: the vertical lines represent the combined set of wavelengths, the spectrum intensities for which make the most significant contribution to the prediction accuracy (26 wavelengths out of 10 · 5 = 50 possible if the values did not intersect). Interestingly, the most significant features lie in

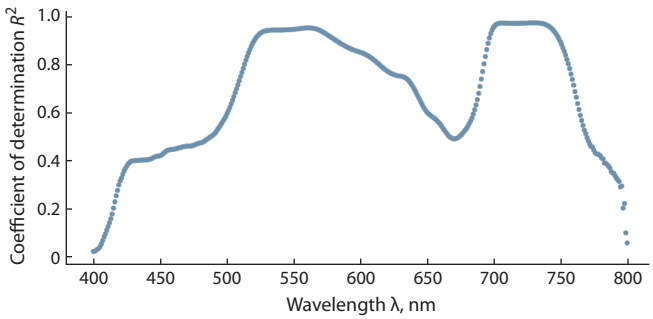


Fig. 1. Determination coefficients of the obtained GGF models at $\lambda \in [400; 800]$, which were calculated on the training sample.

Results of a random forest model trained on different sets of input features

No.	Random forest model	Number of input features	*mse*	*mae*	R^2
1	RF-(400–2450)-base	2051	*30.5*	*3.7*	*0.945*
2	RF-(400–800)-base	401	*26.6*	*3.8*	*0.952*
3	RF-(400–800)-base+der	401 + 400 = 801	10.1	2.4	0.981
4	RF-(400–800)-der	400	**9.1**	**2.4**	**0.984**
5	RF-(400–800)-der+der2	400 + 399 = 799	**8.9**	**2.3**	**0.984**
6	RF-(500–600; 680–740)-base+der+der2	101 + 100 + 99 + 61 + 60 + 59 = 380	10.5	2.7	0.981

Note. The numbers in the description of the feature indicate the range of wavelengths. Additional characteristics of the features: base – reflection spectrum; der – values of the first derivative of the spectrum; der2 – values of the second derivative of the spectrum. The values with the worst accuracy are shown in italics, and the values with the best accuracy are highlighted in bold.

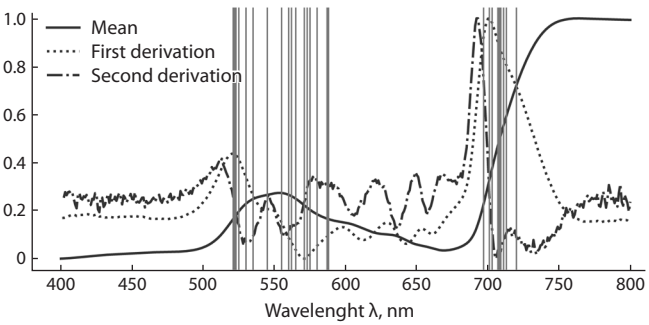


Fig. 2. Characteristics of the reflection spectrum of sycamore maple pigment samples used for model training.

The lines show: the average value of the intensity of the reflection spectrum R_λ (*Y*-axis) for different wavelengths (*X*-axis); the value of the first derivative of the average intensity; the value of the second derivative. The values of the derivatives are normalized to the interval [0; 1]. Vertical lines indicate the wavelengths whose spectrum intensities make the greatest contribution to the prediction accuracy of the RF-(400–2450)-base model.

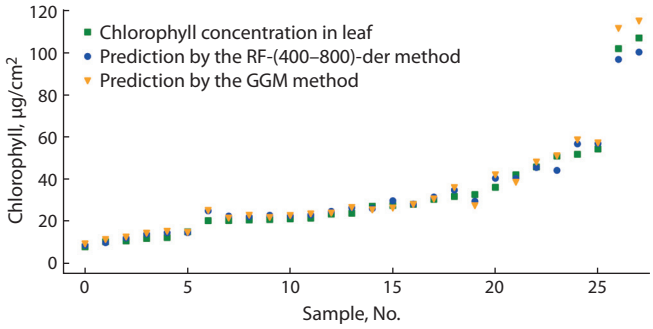


Fig. 3. Comparison of true and predicted values of chlorophyll concentration in sycamore maple leaf tissues for validation sampling.

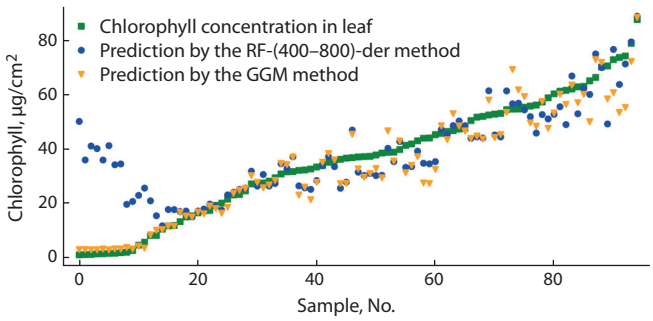


Fig. 4. Comparison of true and predicted values of chlorophyll concentration in leaf tissues of not related to sycamore maple samples.

the visible range; most of these features are in the wavelength range of 500–600 and 680–740 nm. On this basis we have formulated the wavelengths of the input characteristics for the remaining five models for predictions by random forest (see above).

Comparison of the accuracy of the RF and GGM methods. The results of the comparison of the methods for predicting chlorophyll concentrations by the RF-(400–800)-der and GGM methods and their experimentally measured values at different concentrations are shown in Fig. 3 and 4. For sycamore maple samples (the type taken to fit the parameters), the RF-(400–800)-der method shows a better result compared to the GGM method: $\sqrt{mse_{RF}} = 3.01$ μg/cm² versus $\sqrt{mse_{GGM}} = 3.21$ μg/cm². When testing the methods on a sample of plant leaves from other species, the GGM functional dependence method has an advantage $\sqrt{mse_{GGM}} = 6.31$ μg/cm² versus $\sqrt{mse_{RF}} = 12.97$ μg/cm². The GGM method shows high accuracy at low concentrations of chlorophyll, while the RF method shows a large error at these values. However, in the range of chlorophyll concentrations above 20 μg/cm², the RF-(400–800)-der algorithm has the best result: $\sqrt{mse_{RF}} =$ = 5.91 μg/cm² versus $\sqrt{mse_{GGM}} = 7.01$ μg/cm².

Further analysis revealed that for samples with a chlorophyll concentration of less than 7 μg/cm², the reflection coefficients R_{550} (maximum of the reflection spectrum) and R_{680} (minimum of the reflection spectrum) are visually significantly different from all the others (Fig. 5, points in the upper right quarter). The predictions for these samples have a significant error. However, it was not possible to find out what the differences in the reflection spectrum are related to: these samples do not differ from the rest either in the surface density of the leaf or in the equivalent water thickness for the leaf (Jacquemound et al., 2003). Six out of ten plant species from these samples also have samples with normally predicted values. Further analysis of the causes of the anomalous spectrum is difficult, since the data are taken from open sources, and the measurements themselves were carried out more than 17 years ago.

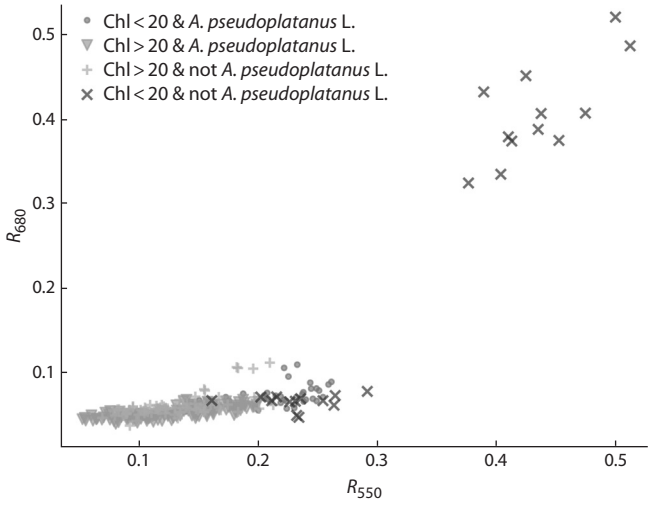


Fig. 5. Scattering diagram of reflection coefficients R_{680} versus R_{550}, with selected categories by chlorophyll concentration (less than/more than 20 mcg/cm^2) and by plant type (*Acer pseudoplatanus* L. or other).

Discussion

Many studies on the application of reflection spectra to estimate pigment concentrations involve neural networks (Golhani et al., 2018), while the decisions founded on tree-based methods are also common in machine learning research tasks. We used the decision tree method to predict chlorophyll concentrations in plant leaves and compared the results with the functional dependence method. We have found the ranges of the spectrum, the intensity of reflection in which most strongly affects the accuracy of the prediction by the random forest method.

The range of 690–750 nm in the literature is called the red edge of photosynthesis (Curran et al., 1990; Gitelson et al., 2003; Croft, Chen, 2018), and the neighborhood of 550 nm, where the maximum of the chlorophyll reflection spectrum is located, is known as the green edge (Gitelson et al., 2003). As it can be seen from Fig. 2, in our study, these regions contain the most important predictors for the random forest method. The choice of a narrower wavelength range of the visible spectrum (400–800 nm) as input features compared to the full source data (400–2450 nm) improved the quality of the model. The explanation is that after dividing the sample into subspaces, some of them are less suitable for training, and the trees trained on these values introduce an error in the total result. The greatest effect was achieved with the use of derived spectral dependences.

The random forest RF method performed well when working with sycamore maple samples, while the functional dependence of GGM performed well when working with different plant species. This is due to the greater generalization ability of the GGM method, as it has fewer configurable parameters. However, the lower accuracy of the RF method on samples from other plant species is partly due to the small size of the training sample and the fact that only one species is represented in it. For example, the best results of the random forest method were achieved with a tree depth of 5 or 6, and this requires a minimum of 32 or 64 objects of the training sample, while the functional method (1) requires a minimum of two points (preferably a point at small values of chlorophyll and a point at large values). Apparently, this feature of the RF method can be eliminated by using more training data with samples from different plant species.

Nevertheless, the procedure for selecting parameters for the RF method showed that the most significant features for prediction lie in the visible region, but the influence of the plant structure was not taken into account in this method. Along with this, in the functional dependence (1), the structure of the plant tissue is taken into account by the R_{NIR} member. If the experiment is performed with different plant species (see Fig. 4), then at low values of chlorophyll, the structure of the plant begins to play a significant role.

Interestingly, both methods work in the range $\lambda \in [525; 555] \cup [695; 725]$. Both methods work on the decline of the derivative of the reflection spectrum, as it is shown in Fig. 2.

The word "random" in the name of the method "random forest" can lead to the idea that when you change the random parameter used by the algorithm, you can get radically different results. We believe that with reasonably selected control parameters, a reasonable division into training and test samples, this probability is low. In our case, 625 models were built for each set of input features (a search of a set of 125 combinations of control parameters, and 5 cross-checked models for each combination). In addition, it follows from the above table that the RF-(400–800)-base+der, RF-(400–800)-der, RF-(400–800)-der+der2 methods have similar results (and, importantly, have less *mse* compared to the GGM method), which indirectly confirms that the results will not change dramatically.

Conclusion

The random forest method is one of the algorithms for constructing functional dependencies using machine learning methods. Therefore, it can be used for mass automatic construction of functions that connect the observed features with the desired ones in monitoring tasks. The results of this work have shown that it is advisable to use the random forest algorithm (and similar ones) in the task of determining the content of chlorophyll in a plant leaf if there is a large sample, at least 32 elements, represented by a wide range of chlorophyll concentrations, while the structure of the leaf tissue changes slightly (for example, the application of the algorithm only on those plants on which it was trained). In other cases, it is better to give preference to methods based on empirical dependencies (such as the GGM method discussed here).

References

Alt V.V., Gurova T.A., Elkin O.V., Klimenko D.N., Maximov L.V., Pestunov I.A., Dubrovskaya O.A., Genaev M.A., Erst T.V., Genaev K.A., Komyshev E.G., Khlestkin V.K., Afonnikov D.A. The use of Specim IQ, a hyperspectral camera, for plant analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2020;24(3):259-266. DOI 10.18699/VJ19.587. (in Russian)

Breiman L. Bagging predictors. *Mach. Learn*. 1996;24:123-140. DOI 10.1023/A:1018054314350.

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

Croft H., Chen J. Leaf pigment content. In: Liang S. (Ed.). Comprehensive Remote Sensing. Oxford, UK: Elsevier, 2018;117-142. DOI 10.1016/B978-0-12-409548-9.10547-0.

Curran P.J., Dungan J.L., Gholz H.L. Exploring the relationship between reflectance red edge and chlorophyll content in slash pine. *Tree Physiol*. 1990;7:33-48. DOI 10.1093/treephys/7.1-2-3-4.33.

Doktor D., Lausch A., Spengler D., Thurner M. Extraction of plant physiological status from hyperspectral signatures using machine learning methods. *Remote Sens*. 2014;6(12):12247-12274. DOI 10.3390/rs61212247.

Du H., Fuh R.-C. A., Li J., Corkan L.A., Lindsey J.S. PhotochemCAD: A computer-aided design and research tool in photochemistry. *Photochem. Photobiol*. 1998;68:141-142. DOI 10.1111/j.1751-1097.1998.tb02480.x.

Feng X., Zhan Y., Wang Q., Yang X., Yu C., Wang H., He Y. Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *Plant J*. 2020;101(6):1448-1461. DOI 10.1111/tpj.14597.

Féret J.-B., François C., Asner G.P., Gitelson A.A., Martin R.E., Bidel L.P.R., Ustin S.L., le Maire G., Jacquemoud S. PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sens. Environ*. 2008;112:3030-3043. DOI 10.1016/j.rse.2008.02.012.

Gitelson A.A., Gritz Y., Merzlyak M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for nondestructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 2003;160(3):271-282. DOI 10.1078/0176-1617-00887.

Gitelson A.A., Merzlyak M.N., Chivkunova O.B. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* 2001;74(1):38-45. DOI 10.1562/0031-8655(2001)074<0038:OPANEO>2.0.CO;2.

Golhani K., Balasundram S.K., Vadamalai G., Pradhan B. A review of neural networks in plant disease detection using hyperspectral data. *Inf. Process. Agric.* 2018;5:354-371. DOI 10.1016/j.inpa.2018.05.002.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, 2009. DOI 10.1007/978-0-387-84858-7.

Ho T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998;20(8):832-844. DOI 10.1109/34.709601.

Horler D.N.H., Dockray M., Barber J. The red edge of plant leaf reflectance. *Int. J. Remote Sens*. 1983;4:273-288. DOI 10.1080/01431168308948546.

Jacquemound S., Bidel L., Francois C., Pavan G. ANGERS Leaf Optical Properties Database. 2003. Data set. Available online [ecosis.org] from the Ecological Spectral Information System (EcoSIS), 2003.

Keskitalo J., Bergquist G., Gardeström P., Jansson S. A cellular timetable of autumn senescence. *Plant Physiol*. 2005;139:1635-1648. DOI 10.1104/pp.105.066845.

Lichtenthaler H.K. Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. *Methods Enzymol*. 1987;148:350-382. DOI 10.1016/0076-6879(87)48036-1.

Louppe G., Wehenkel L., Sutera A., Geurts P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* 2013;26:431-439.

Merzlyak M.N., Gitelson A.A., Chivkunova O.B., Solovchenko A.E., Pogosyan S.I. Application of reflectance spectroscopy for analysis of higher plant pigments. *Rus. J. Plant Physiol*. 2003;50(5):704-710. DOI 10.1023/A:1025608728405.

Młodzińska E. Survey of plant pigments: molecular and environmental determinants of plant colors. *Acta Biol. Crac. Ser. Bot.* 2009;51(1):7-16.

Nikolaev S.V., Urbanovich E.A., Shayapov V.R., Orlova E.A., Afonnikov D.A. A method of evaluating the absorption spectrum of wheat leaf by the spectrum of diffuse reflection. *Sibirskii Vestnik Sel'skokhozyaistvennoi Nauki = Siberian Herald of Agricultural Science*. 2018;48(5):68-76. DOI 10.26898/0370-8799-2018-5-9. (in Russian)

Porra R.J., Thompson W.A., Kriedemann P.E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: Verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *BBA – Bioenergetics.* 1989;975:384-394. DOI 10.1016/S0005-2728(89)80347-0.

Suo X.-M., Jang Y.-T., Yang M., Li S.-K., Wang K.-R., Wang C.-T. Artificial neural network to predict leaf population chlorophyll content from cotton plant images. *Agric. Sci. China*. 2010;9(1):38-45.

Wellburn A.R. The spectral determination of chlorophylls *a* and *b*, as well as total carotenoids, using various solvents with spectrophotometers of different resolution. *J. Plant Physiol.* 1994;144:307-313. DOI 10.1016/S0176-1617(11)81192-2.

ORCID ID
E.A. Urbanovich orcid.org/0000-0003-0602-3097
D.A. Afonnikov orcid.org/0000-0001-9738-1409

Acknowledgements. The work was supported by Russian Foundation for Basic Research No. 17-29-08028 and budget project No. 0259-2021-0009.
Conflict of interest. The authors declare no conflict of interest.
Received October 15, 2020. Revised December 14, 2020. Accepted December 15, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):71-81
DOI 10.18699/VJ21.009

Original Russian text www.bionet.nsc.ru/vogis/

Automatic morphology phenotyping of tetra- and hexaploid wheat spike using computer vision methods

A.Yu. Pronozin[1], A.A. Paulish[2], E.A. Zavarzin[2], A.Yu. Prikhodko[2], N.M. Prokhoshin[2], Yu.V. Kruchinina[1, 3], N.P. Goncharov[1, 4], E.G. Komyshev[1, 2, 3], M.A. Genaev[1, 2, 3]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Kurchatov Genomics Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Novosibirsk State Agrarian University, Novosibirsk, Russia
mag@bionet.nsc.ru

Abstract. Intraspecific classification of cultivated plants is necessary for the conservation of biological diversity, study of their origin and their phylogeny. The modern cultivated wheat species originated from three wild diploid ancestors as a result of several rounds of genome doubling and are represented by di-, tetra- and hexaploid species. The identification of wheat ploidy level is one of the main stages of their taxonomy. Such classification is possible based on visual analysis of the wheat spike traits. The aim of this study is to investigate the morphological characteristics of spikes for hexa- and tetraploid wheat species based on the method of high-performance phenotyping. Phenotyping of the quantitative characteristics of the spike of 17 wheat species (595 plants, 3348 images), including eight tetraploids (*Triticum aethiopicum*, *T. dicoccoides*, *T. dicoccum*, *T. durum*, *T. militinae*, *T. polonicum*, *T. timopheevii*, and *T. turgidum*) and nine hexaploids (*T. compactum*, *T. aestivum*, i:ANK-23 (near-isogenic line of *T. aestivum* cv. Novosibirskaya 67), *T. antiquorum*, *T. spelta* (including cv. Rother Sommer Kolben), *T. petropavlovskyi*, *T. yunnanense*, *T. macha*, *T. sphaerococcum*, and *T. vavilovii*), was performed. Wheat spike morphology was described on the basis of nine quantitative traits including shape, size and awns area of the spike. The traits were obtained as a result of image analysis using the WERecognizer program. A cluster analysis of plants according to the characteristics of the spike shape and comparison of their distributions in tetraploid and hexaploid species showed a higher variability of traits in hexaploid species compared to tetraploid ones. At the same time, the species themselves form two clusters in the visual characteristics of the spike. One type is predominantly hexaploid species (with the exception of one tetraploid, *T. dicoccoides*). The other group includes tetraploid ones (with the exception of three hexaploid ones, *T. compactum*, *T. antiquorum*, *T. sphaerococcum*, and i:ANK-23). Thus, it has been shown that the morphological characteristics of spikes for hexaploid and tetraploid wheat species, obtained on the basis of computer analysis of images, include differences, which are further used to develop methods for plant classifications by ploidy level and their species in an automatic mode.
Key words: wheat spike morphology; wheat; phenomics; image processing; computer vision; machine learning; biotechnology.

For citation: Pronozin A.Yu., Paulish A.A., Zavarzin E.A., Prikhodko A.Yu., Prokhoshin N.M., Kruchinina Yu.V., Goncharov N.P., Komyshev E.G., Genaev M.A. Automatic morphology phenotyping of tetra- and hexaploid wheat spike using computer vision methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):71-81. DOI 10.18699/VJ21.009

Автоматическое фенотипирование морфологии колоса тетра- и гексаплоидных видов пшеницы методами компьютерного зрения

А.Ю. Пронозин[1], А.А. Паулиш[2], Е.А. Заварзин[2], А.Ю. Приходько[2], Н.М. Прохошин[2], Ю.В. Кручинина[1, 3], Н.П. Гончаров[1, 4], Е.Г. Комышев[1, 2, 3], М.А. Генаев[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 Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Новосибирский государственный аграрный университет, Новосибирск, Россия
mag@bionet.nsc.ru

Аннотация. Внутривидовая классификация культурных растений необходима для эффективного сохранения биологического разнообразия видов, изучения их происхождения, определения филогении и проведения межвидовой гибридизации при селекции. Современные возделываемые виды пшениц произошли

от трех диких диплоидных предков в результате гибридизации и нескольких раундов удвоения геномов и представлены ди-, тетра- и гексаплоидными видами. Поэтому идентификация плоидности пшениц и определение их геномного состава являются одними из основных этапов их классификации на основе визуального анализа фенотипических признаков колоса. Цель работы – исследование морфологических характеристик колосьев полиплоидных видов пшеницы методами высокопроизводительного фенотипирования. Выполнено фенотипирование количественных характеристик колоса 17 видов пшеницы (595 растений, 3348 изображений), включая восемь тетраплоидных: *Triticum aethiopicum*, *T. dicoccoides*, *T. dicoccum*, *T. durum*, *T. militinae*, *T. polonicum*, *T. timopheevii*, *T. turgidum* и девять гексаплоидных: *T. compactum*, *T. aestivum* (в том числе изогенная линия сорта Новосибирская 67 АНК-23), *T. antiquorum*, *T. spelta* (включая стародавний сорт *T. spelta* Rother Sommer Kolben), *T. petropavlovskyi*, *T. yunnanense*, *T. macha*, *T. sphaerococcum*, *T. vavilovii*. Морфология колоса описана на основе девяти количественных признаков, включающих форму, размер и остистость. Признаки были получены в результате анализа цифровых изображений с помощью программы WERecognizer. Кластерный анализ растений по характеристикам формы колоса и сравнение их распределений у тетра- и гексаплоидных видов показали более высокую вариабельность признаков у гексаплоидных видов по сравнению с тетраплоидными. При этом сами виды в пространстве характеристик колоса формируют два кластера. К первому относятся преимущественно гексаплоидные виды, за исключением одного тетраплоидного, дикорастущего *T. dicoccoides*, ко второму – тетраплоидные, за исключением трех гексаплоидных, *T. compactum*, *T. antiquorum*, *T. sphaerococcum*, и i:АНК-23. Показано, что морфологические характеристики колосьев для гекса- и тетраплоидных видов, полученные на основе компьютерного анализа изображений, демонстрируют различия, которые в дальнейшем могут быть использованы для разработки методики эффективной классификации растений по плоидности и их видовой принадлежности в автоматическом режиме.
Ключевые слова: пшеница; морфология колоса; феномика; обработка изображений; компьютерное зрение; машинное обучение; биотехнологии.

Introduction

A number of important issues, including aspects of the effective conservation of the biological diversity of cultivated plant species, the study of their origin, and their phylogeny, presupposes a detailed development of intraspecific classifications (Dorofeev et al., 1979; Goncharov, 2011). The producing of such classifications, reflecting the phylogenesis and genetic structure of species, should be considered the main goal of modern taxonomy (Hammer et al., 2011). When developing the classification of cultivated plants, the most complete description of all existing large and small forms (taxons) is assumed (Sinskaya, 1969). On the one hand, this is determined, by the convenience of using such a division in experimental work, on the other hand, it is also determened in the breeding and testing of cultivated plants.

The success and effectiveness of research work is often associated with the detailing and completeness of the experimental study, which depends on what the material is and how much it should be studied. In this regard, it is extremely important that the natural differentiation of one or another genus, the relationship between species, are reflected with high accuracy by a detailed taxonomy (Dorofeev, 1985). It should be noted, that for most of the plants important for agriculture, the volumes of the genus and species have not been unambiguously described yet (Rodionov et al., 2019).

A serious problem in the taxonomy of cultivated plants is the aspect of taxa agregation *vs.* fragmentation, and in cases of cultivation it manifests itself especially in contrast (Golovnina et al., 2009; Goncharov, 2011). At the same time, the effective use of taxonomy of cultivated plants in the work of researchers causes certain difficulties. Both dichotomous tables (Dorofeev et al., 1979; Goncharov, 2009) and ideographic manual book (Zuev et al., 2019) require certain skills; therefore, the producing of a database and software that allows the identification of species by digital images is a very promising direction. The development of these methods is mainly based on technologies for analyzing digital images of plant organs within the framework of computer phenomics (Afonnikov et al., 2016; Zhang et al., 2019; Demidchik et al., 2020; Yang et al., 2020).

Wheat is one of the world's most important food crops. The modern cultivated wheat species evolved from three wild diploid ancestors as a result of their hybridization and several rounds of genome doubling (polyploidization). Currently, cultivated wheat is represented by di- ($2n = 2x = 14$, A^bA^b genome), tetra- ($2n = 4x = 28$, BBA^uA^uDD genome) and hexaploid ($2n = 6x = 42$, BBA^uA^uDD genome) species (Goncharov, Kondratenko, 2008). The main cultivated species, bread wheat (*Triticum aestivum* L.), is a hexaploid (genomic formula BBA^uA^uDD). The ploidy level is one of the main taxonomic and classifying characteristics of wheat species (Dorofeev et al., 1984; van Slageren, Payne, 2013). It can be established by cytogenetic (Rodionov et al., 2020), molecular methods, as well as by comparing the morphological characteristics of plants (Dorofeev et al., 1984). In this work, we studied the morphological traits of the plant spikes of tetraploid and hexaploid wheat species based on the method of high-throughput phenotyping.

The aim of the research was to study the distribution of morphological traits of spikes of tetra- and hexaploid wheat species and compare their distributions.

Material and methods

Biological material. We studied 17 polyploid wheat species, namely, nine hexaploids (*Triticum compactum* Host, *T. aestivum* L., *T. antiquorum* Heer ex Udacz., *T. spelta* L. (includion of *T. spelta* cv. Rother Sommer Kolben), *T. petropavlovskyi* Udacz. et Migusch., *T. yunnanense* King ex S.L. Chen, *T. macha* Dekapr. et Menabde, *T. sphaerococcum* Perciv., *T. vavilovii* (Thum.) Jakubz.), the near-isogenic line ANK-23

Table 1. Characteristics of the studied wheat species

Species	Total					Ploidy	List of vegetation
	photos	plants	accessions	table	pin		
T. compactum Host	472	101	10	177	295	Hexaploid	II18, IX16
T. aestivum L.	456	80	8	166	290		II19, IX16, IX18, X14
T. antiquorum Heer ex Udacz.	184	37	4	116	68		II18, X14
T. spelta L.	164	49	5	40	124		II18
T. petropavlovskyi Udacz. et Migusch.	374	75	6	74	300		II17, IX17, IX18
i:ANK-23	50	10	1	14	36		IX16
T. yunnanense King ex S.L. Chen	191	43	3	43	148		IX17, IX18
T. spelta cv. Rother Sommer Kolben	45	9	1	9	36		IX16, II18
T. macha Dekapr. et Menabde	46	10	1	10	36		IX17, IX18
T. sphaerococcum Perciv.	100	20	2	20	80		IX17
T. vavilovii (Thum.) Jakubz.	15	3	1	3	12		II18
T. aethiopicum Jakubz.	595	119	12	119	476	Tetraploid	X14
T. dicoccoides (Körn. ex Aschers. et Graebn.) Schweinf.	40	8	1	8	32		II16
T. dicoccum (Schrank) Schuebl.	41	9	1	9	32		II17
T. durum Desf.	275	56	5	55	220		II16, II17, II19, IX18
T. militinae Zhuk. et Migusch.	40	8	1	8	32		IX17
T. polonicum L.	95	19	2	19	76		II16, II19
T. timopheevii (Zhuk.) Zhuk.	125	25	3	25	100		II16, IX18
T. turgidum L.	40	8	1	8	32		II15

of bread wheat cv. Novosibirskaya 67 and eight tetraploids (*T. aethiopicum* Jakubz., *T. dicoccoides* (Körn. ex Aschers. et Graebn.) Schweinf., *T. dicoccum* (Schrank) Schuebl., *T. durum* Desf., *T. militinae* Zhuk. et Migusch., *T. polonicum* L., *T. timopheevii* (Zhuk.) Zhuk., *T. turgidum* L.); the sample consists of spikes of 595 individual plants, which was grown in nine vegetation seasons. The plants were grown in 2014–2019 in a greenhouse at the Shared Center 'Laboratory of Artificial Plant Cultivation' of the Institute of Cytology and Genetics SB RAS. A description of the material used is given in Table 1.

It should be noted that none of the large genebanks of the world have typical sets of wheat accessions (collections), so them usually reflect either the researchers view on the methods of selection of such sets (Palmova, 1935) or are determined by the representativeness of the researchers available material (Goncharov, Shumny, 2008). Standard taxonomic descriptions of specimens are given in publicly available databases on genebank websites (http://db.vir.nw.ru/virdb/maindb).

Digital images obtaining. In this work we used two protocols for receiving mature spikes photo. The first is that the spike is placed on the glass of a light table, which is located on a table with a blue top (background). The camera is fixed on a stand above the glass. With this method, the front projection of the spike can be captured. Second, the spike is held vertically in front of the blue background. The spike is supported by clip that are placed on a tripod. With this method, by rotating the spike about its axis, four or more projections of the spike can be captured (Genaev et al., 2018). According to the protocols, a ColorChecker must be present in the photographs. It is needed for colour normalisation and scaling. One plant in our dataset can correspond to up to five pictures of its spike taken with different protocols and in different projections. Examples of spikes images (one for each species) are shown in Fig. 1. In total 3348 spike images in different projections were captured by the two protocols, 2097 of them were of hexaploid species and 1251 were of tetraploid species. Of these, 915 images were obtained using the "on the table" protocol and 2433 "on the clip".

Evaluation of spikes quantitative characteristics. WERecognizer (Genaev et al., 2019) was used to estimate spikes quantitative characteristics based on image analysis. This program describes a wheat spike by geometric model of two quadrangles based on image analysis (Fig. 2). The geometry of this model is described by nine independent parameters. The parameters x_{u1}, x_{u2}, y_{u1}, y_{u2} are for the

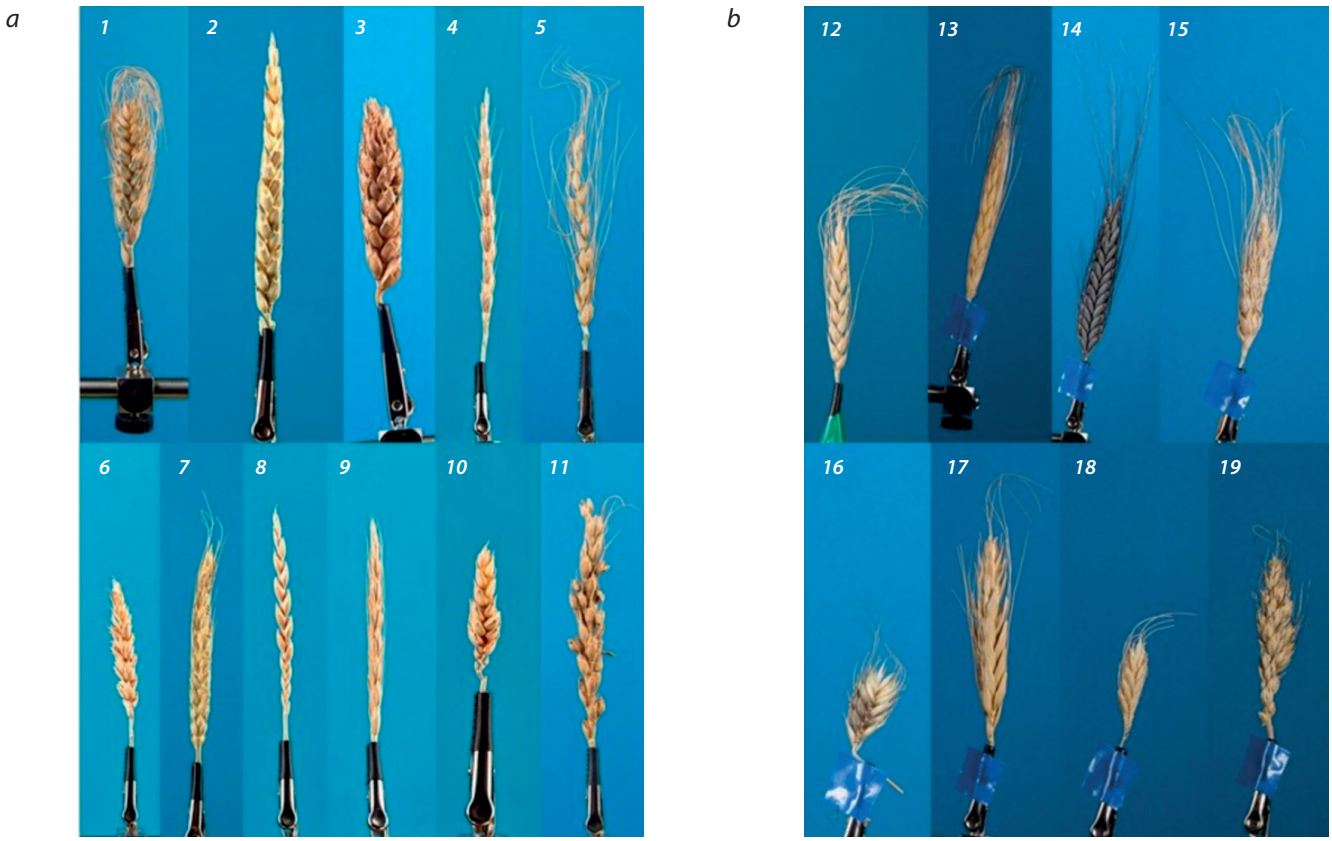


Fig. 1. Spike images of hexaploid (*a*) and tetraploid (*b*) wheat species.

1 – T. compactum; *2 – T. aestivum*; *3 – T. antiquorum*; *4 – T. spelta*; *5 – T. petropavlovskyi*; *6* – i:ANK-23; *7 – T. yunnanense*; *8 – T. spelta* cv. Rother Sommer Kolben; *9 – T. macha*; *10 – T. sphaerococcum*; *11 – T. vavilovii*; *12 – T. aethiopicum*; *13 – T. dicoccoides*; *14 – T. dicoccum*; *15 – T. durum*; *16 – T. militinae*; *17 – T. polonicum*; *18 – T. timopheevii*; *19 – T. turgidum.*

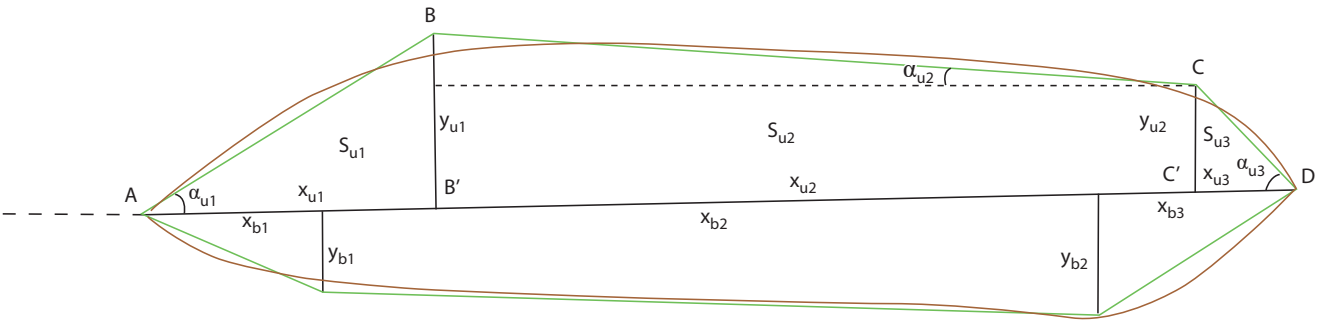


Fig. 2. Wheat spike shape represents in the form of two quadrangles (Genaev et al., 2019).

The black horizontal line shows the spike centerline. Brown line – spike contour. Green lines – the quadrilaterals that approximate the spike contour. The spike base – left dotted line. The figure for the upper quadrangle shows the main parameters that characterize spike geometry. Similar parameters are defined for the lower quadrangle.

upper quadrangle; the parameters x_{b1}, x_{b2}, y_{b1}, y_{b2} are for the lower quadrangle; the common parameter for the two quadrangles is the ear length. The program additionally calculates a number of general features of the shape and size of the spike, as well as the characteristic of its awning. Details of the feature extraction algorithm are given in (Genaev et al., 2019).

In the present study, we used the model traits that we selected as the most informative for predicting spike density index in our previous study (Genaev et al., 2019), as well as the general shape and spike trait characteristics. These traits characterise a complex view of the morphology (phenotype) of the spike by describing its shape (Circularity, Roundness), the physical dimensions of the ear body (Perimeter, Rachis length) and the area of the awns (Awns area), the traits obtained by approximating the ear by two quadrangles are related to the width (x_{u2}, y_{bm}) and length (x_{b2}, y_{u2}) of individual segments of the ear (Table 2).

Table 2. Description of the spike trait characters

Features	Description	Dimension
Awns area	Awns area	mm^2
Circularity	The roundness index is equal to the ratio of the perimeter of a circle with an area equal to the contour area to the perimeter of the contour. The index indicates how close the shape of the contour is to that of a circle. The value varies from 0 to 1	Dimensionless
Roundness	The roundness index is equal to the ratio of the contour area to the area of a circle with a diameter equal to the centreline of the spike	
Perimeter	Perimeter of ear contour without awns	mm
Rachis length	Length of the broken line along the axis of the compound spike (spike axis line)	
x_{u2}	Quadrangle model parameter related to the length of the left (the top one in Fig. 2) centre spike	
x_{b2}	Quadrangle model parameter related to the length of the right (the bottom one in Fig. 2) centre part of the spike	
y_{u2}	Distance of vertex C to its projection C′ on base AD (see Fig. 2)	
y_{bm}	Parameter of the quadrilateral model. Average value of the height of the right (bottom) quadrilateral	

Data analysis. In order to estimate the distribution of spikes in the feature space under study, we used a non-linear t-SNE dimensionality reduction algorithm (t-distributed stochastic neighbor embedding; Maaten, Hinton, 2008). This method allows to visualize multidimensional data by mapping objects in multidimensional space to a smaller (two- or three-dimensional) space. The basic idea behind t-SNE is to reduce the dimensionality of space while maintaining the relative pairwise distances between objects. The advantage of the t-SNE method is its tendency to localize isolated, dense spatial structures of arbitrary geometry. The t-SNE method was applied to ordinate images of spikes; the images of each of the projections of a single ear were treated as separate objects.

In order to assess the similarity of the quantitative characteristics of spikes for different species, we used hierarchical clustering (Johnson, 1967) of 17 wheat species according to the traits obtained by averaging over all spikes of the same species. Each species was characterized by a feature vector of length 9. A value of $1- r$ was used as a metric for the distance between species, where r is the value of the Pearson correlation coefficient between the values of the traits (Müllner, 2011). The linkage (UGMA algorithm) and dendrogram functions from the SciPy library (Virtanen et al., 2020) were used for clustering and dendrogram construction.

To compare the variance of traits in plants belonging to different ploidy types, we used F statistics (Snedecor, Cochran, 1989), which evaluates the significance of differences in the variance of two distributions. The data were normalized by the StandardScaler function of the scikit-learn library (Pedregosa et al., 2011). The test was performed independently for each of the nine traits described in Table 2. In this test, one spike image per plant was used, obtained in the "on the table" projection protocol.

Results and discussions

The mean, median, standard deviation and variance of the nine features calculated for the 17 wheat species are presented in the Supplementary 1[1].

Let's review the distribution of spikes in our sample of plants according to the characteristic "area of the spikes". The higher this parameter, the more awns were identified for the spike in the image. According to this characteristic, spikes of hexaploid wheat can be conditionally divided into three classes: awned (parameter value above 90), moderately awned (parameter value from 30 to 90), and awnless (parameter value below 30). The species *T. compactum*, *T. spelta*, *T. petropavlovskyi* and *T. vavilovii* are considered awned according to this criterion. *T. aestivum*, *T. yunnanense*, *T. macha* are moderately awned. The awnless ones are *T. antiquorum*, i:ANK-23, *T. spelta* cv. Rother Sommer Kolben, and *T. sphaerococcum* (see Supplementary 1). These data agree well with the appearance of the spikes (see Fig. 1, *a*). Thus, representatives of hexaploid wheat show considerable diversity in the presence/absence of awns.

If the classification above is applied to tetraploid wheat, only representatives of *T. militinae* (mean value of the parameter 24.09 mm^2) can be assigned to the awnless category. Four species can be classified as moderately awned: *T. dicoccoides*, *T. polonicum*, *T. timopheevii* and *T. turgidum*. Three species are considered awned: *T. aethiopicum, T. dicoccum*, *T. durum*. In general, the representation of awned species (specimens) in tetraploid species is significantly higher than in hexaploid species.

Analysis of such characteristic as spike length shows that spikes can also be divided into three classes: length less than 60 mm (short), from 60 to 90 mm (medium) and more than 90 mm (long). According to this classification, the hexaploid

[1] Supplementary materials 1–4 are available in the online version of the paper: http://vavilov.elpub.ru/jour/manager/files/SupplPronozinA_Engl.pdf

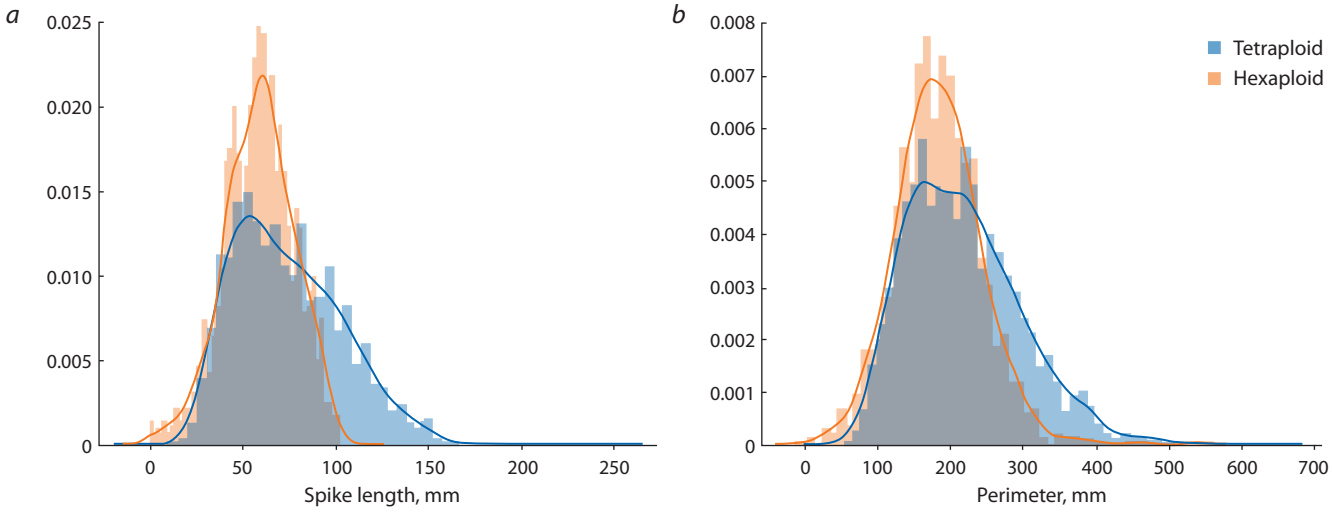


Fig. 3. Length (*a*) and perimeter (*b*) distribution of a wheat spike in tetraploid (blue) and hexaploid (orange) wheat species.

wheat species *T. spelta*, *T. petropavlovskyi* and *T. vavilovii* can be classified as long spikes, *T. aestivum*, *T. yunnanense*, *T. spelta* cv. Rother Sommer Kolben and *T. macha* to medium spikes, and *T. compactum*, *T. antiquorum*, *T. sphaerococcum* and the near-isogenic lineage ANK-23 to short spikes. The boundary between species characterized by long and medium spikes is rather conditional. For tetraploid species we did not find any species which according to this parameter would fall into the category of long-boned. The medium-sized category could include *T. aethiopicum*, *T. dicoccoides*, *T. polonicum*, *T. turgidum*, the short spike category – *T. dicoccum*, *T. durum*, *T. timopheevii* and *T. militinae*.

The spike length distribution of the samples studied for hexaploid and tetraploid species is shown in Fig. 3, *a*. The Fig. 3, *b* shows the distribution of the parameter also characterizing the size of the spikes – the perimeter of the contour of the body of the spike in the image.

Fig. 3 shows that the distributions of both parameters in hexaploid wheat are more scattered, while the variability of these traits in hexaploid wheat is higher mainly due to the higher frequency of occurrence of ears with high values of these traits.

The distribution of the analyzed ears images in the space of nine features was visualized using the t-SNE method, resulting in a two-dimensional parameter space (components 1 and 2). The results of the transformation are shown in Fig. 4. In the resulting diagram, each point represents one of the analysed images of the spike. In Fig. 4, *a* the dots are coloured according to the type of ploidy of the plant (blue colour corresponds to tetraploid wheat species, orange to hexaploid ones). In Fig. 4, *b* the colour and shape of each dot corresponds to a particular wheat spike image.

The diagram in Fig. 4, *a* shows that the areas occupied by hexa- and tetraploid wheat species strongly overlap on the graph. This means that the spikes of these two groups are quite similar in their characteristics. This is consistent with the results presented in the Supplementaries 1 and 2 as well as in Fig. 3. However, it should be noted that in the diagram in Fig. 4, *a* samples of hexaploid species occupy a larger area, primarily due to the predominance of the corresponding points in the right part of the diagram. One can see that orange dots (hexaploid wheat) predominate in the area with values of component 1 more than –20, this predominance is even more pronounced in the upper right corner of the diagram (values of component 1 less than 0 and component 2 more than 20). This means that a number of spike trait characteristics have some values for hexaploid species specific only, but not for tetraploid ones. This agrees well with the result shown in Fig. 3. In particular, such areas may correspond to large values of the parameters "perimeter" and "ear length".

The diagram in Fig. 4, *b* shows that the areas occupied by samples of the different species overlap considerably. For example, *T. aestivum* and *T. durum* species overlap across the entire plot area (dotted line). At the same time, it should be noted that the images of spikes belonging to the same wheat species occupy mostly compact areas on the graph. At the same time, there are species for which the spike samples are divided into several clearly visible clusters according to their characteristics. Such species include *T. compactum* (small blue circle, component 1 from –60 to 0, component 2 from –60 to 0) and *T. petropavlovskyi* (purple triangle, component 1 from –20 to 0, component 2 from 40 to 80).

Fig. 1 shows that hexaploids are represented by plants with two characteristic types of spikes: long and thin (*T. aestivum*, *T. spelta*, *T. petropavlovskyi*, *T. yunnanense*, *T. spelta* cv. Rother Sommer Kolben); short and rounded (*T. compactum*, *T. antiquorum*, i:ANK-23, *T. sphaerococcum*, *T. macha*, *T. vavilovii*). In Fig. 4. *b* the group of plants with short and rounded spikes is located in the component 2 value range from –80 to 0 (lower part of the graph). Plants with long and thin spikes have component 2 values between 0 and 80 (upper part of graph). In Fig. 4, *a*, these two groups of plants correspond roughly to the two clouds of dots in hexaploid

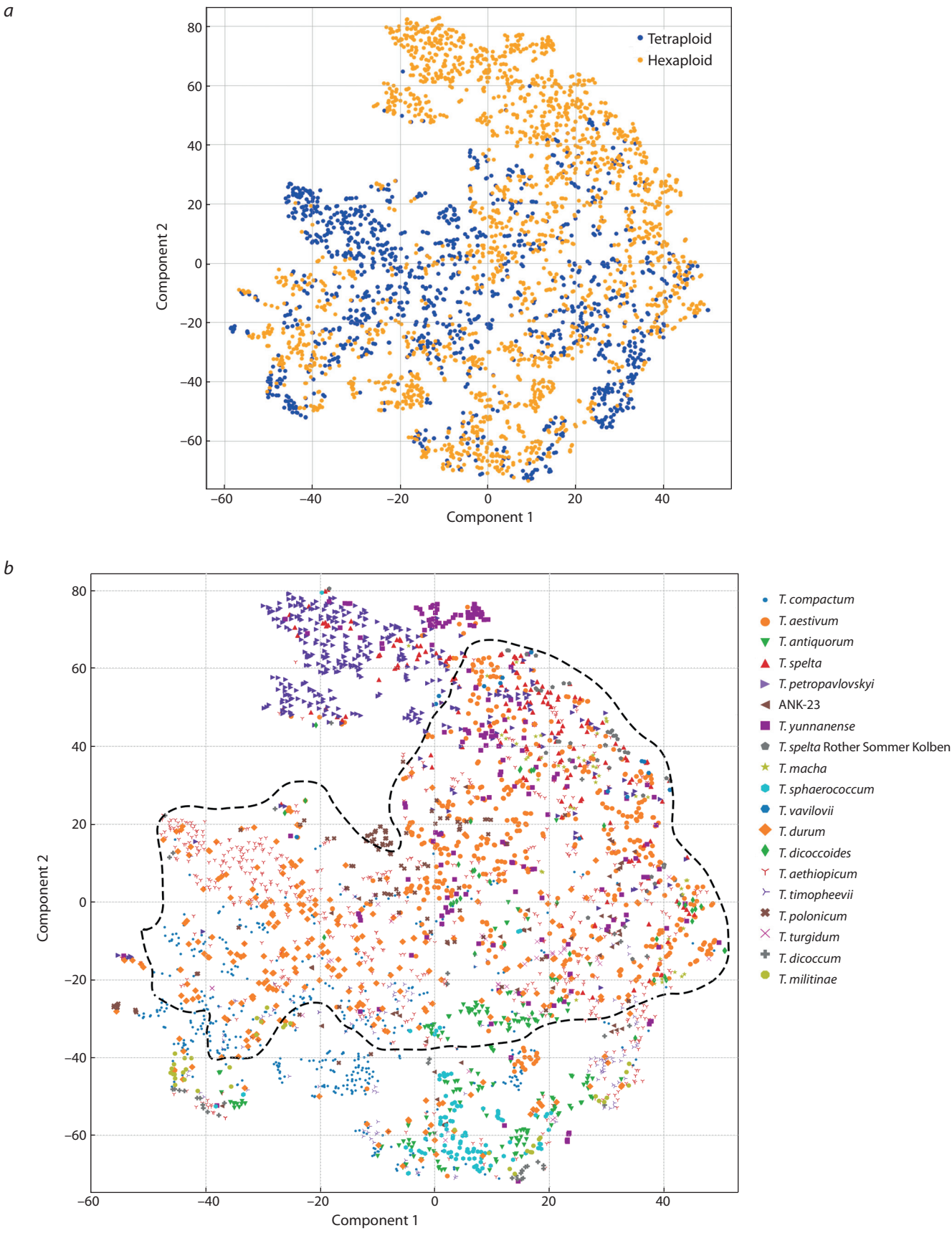


Fig. 4. Clustering of spike digital images of individual genotypes by the t-SNE method, obtained on the basis of quantitative traits from Table 2.

a – blue color corresponds to tetraploid wheat species, orange – hexaploid; *b* – the color and shape of each point corresponds to a specific type. The blue polygon marks the area occupied by *T. aestivum* and *T. durum* species. Clustering is called automatic partitioning into clusters. The automatic arrangement on the plane and in space is called ordination.

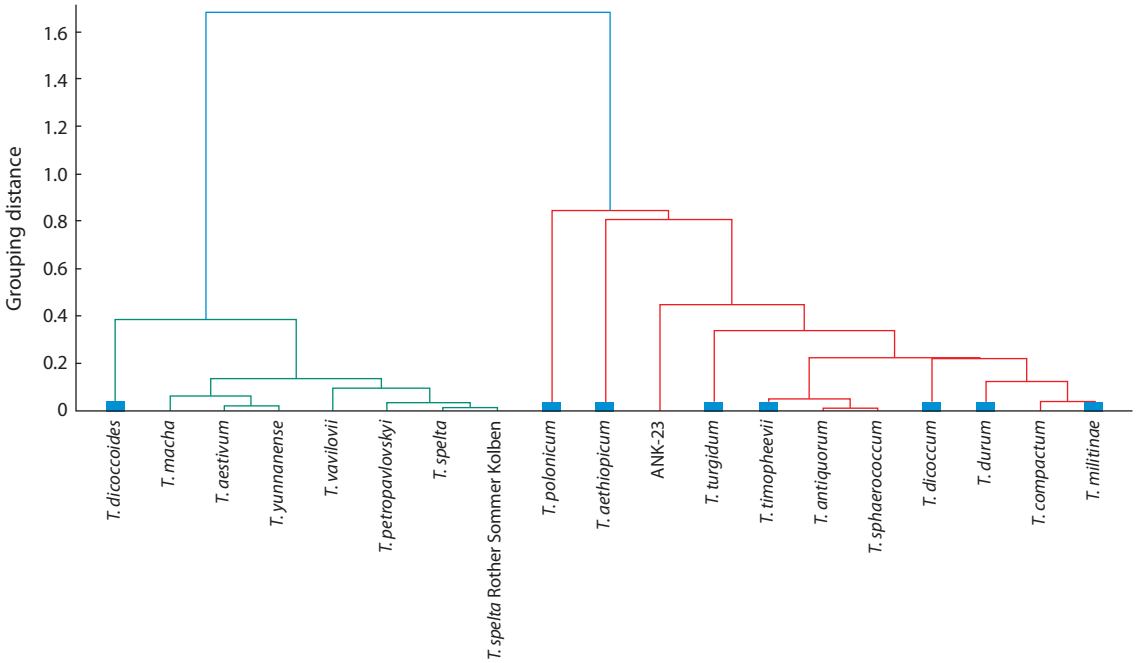


Fig. 5. Results of the hierarchical cluster analysis for nine signs of a wheat spike.
Blue squares correspond to tetraploid.

wheat at the top and bottom of the graph, which overlap slightly in the central part of the graph. Thus, the diagrams in Fig. 4 provide a clear indication of the diversity of spikes in their characteristics within and between species.

To characterize in more detail the similarity of morphometric characteristics of spikes in different wheat species, we conducted a hierarchical cluster analysis for them based on a comparison of the mean values of the studied traits. (Fig. 5).

Fig. 5 shows that the wheat species were divided into two clusters (highlighted in red and green). The first cluster (red) predominantly includes tetraploid species (shown in blue rectangles near the terminal tree nodes). However, wild tetraploid wheat species *T. diccocoides* is not included in this cluster, while among hexaploid species, *T. compactum*, *T. antiquorum* and *T. sphaerococcum* differing from all other species by compact spike shape, i. e. having the shortest spike of all studied hexaploid wheat species are included in it. It should be noted that in the work of A. Zatybekov et al. (2020), using economically important traits, samples of six tetraploid species were clustered arbitrarily, i. e. irrespective of their species identity. It is important to note, that remaining hexaploid species were clearly divided by spike length into two clusters of medium (*T. macha*, *T. aestivum*, and *T. yunnanense*) and long spikes (*T. vavilovii*, *T. petropavlovskyi*, and *T. spelta*).

T. spelta and *T. spelta* cv. Rother Sommer Kolben (a German landrace) occur in the same cluster. This allows us to conclude that the "species" shape of spike during long-term wheat breeding did not change for a long time (in this case, more than fifty years) and may be successfully used for classification of the species.

It should be noted that the only wild tetraploid loose spike species in the genus, *T. dicoccoides*, has fallen to the hexaploids. While hexaploid wheat species with compact ear type – *T. compactum*, *T. antiquorum*, *T. sphaerococcum* and human-made near-isogenic line ANK-23 of spring bread wheat cv. Novosibirskaya 67 (Koval, 1997) – were included into tetraploid species. The latter leads to the conclusion that although near-isogenic lines are produced on a particular (specific) species, nevertheless, their species identity should be treated with caution.

Let's take a look at *T. petropavlovskyi*. The species was founded at the Chinese Pamir – route of the Great silk road. According to the results of the study of gliadins, all accessions of this species were very similar to such hybrid combination obtained from crossing bread wheat with *T. polonicum* (Watanabe et al., 2004). The authors of the "Cultural Flora of the USSR" also considered a possible hybrid origin of this species (Dorofeev et al., 1979). In addition, *T. petropavlovskyi* also resembles bread wheat in a number of taxonomic traits (Goncharov, 2005). Previously, R.L. Boguslavsky (1982) described hybrids from crossing *T. aestivum* with *T. polonicum* produced by CIMMYT breeders as subspecies of *T. petropavlovskyi* ssp. *mexicana* Bogusl. Based on the above, we considered it appropriate to conbided *T. petropavlovskyi* as the subspecies of *T. aestivum*:

***Triticum aestivum* ssp. *petropavlovskyi* comb. et stat. nov. (Udacz. et Migusch.) N.P. Gontsch.** – *T. turanicum* Jakubz. convar. *montanostepposum* Jakubz. f. *aristiforme* Jakubz. 1959. *Bot. Zhur.* 10:1428, nom. illig. – *T. petropavlovskyi* Udacz. et Migusch. 1970. *Vestn. Sel'skokhoz. Nauki.* 9:20.

Table 3. Results of using *F* statistics to confirm the hypothesis of a significant difference in the variance of two distributions

Features	*F*-statistics	*p*-value	Hexaploid dispersion	Tetraploids dispersion	Average hexaploids	Average tetraploids
Awns area	0.376	1.000	1.415	3.763	84.875	160.643
Circularity index	1.188	0.065	0.959	0.807	0.178	0.181
Roundness	1.828	1.110e–07	**1.312**	0.718	0.141	0.172
Perimeter	1.570	4.710e–05	**1.080**	0.688	218.124	185.015
Rachis length	3.500	< 1e–15	**1.320**	0.377	74.136	59.280
x_{u2}	3.928	< 1e–15	**1.336**	0.340	53.837	36.853
x_{b2}	4.437	< 1e–15	**1.331**	0.300	54.004	36.726
y_{u2}	4.275	< 1e–15	**2.491**	0.583	3.844	4.171
y_{bm}	1.081	0.248	0.695	0.643	0.225	0.246

Note. Significant dispersion differences are shown in bold.

Typus: described by an accession from China "Origin: China, Xinjiang Province, village Kurlia, K-48376, 1957. A.M. Gorsky exp[edition]. Reproduction of Central Asia, Tashkent, Central Asian Station of VIR. 08. VII. 1969, Collected/defined: R.A. Udachin & E.F. Migushova" in St. Petersburg (WIR!). (The herbarium specimens of the type and paratype of *Triticum aestivum* ssp. *petropavlovskyi* are given in the Supplementaries 3 and 4).

Note that the results presented in Fig. 3 and 4, *a* show that hexaploid species have a greater variability in spike shape, size and awnness characteristics. Therefore, we hypothesized that the spike trait characteristics of hexaploid species may have a higher variation than those of tetraploid species. To test this assumption, we compared the variance of the estimated parameters using an *F*-distribution (Table 3).

The results presented in Table 3 show that the variance of most of the characters for hexaploids and tetraploids have significant differences ($p < 0.05$). At the same time, the significant differences in variance were not found for such traits as y_{bm} (quadrangle model parameter), Awns area and Circularity index. It is interesting to note that for all significant differences, we observe a higher variance in hexaploids than in tetraploids. Thus, the analysis showed that hexaploid species show higher diversity in spike morphometric trait characteristics compared to tetraploid species.

The data represent plants of 17 wheat species: 9 hexaploids (*T. compactum*, *T. aestivum*, *T. antiquorum*, *T. spelta* (including *T. spelta* cv. Rother Sommer Kolben), *T. petropavlovskyi*, i:ANK-23 (near-isogenic line of bread wheat cv. Novosibirskaya 67), *T. yunnanense*, *T. macha*, *T. sphaerococcum*, *T. vavilovii*) and 8 tetraploids (*T. aethiopicum*, *T. dicoccoides*, *T. dicoccum*, *T. durum*, *T. militinae*, *T. polonicum*, *T. timopheevii*, *T. turgidum*). The results of their clustering are presented so that the colour and shape of each dot corresponds to a particular species (see Fig. 5).

It is well known that genome doubling as a result of duplications (autopoploidy) or hybridization and subsequent polyploidization (alloploidy) leads to marked changes in plant phenotype (Finigan et al., 2012; Romanov, Pimonov, 2018; Rodionov et al., 2019). These changes in plants occur both at the cellular level (Liu et al., 2018) and at the organ level (Robinson et al., 2018). In many cases, in plants, an increase in ploidy leads to an increase in cell and organ size (Comai, 2005; Williams, Oliveira, 2020), increasing resistance to stress (Tan et al., 2015). Currently, researchers suggest that there are four types of molecular mechanisms of such variability: 1) increased gene/allele dosage, 2) increased genetic diversity, 3) altered genetic regulation, and 4) epigenetic rearrangements of the genome (Chen, 2007; Finigan et al., 2012).

The analysis of morphological characteristics of spikes of hexaploid ($2n = 6x = 42$) and tetraploid ($2n = 4x = 28$) wheat has shown, that most of spike characteristics have significantly higher variation in wheat with higher spike ploidy. Our results are in agreement with the ideas about the influence of ploidy on plant phenotype variability.

Conclusion

A large-scale analysis of the spike digital images of 595 plants of 8 tetra- and 9 hexaploid wheat species was carried out. Nine quantitative traits describing the shape, size and awnedness of the spike were studied. The variability among the above genotypes was studied and it was shown that two clusters are formed in the spike characteristic space. The first cluster includes mainly hexaploid species (with the exception of wild tetraploid species *T. dicoccoides*). The second cluster includes tetraploid species (with the exception of three hexaploid species with compact spike shape – *T. antiquorum*, *T. sphaerococcum*, and near-isogenic line ANK-23). Analysis of variance of these characters in hexaploid and tetraploid wheats showed a significant in-

crease in variance for six of nine characters in the sample of hexaploids, i. e. greater ploidy level gives more variability in quantitative characters of spike morphology.

Thus, it is shown that morphological trait characteristics of spikes of hexa- and tetraploid species, obtained on the basis of computer image analysis, demonstrate the differences, which can be used in the future to develop a method of classification of plants by ploidy level and their species affiliation in automatic mode.

References

Afonnikov D.A., Genaev M.A., Doroshkov A.V., Komyshev E.G., Pshenichnikova T.A. Methods of high-throughput plant phenotyping for large-scale breeding and genetic experiments. *Russ. J. Genet*. 2016;52(7):688-701. DOI 10.1134/S1022795416070024.

Boguslavsky R.L. A new botanical form of hexaploid wheat. *Nauchno-Tekhnicheskiy Byulleten VIR = Scientific and Technological Bulletin of the Vavilov Institute of Plant Industry.* 1982;119:73-74. (in Russian)

Chen Z.J. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* 2007;58:377-406. DOI 10.1146/annurev.arplant.58.032806.103835.

Comai L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet*. 2005;6(11):836-846. DOI 10.1038/nrg1711.

Demidchik V.V., Shashko A.Y., Bandarenka V.Y., Smolikova G.N., Przhevalskaya D.A., Charnysh M.A., Pozhvanov G.A., Barkovskyi A.V., Smolich I.I., Sokolik A.I., Yu M., Medvedev S.S., Plant phenomics: fundamental bases, software and hardware platforms, and machine learning. *Russ. J. Plant Physiol*. 2020;67:397-412. DOI 10.1134/S1021443720030061.

Dorofeev V.F. Intraspecific taxonomy of wheat. *Doklady VASKhNIL = Reports of the Academy of Agricultural Sciences.* 1985;9:1-4. (in Russian)

Dorofeev V.F., Filatenko A.A., Migushova E.F., Udachin R.A., Yakubtsiner M.M. Flora of Cultivated Plants of USSR. Vol. 1. Wheat. Leningrad: Kolos Publ., 1979. (in Russian)

Dorofeev V.F., Rudenko M.I., Filatenko A.A., Baras J., Segnalova J., Lemann H. (Compilators). The International Comecon List of Descriptors for the Genus *Triticum* L. Leningrad: VIR Publ., 1984. (in Russian)

Finigan P., Tanurdzic M., Martienssen R.A. Origins of novel phenotypic variation in polyploids. In: Polyploidy and Genome Evolution. Berlin; Heidelberg: Springer Press, 2012;57-76. DOI 10.1007/978-3-642-31442-1_4.

Genaev M.A., Komyshev E.G., Fu Hao, Koval V.S., Goncharov N.P., Afonnikov D.A. SpikeDroidDB: an information system for annotation of morphometric characteristics of wheat spike. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2018;22(1):132-140. DOI 10.18699/VJ18.340. (in Russian)

Genaev M.A., Komyshev E.G., Smirnov N.V., Kruchinina Y.V., Goncharov N.P., Afonnikov D.A. Morphometry of the wheat spike by analyzing 2D images. *Agronomy*. 2019;9(7):390. DOI 10.3390/agronomy9070390.

Golovnina K.A., Kondratenko E.Ya., Blinov A.G., Goncharov N.P. Phylogeny of the A genome of wild and cultivated wheat species. *Russ. J. Genet*. 2009;45(11):1360-1367. DOI 10.1134/S1022795409110106.

Goncharov N.P. Comparative-genetic analysis – a base for wheat taxonomy revision. *Czech J. Genet. Plant Breed*. 2005;41:52-55.

Goncharov N.P. Manual Book of Common and Hard Wheat Varieties. Novosibirsk: SO RAN Publ., 2009. (in Russian)

Goncharov N.P. Genus *Triticum* L. taxonomy: the present and the future. *Plant Syst. Evol*. 2011;295(1-4):1-11. DOI 10.1007/s00606-011-0480-9.

Goncharov N.P., Kondratenko E.Ya. Wheat origin, domestication and evolution. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders.* 2008;12(1/2):159-179. (in Russian)

Goncharov N.P., Shumny V.K. From preservation of genetic collections to organization of National project of plant gene pools conservation in permafrost. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders.* 2008;12(4):509-523. (in Russian)

Hammer K., Filatenko A.A., Pistrick K. Taxonomic remarks on *Triticum* L. and ×*Triticosecale* Wittm. *Genet. Resour. Crop Evol.* 2011;58(1):3-10. DOI 10.1007/s10722-010-9590-4.

Johnson S.C. Hierarchical clustering schemes. *Psychometrika*. 1967;32(3):241-254.

Koval S.F. The catalog of near-isogenic lines of Novosibirskaya-67 common wheat and principles of their use in experiments. *Russ. J. Genet*. 1997;33(8):995-1000.

Liu W., Zheng Y., Song S., Huo B., Li D., Wang J. *In vitro* induction of allohexaploid and resulting phenotypic variation in *Populus*. *Plant Cell Tiss. Organ Cult.* 2018;134(2):183-192. DOI 10.1007/s11240-018-1411-z.

Müllner D. Modern hierarchical, agglomerative clustering algorithms. *arXiv.* 2011;1109.2378.

Palmova E.F. Introduction to Wheat Ecology. Moscow; Leningrad: Selkhozgiz Publ., 1935. (in Russian)

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res*. 2011;12:2825-2830.

Robinson D.O., Coate J.E., Singh A., Hong L., Bush M., Doyle J.J., Roeder A.H. Ploidy and size at multiple scales in the *Arabidopsis* sepal. *Plant Cell.* 2018;30(10):2308-2329. DOI 10.1105/tpc.18.00344.

Rodionov A.V., Amosova A.V., Belyakov E.A., Zhurbenko P.M., Mikhailova Y.V., Punina E.O., Shneyer V.S., Loskutov I.G., Muravenko O.V. Genetic consequences of interspecific hybridization, its role in speciation and phenotypic diversity of plants. *Russ. J. Genet*. 2019;55(3):278-294. DOI 10.1134/S1022795419030141.

Rodionov A.V., Shneyer V.S., Gnutikov A.A., Nosov N.N., Punina E.O., Zhurbenko P.M., Loskutov I.G., Muravenko O.V. Species dialectics: from initial uniformity, through the greatest possible diversity to ultimate uniformity. *Botanicheskiy Zhurnal = Botanical Journal.* 2020;105(9):835-853. DOI 10.31857/S0006813620070091. (in Russian)

Romanov B.V., Pimonov K.I. Phenomogenomics of Production Traits of Wheat Species. Persianovsky: Donskoy GAU Publ., 2018. (in Russian)

Sinskaya E.N. Historical Geography of Cultural Flora (At the Dawn of Agriculture). Leningrad: Kolos Publ., 1969. (in Russian)

Snedecor G.W., Cochran W.G. Statistical Methods. Ames, Iowa: Iowa State University Press, 1989.

Tan F., Tu H., Liang W., Long J.M., Wu X.M., Zhang H.Y., Guo W.W. Comparative metabolic and transcriptional analysis of a doubled diploid and its diploid citrus rootstock (*C. junos* cv. Ziyang xiangcheng) suggests its potential value for stress resistance improvement. *BMC Plant Biol*. 2015;15:89. DOI 10.1186/s12870-015-0450-4.

Udachin R.A., Migushova E.F. New in the knowledge of the genus *Triticum*. *Vestnik Selskokhozyaystvennoy Nauki = Herald of Agricultural Sciences*. 1970;9:20-24. (in Russian)

van der Maaten L., Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res*. 2008;9:2579-2605.

van Slageren M., Payne T. Concepts and nomenclature of the Farro wheats, with special reference to Emmer, *Triticum turgidum* subsp.

dicoccum (Poaceae). *Kew Bull.* 2013;68:477-494. DOI 10.1007/S12225-013-9459-8.

Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S.J., Brett M., Jones E., Kern R., Larson E., Carey C.J., Polat I., Feng Yu, Moore E.W., VanderPlas J., Laxalde D., Perktold J., Cimrman R., Henriksen I., Quintero E.A., Harris C.R., Archibald A.M., Riberio A.H., Pedregosa F., van Mulbregt P. SciPy 1.0 Contributors. SciPy 1.0 – fundamental algorithms for scientific computing in Python. *Nat. Meth*. 2020;17(3):261-272. DOI 10.1038/s41592-019-0686-2.

Watanabe N., Bannikova S.V., Goncharov N.P. Inheritance and chromosomal location of the genes for long glume phenotype found in Portuguese landraces of hexaploid wheat, 'Arrancada'. *J. Genet. Breed*. 2004;58:273-278.

Williams J.H., Oliveira P.E. For things to stay the same, things must change: polyploidy and pollen tube growth rates. *Ann. Bot*. 2020; 125(6):925-935. DOI 10.1093/aob/mcaa007.

Yakubtsiner M.M. More on Chinese wheats. *Botanicheskiy Zhurnal = Botanical Journal*. 1959;44(10):1425-1436. (in Russian)

Yang W., Feng H., Zhang X., Zhang J., Doonan J.H., Batchelor W.D., Xiong L., Yan J. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant*. 2020;13(2):187-214. DOI 10.1016/j.molp.2020.01.008.

Zatybekov A., Anuarbek S., Abugalieva S., Turuspekov Y. Phenotypic and genetic variability of a tetraploid wheat collection grown in Kazakhstan. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2020;24(6):605-612. DOI 10.18699/VJ20.654.

Zhang Y., Zhao C., Du J., Guo X., Wen W., Gu S., Wang J., Fan J. Crop phenomics: current status and perspectives. *Front. Plant Sci.* 2019; 10:714. DOI 10.3389/fpls.2019.00714.

Zuev E.V., Amri A., Brykova A.N., Pyukkenen V.P., Mitrofanova O.P. Atlas of the Diversity of Soft Wheat (*Triticum aestivum* L.) by Ear and Grain Characteristics. St. Petersburg: Kopi-R Publ., 2019. (in Russian)

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
Y.V. Kruchinina orcid.org/0000-0002-1084-9521

Acknowledgements. Preparation of spike samples, phenotyping, development of algorithms for shape analysis and classification were funded by the Kurchatov Genome Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1662. Cultivation of experimental plants and *de visu* determination of their species belonging by the traits determining the architectonics of the spike were supported by grant of Russian Science Foundation 16-16-10021. The data were processed using the resources of the Bioinformatics Center supported by the budget project No. 0259-2021-0009. The authors A.A.P., E.A.Z., A.Yu.P., and N.M.P. were supported by the Mathematical Center in Akademgorodok, agreement with the Ministry of Science and Higher Education of the Russian Federation No. 075-15-2019-1675. The authors are grateful to D.A. Afonnikov for comments and recommendations during the work and I.G. Chukhina (VIR, St. Petersburg) for providing photos of herbarium specimens of *T. petropavlovskyi.*

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

Received October 27, 2020. Revised December 31, 2020. Accepted January 2, 2021.

Вавиловский журнал генетики и селекции. 2021;25(1):82-91
DOI 10.18699/VJ21.010

Original Russian text www.bionet.nsc.ru/vogis/

Sensitivity and identifiability analysis of COVID-19 pandemic models

O.I. Krivorotko[1, 2] ✉, S.I. Kabanikhin[1, 2], M.I. Sosnovskaya[2], D.V. Andornaya[2]

[1] Institute of Computational Mathematics and Mathematical Geophysics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ krivorotko.olya@mail.ru

Abstract. The paper presents the results of sensitivity-based identifiability analysis of the COVID-19 pandemic spread models in the Novosibirsk region using the systems of differential equations and mass balance law. The algorithm is built on the sensitivity matrix analysis using the methods of differential and linear algebra. It allows one to determine the parameters that are the least and most sensitive to data changes to build a regularization for solving an identification problem of the most accurate pandemic spread scenarios in the region. The performed analysis has demonstrated that the virus contagiousness is identifiable from the number of daily confirmed, critical and recovery cases. On the other hand, the predicted proportion of the admitted patients who require a ventilator and the mortality rate are determined much less consistently. It has been shown that building a more realistic forecast requires adding additional information about the process such as the number of daily hospital admissions. In our study, the problems of parameter identification using additional information about the number of daily confirmed, critical and mortality cases in the region were reduced to minimizing the corresponding misfit functions. The minimization problem was solved through the differential evolution method that is widely applied for stochastic global optimization. It has been demonstrated that a more general COVID-19 spread compartmental model consisting of seven ordinary differential equations describes the main trend of the spread and is sensitive to the peaks of confirmed cases but does not qualitatively describe small statistical datasets such as the number of daily critical cases or mortality that can lead to errors in forecasting. A more detailed agent-oriented model has been able to capture statistical data with additional noise to build scenarios of COVID-19 spread in the region.
Key words: parameter sensitivity; identifiability; ordinary differential equations; inverse problems; epidemiology; COVID-19; forecasting; Novosibirsk region.

For citation: Krivorotko O.I., Kabanikhin S.I., Sosnovskaya M.I., Andornaya D.V. Sensitivity and identifiability analysis of COVID-19 pandemic models. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):82-91. DOI 10.18699/VJ21.010

Анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19

О.И. Криворотько[1, 2] ✉, С.И. Кабанихин[1, 2], М.И. Сосновская[2], Д.В. Андорная[2]

[1] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ krivorotko.olya@mail.ru

Аннотация. Разработан алгоритм анализа чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19 в Новосибирской области, основанных на системах дифференциальных уравнений и законе действующих масс. Основу алгоритма составляет анализ матрицы чувствительности методами дифференциальной и линейной алгебры, показывающей степень зависимости неизвестных параметров моделей от заданных измерений. В результате работы алгоритма выявляются наименее и наиболее чувствительные к измерениям параметры, что способствует построению регуляризующего алгоритма решения задачи идентификации параметров для построения более точных сценариев развития эпидемии в регионе. Анализ чувствительности математических моделей

распространения коронавирусной инфекции COVID-19 показал, что параметр контагиозности вируса устойчиво определяется по количеству ежедневно выявляемых заболевших, критических и вылечившихся больных. С другой стороны, прогнозируемая доля госпитализированных больных, находящихся в критическом состоянии и требующих подключения аппарата ИВЛ, а также коэффициент смертности определяются гораздо менее устойчиво. Для построения более реалистичного прогноза необходимо добавить дополнительную информацию о процессе (например, о количестве ежедневных случаев госпитализации). Задачи уточнения идентифицируемых параметров по дополнительной информации о количестве выявленных, критических и смертельных случаев в Новосибирской области были сведены к задачам минимизации соответствующих целевых функционалов. Задача минимизации была решена с помощью метода дифференциальной эволюции, широко используемого в задачах стохастической глобальной оптимизации. Показано, что более общая камерная модель, состоящая из семи обыкновенных дифференциальных уравнений, описывает основную тенденцию распространения коронавирусной инфекции, чувствительна к пикам выявленных случаев, однако некачественно описывает небольшие статистики (количество ежедневных критических, смертельных случаев), что может приводить к ошибочным выводам. Более подробная агентно-ориентированная математическая модель, учитывающая поведение отдельных агентов, позволяет улавливать небольшие шумы в данных и строить сценарии развития распространения эпидемии в регионе.

Ключевые слова: чувствительность параметров; идентифицируемость; обыкновенные дифференциальные уравнения; обратные задачи; эпидемиология; COVID-19; прогнозирование; Новосибирская область.

Introduction

Many mathematical models in biology (epidemiology, immunology, pharmacokinetics, systems biology), medicine (tomography), physics and chemistry (meteorology, chemical kinetics), as well as sociology are described by systems of differential equations, whether they be ordinary (Kermack, McKendrick, 1927), partial (Habtemariam et al., 2008), or stochastic differential ones (Lee et al., 2020). Coefficients in these equations characterize specific features of simulated processes under given conditions. To build an adequate mathematical model, one needs to refine the coefficients of the equations based on the known parameters of the process and any additional information available about it. For example, when considering epidemiological problems, the parameters such as the infection transmission rate in the region; the critical case rates depending on comorbidities, age, and other demographic factors; the proportion of asymptomatic carriers/latent infection cases, etc., are unknown or approximately derived based on statistical data. These parameters are often sensitive to the measurements prone to errors (rounding, instrument, and human factor errors), which leads to unstable solutions of parameter identification problems.

Identifiability analysis of the differential equation systems modeling biological, medical, and physical processes is an important step to undertake before developing computational algorithms (Bellu et al., 2007; Raue et al., 2010, 2014; Miao et al., 2011; Kabanikhin et al., 2016; Voropaeva, Tsgoev, 2019). A classification of identifiability types distinguishing between structural identifiability, practical identifiability, and sensitivity analysis is presented in (Krivorotko et al., 2020a). The authors also analyze the systems of ordinary differential equations (ODE) describing epidemiological and immunological processes in terms of practical identifiability and parameter sensitivity to measurement errors.

A detailed review of methods and case studies of structural identifiability analysis in biological problems described by ODE systems may be found in (Miao et al., 2011; Kabanikhin et al., 2016). The model structure being as follows:

$$\begin{cases} \dfrac{d\bar{x}}{dt} = g(t,\bar{x},q), \quad t \in (0,T),\ x(t) \in C^1(\mathbb{R}^N),\ q \in \mathbb{R}^L, \\ \bar{x}(0) = \bar{x}_0, \\ x_i(t_k) = f_{ik}, \qquad i \in \{1,\ldots,M\},\ k = 1,\ldots,K, \end{cases} \tag{1}$$

The analysis based on this model makes it possible to verify the uniqueness of the solution q of the parameter identification problem and the initial conditions of model $\bar{x}_0$ (or their part) based on available measurements f_{ik}, while also providing recommendations on adding new information and modifying the conditions of the parameter identification problem to ensure the uniqueness of the solution.

In the present paper, the analysis of the semi-relative sensitivity of the mathematical models to describe epidemiological and social processes is presented. This approach, proposed in (Adams et al., 2004) for analyzing ODE systems, shows the degree of parameter sensitivity to measurements and identifies lacking/excessive measurements based on a certain reference set of parameters for solving the stated parameter identification problem. Two mathematical models of the spread of the new coronavirus infection caused by the SARS-CoV-2 virus described by ODE systems are used as examples. A regularization algorithm for numerical solution of the parameter identification problem is developed for SEIR compartmental model and agent-based model using the statistical data from public sources. The modeling results and the scenario of COVID-19 spread in the Novosibirsk region are presented.

Parameter sensitivity analysis in systems of ordinary differential equations

Sensitivity analysis is used for identifiability assessment of the unknown parameters of the model represented by ODE system (1) before developing a numerical solution algorithm

for the parameter identification problem. These methods do not require real experimental data, but the number of measurements and their time may be a necessity. Sensitivity analysis for a mathematical model is performed with regard to a set of nominal parameters q^*, whose values are taken from the literature or statistical data available.

Sensitivity analysis methods are based on a sensitivity matrix. Assume that $t_1 \le t_2 \le \ldots \le t_K$ are the fixed times of measurements f_{ik}. Then, the sensitivity matrix coefficients for parameter vector q^* are calculated as:

$$s_{ij}(t) = \frac{\partial f_i(t, q^*)}{\partial q_j} \cdot q_j^*, \qquad (2)$$

where, f_i, $i = 1, \ldots, M$, is the ith entry of the measurement function vector, and q_j, $j = 1, \ldots, L$, is the jth entry of the parameter vector.

Thus, the sensitivity matrix is determined as follows:

$$S_{M\cdot K\times L} = \begin{pmatrix} s_{11}(t_1) & \cdots & s_{1L}(t_1) \\ \vdots & \ddots & \vdots \\ s_{M1}(t_1) & \cdots & s_{ML}(t_1) \\ \vdots & \vdots & \vdots \\ s_{11}(t_K) & \cdots & s_{1L}(t_K) \\ \vdots & \ddots & \vdots \\ s_{M1}(t_K) & \cdots & s_{ML}(t_K) \end{pmatrix}.$$

The sensitivity matrix is calculated using the conventional sensitivity function:

$$s_{q_j}(t) = \frac{\partial x}{\partial q_j}(t), \qquad j = 1, \ldots, L.$$

When the first equation from (1) is differentiated with respect to q_j, each vector function s_{q_j} should satisfy the Cauchy problem as follows:

$$\begin{cases} \dot{s}_{q_j}(t) = \dfrac{\partial g}{\partial \bar{x}}(t, \bar{x}(t;q), q) \cdot s_{q_j}(t) + \dfrac{\partial g}{\partial q_j}(t, \bar{x}(t;q), q), \\ s_{q_j}(t_0) = \dfrac{\partial \bar{x}_0}{\partial q_j}. \end{cases} \qquad (3)$$

So, $s_{q_j}(t)$ is obtained by numerically solving the Cauchy problem.

First, the assessment is performed for the parameters q, to which the model's solution is most sensitive. These parameters, in turn, are defined by calculating semi-relative sensitivity. Here, sensitivity is considered a time function on the interval of interest. To obtain a general measure of parameter sensitivity of the solution, a time norm (over space L_2) is derived for each state/parameter combination and the obtained scalar quantities are ranked to identify the most sensitive parameters. The lower the value $\left\| \dfrac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$, the less the effect of q_k on f_i. This general measure will be referred to as semi-relative sensitivity.

The orthogonal method is then used for sensitivity analysis. The idea of the method suggested in (Yao et al., 2003) is to investigate linear dependencies of the columns of sensitivity matrix S. In such a way it will be possible to assess parameter sensitivity to the input data and parameter interdependence at the same time.

Sensitivity analysis of COVID-19 spread models

The feature of currently developed COVID-19 spread models is that they analyze the behavior of asymptomatic cases and the effect of the incubation period on the epidemiological situation in the regions. Several open-source suites (Gomez et al., 2020; Tuomisto et al., 2020; Wolfram, 2020) and web services have been developed for modeling COVID-19 spread scenarios:

- on a global scale: https://covid19-scenarios.org/ (University of Basel, Switzerland);
- in Moscow, the Novosibirsk region, and some European countries: https://covid19.biouml.org/ (Institute of Computational Technologies, SB RAS, Novosibirsk);
- in Almaty, the Republic of Kazakhstan: http://covid19.mmay.info/almaty/?fbclid=IwAR20yx7F4MdWRqwUDzripUK29lWAvoyCSkDPafgpj25ummay23e7oFHBdXg.

Two fundamental approaches to epidemic propagation modeling may be distinguished:

1. *Compartmental approach* (top-down modeling). Here, the interaction between the agents within the population grouped by similar attributes (susceptible group, (a)symptomatic carriers, hospital admission cases, critical cases, etc.) is described using the mass balance law within the compartmental model first suggested in 1927 (Kermack, McKendrick, 1927). Agents are distributed in time depending on the assigned transition probabilities between groups such as infection probability, virus contagiousness, mortality rate, etc.
2. *Agent-based approach* (bottom-up modeling) is based on studying the interactions between individuals and their effect on global parameters (e. g. virus contagiousness, mortality rate, severe case probability, etc.). Agent-based models are characterized by random graphs whose arc lengths describe the probabilities of transition to different agent states.

The parameters of transition between groups or agent states are often unknown or broadly defined. For instance, the incubation period of the disease according to WHO data varies from 2 to 14 days, which complicates the analysis of the model and the building of adequate disease spread scenarios.

Let us consider two breakdowns of the population of a particular age (e. g., ages 20–29) into groups. The transition into different agent states in the course of the disease caused by the SARS-CoV-2 virus is presented in Fig. 1. These models do not take into account such factors as sex segregation, annual birth and mortality rates (since the modeling interval of less than a year is analyzed), vaccination, passenger traffic, and comorbidities, which affect the probabilities of transition to different agent states. Our goal here is to demonstrate the correlation between dependences of similar parameters on the same measurements and recommend what parameters can be determined consistently and based on what measurements.

The ODE system (1) describing COVID-19 spread in the population is divided into 10 groups (Kerr et al., 2020) based on the mass balance law and is expressed as follows:

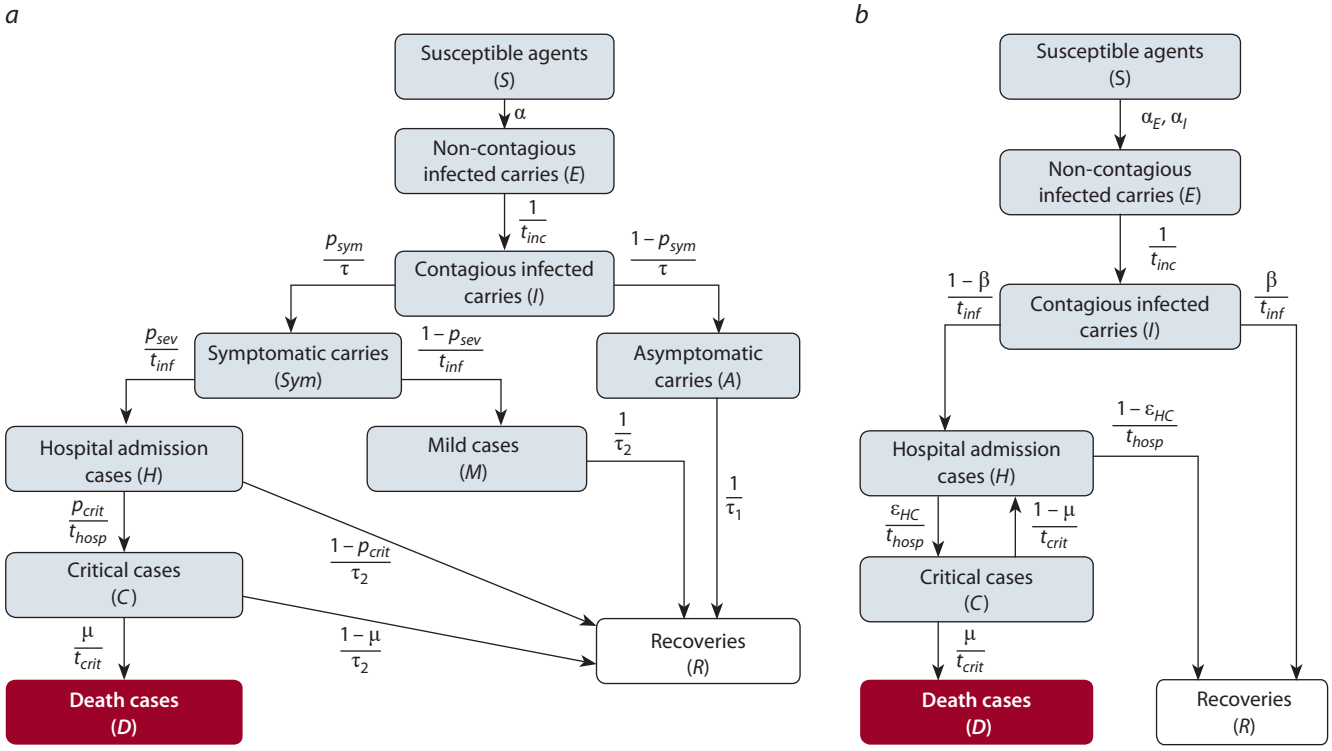


Fig. 1. Agent-state diagram in (*a*) the COVASIM package (Kerr et al., 2020) and (*b*) the SEIR-HCD model (Unlu et al., 2020).

$$
\begin{cases}
\frac{dS}{dt} = -\alpha S(t), \\
\frac{dE}{dt} = \alpha S(t) - \frac{1}{t_{inc}} E(t), \\
\frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{\tau} I(t), \\
\frac{dA}{dt} = \frac{1-p_{sym}}{\tau} I(t) - \frac{1}{\tau_1} A(t), \\
\frac{dSym}{dt} = \frac{p_{sym}}{\tau} I(t) - \frac{p_{sev}}{t_{inf}} Sym(t) - \frac{1-p_{sev}}{\tau} Sym(t), \\
\frac{dR}{dt} = \frac{1}{\tau_1} A(t) + \frac{1-p_{crit}}{\tau_2} H(t) + \frac{1}{\tau_1} M(t) + \frac{1-\mu}{\tau_2} C(t), \\
\frac{dH}{dt} = \frac{p_{sev}}{t_{inf}} Sym(t) - \frac{p_{crit}}{t_{hosp}} H(t) - \frac{1-p_{crit}}{\tau_2} H(t), \\
\frac{dM}{dt} = \frac{1-p_{sev}}{\tau} Sym(t) - \frac{1}{\tau_2} M(t), \\
\frac{dC}{dt} = \frac{p_{crit}}{t_{hosp}} H(t) - \frac{\mu}{t_{crit}} C(t) - \frac{1-\mu}{\tau_2} C(t), \\
\frac{dD}{dt} = \frac{\mu}{t_{crit}} C(t),
\end{cases}
\tag{4}
$$

with the initial conditions:

$$S(0) = S_0,\ E(0) = E_0,\ I(0) = I_0,\ A(0) = A_0,\ Sym(0) = Sym_0,$$

$$R(0) = R_0,\ H(0) = H_0,\ M(0) = M_0,\ C(0) = C_0,\ D(0) = D_0.$$

Model (4) characterizes a class of agent states for an age group within the agent-based model (see Fig. 1, *a*).

The equation system for the SEIR-HCD model, where the population is divided into 7 groups (Krivorotko et al., 2020b; Unlu et al., 2020), is composed in a similar fashion:

$$
\begin{cases}
\frac{dS}{dt} = -\frac{5-a(t-\tau)}{5}\left(\frac{\alpha_I S(t) I(t)}{N(t)} + \frac{\alpha_E S(t) E(t)}{N(t)}\right), \\
\frac{dE}{dt} = \frac{5-a(t-\tau)}{5}\left(\frac{\alpha_I S(t) I(t)}{N(t)} + \frac{\alpha_E S(t) E(t)}{N(t)}\right) - \frac{1}{t_{inc}} E(t), \\
\frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{t_{inf}} I(t), \\
\frac{dR}{dt} = \frac{\beta}{t_{inf}} I(t) + \frac{1-\varepsilon_{HC}}{t_{hosp}} H(t), \\
\frac{dH}{dt} = \frac{1-\beta}{t_{inf}} I(t) + \frac{1-\mu}{t_{crit}} C(t) - \frac{1}{t_{hosp}} H(t), \\
\frac{dC}{dt} = \frac{\varepsilon_{HC}}{t_{hosp}} H(t) - \frac{1}{t_{crit}} C(t), \\
\frac{dD}{dt} = \frac{\mu}{t_{crit}} C(t),
\end{cases}
\tag{5}
$$

with the initial conditions:

$$S(0) = S_0,\ E(0) = E_0,\ I(0) = I_0,\ R(0) = R_0,$$

$$H(0) = H_0,\ C(0) = C_0,\ D(0) = D_0.$$

Here, $S(t)$ is a susceptible agent at time t, $E(t)$ – a noncontagious infected (not transmitting the virus), $I(t)$ – a contagious infected (transmitting the virus), $A(t)$ – an asymptomatic case, $Sym(t)$ – a symptomatic case, $H(t)$ – a severe case, $C(t)$ – a critical case (requiring a ventilator), $M(t)$ – a mild case, $R(t)$ – a recovered case, and $D(t)$ – a mortality case. The averaged parameters of models (4) and (5) for the Novosibirsk region are presented in Table 1 (Lauer et al., 2020; Verity et al., 2020; Wölfel et al., 2020).

Table 1. Averaged parameters used in models (4), (5) for the Novosibirsk region (Kerr et al., 2020; Unlu et al., 2020)

Parameter	Description	Value/interval
τ_1	Days before recovery for asymptomatic and mild cases	6–11
τ_2	Days before recovery for severe and critical cases	12–17
p_{sym}	Probability of symptom expression after infection	0.6
p_{sev}	Probability of a symptomatic case becoming a severe one (requiring hospital admission)	0.0072
p_{crit}	Probability of a severe case becoming a critical one (requiring a ventilator)	0.00036
α	Probability of virus transmission to contact agents	(0.01, 0.025)
$E(0)$	Initial quantity of infected agents in the population	(1, 100)
$a(t)$	Yandex self-isolation index describing the compliance with self-isolation principles in the population on a scale from 0 (lack of isolation) to 5 (complete isolation)	(0, 5)
α_E	Contagion probability between the asymptomatic and susceptible groups ($\alpha_E \gg \alpha_I$)	(0, 1)
α_I	Contagion probability between the infected and susceptible groups linked with virus contagiousness and social factors	(0, 1)
β	Recovery probability for infected cases without complications	(0, 1)
ε_{HC}	Proportion of admitted critical-state patients requiring a ventilator	(0, 1)
μ	Probability of COVID-19 related mortality	(0, 0.5)
τ	Latent period in days (delay in virion isolation)	2
t_{inc}	Incubation period in days	2–14
t_{inf}	Infection period in days	2.5–14
t_{hosp}	Hospitalization period in days	4–5
t_{crit}	Days on a ventilator	10–20

Note that the coefficients t_{inc}^{-1}, t_{inf}^{-1}, t_{hosp}^{-1}, t_{crit}^{-1}, τ^{-1}, τ_1^{-1}, τ_2^{-1} at the respective agent states in models (4) and (5) describe the delay in transition between states (Likhoshvai et al., 2004). Consider the following equation (5):

$$\frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{t_{inf}} I(t),$$

where coefficient t_{inc}^{-1} (in the linear approximation) indicates the delay of t_{inc} days before the transition from non-contagious infected group $E(t)$ to contagious infected group $I(t)$, and coefficient $-t_{inf}^{-1}$ – that the agent stays in the contagious infected group for the infection period of t_{inf} days.

Mathematical model 1 (see the diagram in Fig. 1, *a*). Assume that additional information on recoveries and deaths on fixed days is available for the mathematical model (4):

$$R(t_k) = R_k,\ D(t_k) = D_k,\ k = 1, \ldots, 225. \qquad (6)$$

Here, R_k is the number of recovered agents on day k, D_k is the number of disease-related deaths on day k.

The model analyzes the semi-relative sensitivity of two unknown parameters, i. e. contagiousness α and initial quantity of asymptomatic cases $E(0)$ in the model (4) to the measurements (6). It will allow us to determine the possibility of consistent identification of the unknown parameters based on the data available for an adequate representation of epidemiological situation in the region. The sensitivities of parameters $(q_k) = (\alpha, E(0))$, $k = 1, 2$, to measurements $(f_i) = (R, D)$, $i = 1, 2$, represented by norm $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$ and sorted in descending order are presented in Table 2. The lower the value $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$, the less the effect of q_k on f_i.

Table 2. Semi-relative sensitivities of various model (4) states to the parameters sorted in descending order

Variable, f_i	Parameter, q_k	$\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$
$R(t)$	$E(0)$	$8.9 \cdot 10^6$
	α	$7.6 \cdot 10^{14}$
$D(t)$	$E(0)$	$6.7 \cdot 10^{-14}$
	α	$6.07 \cdot 10^{-6}$

Figure 2 demonstrates how sensitive function $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ changes in time depending on the parameter. Thus, parameters α and $E(0)$ within the model (4) are less sensitive to variable $D(t)$ and are therefore not identifiable by the mortality data alone. On the other hand, these parameters are sensitive to function $R(t)$, and, as a result, are recovered more consistently based on the recovery data.

Mathematical model 2 (see the diagram in Fig. 1, *b*). Let us now investigate the SEIR-HCD mathematical model (5). Assume that additional information on diagnoses, critical cases, and mortality on fixed days is available:

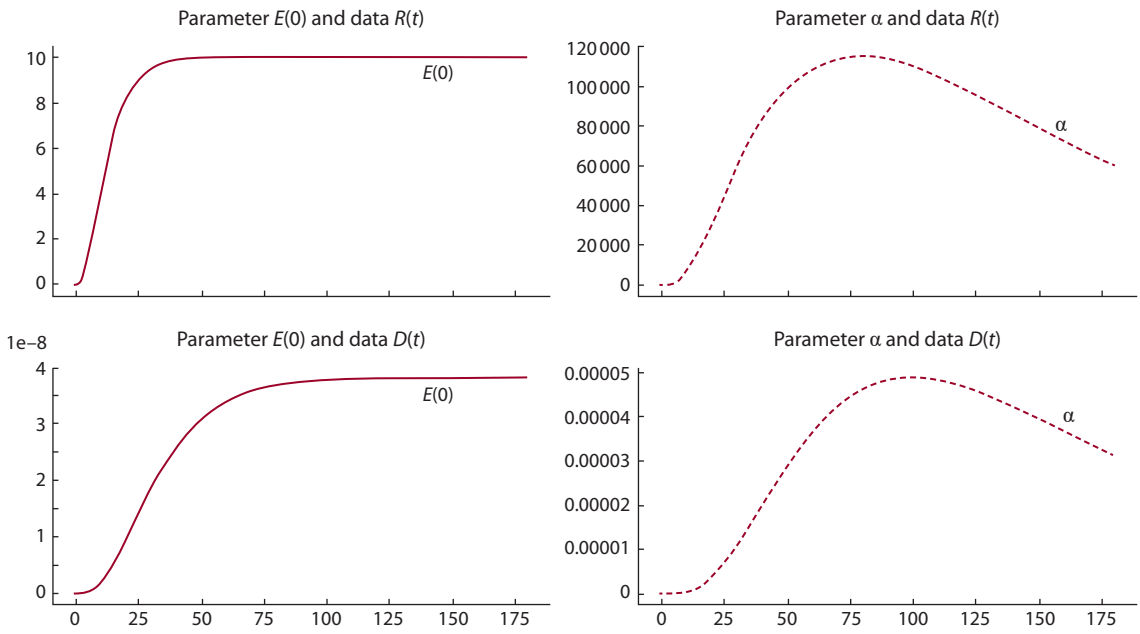


Fig. 2. Sensitive function $\frac{\partial f_i(t)}{\partial q_k} q_k$ for model (4) for the period from 12.03.2020 to 09.09.2020 (182 days).

$$I(t_k) = (1-b_k) f_k, \quad C(t_k) = C_k, \quad D(t_k) = D_k, \\ t_k \in (t_0, T), k = 1, \ldots, 205, \tag{7}$$

where $b(t) \in [0, 1]$ is the proportion of asymptomatic carriers in the diagnoses, f_k – the daily number of diagnoses on day k, C_k – the number of critical cases on day k.

Parameters $q = (\alpha_E, \alpha_I, \beta, \varepsilon_{HC}, \mu, E_0)^T \in \mathbb{R}^6$ are considered unknown. To analyze the semi-relative sensitivity of the parameter vector q to measurements (7) within the mathematical model (5), we derive $\frac{\partial f_i(t)}{\partial q_k} q_k^*$, $(f_i) = (I, C, D)$, $i = 1, 2, 3$, and analyze the values $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$ (Table 3). Consistency of identifying parameters β, ε_{HC}, and μ as the result of solving the inverse problem barely depends on the available measurements of the number of infected carriers $I(t)$; however, it is not the case for the more sensitive coefficients α_E, α_I, E_0.

Figure 3 shows how sensitive function $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ changes in time depending on the parameter. The more the parameter changes in time, the higher its sensitivity to the measurements analyzed, and the more consistently it is identified.

In Fig. 4, the results of parameter sensitivity analysis for the model (5) at various iterations of the orthogonal algorithm are presented (see the description of the algorithm in (Krivorotko et al., 2020a)). Iterations of the orthogonal algorithm, whose total number is one less than the dimension of the unknown parameter vector (i. e. the number of columns in the sensitivity matrix), are plotted along the horizontal axis, and the norms of perpendiculars for the obtained transformations of sensitivity matrices – along the vertical axis. It was shown that the contagion between the asymptomatic and susceptible groups α_E, the contagion between the infected and susceptible groups α_I linked with virus contagiousness and social factors, and the initial quantity of infected carriers and the agents in incubation period E_0 turned out to be the more identifiable parameters. The ranking of the parameters obtained via sensitivity analysis

Table 3. Semi-relative sensitivities of various states of the model (5) to parameters, sorted in descending order

Variable, f_i	Parameter, q_k	$\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$
I	α_E	$2.865 \cdot 10^{14}$
I	α_I	$2.396 \cdot 10^{14}$
I	E_0	$1.854 \cdot 10^{14}$
C	α_E	$2.386 \cdot 10^{12}$
C	α_I	$1.996 \cdot 10^{12}$
C	E_0	$1.544 \cdot 10^{12}$
D	α_E	$7.110 \cdot 10^{11}$
D	α_I	$5.947 \cdot 10^{11}$
D	E_0	$4.601 \cdot 10^{11}$
C	ε_{HC}	$4.833 \cdot 10^{4}$
C	β	$3.428 \cdot 10^{4}$
D	ε_{HC}	$3.041 \cdot 10^{4}$
D	μ	$2.982 \cdot 10^{4}$
D	β	$2.164 \cdot 10^{4}$
C	μ	$3.695 \cdot 10^{2}$
I	β	$2.03 \cdot 10^{-6}$
I	ε_{HC}	$1.6 \cdot 10^{-7}$
I	μ	$3 \cdot 10^{-8}$

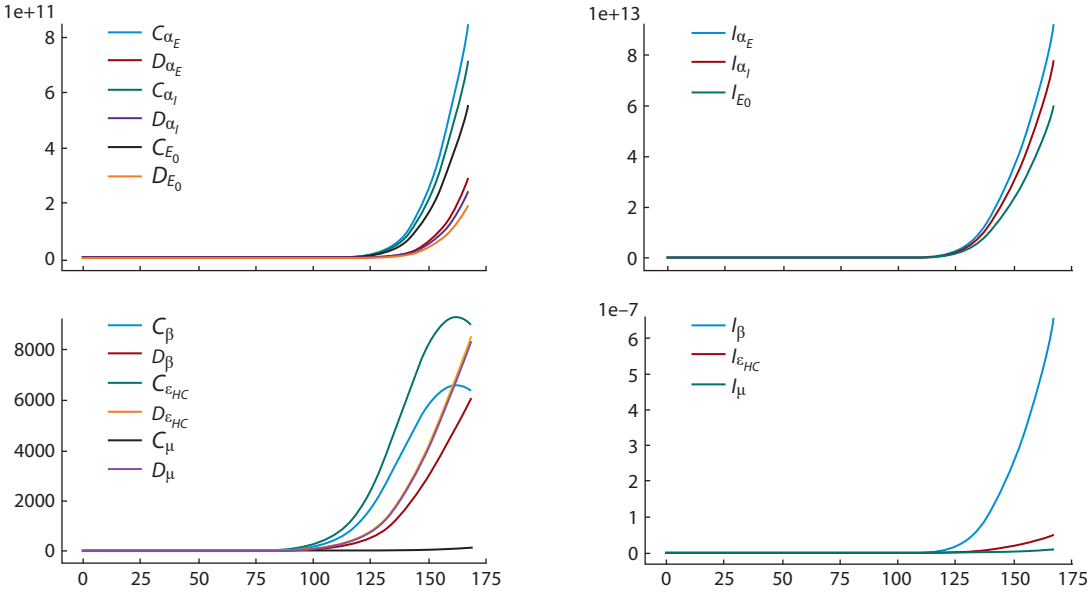


Fig. 3. Semi-relative sensitivity function $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ for the time interval from 15.04.2020 to 01.10.2020 (170 days).

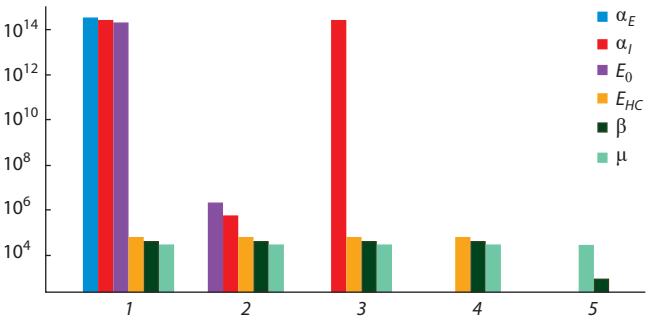


Fig. 4. Normalised perpendiculars for each parameter (different colors) at the different iterations (*1–5*) of the sensitivity-based orthogonal algorithm (5).

for the model (5) from the most sensitive to the least sensitive is as follows: α_E, E_0, α_I, ε_{HC}, μ, β.

As a result of identifiability analysis, a conclusion can be made that model parameters α_E, E_0, and α_I are the least sensitive to data variations (errors), i. e. are more identifiable. In other words, these parameters are identified more consistently as a result of solving the inverse problem (5), (7). In turn, parameters ε_{HC}, μ, and β are the most sensitive to measurement errors, i. e., less identifiable (and have the lowest values of the norms of perpendiculars in the sensitivity matrix). Hence, the regularization algorithm should be developed to ensure the consistent identification of sensitive parameters.

Mathematical modeling of COVID-19 spread in the Novosibirsk region

To build a COVID-19 spread model for the Novosibirsk region, the following publicly reported data were used:

(a) Number of people tested (including the number of diagnoses f and proportion of asymptomatic carriers $b(t)$), recovered cases (R), and COVID-19 related deaths (D);

(b) Duration of incubation period t_{inc}, latent period τ, infection period t_{inf}, hospitalization period t_{hosp}, and duration of ventilation t_{crit};

(c) Recovery time for mild τ_1 and severe τ_2 cases;

(d) Demographic profiles (population size and its age distribution in the region);

(e) Average household size (2.6 people) in the Russian Federation in 2019, according to UN data (https://population.un.org/Household/#/countries/840).

Additional information was regularly obtained from the following websites:

- Ministry of Health of the Novosibirsk region: https://zdrav.nso.ru/ (d).
- Federal State Statistics Service of the Novosibirsk region: https://novosibstat.gks.ru/folder/31729 (c).
- Stopcoronavirus website: https://стопкоронавирус.рф (a).
- World Health Organization: https://www.who.int (b).

The modeling was performed taking into account the measures to contain the COVID-19 spread (Table 4).

Solutions of inverse problems (4), (6) and (5), (7) were reduced to misfit function minimization (Kabanikhin, 2008):

$$J(q) = \sum_s \sum_{i=1}^{T} w_s \cdot G\left(c_d^{i,s}, c_m^{i,s}(q)\right).$$

Here, s is the statistics used for data comparison (cumulative diagnoses, critical cases, and mortality), w_s – the weight coefficient, $c_d^{i,s}$, $c_m^{i,s}$ – data (with subscript d) and model values (with subscript m), T – the modeling interval in days, q – the unknown parameter vector: $q_1 = (\beta, E_0)^T$ for the inverse problem (4), (6) and $q_2 = (\alpha_E(t), \alpha_I(t), \beta, \varepsilon_{HC}, \mu, E_0)^T$ for the inverse problem (5), (7). The absolute norm for computational experiments was set as follows:

$G_1 = \frac{|c_d^{i,s} - c_m^{i,s}|}{M}$, where $M = \max_t \{c_d^{t,s}\}$ was the normalization item; and the standard deviation was $G_2 = (c_d^{i,s} - c_m^{i,s})^2/T$.

Table 4. COVID-19 containment measures in the Novosibirsk region to be used in the models (4), (5)

Date	Containment measures
March 18, 2020	Introduction of distance learning in the schools and higher education institutions of the Novosibirsk region
March 28, 2020	Suspension of all mass gathering and public entertainment events in the region
April 27, 2020	Implementation of the face mask regime for shops by the Governor of the Novosibirsk region
July 6, 2020	Opening of outdoor terraces in cafes and restaurants
September 1, 2020	Reintroduction of contact learning in schools and higher education institutions
September 28, 2020	Implementation of the face mask regime for all premises and restricted containment measures in educational institutions

Minimization of misfit function $J(q)$ was implemented using the differential evolution method from the SciPy.Optimize Python library. The general algorithm of global minimum search was as follows:

1. Creation of the initial generation $\{\vec{q_i}\} \in B$, $i = 1 \dots N$..
2. Creation of a new generation:
   - Mutation:
     For all $\vec{q_i} \in B$ three random vectors were selected as follows: $\vec{v_1}, \vec{v_2}, \vec{v_3} \in B$, $(\vec{v_j} \neq \vec{q_i}, j = 1, 2, 3)$.
     Mutant vector: $\vec{v} = \vec{v_1} + F(\vec{v_2} - \vec{v_3})$, $F \in [0, 2]$.
   - Crossover: trial vector $\vec{u}$ was calculated as follows:

$$u_k = \begin{cases} v_k, \text{ if } rand < p, \\ q_k, \text{ if } rand \geq p, \end{cases} \qquad k = 1 \dots N_q.$$

3. Selection:

$$\vec{q_i} = \begin{cases} \vec{q_i}, \text{ if } J(\vec{x_i}) < J(\vec{u_i}), \\ \vec{u_i}, \text{ else.} \end{cases}$$

The results of COVID-19 spread modeling in the Novosibirsk region with the forecast up to December 10, 2020, are presented in Fig. 5. The model was built using the agent-based approach relying on the investigation of interactions between individuals and their effect on global parameters. The modeling was performed using Covasim, a simulator for developing stochastic agent-based models. A detailed discussion of the model structure may be found in (Kerr et al., 2020). We also used the statistical data on diagnoses and deaths from March 12 to October 23, 2020. The following misfit function was minimized taking into account the identifiability analysis results for the model (4), (6):

$$J(q_1) = \frac{1}{T}\sum_{i=1}^{T}(f_d^i - f_m^i)^2 + 100 \cdot (D_d^i - D_m^i)^2.$$

Here, f_d^i, f_m^i are cumulative diagnoses, and D_d^i, D_m^i are cumulative deaths.

Modeling results f_m^i and statistics f_d^i of cumulative and daily diagnoses are presented in Fig. 5, *a*, *b*. Modeling results D_m^i and statistics D_d^i of cumulative COVID-19 related deaths in the Novosibirsk region are presented in Fig. 5, *c*. Note that the second wave of the epidemic may be observed in the Novosibirsk region in mid-September in both the statistical data and modeling results. Its growth will be insignificant (i.e. it will not exceed 215 new daily diagnoses by mid-December,

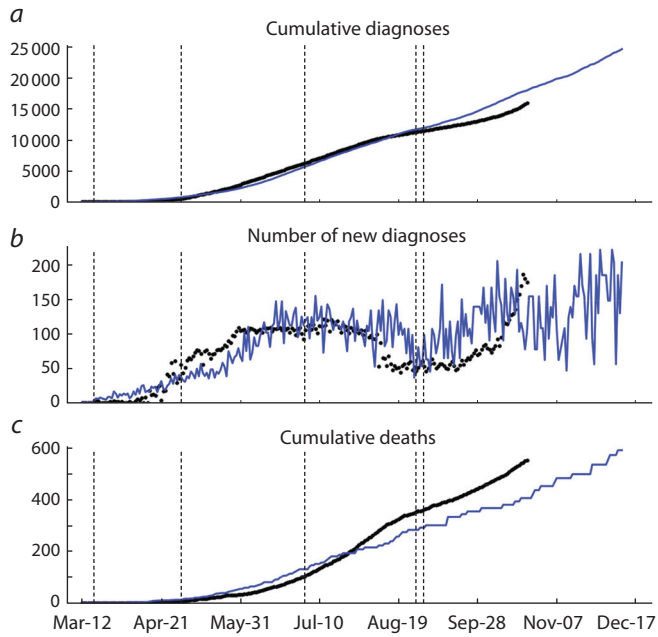


Fig. 5. COVID-19 spread model for the Novosibirsk region (solid blue line) using the agent-based approach and statistical data (black dots) with containment measures (vertical dashed lines).

2020) due to the introduction of stricter containment measures from October 28.

The inverse problem (5), (7) was reduced to the minimization of the following misfit function (Krivorotko et al., 2020b):

$$J(q_2) = \sum_{k=1}^{K}\left(w_1|t_{inc}^{-1}E(t_{k-1};q_2) - (1-b_k)f_k| + \right.$$
$$\left. + w_2|C(t_k;q_2) - C_k| + w_3|D(t_k;q_2) - D_k|\right).$$

Infection rate parameters $\alpha_E(t)$ and $\alpha_I(t)$ linked to virus contagiousness and varying in time were represented as piecewise constant functions depending on the interventions (see Table 4).

Based on the identifiability analysis results for the model (5), (7), more rigid restrictions were imposed for poorly identifiable parameters (see Table 1). The result of solving the inverse problem (5), (7) for the SEIR-HCD model for the period from April 15, 2020, to October 3, 2020, is presented in Fig. 6.

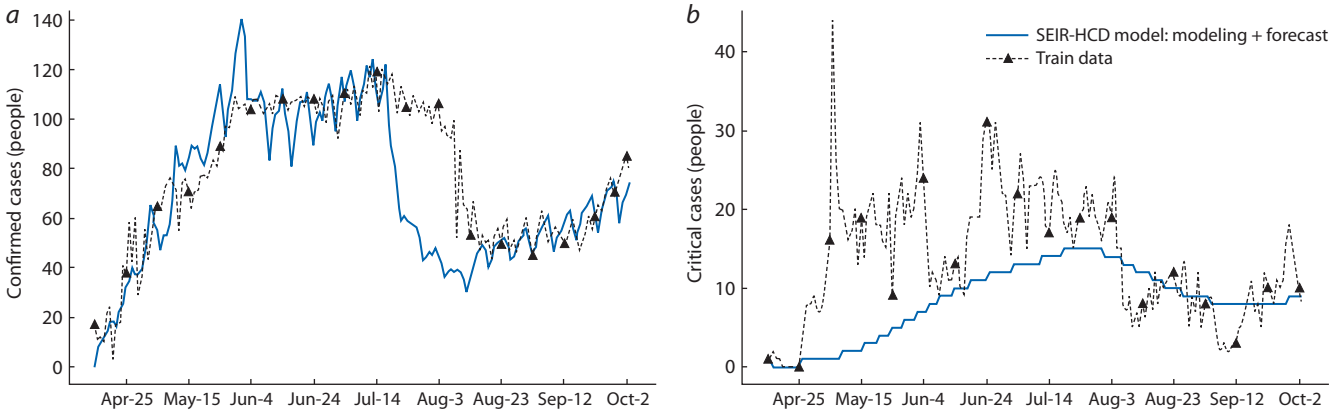


Fig. 6. Modeling COVID-19 spread in the Novosibirsk region (solid blue line) from 15.04.2020 to 03.10.2020 and the statistical data (dashed black line) for (*a*) daily confirmed cases f_k and (*b*) the critical cases C_k requiring a ventilator.

Note that although the rough mathematical model (with ODE system of 7 equations) captures the general trend based on the number of diagnoses (the peak of confirmed cases in the region, see Fig. 6, *a*), it is still unable to qualitatively describe highly variable statistics (critical cases requiring a ventilator, see Fig. 6, *b*). Nonsmooth solutions in Fig. 6 result from the use of the Yandex self-isolation index characterized by weekly seasonality. In this case, smoothing would undermine the use of the tool. A more detailed analysis of modeling and forecasting of COVID-19 spread in the Moscow and Novosibirsk regions is presented in (Krivorotko et al., 2020b). This case requires the agent-based approach capable of detailed description of small statistical datasets.

Conclusions

In the present study, sensitivity-based identifiability analysis has been performed for the COVID-19 pandemic spread models based on systems of differential equations. The algorithm is based on the analysis of the sensitivity matrix using the differential and linear algebra apparatus, which shows the degree of dependence of the unknown model parameters on the given measurements.

The analysis has shown that the virus contagiousness is consistently identified based on the number of new daily diagnoses, critical cases, and recoveries. On the other hand, the predicted proportion of admitted critical state patients requiring a ventilator and the mortality rate are identified much less consistently. It has been demonstrated that developing a more realistic forecast will require additional information about the process such as the number of daily hospital admissions.

The identifiable parameters refinement problems have been reduced to the minimization of the respective misfit functions describing the proximity of the modeling data to the statistics of the diagnoses, critical cases, and mortality in the Novosibirsk region. The use of absolute and quadratic norms as measures of deviation between the data and the modeling results in the minimization procedures has not yielded any significant differences in terms of analyzing the modeling results. It has been shown that a rough compartmental model of seven ODEs describes the general trend of the coronavirus infection propagation, as it is sensitive to peaks of confirmed cases; however, it is unable to qualitatively describe smaller statistics (daily numbers of critical cases t_k and deaths), which may lead to improper conclusions. A more detailed mathematical model using the agent-based approach, where a class of agent states is expressed by a system of ten ODEs, will make it possible to describe noisy statistical datasets in more detail and build adequate scenarios of COVID-19 pandemic spread.

References

Adams B.M., Banks H.T., Davidiana M., Kwona H.D., Trana H.T., Wynnea S.N., Rosenbergb E.S. HIV dynamics: modeling, data analysis, and optimal treatment protocols. *J. Comput. Appl. Math.* 2004; 184:10-49. DOI 10.1016/j.cam.2005.02.004.

Bellu G., Saccomani M.P., Audoly S., D'Angiò L. DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput. Methods Programs Biomed*. 2007;88(1):52-61. DOI 10.1016/j.cmpb.2007.07.002.

Gomez J., Prieto J., Leon E., Rodriguez A. INFEKTA: a general agent-based model for transmission of infectious diseases: studying the COVID-19 propagation in Bogotá – Colombia. *MedRxiv*. 2020. DOI 10.1101/2020.04.06.20056119.

Habtemariam T., Tameru B., Nganwa D., Beyene G., Ayanwale L., Robnett V. Epidemiologic modeling of HIV/AIDS: use of computational models to study the population dynamics of the disease to assess effective intervention strategies for decision-making. *Adv. Syst. Sci. Appl.* 2008;8(1):35-39.

Kabanikhin S.I. Definitions and examples of inverse and ill-posed problems. *J. Inverse Ill-Posed Probl*. 2008;16(4):317-357. DOI 10.1515/JIIP.2008.019.

Kabanikhin S.I., Voronov D.A., Grodz A.A., Krivorotko O.I. Identifiability of mathematical models in medical biology. *Russ. J. Genet. Appl. Res.* 2016;6(8):838-844. DOI 10.1134/S2079059716070054.

Kermack W.O., McKendrick A.G. A contribution of the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*. 1927;115:700-721. DOI 10.1098/rspa.1927.0118.

Kerr C., Stuart R., Mistry D., Abeysuriya R., Hart G., Rosenfeld K., Selvaraj P., Nunez R., Hagedorn B., George L., Izzo A., Palmer A., Delport D., Bennette C., Wagner B., Chang S., Cohen J., Panovska-Griffiths J., Jastrzebski M., Oron A., Wenger E., Famulare M., Klein D. Covasim: an agent-based model of COVID-19 dynamics and interventions. *MedRxiv.* 2020. DOI 10.1101/2020.05.10.20097469.

Krivorotko O.I., Andornaya D.V., Kabanikhin S.I. Sensitivity analysis and practical identifiability of some mathematical models in biology. *J. Appl. Ind. Math.* 2020a;14:115-130. DOI 10.1134/S1990478920010123.

Krivorotko O.I., Kabanikhin S.I., Zyat'kov N.Yu., Prikhod'ko A.Yu., Prokhoshin N.M., Shishlenin M.A. Mathematical modeling and forecasting of COVID-19 in Moscow and Novosibirsk region. *Numer. Analysis Applications*. 2020b;13(4):332-348. DOI 10.1134/S1995423920040047.

Lauer S.A., Grantz K.H., Bi Q., Jones F.K., Zheng Q., Meredith H., Azman A.S., Reich N.G., Lessler J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med*. 2020;172:577-582. DOI 10.7326/m20-0504.

Lee W., Liu S., Tembine H., Li W., Osher S. Controlling propagation of epidemics via mean-field games. *ArXiv*. 2020;arXiv:2006.01249.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. Modeling nonbranching multistage synthesis by an equation with retarded argument. *Sibirskiy Zhurnal Industrialnoy Matematiki = Journal of Applied and Industrial Mathematics*. 2004;7(1):73-94. (in Russian)

Miao H., Xia X., Perelson A.S., Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.* 2011;53(1):3-39. DOI 10.1137/090757009.

Raue A., Becker V., Klingmüller U., Timmer J. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*. 2010;20(4):045105. DOI 10.1063/1.3528102.

Raue A., Karlsson J., Saccomani M.P., Jirstrand M., Timmer J. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics.* 2014;30(10):1440-1448. DOI 10.1093/bioinformatics/btu006.

Tuomisto J.T., Yrjölä J., Kolehmainen M., Bonsdorff J., Pekkanen J., Tikkanen T. An agent-based epidemic model REINA for COVID-19 to identify destructive policies. *MedRxiv*. 2020. DOI 10.1101/2020.04.09.20047498.

Unlu E., Leger H., Motornyi O., Rukubayihunga A., Ishacian T., Chouiten M. Epidemic analysis of COVID-19 outbreak and counter-measures in France. *MedRxiv*. 2020. DOI 10.1101/2020.04.27.20079962.

Verity R., Okell L., Dorigatti I., Winskill P., Whittaker C., Imai N., Cuomo-Dannenburg G., Thompson H., Walker P., Fu H., Dighe A., Griffin J., Baguelin M., Bhatia S., Boonyasiri S., Cori A., Cucunubá Z., FitzJohn R., Gaythorpe K., Green W., Hamlet A., Hinsley W., Laydon D., Nedjati-Gilani G., Riley S., Elsland S., Volz E., Wang H., Wang Y., Xi X., Donnelly C., Ghani A., Ferguson N.M. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* 2020;20(6):669-677. DOI 10.1016/S1473-3099(20)30243-7.

Voropaeva O.F., Tsgoev Ch.A. A numerical model of inflammation dynamics in the core of myocardial infarction. *J. Appl. Ind. Math.* 2019;13(2):372-383. DOI 10.1134/S1990478919020182.

Wolfram C. An agent-based model of COVID-19. *Complex Syst*. 2020;29(1):87-105. DOI 10.25088/ComplexSystems.29.1.87.

Wölfel R., Corman V.M., Guggemos W., Seilmaier M., Zange S., Müller M.A., Niemeyer D., Jones T.C., Vollmar P.V., Rothe C., Hoelscher M., Bleicker T., Brünink S., Schneider J., Ehmann R., Zwirglmaier K., Drosten C., Wendtner C. Virological assessment of hospitalized patients with COVID-2019. *Nature*. 2020;581:465-469. DOI 10.1038/s41586-020-2196-x.

Yao K.Z., Shaw B.M., Kou B., McAuley K.B., Bacon D.W. Modeling ethylene/butene copoly-merization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Engineer*. 2003;11(3):563-588. DOI 10.1081/PRE-120024426.

ORCID ID

O.I. Krivorotko orcid.org/0000-0003-0125-4988

Acknowledgements. The work was supported by the Russian Foundation for Basic Research (project no. 18-31-20019) and the Council for Grants of the President of the Russian Federation (project no. 075-15-2019-1078 (MK-814.2019.1)).

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

Received October 25, 2020. Revised December 17, 2020. Accepted December 18, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):92-100
DOI 10.18699/VJ21.011

Original Russian text www.bionet.nsc.ru/vogis/

The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases

T.M. Khlebodarova[1, 2]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
tamara@bionet.nsc.ru

Abstract. The assumption that chronic mechanical stress in brain cells stemming from intracranial hypertension, arterial hypertension, or mechanical injury is a risk factor for neurodegenerative diseases was put forward in the 1990s and has since been supported. However, the molecular mechanisms that underlie the way from cell exposure to mechanical stress to disturbances in synaptic plasticity followed by changes in behavior, cognition, and memory are still poorly understood. Here we review (1) the current knowledge of molecular mechanisms regulating local translation and the actin cytoskeleton state at an activated synapse, where they play a key role in the formation of various sorts of synaptic plasticity and long-term memory, and (2) possible pathways of mechanical stress intervention. The roles of the mTOR (mammalian target of rapamycin) signaling pathway; the RNA-binding FMRP protein; the CYFIP1 protein, interacting with FMRP; the family of small GTPases; and the WAVE regulatory complex in the regulation of translation initiation and actin cytoskeleton rearrangements in dendritic spines of the activated synapse are discussed. Evidence is provided that chronic mechanical stress may result in aberrant activation of mTOR signaling and the WAVE regulatory complex via the YAP/TAZ system, the key sensor of mechanical signals, and influence the associated pathways regulating the formation of F actin filaments and the dendritic spine structure. These consequences may be a risk factor for various neurological conditions, including autistic spectrum disorders and epileptic encephalopathy. In further consideration of the role of the local translation system in the development of neuropsychic and neurodegenerative diseases, an original hypothesis was put forward that one of the possible causes of synaptopathies is impaired proteome stability associated with mTOR hyperactivity and formation of complex dynamic modes of *de novo* protein synthesis in response to synapse-stimulating factors, including chronic mechanical stress.
Key words: synapse; YAP/TAZ mechanosensor; mTOR; FMRP-dependent translation; complex dynamics; F actin; WAVE regulatory complex; autism spectrum disorders; epileptic encephalopathy.

For citation: Khlebodarova T.M. The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):92-100. DOI 10.18699/VJ21.011

Механический стресс клеток мозга, локальная трансляция и нейродегенеративные заболевания: молекулярно-генетические аспекты

Т.М. Хлебодарова[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
tamara@bionet.nsc.ru

Аннотация. Идея о том, что хронический механический стресс, который испытывают клетки мозга при повышенном внутричерепном давлении, артериальной гипертензии или вследствие травмы, может быть одним из факторов риска в развитии нейродегенеративных заболеваний, появилась еще в 90-е годы прошлого столетия и поддерживается в настоящее время. Однако молекулярно-генетические механизмы реализации событий, ведущих от механического воздействия на клетки к нарушению пластичности синапсов и последующему изменению поведения, когнитивных способностей и памяти, не ясны. В настоящем обзоре рассмотрены существующие данные о молекулярно-генетических механизмах регуляции локальной трансляции и актинового цитоскелета в активированном синапсе, играющих центральную роль в формировании различных видов пластичности синапса и долговременной памяти, и возможных путях влияния механического стресса на их состояние. Обсуждается роль mTOR сигнального каскада, РНК-связывающего белка FMRP, белка CYFIP1, взаимодействующего с FMRP, семейства малых ГТФаз и WAVE регуляторного комплекса

в регуляции инициации локальной трансляции и перестроек актинового цитоскелета в дендритных шипиках активированного синапса. Приводятся факты, свидетельствующие о том, что в условиях хронического механического стресса возможна аберрантная активация mTOR сигнального каскада и WAVE регуляторного комплекса через сенсор механических сигналов – регуляторный фактор YAP/TAZ, следствием которой могут быть нарушения активности системы локальной трансляции, а также связанных с ними механизмов регуляции формирования F-актиновых филаментов и структуры дендритных шипиков. Это может быть одной из причин развития различных неврологических патологий, включая аутистические расстройства и эпилептическую энцефалопатию. Высказывается оригинальная гипотеза, согласно которой одной из возможных причин синаптопатий может быть нарушение стабильности протеома, связанное с гиперактивностью mTOR и формированием сложных динамических режимов синтеза белков *de novo* в ответ на стимуляцию синапса, в том числе и в условиях хронического механического стресса.
Ключевые слова: синапс; механосенсор YAP/TAZ; mTOR; FMRP-зависимая трансляция; сложная динамика; F-актин; WAVE регуляторный комплекс; расстройства аутистического спектра; эпилептическая энцефалопатия.

Mechanical stress and neurodegenerative disorders

Mechanical signals are an important factor that determines the fate of cells, including their proliferation, survival, and differentiation, and takes part in tissue regeneration and wound healing. Mechanotransduction involves the reception of these forces and their conversion to biochemical and molecular signals, in particular, triggering of signaling pathways and expression of certain genes to allow cell adaptation to physical environment. There is ample evidence for the central role of the transcription regulator YAP (yes-associated protein 1) and its paralog TAZ (transcriptional co-activator with PDZ-binding motif), collectively named YAP/TAZ, as mechanical signal sensors and mediators (Dupont et al., 2011; Totaro et al., 2018; Dasgupta, McCollum, 2019). Impaired interaction of a cell and its environment causes aberrant YAP/TAZ activation and eventually a variety of diseases: atherosclerosis, fibrosis, lung hypertension, inflammation, muscle dystrophy, and cancer (Levy Nogueira et al., 2015, 2018; Yu et al., 2015; Panciera et al., 2017; Hong et al., 2019; Zhu et al., 2020). Recent studies indicate that mechanical stress may be among the causes of neurodegenerative processes in the brain, e. g., Alzheimer's disease (Levy Nogueira et al., 2015, 2016a, b, 2018).

The assumption that chronic mechanical stress experienced by brain cells exposed to intracranial hypertension, arterial hypertension, or mechanical injury is a risk factor for Alzheimer's disease and other neurodegenerative conditions was put forward as early as (Wostyn, 1994), and it is supported still now (Levy Nogueira et al., 2018).

What facts point to the existence of mechanisms by which mechanical stress influences nerve cell functions? First, it has been found that YAP/TAZ, being the key sensor and mediator of mechanical signals, activates the mTOR (mammalian target of rapamycin) signaling pathway (Tumaneng et al., 2012; McCarthy, 2013; Hu et al., 2017). This pathway is the central regulator of cap-dependent translation at a synapse. This regulation supports the dynamic plasticity of the synapse in response to external stimuli, and this plasticity underlies learning and memory (Costa-Mattioli et al., 2009; Buffington et al., 2014; Rosenberg et al., 2014; Santini et al., 2014). Disruption of these processes causes synaptic dysfunction and various neuropsychic disorders (Trifonova et al., 2017). YAP/TAZ activates mTOR by two mechanisms, illustrated in Fig. 1: by stimulating the transcription of Rheb GTPase (Ras homologue enriched in brain) (Hu et al., 2017), which is an mTORC1 kinase activator, and by inhibiting the translation of PTEN (phosphatase and tensin homolog) with miR29 microRNA, thereby inducing aberrant PI3K-mediated activation of mTORC1 and mTORC2 kinases (Tumaneng et al., 2012; McCarthy, 2013).

Second, the actin cytoskeleton is the key mediator of mechanical signals (Seo, Kim, 2018). Its rearrangements in dendritic spines contribute much to learning and long-term memory formation (Basu, Lamprecht, 2018; Borovac et al., 2018). They are controlled by Rho GTPases (Tapon, Hall, 1997), whose hypo- or hyperactivity results in dendritic spine structure distortion, defective memory, and poor learnability. It may also cause multiple neurodevelopment disorders of various origins (Ba et al., 2013; Pyronneau et al., 2017; Zamboni et al., 2018; Nishiyama, 2019). The function of Rho GTPases at an activated synapse depends considerably on their *de novo* synthesis, which is determined by mTOR activity (Briz et al., 2015).

The activation of the mTOR signaling pathway by the YAP/TAZ mechanosensor under mechanical stress (Tumaneng et al., 2012; McCarthy, 2013; Hu et al., 2017) also promotes the activation of the heteropentameric WAVE regulatory complex (WASP family verprolin homologue) via S6K kinase and RAC1 GTPase (Derivery et al., 2009) by inducing its breakdown into subcomplexes and interaction of WAVE1 with Arp2/3 (Cory, Ridley, 2002; Millard et al., 2004; Abekhoukh, Bardoni, 2014; Molinie, Gautreau, 2018). These processes result in aberrant actin polymerization and structural anomalies of dendritic spines (see Fig. 1).

Thus, the pathways of the influence of mechanical stimuli on nerve cell functioning may involve the activation of mTOR signaling and rearrangements of the actin cytoskeleton in dendritic spines, which, in turn, depend on the activity of the local translation system at the synapse, controlled by mTOR. Just the disturbances in the local translation system at the synapse, including those caused by enhanced mTOR activity (manifesting themselves as synapse plasticity aberrations in the form of imbalance between synapse excitation and inhibition (Gobert et al., 2020)) are thought to be associated with various neuropsychic conditions, including autism spectrum disorders (ASDs), epilepsy, Parkinson's disease,

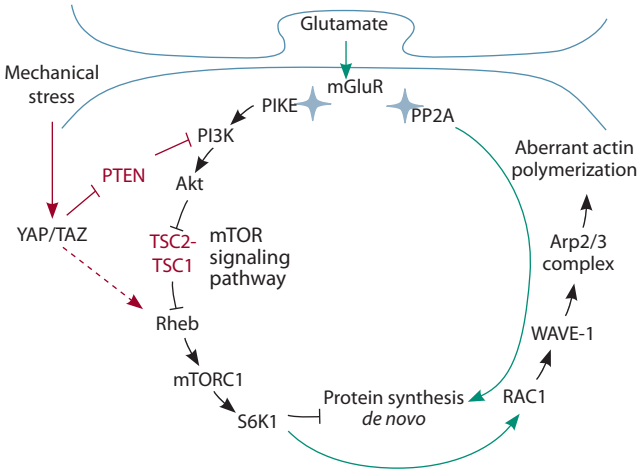


Fig. 1. Possible pathways of the effect of mechanical stress mediated by mTOR signaling on the intensity of local translation and the formation of actin cytoskeleton in dendritic spines of glutamatergic synapses in pyramidal cells of the hippocampus.

mGluR – receptor protein; PIKE (PI3-kinase enhancer), Rheb (Ras homologue enriched in brain), and Rac1 – GTPases; PI3K – phosphatidylinositol-3-kinase; Akt – protein kinase B; TSC1/2 – tuberous sclerosis complex 1/2; mTOR (mechanistic target of rapamycin) – serine/threonine kinase; S6K1 – S6 kinase 1; PTEN – phosphatase and tensin homolog; PP2A – protein phosphatase 2A; YAP/TAZ – mechanosensor; WAVE-1 – WAVE-1 regulatory complex; Arp2/3 – actin binding proteins. Proteins whose gene mutations are associated with neurological disorders are shown in red. Green arrows indicate translation activation via PP2A phosphatase and actin polymerization via S6 kinase and Rac1 GTPase in response to synapse stimulation by glutamate. Red arrows indicate possible mechanisms by which mechanical stress affects mTOR signaling.

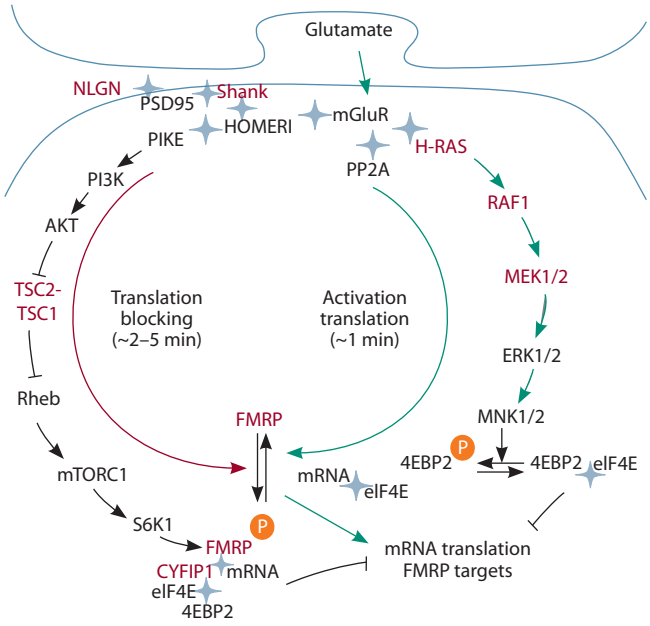


Fig. 2. Schematic presentation of local translation regulation in dendritic spines of glutamatergic synapses of hippocampal pyramidal cells in response to synapse stimulation.

mGluR – receptor protein; NLGNs, Shank, PSD95, HOMERI – postsynaptic proteins; PIKE (PI3-kinase enhancer) and Rheb (Ras homologue enriched in brain) – GTPases; PI3K – phosphatidylinositol-3-kinase; Akt – protein kinase B; S6K1 – S6 kinase 1; TSC1/2 – tuberous sclerosis complex 1/2; mTOR (mechanistic target of rapamycin) – serine/threonine kinase; FMRP (fragile X mental retardation protein) – RNA binding protein; PP2A – protein phosphatase 2A; H-RAS – GTPase; RAF1, MEK1/2, ERK1/2 and MNK1/2 – kinases; eIF4E – factor of translation initiation; 4EBP2 – 4E-binding protein; CYFIP1 – cytoplasmic FMRP interacting protein 1. The proteinaceous products of genes whose mutations are associated with neurological disorders are shown in red. Green arrows indicate pathways of local translation activation via PP2A phosphatase and the RAS/ERK signaling pathway. The red arrow indicates blocking via the mTOR signaling pathway.

and Alzheimer's disease (Gkogkas, Sonenberg, 2013; Meng et al., 2013; Won et al., 2013; Cai et al., 2015; Huber et al., 2015; Pramparo et al., 2015; Klein et al., 2016; Martin, 2016; Onore et al., 2017).

In this regard, the molecular mechanisms regulating local translation and dynamic rearrangements of the actin cytoskeleton in dendritic spines affected by this regulation enjoy close attention (Bramham, 2008).

Local translation and neurodegenerative disorders

There is convincing evidence that local cap-dependent translation in the postsynaptic space of a dendritic spine enables its dynamic plasticity in response to external stimuli, which underlies learning and memory (Huber et al., 2000; Costa-Mattioli et al., 2009; Rosenberg et al., 2014; Santini et al., 2014; Louros, Osterweil, 2016).

Numerous examples have been reported that impaired local translation control at a synapse brings forth various neuropsychic disorders, including ASDs, epilepsy, Parkinson's disease, etc. (Gkogkas, Sonenberg, 2013; Buffington et al., 2014; Klein et al., 2016; Martin, 2016; Trifonova et al., 2017). Figure 2 illustrates the main regulatory events mediating the activation of local protein synthesis in dendritic spines of glutamatergic synapses of hippocampal pyramidal cells in response to the stimulation of metabotropic glutamate receptors (mGluR) on the postsynaptic membrane of excitatory synapses by glutamate. The local translation activity is controlled by the mTOR and RAS/ERK pathways (Huber et al., 2000; Darnell, Klann, 2013; Beggs et al., 2015; Chen, Joseph, 2015).

The key element in the regulation of local cap-dependent translation at a synapse is the RNA-binding fragile X mental retardation protein, or FMRP (Feng et al., 1997). It is the target of S6 kinase and PP2A phosphatase, which are activated in response to the stimulation of mGluR receptors (Narayanan et al., 2007, 2008). When phosphorylated, it arrests translation by binding to mRNA, ribosomes, and the eIF4E translation factor (Brown et al., 1998; Napoli et al., 2008; Chen et al., 2014). Dephosphorylation disrupts FMRP linkage to its targets, resulting in, on the one hand, to accelerated mRNA translation and, on the other hand, rapid degradation of FMRP itself (Nalavadi et al., 2012). FMRP controls translation efficiency through RNA binding sites (Chen, Joseph, 2015). It directly binds to the coding and 3′-UTR mRNA sequences (Brown et al., 1998; Darnell et al., 2011) and to the L5 protein of 80S ribosomes (Chen et al., 2014). In this way, it controls transcription elongation and termination. Translation can also be repressed in 3′-UTR by physical interaction of FMRP with the 43-kDa TAR DNA-binding protein (TDP-43) (Majumder et al., 2016).

FMRP is also involved in translation regulation at its initiation step by interaction with the cytoplasmic FMRP-interacting protein 1 (CYFIP1) (Napoli et al., 2008). The current notions of mechanisms regulating translation by means of FMRP (Napoli et al., 2008; Majumder et al., 2016) presume the interaction of a single molecule of the protein with 3′-UTR via TDP-43 and with translation initiation factor eIF4E via CYFIP1. Thus, FMRP and CYFIP1 are the key regulators of translation regulation at an activated synapse.

FMRP targets are mRNAs for proteinaceous components of the mTOR signaling pathway (PI3K kinase, PTEN phosphatase, tuberous sclerosis complex 2 (TSC2 and mTOR), PP2A phosphatase, receptor proteins (mGluR, NMDAR, and AMPAR), proteins forming the postsynaptic membrane (NLGN, SHANK, and PSD95), the ubiquitin-dependent protein degradation system (E3 ubiquitin ligase), and its own mRNA (FMR1) (Brown et al., 1998; Muddashetty et al., 2007; Gross et al., 2010; Sharma et al., 2010; Darnell et al., 2011; Ascano et al., 2012). Apparently, FMRP plays the key role in dynamic proteome regulation at an activated synapse (Zukin et al., 2009; Iacoangeli, Tiedge, 2013).

It is known that mutations in genes encoding most of these proteins result in synapse malfunction and various disorders. Mutations in the gene for the SHANK3 protein of the postsynaptic membrane cause Phelan–McDermid syndrome; in the gene for PTEN phosphatase, Cowden's disease; for NF1, type 1 neurofibromatosis; in the genes for GTPase, H-RAS, RAF1, and MEK1 kinase, Costello and Noonan syndromes; TSC2-TSC1, tuberous sclerosis; FMRP, fragile X syndrome; UBE3A ubiquitin-protein ligase, Angelman syndrome; and in genes for neuroligins NLGN3/4 and neurexin NRXN1, typical autism (Trifonova et al., 2016). Mutations in the *Shank3* gene and its abnormal expression are also considered to cause autism, schizophrenia, and epilepsy (Peça et al., 2011; Mei et al., 2016; de Sena Cortabitarte et al., 2017; Monteiro, Feng, 2017; Fu et al., 2020). Mutations in the gene for PTEN phosphatase often bring forth various neurological diseases: macrocephaly, epilepsy, mental deficiency, and autism (Zhou, Parada, 2012; Trifonova et al., 2016).

These data suggest that synapse malfunctions are related to anomalies in local translation regulation. One of the possible synaptopathy causes is disturbed proteome stability, which hampers the formation of synapse plasticity and long-term memory (Cajigas et al., 2010). Indeed, just poor proteome stability is reported to be associated with autism and other neuropsychic disorders (Klein et al., 2016; Louros, Osterweil, 2016).

It should be mentioned that the structure-functional organization of the system regulating FMRP activity includes negative and positive feedback loops, which are instability factors in molecular systems (Mackey, Glass, 1977; Decroly, Goldbeter, 1982; Goldbeter et al., 2001; Bastos de Figueiredo et al., 2002; Likhoshvai et al., 2013, 2015, 2016, 2020; Kogai et al., 2015, 2017; Suzuki et al., 2016; Khlebodarova et al., 2017).

These regulatory loops act in different time spans. They are associated with rapid (ca. 1 min) translation activation of FMRP-dependent mRNAs via PP2A phosphatase and its rather rapid (2–5 min) arrest via the activation of S6 kinase (Narayanan et al., 2007, 2008). That is, the normal work of a synapse is supported by fine dynamic interplay among components of these signaling pathways at an activated synapse (see Fig. 2).

Analysis of dynamic features of the local translation system shows that an increase in the rate and efficiency of FMRP-dependent translation may induce instability in the local translation system, in particular, just in the physiological range of its operation (Khlebodarova et al., 2018; Likhoshvai, Khlebodarova, 2019). This result suggests that the known cases of ASDs related to the hyperactivity of the translation system at synapses (Pramparo et al., 2015; Onore et al., 2017) stem from proteome stability impairments associated with the formation of complex dynamic patterns of receptor protein synthesis in response to synapse stimulation (Khlebodarova et al., 2018, 2020). It is a brand-new insight into possible causes of synaptopathies.

It should be added that the elevated activity of mTOR signaling is a feature of not only ASDs but also other psychic and neurological diseases: Alzheimer's disease (Pei, Hugon, 2008), epilepsy (Wong, 2010), and even Down syndrome (Troca-Marin et al., 2012). It is also presumed that elevated mTOR activity causes early senescence and age-related neurodegenerative conditions in humans (Johnson et al., 2013).

In this regard, the hypothesis that the high copy numbers of rRNAs in some individuals are a risk factor for the development of ASDs, schizophrenia, and mental deficiency appears to be reasonable (Chestkov al., 2018; Porokhovnik, 2019; Porokhovnik, Lyapunova, 2019) on the assumption that individual variations in copy numbers of rRNA genes correlate with ribosome concentrations in a cell and the activity of the translational machinery.

The actin cytoskeleton and neurodegenerative diseases

The actin cytoskeleton structure determines the morphology of dendritic spines in nerve cells. Its rearrangements by rapid assembly of actin monomers (G actin) to filaments (F actin) and inverse disassembly are essential for the formation of synaptic plasticity and long-term memory (Penzes, Rafalovich, 2012; Basu, Lamprecht, 2018). Disturbances in the mechanisms regulating the formation of F actin filaments and dendritic spine structure are thought to be associated with neurodegenerative disorders: Alzheimer's disease, schizophrenia, and autism (Bamburg, Bernstein, 2016; Borovac et al., 2018; Forrest et al., 2018; Ben Zablah et al., 2020; Lauterborn et al., 2020). Fig. 3 illustrates the major regulatory events underlying actin cytoskeleton rearrangements in dendritic spines of glutamatergic synapses in hippocampal pyramidal cells, which are activated in response to the action of glutamate on metabotropic glutamate receptors mGluR and ionotropic glutamate receptors NMDAR (N-methyl-d-aspartate receptor) on the postsynaptic membrane of excitatory synapses.

The induction of actin filament formation and filament stabilization at an activated synapse depends substantially on the activity of cofilin and the WAVE regulatory complex, which is controlled by S6K, LIMK1, and PAK1 kinases via signaling pathways mediated by the RAS family of small

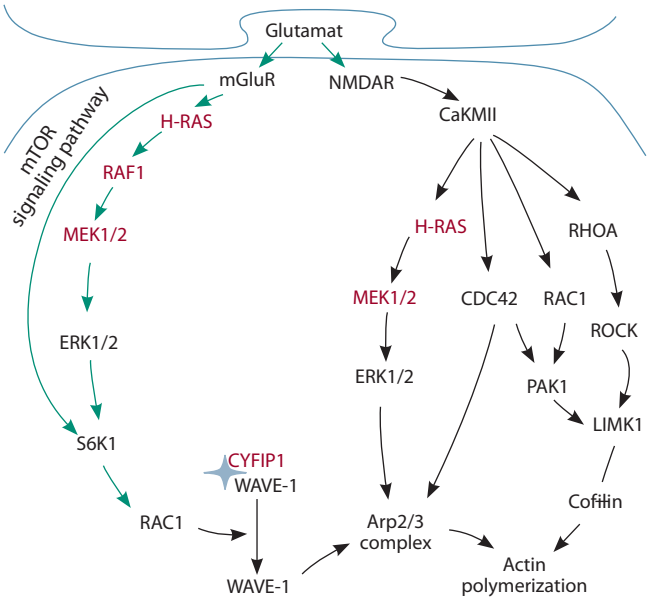


Fig. 3. Schematic presentation of the regulation of actin cytoskeleton formation in dendritic spines of glutamatergic synapses of hippocampal pyramidal cells in response to synapse stimulation.

mGluR, NMDAR – receptor proteins; H-RAS, RHOA, RAC1, CDC42 – RAS GTPases; CaKMII, MEK1/2, RAF1, ERK1/2, S6K1, PAK1, ROCK, LIMK1 – kinases; CYFIP1 – cytoplasmic FMRP interacting protein 1; WAVE-1 – WAVE-1 regulatory complex; Arp2/3 – actin binding proteins; Cofilin – actin depolymerization factor. The proteinaceous products of genes whose mutations are associated with neurological disorders are shown in red. Green arrows indicate the pathways of actin polymerization regulation via mTOR and S6K1, black – via CaKMII signaling.

GTPases: H-RAS, RhoA, Rac1, and Cdc42 (Tapon, Hall, 1997; Rex et al., 2009; Ip et al., 2011; Chen et al., 2017; Schaks et al., 2018). The operation of these signaling pathways at an activated synapse depends greatly on fast *de novo* synthesis of Rho GTPases (Briz et al., 2015). Arrest of protein synthesis in dendritic spines of hippocampal cells completely suppresses the stimulation of RhoA GTPase, cofilin phosphorylation, and actin polymerization (Briz et al., 2015). A mutation in the *Fmr1* gene, which encodes the FMRP protein, the key local transcription regulator, completely suppresses the physiological stimulation of GTPase Rac1 and its effector PAK1 kinase, disrupting the stabilization of actin filaments at hippocampal cell synapses (Chen et al., 2010).

Proceeding from the above, the activity of RAS GTPases controlling the formation and stabilization of actin filaments in dendritic spines depends directly on their *de novo* synthesis, i. e., on the activity of mTOR and FMRP-dependent local translation. It is conjectured that unstable local translation (Khlebodarova et al., 2018, 2020; Likhoshvai, Khlebodarova, 2019), also results in hypo- or hyperactivity of RAS, which, in turn, causes aberrations in the structure of dendritic spines and neurological disorders associated therewith (Ba et al., 2013; Pyronneau et al., 2017; Zamboni et al., 2018; Nishiyama, 2019).

Rac1 GTPase plays the key role in the regulation of the heteropentameric WAVE regulatory complex. The activity of this GTPase depends much on S6K and mTOR1 kinases. Normally, the WAVE complex is inactive, but its interaction with Rac GTPase induces its dissociation into two subcomplexes: CYFIP1-containing and WAVE1-containing (Derivery et al., 2009). The latter interacts with the Arp2/3 (actin-related proteins) complex and induces actin polymerization, as in Fig. 3 (Cory, Ridley, 2002; Millard et al., 2004; Abekhoukh, Bardoni, 2014; Molinie, Gautreau, 2018).

The disintegration of the WAVE complex and aberrant WAVE1 activation cause epileptic encephalopathy (Nakashima et al., 2018; Zhang et al., 2019; Zweier et al., 2019; Schaks et al., 2020). This may happen in cases of abnormal stoichiometric control of WAVE component synthesis (Abekhoukh et al., 2017) or mutations disrupting the interaction between WAVE1 and CYFIP2 (Nakashima et al., 2018; Zhang et al., 2019; Zweier et al., 2019; Schaks et al., 2020).

It should be noted that CYFIP1, being one of the main components of the WAVE regulatory complex, is also involved in translation regulation at its initiation step by interaction with the RNA-binding FMRP protein (Napoli et al., 2011). Thus, the mechanisms regulating local translation and actin cytoskeleton rearrangements in neural dendritic spines are additionally interlinked via the CYFIP1 protein (De Rubeis et al., 2013).

Conclusions

Analysis of presently available data shows the mechanisms regulating the local translation system at synapses and dynamic rearrangements of the actin cytoskeleton in dendritic spines of nerve cells, which play the central role in the formation of various types of synapse plasticity and long-term memory, are closely linked to each other and to the activity of the YAP/TAZ mechanosensor. This sensor can indirectly, via mTOR and S6K kinase, affect both translation efficiency and the state of actin filaments in dendritic spines (Tapon, Hall, 1997; Tumaneng et al., 2012; McCarthy, 2013; Reddy et al., 2013; Briz et al., 2015; Hu et al., 2017; Seo, Kim, 2018).

It is well substantiated that mTOR hyperactivity and functional aberrations in practically every component of the local translation system and of the machinery controlling rearrangements of the actin cytoskeleton in dendritic spines can cause numerous neurodevelopment disorders of various origins (Pei, Hugon, 2008; Wong, 2010; Johnson et al., 2013; Pramparo et al., 2015; Onore et al., 2017; Pyronneau et al., 2017; Trifonova et al., 2017; Nakashima et al., 2018; Nishiyama, 2019; Zhang et al., 2019).

Theoretical analysis of the dynamic features of local translation system operation presented in (Khlebodarova et al., 2018, 2020; Likhoshvai, Khlebodarova, 2019) suggests that one of the possible mechanisms of neurological disorders arising under chronic mechanical stress is the abnormal hyperactivity of mTOR and local translation at the synapse. It induces the dynamic instability of *de novo* protein synthesis at the activated synapse.

Thus, it is obvious that chronic mechanical stress may be one of the risk factors for synaptopathies and neurodegenerative diseases because of mTOR hyperactivation, which disturbs proteome stability, much needed for proper synapse plasticity and long-term memory (Klein et al., 2016; Louros, Osterweil, 2016).

References

Abekhoukh S., Bardoni B. CYFIP family proteins between autism and intellectual disability: links with Fragile X syndrome. *Front. Cell. Neurosci*. 2014;8:81. DOI 10.3389/fncel.2014.00081.

Abekhoukh S., Sahin H.B., Grossi M., Zongaro S., Maurin T., Madrigal I., Kazue-Sugioka D., Raas-Rothschild A., Doulazmi M., Carrera P., Stachon A., Scherer S., Drula Do Nascimento M.R., Trembleau A., Arroyo I., Szatmari P., Smith I.M., Milà M., Smith A.C., Giangrande A., Caillé I., Bardoni B. New insights into the regulatory function of CYFIP1 in the context of WAVE- and FMRP-containing complexes. *Dis. Model. Mech.* 2017;10(4):463-474. DOI 10.1242/dmm.025809.

Ascano M. Jr., Mukherjee N., Bandaru P., Miller J.B., Nusbaum J.D., Corcoran D.L., Langlois C., Munschauer M., Dewell S., Hafner M., Williams Z., Ohler U., Tuschl T. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*. 2012;492:382-386. DOI 10.1038/nature11737.

Ba W., van der Raadt J., Nadif Kasri N. Rho GTPase signaling at the synapse: implications for intellectual disability. *Exp. Cell Res*. 2013;319(15):2368-2374. DOI 10.1016/j.yexcr.2013.05.033.

Bamburg J.R., Bernstein B.W. Actin dynamics and cofilin-actin rods in Alzheimer disease. *Cytoskeleton* (*Hoboken*). 2016;73(9):477-497. DOI 10.1002/cm.21282.

Bastos de Figueiredo J.C., Diambra L., Glass L., Malta C.P. Chaos in two-looped negative feedback systems. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 2002;65:051905.

Basu S., Lamprecht R. The role of actin cytoskeleton in dendritic spines in the maintenance of long-term memory. *Front. Mol. Neurosci*. 2018;11:143. DOI 10.3389/fnmol.2018.00143.

Beggs J.E., Tian S., Jones G.G., Xie J., Iadevaia V., Jenei V., Thomas G., Proud C.G. The MAP kinase-interacting kinases regulate cell migration, vimentin expression and eIF4E/CYFIP1 binding. *Biochem J*. 2015;467(1):63-76. DOI 10.1042/BJ20141066.

Ben Zablah Y., Merovitch N., Jia Z. The role of ADF/cofilin in synaptic physiology and Alzheimer's disease. *Front. Cell Dev. Biol*. 2020;8:594998. DOI 10.3389/fcell.2020.594998.

Borovac J., Bosch M., Okamoto K. Regulation of actin dynamics during structural plasticity of dendritic spines: Signaling messengers and actin-binding proteins. *Mol. Cell. Neurosci.* 2018;91:122-130. DOI 10.1016/j.mcn.2018.07.001.

Bramham C.R. Local protein synthesis, actin dynamics, and LTP consolidation. *Curr. Opin. Neurobiol*. 2008;18(5):524-531. DOI 10.1016/j.conb.2008.09.013.

Briz V., Zhu G., Wang Y., Liu Y., Avetisyan M., Bi X., Baudry M. Activity-dependent rapid local RhoA synthesis is required for hippocampal synaptic plasticity. *J. Neurosci*. 2015;35(5):2269-2282. DOI 10.1523/JNEUROSCI.2302-14.2015.

Brown V., Small K., Lakkis L., Feng Y., Gunter C., Wilkinson K.D., Warren S.T. Purified recombinant Fmrp exhibits selective RNA binding as an intrinsic property of the fragile X mental retardation protein. *J. Biol. Chem.* 1998;273(25):15521-15527. DOI 10.1074/jbc.273.25.15521.

Buffington S.A., Huang W., Costa-Mattioli M. Translational control in synaptic plasticity and cognitive dysfunction. *Annu. Rev. Neurosci*. 2014;37:17-38. DOI 10.1146/annurev-neuro-071013-014100.

Cai Z., Chen G., He W., Xiao M., Yan L.J. Activation of mTOR: a culprit of Alzheimer's disease? *Neuropsychiatr. Dis. Treat*. 2015;11:1015-1030. DOI 10.2147/NDT.S75717.

Cajigas I.J., Will T., Schuman E.M. Protein homeostasis and synaptic plasticity. *EMBO J.* 2010;29:2746-2752. DOI 10.1038/emboj.2010.173.

Chen B., Chou H.T., Brautigam C.A., Xing W., Yang S., Henry L., Doolittle L.K., Walz T., Rosen M.K. Rac1 GTPase activates the WAVE regulatory complex through two distinct binding sites. *Elife.* 2017;6:e29795. DOI 10.7554/eLife.29795.

Chen E., Joseph S. Fragile X mental retardation protein: A paradigm for translational control by RNA-binding proteins. *Biochimie*. 2015;114:147-154. DOI 10.1016/j.biochi.2015.02.005.

Chen E., Sharma M.R., Shi X., Agrawal R.K., Joseph S. Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Mol. Cell.* 2014;54:407-417. DOI 10.1016/j.molcel.2014.03.023.

Chen L.Y., Rex C.S., Babayan A.H., Kramár E.A., Lynch G., Gall C.M., Lauterborn J.C. Physiological activation of synaptic Rac>PAK (p-21 activated kinase) signaling is defective in a mouse model of fragile X syndrome. *J. Neurosci.* 2010;30(33):10977-10984. DOI 10.1523/JNEUROSCI.1077-10.2010.

Chestkov I.V., Jestkova E.M., Ershova E.S., Golimbet V.E., Lezheiko T.V., Kolesina N.Y., Porokhovnik L.N., Lyapunova N.A., Izhevskaya V.L., Kutsev S.I., Veiko N., Kostyuk S.V. Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophr. Res*. 2018;197:305-314. DOI 10.1016/j.schres.2018.01.001.

Cory G.O.C., Ridley A.J. Cell motility: braking WAVEs. *Nature*. 2002;418:732-733. DOI 10.1038/418732a.

Costa-Mattioli M., Sossin W.S., Klann E., Sonenberg N. Translational control of long-lasting synaptic plasticity and memory. *Neuron*. 2009;61:10-26. DOI 10.1016/j.neuron.2008.10.055.

Darnell J.C., Klann E. The translation of translational control by FMRP: therapeutic targets for FXS. *Nat. Neurosci.* 2013;16(11):1530-1536. DOI 10.1038/nn.3379.

Darnell J.C., Van Driesche S.J., Zhang C., Hung K.Y., Mele A., Fraser C.E., Stone E.F., Chen C., Fak J.J., Chi S.W., Licatalosi D.D., Richter J.D., Darnell R.B. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*. 2011;146:247-261. DOI 10.1016/j.cell.2011.06.013.

Dasgupta I., McCollum D. Control of cellular responses to mechanical cues through YAP/TAZ regulation. *J. Biol. Chem*. 2019;294(46):17693-17706. DOI 10.1074/jbc.REV119.007963.

De Rubeis S., Pasciuto E., Li K.W., Fernández E., Di Marino D., Buzzi A., Ostroff L.E., Klann E., Zwartkruis F.J., Komiyama N.H., Grant S.G., Poujol C., Choquet D., Achsel T., Posthuma D., Smit A.B., Bagni C. CYFIP1 coordinates mRNA translation and cytoskeleton remodeling to ensure proper dendritic spine formation. *Neuron*. 2013;79(6):1169-1182. DOI 10.1016/j.neuron.2013.06.039.

de Sena Cortabitarte A., Degenhardt F., Strohmaier J., Lang M., Weiss B., Roeth R., Giegling I., Heilmann-Heimbach S., Hofmann A., Rujescu D., Fischer C., Rietschel M., Nöthen M.M., Rappold G.A., Berkel S. Investigation of SHANK3 in schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet*. 2017;174(4):390-398. DOI 10.1002/ajmg.b.32528.

Decroly O., Goldbeter A. Birhythmicity, chaos, and other patterns of temporal self-organization in a multiply regulated biochemical system. *Proc. Natl. Acad. Sci. USA*. 1982;79:6917-6921. DOI 10.1073/pnas.79.22.6917.

Derivery E., Lombard B., Loew D., Gautreau A. The Wave complex is intrinsically inactive. *Cell Motil. Cytoskeleton*. 2009;66(10):777-790. DOI 10.1002/cm.20342.

Dupont S., Morsut L., Aragona M., Enzo E., Giulitti S., Cordenonsi M., Zanconato F., Le Digabel J., Forcato M., Bicciato S., Elvassore N., Piccolo S. Role of YAP/TAZ in mechanotransduction. *Nature*. 2011;474(7350):179-183. DOI 10.1038/nature10137.

Feng Y., Absher D., Eberhart D.E., Brown V., Malter H.E., Warren S.T. FMRP associates with polyribosomes as an mRNP, and the I304N mutation of severe fragile X syndrome abolishes this association. *Mol. Cell*. 1997;1(1):109-118. DOI 10.1016/s1097-2765(00)80012-x.

Forrest M.P., Parnell E., Penzes P. Dendritic structural plasticity and neuropsychiatric disease. *Nat. Rev. Neurosci*. 2018;19(4):215-234. DOI 10.1038/nrn.2018.16.

Fu Y.J., Liu D., Guo J.L., Long H.Y., Xiao W.B., Xiao W., Feng L., Luo Z.H., Xiao B. Dynamic change of shanks gene mRNA expression and DNA methylation in epileptic rat model and human patients. *Mol. Neurobiol*. 2020;57(9):3712-3726. DOI 10.1007/s12035-020-01968-5.

Gkogkas C.G., Sonenberg N. Translational control and autism-like behaviors. *Cell. Logist*. 2013;3:e24551. DOI 10.4161/cl.24551.

Gobert D., Schohl A., Kutsarova E., Ruthazer E.S. TORC1 selectively regulates synaptic maturation and input convergence in the developing visual system. *Dev. Neurobiol*. 2020. DOI 10.1002/dneu.22782.

Goldbeter A., Gonze D., Houart G., Leloup J.C., Halloy J., Dupont G. From simple to complex oscillatory behavior in metabolic and genetic control networks. *Chaos*. 2001;11:247-260. DOI 10.1063/1.1345727.

Gross C., Nakamoto M., Yao X., Chan C.B., Yim S.Y., Ye K., Warren S.T., Bassell G.J. Excess phosphoinositide 3-kinase subunit synthesis and activity as a novel therapeutic target in fragile X syndrome. *J. Neurosci*. 2010;30:10624-10638. DOI 10.1523/JNEUROSCI.0402-10.2010.

Hong L., Li Y., Liu Q., Chen Q., Chen L., Zhou D. The Hippo signaling pathway in regenerative medicine. *Methods Mol. Biol*. 2019;1893:353-370. DOI 10.1007/978-1-4939-8910-2_26.

Hu J.K., Du W., Shelton S.J., Oldham M.C., DiPersio C.M., Klein O.D. An FAK-YAP-mTOR signaling axis regulates stem cell-based tissue renewal in mice. *Cell Stem Cell.* 2017;21(1):91-106. DOI 10.1016/j.stem.2017.03.023.

Huber K.M., Kayser M.S., Bear M.F. Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science.* 2000;288(5469):1254-1257. DOI 10.1126/science.288.5469.1254.

Huber K.M., Klann E., Costa-Mattioli M., Zukin R.S. Dysregulation of mammalian target of rapamycin signaling in mouse models of autism. *J. Neurosci*. 2015;35(41):13836-13842. DOI 10.1523/JNEUROSCI.2656-15.2015.

Iacoangeli A., Tiedge H. Translational control at the synapse: role of RNA regulators. *Trends Biochem. Sci*. 2013;38(1):47-55. DOI 10.1016/j.tibs.2012.11.001.

Ip C.K., Cheung A.N., Ngan H.Y., Wong A.S. p70 S6 kinase in the control of actin cytoskeleton dynamics and directed migration of ovarian cancer cells. *Oncogene*. 2011;30(21):2420-2432. DOI 10.1038/onc.2010.615.

Johnson S.C., Rabinovitch P.S., Kaeberlein M. mTOR is a key modulator of ageing and age-related disease. *Nature*. 2013;493(7432):338-345. DOI 10.1038/nature11861.

Khlebodarova T.M., Kogai V.V., Fadeev S.I., Likhoshvai V.A. Chaos and hyperchaos in simple gene network with negative feedback and time delays. *J. Bioinform. Comput. Biol*. 2017;15(2):1650042. DOI 10.1142/S0219720016500426.

Khlebodarova T.M., Kogai V.V., Likhoshvai V.A. On the dynamical aspects of local translation at the activated synapse. *BMC Bioinformatics*. 2020;21(Suppl. 11):258. DOI 10.1186/s12859-020-03597-0.

Khlebodarova T.M., Kogai V.V., Trifonova E.A., Likhoshvai V.A. Dynamic landscape of the local translation at activated synapses. *Mol. Psych*. 2018;23(1):107-114. DOI 10.1038/mp.2017.245.

Khlebodarova T.M., Likhoshvai V.A., Kogai V.V. On the chaotic potential of local translation at activated synapses. In: Mathematical Biology and Bioinformatics. Puschino: IMPB RAS, 2018;7:e68.1-e68.6. DOI 10.17537/icmbb18.6. (in Russian)

Klein M.E., Monday H., Jordan B.A. Proteostasis and RNA binding proteins in synaptic plasticity and in the pathogenesis of neuropsychiatric disorders. *Neural Plast*. 2016;2016:3857934. DOI 10.1155/2016/3857934.

Kogai V.V., Khlebodarova T.M., Fadeev S.I., Likhoshvai V.A. Complex dynamics in alternative mRNA splicing: mathematical model. *Vychislitelnye Technologii = Computational Technologies*. 2015;20(1):38-52. (in Russian)

Kogai V.V., Likhoshvai V.A., Fadeev S.I., Khlebodarova T.M. Multiple scenarios of transition to chaos in the alternative splicing model. *Int. J. Bifurcat. Chaos*. 2017;27(2):1730006. DOI 10.1142/S0218127417300063.

Lauterborn J.C., Cox C.D., Chan S.W., Vanderklish P.W., Lynch G., Gall C.M. Synaptic actin stabilization protein loss in Down syndrome and Alzheimer disease. *Brain Pathol*. 2020;30(2):319-331. DOI 10.1111/bpa.12779.

Levy Nogueira M., da Veiga Moreira J., Baronzio G.F., Dubois B., Steyaert J.M., Schwartz L. Mechanical stress as the common denominator between chronic inflammation, cancer, and Alzheimer's disease. *Front. Oncol*. 2015;5:197. DOI 10.3389/fonc.2015.00197.

Levy Nogueira M., Epelbaum S., Steyaert J.M., Dubois B., Schwartz L. Mechanical stress models of Alzheimer's disease pathology. *Alzheimers Dement*. 2016a;12(3):324-333. DOI 10.1016/j.jalz.2015.10.005.

Levy Nogueira M., Hamraz M., Abolhassani M., Bigan E., Lafitte O., Steyaert J.M., Dubois B., Schwartz L. Mechanical stress increases brain amyloid β, tau, and α-synuclein concentrations in wild-type mice. *Alzheimers Dement*. 2018;14(4):444-453. DOI 10.1016/j.jalz.2017.11.003.

Levy Nogueira M., Lafitte O., Steyaert J.M., Bakardjian H., Dubois B., Hampel H., Schwartz L. Mechanical stress related to brain atrophy in Alzheimer's disease. *Alzheimers Dement*. 2016b;12(1):11-20. DOI 10.1016/j.jalz.2015.03.005.

Likhoshvai V.A., Fadeev S.I., Kogai V.V., Khlebodarova T.M. On the chaos in gene networks. *J. Bioinform. Comput. Biol*. 2013;11(1):1340009. DOI 10.1142/S021972001340009X.

Likhoshvai V.A., Golubyatnikov V.P., Khlebodarova T.M. Limit cycles in models of circular gene networks regulated by negative feedbacks. *BMC Bioinformatics*. 2020;21(Suppl. 11):255. DOI 10.1186/s12859-020-03598-z.

Likhoshvai V.A., Khlebodarova T.M. On stationary solutions of delay differential equations: a model of local translation in synapses. *Matematicheskaya Biologiya i Bioinformatika = Mathematical Biology and Bioinformatics.* 2019;14(2):554-569. DOI 10.17537/2019.14.554. (in Russian)

Likhoshvai V.A., Kogai V.V., Fadeev S.I., Khlebodarova T.M. Alternative splicing can lead to chaos. *J. Bioinform. Comput. Biol*. 2015;13(1):1540003. DOI 10.1142/S021972001540003X.

Likhoshvai V.A., Kogai V.V., Fadeev S.I., Khlebodarova T.M. Chaos and hyperchaos in a model of ribosome autocatalytic synthesis. *Sci. Rep*. 2016;6:38870. DOI 10.1038/srep38870.

Louros S.R., Osterweil E.K. Perturbed proteostasis in autism spectrum disorders. *J. Neurochem*. 2016;139(6):1081-1092. DOI 10.1111/jnc.13723.

Mackey M.C., Glass L. Oscillation and chaos in physiological control systems. *Science*. 1977;197(4300):287-289. DOI 10.1126/science.267326.

Majumder P., Chu J.F., Chatterjee B., Swamy K.B., Shen C.J. Co-regulation of mRNA translation by TDP-43 and fragile X syndrome protein FMRP. *Acta Neuropathol*. 2016;132(5):721-738. DOI 10.1007/s00401-016-1603-8.

Martin I. Decoding Parkinson's disease pathogenesis: the role of deregulated mRNA translation. *J. Parkinsons Dis.* 2016;6(1):17-27. DOI 10.3233/JPD-150738.

McCarthy N. Signalling: YAP, PTEN and miR-29 size each other up. *Nat. Rev. Cancer*. 2013;13(1):4-5. DOI 10.1038/nrc3422.

Mei Y., Monteiro P., Zhou Y., Kim J.A., Gao X., Fu Z., Feng G. Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. *Nature.* 2016;530(7591):481-484. DOI 10.1038/nature16971.

Meng X.F., Yu J.T., Song J.H., Chi S., Tan L. Role of the mTOR signaling pathway in epilepsy. *J. Neurol. Sci*. 2013;332(1-2):4-15. DOI 10.1016/j.jns.2013.05.029.

Millard T.H., Sharp S.J., Machesky L.M. Signalling to actin assembly via the WASP (Wiskott-Aldrich syndrome protein)-family proteins

and the Arp2/3 complex. *Biochem. J*. 2004;380(Pt. 1):1-17. DOI 10.1042/ BJ20040176.

Molinie N., Gautreau A. The Arp2/3 regulatory system and its deregulation in cancer. *Physiol. Rev*. 2018;98(1):215-238. DOI 10.1152/physrev.00006.2017.

Monteiro P., Feng G. SHANK proteins: roles at the synapse and in autism spectrum disorder. *Nat. Rev. Neurosci*. 2017;18(3):147-157. DOI 10.1038/nrn.2016.183.

Muddashetty R.S., Kelić S., Gross C., Xu M., Bassell G.J. Dysregulated metabotropic glutamate receptor-dependent translation of AMPA receptor and postsynaptic density-95 mRNAs at synapses in a mouse model of fragile X syndrome. *J. Neurosci*. 2007;27(20):5338-5348. DOI 10.1523/JNEUROSCI.0937-07.2007.

Nakashima M., Kato M., Aoto K., Shiina M., Belal H., Mukaida S., Kumada S., Sato A., Zerem A., Lerman-Sagie T., Lev D., Leong H.Y., Tsurusaki Y., Mizuguchi T., Miyatake S., Miyake N., Ogata K., Saitsu H., Matsumoto N. De novo hotspot variants in CYFIP2 cause early-onset epileptic encephalopathy. *Ann. Neurol*. 2018;83(4):794-806. DOI 10.1002/ana.25208.

Nalavadi V.C., Muddashetty R.S., Gross C., Bassell G.J. Dephosphorylation-induced ubiquitination and degradation of FMRP in dendrites: a role in immediate early mGluR-stimulated translation. *J. Neurosci*. 2012;32(8):2582-2587. DOI 10.1523/JNEUROSCI.5057-11.2012.

Napoli I., Mercaldo V., Boyl P.P., Eleuteri B., Zalfa F., De Rubeis S., Di Marino D., Mohr E., Massimi M., Falconi M., Witke W., Costa-Mattioli M., Sonenberg N., Achsel T., Bagni C. The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell*. 2008;134(6):1042-1054. DOI 10.1016/j.cell.2008.07.031.

Narayanan U., Nalavadi V., Nakamoto M., Pallas D.C., Ceman S., Bassell G.J., Warren S.T. FMRP phosphorylation reveals an immediate-early signaling pathway triggered by group I mGluR and mediated by PP2A. *J. Neurosci*. 2007;27(52):14349-14357. DOI 10.1523/JNEUROSCI.2969-07.2007.

Narayanan U., Nalavadi V., Nakamoto M., Thomas G., Ceman S., Bassell G.J., Warren S.T. S6K1 phosphorylates and regulates fragile X mental retardation protein (FMRP) with the neuronal protein synthesis-dependent mammalian target of rapamycin (mTOR) signaling cascade. *J. Biol. Chem*. 2008;283(27):18478-18482. DOI 10.1074/jbc.C800055200.

Nishiyama J. Plasticity of dendritic spines: Molecular function and dysfunction in neurodevelopmental disorders. *Psychiat. Clin. Neurosci*. 2019;73(9):541-550. DOI 10.1111/pcn.12899.

Onore C., Yang H., Van de Water J., Ashwood P. Dynamic Akt/mTOR signaling in children with autism spectrum disorder. *Front. Pediatr*. 2017;5:43. DOI 10.3389/fped.2017.00043.

Panciera T., Azzolin L., Cordenonsi M., Piccolo S. Mechanobiology of YAP and TAZ in physiology and disease. *Nat. Rev. Mol. Cell Biol*. 2017;18(12):758-770. DOI 10.1038/nrm.2017.87.

Peça J., Feliciano C., Ting J.T., Wang W., Wells M.F., Venkatraman T.N., Lascola C.D., Fu Z., Feng G. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature*. 2011;472(7344):437-442. DOI 10.1038/nature09965.

Pei J.J., Hugon J. mTOR-dependent signalling in Alzheimer's disease. *J. Cell. Mol. Med*. 2008;12(6b):2525-2532. DOI 10.1111/j.1582-4934.2008.00509.x.

Penzes P., Rafalovich I. Regulation of the actin cytoskeleton in dendritic spines. *Adv. Exp. Med. Biol.* 2012;970:81-95. DOI 10.1007/978-3-7091-0932-8_4.

Porokhovnik L. Individual copy number of ribosomal genes as a factor of mental retardation and autism risk and severity. *Cells*. 2019;8(10):1151. DOI 10.3390/cells8101151.

Porokhovnik L.N., Lyapunova N.A. Dosage effects of human ribosomal genes (rDNA) in health and disease. *Chromosome Res*. 2019;27(1-2):5-17. DOI 10.1007/s10577-018-9587-y.

Pramparo T., Pierce K., Lombardo M.V., Carter Barnes C., Marinero S., Ahrens-Barbeau C., Murray S.S., Lopez L., Xu R., Courchesne E. Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practice. *JAMA Psychiatry*. 2015;72:386-394. DOI 10.1001/jamapsychiatry.2014.3008.

Pyronneau A., He Q., Hwang J.Y., Porch M., Contractor A., Zukin R.S. Aberrant Rac1-cofilin signaling mediates defects in dendritic spines, synaptic function, and sensory perception in fragile X syndrome. *Sci. Signal*. 2017;10(504):eaan0852. DOI 10.1126/scisignal.aan0852.

Reddy P., Deguchi M., Cheng Y., Hsueh A.J. Actin cytoskeleton regulates Hippo signaling. *PLoS One*. 2013;8(9):e73763. DOI 10.1371/journal.pone.0073763.

Rex C.S., Chen L.Y., Sharma A., Liu J., Babayan A.H., Gall C.M., Lynch G. Different Rho GTPase-dependent signaling pathways initiate sequential steps in the consolidation of long-term potentiation. *J. Cell Biol*. 2009;186(1):85-97. DOI 10.1083/jcb.200901084.

Rosenberg T., Gal-Ben-Ari S., Dieterich D.C., Kreutz M.R., Ziv N.E., Gundelfinger E.D., Rosenblum K. The roles of protein expression in synaptic plasticity and memory consolidation. *Front. Mol. Neurosci*. 2014;7:86. DOI 10.3389/fnmol.2014.00086.

Santini E., Huynh T.N., Klann E. Mechanisms of translation control underlying long-lasting synaptic plasticity and the consolidation of long-term memory. *Prog. Mol. Biol. Transl. Sci*. 2014;122:131-167. DOI 10.1016/B978-0-12-420170-5.00005-2.

Schaks M., Reinke M., Witke W., Rottner K. Molecular dissection of neurodevelopmental disorder-causing mutations in CYFIP2. *Cells*. 2020;9(6):1355. DOI 10.3390/cells9061355.

Schaks M., Singh S.P., Kage F., Thomason P., Klünemann T., Steffen A., Blankenfeldt W., Stradal T.E., Insall R.H., Rottner K. Distinct interaction sites of RAC GTPase with WAVE regulatory complex have non-redundant functions *in vivo*. *Curr. Biol*. 2018;28(22):3674-3684.e6. DOI 10.1016/j.cub.2018.10.002.

Seo J., Kim J. Regulation of Hippo signaling by actin remodeling. *BMB Rep*. 2018;51(3):151-156. DOI 10.5483/bmbrep.2018.51.3.012.

Sharma A., Hoeffer C.A., Takayasu Y., Miyawaki T., McBride S.M., Klann E., Zukin R.S. Dysregulation of mTOR signaling in fragile X syndrome. *J. Neurosci*. 2010;30(2):694-702. DOI 10.1523/JNEUROSCI.3696-09.2010.

Suzuki Y., Lu M., Ben-Jacob E., Onuchic J.N. Periodic, quasi-periodic and chaotic dynamics in simple gene elements with time delays. *Sci. Rep*. 2016;6:21037. DOI 10.1038/srep21037.

Tapon N., Hall A. Rho, Rac and Cdc42 GTPases regulate the organization of the actin cytoskeleton. *Curr. Opin. Cell Biol*. 1997;9(1):86-92. DOI 10.1016/s0955-0674(97)80156-1.

Totaro A., Panciera T., Piccolo S. YAP/TAZ upstream signals and downstream responses. *Nat. Cell Biol*. 2018;20(8):888-899. DOI 10.1038/s41556-018-0142-z.

Trifonova E.A., Khlebodarova T.M., Gruntenko N.E. Molecular mechanisms of autism as a form of synaptic dysfunction. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2016;20(6):959-967. DOI 10.18699/VJ16.217. (in Russian)

Trifonova E.A., Khlebodarova T.M., Gruntenko N.E. Molecular mechanisms of autism as a form of synaptic dysfunction. *Russ. J. Genet.: Appl. Res.* 2017;7(8):869-877.

Troca-Marin J.A., Alves-Sampaio A., Montesinos M.L. Deregulated mTOR-mediated translation in intellectual disability. *Prog. Neurobiol*. 2012;96(2):268-282. DOI 10.1016/j.pneurobio.2012.01.005.

Tumaneng K., Schlegelmilch K., Russell R.C., Yimlamai D., Basnet H., Mahadevan N., Fitamant J., Bardeesy N., Camargo F.D., Guan K.L. YAP mediates crosstalk between the Hippo and PI(3)K–TOR pathways by suppressing PTEN via miR-29. *Nat. Cell Biol*. 2012;14(12):1322-1329. DOI 10.1038/ncb2615.

Won H., Mah W., Kim E. Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. *Front. Mol. Neurosci*. 2013;6:19. DOI 10.3389/fnmol.2013.00019.

Wong M. Mammalian target of rapamycin (mTOR) inhibition as a potential antiepileptogenic therapy: From tuberous sclerosis to com-

mon acquired epilepsies. *Epilepsia*. 2010;51(1):27-36. DOI 10.1111/j.1528-1167.2009.02341.x.

Wostyn P. Intracranial pressure and Alzheimer's disease: a hypothesis. *Med. Hypotheses*. 1994;43(4):219-222. DOI 10.1016/0306-9877(94)90069-8.

Yu F.X., Zhao B., Guan K.L. Hippo pathway in organ size control, tissue homeostasis, and cancer. *Cell*. 2015;163(4):811-828. DOI 10.1016/j.cell.2015.10.044.

Zamboni V., Jones R., Umbach A., Ammoni A., Passafaro M., Hirsch E., Merlo G.R. Rho GTPases in intellectual disability: from genetics to therapeutic opportunities. *Int. J. Mol. Sci*. 2018;19:1821. DOI 10.3390/ijms19061821.

Zhang Y., Lee Y., Han K. Neuronal function and dysfunction of CYFIP2: from actin dynamics to early infantile epileptic encephalopathy. *BMB Rep*. 2019;52(5):304-311. DOI 10.5483/BMBRep.2019.52.5.097.

Zhou J., Parada L.F. PTEN signaling in autism spectrum disorders. *Curr. Opin. Neurobiol*. 2012;22(5):873-879. DOI 10.1016/j.conb.2012.05.004.

Zhu T., Ma Z., Wang H., Jia X., Wu Y., Fu L., Li Z., Zhang C., Yu G. YAP/TAZ affects the development of pulmonary fibrosis by regulating multiple signaling pathways. *Mol. Cell. Biochem*. 2020;475(1-2):137-149. DOI 10.1007/s11010-020-03866-9.

Zukin R.S., Richter J.D., Bagni C. Signals, synapses, and synthesis: how new proteins control plasticity. *Front. Neural Circuits*. 2009;3:14. DOI 10.3389/neuro.04.014.2009.

Zweier M., Begemann A., McWalter K., Cho M.T., Abela L., Banka S., Behring B., Berger A., Brown C.W., Carneiro M., Chen J., Cooper G.M. Deciphering Developmental Disorders (DDD) Study, Finnila C.R., Guillen Sacoto M.J., Henderson A., Hüffmeier U., Joset P., Kerr B., Lesca G., Leszinski G.S., McDermott J.H., Meltzer M.R., Monaghan K.G., Mostafavi R., Õunap K., Plecko B., Powis Z., Purcarin G., Reimand T., Riedhammer K.M., Schreiber J.M., Sirsi D., Wierenga K.J., Wojcik M.H., Papuc S.M., Steindl K., Sticht H., Rauch A. Spatially clustering *de novo* variants in CYFIP2, encoding the cytoplasmic FMRP interacting protein 2, cause intellectual disability and seizures. *Eur. J. Hum. Genet*. 2019;27(5):747-759. DOI 10.1038/s41431-018-0331-z.

Acknowledgements. This work was supported by the budget project No. 0259-2021-0009.

Conflict of interest. The author declares no conflict of interest.

Received October 19, 2020. Revised December 21, 2020. Accepted December 22, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):101-107
DOI 10.18699/VJ21.012

Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes

K.V. Ustyantsev, E.V. Berezikov ✉

Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ eberez@bionet.nsc.ru

Abstract. In eukaryotes, trans-splicing is a process of nuclear pre-mRNA maturation where two different RNA molecules are joined together by the spliceosomal machinery utilizing mechanisms similar to cis-splicing. In diverse taxa of lower eukaryotes, spliced leader (SL) trans-splicing is the most frequent type of trans-splicing, when the same sequence derived from short small nuclear RNA molecules, called SL RNAs, is attached to the 5′ ends of different non-processed pre-mRNAs. One of the functions of SL trans-splicing is processing polycistronic pre-mRNA molecules transcribed from operons, when several genes are transcribed as one pre-mRNA molecule. However, only a fraction of trans-spliced genes reside in operons, suggesting that SL trans-splicing must also have some other, less understood functions. Regenerative flatworms are informative model organisms which hold the keys to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is fueled by the division and differentiation of the adult somatic stem cell population called neoblasts. *Macrostomum lignano* is a flatworm model organism where substantial technological advances have been achieved in recent years, including the development of transgenesis. Although a large fraction of genes in *M. lignano* were estimated to be SL trans-spliced, SL trans-splicing was not studied in detail in *M. lignano* before. Here, we performed the first comprehensive study of SL trans-splicing in *M. lignano*. By reanalyzing the existing genome and transcriptome data of *M. lignano*, we estimate that 30 % of its genes are SL trans-spliced, 15 % are organized in operons, and almost 40 % are both SL trans-spliced and in operons. We annotated and characterized the sequence of SL RNA and characterized conserved cis- and SL trans-splicing motifs. Finally, we found that a majority of SL trans-spliced genes are evolutionarily conserved and significantly over-represented in neoblast-specific genes. Our findings suggest an important role of SL trans-splicing in the regulation and maintenance of neoblasts in *M. lignano*.
Key words: flatworms; regeneration; splicing; trans-splicing; neoblasts; spliced leader; *Macrostomum lignano*.

For citation: Ustyantsev K.V., Berezikov E.V. Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):101-107. DOI 10.18699/VJ21.012

Биоинформационный анализ сплайс-лидерного транс-сплайсинга у регенерирующего плоского червя *Macrostomum lignano* показал его преобладание среди консервативных генов и генов стволовых клеток

К.В. Устьянцев, Е.В. Березиков ✉

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ eberez@bionet.nsc.ru

Аннотация. Транс-сплайсинг у эукариот – это процесс созревания ядерных пре-мРНК, когда две различные молекулы РНК соединяются с помощью структур сплайсосомы по механизму, схожему с цис-сплайсингом. У различных таксонов низших эукариот наиболее распространенный тип транс-сплайсинга – сплайс-лидерный (СЛ) транс-сплайсинг, при котором одинаковая последовательность, происходящая от коротких малых ядерных РНК молекул, называемых СЛ РНК, присоединяется к 5′-концам различных непроцессированных пре-мРНК. Одна из функций СЛ транс-сплайсинга состоит в процессировании полицистронных молекул пре-мРНК, транскрибируемых с оперонов, когда транскрипция нескольких генов осуществляется как одна молекула пре-мРНК. Однако лишь часть генов, подвергающихся транс-сплайсингу, содержится в оперонах, что говорит о том, что у СЛ транс-сплайсинга должны быть и другие, менее изученные, функции. Регенерирующие плоские черви являются информативными модельными организмами, хранящими ключи к пониманию механизмов регуляции стволовых клеток и их дифференцировки во время регенерации и при гомеостазе. Их способность к регенерации – следствие деления и дифференцировки соматических стволовых клеток, называемых необластами, которые присутствуют у взрослых особей. *Macrostomum lignano* – модельный плоский червь, в исследованиях на котором в

последние годы достигнут существенный технологический прогресс, включая разработку метода трансгенеза. Сплайс-лидерный транс-сплайсинг ранее не был детально изучен у *M. lignano*, хотя известно, что значительная часть генов *M. lignano* подвергается этому типу транс-сплайсинга. В настоящей работе мы осуществили первое обширное исследование СЛ транс-сплайсинга у *M. lignano*. Повторно проанализировав геномные и транскриптомные данные *M. lignano*, мы оцениваем, что 30 % его генов подвергаются СЛ транс-сплайсингу, 15 % расположены в оперонах, а почти 40 % находятся в оперонах и проходят через СЛ транс-сплайсинг. Мы провели аннотацию и охарактеризовали последовательность СЛ РНК и консервативных мотивов цис- и транс-сплайсинга. Обнаружено, что большинство генов, подвергающихся СЛ транс-сплайсингу, эволюционно консервативны и значительно перепредставлены в генах, специфичных для необластов. Наши результаты предполагают важную роль СЛ транс-сплайсинга в регуляции функционирования необластов у *M. lignano*.
Ключевые слова: плоские черви; регенерация; сплайсинг; транс-сплайсинг; необласты; сплайс-лидер; *Macrostomum lignano*.

Introduction

Before being used as templates for protein production, majority of RNA molecules transcribed in the nucleus (pre-mRNA) undergo three major modifications to become mature and fully functional mRNA. This is called RNA processing and it involves capping of the 5′ end, polyadenylation of the 3′ end, and splicing. Two types of splicing are distinguished – cis- and trans-splicing. During cis-splicing all the processing happens with the same pre-mRNA molecule, resulting in the removal of introns and merging of its exons. During trans-splicing, on the other hand, two different pre-mRNA molecules expressed from distinct genomic loci are joint into a new chimeric trans-spliced mRNA (Lasda, Blumenthal, 2011).

Trans-splicing was originally discovered in trypanosomes (Euglenozoa), where it was found that a short 39 bp leader sequence was post-transcriptionally attached to the 5′ ends of variant surface glycoproteins pre-mRNA (Boothroyd, Cross, 1982). Later, 5′ end addition of a 22 bp spliced leader (SL) was also observed in *Caenorhabitis elegans* mRNA of actin gene and some other genes (Krause, Hirsh, 1987). Now this process is well known as SL trans-splicing. A distinct feature of SL trans-splicing is that all such processed transcripts have the same short SL sequence, or its variant, at their 5′ ends. The SL sequence is derived from an exon of a non-coding small nuclear RNA molecule called SL RNA, which is ~100 nt in length and has 2,2,7-trimethylguanosine cap at its 5′ end instead of 7-methylguanosine cap, which is found in non-trans-spliced mRNAs (Liou, Blumenthal, 1990; Lasda, Blumenthal, 2011). SL RNAs have a splicing donor site at the exon 3′ end, while SL trans-spliced pre-mRNAs have a splicing acceptor site at the 5′ end of their first exon. SL trans-splicing results in removal of the 5′ non-exon pre-mRNA part called outron (Lasda, Blumenthal, 2011). It is experimentally shown that the only requirement for a gene to be predominantly SL trans-spliced is an acceptor splicing site close to the 5′ end of the first exon that is not complemented by a donor splicing site upstream in cis (Conrad et al., 1993). Thus comes another important feature of SL trans-splicing, namely that it allows formation and resolving of operons – adjacent genes transcribed as a single pre-mRNA from the same promoter region (Blumenthal, Gleason, 2003). However, apart from a clear function in polycistronic transcripts resolution, the function of SL trans-splicing for monocystronic transcripts is still in debate (Danks et al., 2015). It is hypothesized that the function may be in equalization of 5′ UTRs in length and their clearance from out-of-frame AUG start codons, while at the same time allowing less restricted evolution of 5′ upstream regulatory sequences, and in additional control of translation (Hastings, 2005; Danks, Thompson, 2015). So far, SL trans-splicing was found in several clades of eukaryotes: dinoflagellates, euglenozoans, cnidarians, flatworms, nematodes, and ascidians (Lei et al., 2016). SL trans-splicing is most prominent in trypanosomes (100 % genes are trans-spliced) and in nematodes (70 % genes are trans-spliced) (Allen et al., 2011; Lei et al., 2016).

Regenerative flatworms are informative models to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is driven by the division and differentiation of the adult somatic stem cell population called neoblasts (Wagner et al., 2011; Mouton et al., 2018). *Macrostomum lignano* is the only flatworm species for which a method for stable transgenesis is available so far. The worm also has a number of features allowing for efficient cell lineage tracing and phenotype screening, which makes *M. lignano* an attractive model to study a wide range of biological processes (Grudniewska et al., 2016; Wudarski et al., 2017, 2019, 2020). Well-annotated *M. lignano* genome and transcriptome assemblies were recently published (Wudarski et al., 2017; Grudniewska et al., 2018). It was estimated that almost 21 % of its genes are SL trans-spliced to the same 35 bp SL sequence (Grudniewska et al., 2018). However, trans-splicing was not studied in details in *M. lignano*, and its impact on the genome functioning and maintenance is still unknown. Here, we present the first comprehensive study of SL trans-splicing in *M. lignano* and show that it is strongly connected with genes specific for the neoblasts of the worm.

Materials and methods

Data. The published *M. lignano* genome Mlig_3_7 (Wudarski et al., 2017) and transcriptome Mlig_RNA_3_7_DV1_v3 (Grudniewska et al., 2018) assemblies and the corresponding annotation tracks were obtained from (http://gb.macgenome.org/downloads/Mlig_3_7/).

Genome deduplication. Mlig_3_7 genome assembly was deduplicated using purge_dups software (v1.0.1) with default settings (Guan et al., 2020) and utilizing published PacBio genome sequencing data (Wasik et al., 2015) for the calculation of contig coverages. Contig names from the deduplicated genome assembly were used to extract respective gene annotations from the full Mlig_3_7 genome annotation.

Motif discovery and SL RNA annotation. Presence of the SL sequence at the 5′ end of the *M. lignano* transcripts was established in the previous studies (Wasik et al., 2015; Grudniewska et al., 2016). For the annotation of trans-spliced genes, SL-containing RNA-seq reads were mapped to the

Mlig_3_7 genome assembly and the presence of such reads at the beginning of transcripts was used as an evidence of SL trans-splicing (Wudarski et al., 2017; Grudniewska et al., 2018). Therefore, all the SL trans-spliced transcripts have the corresponding annotation in the Mlig_3_7 genome assembly, and the sequences upstream of their first exon were considered as outrons. Using the deduplicated annotation track of gene coordinates, we retrieved nucleotide sequences of genomic regions corresponding to exon-intron and exon-outron (for the trans-spliced genes) junctions with 50 bp flanks in both directions. All the sequences were converted to forward orientation and split into three groups corresponding to cis-donor, cis-acceptor, and trans-spliced acceptor sites. The sequences then were analysed for the presence of enriched motif using a stand-alone version of the DREME tool (Bailey, 2011).

To determine the SL RNA gene sequence in the genome assembly, we used the 35 bp *M. lignano* SL sequence (CGG TCTCTTACTGCGAAGACTCAATTTA TTGCATG) as a seed for a BLASTn (Altschul et al., 1990) search requiring only 100 % matching hits. Next, we manually investigated genomic sequences surrounding the BLAST hits by matching the SL sequence track in the genome browser to the expected size of SL RNA (~100 bp). The corresponding sequences were then checked for folding into secondary structure canonical for SL RNA folding using Mfold web server (Zuker, 2003), and conserved motifs were then manually identified.

Prediction of operons. Intergenic distances were retrieved from the deduplicated genome annotation file. We only considered distances between immediately adjacent transcripts with the same transcriptional orientation and not interrupted by transcripts in opposite orientation. Distances were split into three categories: between SL trans-spliced genes, between a non-SL trans-spliced gene and an SL trans-spliced gene, and between non-SL trans-spliced genes. To adjust for repetitive element insertions, we retrieved the corresponding coordinates from the genome browser RepeatMasker and TRF tracks (http://gb.macgenome.org/) and subtracted them from the previously identified intergenic distances. Distribution of the distances was visualized as density plots using ggplot2 library in R.

After the analysis of the graphical data of the distances distributions, we selected the threshold value of 1000 bp, below which a pair of adjacent and SL trans-spliced genes were considered as belonging to the same operon. The same applies if the first gene is non-trans-spliced, but the second is SL trans-spliced. The distributions of lengths of operons of various sizes was visualized as violin plots using ggplot2 library in R.

Estimation of gene conservation. Gene annotation and data classifying genes as being specific to neoblasts or germline were retrieved from the previous study (Grudniewska et al., 2018). A gene was considered to be conserved if it has an open reading frame with a detectable homology to a human gene, which is indicated in its annotation, and non-conserved if lacking the homology to human, but has a predicted open reading frame with homology to proteins in other organisms. Otherwise, a gene was considered non-coding.

Results

Deduplication of genome assembly. The published Mlig_3_7 genome assembly is based on the sequencing data from DV1 *M. lignano* line. This line has a $2n = 10$ karyotype (four large and six small chromosomes) and was demonstrated to have undergone a duplication of its large chromosome (Zadesenets et al., 2017), while the karyotype of the basal wild type population is $2n = 8$ (two large and six small chromosomes) (Wudarski et al., 2017). The size of Mlig_3_7 assembly is 764 Mb, which corresponds to the experimental measurement of the genome size in the DV1 line (Wudarski et al., 2017), and the assembly contains the duplicated large chromosome sequences. To avoid gene overcounting due to the presence of these duplicated sequences in the Mlig_3_7 assembly, we removed the most redundant scaffolds by deduplicating Mlig_3_7 assembly using purge_dups software (Guan et al., 2020). This resulted in approximately 46 % drop in the number of scaffolds (from 5270 to 2841) and decreased the genome size to 580 Mb, which is close to the genome size measurements for the NL10 line of *M. lignano*, which does not have the chromosomal duplication (Wudarski et al., 2017). Next, we removed the records from transcriptome annotation which corresponded to the redundant scaffolds.

Motif discovery and SL RNA gene mapping. Investigation of the deduplicated part of the transcriptome shows that a significant fraction of genes, 21754 out of 71499 (30 %), are SL trans-spliced in *M. lignano.* This means that they all have the same 35 bp SL sequence (CGGTCTCTTACTGCG AAGACTCAATTTA TTGCATG) at the 5′ end of their processed transcripts (Wudarski et al., 2017; Grudniewska et al., 2018). Despite this, SL trans-splicing was not characterized in more detail in *M. lignano*. First, we retrieved genomic DNA sequences near the cis-splicing and SL trans-splicing exon-intron/exon-outron junction sites and checked if they are enriched for some motifs using DREME (Fig. 1, *a*) (Bailey, 2011). In total, we obtained 187627 regions around 5′ donor and 3′ acceptor cis-splicing sites and 21754 regions around SL trans-splicing sites. The first most enriched motifs near cis-splicing 5′ donor and 3′ acceptor sites were GT[G/A]AG (found in 122399 regions, p-value: $8.8e^{-23468}$) and CAG (found in 112174 regions, p-value: $1.7e^{-12459}$), respectively, corresponding to canonical cis-splicing motifs. A motif [T/C] TNCAG (found in 9551 regions, p-value: $1.3e^{-1631}$) was the top enriched motif near SL trans-splicing 3′ acceptor sites. All the motifs were positioned right at the exon-intron/exon-outron junctions of the corresponding sites (see Fig. 1, *a*).

Next, to confirm the presence of the SL RNA gene in the genome assembly, we analysed the secondary structure of the previously published sequence of *M. lignano* SL RNA from the ML2 version of the genome (Wasik et al., 2015). However, we found that the reported sequence was erroneously assigned as SL RNA, since it clearly maps to the 5′ end of an SL trans-spliced protein-coding gene (Mlig013257.g1, scaf577:45663-48770) in the Mlig_3_7 assembly, and also does not fold into canonical structure with three hairpin loops (data not shown) (Xie, Hirsh, 1998). Therefore, we decided to identify the actual SL RNA gene in the newer Mlig_3_7 assembly. Using SL sequence as a seed for the genomic BLASTn search, we mapped a 109 bp sequence, which is repeated eight times in the deduplicated genome and has the canonical SL RNA secondary structure predicted by Mfold web server (see Fig. 1, *b*) (Zuker, 2003). Subsequent sequence analysis showed clear signatures of an SL RNA: the SL sequence

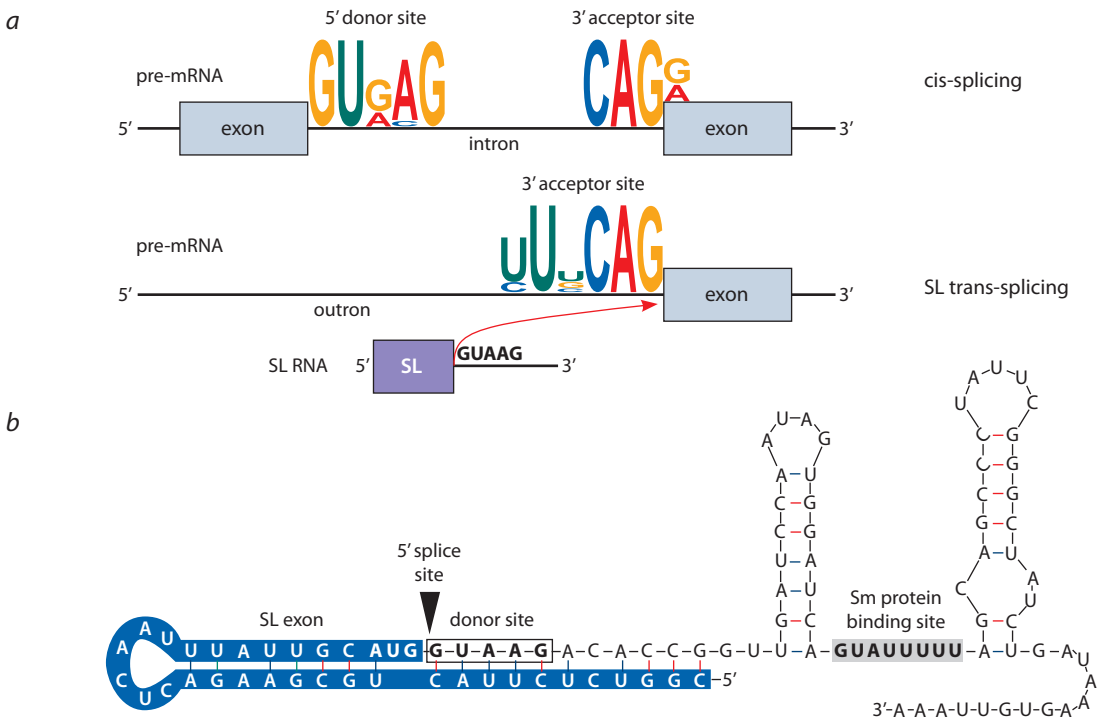


Fig. 1. Features of cis-splicing and SL trans-splicing in *M. lignano.*

a – conserved motifs enriched at the splicing junction sites in cis-spliced and SL trans-spliced genes; *b* – sequence and predicted secondary structure of *M. lignano* SL RNA gene.

is at the 5′ end of the gene and forms the first hairpin loop, immediately after the SL sequence there is a clear 5′ donor splicing site (GTAAG), and between two other hairpin loops there is a motif similar to the binding site of Sm spliceosomal protein (see Fig. 1, *b*) (Ganot et al., 2004; Stover et al., 2006).

Operon analysis. The important feature of SL trans-splicing is that it allows for processing of long polycistronic pre-mRNA molecules expressed from a single promoter region in a way similar to prokaryote operons. In principle, genome-guided transcriptome assembly using RNA-seq data allows identification of such operons and their corresponding pre-mRNA sequences, which we previously annotated as transcriptional units (Wudarski et al., 2017; Grudniewska et al., 2018). However, it is not always possible to fully reconstruct an operon from RNA-seq data alone, since transcriptional units predicted from RNA-seq data tend to split in the repeat-rich intergenic regions of operons, where read coverage depends on both operon expression level and the frequency of repeats in the genome. Instead, to estimate what fraction of *M. lignano* genes are organized in operons based on their genomic organization, we first explored how intergenic distances between trans- and non-trans-spliced genes are distributed in *M. lignano* genome (Fig. 2, *a*). We found that distribution of distances between trans-spliced genes has multimodal distribution, while it is unimodal distribution for non-trans-spliced/trans-spliced and non-trans-spliced/non-trans-spliced intergenic distances (see Fig. 2, *a*). SL trans-splicing is an ancient evolutionary mechanism (Lei et al., 2016), which is mostly abundant in the genomes of simply organized organisms, which have low repetitive content and relatively small genomes (Gregory et al., 2007). We hypothesized that neutral accumulation of repeats could have influenced the distances between genes in the operons. Interestingly, after we adjusted the intergenic distances by subtracting the fraction occupied by repetitive sequences (simple repeats and transposable elements), it had the most impact on the distances between trans-spliced genes, revealing a clear bimodal distribution with the most prominent peak at around 100 bp (see Fig. 2, *a*). This observation indicates that repeats have a substantial contribution to intergenic distances in operons. To classify genes as belonging to the same operon, we decided to use the repeat-adjusted distances with a threshold value of 1 Kb, which separates the two modes of the intergenic distances between trans-spliced genes (see Fig. 2, *a*).

Using these criteria for defining operons, we found that 10458 genes (approx. 15 % of all genes and 40 % of SL trans-spliced genes) can be assigned to operons, of which 1854 (18 %) start from a non-trans-spliced gene (see Fig. 2, *b*, Fig. 3). The vast majority of them are comprised of two and three genes (75 and 18 %), with the maximum operon size reaching nine genes (two operons) (see Fig. 2, *b*). An example of an operon defined in this way is provided in Fig. 2, *c*.

SL trans-splicing is enriched in evolutionary conserved and stem cell genes. We know from a previous study (Grudniewska et al., 2018) that evolutionary conserved protein-coding genes, which still have detectable homology between *M. lignano* and human, are enriched in somatic stem cells – neoblasts (85 % compared to overall 47 %) (see Fig. 3). On the contrary, only 38 % of germline-specific genes in *M. lignano* are conserved in human, suggesting their relatively recent appearance in evolution of flatworms. We investigated whether there is a correlation between gene conservation and

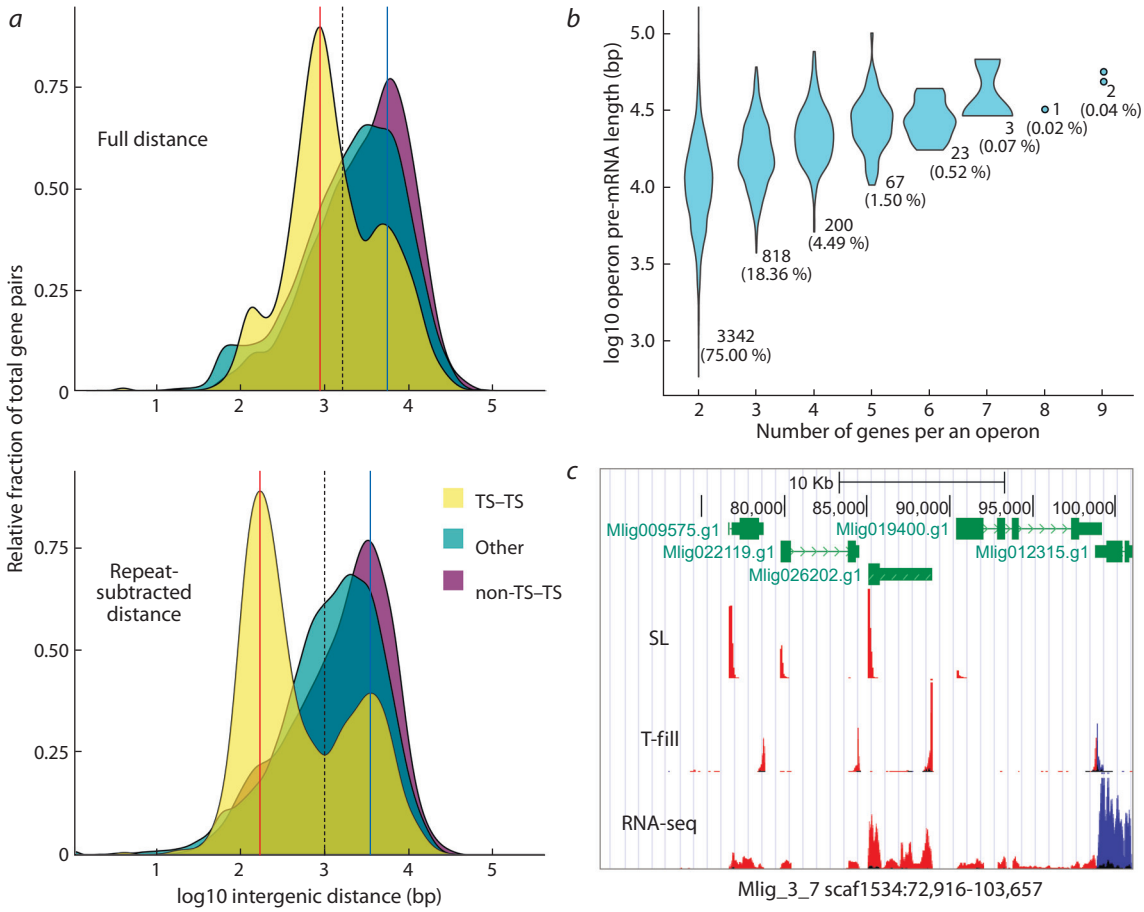


Fig. 2. Identification and characteristics of genes in operons in *M. lignano* genome.

a – distribution of intergenic distances between various gene types. TS – SL trans-spliced, non-TS – non-SL trans-spliced. Red and blue vertical lines indicate modes of the distributions. Vertical black dashed line indicates distance threshold value selected to separate genes in operons; *b* – putative pre-mRNA length and abundance of different operon sizes; *c* – an example of an operon with four genes as depicted in the *M. lignano* genome browser (http://gb.macgenome.org). Genes are in green, with exons as blocks and introns as dashed lines. Non-protein-coding part of the exons are narrower. SL – RNA-seq reads mapped which contained the SL sequence at their 5′ ends (trimmed). T-fill – RNA-seq reads mapped containing mRNA 3′ poly-A ends. Reads mapped in forward orientation are in red, and the reversed reads are in blue.

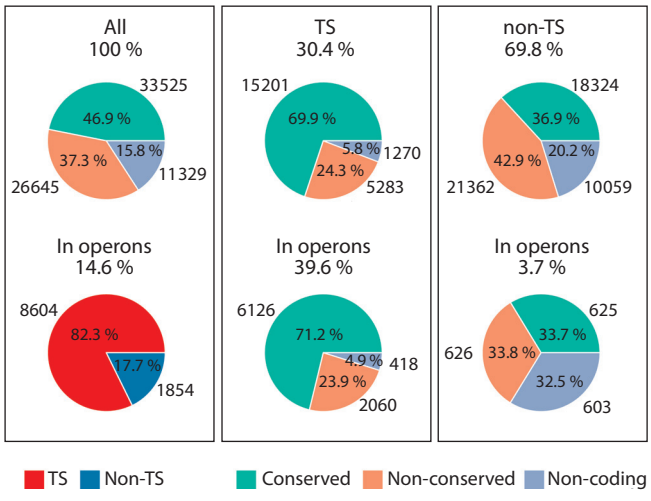


Fig. 3. Evolutionary conservation of *M. lignano* genes.

TS – SL trans-spliced; non-TS – non-SL trans-spliced; conserved – protein-coding genes with a homology to human; non-conserved – protein-coding genes lacking the homology to human; non-coding – genes do not code for a protein.

SL trans-splicing in *M. lignano*. We found that 69.9 % of the SL trans-spliced genes are conserved between *M. lignano* and human, while 24.3 % are not conserved and 5.8 % are non-coding (see Fig. 3). Trans-spliced genes that are located in operons have a very similar distribution of conserved, non-conserved and non-coding genes (see Fig. 3). In contrast, among non-trans-spliced genes only 36.9 % are conserved in human, while 42.9 % are non-conserved and 20.2 % are non-coding (see Fig. 3). Thus, SL trans-spliced genes are strongly enriched for conserved genes but there is no dependence on whether these genes are in operons or not.

Next, we calculated the fraction of SL trans-spliced genes among genes enriched in neoblasts (stem cells) and germline – the only proliferation capable cell types in the worm (Grudniewska et al., 2018). Intriguingly, 85 % of the stem cell genes (746) are SL trans-spliced, and almost 86 % (752) are conserved in human (see the Table), and 728 genes are both conserved in human and SL trans-spliced, which is 96.8 % of all the conserved genes in neoblasts. Given that out of 33 525 conserved genes present in the Mlig_3_7 genome annotation 15 201 (45.3 %) are trans-spliced (see Fig. 3), this represents

Summary of transcripts from *M. lignano* proliferation-capable cell types

Cell type	Total transcripts	Trans-spliced (%)	In operons (%)	Conserved in human (%)	Non-conserved in human (%)	Non-coding (%)
Neoblasts	878	746 (85.0)	343 (39.1)	752 (85.6)	19 (2.2)	107 (12.2)
Germline	1985	362 (18.2)	192 (9.7)	736 (37.1)	248 (12.5)	1001 (50.4)

a 2.13-fold enrichment for conserved SL trans-spliced genes among neoblast genes relative to the expected from the random distribution (*p*-value: 1.98e–7, chi-square test). On the contrary, only 18 % of the germline genes are SL trans-spliced and 37 % are conserved in human. Taken all together, this suggests that SL trans-splicing plays an important role in stem cell regulation in *M. lignano*.

Discussion

SL trans-splicing is widespread in diverse flatworm taxa, including both parasitic and free-living species (Zayas et al., 2005; Protasio et al., 2012; Wudarski et al., 2017; Ershov et al., 2019). However, most of the studies of SL trans-splicing were focused on nematodes and trypanosomes (Lasda, Blumenthal, 2011; Lei et al., 2016). Here, we performed the first study which focuses on SL trans-splicing in the free-living regenerative flatworm model *M. lignano.* By reanalyzing the available genome and transcriptome data, we found that 30 % of the worm genes are SL trans-spliced, and 15 % are estimated to be organized in operons (see Fig. 3). For a comparison, in *C. elegans* 70 % of genes are SL trans-spliced and 17 % are in operons, in ascidian chordate *Ciona intestinalis* it is 58 and 20 %, respectively, and in the parasitic liver fluke *Schistosoma mansoni* 11 % are SL trans-spliced with a few genes in operons (Blumenthal, Gleason, 2003; Satou et al., 2008; Matsumoto et al., 2010; Protasio et al., 2012). Among free-living flatworms, trans-splicing was studied before (Zayas et al., 2005; Rossi et al., 2014), but there is no firm estimation of its abundance and prevalence of genes in operons. The size of operons in *M. lignano* also varies similarly to *C. elegans*, where it ranges from two to eight genes, with the most frequent intergenic distance around 100 bp, and the majority of operons comprised of two genes (see Fig. 2) (Allen et al., 2011).

The most striking finding of our study is that most of *M. lignano* SL trans-spliced genes are evolutionary conserved (see Fig. 3) and, most importantly, that overwhelming majority of neoblast-specific genes (85 %) are SL trans-spliced (see the Table). Interestingly, 39 % of neoblast genes are also clustered in operons (see the Table), suggesting their early evolutionary origin and the necessity for synchronized expression and similar transcriptional regulation. Neoblasts are the key players of outstanding regeneration capacity in free-living flatworms, and thus they are the primary subject of the studies on flatworm regeneration. All the tissue renewal and growth in adult flatworms is due to neoblast proliferation and differentiation (Egger et al., 2006; Ladurner et al., 2008; Wagner et al., 2011). Our data clearly indicates importance of SL trans-splicing for the gene regulation of neoblasts in *M. lignano* and lay ground for further studies of how exactly SL trans-splicing machinery contributes to different stages of neoblast activity.

Conclusion

Spliced leader trans-splicing affects a substantial fraction of *M. lignano* genes. We annotated and characterized the sequence of SL RNA, identified the conserved motifs at the exon-intron/exon-outron junction sites in cis- and SL trans-spliced genes, and provided the first comprehensive analysis of genes comprising operons in *M. lignano*. Most importantly, we found that the SL trans-spliced fraction is over-represented by evolutionary conserved protein-coding genes, in contrast to the non-trans-spliced part of the genome, and that the stem cell-specific genes are predominantly SL trans-spliced. Our findings suggest an important and evolutionary conserved role of SL trans-splicing in regulation and maintenance of neoblasts in *M. lignano.* Thus, a thorough investigation of the molecular mechanism of SL trans-splicing is required to fully understand the regulation of regeneration and stem cell differentiation in flatworms.

References

Allen M.A., Hillier L.W., Waterston R.H., Blumenthal T. A global analysis of *C. elegans* trans-splicing. *Genome Res.* 2011;21(2):255-264. DOI 10.1101/gr.113811.110.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol*. 1990;215(3):403-410. DOI 10.1016/S0022-2836(05)80360-2.

Bailey T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011;27(12):1653-1659. DOI 10.1093/bioinformatics/btr261.

Blumenthal T., Gleason K.S. *Caenorhabditis elegans* operons: form and function. *Nat. Rev. Genet*. 2003;4(2):110-118. DOI 10.1038/nrg995.

Boothroyd J.C., Cross G.A. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5 end. *Gene*. 1982;20(2):281-289. DOI 10.1016/0378-1119(82)90046-4.

Conrad R., Liou R.F., Blumenthal T. Conversion of a trans-spliced *C. elegans* gene into a conventional gene by introduction of a splice donor site. *EMBO J.* 1993;12(3):1249-1255.

Danks G.B., Raasholm M., Campsteijn C., Long A.M., Manak J.R., Lenhard B., Thompson E.M. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol. Biol. Evol*. 2015;32(3):585-599. DOI 10.1093/molbev/msu336.

Danks G., Thompson E.M. Trans-splicing in metazoans: A link to translational control? *Worm*. 2015;4(3):e1046030. DOI 10.1080/21624054.2015. 1046030.

Egger B., Ladurner P., Nimeth K., Gschwentner R., Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. *Dev. Genes Evol*. 2006;216(10):565-577. DOI 10.1007/s00427-006-0069-4.

Ershov N.I., Mordvinov V.A., Prokhortchouk E.B., Pakharukova M.Y., Gunbin K.V., Ustyantsev K., Genaev M.A., Blinov A.G., Mazur A., Boulygina E., Tsygankova S., Khrameeva E., Chekanov N., Fan G., Xiao A., Zhang H., Xu X., Yang H., Solovyev V., Lee S.M.-Y.,

Liu X., Afonnikov D.A., Skryabin K.G. New insights from *Opisthorchis felineus* genome: update on genomics of the epidemiologically important liver flukes. *BMC Genomics*. 2019;20(1):399. DOI 10.1186/s12864-019-5752-8.

Ganot P., Kallesøe T., Reinhardt R., Chourrout D., Thompson E.M. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol. Cell. Biol*. 2004;24(17):7795-7805. DOI 10.1128/MCB.24.17.7795-7805.2004.

Gregory T.R., Nicol J.A., Tamm H., Kullman B., Kullman K., Leitch I.J., Murray B.G., Kapraun D.F., Greilhuber J., Bennett M.D. Eukaryotic genome size databases. *Nucleic Acids Res*. 2007;35(Suppl. 1): D332-D338. DOI 10.1093/nar/gkl828.

Grudniewska M., Mouton S., Grelling M., Wolters A.H.G., Kuipers J., Giepmans B.N.G., Berezikov E. A novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. *Sci. Rep*. 2018; 8(1):1-10. DOI 10.1038/s41598-018-21107-4.

Grudniewska M., Mouton S., Simanov D., Beltman F., Grelling M., de Mulder K., Arindrarto W., Weissert P.M., van der Elst S., Berezikov E. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife*. 2016;5:e20607. DOI 10.7554/eLife.20607.

Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36(9):2896-2898. DOI 10.1093/bioinformatics/btaa025.

Hastings K.E.M. SL trans-splicing: easy come or easy go? *Trends Genet*. 2005;21(4):240-247. DOI 10.1016/j.tig.2005.02.005.

Krause M., Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell*. 1987;49(6):753-761. DOI 10.1016/0092-8674 (87)90613-1.

Ladurner P., Egger B., De Mulder K., Pfister D., Kuales G., Salvenmoser W., Schärer L. The stem cell system of the basal flatworm *Macrostomum lignano*. In: Bosch T.C.G. (Ed.). Stem Cells: From Hydra to Man. Dordrecht: Springer, Netherlands, 2008;75-94. DOI 10.1007/978-1-4020-8274-0_5.

Lasda E.L., Blumenthal T. Trans-splicing. *Wiley Interdiscip. Rev. RNA*. 2011;2(3):417-434. DOI 10.1002/wrna.71.

Lei Q., Li C., Zuo Z., Huang C., Cheng H., Zhou R. Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol*. 2016;8(3):562-577. DOI 10.1093/gbe/evw025.

Liou R.F., Blumenthal T. trans-spliced *Caenorhabditis elegans* mRNAs retain trimethylguanosine caps. *Mol. Cell. Biol*. 1990;10(4):1764-1768.

Matsumoto J., Dewar K., Wasserscheid J., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Satou Y., Hastings K.E.M. High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. *Genome Res.* 2010;20(5):636-645. DOI 10.1101/gr.100271.109.

Mouton S., Grudniewska M., Glazenburg L., Guryev V., Berezikov E. Resilience to aging in the regeneration-capable flatworm *Macrostomum lignano*. *Aging Cell*. 2018;17(3):e12739. DOI 10.1111/acel. 12739.

Protasio A.V., Tsai I.J., Babbage A., Nichol S., Hunt M., Aslett M.A., Silva N.D., Velarde G.S., Anderson T.J.C., Clark R.C., Davidson C., Dillon G.P., Holroyd N.E., LoVerde P.T., Lloyd C., McQuillan J., Oliveira G., Otto T.D., Parker-Manuel S.J., Quail M.A., Wilson R.A., Zerlotini A., Dunne D.W., Berriman M. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis*. 2012;6(1):e1455. DOI 10.1371/journal.pntd.0001455.

Rossi A., Ross E.J., Jack A., Sánchez Alvarado A. Molecular cloning and characterization of SL3: A stem cell-specific SL RNA from the planarian *Schmidtea mediterranea*. *Gene*. 2014;533(1):156-167. DOI 10.1016/j.gene.2013.09.101.

Satou Y., Mineta K., Ogasawara M., Sasakura Y., Shoguchi E., Ueno K., Yamada L., Matsumoto J., Wasserscheid J., Dewar K., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Hastings K.E.M., Lemaire P., Lindquist E., Endo T., Hotta K., Inaba K. Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol*. 2008;9(10):R152. DOI 10.1186/gb-2008-9-10-r152.

Stover N.A., Kaye M.S., Cavalcanti A.R.O. Spliced leader trans-splicing. *Curr. Biol*. 2006;16(1):R8-R9. DOI 10.1016/j.cub.2005.12.019.

Wagner D.E., Wang I.E., Reddien P.W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*. 2011;332(6031):811-816. DOI 10.1126/science.1203983.

Wasik K., Gurtowski J., Zhou X., Ramos O.M., Delás M.J., Battistoni G., Demerdash O.E., Falciatori I., Vizoso D.B., Smith A.D., Ladurner P., Schärer L., McCombie W.R., Hannon G.J., Schatz M. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl. Acad. Sci. USA*. 2015;112(40): 12462-12467. DOI 10.1073/pnas.1516718112.

Wudarski J., Egger B., Ramm S.A., Schärer L., Ladurner P., Zadesenets K.S., Rubtsov N.B., Mouton S., Berezikov E. The free-living flatworm *Macrostomum lignano*. *EvoDevo*. 2020;11(1):5. DOI 10.1186/ s13227-020-00150-1.

Wudarski J., Simanov D., Ustyantsev K., de Mulder K., Grelling M., Grudniewska M., Beltman F., Glazenburg L., Demircan T., Wunderer J., Qi W., Vizoso D.B., Weissert P.M., Olivieri D., Mouton S., Guryev V., Aboobaker A., Schärer L., Ladurner P., Berezikov E. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nat. Commun*. 2017; 8(1):2120. DOI 10.1038/s41467-017-02214-8.

Wudarski J., Ustyantsev K., Glazenburg L., Berezikov E. Influence of temperature on development, reproduction and regeneration in the flatworm model organism, *Macrostomum lignano*. *Zool. Lett*. 2019; 5(1):7. DOI 10.1186/s40851-019-0122-6.

Xie H., Hirsh D. *In vivo* function of mutated spliced leader RNAs in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*. 1998;95(8): 4235-4240.

Zadesenets K.S., Schärer L., Rubtsov N.B. New insights into the karyotype evolution of the free-living flatworm *Macrostomum lignano* (Platyhelminthes, Turbellaria). *Sci. Rep*. 2017;7(1):6066. DOI 10.1038/s41598-017-06498-0.

Zayas R.M., Bold T.D., Newmark P.A. Spliced-leader trans-splicing in freshwater planarians. *Mol. Biol. Evol*. 2005;22(10):2048-2054. DOI 10.1093/molbev/msi200.

Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003;31(13):3406-3415. DOI 10.1093/nar/gkg595.

ORCID ID

K.V. Ustyantsev orcid.org/0000-0003-4346-3868
E.V. Berezikov orcid.org/0000-0002-1145-2884

Acknowledgements. A part of work on SL motifs discovery and SL RNA gene mapping was done by K. Ustyantsev at the Institute of Cytology and Genetics SB RAS and supported by the budget project No. 0259-2021-0009. The rest of the study was performed by K. Ustyantsev and E. Berezikov at the Institute of Cytology and Genetics SB RAS and supported by the Russian Science Foundation grant No. 20-14-00147 to E. Berezikov.

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

Received October 17, 2020. Revised December 3, 2020. Accepted December 8, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):108-116
DOI 10.18699/VJ21.013

Original Russian text www.bionet.nsc.ru/vogis/

Macrostomum lignano as a model to study the genetics and genomics of parasitic flatworms

K.V. Ustyantsev, V.Yu. Vavilova, A.G. Blinov, E.V. Berezikov ✉

Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ eberez@bionet.nsc.ru

Abstract. Hundreds of millions of people worldwide are infected by various species of parasitic flatworms. Without treatment, acute and chronical infections frequently lead to the development of severe pathologies and even death. Emerging data on a decreasing efficiency of some important anthelmintic compounds and the emergence of resistance to them force the search for alternative drugs. Parasitic flatworms have complex life cycles, are laborious and expensive in culturing, and have a range of anatomic and physiological adaptations that complicate the application of standard molecular-biological methods. On the other hand, free-living flatworm species, evolutionarily close to parasitic flatworms, do not have the abovementioned difficulties, which makes them potential alternative models to search for and study homologous genes. In this review, we describe the use of the basal free-living flatworm *Macrostomum lignano* as such a model. *M. lignano* has a number of convenient biological and experimental properties, such as fast reproduction, easy and non-expensive laboratory culturing, optical body transparency, obligatory sexual reproduction, annotated genome and transcriptome assemblies, and the availability of modern molecular methods, including transgenesis, gene knockdown by RNA interference, and *in situ* hybridization. All this makes *M. lignano* amenable to the most modern approaches of forward and reverse genetics, such as transposon insertional mutagenesis and methods of targeted genome editing by the CRISPR/Cas9 system. Due to the availability of an increasing number of genome and transcriptome assemblies of different parasitic flatworm species, new knowledge generated by studying *M. lignano* can be easily translated to parasitic flatworms with the help of modern bioinformatic methods of comparative genomics and transcriptomics. In support of this, we provide the results of our bioinformatics search and analysis of genes homologous between *M. lignano* and parasitic flatworms, which predicts a list of promising gene targets for subsequent research.
Key words: flatworms; parasitic flatworms; model organism.

For citation: Ustyantsev K.V., Vavilova V.Yu., Blinov A.G., Berezikov E.V. *Macrostomum lignano* as a model to study the genetics and genomics of parasitic flatworms. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):108-116. DOI 10.18699/VJ21.013

Macrostomum lignano как модельный объект для исследования генетики и геномики паразитических плоских червей

К.В. Устьянцев, В.Ю. Вавилова, А.Г. Блинов, Е.В. Березиков ✉

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ eberez@bionet.nsc.ru

Аннотация. Инфекциям различных видов паразитических плоских червей подвержены сотни миллионов человек по всему миру. Как острые, так и хронические инфекции в отсутствие лечения с высокой частотой приводят к развитию тяжелых патологий и даже к смерти. Данные о снижении эффективности некоторых важных противогельминтных лекарственных препаратов и развитии резистентности к ним вынуждают исследователей искать альтернативные соединения. Паразитические плоские черви обладают сложным жизненным циклом, трудоемки и дорогостоящи в разведении, а также имеют ряд приспособлений, осложняющих работу с ними стандартными молекулярно-биологическими методами. Напротив, эволюционно близкородственные паразитическим плоским червям свободноживущие виды плоских червей лишены вышеописанных трудностей, что делает их перспективными альтернативными модельными объектами для поиска и исследования гомологичных генов. В этом обзоре мы описываем применение базального свободноживущего плоского червя *Macrostomum lignano* в качестве такой модели. *M. lignano* обладает большим набором удобных биологических и экспериментальных особенностей, таких как быстрое время репродукции, дешевизна и легкость в лабораторном разведении, оптическая прозрачность тела, облигатное половое размножение, аннотированные геномные и транскриптомные сборки, а также доступность современных молекулярных методов исследования, включая трансгенез, генный нокдаун с помощью РНК-интерференции и гибридизацию *in situ*. Все

это делает *M. lignano* пригодным для применения самых современных подходов «прямой» и «обратной» генетики, таких как транспозонный инсерционный мутагенез и методы направленного редактирования генома с использованием системы CRISPR/Cas9. Благодаря растущему количеству доступных сборок геномов и транскриптомов различных видов паразитических плоских червей новые знания, полученные в исследованиях на *M. lignano*, могут быть легко транслированы на паразитических плоских червей с применением современных биоинформационных подходов сравнительной геномики и транскриптомики. В подтверждение этому мы приводим результаты нашего биоинформационного поиска и анализа гомологичных генов *M. lignano* и паразитических плоских червей, которые позволили определить список перспективных генов-мишеней для дальнейшего исследования.
Ключевые слова: плоские черви; паразитические черви; модельный организм.

Introduction

Hundreds of millions of people worldwide are infected by various species of parasitic flatworms (Waikagul et al., 2018). The highest frequency of infections, as well the most severe pathologies, are induced by the species of the class Trematoda, or liver flukes, which cause such well-known diseases as schistosomiasis, clonorchiasis, and opisthorchiasis. Characteristic severe effects of the liver flukes infections are acute and chronic inflammation of liver and biliary tract, which can develop into liver fibrosis and cholangiocarcinoma, respectively (Wongratanacheewin et al., 2003; Kaewpitoon et al., 2008; Andrade, 2009; Pomaznoy et al., 2016; Schwartz, Fallon, 2018). Infections of another class of parasitic flatworms, Cestoda, or tape worms, often do not lead to such severe pathologies and death, but in the long-term perspective and without treatment they can lead to significant aberrations in vital activity and as a consequence a decrease in life quality of sick people (Budke et al., 2009; Waikagul et al., 2018).

In the world, for more than 40 years praziquantel and its derivatives have been the "number one" drugs against helminthiases (Chai, 2013; Pakharukova et al., 2015). However, continuous and widespread use of praziquantel has already resulted in the increasing number of reports on emerging resistance to the drug in different species of helminthes (Botros, Bennett, 2007; Wang et al., 2012; Mwangi et al., 2014; Jesudoss Chelladurai et al., 2018). An induced resistance to praziquantel was experimentally demonstrated in some schistosomes (Mwangi et al., 2014). Inital successes of praziquantel slowed down investments into the development of new anthelmintic drugs, which further complicates the situation. At the same time, the developed alternatives to praziquantel demonstrate analogous or sometimes even lower efficiency, more side effects, and usually are effective only against certain trematode species (Siqueira et al., 2017). Therefore, there is an urgent need for new and more effective anthelmintic drugs.

Parasitic flatworms have complex life cycles with several changes of the hosts (Morand et al., 1995; Poulin, Cribb, 2002), are laborious and expensive in laboratory culturing, and have numerous specific adaptations that complicate their study by standard molecular techniques. All these properties, undoubtedly, slow down fast development of new anthelmintic drugs. Our knowledge on a broad spectrum of biological questions was gained via research on convenient model organisms, such as nematodes, fruit flies, mice, yeast, etc. Similarly, studies of free-living animals help to obtain new information about their parasitic relatives. For example, investigating model free-living roundworm (nematode) *Caenorhabditis elegans*, new data were obtained, which allowed description of a more detailed mechanism of action for some anti-nematode drugs, as well as helped the search for new genes potentially regulating the life cycle of parasitic nematodes. Subsequently, these genes can be used as targets for developing new drugs (Cully et al., 1994; Couthier et al., 2004; Guest et al., 2007; Laing et al., 2010). Among flatworms, free-living species can be used as models to screen for new drugs directed against their parasitic relatives (Collins, Newmark, 2013). Despite fundamental differences in the life cycles, free-living flatworms have a set of evolutionary conserved properties of their physiology and reproduction, which are shared with parasitic species.

In this study, we describe the properties, advantages, and potential application of the free-living flatworm *Macrostomum lignano* as a convenient research model for efficient screening of conserved genes homologous to the genes of parasitic flatworms, which can serve as targets for the development of new anthelmintic drugs.

General properties of *Macrostomum lignano* as a model

Macrostomum lignano is a free-living flatworm (phylum Platyhelminthes, class Rhabditophora) from a basal (the earliest branching) clade – Macrostomorpha (Ladurner et al., 2005; Egger et al., 2015). *M. lignano* can easily tolerate a wide range of different environmental conditions, such as temperature, salinity, and oxygen concentration (Rivera-Ingraham et al., 2013, 2016; Wudarski et al., 2019). It was experimentally demonstrated that the worms can survive at the temperatures between 4 to 37 °C (Wudarski et al., 2019). *M. lignano* is easy to culture in laboratory conditions (Wudarski et al., 2020). The size of adult animals varies from 1 to 3 mm in length and 0.3 mm in width. Worms are maintained in Petri dishes with artificial sea water. A species of unicellular diatom algae *Nitzschia curvilineata*, which is itself easy to culture in laboratory conditions under artificial illumination, is used as food source. In one standard (9 cm) Petri dish, 500–600 individuals can be easily simultaneously maintained. Standard cultivation temperatures are 20 °C and 14/10 hours day/night light cycle.

Free-living flatworms are famous for their high regeneration capacity (Egger et al., 2006; Mouton et al., 2018; Ivankovic et al., 2019). The known champions are planarians, which can restore a full-grown animal from just a few cells (Wagner et al., 2011). *M. lignano* is nearly as regenerative as planarians, and can fully regenerate its body posterior from the pharynx and anterior to the brain (Egger et al., 2006). Flatworm regeneration comes from division and differentiation of somatic stem cell population called neoblasts (Wagner et al., 2011). Neoblasts and their differentiating progenitors are the only dividing cells in flatworms, and, apart from regeneration, they are also responsible for the natural tissue renewal during homeostasis (Nimeth et al., 2002; Ladurner et al., 2008). Importantly, there are also neoblast-like cells in parasitic flatworms, which are morphologically similar to neoblasts described in free-living species (Brehm, 2010; Collins, Newmark, 2013; Collins et al., 2013; McCusker et al., 2016). Neoblast-like cells can differentiate into other cell types and are responsible for regeneration of lost body parts in parasitic flatworms, as well as have similar transcriptional profiles to neoblasts from free-living species. Thus, there is an obvious homology of central systems of homeostasis and regeneration between free-living and parasitic flatworms.

An important advantage of *M. lignano* compared to other popular free-living model flatworms – planarians – is its body transparency (Ivankovic et al., 2019; Wudarski et al., 2020). This substantially facilitates morphological studies of its internal structures with the help of light microscopy. *M. lignano* is an obligatory reciprocal hermaphrodite, favorably distinguishing it from planarians, which in laboratory conditions reproduce predominantly asexually through fission, and are also genetically mosaic even within an individual (Schärer, Ladurner, 2003; Leria et al., 2019). Obligatory sexual reproduction of *M. lignano* allows its application in controlled genetic studies.

Currently, the presence of a simple and efficient method for transgenesis is the unique feature of *M. lignano* among other flatworm species (Wudarski et al., 2017). *M. lignano* lays 1–2 single cell eggs per day. Eggs are large (~100 μm), have relatively hard shells, and can be easily manipulated with the help of plastic microtools. These properties allowed the development of a successful protocol for delivery of various genetic constructs (DNA, mRNA, proteins) inside the eggs by means of microinjection (Wudarski et al., 2017, 2020). To date, there is a range of *M. lignano* transgenic lines which express genes of reporter green and red fluorescent proteins in different organs and tissues, allowing to study the place and dynamics of expression of a gene of interest *in vivo* (Wudarski et al., 2017, 2019).

Apart from transgenesis, other classical molecular and cytological methods are successfully applied in *M. lignano*. Localization of a gene of interest expression can be studied by means of *in situ* hybridization (Pfister et al., 2007; Grudniewska et al., 2016; Wudarski et al., 2017; Lengerer et al., 2018). To identify gene function, there is a very simple and efficient protocol for knockdown of gene expression by RNA interference, and there is no need for special delivery of double-stranded (dsRNA) constructs – worms are simply soaked in dsRNA solution and after 1–3 weeks, due to the transparency of *M. lignano*, it is possible to observe occurred morphological, physiological, or behavior changes (Grudniewska et al., 2016, 2018; Lengerer et al., 2018; Wudarski et al., 2019). Thus, the available experimental methods allow implementation of complex studies on the expression and gene function in *M. lignano*.

Any modern model organism needs a well-assembled genome and transcriptome assembly with annotation of genes and repetitive sequences, transposons and simple/tandem repeats. *M. lignano* is not an exception (Wasik et al., 2015; Grudniewska et al., 2016, 2018; Wudarski et al., 2017; Biryukov et al., 2020). *M. lignano* has a relatively compact genome of ~500 Mb. Genome and transcriptome assemblies can be openly accessed and viewed using the convenient web-interface http://gb.macgenome.org/ (Wudarski et al., 2017; Grudniewska et al., 2018). We already know genes that are differentially expressed specifically in neoblasts and the worm germline (Grudniewska et al., 2016, 2018). Thus, *M. lignano* can be used for computational analysis of evolution, comparative genomics and transcriptomics to search for conserved genes homologous to parasitic flatworms. Main properties of *M. lignano*, planarians, and parasitic flatworms are summarized in the Table.

Specific features of *M. lignano* as a model to search for gene targets regulating germline development and function in parasitic flatworms

Development of acute and chronic inflammation is an important hallmark of trematode-caused pathologies, which are caused by constant egg laying of the parasites, leading to the activation of the immunological response, which is especially relevant to schistosomiasis (Wongratanacheewin et al., 2003; Kaewpitoon et al., 2008; Collins, Newmark, 2013; Schwartz, Fallon, 2018). Thus, the germline of helminths and genes that control its development and homeostasis appear as promising targets for the development of new drugs directed to suppress their expression.

In a recent work on *M. lignano* (Grudniewska et al., 2018) it was shown that the majority of its genes classified as germline-specific are flatworm-specific (both for free-living and parasitic species) and lack a homolog in human and other model organisms. Investigation of flatworm-specific genes can be the key to search for new anthelmintic drugs with fewer side effects due to their target action on the gene products absent in humans. *M. lignano* is a convenient model to screen for such targets. As mentioned earlier, all organs of its reproductive system are clearly distinguishable under a common light dissecting microscope. This significantly facilitates the screening of phenotypes linked to the disruption of genes active in gonads and/or copulative organs (Grudniewska et al., 2018). Importantly, the worm

Comparison of key properties of free-living flatworms *M. lignano* and planarians, and parasitic flatworms as model organisms

Properties	*M. lignano*	Planarians	Parasitic flatworms
General properties			
Cost of culturing	Cheap	Cheap	Expensive
Laboriousness of culturing	Easy	Easy	Hard
In vitro culturing	Yes	Yes	Possible, but hard
Life cycle	Simple, no metamorphosis	Simple, no metamorphosis	Complex, with changing of several hosts and larvae stages
Reproduction type	Only sexual, cross fertilization	Asexual and sexual	Asexual and sexual
Suitable for controlled genetic studies	Yes	No, laboratory lines mostly reproduce asexually	No, sexual reproduction occurs within the host and uncontrollable (Richards, 1975)
Body transparency	Yes	No, strong pigmentation	Varies between species and different stages of the life cycle
Availability of annotated genome and transcriptome assemblies	Yes (Wudarski et al., 2017; Grudniewska et al., 2018)	Yes (Grohme et al., 2018)	Yes (Berriman et al., 2009; Zheng et al., 2013; Ershov et al., 2019)
Available research methods			
Transgenesis	Yes, microinjections into single-cell stage eggs (Wudarski et al., 2017, 2020)	No	Hard and inefficient, transgene inheritance was never shown: electroporation or microinjections into adults (Beckmann, Grevelding, 2012; Moguel et al., 2015)
RNA interference	Yes, immersion in dsRNA solution (Wudarski et al., 2020a, b)	Yes, injection of dsRNA, feeding with dsRNA-containing food (Rouhana et al., 2013)	Yes, efficient dsRNA delievery by electroporation, microbombardment, lipofection at all stages of the life cycle (McGonigle et al., 2008; Pierson et al., 2010; Da'dara, Skelly, 2015)
In situ hybridization	Yes (Wudarski et al., 2020)	Yes (Rouhana et al., 2013)	Yes (Cogswell et al., 2011)

hermaphroditism will allow maintaining in populations genetic aberrations linked to the activity of either male or female reproductive systems. Disturbances in fertility will already be detectable within a week at 25 °C (Wudarski et al., 2019), which will help not to miss mutations in the absence of a clear morphological phenotype.

Main methods and application of *M. lignano* for comparative genomics

Now we are already at the beginning of the era of targeted genome editing that started with the wide spread of CRISPR/Cas9 technology (Anzalone et al., 2020). Given a well-annotated genome assembly, it is possible to introduce mutations to a certain gene of interest, which would lead to complete disruption of its function (knockout) (Chen et al., 2014). Of particular interest is insertion of marker reporter sequences (e. g. fluorescent proteins) directly in the open reading frame of a target gene (knockin), which allows direct visualization of the gene expression pattern by the localization of the encoded protein (Albadri et al., 2017; Artegiani et al., 2020). For example, by combining labeling of several proteins by different fluorescent proteins, interactome studies are possible.

The function of CRISPR/Cas9 depends on only two (in the case of knockouts) or three (in the case of knockins) components: guide RNA, Cas9 nuclease protein, and a matrix for homologous recombination. In the simplest scenario, these are two plasmid vectors, one of which encodes guide RNA and Cas9, and the other is the matrix for homologous recombination (Hsu et al., 2014). Alternatively, this can be a combination of *in vitro* synthesized guide RNA and Cas9 in the form of mRNA or Cas9 protein in the complex with the guide RNA, which eliminates the possibility for unwanted insertion of the plasmid vector (Hsu et al., 2014; Kim et al., 2014). Successful and reproducible application of CRISPR/Cas9 is impossible without an efficient delivery

of genetic constructs (DNA, mRNA or proteins). Currently, *M. lignano* is the only flatworm for which this is possible by means of microinjection into single-cell stage eggs of the worm (Wudarski et al., 2017). Such an approach is certainly the most effective, since all the components of the systems are delivered simultaneously in the required molar ratio at the single-cell stage, which decreases chances for mosaic progeny. Although currently there are no published data on the application of CRISPR/Cas9 in *M. lignano*, our preliminary experiments show that this approach can be efficiently applied for a knockin introduction in the *M. lignano* genome.

Studies of phenotypes after targeted disruption/labeling of a gene of interest are characteristic of reverse genetics methods (Pareek et al., 2018). The main disadvantage of this approach is that a high-quality assembly and the annotation of the genome are required for the correct selection of the modification site and the preliminary assessment of the gene function based on its homology to already known proteins (Skromne, Prince, 2008). Moreover, genome editing by CRISPR/Cas9 depends on how frequently a GG pattern occurs in the genome, as the Cas9 protein must first detect a PAM-site (Protospacer Adjacent Motif) NGG in the target sequence (Hsu et al., 2014). An additional problem is that different guide RNAs vary significantly in their efficiency of double-strand break induction, and it is rarely possible to exactly predict the efficiency during the *in silico* design (Chuai et al., 2017). While classical models, such as human cell lines, mouse, *Drosophila*, the nematode *C. elegans*, and yeasts are thoroughly studied and there are enough data on their gene function to predict a phenotype, and their genomic GC-content is optimal, the situation with alternative models is different.

The function of a gene is rarely known, as it can be conserved only within a certain evolutionary taxon (e. g. the case of flatworm germline-specific genes). The genome can have a low GC-content, less than 40 %, which lowers the probability to meet a GG in the target regions that could be mutated to result in the target gene knockout (Casandra et al., 2018). In such cases, one should follow a historically earlier approach of forward genetics: from a phenotype to the gene (Pareek et al., 2018).

Transposon insertional mutagenesis is the most developed tool among the methods of forward genetics. Compared to chemical mutagens, which induce mutations throughout the genome but require significant time to map the mutation, a transposon movement and its insertion place can be easily detected by modern methods within one-two days (Potter, Luo, 2010; Frøkjær-Jensen et al., 2012; Stefano et al., 2016; Kalendar et al., 2019). This is achieved because the transposon sequence is originally not present in the studied genome; various promoters, enhancers, and gene trapping reporter constructs can be put in the transposon to additionally report on its insertion as well (Bonin, Mann, 2004; Song et al., 2012; Chang et al., 2019). In a recent study on the malaria parasite, it was transposon mutagenesis using the *piggyBac* DNA transposon that allowed to create 38,000 mutants of the plasmodium, and in these mutants 2680 genes regulating the parasite reproduction in blood cells were identified (Casandra et al., 2018). The authors note that it was not possible to apply CRISPR/Cas9 due to anomalously low GC-content (< 20 %) of the plasmodium genome. *M. lignano* and other flatworms, including parasitic ones, are now far from being classical and ubiquitously used model objects. As mentioned above, genes specific to the germline of flatworms mostly lack a homolog in other animals, eliminating the predictive power of the reverse genetics methods. Thus, transposon mutagenesis appears to be the most promising approach to search for the genes regulating flatworm germline, as well as other flatworm-specific genes controlling other functions, and the development of an efficient protocol for transposon mutagenesis in *M. lignano* is warranted.

Importantly, new knowledge gained from experiments on *M. lignano* can be transferred to parasitic flatworms due to availability of numerous assemblies of genomes and transcriptomes for the most significant parasitic species, which are accessible at the WormBase ParaSite (https://parasite.wormbase.org/index.html) database (Berriman et al., 2009; Zheng et al., 2013; Cwiklinski et al., 2015; Ershov et al., 2019). By using modern computational tools of comparative genomics and transcriptomics, it is possible to readily identify the sequences of potential target genes revealed in *M. lignano*, which are homologous in different parasitic flatworm species, and to perform their comparative and phylogenetic analyses *in silico*. This will allow to select candidate genes that will be the most conserved throughout all parasitic flatworm genomes, and (preferably) have weak homology to human genes.

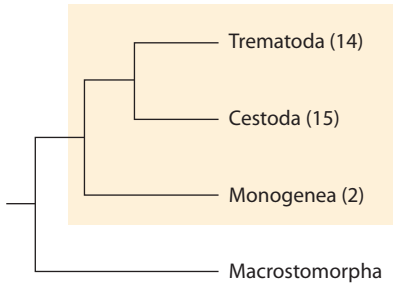
Computational analysis of conserved genes between *M. lignano* and parasitic flatworms

From the WormBase ParaSite database, amino acid sequences of protein-coding genes from 31 parasitic flatworm species were retrieved: 14 species from the class Trematoda, 15 species from Cestoda, and 2 species from Monogenea (see Figure, *a*, Supplementary 1)[1].

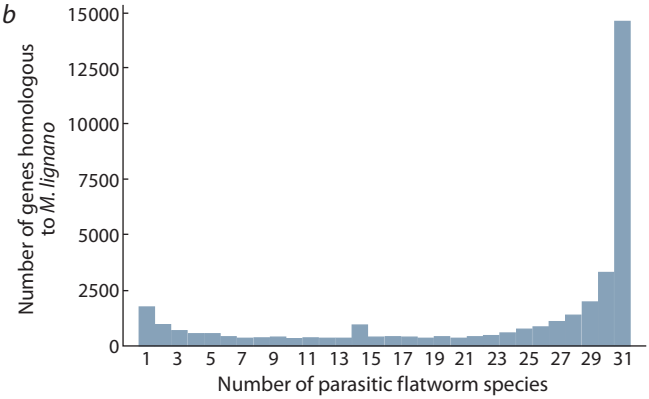
Among 60,170 protein-coding sequences from *M. lignano*, 37,113 homologs to at least one species of parasitic flatworms were found, and 14,576 homologs were identified for all the 31 species (median – 29 species) (see Figure, *b*, Supplementary 1 and 2). The summary of *M. lignano* homologs distribution among the species of parasitic flatworm classes is shown in Figure, *c* and in the Supplementary 2. We found 2887 protein-coding genes conserved between all three classes of parasitic flatworms, but lacking a human homolog, among which 18 genes are specific for neoblasts and 56 genes are specific for the germline of *M. lignano*, respectively (Grudniewska et al., 2018). These genes appear as the most promising candidates for further studies by experimental methods of reverse genetics.

[1] Supplementary materials 1–2 are available in the online version of the paper: http://www.macgenome.org/download/pdf/Ustyantsev_2021/

a



b



c

Categories of *M. lignano* homologs	Trematoda	Cestoda	Monogenea	All taxa
Total homologs	34 844	34 035	29 307	27 889
Conservative in flatworms	5895	5527	3525	2887
In neoblasts	37	34	19	18
In gonads	129	130	70	56
Conservative in human	28 949	28 508	25 782	25 002
In neoblasts	693	695	644	637
In gonads	554	561	502	486

Homology of genes in *M. lignano* and parasitic flatworm species.

a – phylogenetic relationships between *M. lignano* (Macrostomorpha) and parasitic flatworm classes according to (Park et al., 2007). Number of species in WormBase ParaSite database used in the analysis is shown in parentheses next to the taxa names; *b* – distribution of homologous genes among the number of the studied parasitic flatworm species; *c* – distribution of *M. lignano* homologous genes among parasitic flatworm classes. Number of homologs found at least in one species of each class is shown in the "All taxa" column.

Conclusion

In this study, we highlighted the key properties of free-living flatworm *M. lignano* as a model organism, and those that make it a promising object for fast and efficient screening of potential anthelmintic drugs. The availability of easy to implement transgenesis in *M. lignano* opens access to the whole arsenal of the modern methods in molecular biology to study gene functions, and its body transparency allows *in vivo* monitoring of phenotypical changes caused by gene disruption or labeling by methods of forward and reverse genetics without additional manipulations. Genes regulating development and germline functioning in flatworms appear as the most promising targets, since they are conserved among flatworms and have no homologs in human.

References

Albadri S., Del Bene F., Revenu C. Genome editing using CRISPR/Cas9-based knock-in approaches in zebrafish. *Methods*. 2017;121-122:77-85. DOI 10.1016/j.ymeth.2017.03.005.

Andrade Z.A. Schistosomiasis and liver fibrosis. *Parasite Immunol*. 2009;31:656-663. DOI 10.1111/j.1365-3024.2009.01157.x.

Anzalone A.V., Koblan L.W., Liu D.R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol*. 2020;38:824-844. DOI 10.1038/s41587-020-0561-9.

Artegiani B., Hendriks D., Beumer J., Kok R., Zheng X., Joore I., Chuva de Sousa Lopes S., van Zon J., Tans S., Clevers H. Fast and efficient generation of knock-in human organoids using homology-independent CRISPR-Cas9 precision genome editing. *Nat. Cell Biol*. 2020;22:321-331. DOI 10.1038/s41556-020-0472-5.

Beckmann S., Grevelding C.G. Paving the way for transgenic schistosomes. *Parasitology*. 2012;139:651-668. DOI 10.1017/S0031182011001466.

Berriman M., Haas B.J., LoVerde P.T., Wilson R.A., Dillon G.P., Cerqueira G.C., Mashiyama S.T., Al-Lazikani B., Andrade L.F., Ashton P.D., Aslett M.A., Bartholomeu D.C., Blandin G., Caffrey C.R., Coghlan A., Coulson R., Day T.A., Delcher A., DeMarco R., Djikeng A., Eyre T., Gamble J.A., Ghedin E., Gu Y., Hertz-Fowler C., Hirai H., Hirai Y., Houston R., Ivens A., Johnston D.A., Lacerda D., Macedo C.D., McVeigh P., Ning Z., Oliveira G., Overington J.P., Parkhill J., Pertea M., Pierce R.J., Protasio A.V., Quail M.A., Rajandream M.-A., Rogers J., Sajid M., Salzberg S.L., Stanke M., Tivey A.R., White O., Williams D.L., Wortman J., Wu W., Zamanian M., Zerlotini A., Fraser-Liggett C.M., Barrell B.G., El-Sayed N.M. The genome of the blood fluke *Schistosoma mansoni*. *Nature*. 2009;460:352-358. DOI 10.1038/nature08160.

Biryukov M., Berezikov E., Ustyantsev K. Classification of LTR retrotransposons in the flatworm *Macrostomum lignano*. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Letters to Vavilov Journal of Genetics and Breeding*. 2020;6(2):54-59. DOI 10.18699/Letters2020-6-12.

Bonin C.P., Mann R.S. A piggyBac transposon gene trap for the analysis of gene expression and function in *Drosophila*. *Genetics*. 2004;167:1801-1811. DOI 10.1534/genetics.104.027557.

Botros S.S., Bennett J.L. Praziquantel resistance. *Expert Opin. Drug Discov.* 2007;S35-S40. DOI 10.1517/17460441.2.S1.S35.

Brehm K. *Echinococcus multilocularis* as an experimental model in stem cell research and molecular host-parasite interaction. *Parasitology*. 2010;137:537-555. DOI 10.1017/S0031182009991727.

Budke C.M., White A.C., Jr., Garcia H.H. Zoonotic Larval cestode infections: neglected, neglected tropical diseases? *PLoS Negl. Trop. Dis*. 2009;3:e319. DOI 10.1371/journal.pntd.0000319.

Casandra D., Oberstaller J., Jiang R.H.Y., Bronner I.F., Adams J.H., Rayner J.C., Brown J., Mayho M., Swanson J., Otto T.D., Li S., Zhang M., Liao X., Wang C., Udenze K., Adapa S.R. Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. *Science*. 2018;360:eaap7847. DOI 10.1126/science.aap7847.

Chai J.-Y. Praziquantel treatment in trematode and cestode infections: an update. *Infect. Chemother*. 2013;45:32-43. DOI 10.3947/ic.2013.45.1.32.

Chang H., Pan Y., Landrette S., Ding S., Yang D., Liu L., Tian L., Chai H., Li P., Li D.-M., Xu T. Efficient genome-wide first-generation phenotypic screening system in mice using the piggyBac transposon. *Proc. Natl. Acad. Sci. USA*. 2019;116:18507-18516. DOI 10.1073/pnas.1906354116.

Chen X., Xu F., Zhu C., Ji J., Zhou X., Feng X., Guang S. Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in *Caenorhabditis elegans*. *Sci. Rep.* 2014;4:7581. DOI 10.1038/srep07581.

Chuai G., Wang Q.-L., Liu Q. *In silico* meets *in vivo*: towards computational CRISPR-based sgRNA design. *Trends Biotechnol*. 2017;35:12-21. DOI 10.1016/j.tibtech.2016.06.008.

Cogswell A.A., Collins J.J., Newmark P.A., Williams D.L. Whole mount *in situ* hybridization methodology for *Schistosoma mansoni*. *Mol. Biochem. Parasitol*. 2011;178:46-50. DOI 10.1016/j.molbiopara.2011.03.001.

Collins J.J., Newmark P.A. It's no fluke: the planarian as a model for understanding schistosomes. *PLoS Pathog*. 2013;9:e1003396. DOI 10.1371/journal.ppat.1003396.

Collins J.J., Wang B., Lambrus B.G., Tharp M., Iyer H., Newmark P.A. Adult somatic stem cells in the human parasite, *Schistosoma mansoni*. *Nature*. 2013;494:476-479. DOI 10.1038/nature11924.

Couthier A., Smith J., McGarr P., Craig B., Gilleard J.S. Ectopic expression of a *Haemonchus contortus* GATA transcription factor in *Caenorhabditis elegans* reveals conserved function in spite of extensive sequence divergence. *Mol. Biochem. Parasitol*. 2004;133:241-253.

Cully D.F., Vassilatis D.K., Liu K.K., Paress P.S., Van der Ploeg L.H.T., Schaeffer J.M., Arena J.P. Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature*. 1994;371:707-711. DOI 10.1038/371707a0.

Cwiklinski K., Dalton J.P., Dufresne P.J., La Course J., Williams D.J., Hodgkinson J., Paterson S. The *Fasciola hepatica* genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biol*. 2015;16:71. DOI 10.1186/s13059-015-0632-2.

Da'dara A.A., Skelly P.J. Gene suppression in schistosomes using RNAi. In: Peacock C. (Ed.). Parasite Genomics Protocols, Methods in Molecular Biology. New York: Springer, 2015;143-164. DOI 10.1007/978-1-4939-1438-8_8.

Egger B., Ladurner P., Nimeth K., Gschwentner R., Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. *Dev. Genes Evol*. 2006;216:565-577. DOI 10.1007/s00427-006-0069-4.

Egger B., Lapraz F., Tomiczek B., Müller S., Dessimoz C., Girstmair J., Škunca N., Rawlinson K.A., Cameron C.B., Beli E., Todaro M.A., Gammoudi M., Noreña C., Telford M.J. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr. Biol*. 2015;25:1347-1353. DOI 10.1016/j.cub.2015.03.034.

Ershov N.I., Mordvinov V.A., Prokhortchouk E.B., Pakharukova M.Y., Gunbin K.V., Ustyantsev K., Genaev M.A., Blinov A.G., Mazur A., Boulygina E., Tsygankova S., Khrameeva E., Chekanov N., Fan G., Xiao A., Zhang H., Xu X., Yang H., Solovyev V., Lee S.M.-Y., Liu X., Afonnikov D.A., Skryabin K.G. New insights from *Opisthorchis felineus* genome: update on genomics of the epidemiologically important liver flukes. *BMC Genomics*. 2019;20:399. DOI 10.1186/s12864-019-5752-8.

Frøkjær-Jensen C., Davis M.W., Ailion M., Jorgensen E.M. Improved Mos1-mediated transgenesis in *C. elegans*. *Nat. Methods*. 2012;9:117-118. DOI 10.1038/nmeth.1865.

Grohme M.A., Schloissnig S., Rozanski A., Pippel M., Young G.R., Winkler S., Brandl H., Henry I., Dahl A., Powell S., Hiller M., Myers E., Rink J.C. The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature*. 2018;554:56-61. DOI 10.1038/nature25473.

Grudniewska M., Mouton S., Grelling M., Wolters A.H.G., Kuipers J., Giepmans B.N.G., Berezikov E. A novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. *Sci. Rep*. 2018;8:1-10. DOI 10.1038/s41598-018-21107-4.

Grudniewska M., Mouton S., Simanov D., Beltman F., Grelling M., de Mulder K., Arindrarto W., Weissert P.M., van der Elst S., Berezikov E. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife*. 2016;5:e20607. DOI 10.7554/eLife.20607.

Guest M., Bull K., Walker R.J., Amliwala K., O'Connor V., Harder A., Holden-Dye L., Hopper N.A. The calcium-activated potassium channel, SLO-1, is required for the action of the novel cyclo-octadepsipeptide anthelmintic, emodepside, in *Caenorhabditis elegans*. *Int. J. Parasitol*. 2007;37:1577-1588. DOI 10.1016/j.ijpara.2007.05.006.

Hsu P.D., Lander E.S., Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. 2014;157:1262-1278. DOI 10.1016/j.cell.2014.05.010.

Ivankovic M., Haneckova R., Thommen A., Grohme M.A., Vila-Farré M., Werner S., Rink J.C. Model systems for regeneration: planarians. *Development*. 2019;146. DOI 10.1242/dev.167684.

Jesudoss Chelladurai J., Kifleyohannes T., Scott J., Brewer M.T. Praziquantel resistance in the zoonotic cestode *Dipylidium caninum*. *Am. J. Trop. Med. Hyg*. 2018;99:1201-1205. DOI 10.4269/ajtmh.18-0533.

Kaewpitoon N., Kaewpitoon S.J., Pengsaa P., Sripa B. *Opisthorchis viverrini*: The carcinogenic human liver fluke. *World J. Gastroenterol*. 2008;14:666-674. DOI 10.3748/wjg.14.666.

Kalendar R., Shustov A.V., Seppänen M.M., Schulman A.H., Stoddard F.L. Palindromic sequence-targeted (PST) PCR: a rapid and efficient method for high-throughput gene characterization and genome walking. *Sci. Rep*. 2019;9:1-11. DOI 10.1038/s41598-019-54168-0.

Kim S., Kim D., Cho S.W., Kim J., Kim J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res*. 2014;24:1012-1019. DOI 10.1101/gr.171322.113.

Ladurner P., Egger B., De Mulder K., Pfister D., Kuales G., Salvenmoser W., Schärer L. The stem cell system of the basal flatworm *Macrostomum lignano*. In: Bosch T.C.G. (Ed.). Stem Cells: From Hydra to Man. Dordrecht: Springer, 2008;75-94. DOI 10.1007/978-1-4020-8274-0_5.

Ladurner P., Schärer L., Salvenmoser W., Rieger R.M. A new model organism among the lower Bilateria and the use of digital microscopy in taxonomy of meiobenthic Platyhelminthes: *Macrostomum lignano*, n. sp. (Rhabditophora, Macrostomorpha). *J. Zool. Syst. Evol. Res*. 2005;43:114-126. DOI 10.1111/j.1439-0469.2005.00299.x.

Laing S.T., Ivens A., Laing R., Ravikumar S., Butler V., Woods D.J., Gilleard J.S. Characterization of the xenobiotic response of *Cae-*

norhabditis elegans to the anthelmintic drug albendazole and the identification of novel drug glucoside metabolites. *Biochem. J.* 2010;432:505-516. DOI 10.1042/BJ20101346.

Lengerer B., Wunderer J., Pjeta R., Carta G., Kao D., Aboobaker A., Beisel C., Berezikov E., Salvenmoser W., Ladurner P. Organ specific gene expression in the regenerating tail of *Macrostomum lignano*. *Dev. Biol*. 2018;433(2):448-460. DOI 10.1016/j.ydbio.2017.07.021.

Leria L., Vila-Farré M., Solà E., Riutort M. Outstanding intraindividual genetic diversity in fissiparous planarians (*Dugesia*, Platyhelminthes) with facultative sex. *BMC Evol. Biol*. 2019;19:130. DOI 10.1186/s12862-019-1440-1.

McCusker P., McVeigh P., Rathinasamy V., Toet H., McCammick E., O'Connor A., Marks N.J., Mousley A., Brennan G.P., Halton D.W., Spithill T.W., Maule A.G. Stimulating neoblast-like cell proliferation in juvenile *Fasciola hepatica* supports growth and progression towards the adult phenotype *in vitro*. *PLoS Negl. Trop. Dis.* 2016;10:e0004994. DOI 10.1371/journal.pntd.0004994.

McGonigle L., Mousley A., Marks N.J., Brennan G.P., Dalton J.P., Spithill T.W., Day T.A., Maule A.G. The silencing of cysteine proteases in *Fasciola hepatica* newly excysted juveniles using RNA interference reduces gut penetration. *Int. J. Parasitol*. 2008;38:149-155. DOI 10.1016/j.ijpara.2007.10.007.

Moguel B., Moreno-Mendoza N., Bobes R.J., Carrero J.C., Chimal-Monroy J., Díaz-Hernández M.E., Herrera-Estrella L., Laclette J.P. Transient transgenesis of the tapeworm *Taenia crassiceps. SpringerPlus.* 2015;4:496. DOI 10.1186/s40064-015-1278-y.

Morand S., Robert F., Connors V.A. Complexity in parasite life cycles: population biology of cestodes in fish. *J. Anim. Ecol*. 1995;64:256-264. DOI 10.2307/5760.

Mouton S., Grudniewska M., Glazenburg L., Guryev V., Berezikov E. Resilience to aging in the regeneration-capable flatworm *Macrostomum lignano*. *Aging Cell*. 2018;17:e12739. DOI 10.1111/acel.12739.

Mwangi I.N., Sanchez M.C., Mkoji G.M., Agola L.E., Runo S.M., Cupit P.M., Cunningham C. Praziquantel sensitivity of Kenyan *Schistosoma mansoni* isolates and the generation of a laboratory strain with reduced susceptibility to the drug. *Int. J. Parasitol. Drugs Drug Resist*. 2014;4:296-300. DOI 10.1016/j.ijpddr.2014.09.006.

Nimeth K., Ladurner P., Gschwentner R., Salvenmoser W., Rieger R. Cell renewal and apoptosis in *Macrostomum* sp. [*Lignano*]. *Cell Biol. Int*. 2002;26:801-815. DOI 10.1006/cbir.2002.0950.

Pakharukova M.Y., Shilov A.G., Pirozhkova D.S., Katokhin A.V., Mordvinov V.A. The first comprehensive study of praziquantel effects *in vivo* and *in vitro* on European liver fluke *Opisthorchis felineus* (Trematoda). *Int. J. Antimicrob. Agents.* 2015;46:94-100. DOI 10.1016/j.ijantimicag.2015.02.012.

Pareek A., Arora A., Dhankher O.P. Stepping forward and taking reverse as we move ahead in genetics. *Ind. J. Plant Physiol*. 2018;23:609-611. DOI 10.1007/s40502-018-0428-y.

Park J.-K., Kim K.-H., Kang S., Kim W., Eom K.S., Littlewood D. A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). *BMC Evol. Biol*. 2007;7:11. DOI 10.1186/1471-2148-7-11.

Pfister D., De Mulder K., Philipp I., Kuales G., Hrouda M., Eichberger P., Borgonie G., Hartenstein V., Ladurner P. The exceptional stem cell system of *Macrostomum lignano*: Screening for gene expression and studying cell proliferation by hydroxyurea treatment and irradiation. *Front. Zool*. 2007;4:9. DOI 10.1186/1742-9994-4-9.

Pierson L., Mousley A., Devine L., Marks N.J., Day T.A., Maule A.G. RNA interference in a cestode reveals specific silencing of selected highly expressed gene transcripts. *Int. J. Parasitol*. 2010;40:605-615. DOI 10.1016/j.ijpara.2009.10.012.

Pomaznoy M.Y., Logacheva M.D., Young N.D., Penin A.A., Ershov N.I., Katokhin A.V., Mordvinov V.A. Whole transcriptome profiling of adult and infective stages of the trematode *Opisthorchis felineus. Parasitol. Int*. 2016;65:12-19. DOI 10.1016/j.parint.2015.09.002.

Potter C.J., Luo L. Splinkerette PCR for mapping transposable elements in *Drosophila. PLoS One*. 2010;5:e10168. DOI 10.1371/journal.pone.0010168.

Poulin R., Cribb T.H. Trematode life cycles: Short is sweet? *Trends Parasitol*. 2002;18:176-183. DOI 10.1016/S1471-4922(02)02262-6.

Richards C.S. Genetic studies on variation in infectivity of *Schistosoma mansoni*. *J. Parasitol*. 1975;61:233-236. DOI 10.2307/3278999.

Rivera-Ingraham G.A., Bickmeyer U., Abele D. The physiological response of the marine platyhelminth *Macrostomum lignano* to different environmental oxygen concentrations. *J. Exp. Biol*. 2013;216:2741-2751. DOI 10.1242/jeb.081984.

Rivera-Ingraham G.A., Nommick A., Blondeau-Bidet E., Ladurner P., Lignot J.-H. Salinity stress from the perspective of the energy-redox axis: Lessons from a marine intertidal flatworm. *Redox Biol*. 2016;10:53-64. DOI 10.1016/j.redox.2016.09.012.

Rouhana L., Weiss J.A., Forsthoefel D.J., Lee H., King R.S., Inoue T., Shibata N., Agata K., Newmark P.A. RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev. Dyn*. 2013;242:718-730. DOI 10.1002/dvdy.23950.

Schärer L., Ladurner P. Phenotypically plastic adjustment of sex allocation in a simultaneous hermaphrodite. *Proc. Biol. Sci*. 2003;270:935-941. DOI 10.1098/rspb.2002.2323.

Schwartz C., Fallon P.G. *Schistosoma* "eggs-iting" the host: granuloma formation and egg excretion. *Front. Immunol*. 2018;9. DOI 10.3389/fimmu.2018.02492.

Siqueira L.D.P., Fontes D.A.F., Aguilera C.S.B., Timoteo T.R.R., Angelos M.A., Silva L.C.P.B.B., de Melo C.G., Rolim L.A., da Silva R.M.F., Neto P.J.R. Schistosomiasis: Drugs used and treatment strategies. *Acta Trop.* 2017;176:179-187. DOI 10.1016/j.actatropica.2017.08.002.

Skromne I., Prince V.E. Current perspectives in zebrafish reverse genetics: moving forward. *Dev. Dyn*. 2008;237:861-882. DOI 10.1002/dvdy.21484.

Song G., Li Q., Long Y., Gu Q., Hackett P.B., Cui Z. Effective gene trapping mediated by sleeping beauty transposon. *PLoS One*. 2012;7:e44123. DOI 10.1371/journal.pone.0044123.

Stefano B., Patrizia B., Matteo C., Massimo G. Inverse PCR and quantitative PCR as alternative methods to southern blotting analysis to assess transgene copy number and characterize the integration site in transgenic woody plants. *Biochem. Genet*. 2016;54:291-305. DOI 10.1007/s10528-016-9719-z.

Wagner D.E., Wang I.E., Reddien P.W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*. 2011;332:811-816. DOI 10.1126/science.1203983.

Waikagul J., Kobayashi J., Pongvongsa T., Sato M.O., Adsakwattana P., Fontanilla I.K.C., Sato M., Fornillos R.J.C. Odds, challenges and new approaches in the control of helminthiasis, an Asian study. *Parasite Epidemiol. Control*. 2018;4:e00083. DOI 10.1016/j.parepi.2018.e00083.

Wang W., Wang L., Liang Y.-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. *Parasitol. Res*. 2012;111:1871-1877. DOI 10.1007/s00436-012-3151-z.

Wasik K., Gurtowski J., Zhou X., Ramos O.M., Delás M.J., Battistoni G., Demerdash O.E., Falciatori I., Vizoso D.B., Smith A.D., Ladurner P., Schärer L., McCombie W.R., Hannon G.J., Schatz M. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl. Acad. Sci. USA*. 2015;112:12462-12467. DOI 10.1073/pnas.1516718112.

Wongratanacheewin S., Sermswan R.W., Sirisinha S. Immunology and molecular biology of *Opisthorchis viverrini* infection. *Acta Trop*. 2003;88:195-207. DOI 10.1016/j.actatropica.2003.02.002.

Wudarski J., Egger B., Ramm S.A., Schärer L., Ladurner P., Zadesenets K.S., Rubtsov N.B., Mouton S., Berezikov E. The free-living flatworm *Macrostomum lignano*. *EvoDevo*. 2020;11:5. DOI 10.1186/s13227-020-00150-1.

Wudarski J., Simanov D., Ustyantsev K., de Mulder K., Grelling M., Grudniewska M., Beltman F., Glazenburg L., Demircan T., Wunderer J., Qi W., Vizoso D.B., Weissert P.M., Olivieri D., Mouton S., Guryev V., Aboobaker A., Schärer L., Ladurner P., Berezikov E. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nat. Commun*. 2017;8:2120. DOI 10.1038/s41467-017-02214-8.

Wudarski J., Ustyantsev K.V., Berezikov E.V. Approaches to efficient genome editing in the regenerating free-living flatworm *Macrostomum lignano*. In: Methods for Editing Genes and Genomes. Novosibirsk, 2020;101-115. (in Russian)

Wudarski J., Ustyantsev K., Glazenburg L., Berezikov E. Influence of temperature on development, reproduction and regeneration in the flatworm model organism, *Macrostomum lignano*. *Zool. Lett*. 2019;5:7. DOI 10.1186/s40851-019-0122-6.

Zheng H., Zhang W., Zhang L., Zhang Z., Li J., Lu G., Zhu Y., Wang Y., Huang Y., Liu J., Kang H., Chen J., Wang L., Chen A., Yu S., Gao Z., Jin L., Gu W., Wang Z., Zhao L., Shi B., Wen H., Lin R., Jones M.K., Brejova B., Vinar T., Zhao G., McManus D.P., Chen Z., Zhou Y., Wang S. The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat. Genet*. 2013;45:1168-1175. DOI 10.1038/ng.2757.

ORCID ID

K.V. Ustyantsev orcid.org/0000-0003-4346-3868
E.V. Berezikov orcid.org/0000-0002-1145-2884

Acknowledgements. The work on comparative analysis of the characteristics of *M. lignano*, planarians, and parasitic flatworms was supported by the budget project No. 0259-2021-0009 and done by V.V., A.B., and E.B. The search and analysis of homologous genes between *M. lignano* and parasitic flatworms, as well as the analysis of prospective methods and gene targets was done by K.U. in the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences by financial support from by Russian Science Foundation, grant No. 19-74-00029.

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

Received October 17, 2020. Revised December 3, 2020. Accepted December 8, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):117-124
DOI 10.18699/VJ21.014

Original Russian text www.bionet.nsc.ru/vogis/

A transgenic cell line with inducible transcription for studying (CGG)n repeat expansion mechanisms

I.V. Grishchenko[1], A.A. Tulupov[2, 3], Y.M. Rymareva[2], E.D. Petrovskiy[2], A.A. Savelov[2], A.M. Korostyshevskaya[2], Y.V. Maksimova[4, 5], A.R. Shorina[5], E.M. Shitik[1], D.V. Yudkin[1] ✉

[1] State Research Center of Virology and Biotechnology "Vector", Rospotrebnadzor, Koltsovo, Novosibirsk region, Russia
[2] International Tomography Center of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Novosibirsk State Medical University, Novosibirsk, Russia
[5] Novosibirsk City Clinical Hospital No. 1, Novosibirsk, Russia
✉ yudkin_dv@vector.nsc.ru

Abstract. There are more than 30 inherited human disorders connected with repeat expansion (myotonic dystrophy type I, Huntington's disease, Fragile X syndrome). Fragile X syndrome is the most common reason for inherited intellectual disability in the human population. The ways of the expansion development remain unclear. An important feature of expanded repeats is the ability to form stable alternative DNA secondary structures. There are hypotheses about the nature of repeat instability. It is proposed that these DNA secondary structures can block various stages of DNA metabolism processes, such as replication, repair and recombination and it is considered as the source of repeat instability. However, none of the hypotheses is fully confirmed or is the only valid one. Here, an experimental system for studying (CGG)n repeat expansion associated with transcription and TCR-NER is proposed. It is noteworthy that the aberrations of transcription are a poorly studied mechanism of (CGG)n instability. However, the proposed systems take into account the contribution of other processes of DNA metabolism and, therefore, the developed systems are universal and applicable for various studies. Transgenic cell lines carrying a repeat of normal or premutant length under the control of an inducible promoter were established and a method for repeat instability quantification was developed. One type of the cell lines contains an exogenous repeat integrated into the genome by the Sleeping Beauty transposon; in another cell line, the vector is maintained as an episome due to the SV40 origin of replication. These experimental systems can serve for finding the causes of instability and the development of therapeutic agents. In addition, a criterion was developed for the quantification of exogenous (CGG)n repeat instability in the transgenic cell lines' genome.
Key words: hereditary intellectual disability; fragile X syndrome; repeat expansion; transcription; replication; transgenic cell lines; somatic instability.

For citation: Grishchenko I.V., Tulupov A.A., Rymareva Y.M., Petrovskiy E.D., Savelov A.A., Korostyshevskaya A.M., Maksimova Y.V., Shorina A.R., Shitik E.M., Yudkin D.V. A transgenic cell line with inducible transcription for studying (CGG)n repeat expansion mechanisms. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):117-124. DOI 10.18699/VJ21.014

Трансгенная клеточная линия с индуцируемой транскрипцией для исследования механизмов экспансии (CGG)n повторов

И.В. Грищенко[1], А.А. Тулупов[2, 3], Ю.М. Рымарева[2], Е.Д. Петровский[2], А.А. Савелов[2], А.М. Коростышевская[2], Ю.В. Максимова[4, 5], А.Р. Шорина[5], Е.М. Шитик[1], Д.В. Юдкин[1] ✉

[1] Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р.п. Кольцово, Новосибирская область, Россия
[2] Институт «Международный томографический центр» Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Новосибирский государственный медицинский университет, Новосибирск, Россия
[5] Городская клиническая больница № 1, Новосибирск, Россия
✉ yudkin_dv@vector.nsc.ru

Аннотация. Существует ряд наследственных заболеваний человека, причиной которых является экспансия тандемных повторов. К ним относятся миотоническая дистрофия первого типа, болезнь Хантингтона, заболевания, ассоциированные с ломкой Х-хромосомой. Синдром ломкой Х-хромосомы – наиболее распространенная причина наследственной умственной отсталости у человека. На сегодняшний день причины развития экспансии остаются неисследованными. Важная особенность протяженных повторов – их способность формировать альтернативные вторичные структуры ДНК. Существуют гипотезы, объясняю-

щие природу нестабильности повторов, однако все они предполагают возникновение устойчивых вторичных структур ДНК на различных этапах клеточного цикла. Источником нестабильности считаются нарушения в различных процессах метаболизма ДНК (репликация, репарация и рекомбинация), вызванные образованием вторичных структур. Однако ни одна из гипотез до конца не подтверждена и, видимо, не является единственно верной. Вероятно, в различных типах клеток и на определенных стадиях клеточного цикла источником нестабильности выступает множество процессов. В настоящей работе мы предлагаем экспериментальную систему для изучения вклада транскрипции и ассоциированной с ней репарации в нестабильность повтора (CGG)n, поскольку это наименее изученный механизм возникновения нестабильности. Однако предложенные модели могут учитывать вклад и других процессов метаболизма ДНК, например репликации, что делает полученные системы универсальными и применимыми в разных исследованиях. Нами были созданы трансгенные клеточные линии, несущие повтор нормальной и премутантной длины под тетрациклин-индуцируемым промотором. Один тип линий содержит плазмиду с экзогенным повтором, интегрированным в геном посредством транспозона Sleeping Beauty, в другой клеточной линии вектор поддерживается в виде эписомы благодаря ориджину репликации SV40. Такие трансгенные клеточные линии могут служить экспериментальной системой для поиска причин нестабильности и создания терапевтических средств. Кроме того, был разработан критерий для оценки нестабильности экзогенного (CGG)n повтора в геноме трансгенных клеточных линий, расчет которого не зависит от эффективности синтеза протяженных повторов.

Ключевые слова: наследственная умственная отсталость; синдром ломкой Х-хромосомы; экспансия повторов; транскрипция; репликация; трансгенная клеточная линия; соматическая нестабильность.

Introduction

Repeat expansion is a unique type of mutation that is characterized by a dramatic increase of the number of triplet repeats in DNA. Triplet repeats are more prone to expansion: to date, more than 30 diseases associated with their instability are known (Grishchenko et al., 2020). Fragile X syndrome, as the most common form of hereditary intellectual disability is also based on triplet repeat expansion. The cause of the disease is the expansion of the CGG repeat located in the 5′-untranslated region of the *FMR1* gene. Normally, the repeats number is relatively stable and does not exceed 54 triplets; if the (CGG)n expansion increases up to 200 triplets, the FMR1 allele becomes premutant, and the ataxia/tremor syndromes and primary ovarian insufficiency syndrome associated with a Fragile X syndrome develop. The premutant allele frequency in the population is 1:100. Even though the clinical manifestations are often not observed, the expanded repeat can be transmitted over generations. Full mutation develops when triplets numbers increase over 200: the *FMR1* gene promoter becomes methylated, the locus is heterochromatinized, and the FMRP protein is completely lost, which leads to the development of Fragile X syndrome. FMRP is necessary for normal neuron activity and its absence causes pronounced phenotypic manifestations: macroorchidism, endocrine pathologies, cerebellum morphological changes, and intellectual disability characterized by behavior and learning problems (Roberts et al., 2003; Martin et al., 2012; Heulens et al., 2013). The full mutation frequency varies from 1:4,000 in men, and up to 1:6,000 in women.

Despite understanding the syndrome pathogenesis details, the expansion mechanism has not yet been studied. Different processes of DNA metabolism are probably able to increase the CGG-repeat instability. Therefore, the contribution of replication to the expansion processes has been established: the formed hairpin on the newly synthesized DNA strand leads to the additional replication of the region containing the (CGG)n repeat and, therefore, to its increasing (Fouche et al., 2006). However, in people suffering from repeat expansion disorders, and in model mice, expansion is also observed in tissues with low proliferative activity, including the brain lobes, oocytes, liver and muscles (Lokanga et al., 2013); it confirms the theory that the repeat expansion can also depend on other processes affecting DNA. Indeed, for many proteins of the DNA repair and recombination pathways, their probable participation in the repeat expansion process has been shown. Some experimental data indicate the MMR system components involvement in the expansion (Kovalenko et al., 2012; Zhao et al., 2016). Another possible source of instability can be transcription and transcription-coupled repair (TCR-NER), since many repeat tracts are characterized by the R-loops formation – RNA: DNA-resistant duplexes forming during RNA synthesis as well as the disruption of the initiation of PolII transcription (Krasilnikova et al., 2007). The lesions during transcription initiate TCR, a form of excisional nucleotide repair (NER). For some proteins of this cascade, correlations with the (CGG)n instability level were found. It should be noted that for the *FMR1* premutant alleles, which rapidly accumulate repeated units, a significant increase of the FMR1 transcription level was found, which probably indicates the involvement of the TCR system in the repeat instability development. However, there is no unequivocal confirmation of this hypothesis.

To study the details of all the described cascades, it is necessary to have a model in which it is possible to track all the changes occurring with repeat and surrounding regions in response to the induction of a certain DNA metabolism process. To date, similar models have already been proposed (Gorbunova et al., 2003; Kononenko et al., 2020), but none of them can directly assess the contribution of transcription to (CGG)n instability. In this study, we describe the experimental models for repeat instability research based on two types of plasmids: integrated and not integrated into the genome. These models will allow taking into account the contribution of replication, transcription, TCR-NER, and genome location to the CGG-repeat instability. In addition, this model can be used to study the repeat-induced mutagenesis observed in cells with an expanded repeat in the FMR1 promoter region (Shah, Mirkin, 2015).

Materials and methods

Ethics statements. The procedure of involving the patients in the study was strictly designed in accordance with international standards, which include the awareness of the subject, their consent to participate in the study in its entirety, and the guarantee of confidentiality. All of the studies conformed to the ethical standards developed in accordance with the Helsinki Declaration of the World Medical Association as amended in 2000. In addition, the studies were supervised by the Institutional Review Board. The written consent of the study participants was also obtained.

DNA purification and repeat sizing. Peripheral venous blood from all of the patients was collected in Novosibirsk City Clinical Hospital No. 1 into EDTA-containing tubes and frozen before DNA purification. The DNA was purified from whole venous blood and cell cultures using a Wizard® Genomic DNA Purification Kit (Promega, USA).

CGG repeats were sized using a special protocol for the GC-rich DNA amplification proposed earlier (Hayward et al., 2016). For PCR primers NewFraxC (5′-d6RG-tgctttc tagactcagctccgtttcggtttcacttccggt-3′) and NewFraxR4 (5′-taa gcagaattcccttgtagaaagcgccattggagccccgca-3′) and 0.02 units of Q5-DNA polymerase were used. The resulting fragments was separated by agarose gel electrophoresis. To assess the accurate size of the repeat, capillary electrophoresis using a 1200 LIZ length standard (AppliedBiosystems, USA) was performed. The flanking region in the PCR product is a total of 269 bp, thereby the repeat length was determined by the following equation

$$N = \frac{\text{PCR product size} - 269}{3},$$

where N – CGG-triplets number.

Cloning CGG repeats of various lengths into vector systems. The control plasmid pCDH containing no CGG-repeat consisted of the following elements: (1) doxycycline-inducible Tet-O-minimal CMV promoter, IRES sequence, open reading frame (ORF) of the GFP protein, (2) constitutive promoter EF1alpha, transactivator for Tet-O-element rTtA ORF, T2A peptide, the DsRedExpress protein ORF, (3) beta-lactamase promoter, beta-lactamase protein ORF for transformed bacterial cells selection, origin of replication, and (4) SV40 origin of replication. The PCR product carrying the CGG repeat was cloned into the pCDH plasmid at the XbaI and EcoRI restriction endonuclease sites (SibEnzyme, Russia) between the CMV minimal promoter and the IRES sequence.

Plasmid pSBi for CGG-repeat cloning was assembled from the following components: (1) beta-lactamase promoter, beta-lactamase protein ORF for transformed bacterial cells selection, origin of replication, (2) Sleeping Beauty transposon terminal repeats, (3) cassette containing a PGK promoter and a puromycin-N-acetyl transferase ORF, (4) an hPGK promoter, an rTta ORF, (5) an inducible TRE3GS promoter, and an mGFP ORF.

Cloning of the PCR product containing CGG-repeat driven by an inducible promoter was also carried out at the restriction endonuclease sites XbaI and EcoRI (SibEnzyme). For the plasmids production, the electrocompetent cells of the *E. coli* strain NebStable (NEB, USA) were transformed. It was shown that an extended repeat during the transformation of bacterial cells and their cultivation is prone to a dramatically repeat length contraction, which is consistent with the literature data (Bontekoe, 2001); therefore, the NebStable cells were cultured for a day at 20 °C to avoid the repeat size decrease. For HEK293A and HEK293T cells transfection, plasmids were isolated and purified using the QIAGEN® Plasmid Plus Maxi Kit (QIAGEN, Germany).

Eukaryotic cell transfection. HEK293A and HEK293T cells transfection was performed using the Lipofectamine 3000 reagent (Thermo Fisher Scientific, USA). The induction of the Tet-O-minimal CMV promoter and TRE3GS promoter was performed using the doxycycline with a concentration of 1 μg/ml in the cultural media.

Results

Assembly of experimental plasmids carrying CGG-repeat of normal and premutant lengths

We obtained a set of plasmids based on eukaryotic expression vectors with an inducible promoter that regulates the CGG-repeat transcription level of CGG repeat of normal or premutant length and GFP ORF. These plasmids serve as the core of the model system for studying (CGG)n repeat instability. The pCDH plasmid was used as a vector for transient expression and exogenous CGG-repeat maintenance in a non-integrated state in the genome (Fig. 1, *a*). For the integration of the exogenous CGG-repeat into the genome, a construct based on the Sleeping Beauty pSBi transposon/transposase system was assembled (see Fig. 1, *b*).

The pCDH plasmid encodes two reporter proteins: DsRedExpress driven by EF1 promoter and EGFP, whose expression is regulated by the inducible Tet-O-CMV promoter. Downstream of the Tet-O-CMV promoter, a multiple cloning site for CGG-repeat cloning is located. Due to this mutual arrangement of the inducible promoter and the site of the repeat cloning, it can be established that transcription goes through the inserted CGG-repeat due to the synthesis of EGFP mRNA. After several transcription rounds, the influence of transcription on the repeat instability can be detected. In addition, pCDH plasmid contains the SV40 origin of replication and thereby it is able to replicate in HEK293T cells that produce the SV40 large T antigen. In this case, it is possible to assess not only the contribution of transcription, but also the role of replication processes during the maintenance of the pCDH in the form of an episome.

The pSBi vector encodes an mGFP protein driven by an inducible TRE3GS promoter. Before the mGFP ORF are sites for cloning the CGG repeat. Therefore, it is possible to analyze the effect of transcription on changes in the CGG repeat length. Since this vector is based on a transposon, a part of the plasmid flanked by specific repeated sequences recognized by SB transposase and this part of plasmid can be inserted into different regions of the genome by transposase. It is possible to assess the potential influence of the integration sites on the instability of the CGG repeat by the determination of the insertion sites.

To obtain fragments containing a CGG-repeat, we used DNA samples isolated from continuous human B-lymphocytes

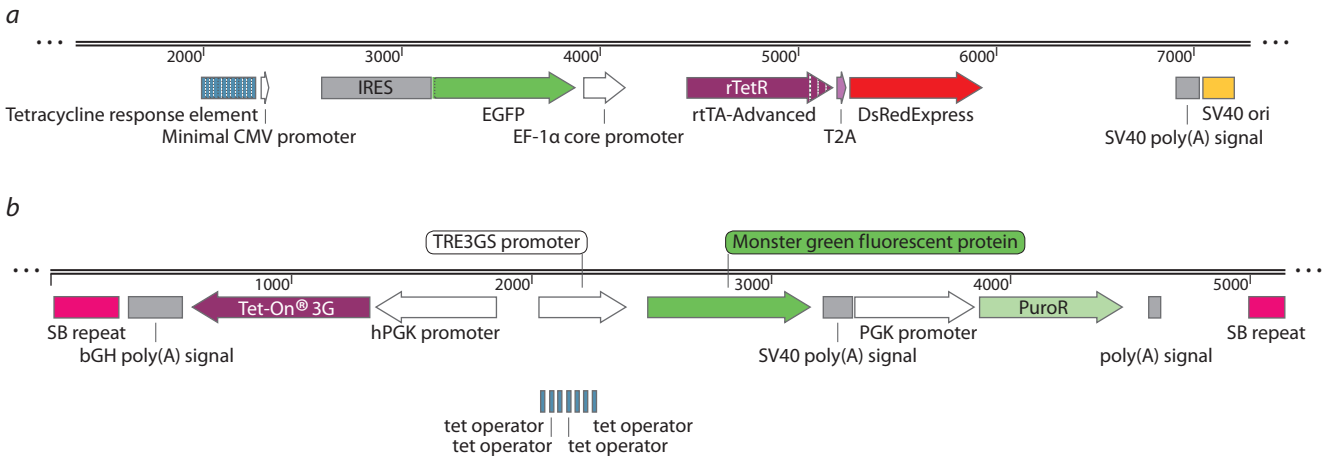


Fig. 1. Vector maps used to generate model cell lines.

a – pCDH plasmid map. IRES – internal ribosome entry site; EGFP – green fluorescent protein ORF; EF1 – constitutive promoter EF1alpha; rtTA – tetracycline/doxycycline-interacting transactivator for tetracycline response element; DsRedExpress – red fluorescent protein ORF; SV40 ori – SV40 viral origin of replication; *b* – pSBi plasmid map. SB repeat – repeat that is recognized by the Sleeping Beauty transposase; Tet-On® 3G – tetracycline/doxycycline-interacting transactivator for TRE3GS promoter; hPGK – constitutive promoter; TRE3GS – inducible promoter; PGK – constitutive promoter; PuroR – puromycin-N-acetyl transferase ORF.

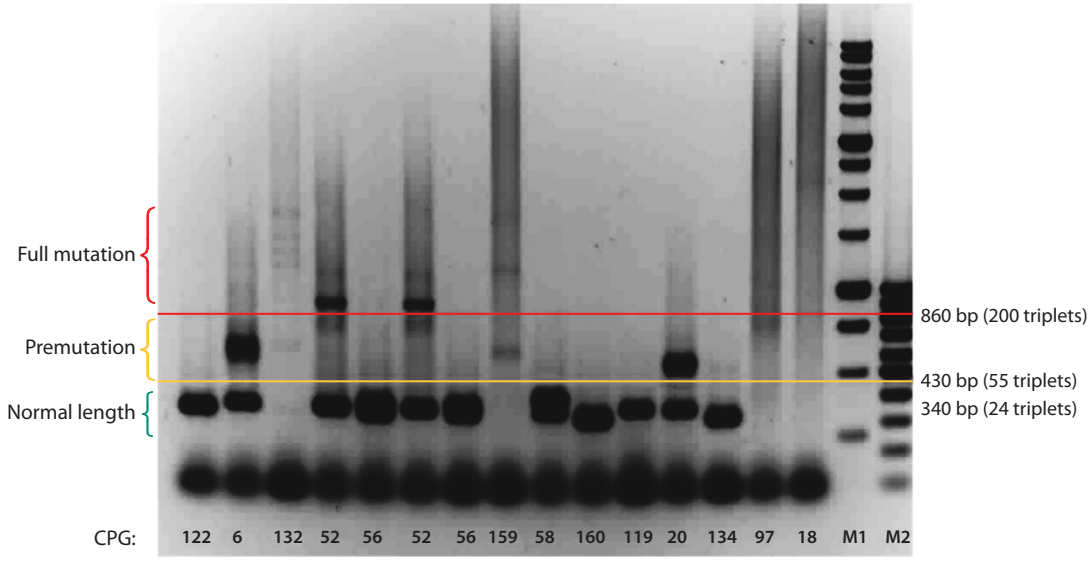


Fig. 2. Sample of CGG-repeat amplification.

CPG – samples of DNA from patients; M1 –1 kb DNA ladder; M2 – 100 bp DNA ladder.

cultures and whole blood samples of patients with Fragile X syndrome (Fig. 2).

To create a construct with exogenous CGG-repeat it was decided to use repeats of normal and premutant lengths. It is expected that the instability of these types of repeats will be significantly different, since the premutant allele is the most unstable, and the normal allele, on the contrary, is prone to only insignificant polymorphism (Lokanga et al., 2013).

As a result, five types of plasmids were obtained. These plasmids carry 5 (pCDH-5), 25 (pSBi-25), 59 (pCDH59), 85 (pCDH85), and 160 repeats (pSBi-160), respectively. The structures of all plasmids were confirmed by Sanger sequencing (Fig. 3).

Study of the experimental plasmid functionality

The eukaryotic cells transfection efficiency by the assembled constructs was evaluated to confirm the correct expression of reporter proteins in the presence of an extended repeat (CGG)n. It has been shown that transfection and reporter protein synthesis after transfections by plasmids carrying CGG-repeat of normal or premutant length occurs with the same efficiency as transfection with control plasmids (without the repeated sequence). When cells were transfected with pCDH plasmids, the expression of DsRedExpress was observed. It was also possible to carry out selection on a medium with the puromycin of cells transfected with pSBi vectors. The ability of the tetracycline/doxycycline-inducible promoters regulating

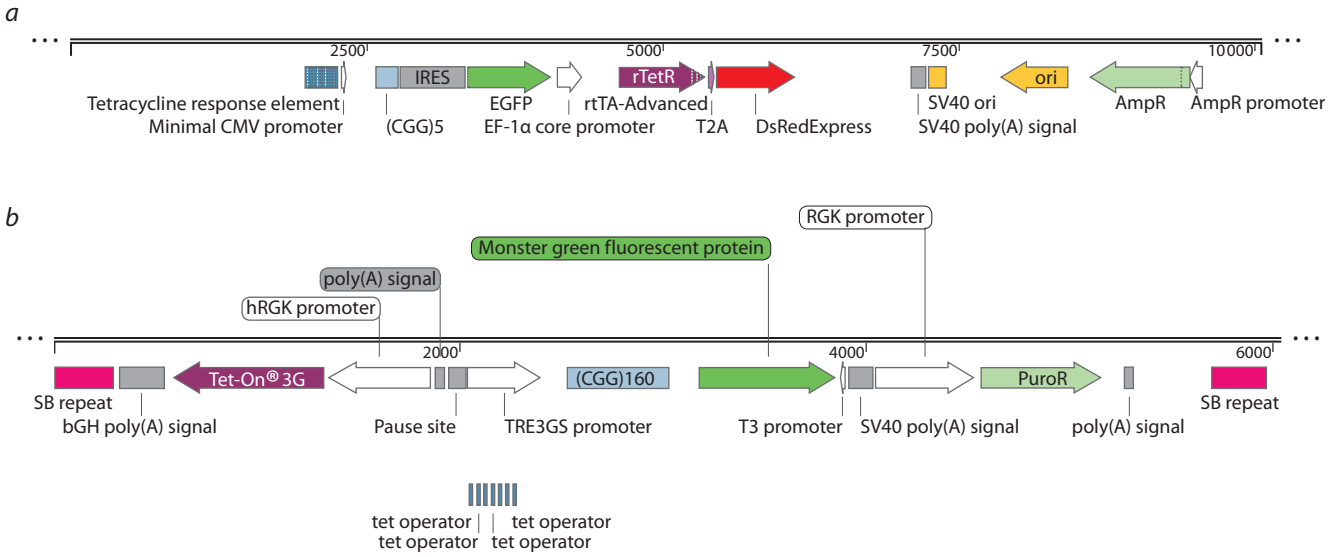


Fig. 3. Maps of plasmids with cloned CGG repeat.

a – pCDH plasmid with CGG-repeat and *b* – pSBi plasmid with CGG-repeat.

green fluorescent protein expression to spontaneous activation was investigated. It is important that the transgenic cells do not have a background EGFP expression because spontaneous promoter activation can interfere with the accurate assessment of the CGG-repeat instability level during transcription and transcription-coupled repair.

Cells transfected with pCDH the showed active expression of the red protein (driven by constitutive promoter EF1) and the absence of the green protein expression (regulated by inducible promoter) without the promoter induction. For induction, doxycycline was added daily to the cells, resulting in a high level of green protein fluorescence (Fig. 4, *a*). When plasmid pSBi containing the TRE3GS promoter was used, no background induction was observed. It allows for performing a selection using puromycin to obtain stable transformants and to avoid the background transcription level influence on the inserted CGG-repeat (see Fig. 4, *b*).

Development of the method for analyzing repeat instability in model cell lines

By using the obtained transgenic cell lines, we expect that the expansion in different cells of the culture will occur at different rates, and, as a result, we will receive a mosaic culture. In this regard, it is necessary to use a value allowing the

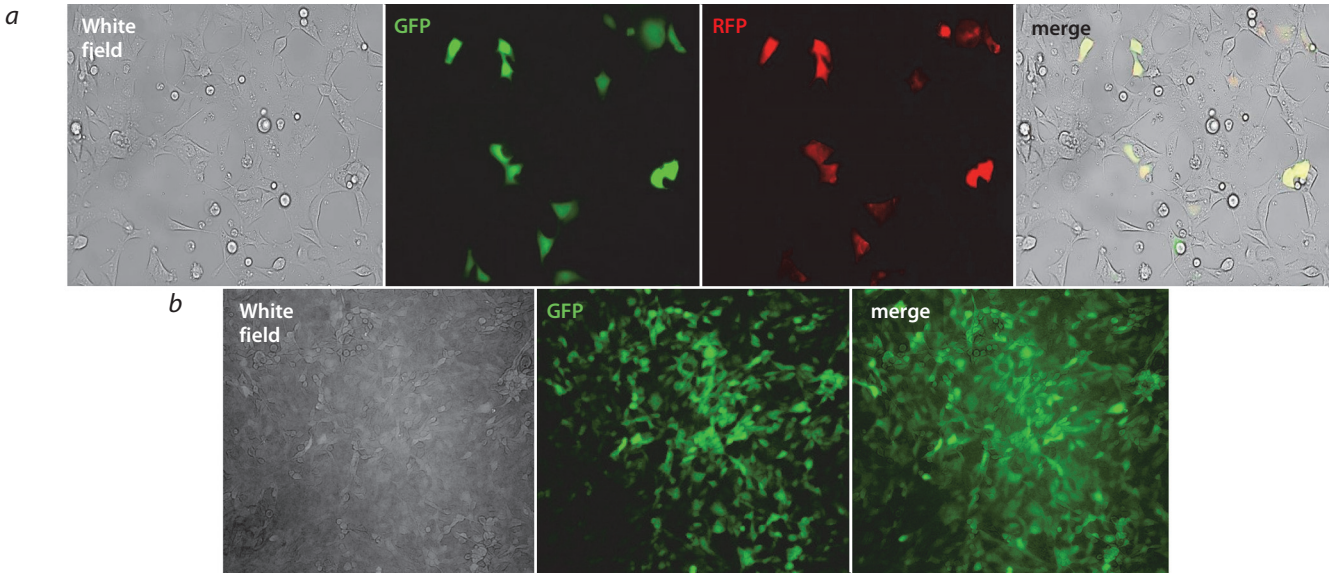


Fig. 4. Induction of tetracycline-dependent promoters in the designed plasmids.

a – induction of Tet-O-promoter in HEK293T cells transfected by Pcdh; *b* – induction of TRE3GS in HEK293A cells transfected by pSBi and selected on puromycin.

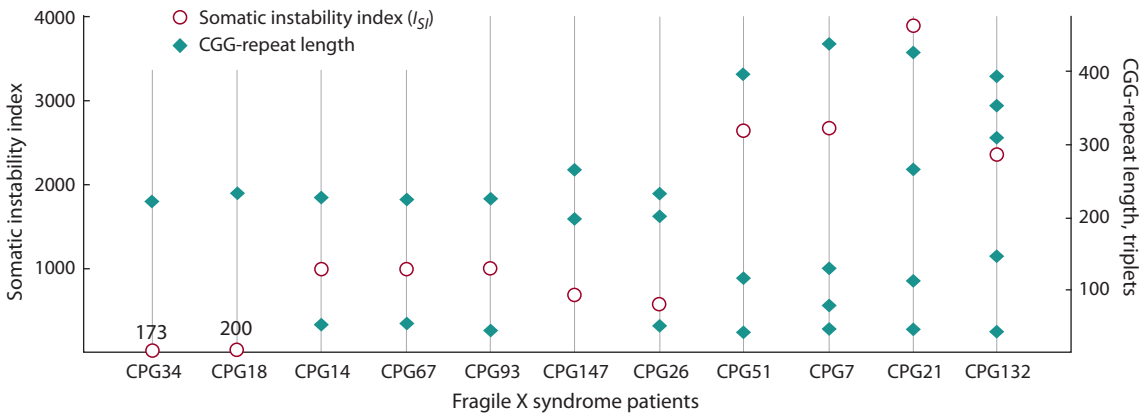


Fig. 5. Repeat length and indexes of somatic instability in FXS patients.
The values on the graph above the markers are the somatic instability indexes.

grading of somatic instability and allowing the comparison of cell lines carrying different alleles of exogenous CGG-repeat. Previously, different approaches were proposed to assess the trinucleotide repeats somatic instability in patients with repeats expansion disorders. For example, the method for the assessment of CAG-repeat instability in Huntington's disease is based on the main allele determination by the maximum peak as a result of fragment analysis and additional peaks, followed by normalization to the summed values of the heights of all peaks (Lee et al., 2010). Another method for the repeat instability level assessment is based on the serial dilution of the template followed by PCR – small-pool PCR (Monckton et al., 1995; Morales et al., 2012).

Amplification by dilution enables the detection of mosaicism, which cannot be detected by conventional PCR due to the low synthesis efficiency of less represented or very large alleles. However, these methods are insufficiently applicable to assess the instability (CGG)n, since the amplification of the larger allele occurs with much less efficiency than the shorter allele synthesis (Usdin, Woodford, 1995; Woodford et al., 1995; Jensen et al., 2010). To quantify the CGG-repeat instability in developed cell lines, we proposed an analysis method based on the calculation of the somatic instability index (I_{SI}) after the amplification of GC-rich templates according to the method of B.E. Hayward et al. (2016). This value enables one to take into account not only the repeat size but also the spread of values between alleles, regardless of the efficiency of their synthesis. For (CGG)n repeats located in the FMR1 gene, we propose the calculation of I_{SI} using the following equation

$$I_{SI} = Me \cdot (N_{max} - Me),$$

where Me – median and N_{max} – maximal length of CGG repeats in any sample.

The median is a value separating the raw data into two halves, and it considers the number of alleles. This value takes into account sample heterogeneity, and not sensitive to the detection of repeat lengths that are too long or too short, unlike using an arithmetic mean. When using the arithmetic mean in the index calculating, the contribution of larger alleles will be taken into account more than the contribution of shorter ones. As a result, cell lines with different degrees of the exogenous repeat instability can have similar values of the magnitude of somatic instability, which will lead to the false results interpretation. The value ($N_{max} - Me$) takes into account the diversity and scatter of values in samples, where large Me values indicate a large median repeat length. The I_{SI} calculation does not take into account the amount of PCR product (peak height) for each allele, i. e. PCR efficiency does not affect the final value. To determine the somatic instability index, the DNA of eleven patients with Fragile X syndrome was isolated from whole blood, which served as a starting material for the synthesis of extended repeats (CGG)n (Fig. 5).

As can be seen from the calculation of the I_{SI}, index increases with an increase in the number and spread of repeat values. It should be noted that the method of analysis of instability works for two or more alleles in patients with mosaicism. In the case of one allele, we take the index of somatic instability equal to the size of the CGG-repeat, since the patient with one allele has $N_{max} - Me = 0$. We cannot accept $I_{SI} = 0$ because the CGG-repeat is unstable by nature.

Discussion

Fragile X syndrome is one of the most common causes of hereditary intellectual disability (Yudkin et al., 2015). The frequency of full mutation in the human population varies from 1:6,000 in women to 1:4,000 in men, while the premutant allele, as the most unstable allele of the FMR1 gene promoter region, occurs in 1:100 cases. The instability of the CGG repeat is expressed in its tendency toward expansion – a multiple and rapid increase of the tract repeated sequence length. In addition, in the cells and tissues of patients as well as the tissues of model animals, repeat contractions are observed, that lead to somatic mosaicism and its degree correlates with the severity of symptoms (Mailick et al., 2018). However, the probability of expansion is in tenfold higher than contraction (Bontekoe, 2001; DeJesus-Hernandez et al., 2011), which may be the reason for the increased severity of the diseases manifestations during transmission in a number of generations.

There are a number of hypotheses explaining the expansion mechanism, but none of them have been sufficiently supported by experimental data. All of the hypotheses – assume as a main reason for repeat instability – the formation of alternative DNA secondary structures at a certain site in DNA during the different processes of DNA metabolism, which can disrupt

these processes. *In vitro* and *in vivo* experiments have shown the formation of alternative DNA secondary structures, such as hairpins, R-loops, and G-quadruplexes (Usdin, Woodford, 1995; Groh et al., 2014; Lam et al., 2014). Such structures can significantly violate these processes of DNA metabolism, which in turn affects the instability of repeats. One of the possible reasons is associated with the slippage of a DNA strand during replication (Pearson, Sinden, 1996; Fouche et al., 2006). Today, it is absolutely clear that the slippage of DNA strands can occur in various cases: during DNA replication in dividing cells as well as during repair processes. However, this model cannot reliably explain why not all of the repeats expand or and why the threshold value for the length of the repeat sequence is similar for different diseases. There is evidence for the contribution of some repair cascade proteins, which are proteins that are necessary for recombination and transcription to repeat instability. However, all of the hypotheses have certain drawbacks and contradictions; therefore, it is necessary to continue the search for the molecular mechanism of repeat expansion.

An expansion model based on a transgenic cell line containing exogenous (CGG)n repeat can serve as a convenient experimental system. In such systems it is possible to track changes in repeat length in response to the induction of different cascades of DNA metabolism. Using the transgenic cell lines obtained in this study make it possible to assess the contribution of replication, transcription, and repair in the cell to CGG-repeat instability. We have assembled two types of plasmids: based on the SV40 origin plasmid, capable of replicating in cell cultures expressing the SV40 T antigen, and based on the Sleeping Beauty transposon-based vector system for integrating the cassette with CGG-repeat and reporter proteins into various genomic loci. The transfection efficiency and the initial expression level of reporter proteins were comparable to those of the control plasmid without the (CGG)n repeat. It is also possible to obtain a transgenic cell culture with single genotype using different approaches such as sorting or limiting dilutions with antibiotic selection. Changes in the length of an exogenous repeat and, therefore, mosaicism that will take place in transgenic culture over time can be detected and estimated using the developed I_{SI} index. This method is useable and reflects the correlation between repeat instability and phenotypic manifestations of the diseases observed in the different brain lobes of patients with Fragile X syndrome and associated disorders.

In the created experimental cell lines it is possible to directly assess the level of repeat expansion or contraction as well as the changes caused by repeat instability. The design of vector systems makes it possible to detect changes in the length of the exogenous CGG-repeat at different genome loci, during cultivation for a long or short time, with or without promoter induction. Measuring the fluorescent proteins expression levels can serve as basis for tracking the possible increase of instability and mutations accumulation mediated by repeat-induced mutagenesis. To determine the contribution of specific proteins from various cascades to the development of instability, it is possible to carry out chromatin immunoprecipitation using transformed cells. In addition, the level of instability in the created cellular models of CGG-repeat expansion can be assessed by the proposed index of somatic instability. Index I_{SI} should also have a biological meaning, i. e. reflect the degree of phenotypic changes in patients with Fragile X associated disorders. To test this hypothesis, a study of the dependences of I_{SI} values in patients with changes in the brain according to FMRI data was started. The preliminary data indicate some correlations, but more research is needed.

Conclusion

To date, the mechanism of the instability of trinucleotide repeats remains not fully understood. At the same time, this research area remains extremely urgent due to the fact that the diseases caused by this mutation are socially significant. To search for the repeat instability reasons, it is necessary to develop cellular models for tracking all of the changes caused by expansion, as well as to evaluate the contribution of various proteins and DNA metabolism pathways to this process. The constructs developed in this work for instability assessing can be used in such studies.

Various cell lines can be transfected with the assembled plasmids. We tested the efficiency of the constructs in two cell lines: HEK293A and HEK293T. After cell transfection and the induction of reporter protein expression, at various passages, it is possible to accurately determine the repeat size (CGG)n, as well as other parameters and show the presence or absence of CGG-repeat expansion, depending on its initial length and the number of passages. In the future, our model can be used in studies for the determination of all the aspects of repeat instability in the human genome and it will help form a more complete understanding of the mechanisms of this mutation.

References

Bontekoe C.J.M. Instability of a (CGG)98 repeat in the Fmr1 promoter. *Hum. Mol. Genet.* 2001;10(16):1693-1699. DOI 10.1093/hmg/10.16.1693.

DeJesus-Hernandez M., Mackenzie I.R., Boeve B.F., Boxer A.L., Baker M., Rutherford N.J., Nicholson A.M., Finch N.A., Flynn H., Adamson J., Kouri N., Wojtas A., Sengdy P., Hsiung G.Y.R., Karydas A., Seeley W.W., Josephs K.A., Coppola G., Geschwind D.H., Wszolek Z.K., Feldman H., Knopman D.S., Petersen R.C., Miller B.L., Dickson D.W., Boylan K.B., Graff-Radford N.R., Rademakers R. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*. 2011;72(2):245-256. DOI 10.1016/j.neuron.2011.09.011.

Fouche N., Ozgur S., Roy D., Griffith J.D. Replication fork regression in repetitive DNAs. *Nucleic Acids Res*. 2006;34(20):6044-6050. DOI 10.1093/nar/gkl757.

Gorbunova V., Seluanov A., Dion V., Sandor Z., Meservy J.L., Wilson J.H. Selectable system for monitoring the instability of CTG/CAG triplet repeats in mammalian cells. *Mol. Cell. Biol*. 2003;23(13):4485-4493. DOI 10.1128/mcb.23.13.4485-4493.2003.

Grishchenko I.V., Purvinsh Y.V., Yudkin D.V. Mystery of expansion: DNA metabolism and unstable repeats. In: Zharkov D.O. (Ed.). Mechanisms of Genome Protection and Repair. Cham: Springer International Publishing, 2020;101-124. DOI 10.1007/978-3-030-41283-8_7.

Groh M., Lufino M.M.P., Wade-Martins R., Gromak N. R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. *PLoS Genet*. 2014;10(5):e1004318. DOI 10.1371/journal.pgen.1004318.

Hayward B.E., Zhou Y., Kumari D., Usdin K. A Set of assays for the comprehensive analysis of FMR1 alleles in the Fragile X-related disorders. *J. Mol. Diagn.* 2016;18(5):762-774. DOI 10.1016/j.jmoldx.2016.06.001.

Heulens I., Suttie M., Postnov A., De Clerck N., Perrotta C.S., Mattina T., Faravelli F., Forzano F., Kooy R.F., Hammond P. Craniofacial characteristics of fragile X syndrome in mouse and man. *Eur. J. Hum. Genet*. 2013;21(8):816-823. DOI 10.1038/ejhg.2012.265.

Jensen M.A., Fukushima M., Davis R.W. DMSO and betaine greatly improve amplification of GC-rich constructs in *de novo* synthesis. *PLoS One*. 2010;5:e11024. DOI 10.1371/journal.pone.0011024.

Kononenko A.V., Ebersole T., Mirkin S.M. Experimental system to study instability of (CGG)n repeats in cultured mammalian cells. In: Richard G.-F. (Ed.). Trinucleotide Repeats: Methods and Protocols. New York: Springer, 2020;137-150. DOI 10.1007/978-1-4939-9784-8_9.

Kovalenko M., Dragileva E., St Claire J., Gillis T., Guide J.R., New J., Dong H., Kucherlapati R., Kucherlapati M.H., Ehrlich M.E., Lee J.M., Wheeler V.C. *Msh2* acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. *PloS One*. 2012;7(9): e44273. DOI 10.1371/journal.pone.0044273.

Krasilnikova M.M., Kireeva M.L., Petrovic V., Knijnikova N., Kashlev M., Mirkin S.M. Effects of Friedreich's ataxia (GAA)n*(TTC)n repeats on RNA synthesis and stability. *Nucleic Acids Res*. 2007; 35(4):1075-1084. DOI 10.1093/nar/gkl1140.

Lam E.Y.N., Beraldi D., Tannahill D., Balasubramanian S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun*. 2014;4(1)1-8. DOI 10.1038/ncomms2792.

Lee J.M., Zhang J., Su A.I., Walker J.R., Wiltshire T., Kang K., Dragileva E., Gillis T., Lopez E.T., Boily M.J., Cyr M., Kohane I., Gusella J.F., MacDonald M.E., Wheeler V.C. A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst. Biol*. 2010;4(1):29. DOI 10.1186/1752-0509-4-29.

Lokanga R.A., Entezam A., Kumari D., Yudkin D., Qin M., Smith C.B., Usdin K. Somatic expansion in mouse and human carriers of fragile X premutation alleles. *Hum. Mutat*. 2013;34(1):157-166. DOI 10.1002/humu.22177.

Mailick M.R., Movaghar A., Hong J., Greenberg J.S., DaWalt L.S., Zhou L., Jackson J., Rathouz P.J., Baker M.W., Brilliant M., Page D., Berry-Kravis E. Health profiles of mosaic versus non-mosaic FMR1 premutation carrier mothers of children with fragile X syndrome. *Front. Genet*. 2018;9:173. DOI 10.3389/fgene.2018.00173.

Martin G.E., Roberts J.E., Helm-Estabrooks N., Sideris J., Vanderbilt J., Moskowitz L. Perseveration in the connected speech of boys with fragile X syndrome with and without autism spectrum disorder. *Am. J. Intellect. Dev. Disab*. 2012;117(5):384-399. DOI 10.1352/1944-7558-117.5.384.

Monckton D.G., Wong L.J.C., Ashizawa T., Caskey C.T. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet*. 1995;4(1):1-8. DOI 10.1093/hmg/4.1.1.

Morales F., Couto J.M., Higham C.F., Hogg G., Cuenca P., Braida C., Wilson R.H., Adam B., Del Valle G., Brian R., Sittenfeld M., Ashizawa T., Wilcox A., Wilcox D.E., Monckton D.G. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet*. 2012;21(16):3558-3567. DOI 10.1093/hmg/dds185.

Pearson C.E., Sinden R.R. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*. 1996;35(15):5041-5053. DOI 10.1021/bi9601013.

Roberts J., Hennon E.A., Anderson K. Fragile X syndrome and speech and language. *ASHA Leader*. 2003;8(19):6-27. DOI 10.1044/leader.FTR2.08192003.6.

Shah K.A., Mirkin S.M. The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair*. 2015;32:106-112. DOI 10.1016/j.dnarep.2015.04.020.

Usdin K., Woodford K.J. CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucleic Acids Res*. 1995;23(20):4202-4209.

Woodford K., Weitzmann M.N., Usdin K. The use of K(+)-free buffers eliminates a common cause of premature chain termination in PCR and PCR sequencing. *Nucleic Acids Res*. 1995;23(3):539. DOI 10.1093/nar/23.3.539.

Yudkin D.V., Lemskaya N.A., Grischenko I.V., Dolskiy A.A. Chromatin changes caused by expansion of CGG repeats in *fmr1* gene. *Mol. Biol*. 2015;49(2):179-184.

Zhao X.-N., Lokanga R., Allette K., Gazy I., Wu D., Usdin K. A MutSbeta-dependent contribution of MutSalpha to repeat expansions in fragile X premutation mice? *PLoS Genet*. 2016;12(7): e1006190. DOI 10.1371/journal.pgen.1006190.

ORCID ID

I.V. Grishchenko orcid.org/0000-0002-2227-8500
A.A. Tulupov orcid.org/0000-0002-1277-4113
E.D. Petrovskiy orcid.org/0000-0003-4325-4062
A.A. Savelov orcid.org/0000-0002-5332-2607
A.M. Korostyshevskaya orcid.org/0000-0002-0095-8994
E.M. Shitik orcid.org/0000-0001-8529-9176
D.V. Yudkin orcid.org/0000-0002-8940-9173

Acknowledgements. This study was funded by Russian Science Foundation through research project No. 18-15-00099 for the molecular biological research part and through research project No. 19-75-20093 for the theoretical part.
The authors are grateful to Ph.D. V.S. Fishman (Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Sector of Genomic Mechanisms of Ontogenesis) for the plasmids that were provided.

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

Received October 23, 2020. Revised December 16, 2020. Accepted December 17, 2020.

Вавиловский журнал генетики и селекции. 2021;25(1):125-134
DOI 10.18699/VJ21.015

Original Russian text www.bionet.nsc.ru/vogis/

Production of subtilisin proteases in bacteria and yeast

A.S. Rozanov[1, 2] ✉, S.V. Shekhovtsov[1, 2], N.V. Bogacheva[1, 2], E.G. Pershina[1, 2], A.V. Ryapolova[3], D.S. Bytyak[3], S.E. Peltek[1, 2]

[1] Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Laboratory of Molecular Biotechnologies, Novosibirsk, Russia
[3] Innovation Centre "Biruch-NT", Malobykovo village, Belgorod region, Russia
✉ rozanov@bionet.nsc.ru

Abstract. In this review, we discuss the progress in the study and modification of subtilisin proteases. Despite long-standing applications of microbial proteases and a large number of research papers, the search for new protease genes, the construction of producer strains, and the development of methods for their practical application are still relevant and important, judging by the number of citations of the research articles on proteases and their microbial producers. This enzyme class represents the largest share of the industrial production of proteins worldwide. This situation can explain the high level of interest in these enzymes and points to the high importance of designing domestic technologies for their manufacture. The review covers subtilisin classification, the history of their discovery, and subsequent research on the optimization of their properties. An overview of the classes of subtilisin proteases and related enzymes is provided too. There is a discussion about the problems with the search for (and selection of) subtilases from natural strains of various microorganisms, approaches to (and specifics of) their modification, as well as the relevant genetic engineering techniques. Details are provided on the methods for expression optimization of industrial subtilases of various strains: the details of the most important parameters of cultivation, i. e., composition of the media, culture duration, and the influence of temperature and pH. Also presented are the results of the latest studies on cultivation techniques: submerged and solid-state fermentation. From the literature data reviewed, we can conclude that native enzymes (i. e., those obtained from natural sources) currently hardly have any practical applications because of the decisive advantages of the enzymes modified by genetic engineering and having better properties: e. g., thermal stability, general resistance to detergents and specific resistance to various oxidants, high activity in various temperature ranges, independence from metal ions, and stability in the absence of calcium. The vast majority of subtilisin proteases are expressed in producer strains belonging to different species of the genus *Bacillus*. Meanwhile, there is an effort to adapt the expression of these enzymes to other microbes, in particular species of the yeast *Pichia pastoris*.

Key words: subtilisin; subtilase; protease; alkaline serine protease; *Pichia pastoris*; *Bacillus subtilis*; biotechnology; genetic engineering; cultivation.

For citation: Rozanov A.S., Shekhovtsov S.V., Bogacheva N.V., Pershina E.G., Ryapolova A.V., Bytyak D.S., Peltek S.E. Production of subtilisin proteases in bacteria and yeast. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):125-134. DOI 10.18699/VJ21.015

Продукция субтилизиновых протеаз в бактериях и дрожжах

А.С. Розанов[1, 2] ✉, С.В. Шеховцов[1, 2], Н.В. Богачева[1, 2], Е.Г. Першина[1, 2], А.В. Ряполова[3], Д.С. Бытяк[3], С.Е. Пельтек[1, 2]

[1] Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, лаборатория молекулярных биотехнологий, Новосибирск, Россия
[3] Инновационный центр «Бирюч-НТ», с. Малобыково, Белгородская область, Россия
✉ rozanov@bionet.nsc.ru

Аннотация. В настоящей работе мы рассматриваем прогресс в изучении и модификации субтилизиновых протеаз. Несмотря на длительное время применения микробиальных протеаз и значительное число работ, посвященных их исследованию, поиск новых генов протеаз, создание продуцентов и развитие методов их применения остаются актуальными, о чем говорит высокий уровень цитирования публикаций, описывающих протеазы и их продуценты. На данный класс ферментов приходится максимальный объем производства промышленных белков в мире, что объясняет большой интерес к нему. Это говорит о чрезвычайно высокой важности получения собственных технологий их производства. В статье представлены сведения о классификации субтилизинов, истории их открытия и дальнейших работ по оптимизации их свойств. Дан обзор классов субтилизиновых протеаз и родственных им ферментов. Проанализированы проблемы поиска и отбора субтилаз из природных штаммов различных микроорганизмов, пути и особенности их модификации и используемые при этом методы генетической инженерии. Детально изучены методы оптимизации продукции промышленных субтилаз у различных штаммов, касающихся важнейших аспектов культивирования: состава

среды, времени культивирования, влияния температуры и pH. Приводятся результаты последних исследований по техникам культивирования – глубинному и твердофазному культивированию. На основании рассмотренных литературных данных можно заключить, что в настоящее время практически не применяются нативные, т. е. обнаруженные в природе ферменты, в связи с решающими преимуществами, предоставляемыми белками, модифицированными при помощи генной инженерии и обладающими улучшенными свойствами: термостабильностью, общей устойчивостью к детергентам и специфической – к различным окислителям, высокой активностью в разных диапазонах температур, независимостью от ионов, стабильностью в отсутствие кальция и т. д. Большинство субтилизиновых протеаз синтезируется в штаммах-продуцентах, относящихся к разным видам рода *Bacillus*. В то же время ведутся работы по адаптации синтеза этих ферментов в других микроорганизмах, в частности дрожжей *Pichia pastoris*.
Ключевые слова: субтилизин; субтилаза; протеаза; щелочная сериновая протеаза; *Pichia pastoris*; *Bacillus subtilis*; биотехнология; генетическая инженерия; культивирование.

Introduction

Proteases are enzymes that degrade proteins via the hydrolysis of peptide bonds. Proteases correspond to the general enzyme class designated as EC 3.4.X.X (Garcia-Carreno, Del Toro, 1997). Endopeptidases act most strongly on intact proteins; they cleave peptide bonds of nonterminal amino acid residues. Exopeptidases sever peptide bonds between amino acid residues at the end of a polypeptide chain. They are categorized into amino- and carboxy-peptidases depending on which end (N- or C-terminus) they remove amino acids from (Barrett, McDonald, 1986). Proteases are subdivided into families in accordance with their mechanism of action. According to database MEROPS (http://merops.sanger.ac.uk) (Rawlings et al., 2014), the following protease families are known: asparagine, cysteine, glutamine, serine and threonine peptidases, metalloproteinases, mixed peptidases, and peptidases with an unknown mechanism of action.

Peptidases are present in all life forms. Today, the most popular proteases are those from prokaryotes, mainly bacteria, because of their excellent potential for various technological applications. Given that proteases are needed in large amounts, the cost of production is as important as protease characteristics; as a consequence, in most cases, proteases are manufactured by means of bacteria. Microorganisms can produce proteases faster and more cheaply than mammalian and plant cells can; the enzyme manufacture is not affected by the climate or changes of seasons or by regulatory or ethical problems. Besides, extracellular enzymes expressed by microorganisms are usually preferred because subsequent processing is simpler, meaning even lower costs (Tufvesson et al., 2010). In terms of a combination of characteristics (activity, pH and temperature ranges, and production costs), subtilisins or subtilases have turned out to be the most popular class of proteases.

Subtilases are one of the largest classes of serine proteases that are encoded in the genomes of all life forms including viruses. By amino acid sequence, subtilases are subdivided into six families: subtilisins, thermitases, proteinases K, lantibiotic peptidases, kexins, and pyrolisins. Subtilisins in turn are categorized into several subfamilies: true subtilisins, highly alkaline proteases, intracellular proteases, intermediate subtilisins, and high-molecular-weight subtilisins.

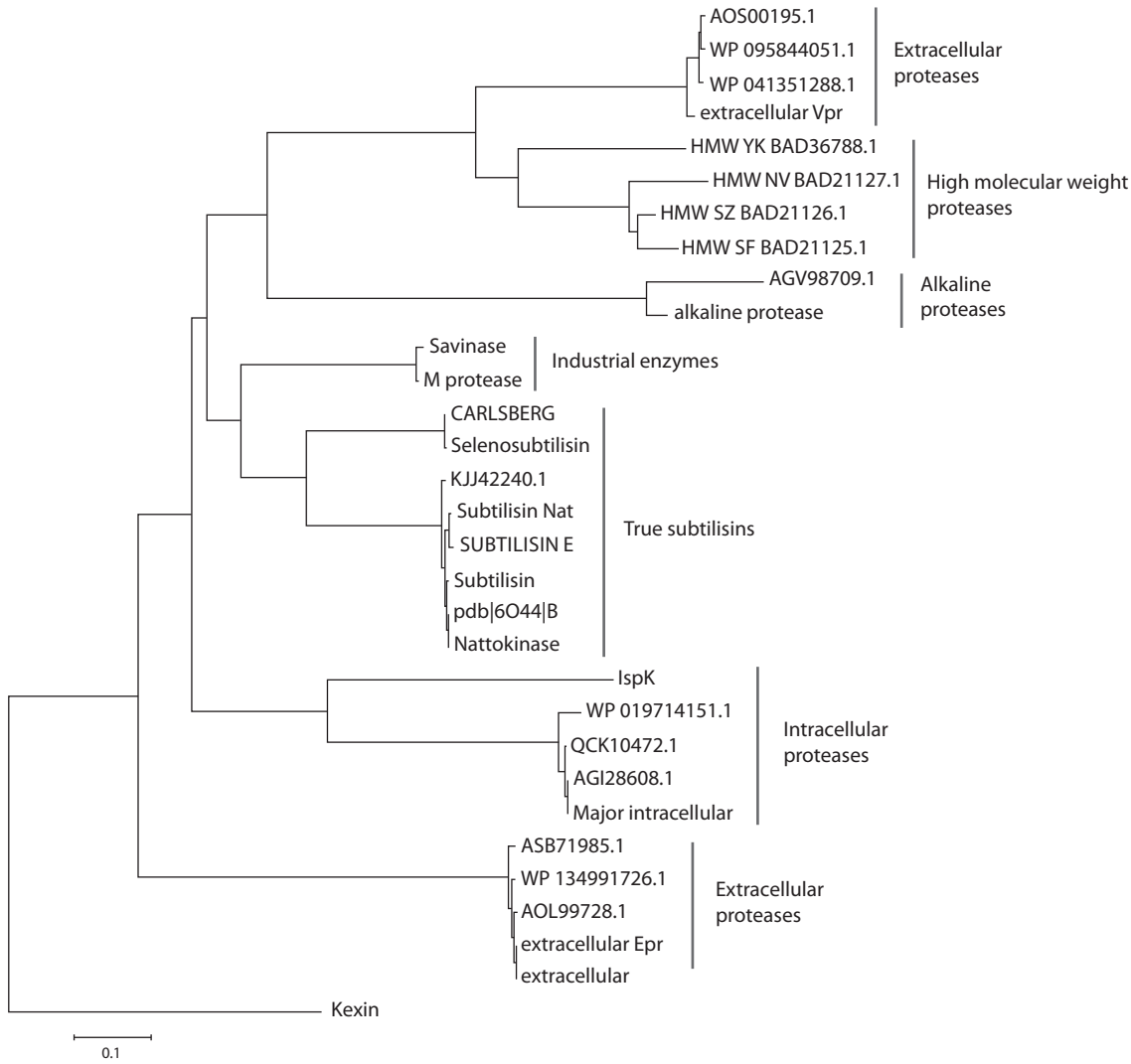
All the subfamilies of subtilisins hold promise for biotechnology. The first alkaline serine protease that gained widespread use was subtilisin A (EC 3.4.21.62), which is an alkaline serine protease from *Bacillus subtilis*. The enzyme owes its name to the species of its bacterial producer (Ottesen, Svendsen, 1970; Ikemura et al., 1987). The history of discovery and study of subtilisins started at a research center of a beer-brewing company called Carlsberg, and the first enzyme to be described is named "subtilisin Carlsberg" (Smith et al., 1966).

The catalytic center of serine proteases is formed by three amino acid residues: Asp-32, His-64, and Ser-221. Because the amino acid residue carrying out the nucleophilic attack is Ser-221, subtilisins and the related proteolytic enzymes are called serine proteinases. Among the highly alkaline proteases, there is an enzyme isolated from strain *Bacillus* sp. KSM-K16 (Kobayashi et al., 1995). Its optimum of activity is at 55 °C and pH 12.3. This enzyme is employed in the industry in complex with a detergent, as is the case for related highly alkaline proteases, Savinase and Maxacal. Intermediate subtilisins are somewhere between true subtilisins and highly alkaline proteases and include some promising enzymes. For instance, the ALTP enzyme isolated from *Alkaliphilus transvaalensis* (Kobayashi et al., 2007) shows maximal activity at very high temperatures and pH, namely, at 70 °C and pH > 12.6. Nonetheless, ALTP can also perform a catalytic function at lower temperatures and pH. The phylogenetic tree based on the amino acid sequences of subtilisin proteases is presented in the Figure.

Intracellular proteases are rather poorly studied in comparison with the above subfamilies. The reason is that they are active at lower pH, which is characteristic of the cytoplasm. For example, the intracellular protease from *B. megaterium* (Jeong et al., 2018) at 50 °C shows an optimum of activity at pH 6.0–7.0.

From alkalophilic *Bacillus* spp., researchers isolated a set of high-molecular-weight subtilisins (Okuda et al., 2004) ~650 amino acid residues long (size of the precursor: 800 amino acid residues). Their optimal pH is 10.5–11.0, and optimal temperatures for activity are 40–45 °C.

Bacteria are most widely used as a microbial producer of proteases, and the genus *Bacillus* is the most famous source among them. Primarily, the reason is the strong ability to secrete proteins, which allows to obtain > 20 g of protein per liter of a medium (Harwood, Cranenburgh, 2008). Furthermore, various *Bacillus* species produce neutral and alkaline proteases (Anandharaj et al., 2016; Rehman et al., 2017), and this property is important for the industry. Proteases of *Bacil-*



Subtilisin proteases and the related enzymes of *B. subtilis* as well as some industrial enzymes.

lus members have unique characteristics enabling their use in many industrial sectors. Consequently, proteases from various *Bacillus* species are responsible for ~60 % of all the sales of enzymes worldwide. Because of the wide ranges of pH and temperature corresponding to good activity and stability, these enzymes are used in the detergent industry (Porres et al., 2002). For this purpose, enzymes should be resistant to an alkaline medium and retain their activity in the presence of inhibitors, including oxidants and surfactants. In addition, the proteases isolated from the strains of *Bacillus* are suitable for the food industry for preparation of biologically active peptides and processing of various food products (Latiffi et al., 2013; Ke et al., 2018). Another feature of these proteases is stability in organic solvents and the consequent suitability for organic synthesis (Hu et al., 2013). Owing to the high commercial significance, a large number of patents deal with the strains of *Bacillus* (see the Table).

The widespread manufacture of proteases by means of *Bacillus* strains is due to the economic efficiency of these strains. Additionally, it is possible to utilize the byproducts of agricultural production as a substrate for these strains, e. g. molasses of sugarcane and corn starch for submerged fermentation (Shikha et al., 2007) or various types of bran and solid residues for solid-state fermentation (Shivasharana, Naik, 2012).

The search for alkaline serineproteases in nature

Proteases are commercially important proteins responsible for the lion's share of protein manufacture. They have many applications, and each technological process has its specific features and requirements for the enzymes used. Besides, the unrelenting interest in these proteins is due to the search for enzymes that are not covered by patents, albeit with properties that are not necessarily better than those of the existing enzymes. Accordingly, a huge number of research papers on this topic is published every year. The largest number of genes of alkaline serine proteases has been discovered in the genomes of bacteria from the genus *Bacillus*. The second place in terms of isolation of proteases belongs to *Actinomycetes*. A substantial number of researchers also seek alkaline proteases of fungal origin (Sharma et al., 2017). In the latest articles, the emphasis is on the search for enzymes having a

Industrial subtilases obtained from *Bacillus* species

Enzyme	Class	Species
Dispase I VR	Proteases	*B. polymyxa*
Dispase II VR		
Proteinase	Subtilisin A	*B. licheniformis*
Neutrase	Metalloproteinases	*B. amyloliquefaciens*
Esperase	Serine endopeptidases (mainly, subtilisin A)	*Bacillus* sp.
Everlase	Subtilisin A	
Protamex	Proteases	
Savinase	Serine endopeptidases (mainly, subtilisin A)	
Alcalase		*B. licheniformis*
Optimase PR	Serine endopeptidases	*B. subtilis*
GenencorVR Protease 899	Neutral metallopeptidase	
ProtexTM 6L	Subtilisins, serine endopeptidases	*B. licheniformis*
Multifect	Neutral serine endopeptidases	*B. amyloliquefaciens*

keratinase activity because of increased interest in the processing of keratin-containing residues, e. g., feathers.

The source of one of the promising genes encoding a serine protease is strain *Bacillus licheniformis* NMS-1 extracted from soil near a natural hot spring in Sri Lanka (Mathew, Gunathilaka, 2015). This protein serves for the creation of detergents. Closely related strain *B. licheniformis* K7A expressing an alkaline protease was obtained in another study (Hadjidj et al., 2018). Analysis of the expressed protein revealed that it has the highest activity at pH 10 and 70 °C. The enzymatic activity is higher than that of commercial preparations of Alcalase and Thermolysin. Another serine protease was found in the genome of the bacterium *Bacillus amyloliquefaciens* FSE-68 isolated from a starter culture for soy fermentation in South Korea. Its amino acid sequence was determined by liquid chromatography with electrospray ionization tandem mass spectrometry (LC/ESI-MS/MS) and by whole-genome sequencing. In comparison with a homolog, i. e., well-studied subtilisin BPN from *B. amyloliquefaciens*, that enzyme showed slightly higher stability in the absence of calcium ions (Cho, 2019). The protein isolated from alkaliphilic strain *Bacillus luteus* H11 manifested proteolytic activity at NaCl concentration up to 5 M, temperature 45 °C, and pH 10.5 (Kalwasińska et al., 2018). In China, during a screening of the bacteria obtained from industrial fermentation of soy, investigators isolated strain *B. subtilis* MX-6, which overexpresses a nattokinase-like protein (Gulmez et al., 2018).

Numerous recent studies on the search for new versions of proteases of bacterial or fungal origin can be discussed *ad infinitum*. Therefore, for a variety of reasons, advances in the manufacture of proteolytic enzymes are still relevant today. This is especially true for developing countries, which strongly wish to increase the proportion of industrial products, including biotechnological ones, on their domestic markets. An especially large number of studies in this field has been published by research groups from India. At present, such studies are virtually absent in Russia.

Genetic engineering of subtilisin

Subtilisin is the industrial enzyme that has probably been studied the most by both statistical and directed mutagenesis. The applications of subtilisin have expanded constantly since the start of its manufacture. To meet the needs of the industry, subtilisin characteristics had to be improved. In the early 1980s, the methods for directed engineering of proteins started to develop rapidly. As a result of application of these methods to subtilisin, mutations of more than a half of its 275 amino acid residues had been described before the year 2000 in scientific literature. Patents contain even more such accomplishments, and, undoubtedly, an even greater number of findings is buried in the freezers of biotech companies. Subtilisins represent a large class of microbial serine proteases, but the most mutagenized proteases are those from *B. amyloliquefaciens* (BPNP), *B. subtilis* (subtilisin E), and *Bacillus lentus* (Savinase).

Protein engineering involves several effective methods that include rational design and directed evolution. The former usually implies the methods of site-directed mutagenesis for replacing specific amino acid residues in a protein; this approach can help to obtain proteins with desired properties, such as higher thermal stability (Jaouadi et al., 2010; Huang et al., 2015). Besides, protein engineering can help to elucidate substrate recognition and point to possible applications of an enzyme (Jaouadi et al., 2014). On the other hand, directed evolution is based on the execution of sequential cycles of mutagenesis and selection (Liu et al., 2014). Thus, researchers may obtain enzymes with higher activity and stability under various conditions, including extreme pH and temperatures, nontraditional media, and modified specificity toward given substrates.

Stability of subtilisin

Its stability has been urgently needed for its manufacture; in this regard, such studies have become widespread. An interesting feature of subtilisin is that its biosynthesis requires participation of its N-terminal pro-domain (Ikemura et al., 1987). Folding of mature subtilisin without the pro-domain is possible theoretically but will take thousands of years.

An important characteristic of subtilisin is its tremendous dependence on calcium (Voordouw et al., 1976; Genov et al., 1995). The universal characteristic of subtilisins is the presence of one or more sites for calcium binding. High-resolution X-ray structures of subtilisin BPNP and of several homologs (Bode et al., 1987; Betzel et al., 1992) have uncovered the details of a conserved calcium-binding site, which is called site A. Calcium in this site is coordinately bound by five carbonyl atoms of oxygen and an asparagine acid residue. Four of the oxygen atoms are provided by the loop containing amino acid residues 75–83. The geometry of the ligands is a pentagonal bipyramid, whose axis crosses the carbonyl groups of amino acid residues 75 and 79. On one side of the loop, bidentate carboxylate (D41) is located, and on the other, the N terminus of the protein and side chain Q2. Seven coordinate distances vary from 2.3 to 2.6 A, the shortest of which involves aspartyl carboxylate.

The second ion-binding site (site B) is located 32 A away from site A in a shallow cleft between two segments of the polypeptide chain near the molecule's surface. The coordination geometry of this region bears a striking resemblance to a distorted pentagonal bipyramid. Three of the formal ligands are derivatives of a protein and include an oxygen atom of carbonyl group E195 and two oxygen atoms from the carboxylate of the D197 side chain. Four water molecules complete the first coordination sphere.

Given that the dependence on calcium is undesirable, some research has been conducted to obtain stable subtilisins that do not depend on the presence or absence of calcium in solution. One research group (Strausberg et al., 2005) modified the amino acid sequence of subtilisin with a damaged calcium-binding site for increasing this enzyme's stability. As a result, they obtained a mutant enzyme that is 15,000-fold more stable than the original protein. To this end, 12 mutations were introduced into the gene of this enzyme.

The latest research on the modification of alkaline serine proteases

In spite of substantial progress in the development of customized properties of alkaline serine proteases, the work on their modification continues to this day. For instance, in one study (Zhao, Feng, 2018), via directed evolution, the authors obtained seven mutants (P9S, A1G/K27Q, A38V, A116T, T162I, S182R, and T243S) of a protease extracted from *Bacillus pumilus* BA06. They all possessed a higher proteolytic activity toward casein and a synthetic peptide substrate at 15 °C. Except for T243S, thermal stability of these mutant enzymes did not decrease relative to the wild-type enzyme. Combinations of mutations further increased the specific caseinolytic activity. Double mutants P9S/K27Q and P9S/T162I showed approximately a fivefold increase in the caseinolytic activity at 15 °C almost without a loss of thermal stability (Zhao, Feng, 2018). In another study by the same group (Zhao et al., 2016), directed mutagenesis was performed on the alkaline protease of *B. pumilus*. The resultant double mutant (W106K/V149I and W106K/M124L) possessed 2.5-fold higher activity in comparison with the original enzyme at 15 °C, whereas its stability at 60 and 70 °C was 2.7-fold and 5-fold higher, respectively (Zhao et al., 2016).

During a comparison of halotolerant subtilisins with unstable ones, researchers discovered six amino acid positions where polar amino acid residues were replaced with nonpolar ones. The researchers hypothesized that these substitutions may lead to higher thermal stability. To test this hypothesis, they carried out mutagenesis of the alcalase from strain *B. subtilis* No. 16 and subtilisin Carlsberg. As a result, there was respectively 1.2-fold and 1.8-fold greater resistance of the enzymes to higher salt concentrations (125 g/L) (Takenaka et al., 2018). In another work (Ashraf et al., 2019), a serine protease from *Pseudomonas aeruginosa* was modified at two positions (A29G and V336I); as a consequence, they achieved a 5 °C increase in the temperature of observed residual activity and 1.4-fold enhancement of the catalytic activity (Ashraf et al., 2019). In yet another study (Gong et al., 2017), statistical mutagenesis of an alkaline-protease gene discovered during a metagenomic analysis increased the enzymatic activity by 6.6-fold.

Preparation of proteases in the strains of *Bacillus* spp.

The *Bacillus* bacteria have been the main microbial producers of serine proteases throughout the whole period of their practical use. Cultivation conditions and composition of the media play an important role in the production of enzymes by microbes (Abidi et al., 2008). To achieve high and commercially significant expression of proteases, it is crucial to find the conditions for growth and induction (Sharma et al., 2015). There is no universal medium suitable for all producer strains. Each microorganism or strain has unique specific conditions for maximal production of a given enzyme. Let us review various parameters of cultivation in more detail.

Composition of media

Carbon and nitrogen are the main components of a medium and act as major stimulators of microbial growth and synthesis of enzymes. The most widespread source of carbon and often the cheapest (after starch) is glucose; however, during its consumption, the effect of catabolic repression of many biosynthetic processes may emerge in the cell. The highest production of the enzyme by bacterial strain AKS-4 is observed at a glucose concentration of 1 %. Under these conditions, the level of expression of the protease reaches 59.10 U/mL (Sharma et al., 2015). Higher production of proteases in *Bacillus pseudofirmus* AL-89 was observed after glucose addition, whereas for *Nesterenkonia* sp., the synthesis of protease AL-20 was found to be suppressed in the presence of glucose (Gessesse et al., 2003).

The highest production of alkaline protease (2450 U/mL) in *B. licheniformis* was achieved in a medium containing 60 g/L

glucose. A further increase in its concentration led to an insignificant decrease in the production of the enzyme. Glucose at a high concentration inhibited the synthesis of the enzyme in *Streptomyces* ssp., and concentration 0.5 % was optimal for the production of the enzyme, whereas 1 % was optimal for growth (Mehta et al., 2006). Production of a protease in *P. aeruginosa* MCMB-327 in a soy-tryptic medium decreased by 95 and 60 % after the addition of glucose and fructose, respectively (Zambare et al., 2011). In another work (Sharma et al., 2014), investigators tried various sources of carbon, such as glucose, lactose, galactose, and starch, for the production of a protease by *Bacillus aryabhattai* K3. The highest production of the protease (622.64 U/mL) was observed with lactose (10 g/L) as a carbon source (Sharma et al., 2014). Similarly, in yet another study (Dodia et al., 2006), researchers found that for most of the analyzed isolates, secretion of the enzyme is optimal with lactose as a carbon source. *B. licheniformis* BBRC 100053 also manifested higher productivity in terms of a protease in culture media containing lactose as a carbon source (Nejad et al., 2010).

Aside from simple sugars, investigators tried other carbon sources for the production of proteases. The addition of 5 % of starch resulted in the highest production of a protease by *Bacillus* sp. 2–5 (Khosravi-Darani et al., 2008). Strain *Bacillus clausii* No. 58 grew well on various carbon sources based on starch (Kumar et al., 2004). Corn starch at 0.5 % yielded the highest productivity in terms of the protease, followed by wheat flour and wheat bran. Nonetheless, the addition of potato starch lowered the titer of the protease, possibly because of the presence of protease inhibitors in potato (Kumar et al., 2004). Wheat flour as a sugar source gave good results on the production of proteases by Bacillus sp. (Chu, 2007). *Bacillus laterosporus* synthesizes proteases while utilizing various carbon sources; the best sources of carbon for the secretion of the protease are soluble starch, trisodium citrate, citric acid, and glycerol (Usharani, Muthuraj, 2010).

Nitrogen sources also significantly affect the yield of a desired protein, and optimal sources vary among different strains. The highest level of protease production by strain *Bacillus cereus* 146 was observed in the presence of a beef extract as a nitrogen source. The presence of a yeast extract, peptone, and tryptone improved the growth parameters of cultures, but the amount of the desired protein was still modest (Shafee et al., 2005). It was demonstrated in another study that tryptone increases the protease synthesis by strain *Bacillus* sp. (Srinivasan et al., 2009). Peptone was found to be optimal for the production of a protease by *B. licheniformis* BBRC 100053 (Nejad et al., 2010). The yeast extract causes the biggest increase in the production of enzymes by *Bacillus* sp. (Prakasham et al., 2006). In case of *Bacillus* sp. APP1, among all the tested sources of organic nitrogen, soy protein meal noticeably raised the synthesis of an extracellular protease (Chu, 2007). Some authors (Jaswal et al., 2008) also reported that the addition of soy protein meal gave the best results in comparison with casein, gelatin, and peptone for the expression of a protease by *Bacillus circulans*. When casein, peptone, the yeast extract, and a beef extract were tested as a nitrogen source for the synthesis of a protease by bacterial strain AKS-4, the highest expression was observed in the presence of casein. Among the various sources of organic nitrogen, nonfat milk gave the highest yield of a protease in the case of *Bacillus caseinilyticus*, followed by a malt extract, peptone, and the yeast extract. Ammonium chloride as an inorganic source of nitrogen inhibits the synthesis of a proteinase (Mothe, 2016).

The influence of pH and temperature on the expression levels of proteases

The impact of pH on the expression level of a desired product is unique for each producer strain. For example, for the expression of proteases in *Bacillus* sp. MIG (Gouda, 2006) and *B. cereus* SIU1(Singh et al., 2010), weakly acidic pH (6.3–6.5) was found to be optimal. In a weakly alkaline medium (pH 8.0–8.5), researchers noted the highest levels of expression for *B. licheniformis* IKBC-17 (Olajuyigbe et al., 2005), *B. subtilis* IKBS 10 (Olajuyigbe et al., 2005), *Bacillus macerans* IKBM-11 (Olajuyigbe et al., 2005), and *B. amovivorus* (Sharmin et al., 2005). In one study on eight isolates of *Bacillus* (Dodia et al., 2006), it was revealed that the best conditions for the growth of bacteria involve pH 9.0, whereas the optimal pH value for the secretion of the enzyme varies between 8.0 and 10.0. pH 9 was found to be optimal for the production of proteases in *Bacillus* sp. (Prakasham et al., 2006), *Bacillus* sp. APP1 (Chu, 2007), and *B. proteolyticus* CFR3001 (Bhaskar et al., 2007). Higher starting pH was set up for the production of a protease by *B. licheniformis* TISTR 1010 (pH 10.0) (Vaithanomsat et al., 2008), for *B. circulans* (pH 10.5) (Jaswal et al., 2008), and for *Bacillus* sp. 2–5 (pH 10.7) (Khosravi-Darani et al., 2008).

Temperature is also a crucial parameter, and the optimal temperature is unique for each strain. For *P. aeruginosa* PseA (Gupta, Khare, 2007), *B. licheniformis* (Asokan, Jayanthi, 2010), *Bacillus coagulans* (Asokan, Jayanthi, 2010), *B. cereus* (Kebabcı, Cihangir, 2010), *P. aeruginosa* MCMB-327 (Zambare et al., 2011), *P. chrysogenum* IHH5 (Ikram-Ul-Haq et al., 2006), and *A. oryzae* 637 (Srinubabu et al., 2007), the optimal temperature for the synthesis of proteases is 30 °C. A lower optimal temperature (25 °C) characterizes *B. circulans* (Jaswal et al., 2008) and *Microbacterium* sp. (Thys et al., 2006), whereas in *B. cinerea*, the highest expression was documented at 28 °C (Abidi et al., 2008). At 37 °C, the maximal level of expression was observed for the strains of *Bacillus amovivorus* (Sharmin et al., 2005), *B. proteolyticus* CFR3001 (Bhaskar et al., 2007), *Bacillus aquimaris* VITP4 (Shivanand, Jayaraman, 2009), and *B. subtilis* Rand (Abusham et al., 2009); at 40 °C for *Bacillus* sp. 2–5 (Khosravi-Darani et al., 2008), *Vibrio pantothenticus* (Gupta et al., 2008), and *Streptomyces roseiscleroticus* (Shivanand, Jayaraman, 2009); and at 50 °C for *Bacillus* sp. APP1 (Porres et al., 2002) and *B. subtilis* BS1 (Shaheen et al., 2008).

Expression of alkaline serine proteases in yeast

The synthesis of proteases is possible not only in the strains of *Bacillus* but also in other bacteria and in yeast, e. g., in the strains of *Pichia pastoris*. These strains naturally do not have a specific activity; for this reason, they require modification

by genetic engineering. There are few such studies, and for the most part, they are aimed at obtaining fungal proteases or medically important proteases.

In one study, B. Liu et al. (2014) performed an analysis of expression of the keratinase gene in *B. licheniformis* BBE11-1 in three heterologous expression systems: in *Escherichia coli*, *B. subtilis*, and *P. pastoris*. The highest (best) level of expression was seen in *B. subtilis* (3010 U/mL); this level was threefold higher than that in *P. pastoris*. It should be noted that the cultivation of *B. subtilis* does not involve methanol, and cultivation duration is twofold shorter. In another study (Radha, Gunasekaran, 2009), there is a description of comparative cloning of the keratinase from *B. licheniformis* MKU3 in *B. megaterium* and *P. pastoris*. As a result, those authors obtained comparable activities of the final culture with the concentration of the desired protein at ~0.35 g/L. The protein from *P. pastoris* was subject to glycosylation. It should be mentioned that cultivation in a bioreactor was not described in that work. Similar results were published about the expression of the keratinase from *B. licheniformis* PWD-1 (Cheng et al., 1995).

In one work (Lin et al., 2009), researchers investigated the expression of the keratinase from *P. aeruginosa* in *P. pastoris*. The expression level was approximately 0.5 g of the protein per liter. In this case, the protein did not undergo glycosylation. In another work (Zhou et al., 2017), protein subtilisin QK (from *B. subtilis* QK02), which is highly similar to nattokinase, was cloned in *P. pastoris* GS115. Their objective was to obtain a protein with thrombolytic effects. As a result, they achieved a high concentration of total protein in the final supernatant (7.6 g/L). In this work, pH was maintained at 5.0, whereas in other studies (Liu et al., 2014) and (Porres et al., 2002) – the absence of pH control caused a pH increase, resulting in inhibition of microbial growth and a drop in keratinase concentration in solution. A similar picture was observed in a study by H.H. Lin et al. (2009).

Cloning of the alkaline protease from thermophilic bacterium *B. stearothermophilus* F1 was also conducted in *P. pastoris* GS115 (Latiffi et al., 2013). The resultant activity was 4.13 U/mL; judging by the obtained molecular weight, the protein was not glycosylated. In one study (Ke et al., 2018), the gene of the alkaline protease from fungus *Aspergillus sojae* was expressed in *P. pastoris*, and the final activity reached 400 U/mL.

The level of expression of a desired protein is strongly affected by codon usage too. In one work (Hu et al., 2013), as a result of optimization of codon usage in a gene, the level of expression of the desired protein was raised relative to the original gene. That study, however, does not present the data on the cultivation under controlled conditions of a bioreactor. An increased copy number of the expression cassette also allows for improving the yield of a desired protein, as exemplified by a serine protease from the fungus *Trichoderma koningii* (Shu et al., 2016).

In conclusion of this section, it is worth noting that the highest accumulation of alkaline serine proteases is greater in the *P. pastoris* expression system than in the *E. coli* expression system, but lower than that in standard *B. subtilis* strains. At the same time, the industrial strains of *Bacillus* spp. outperform both *P. pastoris* expression systems and *B. subtilis* by more than an order of magnitude. A 2005 patent (Shih, 2005) describes strain *B. licheniformis* T1, which ensures the expression level of a protein at 16 g/L, whereas the highest concentration of keratinase produced in *P. pastoris* is approximately 0.1–0.2 g of the desired protein per liter.

Conclusion

Alkaline serine proteases of the subtilisin family are widely applied in various industrial sectors. Proteases isolated from *Bacillus* bacteria constitute approximately 60 % of all enzyme sales across the globe.

Currently, native enzymes, i. e., those found in nature, are hardly used and have been ousted by the proteins modified via genetic engineering and thus possessing better properties, e. g., thermal stability, general resistance to detergents and specific resistance to various oxidants, high activity in various temperature ranges, independence from metal ions, and stability in the absence of calcium.

At present, diverse strains of *Bacillus* serve as microbial producers of alkaline serine proteases. Most of them originally had the desired activity, which has been enhanced by mutagenesis or genetic engineering. Among the producer strains, the species having the GRAS (generally regarded as safe) status dominate, that is, those that are even considered safe to eat: mostly *B. subtilis* and *B. licheniformis*. The strains that originally did not possess a protease activity still cannot be brought to the level of the native producers, even by means of genetic engineering technologies.

In literature, there are reports of the efforts to construct microbial producers of alkaline serine proteases on the basis of a methylotrophic strain of *P. pastoris*. In comparison with the expression of the same genes in the genetically engineered strains of *B. subtilis*, the results have turned out to be noticeably worse. From the above observations, it can be concluded that for constructing the strains effectively producing desired alkaline proteases, it is necessary to employ *Bacillus*-based expression systems. These strains need optimization of the properties of the expressed enzyme and of its expression level by methods of directed and statistical mutagenesis. Patentable microbial producers of an alkaline serine protease (subtilisin A) can be created via a search for new natural strains or may be constructed from the strains that lost patent protection.

References

Abidi F., Limam F., Nejib M.M. Production of alkaline proteases by *Botrytis cinerea* using economic raw materials: Assay as biodetergent. *Process Biochem*. 2008;43(11):1202-1208. DOI 10.1016/j.procbio.2008.06.018.

Abusham R.A., Rahman R.N.Z.R.A., Salleh A., Basri M. Optimization of physical factors affecting the production of thermo-stable organic solvent-tolerant protease from a newly isolated halo tolerant *Bacillus subtilis* strain Rand. *Microb. Cell Fact.* 2009;8(1):20. DOI 10.1186/1475-2859-8-20.

Anandharaj M., Sivasankari B., Siddharthan N., Rani R.P., Sivakumar S. Production, purification, and biochemical characterization of thermostable metallo-protease from novel *Bacillus alkalitelluris* TWI3 isolated from tannery waste. *Appl. Biochem. Biotechnol.* 2016;178(8):1666-1686. DOI 10.1007/s12010-015-1974-7.

Ashraf N.M., Krishnagopal A., Hussain A., Kastner D., Sayed A.M.M., Mok Y.-K., Swaminathan K., Zeeshan N. Engineering of serine protease for improved thermostability and catalytic activity using rational design. *Int. J. Biol. Macromol.* 2019;126:229-237. DOI 10.1016/j.ijbiomac.2018.12.218.

Asokan S., Jayanthi C. Alkaline protease production by *Bacillus licheniformis* and *Bacillus coagulans*. *J. Cell Tissue Res*. 2010;10(1): 2119-2123.

Barrett A.J., McDonald J.K. Nomenclature: protease, proteinase and peptidase. *Biochem. J.* 1986;237(3):935. DOI 10.1042/bj2370935.

Betzel C., Klupsch S., Papendorf G., Hastrup S., Branner S., Wilson K.S. Crystal structure of the alkaline proteinase Savinase™ from *Bacillus lentus* at 1.4 Å resolution. *J. Mol. Biol.* 1992;223(2):427-445. DOI 10.1016/0022-2836(92)90662-4.

Bhaskar N., Sudeepa E.S., Rashmi H.N., Tamil Selvi A. Partial purification and characterization of protease of *Bacillus proteolyticus* CFR3001 isolated from fish processing waste and its antibacterial activities. *Bioresour. Technol.* 2007;98(14):2758-2764. DOI 10.1016/j.biortech.2006.09.033.

Bode W., Papamokos E., Musil D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. Structural analysis, subtilisin structure and interface geometry. *Eur. J. Biochem.* 1987;166(3):673-692. DOI 10.1111/j.1432-1033.1987.tb 13566.x.

Cheng S.-W., Hu H.-M., Shen S.-W., Takagi H., Asano M., Tsai Y.-C. Production and characterization of keratinase of a feather-degrading *Bacillus licheniformis* PWD-1. *Biosci. Biotechnol. Biochem.* 1995; 59(12):2239-2243. DOI 10.1271/bbb.59.2239.

Cho S.J. Primary structure and characterization of a protease from *Bacillus amyloliquefaciens* isolated from *meju*, a traditional Korean soybean fermentation starter. *Process Biochem.* 2019;80:52-57. DOI 10.1016/j.procbio.2019.02.011.

Chu W.H. Optimization of extracellular alkaline protease production from species of *Bacillus*. *J. Ind. Microbiol. Biotechnol.* 2007; 34(3):241-245. DOI 10.1007/s10295-006-0192-2.

Dodia M.S., Joshi R.H., Patel R.K., Singh S.P. Characterization and stability of extracellular alkaline proteases from halophilic and alkaliphilic bacteria isolated from saline habitat of coastal Gujarat, India. *Braz. J. Microbiol.* 2006;37(3):276-282. DOI 10.1590/S1517-83822006000300015.

Garcia-Carreno F.L., Navarrete Del Toro M.A. Classification of proteases without tears. *Biochem. Educ.* 1997;25(3):161-167. DOI 10.1016/S0307-4412(97)00005-8.

Genov N., Filippi B., Dolashka P., Wilson K.S., Betzel C. Stability of subtilisins and related proteinases (subtilases). *Int. J. Pept. Protein Res.* 1995;45(4):391-400. DOI 10.1111/j.1399-3011.1995. tb01054.x.

Gessesse A., Hatti-Kaul R., Gashe B.A., Mattiasson B. Novel alkaline proteases from alkaliphilic bacteria grown on chicken feather. *Enzyme Microb. Technol.* 2003;32(5):519-524. DOI 10.1016/S0141-0229(02)00324-1.

Gong B.L., Mao R.Q., Xiao Y., Jia M.L., Zhong X.L., Liu Y., Xu P.-L., Li G. Improvement of enzyme activity and soluble expression of an alkaline protease isolated from oil-polluted mud flat metagenome by random mutagenesis. *Enzyme Microb. Technol.* 2017;106:97-105. DOI 10.1016/j.enzmictec.2017.06.015.

Gouda M.K. Optimization and purification of alkaline proteases produced by marine *Bacillus* sp. MIG newly isolated from eastern harbour of Alexandria. *Pol. J. Microbiol.* 2006;55(2):119-126.

Gulmez C., Atakisi O., Dalginli K.Y., Atakisi E. A novel detergent additive: Organic solvent- and thermo-alkaline-stable recombinant subtilisin. *Int. J. Biol. Macromol.* 2019;108:436-443. DOI 2018;108: 436-443. https://doi.org/10.1016/j.ijbiomac.2017.11.133.

Gupta A., Joseph B., Mani A., Thomas G. Biosynthesis and properties of an extracellular thermostable serine alkaline protease from *Virgibacillus pantothenticus*. *World J. Microbiol. Biotechnol*. 2008; 24(2):237-243. https://doi.org/10.1007/s11274-007-9462-z.

Gupta A., Khare S.K. Enhanced production and characterization of a solvent stable protease from solvent tolerant *Pseudomonas aeruginosa* PseA. *Enzyme Microb. Technol.* 2007;42(1):11-16. DOI 10.1016/j.enzmictec.2007.07.019.

Hadjidj R., Badis A., Mechri S., Eddouaouda K., Khelouia L., Annane R., Hattab M.E., Jaouadi B. Purification, biochemical, and molecular characterization of novel protease from *Bacillus licheniformis* strain K7A. *Int. J. Biol. Macromol.* 2018;114:1033-1048. DOI 10.1016/j.ijbiomac.2018.03.167.

Harwood C.R., Cranenburgh R. *Bacillus* protein secretion: an unfolding story. *Trends Microbiol.* 2008;16(2):73-79. DOI 10.1016/j.tim. 2007.12.001.

Hu H., Gao J., He J., Yu B., Zheng P., Huang Z., Mau X., Yu J., Han G., Chen D. Codon optimization significantly improves the expression level of a keratinase gene in *Pichia pastoris*. *PLoS One*. 2013; 8(3):e58393. https://doi.org/10.1371/journal.pone.0058393.

Huang R., Yang Q., Feng H. Single amino acid mutation alters thermostability of the alkaline protease from *Bacillus pumilus*: Thermodynamics and temperature dependence. *Acta Biochim. Biophys. Sin.* 2015;47(2):98-105. DOI 10.1093/abbs/gmu120.

Ikemura H., Takagi H., Inouye M. Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J. Biol. Chem.* 1987;262(16):7859-7864.

Ikram-Ul-haq H.M., Umber H. Production of protease by *Penicillium chrysogenum* through optimization of environmental conditions. *J. Agric. Soc. Sci.* 2006;2(1):23-25.

Jaouadi B., Aghajari N., Haser R., Bejar S. Enhancement of the thermostability and the catalytic efficiency of *Bacillus pumilus* CBS protease by site-directed mutagenesis. *Biochimie*. 2010;92(4):360-369. DOI 10.1016/j.biochi.2010.01.008.

Jaouadi N.Z., Jaouadi B., Hlima H.B., Rekik H., Belhoul M., Hmidi M., Bejar S. Probing the crucial role of Leu31 and Thr33 of the *Bacillus pumilus* CBS alkaline protease in substrate recognition and enzymatic depilation of animal hide. *PLoS One*. 2014;9(9). DOI 10.1371/ journal.pone.0108367.

Jaswal R.K., Kocher G.S., Virk M.S. Production of alkaline protease by *Bacillus circulans* using agricultural residues: A statistical approach. *Ind. J. Biotechnol. (IJBT).* 2008;7(3):356-360.

Jeong Y.J., Baek S.C., Kim H. Cloning and characterization of a novel intracellular serine protease (IspK) from *Bacillus megaterium* with a potential additive for detergents. *Int. J. Biol. Macromol.* 2018;108: 808-816. DOI 10.1016/j.ijbiomac.2017.10.173.

Kalwasińska A., Jankiewicz U., Felföldi T., Burkowska-But A., Brzezinska M.S. Alkaline and halophilic protease production by *Bacillus luteus* H11 and its potential industrial applications. *Food Technol. Biotechnol.* 2018;56(4):553-561. DOI 10.17113/ftb.56.04.18.5553.

Ke Y., Yuan X.M., Li J.S., Zhou W., Huang X.H., Wang T. High-level expression, purification, and enzymatic characterization of a recombinant *Aspergillus sojae* alkaline protease in *Pichia pastoris*. *Protein Expr. Purif.* 2018;148:24-29. DOI 10.1016/j.pep.2018.03.009.

Kebabcı Ö., Cihangir N. Isolation of protease producing novel *Bacillus cereus* and detection of optimal conditions. *Afr. J. Biotechnol.* 2010; 10(7):1160-1164. DOI 10.5897/AJB10.164.

Khosravi-Darani K., Falahatpishe H.R., Jalali M. Alkaline protease production on date waste by an alkalophilic *Bacillus* sp. 2-5 isolated from soil. *Afr. J. Biotechnol.* 2008;7(10):1536-1542.

Kobayashi T., Hakamada Y., Adachi S., Hitomi J., Yoshimatsu T., Koike K., Ito S. Purification and properties of an alkaline protease from alkalophilic *Bacillus* sp. KSM-K16. *Appl. Microbiol. Biotechnol.* 1995;43(3):473-481. DOI 10.1007/BF00218452.

Kobayashi T., Lu J., Li Z., Hung V.S., Kurata A., Hatada Y., Takai K., Ito S., Horikoshi K. Extremely high alkaline protease from a deep-subsurface bacterium, *Alkaliphilus transvaalensis. Appl. Microbiol. Biotechnol.* 2007;75(1):71-80. DOI 10.1007/s00253-006-0800-0.

Kumar C.G., Joo H.S., Koo Y.M., Paik S.R., Chang C.S. Thermostable alkaline protease from a novel marine haloalkalophilic *Bacillus clausii* isolate. *World J. Microbiol. Biotechnol.* 2004;20(4):351-357. DOI 10.1023/B:WIBI.0000033057.28828.a7.

Latiffi A.A., Salleh A.B., Rahman R.N.Z.R.A., Oslan S.N., Basri M. Secretory expression of thermostable alkaline protease from *Bacillus stearothermophilus* FI by using native signal peptide and α-factor secretion signal in *Pichia pastoris*. *Genes Genet. Syst.* 2013; 88(2):85-91. DOI 10.1266/ggs.88.85.

Lin H.H., Yin L.J., Jiang S.T. Functional expression and characterization of keratinase from *Pseudomonas aeruginosa* in *Pichia pastoris*. *J. Agric. Food Chem.* 2009;57(12):5321-5325. DOI 10.1021/jf 900417t.

Liu B., Zhang J., Gu L., Du G., Chen J., Liao X. Comparative analysis of bacterial expression systems for keratinase production. *Appl. Biochem. Biotechnol.* 2014;173(5):1222-1235. DOI 10.1007/s12010-014-0925-z.

Liu Y., Zhang T., Zhang Z., Sun T., Wang J., Lu F. Improvement of cold adaptation of *Bacillus alcalophilus* alkaline protease by directed evolution. *J. Mol. Catalys. B: Enzymatic*. 2014;106:117-123. DOI 10.1016/j.molcatb.2014.05.005.

Mathew C.D., Gunathilaka R.M.S. Production, purification and characterization of a thermostable alkaline serine protease from *Bacillus lichniformis* NMS-1. *Int. J. Biotechnol. Mol. Biol. Res*. 2015;6(3): 19-27. DOI 10.5897/IJBMBR2014.0199.

Mehta V.J., Thumar J.T., Singh S.P. Production of alkaline protease from an alkaliphilic actinomycete. *Bioresour. Technol*. 2006;97(14): 1650-1654. DOI 10.1016/j.biortech.2005.07.023.

Mothe T., Sultanpuram V.R. Production, purification and characterization of a thermotolerant alkaline serine protease from a novel species *Bacillus caseinilyticus*. *3 Biotech*. 2016;6(1):1-10. DOI 10.1007/ s13205-016-0377-y.

Nejad Z., Yaghmaei S., Hosseini R. Production of extracellular protease and determination of optimal condition by *Bacillus licheniformis* BBRC 100053. *Chem. Eng. Trans.* 2010;1(3):1447-1452. DOI 10.3303/CET1021242.

Okuda M., Sumitomo N., Takimura Y., Ogawa A., Saeki K., Kawai S., Kobayashi T., Ito S. A new subtilisin family: Nucleotide and deduced amino acid sequences of new high-molecular-mass alkaline proteases from *Bacillus* spp. *Extremophiles*. 2004;8(3):229-235. DOI 10.1007/s00792-004-0381-8.

Olajuyigbe F.M., Ajele J.O., Ajele J.O. Production dynamics of extracellular protease from *Bacillus* species. *Afr. J. Biotechnol.* 2005; 4(8):776-779.

Ottesen M., Svendsen I. The Subtilisins. In: Methods in Enzymology. Academic Press, 1970;19:199-215. DOI 10.1016/0076-6879(70) 19014-8.

Porres J.M., Benito M.J., Lei X.G. Functional expression of keratinase (kerA) gene from *Bacillus licheniformis* in *Pichia pastoris*. *Biotechnol. Lett.* 2002;24(8):631-636. DOI 10.1023/A:1015083007746.

Prakasham R.S., Rao C.S., Sarma P.N. Green gram husk-an inexpensive substrate for alkaline protease production by *Bacillus* sp. in solid-state fermentation. *Bioresour. Technol.* 2006;97(13):1449-1454. DOI 10.1016/j.biortech.2005.07.015.

Radha S., Gunasekaran P. Purification and characterization of keratinase from recombinant *Pichia* and *Bacillus* strains. *Protein Expr. Purif.* 2009;64(1):24-31. DOI 10.1016/j.pep.2008.10.008.

Rawlings N.D., Waller M., Barrett A.J., Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 2014;42(D1):D503-D509. DOI 10.1093/nar/ gkt953.

Rehman R., Ahmed M., Siddique A., Hasan F., Hameed A., Jamal A. Catalytic role of thermostable metalloproteases from *Bacillus subtilis* KT004404 as dehairing and destaining agent. *Appl. Biochem. Biotechnol.* 2017;181(1):434-450. DOI 10.1007/s12010-016-2222-5.

Shafee N., Aris S., Rahman R., Basri M., Salleh A. Optimization of environmental and nutritional conditions for the production of alkaline protease by a newly isolated bacterium *Bacillus* cereus strain 146. *J. Appl. Sci. Res.* 2005;1(1):1-8.

Shaheen M., Shah A., Hameed A., Hasan F. Influence of culture conditions on production and activity of protease from *Bacillus subtilis* BS1. *Pak. J. Bot.* 2008;40(5):2161-2169.

Sharma A., Sharma V., Saxena J., Yadav B., Alam A., Prakash A. Optimization of protease production from bacteria isolated from soil. *Appl. Res. J.* 2015;1(7):388-394.

Sharma K., Kumar R., Vats S., Gupta A. Production, partial purification and characterization of alkaline protease from *Bacillus aryabhattai* K3. *Int. J. Adv. Pharm. Biol. Chem*. 2014;3(2):290-298.

Sharma K.M., Kumar R., Panwar S., Kumar A. Microbial alkaline proteases: Optimization of production parameters and their properties. *J. Genet. Eng. Biotechnol.* 2017;15:115-126. DOI 10.1016/ j.jgeb.2017.02.001.

Sharmin S., Hossain T., Anwar M. Isolation and characterization of a protease producing bacteria *Bacillus amovivorus* and optimization of some factors of culture conditions for protease production. *J. Biol. Sci.* 2005;5(3):358-362. DOI 10.3923/jbs.2005.358.362.

Shih J. Construction of bacillus licheniformis t1 strain and fermentation production of crude enzyme extract therefrom. Patent No. US20050032188A1, 2005.

Shikha, Sharan A., Darmwal N.S. Improved production of alkaline protease from a mutant of alkalophilic *Bacillus pantotheneticus* using molasses as a substrate. *Bioresour. Technol.* 2007;98(4):881-885. DOI 10.1016/j.biortech.2006.03.023.

Shivanand P., Jayaraman G. Production of extracellular protease from halotolerant bacterium, *Bacillus aquimaris* strain VITP4 isolated from Kumta coast. *Process Biochem*. 2009;44(10):1088-1094. DOI 10.1016/j.procbio.2009.05.010.

Shivasharana C.T., Naik G.R. Ecofriendly applications of thermostable alkaline protease produced from a *Bacillus* sp. JB-99 under solid state fermentation. *Int. J. Environ. Sci*. 2012;3(3):956-964. DOI 10.6088/ijes.2012030133003.

Shu M., Shen W., Yang S., Wang X., Wang F., Wang Y., Ma L. High-level expression and characterization of a novel serine protease in *Pichia pastoris* by multi-copy integration. *Enzyme Microb. Technol*. 2016;92:56-66. DOI.10.1016/j.enzmictec.2016.06.007.

Singh S.K., Tripathi V.R., Jain R.K., Vikram S., Garg S.K. An antibiotic, heavy metal resistant and halotolerant *Bacillus cereus* SIU1 and its thermoalkaline protease. *Microb. Cell Fact*. 2010;9(1):59. DOI 10.1186/1475-2859-9-59.

Smith E.L., Markland F.S., Kasper C.B., DeLange R.J., Landon M., Evans W.H. The complete amino acid sequence of two types of subtilisin, BPN' and Carlsberg. *J. Biol. Chem.* 1966;241(24):5974-5976.

Srinivasan T., Das S., Balakrishnan V., Philip R., Kannan N. Isolation and characterization of thermostable protease producing bacteria from tannery industry effluent. *Recent Res. Sci. Technol.* 2009;1(2): 63-66.

Srinubabu G., Lokeswari N., Jayaraju K. Screening of nutritional parameters for the production of protease from *Aspergillus oryzae*. *Electr. J. Chem*. 2007;4(2):208-215. DOI 10.1155/2007/915432.

Strausberg S.L., Ruan B., Fisher K.E., Alexander P.A., Bryan P.N. Directed coevolution of stability and catalytic activity in calcium-free subtilisin. *Biochemistry*. 2005;44(9):3272-3279. DOI 10.1021/ bi047806m.

Takenaka S., Yoshinami J., Kuntiya A., Techapun C., Leksawasdi N., Seesuriyachan P., Chaiyaso I., Watanabe M., Tanaka K., Yoshida K. Characterization and mutation analysis of a halotolerant serine protease from a new isolate of *Bacillus subtilis. Biotechnol. Lett.* 2018; 40(1):189-196. DOI 10.1007/s10529-017-2459-2.

Thys R.C.S., Guzzon S.O., Cladera-Olivera F., Brandelli A. Optimization of protease production by *Microbacterium* sp. in feather

meal using response surface methodology. *Process Biochem*. 2006; 41(1):67-73. DOI 10.1016/j.procbio.2005.03.070.

Tufvesson P., Lima-Ramos J., Nordblad M., Woodley J.M. Guidelines and cost analysis for catalyst production in biocatalytic processes. *Org. Process Res. Dev*. 2010;15(1):266-274. DOI 10.1021/op1002165.

Usharani B., Muthuraj M. Production and characterization of protease enzyme from *Bacillus laterosporus. Afr. J. Microbiol. Res.* 2010; 4(11):1057-1063.

Vaithanomsat P., Malapant T., Apiwattanapiwat W. Silk degumming solution as substrate for microbial protease production. *Nat. Sci*. 2008;42:543-551.

Voordouw G., Milo C., Roche R.S. Role of bound calcium ions in thermostable, proteolytic enzymes. Separation of intrinsic and calcium ion contributions to the kinetic thermal stability. *Biochemistry*. 1976;15(17):3716-3724. DOI 10.1021/bi00662a012.

Zambare V., Nilegaonkar S., Kanekar P. A novel extracellular protease from *Pseudomonas aeruginosa* MCM B-327: enzyme production and its partial characterization. *New Biotechnol*. 2011;28(2): 173-181. DOI 10.1016/j.nbt.2010.10.002.

Zhao H.Y., Feng H. Engineering *Bacillus pumilus* alkaline serine protease to increase its low-temperature proteolytic activity by directed evolution. *BMC Biotechnol*. 2018;18(1):34. DOI 10.1186/s12896-018-0451-0.

Zhao H.Y., Wu L.Y., Liu G., Feng H. Single-site substitutions improve cold activity and increase thermostability of the dehairing alkaline protease (DHAP). *Biosci. Biotechnol. Biochem.* 2016;80(12):2480-2485. DOI 10.1080/09168451.2016.1230005.

Zhou K., Dong Y., Zheng H., Chen B., Mao R., Zhou L., Wang Y. Expression, fermentation, purification and lyophilisation of recombinant Subtilisin QK in *Pichia pastoris*. *Process Biochem*. 2017; 54:1-8. DOI 10.1016/j.procbio.2016.12.028.

ORCID ID

S.V. Shekhovtsov orcid.org/0000-0001-5604-5601
E.G. Pershina orcid.org/0000-0003-2658-7906
S.E. Peltek orcid.org/0000-0002-3524-0456

Acknowledgements. The study was supported by the budget project No. 0259-2019-0005. This work was done within the framework of State Assignment Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences (075-15-2019-1662).

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

Received November 17, 2020. Revised December 17, 2020. Accepted December 21, 2020.