2026 • 30 • 1

Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

VAVILOV JOURNAL OF GENETICS AND BREEDING

Основан в 1997 г.
Периодичность 8 выпусков в год
doi 10.18699/vjgb-26-01

Учредители

Сибирское отделение Российской академии наук

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

Главный редактор

А.В. Кочетов – академик РАН, д-р биол. наук, профессор РАН (Россия)

Заместители главного редактора

Н.А. Колчанов – академик РАН, д-р биол. наук, профессор (Россия)
И.Н. Леонова – д-р биол. наук (Россия)
Н.Б. Рубцов – д-р биол. наук, профессор (Россия)
В.К. Шумный – академик РАН, д-р биол. наук, профессор (Россия)

Ответственный секретарь

Г.В. Орлова – канд. биол. наук (Россия)

Редакционная коллегия

Е.Е. Андронов – канд. биол. наук (Россия)
Ю.С. Аульченко – д-р биол. наук (Нидерланды)
О.С. Афанасенко – академик РАН, д-р биол. наук (Россия)
Д.А. Афонников – д-р биол. наук, доцент (Россия)
Л.И. Афтанас – академик РАН, д-р мед. наук (Россия)
Л.А. Беспалова – академик РАН, д-р с.-х. наук (Россия)
А. Бёрнер – д-р наук (Германия)
Н.П. Бондарь – канд. биол. наук (Россия)
С.А. Боринская – д-р биол. наук (Россия)
П.М. Бородин – д-р биол. наук, проф. (Россия)
А.В. Васильев – чл.-кор. РАН, д-р биол. наук (Россия)
М.И. Воевода – академик РАН, д-р мед. наук (Россия)
Т.А. Гавриленко – д-р биол. наук (Россия)
Н.Е. Грунтенко – д-р биол. наук (Россия)
С.А. Демаков – д-р биол. наук (Россия)
И.К. Захаров – д-р биол. наук, проф. (Россия)
И.А. Захаров-Гезехус – чл.-кор. РАН, д-р биол. наук (Россия)
С.Г. Инге-Вечтомов – академик РАН, д-р биол. наук (Россия)
А.В. Кильчевский – чл.-кор. НАНБ, д-р биол. наук (Беларусь)
А.М. Кудрявцев – чл.-кор. РАН, д-р биол. наук (Россия)
Д.М. Ларкин – канд. биол. наук (Великобритания)
С.А. Лашин – д-р биол. наук (Россия)
Ж. Ле Гуи – д-р наук (Франция)
И.Н. Лебедев – чл.-кор. РАН, д-р биол. наук, проф. (Россия)
Л.А. Лутова – д-р биол. наук, проф. (Россия)
Б. Люгтенберг – д-р наук, проф. (Нидерланды)
В.Ю. Макеев – чл.-кор. РАН, д-р физ.-мат. наук (Россия)
И.В. Максимов – д-р биол. наук (Россия)
Б.А. Малярчук – д-р биол. наук (Россия)
Ю.Г. Матушкин – канд. биол. наук (Россия)
В.И. Молодин – академик РАН, д-р ист. наук (Россия)
М.П. Мошкин – д-р биол. наук, проф. (Россия)
Л.Ю. Новикова – д-р с.-х. наук (Россия)
Е.К. Потокина – д-р биол. наук (Россия)
В.П. Пузырев – академик РАН, д-р мед. наук (Россия)
Д.В. Пышный – чл.-кор. РАН, д-р хим. наук (Россия)
Е.Ю. Рыкова – д-р биол. наук (Россия)
Е.А. Салина – чл.-кор. РАН, д-р биол. наук, проф. (Россия)
В.А. Степанов – академик РАН, д-р биол. наук (Россия)
И.А. Тихонович – академик РАН, д-р биол. наук (Россия)
Е.К. Хлесткина – чл.-кор. РАН, д-р биол. наук, проф. РАН (Россия)
Э.К. Хуснутдинова – д-р биол. наук, проф. (Россия)
М. Чен – д-р биол. наук (Китайская Народная Республика)
Е.В. Шахтшнейдер – д-р мед. наук (Россия)
С.В. Шестаков – академик РАН, д-р биол. наук (Россия)
С.В. Шеховцов – д-р биол. наук (Россия)
Н.С. Юдин – канд. биол. наук (Россия)
Н.К. Янковский – академик РАН, д-р биол. наук (Россия)

2026 • 30 • 1

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Publication frequency: 8 issues a year
doi 10.18699/vjgb-26-01

Founders

Siberian Branch of the Russian Academy of Sciences

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

The Vavilov Society of Geneticists and Breeders

Editor-in-Chief

A.V. Kochetov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Professor of the RAS, Russia

Deputy Editor-in-Chief

N.A. Kolchanov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia
I.N. Leonova, Dr. Sci. (Biology), Russia
N.B. Rubtsov, Professor, Dr. Sci. (Biology), Russia
V.K. Shumny, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia

Executive Secretary

G.V. Orlova, Cand. Sci. (Biology), Russia

Editorial board

O.S. Afanasenko, Full Member of the RAS, Dr. Sci. (Biology), Russia
D.A. Afonnikov, Associate Professor, Dr. Sci. (Biology), Russia
L.I. Aftanas, Full Member of the RAS, Dr. Sci. (Medicine), Russia
E.E. Andronov, Cand. Sci. (Biology), Russia
Yu.S. Aulchenko, Dr. Sci. (Biology), The Netherlands
L.A. Bespalova, Full Member of the RAS, Dr. Sci. (Agricul.), Russia
N.P. Bondar, Cand. Sci. (Biology), Russia
S.A. Borinskaya, Dr. Sci. (Biology), Russia
P.M. Borodin, Professor, Dr. Sci. (Biology), Russia
A. Börner, Dr. Sci., Germany
M. Chen, Dr. Sci. (Biology), People's Republic of China
S.A. Demakov, Dr. Sci. (Biology), Russia
T.A. Gavrilenko, Dr. Sci. (Biology), Russia
N.E. Gruntenko, Dr. Sci. (Biology), Russia
S.G. Inge-Vechtomov, Full Member of the RAS, Dr. Sci. (Biology), Russia
E.K. Khlestkina, Corr. Member of the RAS, Professor of the RAS, Dr. Sci. (Biology), Russia
E.K. Khusnutdinova, Professor, Dr. Sci. (Biology), Russia
A.V. Kilchevsky, Corr. Member of the NAS of Belarus, Dr. Sci. (Biology), Belarus
A.M. Kudryavtsev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
D.M. Larkin, Cand. Sci. (Biology), Great Britain
S.A. Lashin, Dr. Sci. (Biology), Russia
J. Le Gouis, Dr. Sci., France
I.N. Lebedev, Corr. Member of the RAS, Professor, Dr. Sci. (Biology), Russia
B. Lugtenberg, Professor, Dr. Sci., Netherlands
L.A. Lutova, Professor, Dr. Sci. (Biology), Russia
V.Yu. Makeev, Corr. Member of the RAS, Dr. Sci. (Physics and Mathem.), Russia
I.V. Maksimov, Dr. Sci. (Biology), Russia
B.A. Malyarchuk, Dr. Sci. (Biology), Russia
Yu.G. Matushkin, Cand. Sci. (Biology), Russia
V.I. Molodin, Full Member of the RAS, Dr. Sci. (History), Russia
M.P. Moshkin, Professor, Dr. Sci. (Biology), Russia
L.Yu. Novikova, Dr. Sci. (Agricul.), Russia
E.K. Potokina, Dr. Sci. (Biology), Russia
V.P. Puzyrev, Full Member of the RAS, Dr. Sci. (Medicine), Russia
D.V. Pyshnyi, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia
E.Y. Rykova, Dr. Sci. (Biology), Russia
E.A. Salina, Corr. Member of the RAS, Professor, Dr. Sci. (Biology), Russia
E.V. Shakhtshneider, Dr. Sci. (Medicine), Russia
S.V. Shekhovtsov, Dr. Sci. (Biology), Russia
S.V. Shestakov, Full Member of the RAS, Dr. Sci. (Biology), Russia
V.A. Stepanov, Full Member of the RAS, Dr. Sci. (Biology), Russia
I.A. Tikhonovich, Full Member of the RAS, Dr. Sci. (Biology), Russia
A.V. Vasiliev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
M.I. Voevoda, Full Member of the RAS, Dr. Sci. (Medicine), Russia
N.K. Yankovsky, Full Member of the RAS, Dr. Sci. (Biology), Russia
N.S. Yudin, Cand. Sci. (Biology), Russia
I.K. Zakharov, Professor, Dr. Sci. (Biology), Russia
I.A. Zakharov-Gezekhus, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

СОДЕРЖАНИЕ • 2026 • 30 • 1

Биоинформатика и системная биология

5 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Рациональный биоинформатический подход к анализу функциональных свойств метаболитов пробиотических микроорганизмов на основе реконструкции генных сетей. *В.А. Иванисенко, Т.М. Хлебодарова, М.А. Клещев, А.Р. Волянская, И.В. Яцык, А.В. Адамовская, Т.В. Иванисенко, П.С. Деменков, Н.А. Колчанов*

15 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Развитие метода филогенетического футпринтинга для распознавания сайтов связывания транскрипционных факторов на основе использования bootstrap-испытаний для анализа больших бактериальных геномных данных. *А.М. Мухин, Т.М. Хлебодарова, Д.Ю. Ощепков*

27 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Детекция колосков в колосе пшеницы на RGB-изображениях с использованием глубокого машинного обучения. *М.А. Генаев, И.Д. Бусов, Ю.В. Кручинина, В.С. Коваль, Н.П. Гончаров*

Генетика животных

36 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Паттерн распределения *P*-транспозона и *blood* ретротранспозона в отводках референсных линий Harwich и Canton-S из разных лабораторий не влияет на проявление внутривидового PM гибридного дисгенеза у *Drosophila melanogaster*. *Л.П. Захаренко, Ю.Ю. Илинский (на англ. языке)*

Генетика и селекция растений

43 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Локус сои qDTF-7 как пример генетической гетерогенности, ассоциированной с продолжительностью цветения и созревания. *Р.Н. Перфильев, М.И. Шматова, А.Б. Щербань, Е.А. Салина*

53 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Гистохимический тест для определения профиля полифенольных соединений в зерне злаков. *С.Р. Мурсалимов, О.Ю. Шоева (на англ. языке)*

Молекулярная и клеточная биология

61 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Модификация оптогенетической системы BphP1-QPAS1 для регуляции экспрессии генов в листьях табака *Nicotiana benthamiana* с помощью ближнего инфракрасного света. *Э.С. Суркова, Ю.А. Галимова, Н.В. Баттулина, Д.М. Моторина, Е.С. Омелина (на англ. языке)*

72 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Редактирование оснований в гене *AUTS2* и высокопроизводительное NGS-генотипирование клонов: стратегия создания клеточной модели. *А.П. Ян, П.А. Сальников, А.А. Буздин, В.А. Ковальская, Е.В. Мусатова, П.С. Орлов, О.П. Рыжкова, А.И. Субботовская, М.В. Сунцова, А.Ю. Христиченко, А.А. Хабарова*

85 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Биоинформатический анализ механизмов жизнеспособности линий опухолевых клеток при делеции генов-супермишеней. *Д.А. Четверина, Н.Я. Козельчук, Д.В. Ломаев, А.А. Штиль, М.М. Ерохин*

Генетика человека

94 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Особенности полиморфизма гена рецептора тиреоидных гормонов *THRB* у коренного населения Сибири. *Б.А. Малярчук, Н.В. Похилюк, Г.А. Денисова, А.Н. Литвинов*

101 ОБЗОР
Транскриптомика тяжелой формы COVID-19. *А.А. Гусарова, Е.А. Трифонова, А.А. Бабовская, М.М. Гавриленко, В.А. Степанов*

Генетика микроорганизмов

117 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Однонуклеотидные полиморфизмы в геномах вирусов клещевого энцефалита и Западного Нила при тройной природной инфекции у садовой камышовки (*Acrocephalus dumetorum*). *Е.П. Пономарева, В.А. Терновой, Е.В. Протопопова, Н.Л. Тупота, В.Б. Локтев*

126 ОБЗОР
Новые аргументы в дискуссии о природе пикобирнавирусов. *А.Ю. Кашников, Н.В. Епифанова, Н.А. Новикова*

Популяционная генетика

136 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Эволюция порядка генов в мтДНК байкальских эндемичных амфипод и ее возможные механизмы. *Е.А. Сиротинина, Д.Ю. Щербаков, Е.В. Романова*

146 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Морфологический и молекулярно-генетический анализ полиморфизма рода *Iris* L. в Республике Башкортостан и Оренбургской области. *Э.А. Аухадиева, А.Г. Блинов, Е.И. Вшивцева, З.Х. Шигапов, А.Р. Ишбирдин, Я.М. Голованов, А.В. Крюкова*

© Сибирское отделение Российской академии наук, 2026
© Институт цитологии и генетики СО РАН, 2026
Вавиловский журнал генетики и селекции, 2026

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

CONTENTS • 2026 • 30 • 1

Bioinformatics and systems biology

5 ORIGINAL ARTICLE
Rational bioinformatic approach to the analysis of functional properties of metabolites of probiotic microorganisms based on gene network reconstruction.
V.A. Ivanisenko, T.M. Khlebodarova, M.A. Kleshchev, A.R. Volyanskaya, I.V. Yatsyk, A.V. Adamovskaya, T.V. Ivanisenko, P.S. Demenkov, N.A. Kolchanov

15 ORIGINAL ARTICLE
Improvement of a phylogenetic footprinting method for transcription factor binding sites recognition based on the use of bootstrap trials for the analysis of large bacterial genomic data. *A.M. Mukhin, T.M. Khlebodarova, D.Yu. Oshchepkov*

27 ORIGINAL ARTICLE
Wheat spikelet detection on RGB images using deep machine learning. *M.A. Genaev, I.D. Busov, Yu.V. Kruchinina, V.S. Koval, N.P. Goncharov*

Animal genetics

36 ORIGINAL ARTICLE
Different patterns of *P* transposon and *blood* retrotransposon distribution in Harwich and Canton-S sub-strains do not affect the manifestation of *Drosophila melanogaster* intraspecific PM hybrid dysgenesis. *L.P. Zakharenko, Y.Y. Ilinsky*

Plant genetics and breeding

43 ORIGINAL ARTICLE
Soybean locus qDTF-7 as an example of genetic heterogeneity associated with flowering and maturity time.
R.N. Perfil'ev, M.I. Shmatova, A.B. Shcherban, E.A. Salina

53 ORIGINAL ARTICLE
A histochemical assay for polyphenolic profiling in cereal grains.
S.R. Mursalimov, O.Yu. Shoeva

Molecular and cell biology

61 ORIGINAL ARTICLE
Modification of the BphP1-QPAS1 optogenetic system for gene expression regulation in *Nicotiana benthamiana* tobacco leaves using near-infrared light. *E.S. Surkova, Y.A. Galimova, N.V. Battulina, D.M. Motorina, E.S. Omelina*

72 ORIGINAL ARTICLE
Base editing in the *AUTS2* gene and high-throughput NGS genotyping of clones: a strategy for generating a cellular model.
A.P. Yan, P.A. Salnikov, A.A. Buzdin, V.A. Kovalskaia, E.V. Musatova, P.S. Orlov, O.P. Ryzhkova, A.I. Subbotovskaia, M.V. Suntsova, A.U. Khristichenko, A.A. Khabarova

85 ORIGINAL ARTICLE
The computational analysis of tumor cell sensitivity to supertarget deletion.
D.A. Chetverina, N.Y. Kozelchuk, D.V. Lomaev, A.A. Shtil, M.M. Erokhin

Human genetics

94 ORIGINAL ARTICLE
The specific features of the thyroid hormone receptor gene *THRB* polymorphism in indigenous populations of Siberia.
B.A. Malyarchuk, N.V. Pokhilyuk, G.A. Denisova, A.N. Litvinov

101 REVIEW
Transcriptomics of severe COVID-19.
A.A. Gusarova, E.A. Trifonova, A.A. Babovskaya, M.M. Gavrilenko, V.A. Stepanov

Microbial genetics

117 ORIGINAL ARTICLE
Single nucleotide polymorphisms are typical for tick-borne encephalitis and West Nile viruses during triple natural mixed infections in Blyth's reed warbler (*Acrocephalus dumetorum*). *E.P. Ponomareva, V.A. Ternovoi, E.V. Protopopova, N.L. Tupota, V.B. Loktev*

126 REVIEW
New arguments in the discussion about the nature of picobirnaviruses.
A.Yu. Kashnikov, N.V. Epifanova, N.A. Novikova

Population genetics

136 ORIGINAL ARTICLE
Evolution of gene order in mtDNA of Baikal endemic amphipods and its possible mechanisms.
E.A. Sirotinina, D.Yu. Sherbakov, E.V. Romanova

146 ORIGINAL ARTICLE
Morphological and molecular genetic analysis of the genus *Iris* L. polymorphism in the Republic of Bashkortostan and the Orenburg Region. *E.A. Aukhadieva, A.G. Blinov, E.I. Vshivtseva, Z.Kh. Shigapov, A.R. Ishbirdin, Ya.M. Golovanov, A.V. Kryukova*

© Siberian Branch RAS, 2026
© Institute of Cytology and Genetics, SB RAS, 2026
Vavilov Journal of Genetics and Breeding, 2026

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):5-14

doi 10.18699/vjgb-26-11

Rational bioinformatic approach to the analysis of functional properties of metabolites of probiotic microorganisms based on gene network reconstruction

V.A. Ivanisenko, T.M. Khlebodarova, M.A. Kleshchev, A.R. Volyanskaya, I.V. Yatsyk, A.V. Adamovskaya, T.V. Ivanisenko, P.S. Demenkov, N.A. Kolchanov

Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
salix@bionet.nsc.ru

Abstract. An important direction in industrial microbiology is the development of probiotic strains with valuable consumer properties. The probiotic industry is currently one of the most rapidly developing segments of the food and pharmaceutical sectors. Stearic (octadecanoic) acid C18:0 is one of the major metabolites present in the cell-free supernatant of the bacterium *Streptococcus thermophilus*, which is widely used in the production of fermented dairy products, including yogurt and cheese. *S. thermophilus* affects not only the texture and sensory properties of products, but also exhibits various probiotic effects, including antioxidant activity, modulation of the gut microbiota, inhibition of certain pathogens, and others. It is assumed that a number of probiotic effects exerted by *S. thermophilus* may be mediated through octadecanoic acid as one of its main metabolites. Octadecanoic acid C18:0, like other long-chain fatty acids, enters the human body via several mechanisms, including protein-mediated transport and passive diffusion across cell membranes. Inside the cell, octadecanoic acid serves not only as a substrate for the synthesis of triglycerides and other complex lipids, but, as shown in cell-based and *in vivo* models, also acts as a modulator of signaling and stress responses, including those associated with apoptosis. This is an important aspect of the influence of stearic acid on organism functioning, underpinning its anti-inflammatory and potentially anti-tumor effects. However, the molecular genetic mechanisms by which octadecanoic acid acts as a probiotic on the human organism remain insufficiently understood. In the present study, using our previously developed information – software system ANDSystem (employing machine learning and artificial intelligence for automatic extraction of knowledge from scientific texts and databases), we reconstructed gene networks regulating the intrinsic (mitochondrial) and extrinsic (death receptor-mediated) apoptotic pathways in human cells under the influence of stearic (octadecanoic) acid. To search for metabolites produced by probiotic microorganisms that may have beneficial therapeutic properties, we propose an approach that combines gene network reconstruction with differential gene expression analysis. Using this approach, we show that octadecanoic acid produced by *S. thermophilus* can control the intrinsic and extrinsic apoptotic pathways primarily via regulation of PTGS2 expression; the results indicate that cyclooxygenase-2 is a key regulator mediating the effect of octadecanoic acid on apoptosis-related genes.

Key words: industrial microbiology; functional microorganisms; probiotics; producer strains; metabolites; gene networks; ANDSystem

For citation: Ivanisenko V.A., Khlebodarova T.M., Kleshchev M.A., Volyanskaya A.R., Yatsyk I.V., Adamovskaya A.V., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. Rational bioinformatic approach to the analysis of functional properties of metabolites of probiotic microorganisms based on gene network reconstruction. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):5-14. doi 10.18699/vjgb-26-11

Funding. This study was supported by the Ministry of Science and Higher Education of the Russian Federation (Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019–2030), agreement No. 075-15-2025-516.

© Ivanisenko V.A., Khlebodarova T.M., Kleshchev M.A., Volyanskaya A.R., Yatsyk I.V., Adamovskaya A.V., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Рациональный биоинформатический подход к анализу функциональных свойств метаболитов пробиотических микроорганизмов на основе реконструкции генных сетей

В.А. Иванисенко, Т.М. Хлебодарова, М.А. Клещев, А.Р. Волянская, И.В. Яцык, А.В. Адамовская, Т.В. Иванисенко, П.С. Деменков, Н.А. Колчанов

Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
salix@bionet.nsc.ru

Аннотация. Важное направление промышленной микробиологии – создание штаммов пробиотиков, обладающих ценными потребительскими свойствами. Индустрия пробиотиков является одним из наиболее динамично развивающихся сегментов пищевой и фармацевтической промышленности. Стеариновая (октадекановая) кислота C18:0 – один из основных метаболитов в клеточно-свободном супернатанте бактерии *Streptococcus thermophilus*, широко используемой в производстве различных ферментированных молочнокислых продуктов, включая йогурт и сыр. *S. thermophilus* влияет не только на текстуру и вкусовые свойства продуктов, но и обладает различными пробиотическими эффектами, в том числе антиоксидантной активностью, модуляцией кишечной микробиоты, ингибированием определенных патогенов и др. Предполагается, что ряд пробиотических эффектов, которыми обладает *S. thermophilus*, может быть опосредован именно через октадекановую кислоту, как основной метаболит. Октадекановая кислота C18:0, как и другие длинноцепочечные жирные кислоты, поступает в организм человека с использованием различных механизмов белок-опосредованного транспорта и пассивной диффузии через мембрану клеток. В клетке стеариновая кислота не только служит субстратом для синтеза триглицеридов и других сложных липидов, но и, как показано на клеточных и *in vivo* моделях, является модулятором сигнальных и стресс-ответов, связанных в том числе с апоптозом. Это один из важных аспектов влияния стеариновой кислоты на функционирование организма, определяющий противовоспалительный и потенциально противоопухолевый эффекты. Однако молекулярно-генетические механизмы влияния октадекановой кислоты как пробиотика на организм человека в этом отношении остаются недостаточно изученными. В настоящем исследовании с помощью разработанной нами ранее информационно-программной системы ANDSystem, использующей методы машинного обучения и искусственного интеллекта и предназначенной для автоматического извлечения знаний из научных текстов и баз данных, были реконструированы генные сети регуляции внутреннего (митохондриального) и внешнего (индуцируемого рецепторами смерти) пути апоптоза клеток человека под влиянием стеариновой кислоты. Для поиска метаболитов, продуцируемых пробиотическими микроорганизмами, обладающих полезными терапевтическими свойствами, разработан новый поход, включающий реконструкцию генных сетей и анализ дифференциально экспрессирующихся генов. На его основе было показано, что стеариновая кислота, продуцируемая *S. thermophilus*, контролирует как внешний, так и внутренний пути апоптоза через регуляцию экспрессии гена *PTGS2*, кодирующего фермент циклооксигеназу-2. Полученные данные позволяют рассматривать циклооксигеназу-2 как один из центральных регуляторов, опосредующих влияние стеариновой кислоты на экспрессию генов апоптоза. В работе предложен рациональный биоинформатический подход к поиску новых штаммов, обладающих пробиотическим потенциалом, на основе оценки действия продуцируемых ими метаболитов на целевые биологические процессы в клетках человека через реконструкцию генных сетей.

Ключевые слова: индустриальная микробиология; функциональные микроорганизмы; пробиотики; штаммы-продуценты; метаболиты; генные сети; ANDSystem

Introduction

An important direction in industrial microbiology is the development of probiotic strains with valuable consumer properties. The probiotic industry is one of the most dynamically developing segments of the food sector with a large global market. Probiotic microorganisms are widely used in the production of fermented foods, dietary supplements, and specialized foods; industrial strains are required not only to be safe and technologically suitable, but also to provide pronounced functional effects (immunomodulatory, anti-inflammatory, metabolically mediated, etc.) (Terpou et al., 2019; Lau, Quek, 2024; Grujović et al., 2025).

Streptococcus thermophilus is a streptococcal species used in industrial biotechnology for food fermentation, particularly in yogurt and cheese production. It affects the rate of acidification of dairy products, their texture, and sensory properties, and also exhibits a number of probiotic effects including antioxidant activity, modulation of the gut microbiota, inhibition of certain pathogens, and others, which makes it attractive for industrial use (Cui et al., 2016). In a recent study, strains

of *S. thermophilus* isolated from homemade and commercial fermented dairy product dahi were analyzed using genomics and gas chromatography – mass spectrometry profiling of cell-free supernatants. It was shown that a long-chain fatty acid – octadecanoic (stearic) acid (C18:0) – is among the major components of the supernatant (Sudheer et al., 2025).

This observation suggests that some functional effects of *S. thermophilus* may be mediated by octadecanoic acid as one of its principal metabolites synthesized by this microorganism.

Like other long-chain fatty acids, stearic acid can enter human cells via multiple mechanisms, including protein-mediated uptake and passive diffusion across membranes. Key proteins involved in uptake include the CD36 translocase (SR-B2), a master regulator of cellular fatty-acid homeostasis (Chen et al., 2022; Glatz et al., 2022), and members of the SLC27/FATP transporter family, which import fatty acids coupled to their acylation by long-chain acyl-CoA synthetases (Mashek et al., 2007; Anderson, Stahl, 2013). *In vivo*, stearic acid delivery to the membrane is ensured by its reversible binding to albumin (Kamp, Hamilton, 1992; Richieri et al., 1993; Richieri, Kleinfeld, 1995).

In human cells, stearic acid not only serves as a substrate for the synthesis of triglycerides and other complex lipids (Paton, Ntambi, 2009; Minville-Walz et al., 2010; Houten et al., 2016), but also – as demonstrated in cell-based and *in vivo* models – acts as a modulator of signalling and stress responses associated with apoptosis, tumor cell proliferation, leukotoxicity, as well as pro-inflammatory responses of macrophages and microglia (Evans et al., 2009; Yang et al., 2020; Hung et al., 2023). A review (Shen X. et al., 2025) discusses in considerable detail the functions of this vital molecule, including its role in such pathological processes as cardiovascular diseases, diabetes development, and liver damage. According to current evidence, stearic acid can influence cellular function by interacting with the CD36 receptor on the plasma membrane, followed by modulation of intracellular signalling pathways associated with this receptor (Chen et al., 2022; Glatz et al., 2022). Moreover, stearic acid can exert regulatory effects on the expression of several genes – in particular, through modulation of microRNA activity (Shen X. et al., 2025). This may be characterised as an intracellular mode of action of octadecanoic acid. However, the molecular-genetic mechanisms by which stearic acid – as a probiotic – affects the human body remain insufficiently understood. One important aspect of its influence is the modulation of programmed cell death levels (Yang et al., 2020), which constitutes a factor determining anti-inflammatory and potentially anti-tumor effects.

Among the possible approaches to investigate the mechanisms of the potential influence of metabolites produced by microorganisms on the functioning of human cells, reconstruction and analysis of gene networks can be employed. A gene network is a group of coordinately functioning genes that control the phenotypic traits of an organism (Kolchanov et al., 2013). Interactions between genes in a gene network occur via their primary and secondary products – RNA, proteins, and metabolites. Reconstruction of gene networks allows identifying specific molecular pathways in human cells whose functioning is altered under the influence of various factors, including metabolites. It also enables prediction of the molecular-genetic targets of their action and their impact on disease prevention or development (Saik et al., 2019; Bragina et al., 2023; Ivanisenko V. et al., 2024).

In the present study, using the previously developed information–software system ANDSystem – which employs machine learning and artificial intelligence methods (Ivanisenko V. et al., 2015, 2019) – we reconstructed gene networks regulating the process of apoptosis in human cells under the influence of stearic (octadecanoic) acid present in the cell-free supernatant of *S. thermophilus*. Analysis of these gene networks revealed that stearic acid controls apoptosis primarily through the regulation of the expression of the *PTGS2* gene, which encodes the enzyme cyclooxygenase-2. An additional analysis of differential gene expression in HepG2 cell culture (Vendel Nielsen et al., 2013) under exposure to stearic acid showed that the differentially expressed genes were incorporated into the reconstructed gene networks.

The obtained results establish a rational bioinformatic approach to assessing the functional effects of microbial metabolites on target biological processes in human cells through gene network reconstruction. This approach can be applied in the search for new strains with probiotic potential.

Materials and methods

Reconstruction and analysis of gene networks regulating human gene expression under the influence of stearic acid were carried out using the information-software system ANDSystem (Ivanisenko V. et al., 2019, 2024; Ivanisenko T. et al., 2020). ANDSystem is designed for the reconstruction of associative gene networks, and automatically extracts knowledge and facts from scientific publications and biological databases (Ivanisenko V. et al., 2015, 2019). One of the key modules of ANDSystem is a continuously updated knowledge base. At present, this knowledge base contains information on more than 150 million interactions between 12 different types of molecular-genetic entities (genes, RNA, proteins, metabolites, pharmaceutical compounds, etc.). This information has been automatically extracted from the texts of over 30 million scientific publications and patents, as well as from factual databases. The knowledge base encompasses 49 types of interactions between entities, including regulation of gene expression, physical interactions (protein–protein, protein–ligand), chemical interactions (catalytic reactions, post-translational modifications), and other interaction types (Ivanisenko T. et al., 2022). The effectiveness of ANDSystem for studying the molecular mechanisms of diseases, the effects of drugs and metabolites – including through the analysis of omics data (such as metabolomic and transcriptomic datasets) – has been demonstrated in numerous studies. For instance, it has been used for prioritization of apoptosis-related genes in lymphedema (Saik et al., 2019), identification of apoptosis genes as a basis for the comorbidity of Huntington's disease and cancer

(Bragina et al., 2023), and detection of metabolic markers of postoperative delirium (Ivanisenko V. et al., 2024).

Analysis of differential gene expression and functional annotation. To study the effect of stearic acid on gene expression, we analysed available transcriptomic data obtained using DNA microarray technology in an experimental study (Vendel Nielsen et al., 2013) that investigated the impact of fatty acids (elaidic, oleic, and stearic) on the metabolism of HepG2 cell culture (study identifier in the Gene Expression Omnibus (GEO) database: GSE34045). To date, this is the only experimental study that includes an analysis of the transcriptomic profile of cells exposed to stearic acid. Identification of differentially expressed genes (DEGs) was performed using the Limma package (Ritchie et al., 2015). The statistical significance of differences in gene expression levels between the control and stearic acid-treated samples was assessed using the false discovery rate (FDR) method. Differences with FDR values of less than 0.05 were considered statistically significant.

Enrichment analysis of biological processes for the list of DEGs identified as described above was carried out using the DAVID web server, version 2021 (https://david.ncifcrf.gov/; Sherman et al., 2022) with default settings. DAVID evaluates the statistical significance of the overlap between the list of genes under study (in our case, DEGs) and gene lists corresponding to biological processes described in the Gene Ontology (GO) (Sherman et al., 2022).

Results and discussion

Reconstruction of regulatory gene networks of apoptosis controlled by octadecanoic acid

As noted above, stearic acid can affect cell functioning both through the CD36 receptor on the plasma membrane (Chen et al., 2022; Glatz et al., 2022) and by penetrating into the cell and influencing gene expression inside it (Shen X. et al., 2025). In this context, based on the information contained in the knowledge base of the ANDSystem information-software platform, two gene networks regulating apoptosis induced by octadecanoic acid were reconstructed: an intracellular pathway of stearic acid action that does not involve the CD36 receptor protein, according to the ANDSystem knowledge base (gene network GN-VPDO), and a CD36-mediated pathway (gene network GN-CD36).

First, a list of human protein-coding genes involved in the apoptosis process was compiled – 748 genes obtained from the Gene Ontology database (https://geneontology.org/) for the term GO:0006915 (apoptotic process) (Supplementary Table S1)[1]. This list was used as input data in ANDSystem to reconstruct gene networks.

The gene network GN-VPDO, which represents the regulation of apoptosis-related gene expression by stearic (octadecanoic) acid entering the cell, is shown in Figure 1.

Analysis of the GN-VPDO gene network revealed that it included 33 apoptosis-related genes and an additional 11 genes involved in regulating their expression. The network also comprised 48 regulatory interactions between apoptosis genes and regulatory proteins, as well as 2 regulatory interactions between stearic (octadecanoic) acid and apoptosis genes (indicated by turquoise boxes in Figure 1).

Cyclooxygenase-2 protein had the highest number of regulatory connections (25) with apoptosis genes in the GN-VPDO network (Table 1). Additionally, PROX1 and CD276 proteins can be considered key regulators of apoptosis genes in this network, each regulating five genes (Table 1).

Several examples of regulatory pathways through which octadecanoic acid affects apoptosis gene expression in the GN-VPDO network are presented in Figure 2.

Gene *BIRC2*. Figure 2 shows that *BIRC2* gene expression is under the regulatory influence of stearic acid. Suppression of this gene's expression – which has anti-apoptotic activity – by stearic acid in preadipocytes represents one mechanism for reducing adipose tissue under the influence of this metabolite (Shen M.C. et al., 2014).

Gene *SCD*. According to ANDSystem data, stearic acid regulates the expression of the *SCD* gene, which encodes the SCD protein. This protein, in turn, regulates the *DDIT3* gene, which is directly involved in apoptosis (Fig. 2). According to published literature (Aardema et al., 2017), stearic acid activates *SCD* gene expression. The SCD protein suppresses the *DDIT3* gene (Minville-Walz et al., 2010), which is one of the key inducers of apoptosis (Bento et al., 2009).

Gene *BCL2*. The expression of the *PTGS2* gene, which encodes the cyclooxygenase-2 protein (COX-2), is under the regulatory influence of stearic (octadecanoic) acid (Fig. 2). According to published literature (Liu J. et al., 2014), stearic acid induces the expression of the *PTGS2* gene encoding cyclooxygenase-2. In turn, cyclooxygenase-2 – as indicated by ANDSystem – enhances the expression of the *BCL2* gene, a finding that is supported by experimental studies (Lin et al., 2019). It is worth noting that, according to both the ANDSystem knowledge base and published literature (Shen M.C. et al., 2014), the expression of the *BCL2* gene can also be suppressed by stearic acid. Thus, according to gene network reconstruction data, the expression of the *BCL2* gene – which encodes a key inhibitor of apoptosis (Newton et al., 2024) – can be both activated by stearic acid via cyclooxygenase-2 and suppressed by this metabolite, likely without the involvement of the enzyme. Therefore, modulation of either the expression or the activity of cyclooxygenase-2 may alter the balance between pro-apoptotic and anti-apoptotic effects of stearic acid. Available experimental evidence indicates that cyclooxygenase-2 inhibits apoptosis. Moreover, elevated expression of the cyclooxygenase-2 gene is one of the mechanisms by which tumour cells evade apoptosis (Liu C.H. et al., 2001).

Genes *NFKB1*, *NLRP3*, *CASP1*, *MCL1*. According to the ANDSystem data (Fig. 2), the expression of *NFKB1* and *MCL1* genes – which encode the anti-apoptotic proteins NFKB1 and MCL1 (Newton et al., 2024) – as well as the *NLRP3* and *CASP1* genes encoding pro-apoptotic proteins (Yu et al., 2023), is regulated by stearic (octadecanoic) acid with

[1] Supplementary Tables S1–S4 and Figure S1 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Ivanis_Engl_30_1.zip

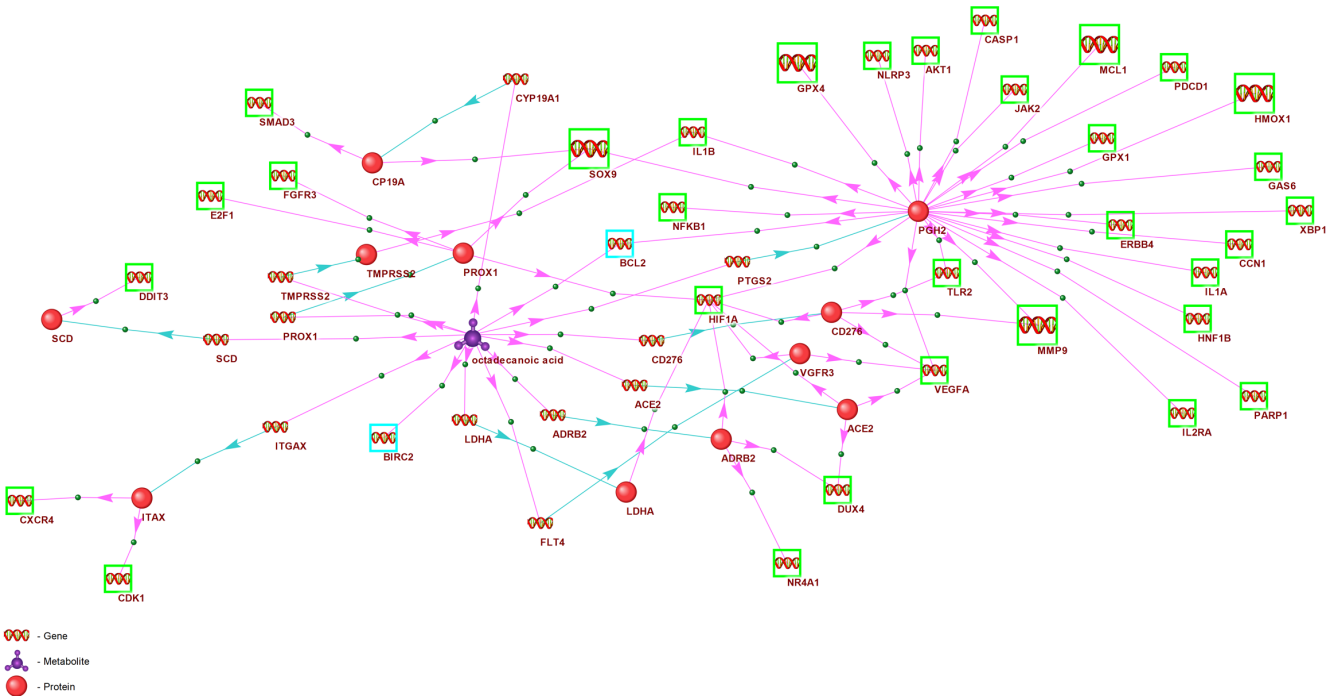


Fig. 1. The GN-VPDO gene network representing the regulation of apoptosis-related gene expression by stearic (octadecanoic) acid.

Green boxes indicate genes involved in the apoptosis process according to the Gene Ontology (GO:0006915, apoptotic process). Turquoise boxes mark apoptosis genes that are directly regulated by octadecanoic acid without intermediaries. Genes without boxes are intermediary genes: their expression is regulated by octadecanoic acid, and their protein products (red spheres in the figure) regulate the expression of apoptosis genes. Large boxes highlight genes that are differentially expressed in the HepG2 cell culture in response to octadecanoic acid (Vendel Nielsen et al., 2013). Stearic (octadecanoic) acid is labelled as "octadecanoic acid". Blue arrows represent gene expression processes leading to the production of their encoded proteins; red arrows indicate the regulatory influence of proteins on the expression of apoptosis genes.

Table 1. List of the five regulatory proteins with the highest number of regulatory connections to apoptosis-related genes in the GN-VPDO gene network

No.	Protein symbol	Protein name	Number of associated apoptosis genes
1	PTSG2	Cyclooxygenase-2	25
2	PROX1	Prospero homeobox 1	5
3	CD276	Protein CD276	5
4	ADRB2	Adrenoreceptor beta 2	4
5	ACE2	Angiotensin converting enzyme 2	4

the involvement of cyclooxygenase-2 (mentioned above). It has been demonstrated that cyclooxygenase-2 activates the expression of the anti-apoptotic genes *NFKB1*, *MCL1*, and *BCL2* (Lin et al., 2019), which may account for the suppression of apoptosis by this enzyme.

The GN-CD36 gene network of apoptosis regulation, reconstructed in our study, comprises 187 molecular-genetic entities, including 98 genes involved in the apoptosis process, the CD36 receptor protein itself, and 44 genes encoding proteins that exert regulatory effects on the expression of apoptosis-related genes. Due to the extensive number of regulatory interactions within the complete GN-CD36 gene network, it is presented in the Supplementary Materials (Fig. S1).

When analysing the complete GN-CD36 gene network (Fig. S1), several key regulatory proteins of apoptosis were identified: PPARG, which regulates 32apoptosis-related genes; cyclooxygenase-2, which regulates 25 genes; and SIRT1, which regulates 26 genes (Table 2).

Figure 3 illustrates several examples of regulatory pathways through which the CD36 receptor influences the expression of apoptosis-related genes, as described below.

Gene *NFKB1*. Two regulatory pathways for the *NFKB1* gene involved in apoptosis were identified. In the first pathway, CD36 regulates the expression of the intermediary gene *SIRT1*, which encodes the SIRT1 protein. SIRT1 then exerts a regulatory effect on the *NFKB1* gene. According to (Yin et

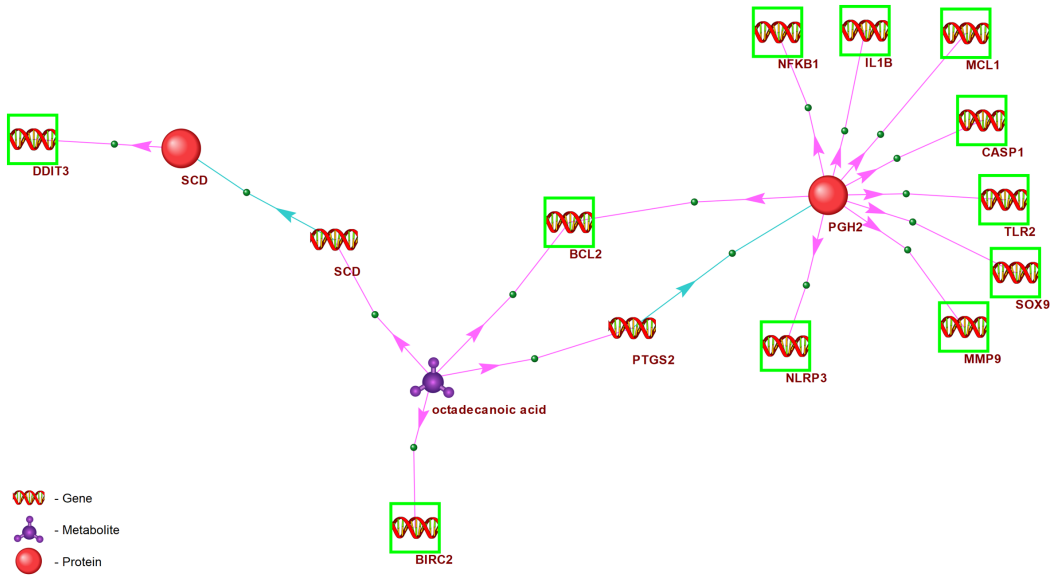


Fig. 2. Fragments of the gene network showing regulatory interactions between stearic (octadecanoic) acid and the apoptosis-related genes *BIRC2*, *SCD*, *BCL2*, *NFKB1*, *NLRP3*, *CASP1*, and *MCL1*.

Red spheres represent proteins encoded by intermediary genes. Green boxes indicate apoptosis genes. Genes without boxes are intermediary genes: their expression is regulated by octadecanoic acid, and their protein products (red spheres in the figure) regulate the expression of apoptosis genes. Stearic (octadecanoic) acid is labelled as "octadecanoic acid". Blue arrows denote gene expression processes leading to the production of their encoded proteins; red arrows illustrate the regulatory influence of proteins on the expression of apoptosis genes.

Table 2. List of the five regulatory proteins with the highest number of regulatory connections to apoptosis-related genes in the GN-CD36 gene network

No.	Protein symbol	Protein name	Number of regulated apoptosis-related genes
1	PPARG	Peroxisome proliferator-activated receptor gamma	32
2	PTGS2	Cyclooxygenase-2	25
3	SIR1	Sirtuin 1	26
4	PPARA	Peroxisome proliferator-activated receptor alpha	19
5	GLI2	GLI family zinc finger 2	19

al., 2020), the SIRT1 protein has an activating effect on the NF-κB regulatory pathway. The second pathway involves the CD36 regulatory effect on the intermediary gene *TRAF6* and its protein product, the TRAF6 transcription factor, which in turn regulates *NFKB1* expression. The TRAF6 protein – whose gene expression is stimulated via CD36 (Cao et al., 2019) – also activates *NFKB1* expression (Saba et al., 2014).

Gene *MYD88*. The CD36 receptor regulates the expression of the *MYD88* gene, which is involved in apoptosis (a short regulatory pathway). Additionally, *MYD88* gene expression is regulated by the CD36 receptor through a longer pathway that involves the intermediary gene *TRAF6* and its encoded protein, which exert a regulatory effect on *MYD88* expression.

Gene *MMP9*. As noted above, the CD36 receptor protein regulates the expression of the *TRAF6* gene, which encodes the TRAF6 protein. In turn, TRAF6 regulates the expression of the metalloproteinase gene *MMP9* (Luo et al., 2016), as well as five other apoptosis-related genes: *IFIT2*, *HIF1A*, *IRF3*, *MYD88*, and *NFKB1*. Thus, the TRAF6 transcription factor acts as a cassette activator of six apoptosis genes.

A comparison of the GN-VPDO and GN-CD36 gene networks revealed that 30 genes involved in the apoptosis process were shared between these networks. Among them, the expression of 25 genes was regulated by cyclooxygenase-2 (Table S2). Consequently, this protein can be regarded as the most critical regulator of apoptosis genes in each of the two gene networks we reconstructed. The list of apoptosis-related genes common to both the GN-VPDO and GN-CD36 gene networks and regulated by cyclooxygenase-2 includes, among others, *BCL2*, *MCL1*, and *NFKB1*. The protein products of

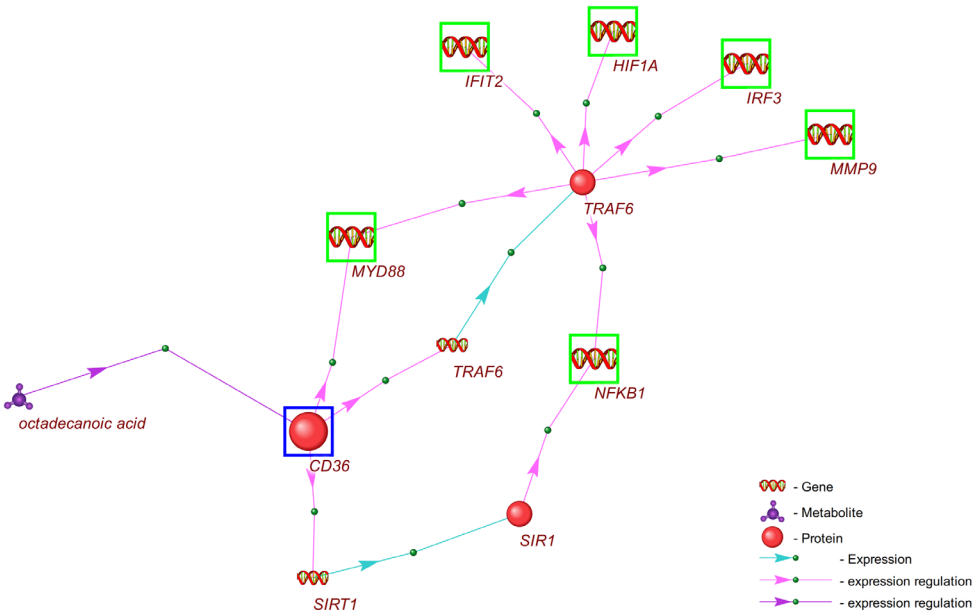


Fig. 3. Fragments of the gene network showing regulatory interactions between the CD36 receptor and the apoptosis-related genes *IRF3*, *IFIT2*, *MMP9*, *HIF1A*, and *NFKB1*.

The CD36 receptor protein is highlighted with a blue box; apoptosis genes are indicated by green boxes. Genes without boxes are intermediary genes: their expression is regulated by octadecanoic acid, and their protein products (red spheres in the figure) regulate the expression of apoptosis genes. Stearic (octadecanoic) acid is labelled as "octadecanoic acid". Blue arrows denote gene expression processes leading to the production of their encoded proteins; red arrows illustrate the regulatory influence of proteins on the expression of apoptosis genes. A purple arrow shows the interaction between octadecanoic acid and the CD36 receptor.

these genes are known to be key anti-apoptotic proteins (Gupta et al., 2023; Newton et al., 2024). On the other hand, the list of genes regulated by cyclooxygenase-2 in these gene networks also includes *NLRP3* and *CASP1*, whose protein products are involved in activating programmed cell death processes, including apoptosis (Yu et al., 2023).

Differential gene expression under exposure to octadecanoic acid

To investigate the potential of stearic acid produced by the bacterium *S. thermophilus* in regulating apoptosis, we analysed experimental data on differential gene expression in hepatocyte-like HepG2 cells treated with stearic acid, as reported by L. Vendel Nielsen et al. (2013). We identified 2,500 differentially expressed genes (DEGs) with a statistical significance level of FDR < 0.05. A Gene Ontology (GO) enrichment analysis of this gene set revealed 40 statistically significant biological processes (FDR < 0.05) (Table S3), including cellular respiration, processes related to mitochondrial electron transport chain function, and apoptosis. The ten most statistically overrepresented biological processes for DEGs in HepG2 cells upon stearic acid treatment are presented in Table 3.

A comparison of the reconstructed gene network for apoptosis-related gene expression regulation involving the CD36 receptor with differential expression data showed that among the apoptosis genes regulated by stearic acid, 16 genes (Table S4) exhibited statistically significant changes in expression in HepG2 cell culture upon exposure to the acid. These included genes such as *MMP9*, *SOX9*, *HMOX1*, *GPX4*, *MCL1*, and *BIRC1*, which play important roles in the apoptosis process. Regarding the GN-VPDO gene network, the number of apoptosis genes whose expression changed in response to octadecanoic acid was substantially lower: only five genes (*GPX4*, *MCL1*, *HMOX1*, *MMP9*, and *SOX9*). These same genes were also present in the GN-CD36 gene network. Thus, the analysis of differential gene expression indicates that the regulatory pathway via the CD36 receptor is more significant for stearic acid-mediated regulation of apoptosis.

Conclusion

In this study, we used the gene network reconstruction method to investigate the regulatory mechanisms by which stearic (octadecanoic) acid – one of the main metabolites in the cell-free supernatant of the bacterium *S. thermophilus* – affects apoptosis in human cells.

We demonstrated that the GN-CD36 gene network described the regulation of 98 apoptosis-related genes, which significantly exceeded the number of apoptosis genes regulated in the GN-VPDO gene network (33 genes). Additionally, cyclooxygenase-2 acted as the primary regulatory protein governing apoptosis gene expression under the influence of stearic acid, both in the GN-CD36 and GN-VPDO gene networks.

These findings provide a foundation for developing an experimental-computational platform to identify functional metabolites produced by probiotic microorganisms that exert

Table 3. Results of Gene Ontology (GO) enrichment analysis for differentially expressed genes in HepG2 cell culture upon exposure to octadecanoic acid

No.	Gene Ontology term	Number of genes in the process	FDR
1	Proton motive force-driven mitochondrial ATP synthesis	28	9.14E-07
2	Aerobic respiration	28	3.34E-06
3	Mitochondrial respiratory chain complex I assembly	28	3.34E-06
4	Mitochondrial electron transport, NADH to ubiquinone	20	1.94E-04
5	G2/M transition of mitotic cell cycle	19	3.73E-03
6	Release of cytochrome c from mitochondria	12	1.62E-02
7	Cellular oxidant detoxification	23	7.98E-03
8	Fatty acid beta-oxidation	17	1.71E-02
9	Autophagy	36	3.92E-02
10	Apoptotic process	108	3.96E-02

Note. FDR (false discovery rate) is a measure of statistical significance of biological process overrepresentation accounting for multiple comparisons.

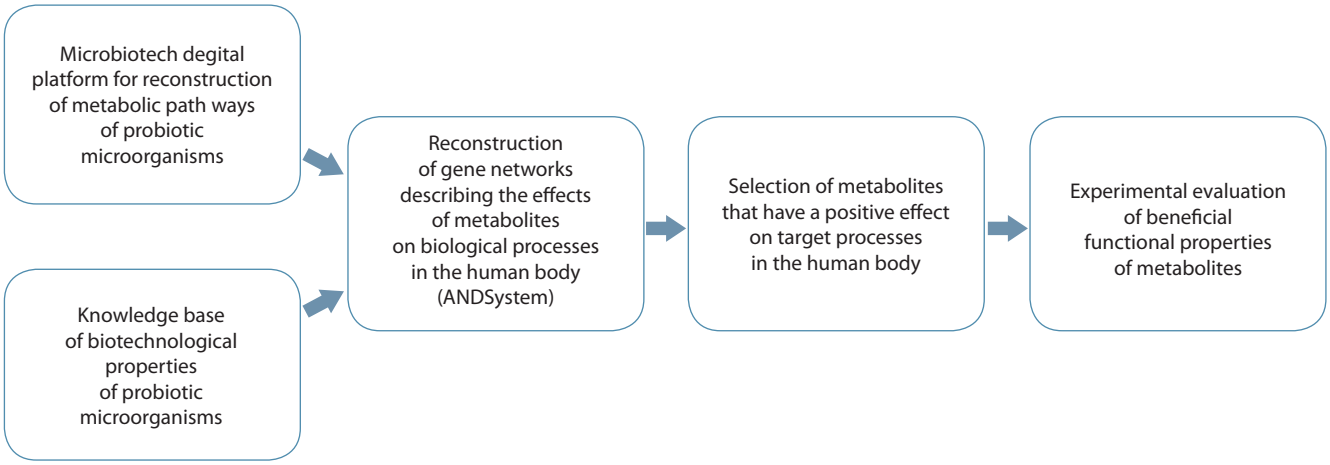


Fig. 4. General scheme of the platform for identifying functional metabolites produced by probiotic microorganisms that exert beneficial effects on target processes in the human body.

beneficial effects on target processes in the human body. The key components of this platform are presented in Figure 4.

The platform receives initial data from two sources. The first is the digital platform "Microbiotech", developed within the framework of the Kurchatov Genomic Center at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (Demenkov et al., 2024). This platform enables the reconstruction of microbial metabolic pathways through computational analysis of their genomes. The second source is the Knowledge Base of Biotechnological Properties of Microorganisms, which is being developed within the same Kurchatov Genomic Center using artificial intelligence methods. These methods allow for automatic extraction of knowledge and facts from scientific publications and patents.

The compiled data are used to reconstruct gene networks that describe the effects of metabolites on biological processes in the human body, employing ANDSystem (Ivanisenko V. et al., 2015, 2019). Analysis of the reconstructed gene networks helps to define selection criteria for metabolites that exert beneficial effects on target processes in the human body. In the final stage, an experimental assessment of the useful functional properties of the selected metabolites is carried out.

References

Aardema H., van Tol H.T.A., Wubbolts R.W., Brouwers J.F.H.M., Gadella B.M., Roelen B.A.J. Stearoyl-CoA desaturase activity in bovine cumulus cells protects the oocyte against saturated fatty acid stress. *Biol Reprod*. 2017;96(5):982-992. doi 10.1095/biolreprod.116.146159

Anderson C.M., Stahl A. SLC27 fatty acid transport proteins. *Mol Aspects Med.* 2013;34(2-3):516-528. doi 10.1016/j.mam.2012.07.010

Bento C., Andersson M.K., Aman P. DDIT3/CHOP and the sarcoma fusion oncoprotein FUS-DDIT3/TLS-CHOP bind cyclin-dependent kinase 2. *BMC Cell Biol*. 2009;10:89. doi 10.1186/1471-2121-10-89

Bragina E.Y., Gomboeva D.E., Saik O.V., Ivanisenko V.A., Freidin M.B., Nazarenko M.S., Puzyrev V.P. Apoptosis genes as a key to identification of inverse comorbidity of Huntington's disease and cancer. *Int J Mol Sci*. 2023;24(11):9385. doi 10.3390/ijms24119385

Cao D., Luo J., Chen D., Xu H., Shi H., Jing X., Zang W. CD36 regulates lipopolysaccharide-induced signaling pathways and mediates the internalization of *Escherichia coli* in cooperation with TLR4 in goat mammary gland epithelial cells. *Sci Rep*. 2016;6:23132. doi 10.1038/srep23132. Erratum in: *Sci Rep*. 2019;9(1):6457. doi 10.1038/s41598-019-42156-3

Chen Y., Zhang J., Cui W., Silverstein R.L. CD36, a signaling receptor and fatty acid transporter that regulates immune cell metabolism and fate. *J Exp Med.* 2022;219(6):e20211314. doi 10.1084/jem.20211314

Cui Y., Xu T., Qu X., Hu T., Jiang X., Zhao C. New insights into various production characteristics of *Streptococcus thermophilus* strains. *Int J Mol Sci*. 2016;17(10):1701. doi 10.3390/ijms17101701

Demenkov P.S., Mukhin A.M., Ivanisenko V.A., Lashin S.A., Kolchanov N.A. The "Microbiotech" digital platform: architecture and purpose. *Problems of Informatics*. 2024;4:27-36. doi 10.24412/2073-0667-2024-4-27-36 (in Russian)

Evans L.M., Cowey S.L., Siegal G.P., Hardy R.W. Stearate preferentially induces apoptosis in human breast cancer cells. *Nutr Cancer.* 2009;61(5):746-753. doi 10.1080/01635580902825597

Glatz J.F.C., Nabben M., Luiken J.J.F.P. CD36 (SR-B2) as master regulator of cellular fatty acid homeostasis. *Curr Opin Lipidol.* 2022;33(2):103-111. doi 10.1097/MOL.0000000000000819

Grujović M.Ž., Semedo-Lemsaddek T., Marković K.G. Application of probiotics in foods: a comprehensive review of benefits, challenges, and future perspectives. *Foods*. 2025;14(17):3088. doi 10.3390/foods14173088

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

Houten S.M., Violante S., Ventura F.V., Wanders R.J.A. The biochemistry and physiology of mitochondrial fatty acid β-oxidation and its genetic disorders. *Annu Rev Physiol.* 2016;78:23-44. doi 10.1146/annurev-physiol-021115-105045

Hung T.-Y., Chen H.-H., Wu C.-Y., Hsu H.-T., Hsueh Y.-P., Kuan Y.-H. Stearic acid activates microglia and impairs memory via NF-κB signaling. *Neurobiol Dis.* 2023;182:106121. doi 10.1016/j.nbd.2023.106121

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

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. New ANDDigest with improved AI-based short-name recognition. *Int J Mol Sci.* 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Likhoshvai V.A., Kolchanov N.A. ANDSystem: an Associative Network Discovery System. *BMC Syst Biol.* 2015;9(Suppl.2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Saik O.V., Kolchanov N.A. A new version of ANDSystem with tissue-specific functionality. *BMC Bioinformatics.* 2019;20(Suppl.1):34. doi 10.1186/s12859-018-2567-6

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

Kamp F., Hamilton J.A. pH gradients across phospholipid membranes caused by fast flip-flop of un-ionized fatty acids. *Proc Natl Acad Sci USA.* 1992;89(23):11367-11370. doi 10.1073/pnas.89.23.11367

Kolchanov N.A., Ignatyeva I.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin V.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2013;17(4/2):833-850 (in Russian)

Lau L.Y.J., Quek S.Y. Probiotics: health benefits, food application, and colonization in the human gastrointestinal tract. *Food Bioeng*. 2024;3(1):41-64. doi 10.1002/fbe2.12078

Lin X.M., Luo W., Wang H., Li R.Z., Huang Y.S., Chen L.K., Wu X.P. The role of prostaglandin-endoperoxide synthase-2 in chemoresistance of non-small cell lung cancer. *Front Pharmacol*. 2019;10:836. doi 10.3389/fphar.2019.00836

Liu C.H., Chang S.H., Narko K., Trifan O.C., Wu M.T., Smith E., Haudenschild C., Lane T.F., Hla T. Overexpression of cyclooxygenase-2 is sufficient to induce tumorigenesis in transgenic mice. *J Biol Chem*. 2001;276(21):18563-18569. doi 10.1074/jbc.M010787200

Liu J., Hu S., Cui Y., Sun M.K., Xie F., Zhang Q., Jin J. Saturated fatty acids up-regulate COX-2 expression in prostate epithelial cells via toll-like receptor 4/NF-κB signaling. *Inflammation*. 2014;37(2):467-477. doi 10.1007/s10753-013-9760-6

Luo Z., Zhang X., Zeng W., Su J., Yang K., Lu L., Lim C.B., Tang W., Wu L., Zhao S., Jia X., Peng C., Chen X. TRAF6 regulates melanoma invasion and metastasis through ubiquitination of Basigin. *Oncotarget.* 2016;7(6):7179-7192. doi 10.18632/oncotarget.6886

Mashek D.G., Li L.O., Coleman R.A. Long-chain acyl-CoA synthetases and fatty acid channeling. *Future Lipidol.* 2007;2(4):465-476. doi 10.2217/17460875.2.4.465

Minville-Walz M., Pierre A.S., Pichon L., Bellenger S., Fèvre C., Bellenger J., Tessier C., Narce M., Rialland M. Inhibition of stearoyl-CoA desaturase 1 expression induces CHOP-dependent cell death in human cancer cells. *PLoS One*. 2010;5(12):e14363. doi 10.1371/journal.pone.0014363

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

Paton C.M., Ntambi J.M. Biochemical and physiological function of stearoyl-CoA desaturase. *Am J Physiol Endocrinol Metab.* 2009;297(1):E28-E37. doi 10.1152/ajpendo.90897.2008

Richieri G.V., Kleinfeld A.M. Unbound free fatty acid levels in human serum. *J Lipid Res.* 1995;36(2):229-240. doi 10.1016/S0022-2275(20)39899-0

Richieri G.V., Anel A., Kleinfeld A.M. Interactions of long-chain fatty acids and albumin: determination of free fatty acid levels using the fluorescent probe ADIFAB. *Biochemistry.* 1993;32(29):7574-7580. doi 10.1021/bi00080a032

Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., Smyth G.K. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi 10.1093/nar/gkv007

Saba R., Sorensen D.L., Booth S.A. MicroRNA-146a: a dominant, negative regulator of the innate immune response. *Front Immunol*. 2014;5:578. doi 10.3389/fimmu.2014.00578

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

Shen M.C., Zhao X., Siegal G.P., Desmond R., Hardy R.W. Dietary stearic acid leads to a reduction of visceral adipose tissue in athy-

mic nude mice. *PLoS One*. 2014;9(9):e104083. doi 10.1371/journal.pone.0104083

Shen X., Miao S., Zhang Y., Guo X., Li W., Mao X., Zhang Q. Stearic acid metabolism in human health and disease. *Clin Nutr*. 2025; 44:222-238. doi 10.1016/j.clnu.2024.12.012

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

Sudheer A., Dastidar D.G., Ghosh G., Taj Z., Nidhin I.K., Chattopadhyay I. Comprehensive genomics, probiotic, and antibiofilm potential analysis of *Streptococcus thermophilus* strains isolated from homemade and commercial dahi. *Sci Rep*. 2025;15(1):7089. doi 10.1038/s41598-025-90999-w

Terpou A., Papadaki A., Lappa I.K., Kachrimanidou V., Bosnea L.A., Kopsahelis N. Probiotics in food systems: significance and emerging strategies towards improved viability and delivery of enhanced beneficial value. *Nutrients*. 2019;11(7):1591. doi 10.3390/nu11071591

Vendel Nielsen L., Krogager T.P., Young C., Ferreri C., Chatgilialoglu C., Jensen O.N., Enghild J.J. Effects of elaidic acid on lipid metabolism in HepG2; dataset GSE34045. *PLoS One.* 2013;8(9): e74283. doi 10.1371/journal.pone.0074283

Yang Y., Huang J., Li J., Yang H., Yin Y. Effects of stearic acid on proliferation, differentiation, apoptosis, and autophagy in porcine intestinal epithelial cells. *Curr Mol Med*. 2020;20(2):157-166. doi 10.2174/1566524019666190917144127

Yin Q., Shen L., Qi Y., Song D., Ye L., Peng Y., Wang Y., Jin Z., Ning G., Wang W., Lin D., Wang S. Decreased SIRT1 expression in the peripheral blood of patients with Graves' disease. *J Endocrinol*. 2020;246(2):161-173. doi 10.1530/JOE-19-0501

Yu C., Chen P., Miao L., Di G. The role of the NLRP3 inflammasome and programmed cell death in acute liver injury. *Int J Mol Sci*. 2023;24(4):3067. doi 10.3390/ijms24043067

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

Received November 15, 2025. Revised November 24, 2025. Accepted December 8, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):15-26

doi 10.18699/vjgb-26-10

Improvement of a phylogenetic footprinting method for transcription factor binding sites recognition based on the use of bootstrap trials for the analysis of large bacterial genomic data

A.M. Mukhin [1, 2, 3], T.M. Khlebodarova[1, 2], D.Yu. Oshchepkov [1, 2]

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

Abstract. The rapid development of high-throughput sequencing technologies has led to an explosive accumulation of high-quality bacterial genome sequence data – their number is approaching three million, and this growth continues. This, in turn, provides additional impetus for the development of technologies for more efficient annotation using analytical methods designed to utilize such large-scale genomic data, as well as for achieving new levels of annotation quality. One such analytical approach is phylogenetic footprinting, which aims to identify motifs corresponding to transcription factor binding sites in the promoter regions of bacterial genomes by comparing corresponding sets of regulatory sequences of orthologous genes in related organisms. The continued accumulation of genomic data has served as the basis for further development of this approach. It has been found that an excessive number of sequences in a set analyzed using phylogenetic footprinting only reduces the accuracy of the method, whereas the inclusion of a sequence selection step in the analyzed set based on data on mutual evolutionary distances improves the method's performance. In this paper, we propose and implement a further step in the development of the phylogenetic footprinting method. This step involves multiple runs of the selection step described above to generate distinct subsamples, subsequent pipeline runs for each subsample, and statistical analysis of the results obtained from multiple pipeline runs. The proposed approach, implemented in the MotifsOnFly method, improves the robustness of motif recognition results obtained from multiple pipeline runs. The effectiveness of the MotifsOnFly method is demonstrated using the analysis of the well-annotated promoter of the *Escherichia coli OmpW* gene.

Key words: phylogenetic footprinting; bacterial genome; transcription factor binding sites; motifs; bootstrap; Python

For citation: Mukhin A.M., Khlebodarova T.M., Oshchepkov D.Yu. Improvement of a phylogenetic footprinting method for transcription factor binding sites recognition based on the use of bootstrap trials for the analysis of large bacterial genomic data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):15-26. doi 10.18699/vjgb-26-10

Funding. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019–2030, Agreement No. 075-15-2025-516).

Acknowledgements. This research was supported in part by computational resources of HPC facilities at the multi-access center "Bioinformatics" of the ICG SB RAS. During the preparation of this manuscript, the authors used Qwen3-Max (https://chat.qwen.ai) for the purposes of improvement of readability of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Развитие метода филогенетического футпринтинга для распознавания сайтов связывания транскрипционных факторов на основе использования bootstrap-испытаний для анализа больших бактериальных геномных данных

А.М. Мухин [1, 2, 3], Т.М. Хлебодарова[1, 2], Д.Ю. Ощепков [1, 2]

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

© Mukhin A.M., Khlebodarova T.M., Oshchepkov D.Yu., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Аннотация. Активное развитие технологий высокопроизводительного секвенирования привело к взрывообразному накоплению высококачественных данных по последовательностям бактериальных геномов – их число приближается к трем миллионам, и дальнейший рост продолжается. Это, в свою очередь, дает дополнительный импульс развитию технологий для более эффективной их аннотации аналитическими методами с прицелом на применение таких больших геномных данных, а также получение нового качества аннотаций. Одним из таких аналитических подходов стал метод филогенетического футпринтинга, направленный на выявление мотивов, соответствующих сайтам связывания транскрипционных факторов в промоторных областях бактериальных геномов путем сравнения соответствующих выборок регуляторных последовательностей генов-ортологов для родственных организмов. Дальнейшее накопление геномных данных стало стимулом для развития подхода. Так, было обнаружено, что избыточное число последовательностей в выборке, анализируемой с использованием филогенетического футпринтинга, лишь ухудшает точность метода, тогда как включение этапа отбора последовательностей в анализируемую выборку с учетом данных о взаимных эволюционных расстояниях повышает качество работы метода. В настоящей статье нами предложен и реализован следующий шаг развития метода филогенетического футпринтинга, основанный на множественном запуске описанного выше этапа отбора для формирования различающихся подвыборок, последующего запуска конвейера для каждой из подвыборок и на статистическом анализе получаемых результатов множественных запусков конвейера. Предложенный подход, реализованный в методе MotifsOnFly, позволяет повысить устойчивость получаемых результатов распознавания мотивов, выявляемых в многократных запусках конвейера. Эффективность метода MotifsOnFly продемонстрирована на примере анализа хорошо аннотированного промотора гена *OmpW Escherichia coli*.
Ключевые слова: филогенетический футпринтинг; бактериальный геном; сайты связывания транскрипционных факторов; мотивы; бутстреп; Python

Introduction

The rapid development and widespread application of high-throughput sequencing technologies in molecular genetics have stimulated advances not only in biotechnology – enabling large-scale assembly of bacterial genomes for analysis, modification, and subsequent utilization of bacterial strains to address biotechnological challenges – but also in bioinformatics methods for increasingly accurate genome annotation. Annotation of bacterial genomes with transcription factor binding sites (TFBSs) represents one of the most critical steps in biotechnology and microbiology, as the binding of transcription factors to their specific sites within gene promoters constitutes a fundamental mechanism regulating gene expression in bacteria (Browning, Busby, 2004).

The phylogenetic footprinting strategy, initially proposed in 1988 (Tagle et al., 1988; Katara et al., 2012) for identifying TFBSs, proved highly productive for *de novo* motif discovery, especially given the growing number of sequenced genomes available at the time. This strategy is based on the general principle that regulatory elements in promoters – such as TFBSs – are, in the vast majority of cases, evolutionarily more conserved and evolve at a slower rate at the DNA sequence level compared to surrounding non-functional sequences (Levy et al., 2001). The subsequent surge in the number of sequenced bacterial genomes further enhanced the effectiveness of phylogenetic footprinting (Blanchette, Tompa, 2002), spurring the development of numerous algorithms optimized for the volume of genomic data then available. These include MotifSuite, FootPrinter, AlignACE, BioProspector, CONSENSUS, MDscan, MEME, CUBIC, and BoBro (Hertz, Stormo, 1999; Liu X. et al., 2001, 2002; Blanchette, Tompa, 2003; Olman et al., 2003; Chen et al., 2008; Bailey et al., 2009; Li et al., 2011a; Claeys et al., 2012).

Later studies revealed that a limited number of properly selected reference promoters could be sufficient for identifying TFBSs within a gene (McCue et al., 2002). This is because closely related sequences with very short evolutionary distances provide little informative value for phylogenetic footprinting due to the insufficient accumulation of mutations in non-functional sequences flanking the TFBSs. Consequently, the continued accumulation of bacterial genomic data enabled refinement of the phylogenetic footprinting method by incorporating a step to pre-select promoter sequences for analysis based on their mutual evolutionary distances. This approach yields more informative sets of orthologous promoters for functional motif recognition. As demonstrated in (Liu B. et al., 2016), this refined methodology outperformed earlier popular motif-finding tools listed above in terms of TFBS prediction accuracy.

Building upon the approach proposed by B. Liu and colleagues (2016), we previously developed a computational pipeline for identifying TFBSs in bacterial genomes (Mukhin et al., 2024). This pipeline integrates a comprehensive suite of necessary databases and algorithms, enabling rapid annotation of selected bacterial genomes with transcription factor binding sites. However, during its application to identify TFBSs in the genome of *Geobacillus icigianus* (Peltek et al., 2024), we observed that the prediction outcome was sensitive to the order, composition, and selection method of the input promoter set. This dependency hindered definitive conclusions regarding the most probable TFBS location and the corresponding transcription factor likely interacting with it.

In the present study, we propose and implement a modification of our previously developed pipeline (Mukhin et al., 2024), which we call MotifsOnFly. This modification incorporates three key components: multiple runs of the pipeline's sub-sampling stage, generating distinct promoter subsets from the full dataset while accounting for pairwise evolutionary distances to ensure diversity among subsets; execution of the full pipeline on each subset to identify overrepresented *de novo* motifs within each; and statistical analysis of the aggregated results. This bootstrap-like approach (with replacement)

enables far more comprehensive utilization of the original data and significantly enhances the robustness and reliability of the results through statistical evaluation of motifs detected across multiple independent runs. Using the well-annotated promoter of the *ompW* gene in *Escherichia coli* as a case study, we demonstrate the advantages of the newly developed MotifsOnFly method, which stems directly from our proposed advancement of the phylogenetic footprinting strategy.

Materials and methods

Computational pipeline. To identify potential transcription factor binding sites (TFBSs) by detecting *de novo* motifs in bacterial genomes, we employed a computational pipeline that leverages operon annotations for 3,850 bacterial genomes available in the DOOR2 database (Mao et al., 2014). The corresponding genomic sequences were retrieved from the NCBI database (Sayers et al., 2021).

The pipeline implements an integrative approach based on the phylogenetic footprinting method originally proposed by B. Liu and colleagues (2016). The implemented pipeline includes the following stages: for a given target gene, orthologous genes are identified across all 3,850 annotated genomes in the database by assessing protein sequence similarity using the GOST software module (Li et al., 2011b). For each identified orthologous gene, its promoter sequence is extracted based on the known operon structure of its host genome, thereby forming a complete set of orthologous promoter sequences for the target gene. Pairwise evolutionary distances between all promoters in this full set are estimated using a phylogenetic tree constructed with ClustalW2 (Larkin et al., 2007). Diverse promoter subsets are then generated by selecting sequences according to their mutual evolutionary distances, following the principles established in B. Liu et al. (2016). For each subset, *de novo* motif discovery is performed using a consensus ("voting") approach that integrates results from multiple established motif-finding algorithms: AlignACE, BioProspector, CONSENSUS, MDscan, MEME, CUBIC, and BoBro (Hertz, Stormo, 1999; Liu X. et al., 2001, 2002; Olman et al., 2003; Chen et al., 2008; Bailey et al., 2009; Li et al., 2011a).

The discovered *de novo* motifs are compared against known TFBSs using the Tomtom algorithm (Gupta et al., 2007) and two reference databases: SwissRegulon (Pachkov et al., 2013), which contains curated TFBS data specifically for *E. coli*, and PRODORIC (Dudek, Jahn, 2022), a comprehensive resource for bacterial regulatory elements. Based on Tomtom's statistical similarity scores, the best-matching known TFBS is identified, thereby predicting the most likely transcription factor (TF) interacting with the motif.

The use of both SwissRegulon and PRODORIC enables dual validation: for *E. coli* promoters, motifs can be matched against organism-specific TFBSs (SwissRegulon) as well as broader bacterial regulatory motifs (PRODORIC). Results from multiple pipeline runs undergo statistical analysis and visualization.

All analytical scripts were implemented in Python versions 3.12, 3.6, and 2.7, with environment and dependency management handled via the Anaconda platform (https://anaconda.com). Genomic data and structured metadata (genes, sequences, operons) were stored and managed using the PostgreSQL database system (https://www.postgresql.org/), deployed on the infrastructure of the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), by the multi-access center "Bioinformatics" ICG SB RAS.

Visualization and analysis. Known TFBS annotations for the *E. coli* genome were obtained from RegulonDB, the most comprehensive knowledgebase on transcription initiation regulation in *Escherichia coli* K-12 (Salgado et al., 2024). All analyses used the reference genome assembly NC_000913.3 from NCBI.

Genome annotation data (from NCBI), experimentally validated TFBSs (from RegulonDB), and projections of pipeline results onto the target promoter region were visualized using the modular genome browser JBrowse2 (Diesh et al., 2023), which supports a wide range of standard genomic file formats. The taxonomic relationships among the analyzed genomes – based on NCBI taxonomy (Sayers et al., 2021) – were visualized using the ETE3 toolkit (Huerta-Cepas et al., 2016). Statistical summaries and visual representations of multiple pipeline runs were generated using the Matplotlib library (Hunter, 2007).

Results

Modification of the phylogenetic footprinting approach implemented in the MotifsOnFly method

The modification of the approach implemented in this study is based on multiple runs of the pipeline stage responsible for generating diverse subsets of promoter sequences, while preserving the selection principles previously demonstrated to be effective (Liu B. et al., 2016), and accumulation and statistical analysis of the results obtained from these repeated runs on the generated subsets. The conceptual workflow of the analysis implemented in the developed MotifsOnFly pipeline is shown in Figure 1. In accordance with the pipeline steps described in detail in the "Materials and methods" section, the protein sequence of the target gene is provided as input to the pipeline (1). This triggers the identification of orthologous genes using the GOST software module (2) across the set of annotated genomes available in the database (3). Subsequently, utilizing the same genomic sequences along with their operon structure annotations (3), promoter sequences of all identified orthologous genes are extracted (4).

The core of the developed modification is the stage of generating multiple promoter subsets (5), while preserving the selection criteria based on pairwise evolutionary distances among promoters. The effectiveness of using such selection criteria in phylogenetic footprinting was previously demonstrated (Liu B. et al., 2016). The full set of promoters (4), ordered by their evolutionary distances from the target promoter (calculated using ClustalW2), was divided into three subgroups with distances ranging from 0.05 to 0.31, from 0.31 to 0.55, and from 0.55 to 0.73, respectively. Based on the sizes of these subgroups, the maximum number M of subsets was determined

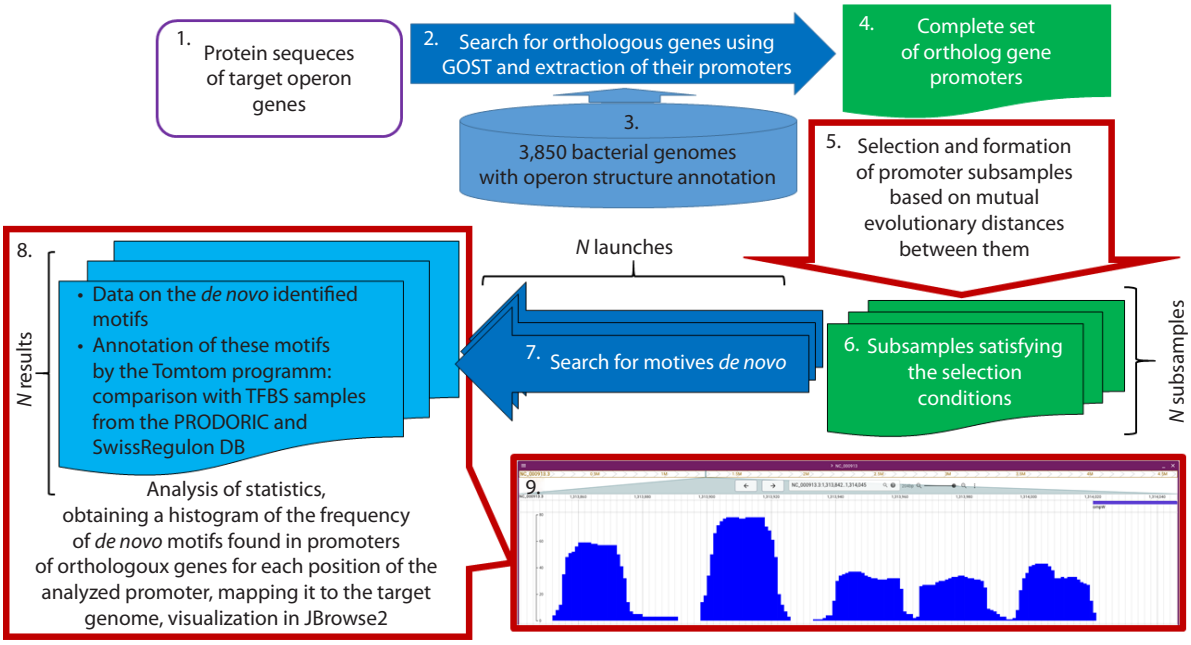


Fig. 1. Schematic of the pipeline for *de novo* identification of functional motifs corresponding to transcription factor binding sites, based on the phylogenetic footprinting approach and implemented in the MotifsOnFly method.

User-provided input data are indicated by a purple frame (1). Software modules (2 and 7) and external data sources (3) used in the pipeline are shown in blue. Intermediate data generated during pipeline execution are marked in green (4, 6). Steps specifically developed within the MotifsOnFly method are highlighted with a red border (5, 8, and 9). The stage involving selection and formation of promoter subsets based on pairwise evolutionary distances (5) enables multiple runs of the pipeline to perform *de novo* motif discovery (7) on these distinct subsets. The resulting data – including identified *de novo* motifs (8) and statistics on their positional distribution across aligned orthologous promoters – allow generation of a histogram showing the frequency of detected *de novo* motifs, which is visualized in the JBrowse2 genome browser (9). Furthermore, each *de novo* motif obtained from individual runs is compared using the Tomtom tool against known TFBS databases – SwissRegulon and PRODORIC (8) – and subsequent statistical analysis of these comparisons enables prediction of the most likely transcription factor regulators of the target promoter.

such that all promoters from the full set could be distributed into subsets of 12 promoters each, maintaining a fixed proportion of 3:3:6 (i. e. 3 promoters from the first subgroup, 3 from the second, and 6 from the third). In the final step, $N = 2M$ subsets were generated through random sampling according to the 3:3:6 proportion, along with additional selection rules that consider pairwise distances among sequences within each subset, thereby preventing the inclusion of overly similar promoters. This bootstrap-like approach (with replacement) ensures maximal – but not exhaustive – inclusion of promoters from the original set into the analysis, as adherence to selection principles that exclude low-information promoters remains a priority for method efficacy (McCue et al., 2002).

The resulting subsets, each supplemented with the target promoter (5), are used for repeated runs of the *de novo* motif discovery stage (7), as detailed in the "Materials and methods" section. Across these multiple runs, statistical data (8) are accumulated, recording which *de novo* motifs were identified at which positions in the target promoter and simultaneously detected in orthologous promoters. These aggregated data yield a histogram showing the frequency of detected *de novo* motifs at each position of the target promoter across orthologous sequences, enabling assessment of motif reliability at each site (Tompa et al., 2005). The histogram is saved in BigWig format for subsequent visualization in the JBrowse2 genome browser (9).

Analysis of the *ompW* gene promoter in *Escherichia coli* K-12 using the developed MotifsOnFly method

The best-studied and most comprehensively annotated bacterial genome to date remains that of *E. coli* K-12 (Salgado et al., 2024). To demonstrate the capabilities of our newly developed MotifsOnFly method – which extends the approach originally implemented by B. Liu and colleagues (2016) – we selected the *ompW* gene of *E. coli*. The expression of this gene is regulated by six transcription factors acting through five known binding sites. The *ompW* gene encodes an outer membrane protein and exhibits a broad range of physiological functions, including bacterial resistance to various antibiotics and herbicides, tolerance to osmotic stress, and support of bacterial growth under harsh environmental conditions such as hypoxia and elevated temperature (Zhang et al., 2020). A distinctive feature of the promoter of this operon –

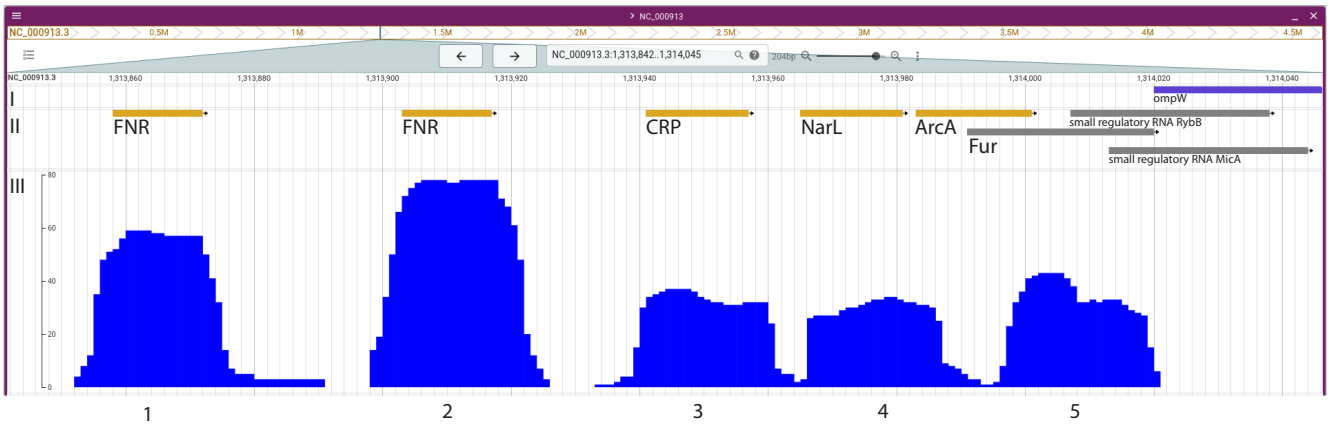


Fig. 2. Screenshot of the JBrowse2 genome browser.

I – genome annotation in the *ompW* promoter region according to NCBI data; II – experimentally validated transcription factor binding site (TFBS) locations from the RegulonDB database; III – pipeline output: a histogram showing the frequency of *de novo* motifs detected at each position of the analyzed promoter across orthologous sequences. The histogram displays distinct peaks, numbered in accordance with their reference in the main text.

which contains only the *ompW* gene – is the presence of both isolated transcription factor binding sites (TFBSs) and clusters of overlapping TFBSs for different transcription factors. As input for the pipeline, we used the protein and promoter sequences of the gene, as described in the "Materials and methods" section.

Figure 2 shows the pipeline results visualized using JBrowse2, including: (I) genome annotation in the *ompW* gene region according to NCBI data; (II) experimentally validated TFBS locations from the RegulonDB database; and (III) a histogram displaying the frequency of *de novo* motifs detected at each position of the analyzed promoter across orthologous sequences. The histogram reveals distinct peaks (Fig. 2), indicating positions in the *ompW* promoter where *de novo* motifs were consistently identified across a substantial proportion of pipeline runs. Specifically: peaks 1 and 2 coincide with experimentally confirmed binding sites for the transcription factor FNR; peak 3 coincides with the known binding site for the transcription factor CRP; peak 4 overlaps with the experimentally verified binding site for the transcription factor NarL; peak 5 overlaps with binding sites for both ArcA and Fur, illustrating the complexity of deciphering regulatory signals near the transcription start site and warranting further dedicated analysis. Additionally, a noticeable "tail" to the right of peak 1 indicates that *de novo* motifs were also detected in this region during some pipeline runs. This observation confirms the instability of results obtained from single-run analyses and underscores the value of the MotifsOnFly method's multi-run, statistically robust framework.

Comparison of the identified *de novo* motifs with known transcription factor binding sites (TFBSs) from the SwissRegulon and PRODORIC databases using the Tomtom tool allows prediction of the most likely transcription factor regulators of the analyzed promoter. For each such comparison, an E-value is calculated to assess the statistical significance of the similarity between the discovered motif and each known TFBS. We computed E-value statistics for all candidate transcription factors corresponding to all peaks (Fig. 3). As evident from the charts, this statistical analysis enables unambiguous identification of FNR as the most probable transcription factor binding to peaks 1 (Fig. 3*a*, *b*) and 2 (Fig. 3*c*, *d*), and CRP for peak 3 (Fig. 3*d*, *e*), regardless of which database is used.

For peak 4, significant matches were found only with sites from the SwissRegulon database (Fig. 4), specifically for the transcription factors NanR and NarL. For peak 5, no significant matches with known TFBSs were obtained from either database, suggesting the need for further in-depth analysis.

It should be noted that for peak 1, in the case of a single pipeline run, FNR would be identified as the top candidate in only 72 % of cases when using the SwissRegulon database and in just 43 % of cases with PRODORIC. Similarly, for peak 2, correct identification of FNR occurs in 90 and 65 % of single runs (SwissRegulon and PRODORIC, respectively), while for peak 3, CRP is correctly prioritized in only 49 and 54 % of cases, respectively. Our proposed modified approach overcomes this instability inherent in single-run phylogenetic footprinting analyses and enables unambiguous prioritization of results that fully align with experimental data.

Moreover, the method allows tracking the presence of *de novo* motifs corresponding to each peak across all orthologous promoters and visualizing this information on the taxonomic tree of the genomes analyzed. Specifically, if a similar motif is detected in the promoter of an orthologous gene from a given bacterial species at the position aligned with a peak in the target promoter, this is indicated by a red dot on the corresponding node of the tree. As an example, Figure 5 illustrates the occurrence of such similar motifs in orthologous promoters at positions aligned with peak 1 of the target promoter.

The information enabling tracking of similar motifs across all orthologous promoters (Fig. 5) within a single peak also allows the construction of a consolidated alignment that includes

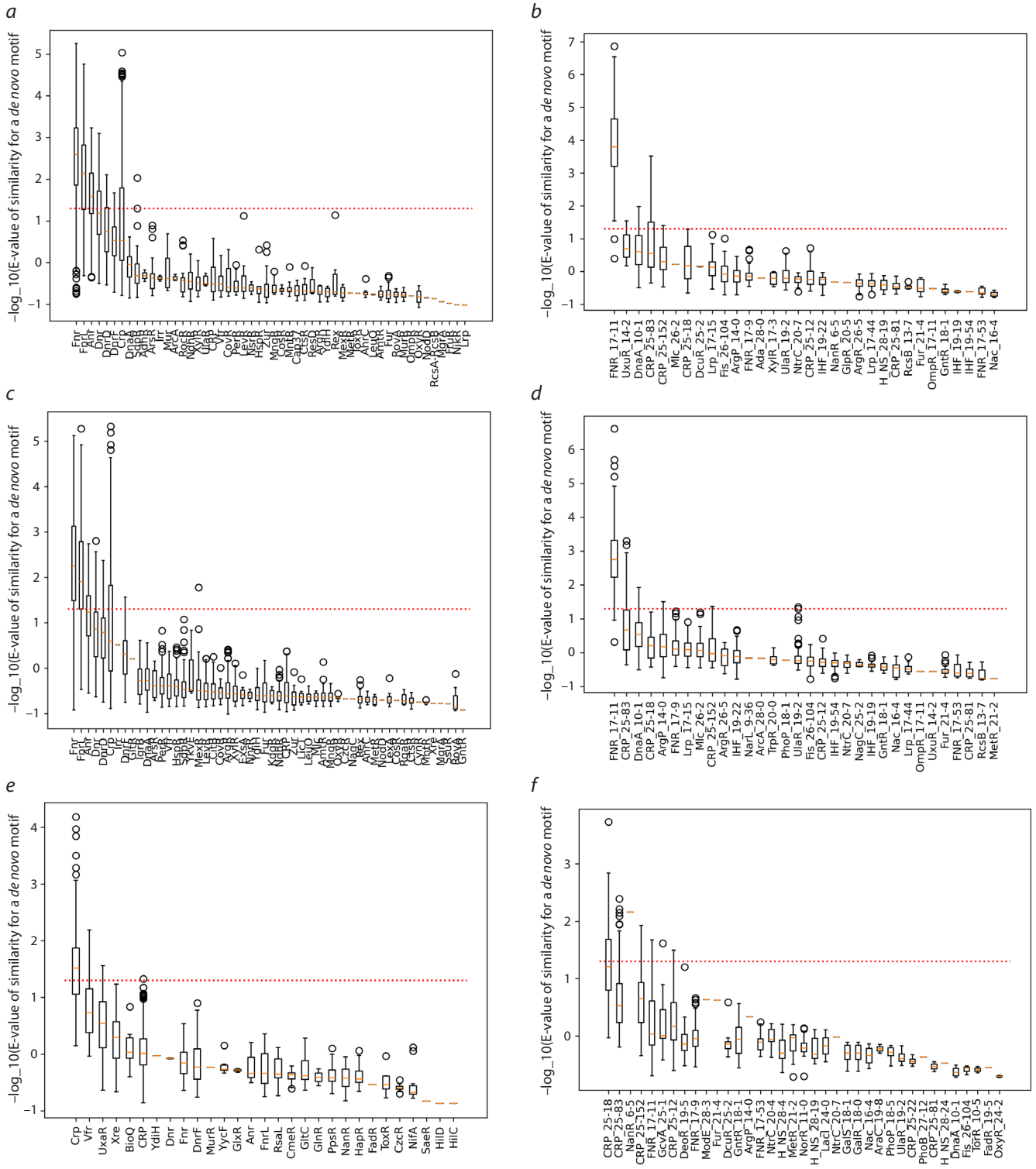


Fig. 3. Box-and-whisker plots showing the statistics of comparisons between the identified motifs for peak 1 (*a*, *b*), peak 2 (*c*, *d*), and peak 3 (*e*, *f*) against known TFBS databases PRODORIC (*a*, *c*, *e*) and SwissRegulon (*b*, *d*, *f*).

The red dashed line indicates the 5 % significance threshold. Statistical analysis results from multiple pipeline runs were visualized using the Matplotlib software package.

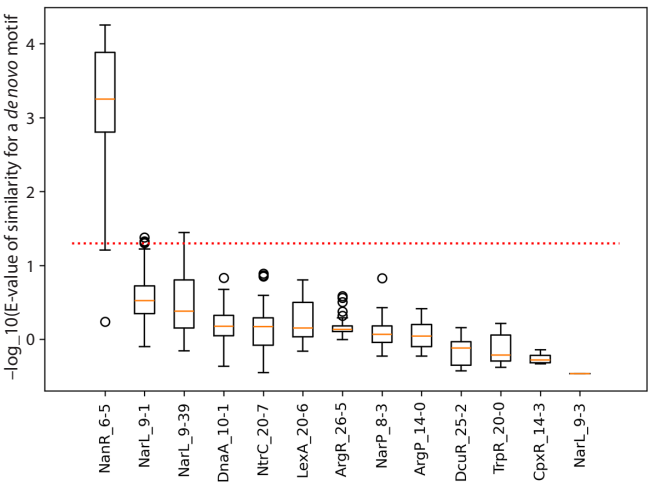


Fig. 4. Box-and-whisker plot showing the statistics of comparisons between the identified motifs for peak 4 and known transcription factor binding sites (TBFSs) from the SwissRegulon database.

The red dashed line indicates the 5 % significance threshold. Results of the statistical analysis from multiple pipeline runs were visualized using the Matplotlib software package.

only those orthologous promoters in which a similar motif was detected simultaneously with the target promoter in at least one pipeline run. Thus, this approach yields a consolidated alignment of all conserved DNA segments corresponding to a given peak, ensuring maximal representation of biologically relevant sequences in the resulting alignment. The boundaries of these informative aligned regions are naturally defined by the width of each peak, which served as the criterion for selecting the length of the consolidated alignment. This strategy also incorporates flanking sequences adjacent to the core binding sites, thereby improving the quality and sensitivity of comparisons with known TFBS databases. These consolidated alignments for each peak were subsequently analyzed using the Tomtom tool. Comparison of the consolidated alignments for peaks 1–3 against known TFBSs yielded the highest confidence matches when using the SwissRegulon database (Fig. 6). In full agreement with experimental data, the top-scoring candidates were again FNR for peak 1 (Fig. 6*a*) and peak 2 (Fig. 6*b*), and CRP for peak 3 (Fig. 6*c*), with E-values of 3.13×10^{-8}, 9.17×10^{-8}, and 1.80×10^{-4}, respectively. These results are consistent with the earlier statistical analysis of individual motif comparisons (Fig. 3) but demonstrate substantially higher confidence in motif similarity due to the increased representativeness and signal-to-noise ratio achieved through the consolidated alignments.

Tomtom analysis of the consolidated alignment for peak 4 yielded significant results only when using the SwissRegulon database: a statistically significant similarity was found with the NanR binding site (Fig. 7*a*), while similarity to the NarL binding site was not statistically significant (Fig. 7*b*). No other transcription factors with binding sites resembling peak 4 were identified.

The consolidated alignment for the peak 5 region, when analyzed with Tomtom, yielded results only upon comparison with the PRODORIC database. However, none of the transcription factor binding sites showed statistically significant similarity to this alignment. Nevertheless, weak (non-significant) similarities were observed with the binding sites of ArcA and Fur (Fig. 8*a*, *b*), the known binding sites of which substantially overlap in this genomic region.

Discussion

We have developed a computational pipeline for identifying functional motifs corresponding to transcription factor binding sites (TFBSs), based on a modification of the phylogenetic footprinting method originally proposed by B. Liu and colleagues (2016). The core idea of our modification lies in repeated execution of an effective promoter selection procedure that accounts for pairwise evolutionary distances, thereby generating diverse promoter subsets. The pipeline is then run independently on each subset, and the results from all runs undergo statistical aggregation (Fig. 1). The approach implemented in our MotifsOnFly method employs a bootstrap-with-replacement scheme, enabling maximal utilization of currently available genomic data while simultaneously leveraging statistical robustness across varying promoter subsets.

The effectiveness of the pipeline was demonstrated through the analysis of the *ompW* gene promoter – one of the most comprehensively annotated promoters in *Escherichia coli* K-12. This gene encodes an outer membrane protein and is known to be regulated by six transcription factors, with five experimentally validated binding sites (Salgado et al., 2024). A detailed regulatory map of this promoter was previously described (Xiao et al., 2016). Our pipeline generated a histogram showing the frequency of *de novo* motifs detected at each position of the target promoter across orthologous sequences (Fig. 2). This histogram revealed distinct peaks that correspond precisely to experimentally confirmed TFBSs, thereby providing reliable identification of conserved, motif-enriched promoter regions. Thus, our approach vividly illustrates the essence of phylogenetic footprinting – highlighting the most evolutionarily conserved segments within alignments of orthologous promoters.

Statistical analysis of *de novo* motifs identified across multiple pipeline runs, followed by comparison against known TFBSs from SwissRegulon and PRODORIC, enabled confident assignment of the most likely regulatory factors. Predictions made by MotifsOnFly fully aligned with experimental data for peaks 1, 2, and 3: FNR (Constantinidou et al., 2006; Myers et al., 2013; Xiao et al., 2016) was correctly identified as the regulator for peaks 1 and 2, and CRP (Gaston et al., 1990; Ushida, Aiba, 1990; Xiao et al., 2016) for peak 3 (Fig. 3). Importantly, however, single-run analyses often failed to yield the correct top candidate – only 72 (SwissRegulon) or 43 % (PRODORIC) of single runs correctly prioritized FNR for peak 1, for example. This underscores the instability of single-run phylogenetic footprinting, which our multi-run framework successfully overcomes. Moreover, by constructing consolidated alignments of all conserved motif instances corresponding to

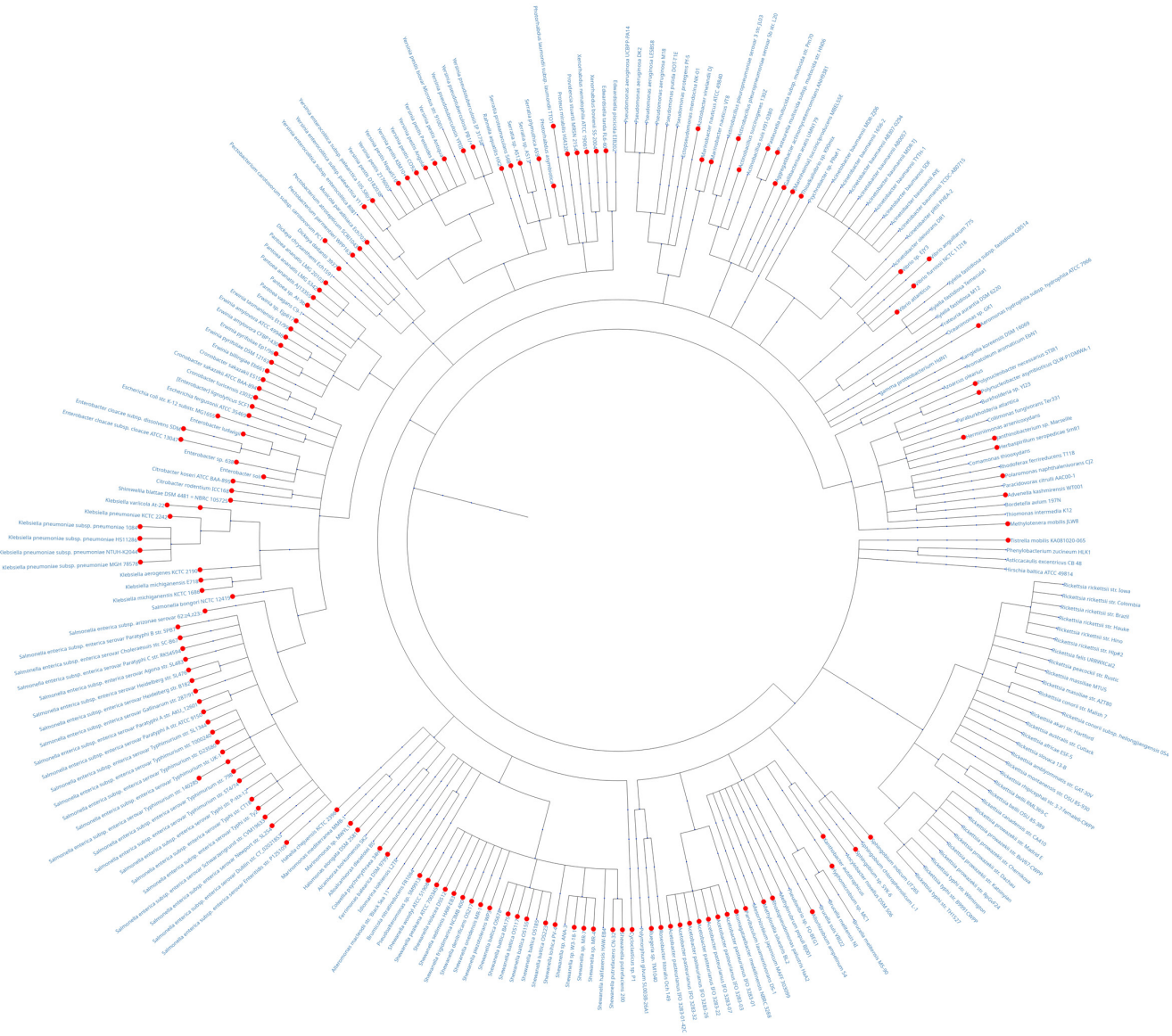


Fig. 5. Visualization of the presence of similar motifs (indicated by red dots) in the promoters of orthologous genes from various bacterial species at positions corresponding to peak 1 in the alignment with the target promoter.

The tree was constructed using taxonomic information for the analyzed genomes as provided by NCBI. Tree visualization was performed with the ETE3 package (Huerta-Cepas et al., 2016).

each peak and comparing them to reference databases (Fig. 6), we achieved extremely high-confidence matches, with E-values as low as 3.13×10^{-8}. Such high-quality alignments not only confirm known regulatory interactions but also open the possibility of generating reliable motif models for TFBSs that are underrepresented in current databases. The pipeline also includes functionality to track the presence of *de novo* motifs across all analyzed orthologous promoters and visualize this information on a taxonomic tree (Fig. 5) – a feature valuable for studying the evolutionary conservation and distribution of regulatory signals.

An intriguing result emerged for peak 4 (Fig. 4 and 7). Although experimental data indicate a NarL binding site in this region (Tyson et al., 1993, 1994; Xiao et al., 2016), both the statistical analysis of multiple runs and the consolidated alignment point to NanR as the top-scoring TF, with NarL ranked second. The consensus NarL site "TACYYMT" does show similarity to the peak 4 alignment, yet our highest-confidence match corresponds to the NanR motif GGTATA (Fig. 7). However, this finding warrants caution: NanR functions as a dimer, and high-affinity DNA binding requires three adjacent GGTATA repeats for cooperative binding of three NanR dimers (Kalivoda et al., 2013; Horne et al., 2021). The presence of only a single GGTATA instance at peak 4 casts doubt on NanR's biological relevance in regulating *ompW*. Furthermore, NanR's regulatory role in *E. coli* is currently considered to be minor,

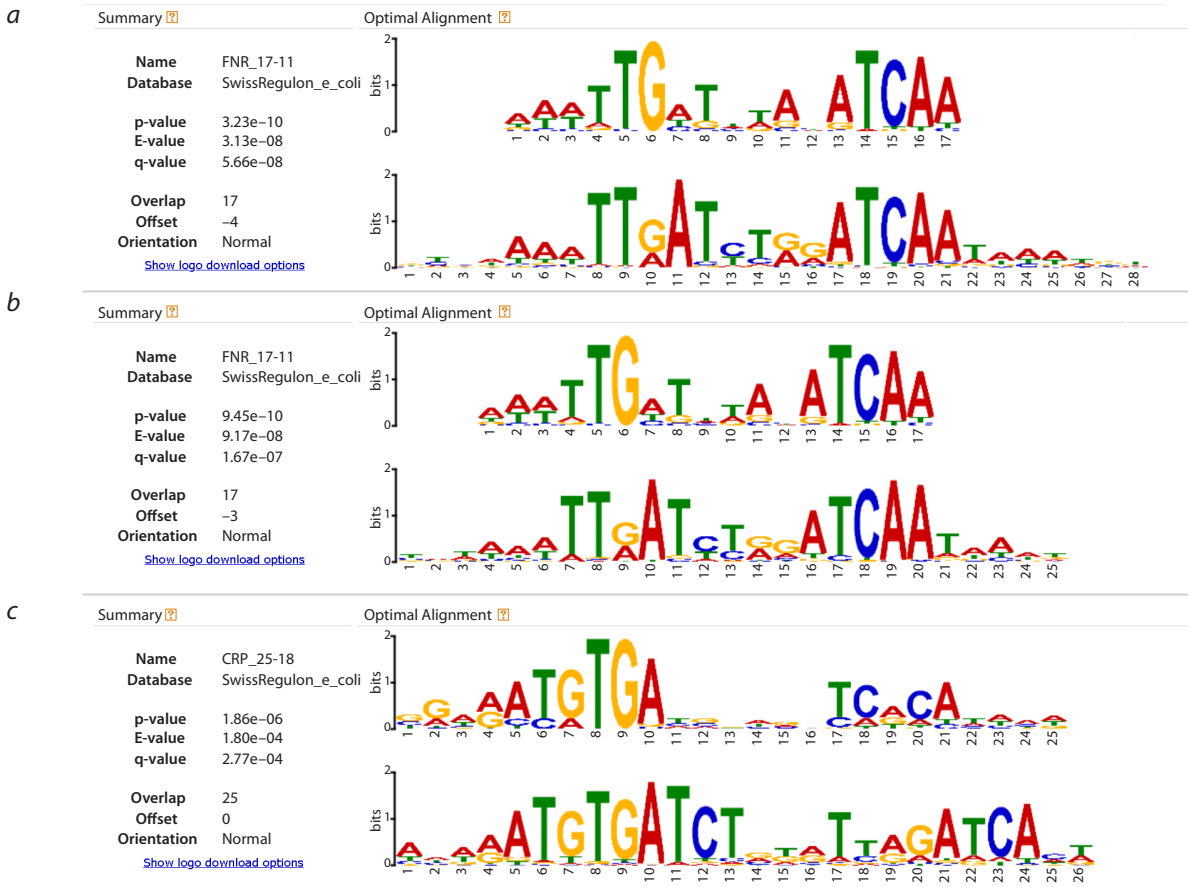


Fig. 6. Results of the comparison between the consolidated alignments of peak regions 1–3 and known transcription factor binding sites.

For each peak, a screenshot of the best match from the SwissRegulon database, as reported by the Tomtom tool, is shown. The alignments exhibit statistically significant similarity to FNR binding sites for peak 1 (*a*, E-value = 3.13×10^{-8}), FNR binding sites for peak 2 (*b*, E-value = 9.17×10^{-8}), and CRP binding site for peak 3 (*c*, E-value = 1.80×10^{-4}). These results are in complete agreement with experimental data.

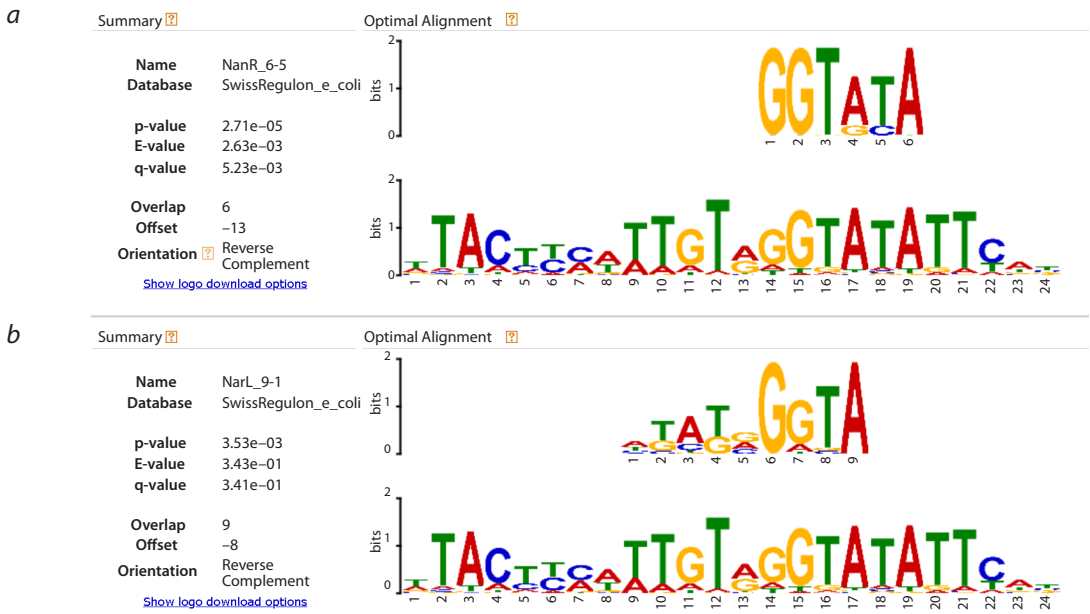


Fig. 7. Comparison of the consolidated alignment for the peak 4 region.

Screenshots from the Tomtom output against the SwissRegulon database are shown. The alignment exhibits significant similarity to the NanR binding site (*a*) and non-significant similarity to the NarL binding site (*b*).

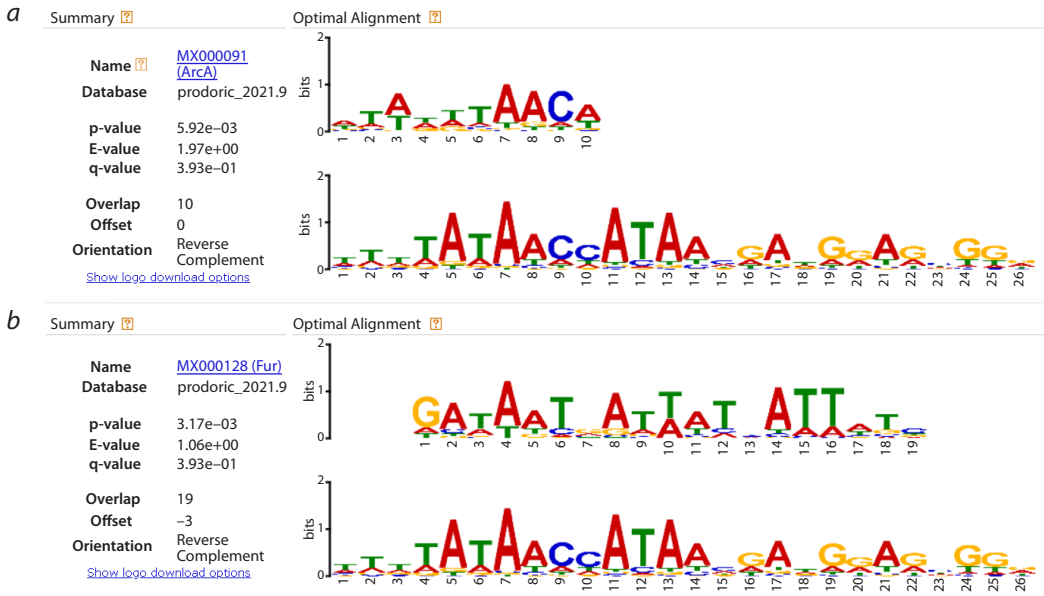


Fig. 8. Comparison of the consolidated alignment for the peak 5 region.

Screenshot from the Tomtom output against the PRODORIC database is shown. Weak (non-significant) similarities are observed between parts of the consolidated alignment for peak 5 and the binding sites of ArcA (*a*) and Fur (*b*). Notably, the predicted binding sites for these two transcription factors exhibit substantial overlap in this region.

limited to four operons involved in sialic acid catabolism (Kalivoda et al., 2013; Shimada et al., 2018). Nevertheless, the possibility of *ompW* (or some orthologs) being regulated via this NanR-like site merits further experimental investigation.

Regarding peak 5, it fully overlaps with the σ70 RNA polymerase binding region, encompassing the transcription start site and canonical −35/−10 promoter elements. This area is densely packed with regulatory signals: RegulonDB (Salgado et al., 2024) annotates overlapping binding sites for ArcA (Park et al., 2013; Xiao et al., 2016) and Fur (Zhang et al., 2020). The structural plasticity of ArcA binding sites allows them to coexist with other TF motifs in compact sequence space (Park et al., 2013). Meanwhile, the Fur site in this region is considered weak (Zhang et al., 2020), possibly due to competitive interactions with SoxS, which can bind the −35 element under oxidative stress conditions induced by iron-bound Fur (Graham et al., 2012; Taliaferro et al., 2012). Thus, the promoter region of the *ompW* gene in the vicinity of peak 5 possesses substantial regulatory potential. Notably, our analysis did not identify any strong similarity between the peak 5 consolidated alignment and known TFBSs in either database. This may reflect the absence of strong, obstructive TFBSs that would interfere with σ70 RNA polymerase binding – consistent with the need for basal transcription. At the same time, the alignment shows comparable, low-significance similarity to numerous TFBSs in PRODORIC (Fig. 8), suggesting the presence of multiple weak, overlapping sites, including those for ArcA and Fur. This interpretation aligns well with existing literature and highlights the complex, layered regulatory logic operating near the core promoter.

Conclusion

The modification of the phylogenetic footprinting approach proposed in our work and implemented in the MotifsOnFly method represents a natural evolution of the strategy developed by B. Liu and colleagues (2016), designed to address the continuously growing volume of large-scale genomic data. MotifsOnFly enables multiple pipeline runs on diverse subsets of orthologous promoter sequences, yielding refined localization of transcription factor binding sites (TFBSs) and facilitating robust statistical analysis. The forms of such statistical analysis are not limited to those described in this article and can be further extended according to the specific goals and requirements of individual research projects. As demonstrated here, the MotifsOnFly method produces reliable and stable TFBS predictions, making it a valuable tool for a broad community of researchers engaged in the annotation and analysis of regulatory sequences in bacterial genomes.

References

Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., Noble W.S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(W):W202-W208. doi 10.1093/nar/gkp335

Blanchette M., Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*. 2002; 12(5):739-748. doi 10.1101/gr.6902

Blanchette M., Tompa M. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res*. 2003;31(13):3840-3842. doi 10.1093/nar/gkg606

Browning D.F., Busby S.J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*. 2004;2(1):57-65. doi 10.1038/nrmicro787

Chen X., Guo L., Fan Z., Jiang T. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics*. 2008;24(9):1121-1128. doi 10.1093/bioinformatics/btn088

Claeys M., Storms V., Sun H., Michoel T., Marchal K. MotifSuite: workflow for probabilistic motif detection and assessment. *Bioinformatics*. 2012;28(14):1931-1932. doi 10.1093/bioinformatics/bts293

Constantinidou C., Hobman J.L., Griffiths L., Patel M.D., Penn C.W., Cole J.A., Overton T.W. A reassessment of the FNR regulon and transcriptomic analysis of the effects of nitrate, nitrite, NarXL, and NarQP as *Escherichia coli* K12 adapts from aerobic to anaerobic growth. *J Biol Chem*. 2006;281(8):4802-4815. doi 10.1074/jbc.M512312200

Diesh C., Stevens G.J., Xie P., De Jesus Martinez T., Hershberg E.A., Leung A., Guo E., … Haw R., Cain S., Buels R.M., Stein L.D., Holmes I.H. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol*. 2023;24(1):74. doi 10.1186/s13059-023-02914-z

Dudek C.-A., Jahn D. PRODORIC: state-of-the-art database of prokaryotic gene regulation. *Nucleic Acids Res*. 2022;50(D1):D295-D302. doi 10.1093/nar/gkab1110

Gaston K., Bell A., Kolb A., Buc H., Busby S. Stringent spacing requirements for transcription activation by CRP. *Cell*. 1990;62(4):733-743. doi 10.1016/0092-8674(90)90118-x

Graham A.I., Sanguinetti G., Bramall N., McLeod C.W., Poole R.K. Dynamics of a starvation-to-surfeit shift: a transcriptomic and modelling analysis of the bacterial response to zinc reveals transient behaviour of the Fur and SoxS regulators. *Microbiology* (*Reading*). 2012;158(Pt.1):284-292. doi 10.1099/mic.0.053843-0

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

Hertz G.Z., Stormo G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. 1999;15(7-8):563-577. doi 10.1093/bioinformatics/15.7.563

Horne C.R., Venugopal H., Panjikar S., Wood D.M., Henrickson A., Brookes E., North R.A., Murphy J.M., Friemann R., Griffin M.D.W., Ramm G., Demeler B., Dobson R.C.J. Mechanism of NanR gene repression and allosteric induction of bacterial sialic acid metabolism. *Nat Commun*. 2021;12(1):1988. doi 10.1038/s41467-021-22253-6

Huerta-Cepas J., Serra F., Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33(6):1635-1638. doi 10.1093/molbev/msw046

Hunter J.D. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95. doi 10.1109/MCSE.2007.55

Kalivoda K.A., Steenbergen S.M., Vimr E.R. Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *J Bacteriol*. 2013;195(20):4689-4701. doi 10.1128/JB.00692-13

Katara P., Grover A., Sharma V. Phylogenetic footprinting: a boost for microbial regulatory genomics. *Protoplasma*. 2012;249(4):901-907. doi 10.1007/s00709-011-0351-9

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947-2948. doi 10.1093/bioinformatics/btm404

Levy S., Hannenhalli S., Workman C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*. 2001;17(10):871-877. doi 10.1093/bioinformatics/17.10.871

Li G., Liu B., Ma Q., Xu Y. A new framework for identifying cis-regulatory motifs in prokaryotes. *Nucleic Acids Res*. 2011a;39(7):e42. doi 10.1093/nar/gkq948

Li G., Ma Q., Mao X., Yin Y., Zhu X., Xu Y. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. *Nucleic Acids Res*. 2011b;39(22):e150. doi 10.1093/nar/gkr766

Liu B., Zhang H., Zhou C., Li G., Fennell A., Wang G., Kang Y., Liu Q., Ma Q. An integrative and applicable phylogenetic footprinting framework for *cis*-regulatory motifs identification in prokaryotic genomes. *BMC Genomics*. 2016;17:578. doi 10.1186/s12864-016-2982-x

Liu X., Brutlag D.L., Liu J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*. 2001;6:127-138

Liu X.S., Brutlag D.L., Liu J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*. 2002;20(8):835-839. doi 10.1038/nbt717

Mao X., Ma Q., Zhou C., Chen X., Zhang H., Yang J., Mao F., Lai W., Xu Y. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res*. 2014;42(D1):D654-D659. doi 10.1093/nar/gkt1048

McCue L.A., Thompson W., Carmack C.S., Lawrence C.E. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* 2002;12(10):1523-1532. doi 10.1101/gr.323602

Mukhin A., Oschepkov D., Lashin S. A computational pipeline for *de novo* recognition of transcription factor binding sites in bacterial genomes. *Problems of Informatics*. 2024;4(65):69-83. doi 10.24412/2073-0667-2024-4-69-83 (in Russian)

Myers K.S., Yan H., Ong I.M., Chung D., Liang K., Tran F., Keleş S., Landick R., Kiley P.J. Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. *PLoS Genet*. 2013;9(6):e1003565. doi 10.1371/journal.pgen.1003565

Olman V., Xu D., Xu Y. CUBIC: identification of regulatory binding sites through data clustering. *J Bioinform Comput Biol*. 2003;1(1):21-40. doi 10.1142/s0219720003000162

Pachkov M., Balwierz P.J., Arnold P., Ozonov E., van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*. 2013;41(D1):D214-D220. doi 10.1093/nar/gks1145

Park D.M., Akhtar M.S., Ansari A.Z., Landick R., Kiley P.J. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. *PLoS Genet*. 2013;9(10):e1003839. doi 10.1371/journal.pgen.1003839

Peltek S., Bannikova S., Khlebodarova T.M., Uvarova Y., Mukhin A.M., Vasiliev G., Scheglov M., Shipova A., Vasilieva A., Oshchepkov D., Bryanskaya A., Popik V. The transcriptomic response of cells of the thermophilic bacterium *Geobacillus icigianus* to terahertz irradiation. *Int J Mol Sci*. 2024;25(22):12059. doi 10.3390/ijms252212059

Salgado H., Gama-Castro S., Lara P., Mejia-Almonte C., Alarcón-Carranza G., López-Almazo A.G., Betancourt-Figueroa F., … Hernández-Alvarez A.J., Santos-Zavaleta A., Capella-Gutierrez S., Gelpi J.L., Collado-Vides J. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *E. coli* K-12. *Nucleic Acids Res*. 2024;52(D1):D255-D264. doi 10.1093/nar/gkad1072

Sayers E.W., Beck J., Bolton E.E., Bourexis D., Brister J.R., Canese K., Comeau D.C., … Wang J., Ye J., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021;49(D1):D10-D17. doi 10.1093/nar/gkaa892

Shimada T., Ogasawara H., Ishihama A. Single-target regulators form a minor group of transcription factors in *Escherichia coli* K-12. *Nucleic Acids Res*. 2018;46(8):3921-3936. doi 10.1093/nar/gky138

Tagle D.A., Koop B.F., Goodman M., Slightom J.L., Hess D.L., Jones R.T. Embryonic epsilon and gamma globin genes of a pro-

simian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*. 1988;203(2):439-455. doi 10.1016/0022-2836(88)90011-3

Taliaferro L.P., Keen E.F., Sanchez-Alberola N., Wolf R.E. Transcription activation by *Escherichia coli* Rob at class II promoters: protein-protein interactions between Rob's N-terminal domain and the σ^{70} subunit of RNA polymerase. *J Mol Biol*. 2012;419(3-4):139-157. doi 10.1016/j.jmb.2012.03.019

Tompa M., Li N., Bailey T.L., Church G.M., De Moor B., Eskin E., Favorov A.V., … Vandenbogaert M., Weng Z., Workman C., Ye C., Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*. 2005;23(1):137-144. doi 10.1038/nbt1053

Tyson K.L., Bell A.I., Cole J.A., Busby S.J. Definition of nitrite and nitrate response elements at the anaerobically inducible *Escherichia coli nirB* promoter: interactions between FNR and NarL. *Mol Microbiol*. 1993;7(1):151-157. doi 10.1111/j.1365-2958.1993.tb01106.x

Tyson K.L., Cole J.A., Busby S.J. Nitrite and nitrate regulation at the promoters of two *Escherichia coli* operons encoding nitrite reductase: identification of common target heptamers for both NarP- and NarL-dependent regulation. *Mol Microbiol*. 1994;13(6):1045-1055. doi 10.1111/j.1365-2958.1994.tb00495.x

Ushida C., Aiba H. Helical phase dependent action of CRP: effect of the distance between the CRP site and the –35 region on promoter activity. *Nucleic Acids Res*. 1990;18(21):6325-6330. doi 10.1093/nar/18.21.6325

Xiao M., Lai Y., Sun J., Chen G., Yan A. Transcriptional regulation of the outer membrane porin gene *ompW* reveals its physiological role during the transition from the aerobic to the anaerobic lifestyle of *Escherichia coli*. *Front Microbiol*. 2016;7:799. doi 10.3389/fmicb.2016.00799

Zhang P., Ye Z., Ye C., Zou H., Gao Z., Pan J. OmpW is positively regulated by iron via Fur, and negatively regulated by SoxS contribution to oxidative stress resistance in *Escherichia coli*. *Microb Pathog*. 2020;138:103808. doi 10.1016/j.micpath.2019.103808

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

Received November 15, 2025. Revised December 9, 2025. Accepted December 10, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):27-35

doi 10.18699/vjgb-26-09

Wheat spikelet detection on RGB images using deep machine learning

M.A. Genaev [1, 2], I.D. Busov[1, 3], Yu.V. Kruchinina [1, 2], V.S. Koval[1, 2], N.P. Goncharov[1, 3]

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

Abstract. This study addresses the challenge of automated high-throughput phenotyping of wheat spike characteristics using modern computer vision and deep learning methods. Accurate estimation of spikelet number is a key indicator of plant productivity, yet traditional manual counting approaches are labor-intensive, slow, and difficult to scale to large breeding datasets. To overcome these limitations, we propose a spikelet detection strategy based on simplified point annotations, where an expert marks only the centers of spikelets rather than drawing detailed segmentation masks or bounding boxes. This significantly reduces annotation time and lowers the overall cost of preparing training datasets for machine learning models. To determine the most effective way of utilizing such simplified annotations, three computational methods were explored: segmentation of binary masks using a U-Net architecture, density regression based on two-dimensional Gaussian distributions optimized via Kullback–Leibler divergence, and detection of fixed-size bounding regions using the YOLOv8 object detection framework. The models were evaluated on dedicated test datasets using both quantitative metrics (MAE, MAPE) and spatial localization metrics (Precision, Recall, F1 score). The results demonstrate that U-Net-based approaches provide consistently high accuracy in spikelet localization and counting while maintaining robustness to annotation imperfections. In contrast, the YOLOv8-based method showed reduced performance, likely due to the geometric mismatch between fixed-size boxes and the natural elongated shape of spikelets. Overall, the proposed methodology highlights the effectiveness of combining minimalistic point-level annotation with advanced segmentation models for automating phenotyping workflows. This approach has the potential to accelerate breeding programs, enhance the efficiency of large-scale phenotypic data collection, and support further development of robust computer-vision tools for plant science applications.

Key words: computer vision; deep learning; wheat; spike; spikelets per spike; phenotyping; object detection

For citation: Genaev M.A., Busov I.D., Kruchinina Yu.V., Koval V.S., Goncharov N.P. Wheat spikelet detection on RGB images using deep machine learning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):27-35. doi 10.18699/vjgb-26-09

Funding. Data preparation, development and verification of the algorithm were carried out with the support of the Russian Science Foundation, project No. 23-14-00150.

Acknowledgements. The calculations were performed using the computing resources of the Bioinformatics Computing Center with the support of budget project No. FWNR-2026-0023.

Детекция колосков в колосе пшеницы на RGB-изображениях с использованием глубокого машинного обучения

М.А. Генаев [1, 2], И.Д. Бусов[1, 3], Ю.В. Кручинина [1, 2], В.С. Коваль[1, 2], Н.П. Гончаров[1, 3]

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

Аннотация. В работе рассматривается задача автоматизированного высокопроизводительного фенотипирования признаков колоса пшеницы с использованием современных методов компьютерного зрения и глубокого обучения. Точная оценка числа колосков является важным компонентом анализа продуктивности растения, однако традиционные методы ручной разметки и подсчета крайне трудоемки, плохо масштабируются и требуют значительных временных затрат. В исследовании предложен подход для эффективной детекции колосков, основанный на использовании упрощенной точечной разметки, при которой эксперт отмечает только центры колосков, без необходимости формировать трудоемкие пиксельные маски или ограничивающие рамки. Такая схема позволяет

© Genaev M.A., Busov I.D., Kruchinina Yu.V., Koval V.S., Goncharov N.P., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

существенно снизить стоимость подготовки обучающей выборки и ускорить процесс аннотации. Для определения оптимального способа обработки упрощенной разметки были исследованы три метода: сегментация бинарных масок с помощью архитектуры U-Net, регрессия плотностных карт на основе двумерного нормального распределения и функции дивергенции Кульбака–Лейблера, а также детекция областей фиксированного размера с использованием модели YOLOv8. Проведено сравнение точности методов по количественным (MAE, MAPE) и пространственным метрикам (Precision, Recall, F1) на тестовых наборах изображений. Анализ результатов показал, что подходы, основанные на U-Net, обеспечивают высокую точность локализации и подсчета колосков при минимальных затратах на разметку данных, тогда как метод YOLOv8 менее устойчив к геометрической вариативности реальных объектов. Предложенный подход демонстрирует, что комбинация точечной разметки и современных моделей сегментации является эффективным инструментом для автоматизации фенотипирования, что может значительно ускорить селекционные исследования и расширить возможности высокопроизводительного анализа морфологических признаков растений.

Ключевые слова: компьютерное зрение; глубокое обучение; пшеница; колос; число колосков; фенотипирование; детекция объектов

Introduction

Population growth and local and global climate change necessitate accelerated breeding of agricultural crops, such as wheat, to increase yield and resistance to biotic and abiotic environmental factors (Efimov et al., 2024). One of the key stages of breeding is the phenotyping of spike parameters – the assessment of phenotypic traits, among which the number of spikelets in a spike is one of the most important indicators of wheat plant productivity (Afonnikov et al., 2016; Skripka et al., 2016; Maslova et al., 2018; Vahamidis et al., 2019).

The number of spikelets in a wheat spike has complex genetic control (Zhang B. et al., 2015) and is often associated with spike density, another important breeding and taxonomic trait (Vavilova et al., 2017, 2019; Savin, 2019).

Traditional methods of phenotyping spike parameters based on visual assessment by experts are slow, expensive, and subjective (Konopatskaia et al., 2016). This has stimulated the development of automated, computer vision-based solutions for high-throughput analysis (Li et al., 2017; Liu et al., 2017; Genaev et al., 2019).

Deep learning-based computer vision methods have proven effective in automating the phenotyping of agricultural plants (Artemenko et al., 2024). Existing approaches to solving the problem of counting spikelets can be divided into several types. The first type is based on object detection. For example, F. Khoroshevsky et al. (2021) used the RetinaNet architecture to detect and count spikelets directly in the field, achieving mean absolute percentage error (MAPE) values ranging from 9.2 to 11.5 %. L. Shi et al. (2023) applied the YOLOv5s model to detect the number of spikelets in images of spikes, obtaining an average absolute error (MAE) of 0.43 for the number of spikelets in a test sample of mature wheat. The second type of approach uses semantic segmentation. T. Misra et al. (2020) proposed the SpikeSegNet architecture based on U-Net, which achieved an accuracy of 95 %. However, what these methods have in common is the need for labor-intensive and costly full data annotation (using bounding boxes or pixel masks), which becomes a key limitation when scaling up research.

As an alternative to reduce annotation costs, point annotation is proposed, where the expert only needs to mark the center of the object. F. Chen et al. (2021) demonstrated that training with incomplete point annotation, where only 50 % of spikelets are annotated, results in a 6.5 % loss in accuracy (F1 drops from 84.15 to 78.65 %) compared to full annotation. Reducing the proportion of labeled objects to 10 % leads to a 16.5 % decrease in accuracy. Thus, reducing the labor costs of labeling by a factor of 10 preserves 83.5 % of the original accuracy, which confirms the promise of this approach. Another way to simplify data preparation is semi-automatic labeling (Alkhudaydi et al., 2019). R. Qiu et al. (2022) proposed a method where the initial model is trained on sparsely labeled data and then used to automatically generate labels, on which a new model is retrained.

These studies demonstrate progress in the automation of spike parameter phenotyping, but also point to a major problem: the complexity and cost of data preparation. Despite this progress, the task of developing an accurate and robust algorithm that works effectively with simplified annotation and maintains high accuracy remains relevant.

In this paper, we explore an alternative approach based on the use of simplified point annotation of only the centers of the spikes, which significantly reduces the time and cost of data preparation (Chen et al., 2021) and may be an effective and practical solution for widespread use in agricultural research.

Materials and methods

Biological material and data set

The study used wheat spikes from the collection of N.P. Goncharov. The sample of plants included representatives of various species of diploids ($2n = 2x = 14$), tetraploid ($2n = 4x = 14$), and hexaploid ($2n = 6x = 42$) wheat. Images of the spikes were obtained in the laboratory according to the protocol described earlier (Genaev et al., 2018). A Canon 350D digital camera with an EF-S 18–55 mm f/3.5–5.6 lens was used for shooting. Shooting parameters: exposure 1/160, aperture 11, ISO 100, focal length 55 mm. The wheat spike was placed on a blue background next to the X-Rite Mini ColorChecker Classic color palette card (http://xritephoto.com/colorchecker-targets). Images of the spike obtained both on a clothespin and on a table were used. An example of the images used for analysis is shown in Figure 1.

In total, the dataset included 1,745 digital images of wheat spikes (82 % of images were taken on a table and 18 % on a

clothespin). To evaluate the generalizability of the models, a separate holdout sample of 14 images obtained using the "on the table" protocol was used, which we did not use in either training or testing of the algorithms.

Marking spikelets in images

Three types of marking were used in the study: based on binary masks, on Gaussian masks, and marking based on bounding boxes. Initially, images were marked manually in ImageJ (Schneider et al., 2012). The centers of the spikelets were marked with dots on the image, and their coordinates were saved in a separate file for each image. Based on the coordinates of the spikelet centers, three options for generating markings for machine learning were applied (Fig. 2).

Annotation in the form of binary masks (Fig. 2*b*). A circular area with a radius of 2 mm was generated for each spikelet. In this case, we used an image segmentation algorithm: a neural network model was trained to predict areas in the image corresponding to binary masks.

Markup in the form of a Gaussian mask (Fig. 2*c*). It can be assumed that the centers of the spikelets marked manually do not correspond exactly to their geometric centers, but deviate randomly. We assumed that the density of this random arrangement of marks around the center has the form of a radial Gaussian distribution. Therefore, as a second marking option, a two-dimensional Gaussian distribution was generated, the center of which coincided with the marked center of the spikelet.

Bounding boxes (Fig. 2*d*). A square of fixed size (6 × 6 mm) was automatically generated around each manually marked spike center. The model was trained to detect these squares.

Model architectures and training

For training purposes, the images were randomly divided into training (~60 %), validation (~20 %), and test (~20 %) samples.

Two approaches were used to determine the centers of spikelets in the image using deep machine learning algorithms. The first involved segmenting the image pixels into two types: those belonging to the mask and those not belonging to the mask. For this, semantic image segmentation models based on the U-Net network (Ronneberger et al., 2015) were used. Two types of masks were used, as described above: binary and Gaussian. In the case of a binary mask, the predicted centers of the spikelets were considered to be the geometric centers of the areas corresponding to the predicted masks. In the case of a Gaussian mask, the center of the spikelet was defined as the pixel with the maximum probability within the area satisfying the condition: the ratio of the probability of each pixel to the maximum probability in a given area must not be lower than the threshold C.

The U-Net network (Ronneberger et al., 2015) can use layer encoders of different architectures. In our work, several encoder options were tested: efficientnet-b3, efficientnet-b4, mit_b2, mit_b1, timm-resnest26d, timm-regnetx_032, timm-res2next50, timm-gernet_m, timm-efficientnet-b4, timm-efficientnet-b3 (Tan, Le, 2019; Wightman, 2019; Radosavovic et al., 2020; Zhang H. et al., 2020; Gao et al., 2021; Xie et al., 2021).

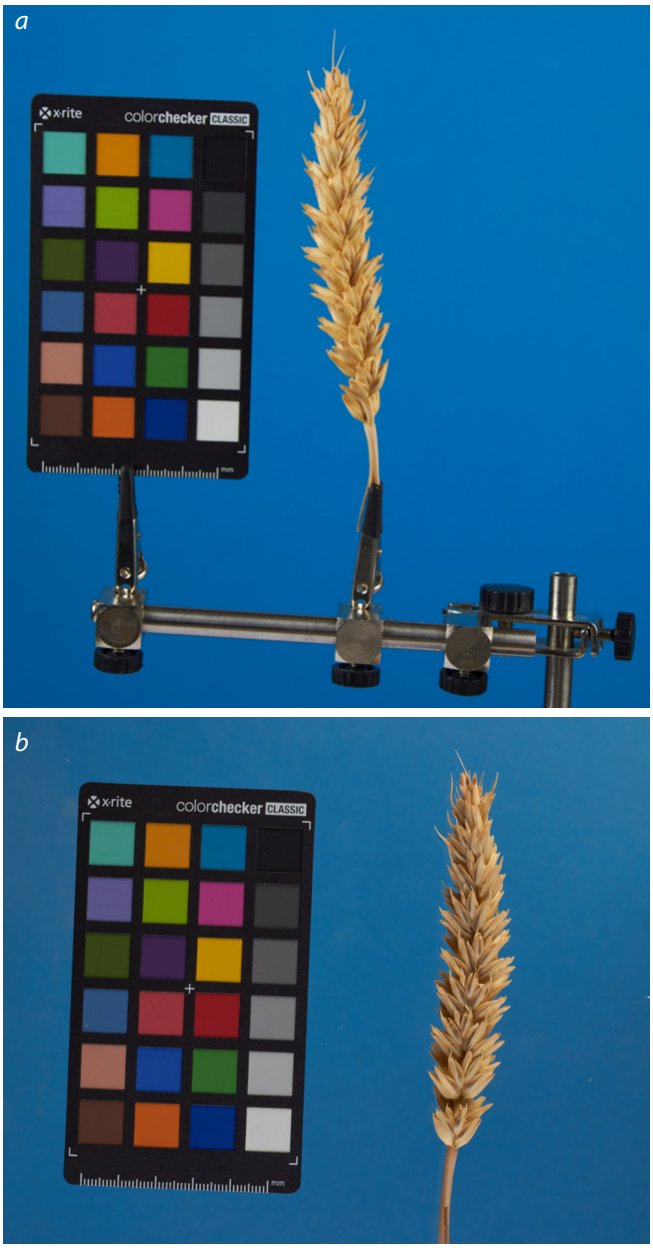


Fig. 1. Examples of images of a wheat spike attached to a clothespin (*a*) and placed on a table (*b*).

The second approach was to detect image regions bounded by squares and corresponding to the central regions of the spikelets. For this, the YOLO network architecture was used, which determines the bounding rectangle for each spikelet in the image. The YOLOv8m architecture (Redmon et al., 2016) was used for this approach.

Thus, we applied three methods to determine the centers: segmentation using the U-Net network for binary (hereinafter referred to as U-Net-BIN) and Gaussian (hereinafter referred to as U-Net-GAUSS) masks that bound the center position,

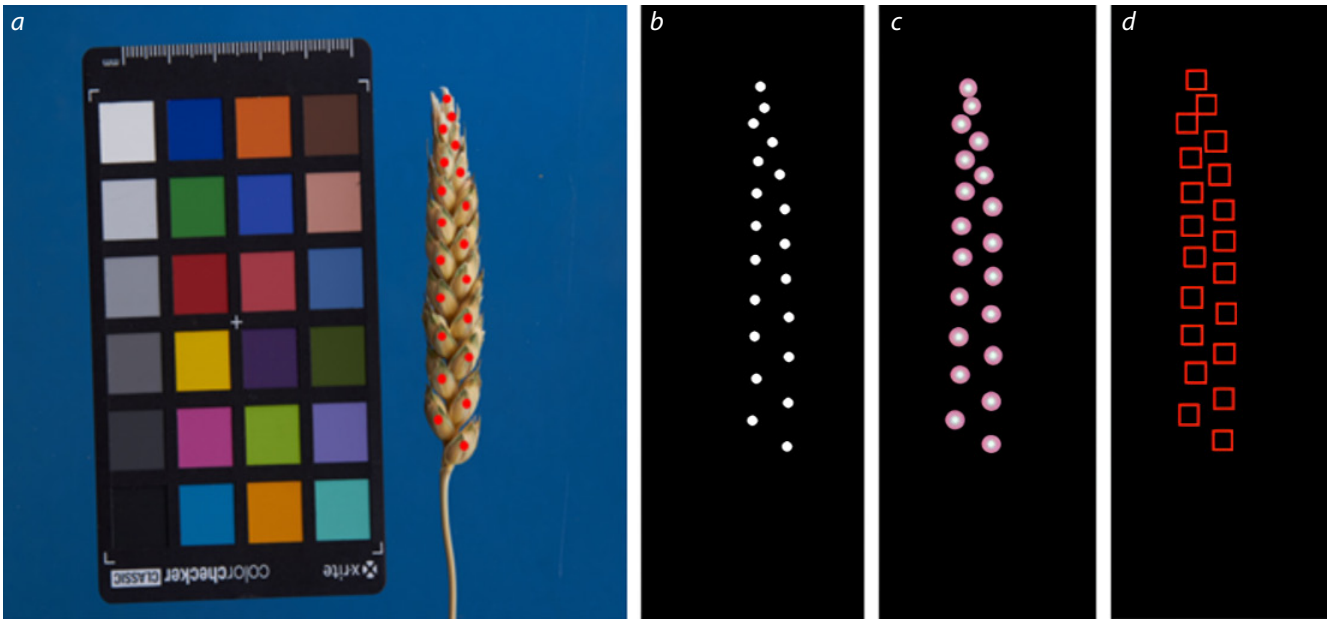


Fig. 2. Methods for marking spikelets based on their centers, indicated manually.

a – the original image of the spike with the centers of the spikelets marked with red dots; *b* – binary mask for spikelets in the form of circles, the centers of which coincide with the marking of spikelet centers; *c* – Gaussian mask, the distribution centers of which coincide with the centers of the marked spikelets, and the probability density has a radial distribution; *d* – a set of bounding boxes, the centers of which coincide with the centers of the spikelets.

and detection of the center area of the spikelet as a square using the YOLO architecture network (hereinafter referred to as YOLOv8).

Loss function type. For the algorithm with binary masks, the binary cross-entropy function (torch.nn.BCEWithLogitsLoss) was used. For Gaussian masks, the Kullback–Leibler divergence (torch.nn.KLDivLoss) was used as the loss function.

Organization of the training process. All three algorithms were trained for 500 epochs. The model weights were initialized based on pre-trained parameters obtained from the ImageNet dataset (Deng et al., 2009). The Adam algorithm was used as the parameter optimization method. During training, a set of augmentations implemented in the albumentations library (Buslaev et al., 2020) was used. The image was normalized to a size of 512×224 (Resize), then the following were applied: horizontal reflection with a probability of 0.5 (HorizontalFlip(p=0.5)); vertical reflection with a probability of 0.277 (VerticalFlip(p=0.277)); rotation by a random angle in the range −30…+30° with a probability of 0.735 (Rotate(limit=30, p=0.735)); Gaussian blur with a kernel size randomly selected in the range from 1 to 3 with a probability of 0.25 (GaussianBlur(blur_limit=(1,3), p=0.25)); addition of Gaussian noise with a probability of 0.15 (GaussNoise(p=0.15)); random brightness and contrast adjustment with a probability of 0.5 (RandomBrightnessContrast(p=0.5)); random change in the intensity of the R, G, and B channels in the range ±15 with a probability of 0.5 (RGBShift(r_shift_limit=15, g_shift_limit=15, b_shift_limit=15, p=0.5)); color transformations: random changes in brightness, contrast, saturation, and hue (ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2, hue=0.2, p=0.703)); conversion of the image to grayscale with a probability of 0.1 (ToGray(p=0.1)).

The neural network architectures/models we investigated for identifying wheat spikelet centers depended on a number of parameters. These included: the type of encoder architecture for the U-Net network, the radius of the circular area r for binary masks, the threshold C of the ratio of the probability of a pixel belonging to the spikelet center to the maximum probability in the case of a Gaussian mask, the parameters of the optimization algorithm, and some others. Their complete list and value ranges are given in Table 1.

During the process, the algorithms required the selection of the most optimal parameter values from Table 1. To do this, a Bayesian optimization algorithm implemented in the Optuna library (Akiba et al., 2019) was applied using images from the validation sample. The F1 accuracy metric described below was used as the target optimization parameter.

Algorithm accuracy metrics

Two types of metrics were used to evaluate the effectiveness of spike center detection. Quantitative metrics evaluated the error in counting the number of spikes in the spike cluster in the image. The mean absolute error (MAE) and mean absolute percentage error (MAPE) of spike counting were evaluated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i|,$$

$$\mathrm{MAPE} = \frac{100\,\%}{n}\sum_{i=0}^{n}\left|\frac{t_i - p_i}{t_i}\right|,$$

Table 1. Hyperparameters used in methods for determining the position of the center of spikelets in an image

Method	Hyperparameters	Value ranges
U-Net-BIN	Encoder architecture	efficientnet-b3, efficientnet-b4, mit_b2, mit_b1, timm-resnest26d, timm-regnetx_032, timm-res2next50, timm-gernet_m, timm-efficientnet-b4, timm-efficientnet-b3
	Gradient descent coefficient at the initial stage of optimization	[0.0001, 0.001]
	Gradient descent coefficient at the final stage of optimization	[0.000001, 0.00005]
	Binary mask radius (*r*, mm)	[0.5, 3]
U-Net-GAUSS	Encoder architecture	efficientnet-b3, efficientnet-b4, mit_b2, mit_b1, timm-resnest26d, timm-regnetx_032, timm-res2next50, timm-gernet_m, timm-efficientnet-b4, timm-efficientnet-b3
	Gradient descent coefficient at the initial stage of optimization	[0.0001, 0.001]
	Gradient descent coefficient at the final stage of optimization	[0.000001, 0.00005]
	Parameter σ for Gaussian distribution	[0.000002, 0.0002]
	Probability threshold for selecting the position of the spikelet center (*C*)	[1/9, 1/2]
YOLOv8	Gradient descent coefficient at the initial stage of optimization	[0.01, 0.00005]
	Gradient descent coefficient at the final stage of optimization	[0.000005, 0.000001]
	DropBlock parameter	[0, 0.5]

where t_i – the true number of spikelets in the image, counted manually; p_i – the number of spikelets determined based on machine learning methods; n – the number of images in the sample used to evaluate accuracy.

To evaluate the accuracy of determining the position of the center of the spikelets in the image, the metrics of precision, recall, and F1-measure were calculated.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{F}_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where TP (true positive) is the number of correct positive predictions of the position of the spikelet, FP (false positive) is the number of incorrect predictions of the position of the spikelet, and FN (false negative) is the number of false predictions of the position of the spikelet. The following rule was used to determine the TP, FP, and FN parameters (Fig. 3): the predicted center position was considered a true positive (TP) if it was within a radius of 2 mm from the true center; a false positive prediction (FP) if the spikelet center was predicted incorrectly; a false negative prediction (FN) if there were no predicted centers within a radius of 2 mm from the true center.

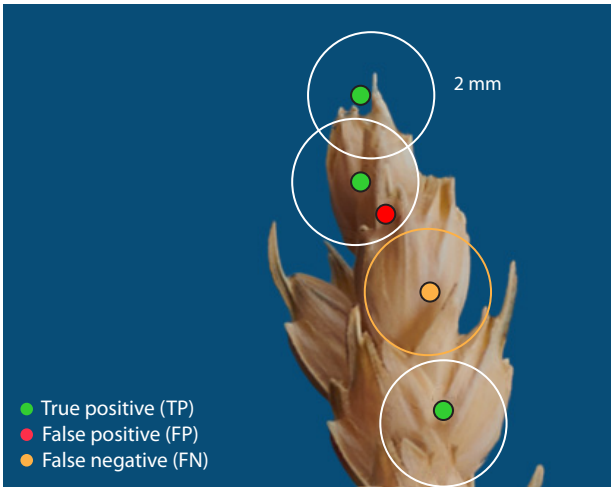


Fig. 3. Visualization of true positives (green dots), false positives (red dots), and false negatives (yellow dots) in the calculation of spatial metrics for assessing the quality of spikelet detection.

The programs we used were implemented in Python 3.11 using the PyTorch 2.3.0, OpenCV 4.9.0, and scikit-learn 1.5.0

libraries. The calculations were performed on a workstation with an AMD Ryzen 9 7950X processor, 64 GB of RAM, and an NVIDIA GeForce RTX 2080 Ti (11 GB) graphics accelerator.

Results and discussion

The optimal values of hyperparameters for three models predicting the position of spikelets in the image, selected using Bayesian optimization, are listed in Table 2.

The best encoder architecture for determining the center of the spikelet based on segmentation and a binary mask was MiT-B2 (Mix Transformer-B2) (Xie et al., 2021). This architecture combines convolutional layers with a self-attention mechanism, which allows it to simultaneously capture local features and global contextual dependencies in the image. This is especially important for the accurate positioning of small objects, such as spikelet centers.

For the Gaussian mask, the EfficientNet-B4 architecture (Tan, Le, 2019) showed the best results. The model uses a compound scaling method that balances the depth, width, and resolution of the input data. This approach provides a good balance between accuracy and computational efficiency, making the architecture well suited for density distribution regression tasks.

The evaluation results on the test sample (Table 3) showed that both U-Net-based approaches significantly outperform the YOLOv8-based method.

As shown in Table 3, both approaches based on the U-Net architecture demonstrated high and comparable accuracy in terms of both quantitative (MAE ~0.5) and spatial (F1 > 0.96) metrics on the test sample. The insignificant superiority of the U-Net-GAUSS model in terms of MAE (0.502 *vs.* 0.512) and the U-Net-BIN model in terms of F1 (0.965 *vs.* 0.962) allows us to consider them equally effective for data, the distribution of which corresponds to the training sample. At the same time, the YOLOv8 model showed significantly lower results (MAE = 3.641, F1 = 0.679), which indicates its unsuitability for solving the problem in the current configuration with automatically generated square bounding boxes.

On the deferred sample (Table 4), the U-Net-BIN model showed the best accuracy, while the U-Net-GAUSS model showed lower accuracy. The largest error was observed for the YOLOv8 model.

The results on the holdout sample, collected at a different time and having a slightly different distribution, demonstrate more pronounced differences between the methods. The U-Net-BIN model maintained high accuracy, showing only a slight increase in error (MAE increased from 0.512 to 0.538),

Table 2. Hyperparameters for neural network models for identifying the position of spikelet centers in an image, selected using Bayesian optimization

Model	Hyperparameters	Optimal values
U-Net-BIN	Encoder architecture	mit_b2
	Gradient descent coefficient at the initial stage of optimization	0.001
	Gradient descent coefficient at the final stage of optimization	0.000001
	Binary mask radius (*r*, mm)	1.5999
U-Net-GAUSS	Encoder architecture	efficientnet-b4
	Gradient descent coefficient at the initial stage of optimization	0.0001
	Gradient descent coefficient at the final stage of optimization	0.0000063723
	Parameter σ for Gaussian distribution	0.00001121
	Probability threshold for selecting the position of the spikelet center (*C*)	0.214564
YOLOv8	Gradient descent coefficient at the initial stage of optimization	0.00042
	Gradient descent coefficient at the final stage of optimization	0.000001
	DropBlock parameter	0.48

Table 3. Evaluation of the effectiveness of methods for determining the position of the spikelet center on the test sample

Model	MAE	MAPE	F1
U-Net-BIN	0.512	0.032	**0.965**
U-Net-GAUSS	**0.502**	0.032	0.962
YOLOv8	3.641	0.280	0.679

Table 4. Evaluation of the effectiveness of methods for determining the position of the spikelet center on a deferred sample (n = 14)

Model	MAE	MAPE	F1
U-Net-BIN	**0.538**	**0.035**	**0.972**
U-Net-GAUSS	0.846	0.062	0.942
YOLOv8	3.000	0.231	0.704

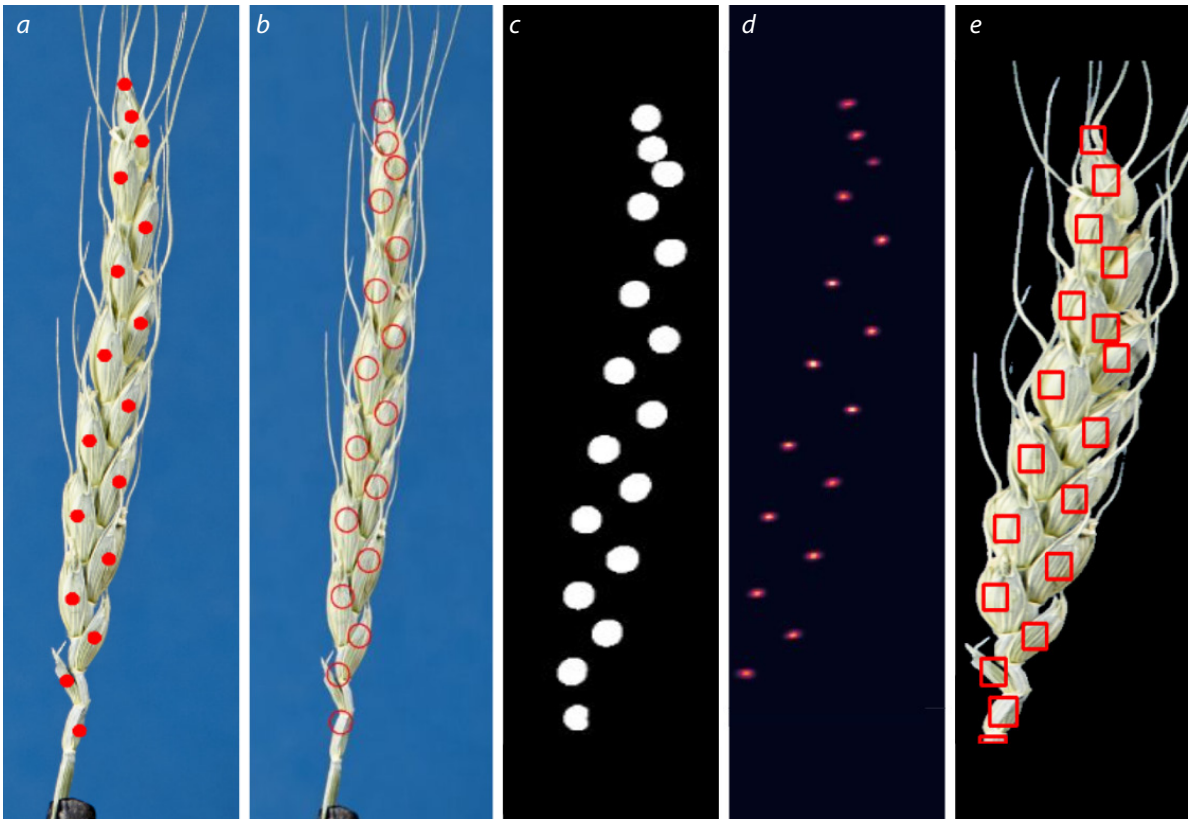


Fig. 4. Model results.

a – manually labeled centers of spikelets in a spike; *b* – the result of the model trained on binary masks superimposed on the actual image of the spike; *c* – the result of the model trained on binary masks; *d* – the result of the model trained on radial Gaussian distribution masks; *e* – the result of the model trained on bounding boxes.

which indicates its high generalizability and reliability. In contrast, the accuracy of the U-Net-GAUSS model on the holdout sample decreased significantly (MAE = 0.846 *vs.* 0.502 on the test sample). This suggests that this approach may be more sensitive to data variability. The YOLOv8 model, as in the test sample, showed the worst result (MAE = 3.0), further confirming the conclusion that the selected method of generating bounding boxes is not suitable for describing the shape of spikelets.

Both proposed methods based on U-Net showed good results on the test sample. However, the model trained on binary masks demonstrated significantly better generalization ability on the holdout sample, indicating its greater reliability for working with data with a different distribution. Regression analysis of the predicted and true number of spikelets showed a strong linear relationship for U-Net models ($R^2 > 0.95$ on the test sample), confirming the high accuracy of the count.

Measuring the inference speed on a single graphics accelerator showed that the processing time for a single image for the model trained on Gaussian masks is approximately 0.67 seconds, while for the model on binary masks it is approximately 0.64 seconds. Thus, the U-Net-BIN approach demonstrates not only higher accuracy but also slightly higher speed.

The low performance of YOLOv8 is probably due to the suboptimality of automatically generated square frames for describing the elongated and curved shape of spikelets (Fig. 4). This indicates that detection methods require more careful and, possibly, manual adjustment of the markup.

Conclusion

This study demonstrates that the use of simplified point labeling of spikelet centers in combination with modern deep learning architectures allows for accuracy comparable to the best modern methods using more complex and expensive labeling (Misra et al., 2020; Khoroshevsky et al., 2021; Zhou et al., 2021; Shi et al., 2023). The use of a simplified annotation scheme not only reduces the cost of experiments but also opens up the possibility of rapidly expanding datasets. The proposed approach can be adapted to solve other problems of counting morphological traits of plants and integrated into automated phenotyping systems to accelerate breeding programs.

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

Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA. ACM, 2019;2623-2631. doi 10.1145/3292500.3330701

Alkhudaydi T., Zhou J., De La Iglesia B. SpikeletFCN: counting spikelets from infield wheat crop images using fully convolutional networks. In: Artificial Intelligence and Soft Computing. ICAISC 2019. Lecture Notes in Computer Science. Vol. 11508. Springer, Cham., 2019;3-13. doi 10.1007/978-3-030-20912-4_1

Artemenko N.V., Epifanov R.U., Genaev M.A., Komyshev E.G., Kruchinina Yu.V., Koval V.S., Goncharov N.P., Afonnikov D.A. Image-based classification of wheat spikes by glume pubescence using convolutional neural networks. *Front Plant Sci.* 2024;14: 1336192. doi 10.3389/fpls.2023.1336192

Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A. Albumentations: fast and flexible image augmentations. *Information*. 2020;11(2):125. doi 10.3390/info11020125

Chen F., Pound M.P., French A.P. Learning to localise and count with incomplete dot-annotations. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada. IEEE, 2021;1612-1620. doi 10.1109/ICCVW54120.2021.00186

Deng J., Dong W., Socher R., Li L.J., Li K., Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009, Miami. IEEE, 2009;248-255. doi 10.1109/CVPR.2009.5206848

Efimov V.M., Rechkin D.V., Goncharov N.P. Multivariate analysis of long-term climate data in connection with yield, earliness and the problem of global warming. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2024;28(2):155-165. doi 10.18699/vjgb-24-18

Gao S.-H., Cheng M.-M., Zhao K., Zhang X.-Y., Yang M.-H., Torr P. Res2Net: a new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021;43(2): 652-662. doi 10.1109/TPAMI.2019.2938758

Genaev M.A., Komyshev E.G., Khao F., 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 J Genet Breed*. 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

Khoroshevsky F., Khoroshevsky S., Bar-Hillel A. Parts-per-object count in agricultural images: solving phenotyping problems via a single deep neural network. *Remote Sens*. 2021;13(13):2496. doi 10.3390/rs13132496

Konopatskaia I., Vavilova V., Blinov A., Goncharov N.P. Spike morphology genes in wheat species (*Triticum* L.). *Proc Latv Acad Sci B Nat Exact Appl Sci*. 2016;70(6):345-355. doi 10.1515/prolas-2016-0053

Li Q., Cai J., Berger B., Okamoto M., Miklavcic S.J. Detecting spikes of wheat plants using neural networks with Laws texture energy. *Plant Methods*. 2017;13:83. doi 10.1186/s13007-017-0231-1

Liu T., Chen W., Wang Y., Wu W., Sun C., Ding J., Guo W. Rice and wheat grain counting method and software development based on Android system. *Comput Electron Agric*. 2017;141:302-309. doi 10.1016/j.compag.2017.08.011

Maslova G., Abdryayev M.R., Sharapov I.I., Sharapova Ju.A. Correlation analysis of yield and elements of productivity of winter soft wheat varieties in the arid conditions of steppe zone of Middle Volga region. *Izvestiya of Samara Scientific Center of the Russian Academy of Sciences.* 2018;20(2-4):680-683. doi 10.24411/1990-5378-2018-00153 (in Russian)

Misra T., Arora A., Marwaha S., Chinnusamy V., Rao A., Jain R., Narayan R., Ray M., Kumar S., Raju D., Ranjan R., Nigam A., Goel S. SpikeSegNet – a deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods*. 2020;16:40. doi 10.1186/s13007-020-00582-9

Qiu R., He Y., Zhang M. Automatic detection and counting of wheat spikelet using semi-automatic labeling and deep learning. *Front Plant Sci*. 2022;13:872555. doi 10.3389/fpls.2022.872555

Radosavovic I., Kosaraju R.P., Girshick R., He K., Dollár P. Designing network design spaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, Seattle, WA, USA. IEEE, 2020;10428-10436. doi 10.1109/CVPR42600.2020.01044

Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, Las Vegas, NV, USA. IEEE, 2016;779-788. doi 10.1109/CVPR.2016.91

Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Vol. 9351. Springer, 2015;234-241. doi 10.1007/978-3-319-24574-4_28

Savin V.Yu. Determination of the force required for stripping the wheat ear. *Inzhenernyye Tekhnologii i Sistemy = Engineering Technologies and Systems*. 2019;29(3):456-466. doi 10.15507/2658-4123.029.201903.456-466 (in Russian)

Schneider C.A., Rasband W.S., Eliceiri K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671-675. doi 10.1038/nmeth.2089

Shi L., Sun J., Dang Y., Zhang S., Sun X., Xi L., Wang J. YOLOv5s-T: a lightweight small object detection method for wheat spikelet counting. *Agriculture*. 2023;13(4):872. doi 10.3390/agriculture13040872

Skripka O.V., Samofalov A.P., Podgornyî S.V., Gromova S.N. Productivity and main elements of productivity of winter wheat varieties of intensive type, bred in the All-Russian Research Institute of Grain Crops. *Dostizheniya Nauki i Tekhniki APK = Achievements of Science and Technology in Agro-Industrial Complex*. 2016;30(9): 30-32 (in Russian)

Tan M., Le Q.V. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA. ICML, 2019;6105-6114

Vahamidis P., Karamanos A.J., Economou G. Grain number determination in durum wheat as affected by drought stress: an analysis at spike and spikelet level. *Ann Appl Biol*. 2019;174(2):190-208. doi 10.1111/aab.12487

Vavilova V., Konopatskaia I., Kuznetsova A.E., Blinov A., Goncharov N.P. Genomic characterization of *DEP1* gene in wheats with normal and compact spike shape. *BMC Genetics*. 2017;18 (Suppl.1): 106. doi 10.1186/s12863-017-0583-6

Vavilova V.Yu., Konopatskaia I.D., Blinov A.G., Goncharov N.P. Interspecific polymorphism of *DEP1* genes and the spike shape in wheats. *Russ J Genet.* 2019;55(7):908-913. doi 10.1134/S1022795419070147

Wightman R. PyTorch Image Models (timm). *GitHub Repository*. 2019. doi 10.5281/zenodo.4414861

Xie E., Wang W., Yu Z., Anandkumar A., Alvarez J., Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. *arXiv*. 2021. doi 10.48550/arXiv.2105.15203

Zhang B., Liu X., Xu W., Chang J., Li A., Mao X., Zhang X., Jing R. Novel function of a putative *MOC1* ortholog associated with spikelet number per spike in common wheat. *Sci Rep*. 2015;5(1):12211. doi 10.1038/srep12211

Zhang H., Wu C., Zhang Z., Zhu Y., Zhang Z., Lin H., Sun Y., He T., Mueller J., Manmatha R., Li M., Smola A. ResNeSt: split-attention networks. *arXiv*. 2020. doi 10.48550/arXiv.2004.08955

Zhou H., Riche A.B., Hawkesford M.J., Whalley W.R., Atkinson B.S., Sturrock C.J., Mooney S.J. Determination of wheat spike and spikelet architecture and grain traits using X-ray Computed Tomography imaging. *Plant Methods.* 2021;17(1):26. doi 10.1186/s13007-021-00726-5

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

Received September 10, 2025. Revised October 23, 2025. Accepted November 1, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):36-42

doi 10.18699/vjgb-26-14

Different patterns of *P* transposon and *blood* retrotransposon distribution in Harwich and Canton-S sub-strains do not affect the manifestation of *Drosophila melanogaster* intraspecific PM hybrid dysgenesis

L.P. Zakharenko, Y.Y. Ilinsky

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
zakharlp@bionet.nsc.ru

Abstract. Intraspecific infertility, the nature of which is not always understood, occurs in many eukaryotes. Intraspecific PM hybrid dysgenesis (PM HD) in *Drosophila melanogaster* manifests in one cross direction as offspring infertility and other genetic disorders due to incompatibility between the maternal cytoplasm and the paternal genome. PM HD is believed to result from a massive transposition of the *P*-element when the maternal cytoplasm lacks a repressor to block it. In this work, we have investigated the distribution of the *P* transposon and *blood* retrotransposon in the reference PM HD strains (Canton-S and Harwich), which have been maintained in different laboratories for several decades. *P*-element distribution patterns vary among Harwich sub-strains, indicating that the *P*-element was translocated in these genomes. The rate of movement of the *P*-element, which was not induced by crosses, is comparable to the rate of movement of other DNA transposons. The distribution pattern of the low-active *blood* retrotransposon in Harwich sub-strains is more stable than that of the *P*-element, indicating genetic relatedness between sub-strains. Derivatives of the *P*-element detected in some Canton-S sub-strains possibly indicate genetic contamination. The significant difference in the *blood* transposable element distribution pattern in Canton-S sub-strains also indicates genetic heterogeneity among them. Despite the complex genealogy of the studied sub-strains, including cases of possible genetic contamination, and differences in *P*-element distribution, the ability to express PM HD symptoms is preserved in the studied sub-strains.
Key words: PM hybrid dysgenesis; *Drosophila melanogaster*; *P*-element; *blood* retrotransposon

For citation: Zakharenko L.P., Ilinsky Y.Y. Different patterns of *P* transposon and *blood* retrotransposon distribution in Harwich and Canton-S sub-strains do not affect the manifestation of *Drosophila melanogaster* intraspecific PM hybrid dysgenesis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):36-42. doi 10.18699/vjgb-26-14

Funding. The work was supported by the Budget Project FWNR-2022-0019 from the Ministry of Science and Higher Education of the Russian Federation.

Acknowledgements. The authors thank the Shared Facility Center for Microscopic Analysis of Biological Objects SB RAS, Novosibirsk and Dr. O. Ignatenko for technical assistance and valuable comments. The authors thank the reviewers for their constructive comments.

Паттерн распределения *P*-транспозона и *blood* ретротранспозона в отводках референсных линий Harwich и Canton-S из разных лабораторий не влияет на проявление внутривидового PM гибридного дисгенеза у *Drosophila melanogaster*

Л.П. Захаренко, Ю.Ю. Илинский

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
zakharlp@bionet.nsc.ru

Аннотация. Внутривидовое бесплодие, природа которого не всегда понятна, встречается у многих эукариот. Внутривидовой PM гибридный дисгенез (PM ГД) у *Drosophila melanogaster* проявляется в одном из направлений скрещивания в бесплодии потомства и некоторых других генетических нарушениях в результате несовместимости между материнской цитоплазмой и отцовским геномом. Считается, что PM ГД является результатом массового перемещения *P*-элемента, если материнская цитоплазма не имеет репрессора, блокирующего его перемещение. В данной работе мы исследовали распределение *P*-элемента в отводках референсных для PM ГД линиях (Canton-S и Harwich), которые долгое время содержались в разных лабораториях. Паттерны распределения *P*-элемента различались в разных отводках Harwich. Скорость перемещения *P*-элемента, не индуцированная скрещиваниями, сопоставима со скоростью перемещения других ДНК-транспозонов. Паттерн

© Zakharenko L.P., Ilinsky Y.Y., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

распределения малоактивного ретротранспозона *blood* в производных линии Harwich был более стабильным по сравнению с *P*-элементом, что свидетельствует о генетическом родстве между отводками. В некоторых отводках линии Canton-S были обнаружены следы *P*-элемента, что может указывать на генетическое загрязнение. Значительная разница в паттерне распределения транспозона *blood* в отводках линии Canton-S также говорит о генетической гетерогенности отводков. Несмотря на сложную генеалогию отводков исследованных линий, включающую случаи возможного генетического загрязнения, и различия в распределении *P*-элементов, способность проявлять симптомы РМ ГД у исследованных отводков сохраняется. Мы показали, что все изученные производные Canton-S имеют сильный M-цитотип, а все отводки Harwich – индуцирующий P-цитотип.

Ключевые слова: РМ гибридный дисгенез; *Drosophila melanogaster*; *P*-элемент; ретротранспозон *blood*

Introduction

Progeny of some *Drosophila melanogaster* suffer from the hybrid dysgenesis (HD) phenomenon that occurs in one cross direction (Kidwell et al., 1977) and results in female sterility, male recombination, high mutation rate and other genetic disorders (Bingham et al., 1982). It is commonly considered that HD is caused by massive *P*, *hobo* or *I* element transpositions in PM, HE or IR HD, respectively (Bingham et al., 1982; Bucheton et al., 1984; Blackman et al., 1987).

It is a commonly held opinion that *P*-elements appeared in the *D. melanogaster* genome by horizontal transfer in the middle of the last century. In American populations, *P*-elements were first identified in *D. melanogaster* obtained from nature in 1938, from Russian populations, in 1966, and from French populations, in 1967 (Anxolabehere et al., 1988). However, all *D. melanogaster* strains collected on different continents from 1969 contain *P*-elements in their genomes (Anxolabehere et al., 1988). Thus, the distribution of the *P*-elements throughout the world was extremely rapid despite possible harm.

PM is the most common variant of HD, manifesting with all symptoms listed above. It was postulated to occur in crosses between females with the M cytotype, which lack *P*-elements in their genome, and males that harbor a functional *P*-element (Bingham et al., 1982). However, it was later found that many M cytotype strains contained a complete *P*-element (Bingham et al., 1982), and the presence of a *P*-element in the genome did not guarantee the induction of HD (Itoh et al., 1999, 2001, 2007). For this reason, the nature of PM HD is studied mainly on the reference Canton-S (**M**aternal cytotype) and Harwich (**P**aternal cytotype) strains. Crossing these strains results in HD phenotypes with 100 % probability. Not every strain with a full-length *P*-element in its genome induces HD, and none of them consistently exhibits HD as strongly as the Harwich strain. This fact alone casts doubt on the hypothesis of the *P*-element's involvement in HD induction. A correlation was found between the strains' ability to induce PM HD and the presence of a highly truncated *P*-element variant named "*Har-P*" in the Harwich genome, which is the most frequent *de novo* insertion. However, genomic location may also influence host tolerance of *Har-P*s, as a significant rescue of viable pupae was observed in crosses with *Har-P*s located only on Chromosome 3, which has six *P*-elements, while no pupae survived with *Har-P*s located on Chromosome 2, which has nine *P*-elements (Srivastav et al., 2019). Thus, the reason for the uniqueness of the reference P strains remains not fully understood. According to a widely accepted hypothesis, the *P*-element mobilization is repressed in P cytotype strains by small RNA (Brennecke et al., 2008; Khurana et al., 2011).

The Canton-S strain has been widely used among genetic scientists since 1925. However, *P*-elements have been identified in some Canton-S strains (Ignatenko et al., 2015), possibly as a result of genetic contamination during cultivation in laboratory. "*A golden rule of Drosophila genetics is never trust a stock label, no matter how reputable the source from which it was obtained*" (Ashburner, 1989).

Here, we analyze the pattern of *P* transposon and *blood* retrotransposon distribution in Canton-S and Harwich sub-strains from different laboratories and their ability to produce dysgenic progeny to describe accumulated genetic differences. We selected *blood* retrotransposon, as its distribution pattern remained stable for two decades in spite of the presence of a complete *blood* copy in the isogenic reference y; cn bw sp genome (Ignatenko et al., 2015). According to our data, the rate of movement of the *P*-element in Harwich sub-strains without dysgenic crosses is comparable to the rate of movement of other DNA transposons. All Canton-S derivatives have a strong reactive M cytotype, and all Harwich sub-strains have an inducer P cytotype.

Methods

Strains of *Drosophila melanogaster*. Five Canton-S (CS) and three Harwich (H) sub-strains from different laboratories were used for the study (see the Table). The H2 sub-strain has a visible, spontaneously occurring mutation in the *sepia* gene (66 D5), and the H3 sub-strain has a mutation in the *white* gene (3 B6). There is no information in the literature about the artificial origin of these mutations. The LK-P (1A) strain containing one complete *P*-element per genome (from S. Ronsseray) was used to obtain a DNA probe of a full-sized *P*-element. Reference strain y; cn bw sp (No. 2057, Bloomington Drosophila collection) with a completely sequenced genome was used to obtain a *blood* probe (see the Table). The Harwich strain was taken from nature (Harwich, Massachusetts) in 1967 by M.L. Tracey (Kidwell et al., 1977). According to S. Ronsseray, the Harwich strain descended from two females (Ronsseray et al., 1984). Bloomington Drosophila Stock Center received the Harwich strain (stock number 4264) from M. Kidwell in 1997.

Hybrid dysgenesis assays. To test HD sensitivity, sub-strains were crossed to CS2 or H3, a cross between which results in high penetrance of HD (Ignatenko et al., 2015). Progeny from crosses between H3 males and females of each Canton-S sub-strain, as well as CS2 female and males of each Harwich sub-strains, were cultivated at 29 °C from the egg

Strains description

Strain	Resource	Year of strain appearance in ICG, Novosibirsk
CS1	Bloomington Drosophila Stock Center (BDSC) #ND	2013
CS2	ICG, I. Zakharov from Obninsk	1990
CS3	ICG, B. Chadov	1990
CS4	ICG, N. Gruntenko	2013
CS5	CNRS-Universite´ Pierre et Marie Curie, France, S. Ronsseray	2013
Harwich-1	BDSC #4264	2013
Harwich-2	France, S. Ronsseray	2013
Harwich-3	ICG, I. Zakharov from L. Kaidanov	1996
LK-P (1A)	France, S. Ronsseray	2013
y; cn bw sp	BDSC #2057	2006

stage to adulthood. As a marker of PM HD, ovary morphology was analyzed in 50 females aged 4–5 days (Kidwell, Novy, 1979).

Fluorescence *in situ* hybridization (FISH). The *P*-element probe was obtained by PCR with a primer complementary to terminal inverted repeats 5′-TGATGAAATAACATAAGGTG GTCCCGTCG-3′ (Takasu-Ishikawa et al., 1992; Lapie et al., 1993), and the *blood* probe, with primers 5′-CAGTGGCATAC GCTTCAAGA-3′ and 5′-GGTTCGCGAAATACCAGTGT-3′ (Ignatenko et al., 2015). The probes were labeled by nick-translation with digoxygenin-dUTP for the *P*-element, and with biotin-dUTP for *blood*. FISH analysis was performed for 3–5 larvae for each sub-strain as in Ignatenko et al. (2015). Squashed preparations of polytene chromosomes from the salivary glands of third-instar larvae were made according to standard procedures (Biémont et al., 1994). The hybridization sites were analyzed using an Axio Imager M1 (Carl Zeiss) microscope in the Shared Facility Center for Microscopic Analysis of Biological Objects SB RAS, Novosibirsk.

Results

To test sub-strains for PM HD, they were crossed to CS2 or H3, a cross between which results in high penetrance of HD (Ignatenko et al., 2015). Females of the CS2 sub-strain were crossed to H1, H2 or H3 males to check Harwich sub-strains induction ability. Females of five Canton-S sub-strains were crossed to H3 males to check the ability of Canton-S sub-strains to respond to induction. Indeed, abnormal gonad morphology was observed in progeny of all dysgenic crosses with 100 % penetrance, indicating a strong P cytotype for Harwich and M cytotype for Canton-S sub-strains (Table S1)[1]. Opposite crosses (female Harwich to Canton-S males) produced normal progeny (Table S1).

Blood element pattern in Harwich and Canton-S sub-strains

The pattern of the *blood* retrotransposon was studied to determine the level of similarity among the sub-strains. The reason for choosing the *blood* TE was its low-level mobility observed in the isogenic reference strain y; cn bw sp. The localization of the *blood* retrotransposon in the y; cn bw sp strain has not changed significantly over two decades (1992–2013) (Ignatenko et al., 2015). As expected, unrelated strains (H3, CS4 and y; cn bw sp) display mostly unique *blood* hybridization sites (89 %), and no hybridization sites were found to be common to all three strains. Pairwise matching sites were rare (Fig. 1, Tables S2, S4).

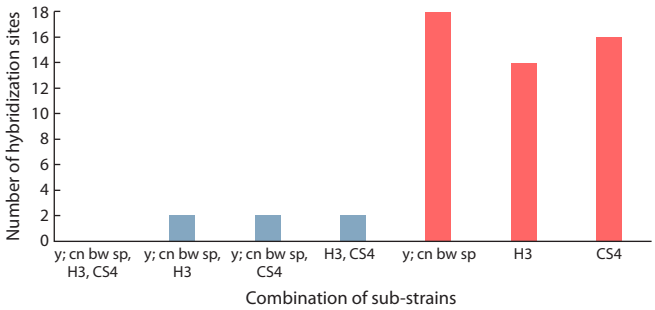


Fig. 1. Distribution of common and unique *blood* hybridization sites in unrelated strains.

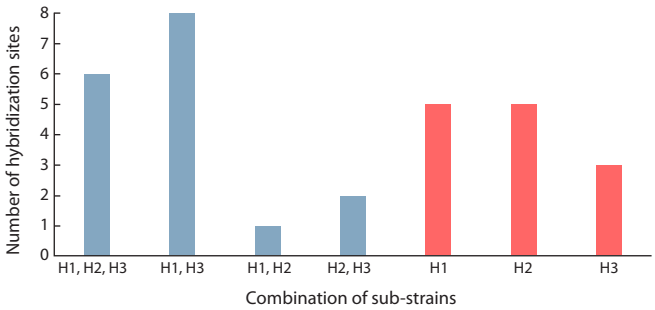


Fig. 2. Distribution of common and unique *blood* hybridization sites in Harwich sub-strains.

Thirty two *blood* hybridization sites were found in the genomes of Harwich sub-strains: 6 sites common for all three sub-strains, 15 unique sites, and 11 sites common for different pairs (Fig. 2). Thus, 20 % of *blood* hybridization sites are common to all three sub-strains (H1, H2, H3), and 47 % are unique.

[1] Supplementary Tables S1–S4 and Figs S1–S3 are available at: https://vavilovj-icg.ru/download/pict-2026-30/appx2.pdf

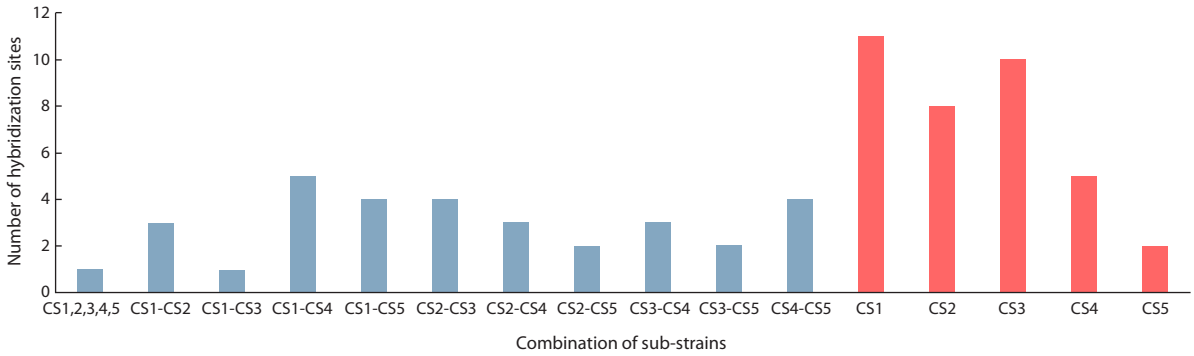


Fig. 3. Distribution of common and unique *blood* hybridization sites in Canton-S sub-strains.

The H1 and H3 sub-strains displayed the greatest similarity, harboring 56 % (14 out of 25) common hybridization sites. It can thus be assumed that of the three investigated Harwich sub-strains, at least H1 and H3 sub-strains have common origin.

Strikingly, the *blood* distribution pattern shows extreme variability among Canton-S sub-strains. Of the total 60 hybridization sites, most are unique (60 %). One site is shared by all five sub-strains, two sites are shared by three sub-strains, and pairwise coincidences range from one to six common sites between different pairs (Fig. 3). Consequently, genetic contamination of at least some sub-strains cannot be excluded. Additional evidence for genetic contamination comes from the presence of traces of the *P*-element in some of these Canton-S sub-strains (Ignatenko et al., 2015) despite the fact that classically Canton-S serves as the gold standard for being *P*-element-negative. It is interesting that traces of the *P*-element (two sites per genome) were found only on the third chromosome and were localized in different regions of the CS1, CS2, CS4 genomes (Ignatenko et al., 2015). Notably, *P*-element appeared in the *D. melanogaster* genome later than the Canton-S strain was isolated from nature. Rahman with coauthors (2015) also reported introgression of *P*-element-containing lab strains into certain stocks of Oregon R, which should be free of *P*-elements originally (Robertson, Engels, 1989).

P-element pattern in Harwich sub-strains

We identified 85 *P*-element hybridization sites in Harwich sub-strains, of which only 5 % were common for all three sub-strains (Fig. 4, Table S3, Figs S1–S3). The percentage of unique *P*-element hybridization sites is 33 % for H1, 69 % for H2 and 28 % for H3 (Fig. 5). This result shows much higher variability in *P*-element distribution compared to *blood* distribution among the Harwich sub-strains. Moreover, common *P*-element hybridization sites, for example in 3F regions, show weak signal in the H3 sub-strain, suggesting a possible difference in fine structure of *P*-elements in different sub-strains (Fig. 4). Although FISH is a semi-quantitative method, the use of a full-length copy of the *P*-element as a probe allows interpreting the intensity of the signal as a reflection of the size of the sequence with which the probe hybridizes on the chromosome. Thus, the difference between sub-strains in the distribution of hybridization sites may be greater than we assume.

The H2 sub-strain harbors the largest number of unique *P*-element sites (44 %), which is consistent with *blood* distribution in this sub-strain also being the most distinct from the other two Harwich sub-strains in pairwise comparisons. In turn, the H1 and H3 sub-strains, which show the greatest similarity in *blood* distribution, shared 35 % (16 out of 47)

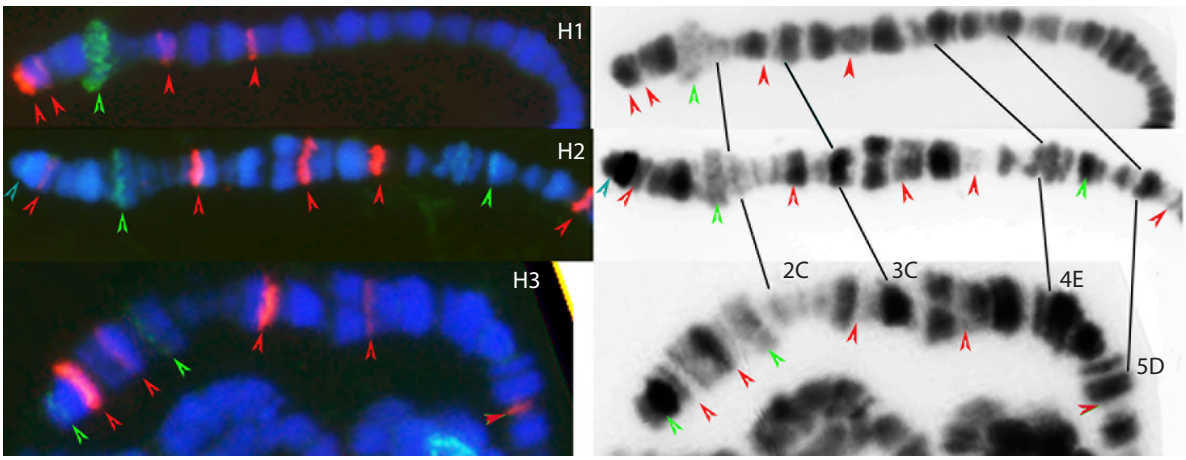


Fig. 4. Localization of *P*-element (red) and *blood* TE (green) hybridization sites on the end of the X chromosomes of Harwich sub-strains from different laboratories. DAPI (blue).

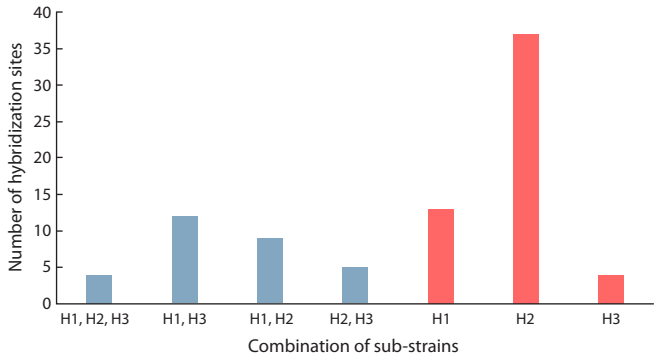


Fig. 5. Distribution of common and unique *P*-element hybridization sites in Harwich sub-strains.

of *P*-element hybridization sites. Thus, although the Harwich sub-strains are more heterogeneous in *P*-element distribution compared to *blood* distribution, a closer relationship between the H1 and H3 sub-strains is suggested by distribution patterns of both retrotransposons. Besides, the H2 and H3 sub-strains have visible genetic markers that help to monitor the purity of the strains. H3 has a mutation in the *white* gene, H2, in the *sepia* gene. The origin of the *sepia* and *white* mutations has not been precisely determined. The H3 strain was obtained from L. Kaidanov in 1996. Similarities in *P*-element distributions at the end of the X chromosomes of different Harwich derivatives support the spontaneous occurrence of the *white* mutation at least (Fig. 4).

Three Canton-S sub-strains (CS1, CS2, CS4) contain *P*-element derivatives as a result of possible genetic contamination. Given that derivatives of *P*-element were found only on the third chromosome, these sub-strains may share a common origin and may have been distributed between laboratories after *P*-element contamination. In terms of *blood* distribution, the CS1 and CS4 sub-strains show the highest degree of similarity among all pairwise comparisons. CS3 and CS5 have no traces of the *P*-element in the genome; this circumstance makes them similar.

Discussion

We found that all Canton-S sub-strains have strong M cytotypes and all Harwich sub-strains exhibit a strong P cytotype independent of possible genetic contamination and a different pattern of *P*-element distribution. A comparison of the Harwich sub-strains indicates that the *P*-element can change its position without crossing with M strains. This is similar to the genomic instability reported for *I*-elements, which also occurs in the absence of dysgenic crosses (Moschetti et al., 2010). The symptoms of IR (Inducer-Reactive) HD differ from those of PM HD: dysgenic females lay eggs, but the offspring die at the embryonic stage. In both cases, the symptoms are consistently reproduced with the same physiological manifestations typical of a given HD type and specific reference strains. All other things being equal, only a limited number of strains can induce HD in both cases.

The Harwich strain was separated from nature (Harwich, Massachusetts) in 1967 by M.L. Tracey (Kidwell et al., 1977). The peak of activity in studying the PM HD phenomenon and the spread of the strain across laboratories occurred in the 90s of the last century (Anxolabehere et al., 1984, 1988; Wang et al., 1993; Simmons et al., 1996; Ronsseray et al., 1998). During this time, at least 300 generations of Drosophila have changed. The H1 and H3 sub-strains differ in 11 out of 25 common *blood* hybridization sites, and in 31 out of 47 *P*-element sites. In our case, the rate of transposition can be estimated as at least as 15×10^{-4} (11/25/300) per site per genome per generation and 22×10^{-4} (31/47/300) for *blood* and for *P*-element, respectively.

According to the FISH analyses, the rate of TE transpositions in natural populations varies by several orders of magnitude, ranging from 10^{-5} per copy per generation to 10^{-2} per copy per generation (Pasyukova et al., 1998; Maside et al., 2000; Díaz-González et al., 2011). Sequencing analysis of parents and offspring from 18 families of full-sib *D. melanogaster* estimated the range of TE insertion rate from 10^{-3} per copy per generation to 10^{-5} per copy per generation (Wang et al., 2023). By comparing the parents and offspring in each family, authors identified 89 new TE insertions across the 89 samples, making it only one insertion per genome per generation on average (Wang et al., 2023). The rate of movement of the *P*-element in our experiment is in the range typical for the movement of other TEs.

Whole genome sequencing was used to estimate the rate of *P*-element movement in dysgenic ovaries and germ cells (Moon et al., 2018; Jansen et al., 2024).

Contrary to predictions based on the insertional mutagenesis model of hybrid dysgenesis, single-cell whole-genome sequencing analysis of DNA from dysgenic and non-dysgenic embryos at late embryonic stages (before dysgenic germ cell death) shows that dysgenic and non-dysgenic germ cells acquire unexpectedly similar, low numbers of new heterozygous *P*-element insertions (Jansen et al., 2024). The authors found double-strand breaks in generative cells and proposed that transposase excises the *P*-element, but does not integrate it into the genome (Jansen et al., 2024). To confirm their assumption, the authors simulate the excision of the *P*-element by using CRISPR Cas9. However, the gaps with blunt ends, induced by CRISPR Cas9, are usually completed by a non-homologous path (Quétier, 2016), whereas transposase recognizes the inverted terminal ends of the *P*-element and cuts them with the formation of sticky ends, which are successfully ligated or completed by reparation enzymes using the homolog. The excision of the *P*-element can occur precisely, without loss of gene functionality (Weinert et al., 2005). Double-strand breaks in dysgenic germline cells can be induced not only by *P*-element translocation.

Moon with coauthors investigated the rate of *P*-element movement in the ovaries of dysgenic females raised at 18 °C, when symptoms of hybrid dysgenesis do not appear and the ovaries have normal morphology (Moon et al., 2018). Nevertheless, the authors found 527 new *P*-element insertion sites in dysgenic ovaries. However, only a single new *P*-element insertion was found in F1 and F2 dysgenic progeny (Eggleston et al., 1988). It means that cells with a huge number of new *P*-elements should be eliminated, but the ovaries of dysgenic females raised at 18 °C, paradoxically, have normal morpho-

logy. Thus, disputable results were obtained about the mass movement of *P*-elements during PM HD.

It should be noted that Moon with coauthors do not see the movement of other TE families in the dysgenic cross direction (Moon et al., 2018).

The Harwich genome contains 132 *P*-elements according to Moon, and 80 (half of which are heterozygous) according to Jansen with coauthors (Moon et al., 2018; Jansen et al., 2024). On the one hand, the annotation of sequenced genomes with repeats is still imperfect and the difference can be explained by a reading error. For example, Moon with coauthors (2018) admit that 34 new insertions per genome may be false positives. On the other hand, it is possible that the issue is not only in the accuracy of the sequencing method or the choice of tissue for sequencing but also that the Harwich sub-strains, although having the same origin, were cultivated in different laboratories for a long time. Jansen with coauthors used the Harwich strain from BDSC. According to Ronsseray, this strain descended from two females since 1967 (Ronsseray et al., 1984). The initial polymorphism of the strain may affect the heterogeneity of the sub-strains. Moon (Moon et al., 2018) received the Harwich strain from Khurana et al. (2011), Khurana (Khurana et al., 2011) received the Harwich strain from Ronsseray. Thus, sub-strains originate from the same source, but have been cultured independently for a long time.

According to our data, Harwich sub-strains from different laboratories also differ in the number and localization of *P*-element hybridization sites. The sub-strain obtained from Ronsseray has more *P*-element hybridization sites than the sub-strains from BDSC. This is less than that revealed by sequencing, because the sensitivity of the FISH method depends on the specific activity of the labeled probe. In addition, weak signals may be missed during FISH analysis. Nevertheless, according to our data, relationship between the Harwich sub-strains is undeniable, since the percentage of common hybridization sites between the Harwich sub-strains is higher than between obviously unrelated strains.

PM HD may result not only from possible massive *P*-element movement, but also from an incompatibility of temperature-dependent metabolic processes of sensitive to PM HD strains, as dysgenic flies, when reared at low temperatures (20 °C), do not display dysgenic symptoms (Engels, Preston, 1979; Dorogova et al., 2017). Additionally, there is a hormonal difference between dysgenic and non-dysgenic flies (Zakharenko et al., 2014). The reference PM HD Harwich and Canton-S strains also differ in a number of physiological characteristics: life expectancy, fertility, locomotor activity, development rate (Zakharenko et al., 2024). The asymmetry in the expression of the dysgenic traits, characteristic of interspecific hybrids, suggests a significant genetic distance between the reference PM HD strains not only in *P*-element absence/presence.

Conclusion

The relationship between Harwich sub-strains is undeniable despite the difference in *P*-element distribution pattern. The rate of movement of the *P*-element in Harwich sub-strains without induction by crossing is in the range typical for the movement of other TEs. All Canton-S sub-strains have strong M cytotypes and all Harwich sub-strains exhibit a strong P cytotype independently of possible genetic contamination and a different pattern of *P*-element distribution.

References

Anxolabehere D., Kai H., Nouaud D., Périquet G., Ronsseray S. The geographical distribution of *P-M* hybrid dysgenesis in *Drosophila melanogaster*. *Genet Sel Evol.* 1984;16(1):15-26. doi 10.1186/1297-9686-16-1-15

Anxolabehere D., Kidwell M.G., Periquet G. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P*-elements. *Mol Biol Evol*. 1988;5(3):252-269. doi 10.1093/oxfordjournals.molbev.a040491

Ashburner M. Drosophila. A Laboratory Handbook. New York: Cold Spring Harbor Laboratory Press, 1989

Biémont C., Lemeunier F., Garcia Guerreiro M.P., Brookfield J.F., Gautier C., Aulard S., Pasyukova E.G. Population dynamics of the copia, mdg1, mdg3, gypsy, and P transposable elements in a natural population of *Drosophila melanogaster*. *Genet Res*. 1994;63(3):197-212. doi 10.1017/s0016672300032353

Bingham P.M., Kidwell M.G., Rubin G.M. The molecular basis of P-M hybrid dysgenesis: the role of the *P*-element, a P-strain-specific transposon family. *Cell*. 1982;29(3):995-1004. doi 10.1016/0092-8674(82)90463-9

Blackman R.K., Grimaila R., Koehler M.M., Gelbart W.M. Mobilization of hobo elements residing within the decapentaplegic gene complex: suggestion of a new hybrid dysgenesis system in *Drosophila melanogaster*. *Cell.* 1987;49(4):497-505. doi 10.1016/0092-8674(87)90452-1

Brennecke J., Malone C.D., Aravin A.A., Sachidanandam R., Stark A., Hannon G.J. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science.* 2008;322(5906):1387-1392. doi 10.1126/science.1165171

Bucheton A., Paro R., Sang H.M., Pelisson A., Finnegan D.J. The molecular basis of I-R hybrid dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. *Cell.* 1984;38(1):153-163. doi 10.1016/0092-8674(84)90536-1

Díaz-González J., Vázquez J.F., Albornoz J., Domínguez A. Long-term evolution of the *roo* transposable element copy number in mutation accumulation lines of *Drosophila melanogaster*. *Genet Res.* 2011;93(3):181-187. doi 10.1017/S0016672311000103

Dorogova N.V., Bolobolova E.U., Zakharenko L.P. Cellular aspects of gonadal atrophy in *Drosophila* P-M hybrid dysgenesis. *Dev Biol*. 2017;424(2):105-112. doi 10.1016/j.ydbio.2017.02.020

Eggleston W.B., Johnson-Schlitz D.M., Engels W.R. P-M hybrid dysgenesis does not mobilize other transposable element families in *D. melanogaster*. *Nature.* 1988;331(6154):368-370. doi 10.1038/331368a0

Engels W.R., Preston C.R. Hybrid dysgenesis in *Drosophila melanogaster*: the biology of female and male sterility. *Genetics*. 1979;92(1):161-174. doi 10.1093/genetics/92.1.161

Ignatenko O.M., Zakharenko L.P., Dorogova N.V., Fedorova S.A. *P*-elements and the determinants of hybrid dysgenesis have different dynamics of propagation in *Drosophila melanogaster* populations. *Genetica*. 2015;143(6):751-759. doi 10.1007/s10709-015-9872-z

Itoh M., Woodruff R.C., Leone M.A., Boussy I.A. Genomic *P*-elements and P-M characteristics of eastern Australian populations of *Drosophila melanogaster*. *Genetica*. 1999;106(3):231-245. doi 10.1023/a:1003918417012

Itoh M., Sasai N., Inoue Y., Watada M. *P*-elements and P-M characteristics in natural populations of *Drosophila melanogaster* in the southernmost islands of Japan and in Taiwan. *Heredity* (*Edinb*). 2001;86:206-212. doi 10.1046/j.1365-2540.2001.00817.x

Itoh M., Takeuchi N., Yamaguchi M., Yamamoto M.T., Boussy I.A. Prevalence of full-size *P* and *KP* elements in North American popu-

lations of *Drosophila melanogaster*. *Genetica*. 2007;131(1):21-28. doi 10.1007/s10709-006-9109-2

Jansen G., Gebert D., Kumar T.R., Simmons E., Murphy S., Teixeira F.K. Tolerance thresholds underlie responses to DNA damage during germline development. *Genes Dev*. 2024;38(13-14):631-654. doi 10.1101/gad.351701.124

Khurana J.S., Wang J., Xu J., Koppetsch B.S., Thomson T.C., Nowosielska A., Li C., Zamore P.D., Weng Z., Theurkaufet W.E. Adaptation to *P* element transposon invasion in *Drosophila melanogaster*. *Cell.* 2011;147(7):1551-1563. doi 10.1016/j.cell.2011.11.042

Kidwell M.G., Novy J.B. Hybrid dysgenesis in *Drosophila melanogaster*: sterility resulting from gonadal dysgenesis in the P-M system. *Genetics.* 1979;92(4):1127-1140. doi 10.1093/genetics/92.4.1127

Kidwell M.G., Kidwell J.F., Sved J.A. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics.* 1977;86(4):813-833. doi 10.1093/genetics/86.4.813

Lapie P., Nasr F., Lepesant J.A., Deutsch J. Deletion scanning of the regulatory sequences of the *Fbp1* gene of *Drosophila melanogaster* using *P* transposase-induced deficiencies. *Genetics.* 1993;135(3): 801-816. doi 10.1093/genetics/135.3.801

Maside X., Assimacopoulos S., Charlesworth B. Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. *Genet Res.* 2000;75:275-284. doi 10.1017/S0016672399004474

Moon S., Cassani M., Lin Y.A., Wang L., Dou K., Zhang Z.Z. A robust transposon-endogenizing response from germline stem cells. *Dev Cell.* 2018;47(5):660-671.e3. doi 10.1016/j.devcel.2018.10.011

Moschetti R., Dimitri P., Caizzi R., Junakovic N. Genomic instability of I elements of *Drosophila melanogaster* in absence of dysgenic crosses. *PLoS One*. 2010;5(10):e13142. doi 10.1371/journal.pone.0013142

Pasyukova E.G., Nuzhdin S.V., Filatov D.A. The relationship between the rate of transposition and transposable element copy number for *copia* and *Doc* retrotransposons of *Drosophila melanogaster*. *Genet Res.* 1998;72(1):1-11. doi 10.1017/S0016672398003358

Quétier F. The CRISPR-Cas9 technology: closer to the ultimate toolkit for targeted genome editing. *Plant Sci*. 2016;242:65-76. doi 10.1016/j.plantsci.2015.09.003

Rahman R., Chirn G.W., Kanodia A., Sytnikova Y.A., Brembs B., Bergman C.M., Lau N.C. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res*. 2015;43(22):10655-10672. doi 10.1093/nar/gkv1193

Robertson H.M., Engels W.R. Modified *P* elements that mimic the P cytotype in *Drosophila melanogaster*. *Genetics*. 1989;123(4):815-824. doi 10.1093/genetics/123.4.815

Ronsseray S., Anxolabéhère D., Périquet G. Hybrid dysgenesis in *Drosophila melanogaster*: influence of temperature on cytotype determination in the P-M system. *Mol Gen Genet*. 1984;196(1):17-23. doi 10.1007/BF00334086

Ronsseray S., Marin L., Lehmann M., Anxolabéhère D. Repression of hybrid dysgenesis in *Drosophila melanogaster* by combinations of telomeric *P*-element reporters and naturally occurring *P* elements. *Genetics*. 1998;149(4):1857-1866. doi 10.1093/genetics/149.4.1857

Simmons M.J., Raymond J.D., Grimes C.D., Belinco C., Haake B.C., Jordan M., Lund C., Ojala T.A., Papermaster D. Repression of hybrid dysgenesis in *Drosophila melanogaster* by heat-shock-inducible sense and antisense *P*-element constructs. *Genetics.* 1996;144(4): 1529-1544. doi 10.1093/genetics/144.4.1529

Srivastav S.P., Rahman R., Ma Q., Pierre J., Bandyopadhyay S., Lau N.C. *Har-P*, a short *P*-element variant, weaponizes *P*-transposase to severely impair *Drosophila* development. *eLife*. 2019;8:e49948. doi 10.7554/elife.49948

Takasu-Ishikawa E., Yoshihara M., Hotta Y. Extra sequences found at *P*-element excision sites in *Drosophila melanogaster*. *Mol Gen Genet*. 1992;232:17-23. doi 10.1007/BF00299132

Wang Y., Balter H., Levitan M., Margulies L. Mutability, sterility and suppression in P-M hybrid dysgenesis: the influence of P subline, cross, chromosome, sex and *P*-element structure. *Genet Res*. 1993; 62(2):111-123. doi 10.1017/s0016672300031700

Wang Y., McNeil P., Abdulazeez R., Pascual M., Johnston S.E., Keightley P.D., Obbard D.J. Variation in mutation, recombination, and transposition rates in *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res*. 2023;33(4):587-598. doi 10.1101/gr.277383.122

Weinert B.T., Min B., Rio D.C. P element excision and repair by nonhomologous end joining occurs in both G_1 and G_2 of the cell cycle. *DNA Repair* (*Amst*). 2005;4(2):171-181. doi 10.1016/j.dnarep.2004.09.004

Zakharenko L.P., Karpova E.K., Rauschenbach I.Y. P-M hybrid dysgenesis affects juvenile hormone metabolism in *Drosophila melanogaster* females. *Russ J Genet*. 2014;50:772-774. doi 10.1134/S1022795414060143

Zakharenko L.P., Bobrovskikh M.A., Gruntenko N.E., Petrovskii D.V., Verevkin E.G., Putilov A.A. Two old wild-type strains of *Drosophila melanogaster* can serve as an animal model of faster and slower aging processes. *Insects.* 2024;15(5):329. doi 10.3390/insects15050329

Conflict of interest. The authors declare no conflict of interest.
Received July 24, 2025. Revised December 12, 2025. Accepted December 15, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):43-52

doi 10.18699/vjgb-26-06

Soybean locus qDTF-7 as an example of genetic heterogeneity associated with flowering and maturity time

R.N. Perfil'ev [1], M.I. Shmatova [1], A.B. Shcherban [1, 2], E.A. Salina [1, 2]

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

Abstract. Genome-wide association studies (GWAS) have become a standard approach for identifying quantitative trait loci associated with diverse phenotypic traits. Further investigation of the locus – specifically, the search for the causal gene and mutation – may present various challenges. One of the challenges is genetic heterogeneity (or locus heterogeneity), when alleles from different closely located genes can influence the same trait. Recently, using GWAS, we found the qDTF-7 locus on soybean chromosome 3, which is associated with flowering time under Novosibirsk conditions. Initially, we identified *GmTOE1*, an ortholog of *TOE1* (*TARGET OF EAT1*), a known flowering-time regulator in Arabidopsis, as the most likely candidate gene for this locus. Four major haplotypes were identified in *GmTOE1*, which are associated with soybean flowering and maturity and are likely to provide soybean adaptation to northern latitudes. However, this gene showed only a very weak association with soybean flowering in the Novosibirsk region compared to the Oryol region, suggesting the presence of another gene within the locus that influences flowering time. We therefore re-analyzed genes in the qDTF-7 locus and identified *GmRVE8c*, an *Arabidopsis RVE8* (*REVEILLE 8*) ortholog, located ~21 kb upstream of *GmTOE1*; *RVE8* is a circadian clock component involved in plant development. After studying the natural variation of the *GmRVE8c* genes, we found four major haplotypes that arose due to three nonsynonymous substitutions and one 19-bp deletion leading to a frameshift. To identify three haplotypes, $GmRVE8c^{hap1,3,4}$, which are predominant in improved soybean cultivars, we developed DNA markers. Using these markers, we genotyped 129 soybean accessions, the developmental time of which had been studied in the Novosibirsk and Oryol regions. Using our data and data from SoyOmics, we found the $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$ haplotypes to be associated with late flowering and maturity in soybean. The early-maturing haplotype $GmRVE8c^{hap1}$ is predominant in cultivars from northern regions and is likely associated with the adaptation of soybean to high latitudes. The $GmRVE8c^{hap4}$ haplotype is in complete linkage with the early-maturing allele $GmTOE1^{C}$, whereas the $GmRVE8c^{hap3}$ haplotype shows strong linkage with the late maturing allele $GmTOE1^{T}$. Furthermore, the ANOVA results indicate an interaction between *GmRVE8c* and *E1*, the major regulator of flowering in soybean. This interaction is manifested as a stronger effect of the $GmRVE8c^{hap3,4}$ haplotypes on flowering and maturity in the genetic background of the *e1-as* allele compared with *E1*. Together, these findings define a complex and intriguing locus, which may serve as a possible example of a genetically heterogeneous locus.
Key words: soybean; flowering; *TOE1*; *RVE8*; genetic heterogeneity

For citation: Perfil'ev R.N., Shmatova M.I., Shcherban A.B., Salina E.A. Soybean locus qDTF-7 as an example of genetic heterogeneity associated with flowering and maturity time. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):43-52. doi 10.18699/vjgb-26-06

Funding. This research was funded by the Russian Science Foundation (RSF project No. 21-76-30003-П). The design of markers and their verification was carried out within the framework of a budget project FWNR-2025-0033.

Локус сои qDTF-7 как пример генетической гетерогенности, ассоциированной с продолжительностью цветения и созревания

Р.Н. Перфильев [1], М.И. Шматова [1], А.Б. Щербань [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 ИЦиГ СО РАН, Новосибирск, Россия
PerfilyevRN@bionet.nsc.ru

Аннотация. В настоящее время полногеномный анализ ассоциаций (GWAS) стал стандартным подходом для идентификации локусов количественных признаков, ассоциированных с различными фенотипическими признаками. При последующем изучении локуса, т. е. поиске гена и мутации, которые влияют на признак, можно столкнуться с разными проблемами. Одна из проблем – генетическая (или локусная) гетерогенность, ситуация, когда аллели из разных близко расположенных генов в одном локусе влияют на один признак. Ранее с помощью

© Perfil'ev R.N., Shmatova M.I., Shcherban A.B., Salina E.A., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

GWAS на хромосоме 3 сои мы обнаружили локус qDTF-7, ассоциированный с продолжительностью цветения в условиях Новосибирска. Первоначально для данного локуса в качестве наиболее вероятного гена-кандидата мы определили *GmTOE1*, ортолог *TOE1* (*TARGET OF EAT 1*), известного регулятора цветения у арабидопсиса. У *GmTOE1* были обнаружены четыре основных гаплотипа, которые ассоциированы с цветением и созреванием сои и, вероятно, связаны с адаптацией сои к северным широтам. Однако этот ген демонстрировал очень слабую ассоциацию с цветением сои в Новосибирской области по сравнению с Орловской, что указывало на наличие в составе локуса другого гена, который может влиять на цветение сои. Повторно проанализировав гены в составе локуса qDTF-7, мы примерно на 21 т.п.н. выше гена *GmTOE1* обнаружили *GmRVE8c*, ортолог гена *RVE8* (*REVEILLE 8*), который вовлечен в процессы развития у арабидопсиса в качестве компонента циркадных ритмов. Изучив естественное нуклеотидное разнообразие *GmRVE8c*, мы выявили четыре основных гаплотипа, которые возникли из-за трех однонуклеотидных замен и одной делеции длиной 19 п.н., приводящей к сдвигу рамки считывания. Для определения трех гаплотипов $GmRVE8c^{hap1, 3, 4}$, преобладающих в современных сортах сои, мы разработали ДНК-маркеры. С помощью этих маркеров мы генотипировали 129 сортообразцов сои, для которых ранее изучили продолжительность развития в условиях Новосибирской и Орловской областей. В результате с помощью наших сведений и данных из SoyOmics мы обнаружили ассоциацию гаплотипов $GmRVE8c^{hap3}$ и $GmRVE8c^{hap4}$ с поздним цветением и созреванием сои. Раннеспелый гаплотип $GmRVE8c^{hap1}$ преобладает в сортообразцах из северных регионов и, вероятно, связан с адаптацией сои к северным широтам. Гаплотип $GmRVE8c^{hap4}$ полностью сцеплен с раннеспелым аллелем $GmTOE1^{C}$, а гаплотип $GmRVE8c^{hap3}$ сильно сцеплен с позднеспелым аллелем $GmTOE1^{T}$. Кроме того, результат ANOVA показывает наличие взаимодействия *GmRVE8c* с главным регулятором цветения сои – геном *E1*. Данное взаимодействие выражается более сильным эффектом гаплотипов $GmRVE8c^{hap3, 4}$ на цветение и созревание на генетическом фоне аллеля *e1-as* в сравнении с *E1*. Все это формирует достаточно комплексный и интересный локус, который может выступать как возможный пример генетически гетерогенного локуса.

Ключевые слова: соя; цветение; *TOE1*; *RVE8*; генетическая гетерогенность

Introduction

Advances in high-throughput plant phenotyping and genotyping methods, as well as the development of mathematical models that enable the identification of links between phenotype and genotype, have led to the discovery of numerous quantitative trait loci in plants (Zhang et al., 2022). The flowering and maturity time of soybean is no exception, and genes controlling these traits have been identified using GWAS (Genome-Wide Association Studies), for example: *Tof5* (Dong et al., 2022), *Tof8* (Li H. et al., 2023a), *Tof13* (Li H. et al., 2023b), *Tof16* (Dong et al., 2021), *qFT13-3* (Li Y.F. et al., 2024), *GmAP1d* (Guo et al., 2024). Thus, the *Tof16* locus was found to encode an ortholog of the *Arabidopsis CCA1* (*CIRCADIAN CLOCK ASSOCIATED 1*) gene, which is a key component of the circadian clock and a transcription factor containing a DNA-binding MYB domain (Dong et al., 2021). Recessive *tof16* alleles that partially or completely lose the original function of the encoded protein delay flowering and maturity time in soybean, which is more favorable for cultivars grown in southern latitudes (Dong et al., 2021).

A total of 54 CCA1-Like proteins have been identified in the soybean genome (Bian et al., 2017), four of which belong to the same phylogenetic clade and show the highest similarity to the *Arabidopsis* genes *RVE4* (*REVEILLE 4*) and *RVE8* (Bian et al., 2017; Shan et al., 2021; Bao et al., 2024). Loss of *RVE8* function in *Arabidopsis* reduces flowering time under both short-day (SD) and long-day (LD) conditions, whereas overexpression delays flowering (Rawat et al., 2011). In addition to reduced flowering time, the *rve4 6 8* and *rve3 4 5 6 8 Arabidopsis* mutants exhibit accelerated development and increased cell size due to disrupted expression of *PIF4* (*PHYTOCHROME INTERACTING FACTOR 4*) and *PIF5*, as well as reduced proteasomes activity (Gray et al., 2017; Scandola et al., 2022). The functional role of *RVE* genes in flowering and maturity time in soybean is still poorly understood (Shan et al., 2021), although only recently it was shown that overexpression of *GmMYB133* (the gene most closely related to *Arabidopsis RVE5*) and *GmRVE8a* in *Arabidopsis* leads to delayed flowering (Shan et al., 2021; Bao et al., 2024). Interestingly, overexpression of *GmPIF4b* in soybean accelerates not only flowering but also subsequent developmental stages (Arya et al., 2021). In soybean, a *GmMYB133-PRR5-PIF4* module has also recently been proposed, as *GmMYB133* in *Arabidopsis* can bind to the *PRR5* (*PSEUDO-RESPONSE REGULATOR 5*) promoter, and *PRR5* can, in turn, bind to *PIF4* (Shan et al., 2021).

At present, genes and mutations affecting traits have been identified for only a relatively small number of quantitative trait loci in plants. Several challenges interfere with fine-mapping, that is, the identification of the gene and mutation within a locus that influence the expression of a trait (Clauw et al., 2024). One of the challenges is allelic heterogeneity, a situation in which different mutations in a single gene lead to the same or similar phenotype (Sasaki et al., 2021; Clauw et al., 2024; Liu H.-J. et al., 2024). One of the classic examples in plants is the multiple alleles of the *FRI* (*FRIGIDA*) gene in *Arabidopsis*, which have a confirmed effect on the expression of *FLC* (*FLOWERING LOCUS C)* but are not detected by GWAS unless these alleles are included as covariates in the statistical model (Atwell et al., 2010). Another less-studied challenge is genetic heterogeneity (or locus heterogeneity) – a situation in which alleles from different closely located genes at the same locus influence the same trait (Sasaki et al., 2021; Clauw et al., 2024; Liu H.-J. et al., 2024). Significantly fewer examples of genetic heterogeneity are currently known. Recently, E. Sasaki et al. (2021) characterized such a locus in *Arabidopsis*, showing that the association of the *AOP2*/*AOP3* (*ALKENYL HYDROXALKYL PRODUCING*) genes with

flowering time can, in fact, be explained by two neighboring genes, *NDX1* (*NODULIN HOMEOBOX GENES 1*) and *GA1* (*GIBBERELLIC ACID REQUIRING 1*). A similar locus was also found in soybean, where pod shattering resistance was shown to arise from mutations in two nearby genes, *Sh1* and *Pdh1* (Li S. et al., 2024).

In this article, we discuss a possible example of a genetically heterogeneous locus in soybean. By reanalyzing genes within the soybean flowering-associated qDTF-7 locus, we identified a new candidate gene, *GmRVE8c*, located ~21 kb upstream of another strong candidate gene, *GmTOE1*. Mutations in *GmRVE8c* were found to form distinct haplotypes associated with soybean flowering and maturity. Furthermore, the effects of *GmRVE8c* haplotypes on flowering and maturity time depend on the allelic state of *E1*, the main regulator of flowering in soybean. Altogether, this forms a complex and intriguing locus, and our work may serve as a theoretical example of the natural mutations underlying associations identified through GWAS.

Materials and methods

Plant materials and phenotypes. As plant material, we used 129 soybean accessions (63 cultivars and 66 breeding lines), which were described and used in our previous works (Perfil'ev et al., 2023, 2024). For these cultivars, the duration from emergence to flowering (DTF, days from emergence to flowering) and to maturity (DTM, days from emergence to maturity) was previously studied in Novosibirsk and Oryol in 2021 and 2022. The details of this field experiment were described in a previous work (Perfil'ev et al., 2023). We used BLUP (Best Linear Unbiased Prediction) values as phenotypes for these 129 accessions, which were calculated separately for each region, as described previously (Perfil'ev et al., 2024).

All available observations for BBD (Beginning bloom date) and MD (Pod maturity date) were downloaded from SoyOmics (https://ngdc.cncb.ac.cn/soyomics/index) (Liu Y. et al., 2023). BLUP values for these phenotypes were calculated using the "lmer" function from the lme4 package (Bates et al., 2015), with genotype, location, and year included as random effects in the model. The resulting BLUP values were used for further analyses.

Characterization of the qDTF-7 locus and natural variation in the *GmRVE8c* gene. The list of genes located within the qDTF-7 locus was obtained using SoyBase (https://www.soybase.org/).

Data on haplotypes of the *GmRVE8c/SoyZH13_03G159500* gene within the coding sequence were obtained using the "HapSnap" module from SoyOmics. The settings in "HapSnap" were kept at default, except for the "Variation type" option, which excluded "Synonymous" and "Unclassed" variation.

The ModiDB database was used to search for the main protein domains of *GmRVE8c*/*Glyma.03G177300* (https://mobidb.org/) (Piovesan et al., 2025).

Co-localization of qDTF-7 with previously identified QTLs was examined using the "BreedingTips" module from the SoyOmics website.

DNA extraction, development of DNA markers, genotyping. Genomic DNA was extracted from 3–4-day-old seedlings grown in Petri dishes using a CTAB buffer (Rogers, Bendich, 1985).

To develop DNA markers for the *GmRVE8c*/*Glyma.03G177300* gene, the nucleotide sequence obtained from the Ensembl Plants database was used (https://plants.ensembl.org/index.html). UGENE v48.0 software was used for primer design (Okonechnikov et al., 2012) (http://ugene.net/ru/). The nucleotide sequences of the primers are provided in the Table.

PCR was performed using BioMaster HS-Taq PCR-Color (2×) (BiolabMix, Novosibirsk, Russia). The mixture with a total volume of 25 µl contained 10 mM Tris-HCl, pH 8.5, 50 mM KCl, 0.1 % Tween 20, 2 mM $MgCl_2$, 0.25 mM of each primer, 50–100 ng of genomic DNA, and 1 U Taq DNA polymerase. The primers used are presented in the Table. PCR conditions: 95 °C, 5 min; (95 °C, 15 s; 55 °C, 10 s; 72 °C, 20 s) for 34 cycles; 72 °C, 1 min.

Digestion of PCR products using the restriction enzyme Mhl I (SibEnzyme, Novosibirsk, Russia) was carried out in a reaction mixture with a total volume of 20 µL, which contained: 8 µl of PCR product; 2 µl 10x buffer recommended by the manufacturer; 1 U enzyme; 10 µl ddH20. Restriction and amplification products were separated on a 2–2.5 % agarose gel with ethidium bromide. Electrophoresis results were photographed using Gel Doc™ XR+ (BioRad, USA). The "Step100+50" DNA ladder (BiolabMix, Novosibirsk, Russia) was used as a molecular size marker.

Genotypes of the *E1* and *GmTOE1* genes for 129 soybean accessions were previously obtained and published (Perfil'ev et al., 2023, 2024). For improved cultivars from the SoyOmics database, genotypes for the mutations soy8699425 for the *E1*/*SoyZH13_06G195900* gene and soy4989324 for the *GmTOE1*/*SoyZH13_03G159700* gene were obtained using the "HapSnap" module, and for the mutation soy8699425, they were additionally recoded as C → *e1-as*, G → *E1* allele.

DNA markers used to determine the main haplotypes of *GmRVE8c* (*Glyma.03G177300/SoyZH13_03G159500*)

Gene	5′–3′ direction	Primer sequences	Haplotypes	Expected amplification/restriction products, bp	Restriction enzyme
GmRVE8c	F	AGATAGGATATGATACCATATATTG	hap1–3	173	–
	R	CGGACCAGCTCTCCCTTGAC	hap4	154	
GmRVE8c	F	AAGAACGGGGCAGTAGCACACG	hap2, 3	251	Mhl I
	R	TTCATCCCATGTGACATACCCA	hap1, 4	221+30	

Statistical analysis. The association of *GmRVE8c* haplotypes with flowering and maturity time was analyzed using the basic function "aov" in R version 4.4.2 (https://www.r-project.org/). Multiple haplotype comparisons were performed using the "TukeyHSD" function. Boxplots and heatmaps illustrating the results of multiple comparisons were generated using the R packages "ggplot2" (Wickham, 2016) and "ggcorplot" (Kassambara, 2023).

Results

Revisiting the qDTF-7 locus identified a new candidate gene, *GmRVE8c*

Previously, using the FarmCPU multilocus model on soybean chromosome 3, we identified a QTN (Gm03_40959110_A_G) that was associated with flowering time in Novosibirsk conditions (Perfil'ev et al., 2024). This QTN was located within a 553 kb LD block, which was designated as the qDTF-7 locus (QTL days from emergence to flowering 7) (Perfil'ev et al., 2024). This locus contains 71 genes (Table S1 in Supplementary Materials)[1], and we initially identified *Glyma.03G177500* (hereafter referred to as *GmTOE1*) as the most likely candidate gene (Perfil'ev et al., 2024). However, this gene exhibited a much weaker association with soybean flowering in Novosibirsk compared to Oryol (Perfil'ev et al., 2024). Therefore, we decided to revisit the genes within this locus more closely and, as a result, found *Glyma.03G177300*. This gene encodes one of four soybean homologs that are most closely related to the *Arabidopsis* genes *RVE8* (*REVEILLE 8*) and *RVE4* (Bao et al., 2024). In this article, the *Glyma.03G177300* gene will be referred to as *GmRVE8c*.

In *Arabidopsis*, the *RVE8* gene is a known component of the circadian clock and also functions to delay flowering (Rawat et al., 2011). Overexpression of one of the soybean *RVE8* orthologs (*GmRVE8a*) in *Arabidopsis* leads to delayed flowering (Bao et al., 2024). Collectively, these findings suggest that *GmRVE8c* is a candidate gene for regulating flowering and maturity in soybean.

Association of *GmRVE8c* haplotypes with flowering and maturity of soybean

Using the SoyOmics database, we studied natural variation in the coding region of the *GmRVE8c* gene. As a result, we found four haplotypes, which arose due to one deletion causing a frameshift and three nonsynonymous substitutions (Table S2). This deletion is predicted to cause a frameshift after four amino acids, likely leading to a complete loss of the gene's original function. All nonsynonymous substitutions found are located outside the predicted MYB DNA-binding domains of the protein (Fig. S1). It is difficult to assume whether these substitutions have consequences for the structure and function of the encoded protein.

We studied the association of three common haplotypes with BLUP_BBD and BLUP_MD in improved soybean cultivars. In cultivated varieties, according to the results of the association analysis for BLUP_BBD, cultivars carrying $GmRVE8c^{hap3}$ flower significantly later than those carrying $GmRVE8c^{hap1}$ and $GmRVE8c^{hap4}$ (Fig. 1*a*). According to the results of the association analysis for BLUP_MD, cultivars carrying $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$ mature significantly later than those carrying $GmRVE8c^{hap1}$ (Fig. 1*a*).

To identify these three haplotypes, we developed DNA markers for the soy32494699 and soy4988787 mutations (see the Table, Fig. S2) and used them to genotype 129 soybean accessions from our collection, in which the qDTF-7 locus was initially identified. As a result, $GmRVE8c^{hap1}$ was found in 109 accessions, while the remaining 12 and 8 carried haplotypes $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$, respectively (Table S3). In addition, *GmRVE8c* shows a significant association with BLUP_DTF and BLUP_MD under the conditions of the Novosibirsk and Oryol regions, and accessions carrying the $GmRVE8c^{hap4}$ haplotype flower and mature significantly later (Fig. 1*b*, *c*).

In cultivars Graciya and VNIIS-1, we previously identified the $GmTOE1^{T}$ allele (Perfil'ev et al., 2024); however, genotyping for *GmRVE8c* revealed the $GmRVE8c^{hap4}$ haplotype in these cultivars, which was unexpected, since $GmRVE8c^{hap4}$ is linked to the $GmTOE1^{C}$ allele. It should be noted that the *GmRVE8c* gene is physically located very close to *GmTOE1*, making the occurrence of crossover genotypes $GmRVE8c^{hap4}$ $GmTOE1^{T}$ extremely unlikely. For example, no such genotype was found in the SoyOmics database. Therefore, we re-extracted DNA from these cultivars and genotyped them using a previously developed DNA marker for the $GmTOE1^{C}$ allele (Perfil'ev et al., 2024), confirming the presence of the $GmTOE1^{C}$ allele in these cultivars (Table S3).

The *GmRVE8c* gene exhibits an epistatic interaction with the *E1* gene and is also located near another flowering regulator, *GmTOE1*

In soybean, *GmCCA1s* (*CIRCADIAN CLOCK ASSOCIATED 1*) genes promote flowering by repressing the expression of the major flowering repressor, *E1* (Dong et al., 2021). *RVE* genes belong to a family of transcription factors with CCA1-like MYB domains, so we suggested that the function of *GmRVE8c* in flowering regulation might also depend on the *E1* gene.

The distance between *GmRVE8c* and *GmTOE1* is approximately 21 kb, and *GmTOE1* is nearly the immediate downstream gene following *GmRVE8c* in the 5′–3′ direction (Table S1). *GmTOE1* harbors natural variation that is associated with flowering and maturity time in soybeans (Perfil'ev et al., 2024). Previously, we identified four main haplotypes, including two early-maturing and two late-maturing haplotypes (hereafter referred to as the $GmTOE1^{C}$ and $GmTOE1^{T}$ alleles, respectively), which differ from each other by the nucleotide substitution T316C, resulting in the replacement of serine with proline at the 106th amino acid (Perfil'ev et al., 2024).

We examined the association of *GmRVE8c* haplotypes, taking into account the *E1* and *GmTOE1* genetic backgrounds, using data from SoyOmics and from our experiment (Fig. S3–S6). The ANOVA results show a significant interaction between *GmRVE8c* and *E1*, based on data from SoyOmics (Tables S4, S5). This indirectly suggests the presence of an

[1] Supplementary Tables S1–S6 and Figures S1–S6 are available at: https://vavilovj-icg.ru/download/pict-2026-30/appx3.zip

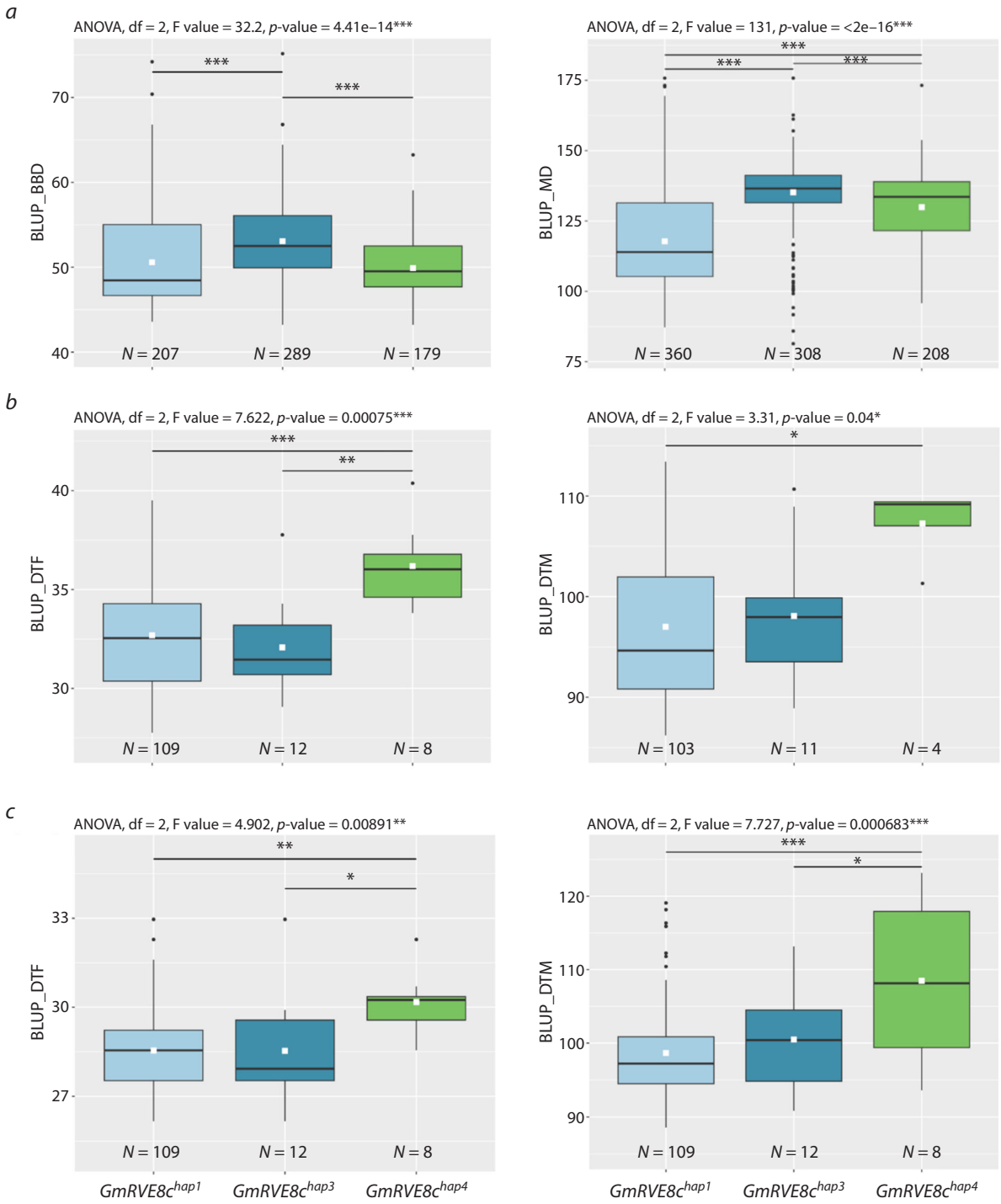


Fig. 1. Association of haplotypes *GmRVE8c* with flowering and maturity time.

a – association of haplotypes *GmRVE8c* with BLUP_BBD and BLUP_MD; *b* – association of haplotypes *GmRVE8c* with BLUP_DTF and BLUP_DTM in Novosibirsk; *c* – association of haplotypes *GmRVE8c* with BLUP_DTF and BLUP_DTM in Oryol. White squares on the boxplots indicate the mean value. Asterisks indicate significant differences between the compared genotypes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

epistatic interaction between *GmRVE8c* and *E1*. Analysis of the established associations using data from SoyOmics showed that, on the *e1-as* allele background, the $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$ haplotypes have a stronger effect on flowering and maturity time (Fig. S3*a*, *b*). In our collection, plants carrying the $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$ haplotypes also carry the *e1-as* or *e1-nl* alleles, and plants with $GmRVE8c^{hap4}$ exhibit later flowering and maturity compared to other genotypes (Fig. S3*c*, *d*).

According to BLUP_BBD data, plants with $GmRVE8c^{hap1}$ $GmTOE1^{T}$ and $GmRVE8c^{hap3}$ $GmTOE1^{T}$ flower significantly later than those with $GmRVE8c^{hap4}$ $GmTOE1^{C}$ and

GmRVE8c^{hap1} *GmTOE1*^{C} (Fig. S4*a*). A similar pattern is observed for BLUP_MD,with the exception of the *GmRVE8c*^{hap4} *GmTOE1*^{C} genotype, which matures later than *GmRVE8c*^{hap1} *GmTOE1*^{C} (Fig. S4*a*).According to data from our collection, plants with *GmRVE8c*^{hap4}*GmTOE1*^{C} flower and mature significantly later than those with *GmRVE8c*^{hap1} *GmTOE1*^{C} in both regions (Fig. S4*b*–*c*). According to BLUP_MD and BLUP_BBD, plants with the *GmRVE8c*^{hap1} *GmTOE1*^{C} genotype flower and mature significantly earlier than those with *GmRVE8c*^{hap1} *GmTOE1*^{T} (Fig. S4*a*). According to the data from our collection, plants with the *GmRVE8c*^{hap1} *GmTOE1*^{C} genotype flower significantly earlier than those with *GmRVE8c*^{hap1} *GmTOE1*^{T}, but only under the conditions of Oryol (Fig. S4*b*). The *GmTOE1*^{C} allele is associated with early flowering and maturity, as we described previously, and *GmRVE8c*^{hap4} appears to partially restore the early-maturing phenotype of *GmTOE1*^{C} plants, making them late-maturing (Fig. S4*a*).

When considering all three genes simultaneously, we interpreted the results by comparing genotypes that differ from each other at only one gene. According to BLUP_BBD and BLUP_MD, plants with the *GmRVE8c*^{hap1} *e1-as GmTOE1*^{C} genotype flower and mature later than those with *GmRVE8c*^{hap4} *e1-as GmTOE1*^{C} (Fig. S5). According to BLUP_BBD, there are no significant differences between genotypes *GmRVE8c*^{hap1}*E1 GmTOE1*^{C} and *GmRVE8c*^{hap1} *E1 GmTOE1*^{T}, or between *GmRVE8c*^{hap1} *e1-as GmTOE1*^{C} and *GmRVE8c*^{hap1} *e1-as GmTOE1*^{T} (Fig. S5*a*). According to BLUP_MD, of the two genotypes, only *GmRVE8c*^{hap1} *E1 GmTOE1*^{C} and *GmRVE8c*^{hap1} *E1 GmTOE1*^{T} differ significantly from each other (Fig. S5*b*). In our collection, plants with the *GmRVE8c*^{hap1} *e1-nl GmTOE1*^{C} genotype flower significantly earlier than those with *GmRVE8c*^{hap1} *e1-nl GmTOE1*^{T} under Oryol conditions (Fig. S6*c*). In addition, genotypes with *GmRVE8c*^{hap4} *e1-as GmTOE1*^{C} flower and mature later than the other genotypes (Fig. S6). Overall (although not completely) this repeats the result of the two-gene analysis from the previous paragraph.

There are also significant differences in BLUP_MD and BLUP_BBD between genotypes *GmRVE8c*^{hap1} *E1 GmTOE1*^{C} and *GmRVE8c*^{hap4} *E1 GmTOE1*^{C}, with plants carrying *GmRVE8c*^{hap4} maturing later but flowering slightly earlier (Fig. S5). In the background of the *e1-as* allele, the *GmRVE8c*^{hap4} haplotype shows a stronger effect on BLUP_BBD (Fig. S5*a*). For the *GmRVE8c*^{hap3} haplotype, a similar pattern is observed, with no significant differences in BLUP_BBD and BLUP_MD between the *GmRVE8c*^{hap3} *E1 GmTOE1*^{T} and *GmRVE8c*^{hap1} *E1 GmTOE1*^{T} genotypes (Fig. S5). However, a significant difference in BLUP_MD is observed between the *GmRVE8c*^{hap3} *e1-as GmTOE1*^{T} and *GmRVE8c*^{hap1} *e1-as GmTOE1*^{T} genotypes (Fig. S5*b*). In our collection, plants carrying the *GmRVE8c*^{hap4} haplotype, even in the presence of the null *e1-nl* allele, show late flowering and maturity (Fig. S5).

Distribution of *GmRVE8c* haplotypes in soybean accessions of different origin

For the *GmRVE8c* gene, we examined haplotype frequencies in soybean cultivars from China across three major soybean growing regions (China I, China II, and China III) and one mixed region (China IV–VI), as well as in our collection, which was divided into three main groups: A – from West Siberia (Novosibirsk and Omsk regions); B – from other Russian regions; and C – from other countries (Fig. 2).

The *GmRVE8c*^{hap1} haplotype predominates in cultivars from northern China, whereas *GmRVE8c*^{hap3} predominates in the other China regions. In our collection, accessions from Western Siberia predominantly carry the *GmRVE8c*^{hap1} haplotype, whereas the late-maturing *GmRVE8c*^{hap4} haplotype is present only in accessions from other regions of Russia and from other countries (Fig. 2*b*). This demonstrates the role of *GmRVE8c* haplotypes in the adaptation of soybean to different cultivation regions.

Discussion

The *GmRVE8c* gene is a new candidate gene for regulating flowering and maturity in soybeans

We identified the *Glyma.03G177300/GmRVE8c* gene as a new candidate gene for the previously found qDTF-7 locus associated with soybean flowering (Perfil'ev et al., 2024). *GmRVE8c* was found to have four major haplotypes resulting from three single nucleotide substitutions and a 19-bp deletion (Table S2). Among the mutations identified in *GmRVE8c*, the soy32494699 substitution (a 19-bp deletion) in the *GmRVE8c*^{hap4} haplotype may lead to a complete loss of the encoded protein's original function due to a frameshift occurring after the fourth amino acid (Table S2). The remaining three *GmRVE8c*$^{hap1–3}$ haplotypes arose due to three nonsynonymous substitutions (Table S2). Although all three substitutions are located outside the predicted DNA-binding domains (Fig. S1), they may still affect the 3D structure of the protein or its interactions with other proteins (Fig. S1). Notably, two of the identified mutations are located within the predicted LIP (Linear Interacting Peptides) regions of the protein (Fig. S1). A complete loss of *GmRVE8c* function appears to result in delayed soybean development, as plants carrying the *GmRVE8c*^{hap4} haplotype exhibit late flowering and maturity (Fig. 1). It is also possible that the *GmRVE8c*^{hap3} haplotype has also lost the original function of the protein, since, according to SoyOmics data, samples carrying this haplotype also exhibit late flowering and maturity (Fig. 1*a*).

The frequency of *GmRVE8c* haplotypes shifts from wild to cultivated soybean; compared with wild soybean, the *GmRVE8c*^{hap2} haplotype is almost absent in landrace and modern cultivars (Table S2). Moreover, the *GmRVE8c*^{hap4} haplotype is not detected in the wild soybean population and likely arose and was selected after soybean domestication (Table S2). The distribution of *GmRVE8c* haplotypes across different cultivation regions shows that the early-maturing haplotype *GmRVE8c*^{hap1} predominates in cultivars from northern China, whereas *GmRVE8c*^{hap3} predominates in other regions (Fig. 2*a*); similarly, in our collection, almost all cultivars from Western Siberia carry the *GmRVE8c*^{hap1} haplotype (Fig. 2*b*). This highlights the role of *GmRVE8c* haplotypes in the adaptation of soybean to different cultivation regions. Due to its association

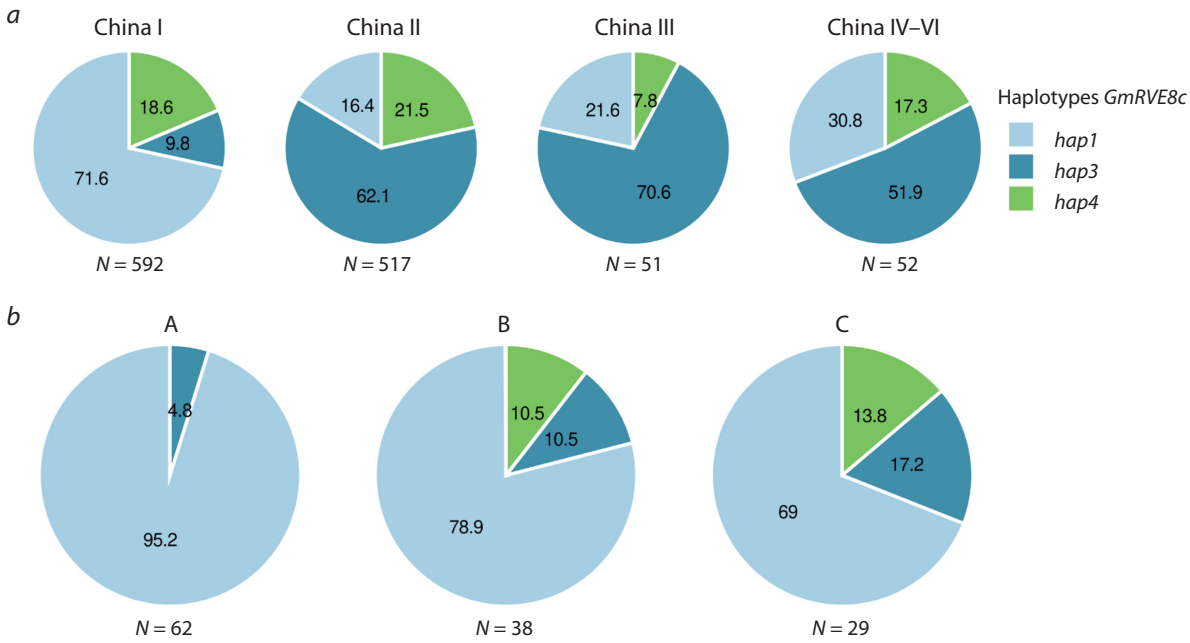


Fig. 2. Distribution of *GmRVE8c* haplotypes: *a* – in improved soybean cultivars from China (China I designates northern provinces of China, and China II, III, IV–VI designate southern provinces); *b* – distribution of *GmRVE8c* haplotypes in three groups of soybean accessions: A – from West Siberia (Novosibirsk and Omsk regions); B – from other Russian regions; and C – from other countries.

Numbers on pie charts indicate percentages; *N* denotes the number of accessions.

with late flowering and maturity, the $GmRVE8c^{hap4}$ haplotype is unfavorable for northern latitudes but, conversely, advantageous for southern latitudes. We believe that $GmRVE8c^{hap4}$ could be valuable in breeding soybean cultivars adapted to southern regions.

The *GmRVE8c* gene shows epistatic interaction with the *E1* gene

The *RVE* gene family encodes transcription factors with CCA1-like MYB domains and is closely related to the *CCA1* (*CIRCADIAN CLOCK ASSOCIATED 1*) and *LHY* (*LATE ELONGATED HYPOCOTYL*) genes (Rawat et al., 2009). In soybean, the *CCA1/LHY* genes are known regulators of flowering, and under SD conditions, *GmLHYs* directly bind to the *E1* promoter and repress its expression, thereby accelerating flowering, whereas loss of *GmLHYs* function delays flowering (Dong et al., 2021). Based on this, we hypothesized that the role of *GmRVE8c* in flowering regulation may also depend on the *E1* gene. ANOVA results based on SoyOmics data show a significant interaction between *E1* and *GmRVE8c*, suggesting that *GmRVE8c* may also act as a repressor of the *E1* gene (Tables S3, S4). Association analysis of these data shows that the $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$ haplotypes more strongly delay flowering and maturity in the background of the semi-functional *e1-as* allele (Fig. S4*a*, *b*). This is an unexpected situation, because if we assume that *GmRVE8c* can suppress the expression of the *E1* gene, then in the presence of the weakly functional *e1-as* allele, loss of *GmRVE8c* function should not lead to a phenotypic change, or at least the effect should be weaker. Similarly, the effect on the flowering time of the loci/genes *Tof16/LHYa*, *Tof8/FKF1b*, and *J/ELF3* under the background of the *e1-as* allele completely disappeared or was reduced (Lu S. et al., 2017; Dong et al., 2021; Li H. et al., 2023a). Phenotypic analysis in our collection shows that the $GmRVE8c^{hap4}$ haplotype affects flowering and maturity even in the presence of the *e1-nl* allele, which has a deletion of the entire *E1* gene (Fig. S3*c*, *d*). It is likely that, under our conditions, *GmRVE8c* may be involved in the transcriptional regulation of the entire *E1* gene family (*E1*, *E1La*, *E1Lb*), or that *GmRVE8c* controls flowering time through an *E1*-independent mechanism.

In *Arabidopsis*, proteins LNK1 (NIGHT LIGHT–INDUCIBLE AND CLOCK-REGULATED 1) and LNK2 recruit RVE4 and RVE8 to activate transcription of *TOC1* (*TIMING OF CAB EXPRESSION 1*) and *PRR5* (*PSEUDO-RESPONSE REGULATOR 5*), and loss of function of *LNK1–4* results in delayed flowering under LD conditions (Xie et al., 2014; De Leone et al., 2018). Knockout of the four *GmLNK2s* orthologs in soybean accelerates flowering under LD conditions due to reduced expression of the *E1* gene (Li Z. et al., 2021). A characteristic feature of *LNK* genes is the absence of their own DNA-binding domains, for which they recruit *CCA1*, *LHY*, *RVE4* and *RVE8*, and switch them from repressors to activators of transcription (Xie et al., 2014). Further studies will clarify whether the function of *GmRVE8c* in flowering control depends on the *E1* gene family and whether GmLNK2s proteins recruit GmRVE8s proteins to regulate *E1* gene transcription.

The qDTF-7 locus as a possible example of a heterogeneous locus in soybean

Locus or genetic heterogeneity is a situation in which alleles from different closely located genes at the same locus influence the same trait. For the qDTF-7 locus, we previously proposed a strong candidate gene, *GmTOE1* (Perfil'ev et al., 2024), which has been shown to play a role in the regulation of flowering in soybean (Wang T. et al., 2016). In this study, we identified a new candidate gene, *GmRVE8c*, located ~21 kb upstream of *GmTOE1*, which may also be involved in the regulation of flowering and maturity in soybean. Using the data available to us, we attempted to analyze how mutations in the *GmRVE8c* and *GmTOE1* genes correspond to each other. In the genetic background of the early-flowering/maturing $GmTOE1^C$ allele, the late-maturing $GmRVE8c^{hap4}$ haplotype arose. We did not find any $GmRVE8c^{hap4}$ $GmTOE1^T$ genotypes in either the SoyOmics database or our collection (Fig. S4). Based on SoyOmics phenotypic data, the effect of the $GmRVE8c^{hap4}$ haplotype appears to partially restore the effect of the early-maturing $GmTOE1^C$ allele on flowering and maturity time (Fig. S4*a*, S5*a*, *b*).

A similar pattern is observed in our collection; however, the $GmRVE8c^{hap4}$ haplotype shows a much stronger effect on flowering and maturity time than either $GmTOE1^C$ or $GmTOE1^T$ (Fig. S4*b*, *c*, S6*a*, *b*). *GmTOE1* alleles have a minor effect on flowering and maturity under our conditions; for example, genotype $GmRVE8c^{hap1}$ $GmTOE1^C$ differs significantly from $GmRVE8c^{hap1}$ $GmTOE1^C$ only for BLUP_DTF in Oryol conditions (Fig. S4*b*, *c*, S6*a*, *b*). Interestingly, the $GmRVE8c^{hap3}$ haplotype predominates in the background of the $GmTOE1^T$ allele, and, for example, in the SoyOmics database and in our collection, we found only two accessions with the $GmRVE8c^{hap3}$ $GmTOE1^C$ genotype (Fig. S4). Therefore, it is difficult to determine whether the delayed flowering and maturity are caused by $GmRVE8c^{hap3}$ or by $GmTOE1^T$, although SoyOmics data suggest that both genes influence these traits (Fig. S4*a*, S5). However, phenotypic analysis in our collection did not reveal any significant effect of $GmRVE8c^{hap3}$ on flowering and maturity time (Fig. S4*b*, *c*, S6*a*, *b*).

It should be mentioned that the mutations detected in the *GmRVE8c* and *GmTOE1* genes require functional validation to confirm whether they actually affect the phenotype. If this proves true, the qDTF-7 locus could serve as an interesting example of how natural mutations shape phenotypic diversity in plants, particularly in soybean.

The influence of the *GmRVE8c* gene on other soybean traits

The early-maturing $GmTOE1^C$ allele likely contributes to soybean adaptation to northern latitudes by reducing flowering time (Perfil'ev et al., 2024). In contrast, the $GmRVE8c^{hap4}$ haplotype is unfavorable for northern latitudes due to its association with late flowering and maturity, whereas in our collection, all accessions from Western Siberia carry the early-maturing $GmRVE8c^{hap1}$ haplotype (Fig. 2). As discussed above, $GmRVE8c^{hap4}$ appears to restore the early-maturing phenotype of $GmTOE1^C$, raising the question of why the $GmRVE8c^{hap4}$ mutation arose in the presence of the early-maturing $GmTOE1^C$ allele. A possible explanation for this is the pleiotropic effect of *GmRVE8c* on other soybean traits. Interestingly, several loci associated with soybean drought tolerance have been identified on chromosome 3 using bi-parental QTL mapping (Table S6). The genes *GmRVE8c* and *GmTOE1* are also positioned within these loci. However, it should be noted that these loci also contain the known drought tolerance regulator *GmSALT3* (Shi et al., 2018; Guan et al., 2021).

Overexpression of *GmRVE8a* in *Arabidopsis* enhances drought and salt tolerance, and GmRVE8a can interact with the GmNAC019 protein (Bao et al., 2024), which is involved in the ABA-mediated drought response (Hoang et al., 2019). *GmLHYs* genes are also involved in the ABA-mediated drought response in soybean, and knockout of all four *GmLHYs* genes enhances drought tolerance (Wang K. et al., 2021). Recently, it has been shown that the soybean circadian clock component *GmPRR3b* directly suppresses the expression of *GmABF3* (*ABSCISIC ACID-RESPONSIVE ELEMENT-BINDING FACTOR 3*), the overexpression of which enhances drought tolerance in soybean (Li C. et al., 2024). We hypothesize that loss of *GmRVE8c* function may also enhance drought tolerance in soybean. Interestingly, in *Arabidopsis*, the *TOE1* and *TOE2* genes have been shown to play key roles in drought escape, and loss of *TOE1/2* function accelerates flowering but reduces drought tolerance (Chen et al., 2025). For the early-maturing $GmTOE1^C$ allele, we assume at least a partial loss of function of the encoded protein (Perfil'ev et al., 2024). It is possible that the $GmRVE8c^{hap4}$ haplotype arose to compensate for drought sensitivity in plants carrying the $GmTOE1^C$ allele. In our view, this is an intriguing hypothesis that warrants further exploration.

According to Table S6, the *GmRVE8c* gene colocalizes with loci associated with plant height and weight, oil and isoflavone content, seed thickness, and resistance to soybean cyst nematode. Theoretically, *GmRVE8c* may also influence some of these traits, such as plant height, and weight, potentially through regulation of the *PRR5–PIF4* module, as has been proposed for *GmMYB133* (Shan et al., 2021). *GmMYB133* has also been shown to be a positive regulator of isoflavone biosynthesis in soybean (Bian et al., 2018). In *Arabidopsis*, *rve 4 6 8* mutants exhibit altered metabolism of carbohydrates, fatty acids, and lipids, as well as reduced proteasome activity, which leads to an increase in cell size (Scandola et al., 2022). Various circadian clock genes have been studied and shown to contribute to plant defense against pathogens (Lu H. et al., 2017). In particular, the role of *GmCCA1-1* in soybean response to infection by the nematode *Heterodera glycines* has recently been demonstrated (Niraula et al., 2022). All of this makes *GmRVE8c* an interesting gene for further study, not only in the context of soybean flowering and maturity but also for other agriculturally important traits.

Conclusion

In this study, we identified a new candidate gene, *GmRVE8c*, and two of its haplotypes, $GmRVE8c^{hap3}$ and $GmRVE8c^{hap4}$, which are associated with late flowering and maturity in soybean. In addition, to distinguish these haplotypes from each other, we developed DNA markers that can be used for marker-assisted selection in soybean. The *GmRVE8c* gene is located

in close proximity to another soybean flowering regulator, *GmTOE1*, which also has mutations associated with flowering and maturity time. *GmTOE1* and *GmRVE8c* are likely to form a genetically heterogeneous locus, i. e. a locus in which alleles of different nearby genes influence the same trait. Also, the influence of the *GmRVE8c*^{hap3} and *GmRVE8c*^{hap4} haplotypes on the flowering and maturity time shows partial dependence on the allelic state of the *E1* gene. Altogether, these findings make qDTF-7 a highly complex and intriguing locus, and we plan to further validate the mutations in *GmRVE8c* and *GmTOE1* using hybrid soybean populations.

References

Arya H., Singh M.B., Bhalla P.L. Overexpression of *PIF4* affects plant morphology and accelerates reproductive phase transitions in soybean. *Food Energy Secur*. 2021;10(3):e291. doi 10.1002/fes3.291

Atwell S., Huang Y.S., Vilhjálmsson B.J., Willems G., Horton M., Li Y., Meng D., … Weigel D., Marjoram P., Borevitz J.O., Bergelson J., Nordborg M. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*. 2010;465(7298):627-631. doi 10.1038/nature08800

Bao G., Sun G., Wang J., Shi T., Xu X., Zhai L., Bian S., Li X. Soybean *RVE8a* confers salt and drought tolerance in Arabidopsis. *Biochem Biophys Res Communs*. 2024;704:149660. doi 10.1016/j.bbrc.2024.149660

Bates D., Mächler M., Bolker B., Walker S. Fitting linear mixed-effects models using lme4. *J Stat Software*. 2015;67(1):1-48. doi 10.18637/jss.v067.i01

Bian S., Jin D., Li R., Xie X., Gao G., Sun W., Li Y., Zhai L., Li X. Genome-wide analysis of CCA1-like proteins in soybean and functional characterization of GmMYB138a. *Int J Mol Sci.* 2017;18(10):2040. doi 10.3390/ijms18102040

Bian S., Li R., Xia S., Liu Y., Jin D., Xie X., Dhaubhadel S., Zhai L., Wang J., Li X. Soybean *CCA1-like MYB* transcription factor *GmMYB133* modulates isoflavonoid biosynthesis. *Biochem Biophys Res Commun*. 2018;507(1-4):324-329. doi 10.1016/j.bbrc.2018.11.033

Chen W., Wang T., Li X., Feng J., Liu Q., Xu Z., You Q., Yang L., Liu L., Chen S., Yue Z., Wang H., Yu D. *Arabidopsis* RGLG1/2 regulate flowering time under different soil moisture conditions by affecting the protein stability of TOE1/2. *New Phytol*. 2025;246(4):1609-1626. doi 10.1111/nph.70073

Clauw P., Ellis T.J., Liu H.-J., Sasaki E. Beyond the standard GWAS – a guide for plant biologists. *Plant Cell Physiol.* 2025;66(4):431-443. doi 10.1093/pcp/pcae079

De Leone M.J., Hernando C.E., Romanowski A., García-Hourquet M., Careno D., Casal J., Rugnone M., Mora-García S., Yanovsky M.J. The *LNK* gene family: at the crossroad between light signaling and the circadian clock. *Genes*. 2019;10(1):2. doi 10.3390/genes10010002

Dong L., Fang C., Cheng Q., Su T., Kou K., Kong L., Zhang C., … Yuan X., Weller J.L., Lu S., Kong F., Liu B. Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nat Commun*. 2021;12:5445. doi 10.1038/s41467-021-25800-3

Dong L., Cheng Q., Fang C., Kong L., Yang H., Hou Z., Li Y., … Tang Y., Zhao X., Lu S., Liu B., Kong F. Parallel selection of distinct Tof5 alleles drove the adaptation of cultivated and wild soybean to high latitudes. *Mol Plant*. 2022;15(2):308-321. doi 10.1016/j.molp.2021.10.004

Gray J.A., Shalit-Kaneh A., Chu D.N., Hsu P.Y., Harmer S.L. The REVEILLE clock genes inhibit growth of juvenile and adult plants by control of cell size. *Plant Physiol*. 2017;173(4):2308-2322. doi 10.1104/pp.17.00109

Guan R., Yu L., Liu X., Li M., Chang R., Gilliham M., Qiu L. Selection of the salt tolerance gene *GmSALT3* during six decades of soybean breeding in China. *Front Plant Sci*. 2021;12:794241. doi 10.3389/fpls.2021.794241

Guo S., Li Y., Qiu H., Hu G., Zhao C., Wang R., Zhang H., Tian Y., Li X., Liu B., Li Ying-hui, Qiu L. *GmAP1d* regulates flowering time under long-day photoperiods in soybean. *Crop J.* 2024;12(3):845-855. doi 10.1016/j.cj.2024.03.004

Hoang X., Nguyen N., Nguyen Y.-N., Watanabe Y., Tran L.-S. The soybean *GmNAC019* transcription factor mediates drought tolerance in *Arabidopsis* in an abscisic acid-dependent manner. *Int J Mol Sci.* 2019;21(1):286. doi 10.3390/ijms21010286

Kassambara A. ggcorrplot: Visualization of a Correlation Matrix Using "ggplot2". 2023. R package version 0.1.4.999. Available: https://github.com/kassambara/ggcorrplot

Li C., Chen Y., Hu Q., Yang X., Zhao Y., Lin Y., Yuan J., Gu J., Li Y., He J., Wang D., Liu B., Wang Z.-Y. PSEUDO-RESPONSE REGULATOR 3b and transcription factor ABF3 modulate abscisic acid-dependent drought stress response in soybean. *Plant Physiol*. 2024;195(4):3053-3071. doi 10.1093/plphys/kiae269

Li H., Du H., He M., Wang J., Wang F., Yuan W., Huang Z., … Liu B., Kong F., Fang C., Zhao X., Yu D. Natural variation of *FKF1* controls flowering and adaptation during soybean domestication and improvement. *New Phytol*. 2023a;238(4):1671-1684. doi 10.1111/nph.18826

Li H., Du H., Huang Z., He M., Kong L., Fang C., Chen L., Yang H., Zhang Y., Liu B., Kong F., Zhao X. The AP2/ERF transcription factor *TOE4b* regulates photoperiodic flowering and grain yield per plant in soybean. *Plant Biotechnol J.* 2023b;21(8):1682-1694. doi 10.1111/pbi.14069

Li S., Wang W., Sun L., Zhu H., Hou R., Zhang H., Tang X., Clark C.B., Swarm S.A., Nelson R.L., Ma J. Artificial selection of mutations in two nearby genes gave rise to shattering resistance in soybean. *Nat Commun.* 2024;15:7588. doi 10.1038/s41467-024-52044-8

Li Y.F., Zhang L., Wang J., Wang X., Guo S., Xu Z.J., Li D., Liu Z., Li Y.H., Liu B., Qiu L.J. Flowering time regulator *qFT13-3* involved in soybean adaptation to high latitudes. *Plant Biotechnol J.* 2024;22(5):1164-1176. doi 10.1111/pbi.14254

Li Z., Cheng Q., Gan Z., Hou Z., Zhang Y., Li Y., Li H., … Kou K., Wang L., Kong F., Liu B., Dong L. Multiplex CRISPR/Cas9-mediated knockout of soybean *LNK2* advances flowering time. *Crop J*. 2021;9(4):767-776. doi 10.1016/j.cj.2020.09.005

Liu H.-J., Swarts K., Xu S., Yan J., Nordborg M. On the contribution of genetic heterogeneity to complex traits. *bioRxiv.* 2024. doi 10.1101/2024.03.27.586967

Liu Y., Zhang Y., Liu X., Shen Y., Tian D., Yang X., Liu S., Ni L., Zhang Z., Song S., Tian Z. SoyOmics: a deeply integrated database on soybean multi-omics. *Mol Plant*. 2023;16(5):794-797. doi 10.1016/j.molp.2023.03.011

Lu H., McClung C.R., Zhang C. Tick Tock: circadian regulation of plant innate immunity. *Annu Rev Phytopathol*. 2017;55:287-311. doi 10.1146/annurev-phyto-080516-035451

Lu S., Zhao X., Hu Y., Liu S., Nan H., Li X., Fang C., … Weller J.L., Liu B., Hou X., Tian Z., Kong F. Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. *Nat Genet*. 2017;49(5):773-779. doi 10.1038/ng.3819

Niraula P.M., McNeece B.T., Sharma K., Alkharouf N.W., Lawrence K.S., Klink V.P. The central circadian regulator *CCA1* functions in *Glycine max* during defense to a root pathogen, regulating the expression of genes acting in effector triggered immunity (ETI) and cell wall metabolism. *Plant Physiol Biochem*. 2022;185:198-220. doi 10.1016/j.plaphy.2022.05.028

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

Perfil'ev R., Shcherban A., Potapov D., Maksimenko K., Kiryukhin S., Gurinovich S., Panarina V., Polyudina R., Salina E. Impact of allelic variation in maturity genes *E1–E4* on soybean adaptation to Central

and West Siberian regions of Russia. *Agriculture*. 2023;13(6):1251. doi 10.3390/agriculture13061251

Perfil'ev R., Shcherban A., Potapov D., Maksimenko K., Kiryukhin S., Gurinovich S., Panarina V., Polyudina R., Salina E. Genome-wide association study revealed some new candidate genes associated with flowering and maturity time of soybean in Central and West Siberian regions of Russia. *Front Plant Sci.* 2024;15:1463121. doi 10.3389/fpls.2024.1463121

Piovesan D., Del Conte A., Mehdiabadi M., Aspromonte M.C., Blum M., Tesei G., von Bülow S., Lindorff-Larsen K., Tosatto S.C.E. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Res*. 2025;53(D1):D495-D503. doi 10.1093/nar/gkae969

Rawat R., Schwartz J., Jones M.A., Sairanen I., Cheng Y., Andersson C.R., Zhao Y., Ljung K., Harmer S.L. REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways. *Proc Natl Acad Sci USA*. 2009;106(39):16883-16888. doi 10.1073/pnas.0813035106

Rawat R., Takahashi N., Hsu P.Y., Jones M.A., Schwartz J., Salemi M.R., Phinney B.S., Harmer S.L. REVEILLE8 and PSEUDO-REPONSE REGULATOR5 form a negative feedback loop within the Arabidopsis circadian clock. *PLoS Genet*. 2011;7(3):e1001350. doi 10.1371/journal.pgen.1001350

Rogers S.O., Bendich A.J. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol*. 1985;5(2):69-76. doi 10.1007/BF00020088

Sasaki E., Köcher T., Filiault D.L., Nordborg M. Revisiting a GWAS peak in *Arabidopsis thaliana* reveals possible confounding by genetic heterogeneity. *Heredity.* 2021;127(3):245-252. doi 10.1038/s41437-021-00456-3

Scandola S., Mehta D., Li Q., Rodriguez Gallo M.C., Castillo B., Uhrig R.G. Multi-omic analysis shows *REVEILLE* clock genes are involved in carbohydrate metabolism and proteasome function. *Plant Physiol.* 2022;190(2):1005-1023. doi 10.1093/plphys/kiac269

Shan B., Wang W., Cao J., Xia S., Li R., Bian S., Li X. Soybean *GmMYB133* inhibits hypocotyl elongation and confers salt tolerance in *Arabidopsis*. *Front Plant Sci*. 2021;12:764074. doi 10.3389/fpls.2021.764074

Shi X., Yan L., Yang C., Yan W., Moseley D.O., Wang T., Liu B., Di R., Chen P., Zhang M. Identification of a major quantitative trait locus underlying salt tolerance in 'Jidou 12' soybean cultivar. *BMC Res Notes.* 2018;11:95. doi 10.1186/s13104-018-3202-3

Wang K., Bu T., Cheng Q., Dong L., Su T., Chen Z., Kong F., Gong Z., Liu B., Li M. Two homologous *LHY* pairs negatively control soybean drought tolerance by repressing the abscisic acid responses. *New Phytol*. 2021;229(5):2660-2675. doi 10.1111/nph.17019

Wang T., Sun M.-Y., Wang X.-S., Li W.-B., Li Y.-G. Over-expression of GmGIa-regulated soybean *miR172a* confers early flowering in transgenic *Arabidopsis thaliana*. *Int J Mol Sci.* 2016;17(5):645. doi 10.3390/ijms17050645

Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer, 2016. doi 10.1007/978-3-319-24277-4

Xie Q., Wang P., Liu X., Yuan L., Wang L., Zhang C., Li Y., Xing H., Zhi L., Yue Z., Zhao C., McClung C.R., Xu X. LNK1 and LNK2 are transcriptional coactivators in the *Arabidopsis* circadian oscillator. *Plant Cell*. 2014;26(7):2843-2857. doi 10.1105/tpc.114.126573

Zhang M., Liu S., Wang Z., Yuan Y., Zhang Z., Liang Q., Yang X., Duan Z., Liu Y., Kong F., Liu B., Ren B., Tian Z. Progress in soybean functional genomics over the past decade. *Plant Biotechnol J.* 2022;20(2):256-282. doi 10.1111/pbi.13682

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

Received August 14, 2025. Revised November 25, 2025. Accepted December 23, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):53-60

doi 10.18699/vjgb-26-05

A histochemical assay for polyphenolic profiling in cereal grains

S.R. Mursalimov, O.Yu. Shoeva

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
mursalimovsr@gmail.com

Abstract. In different cell layers, cereal grains may accumulate various economically important polyphenols such as colored anthocyanins and melanins and colorless proanthocyanidins. To effectively create new cultivars with different combinations of these compounds, a simple, fast, and precise screening method is required. Here, a histochemical assay that includes a combination of hot ethanolic, acidic, alkaline, and ammoniacal silver treatments of grain cryosections followed by microscopy was successfully applied to distinguish these substances in cereal grains. Barley lines previously characterized chemically for the presence of anthocyanins, proanthocyanidins, and melanins in grains were used as a model. In black barley grains, this approach allowed to visually distinguish insoluble melanins that do not react to a pH change from anthocyanins, which can be insoluble or soluble but always react to changing pH. For the first time, ammoniacal silver staining commonly used for melanin identification in human and animal tissues was adapted for melanin identification in plant tissues. Along with melanins, this reagent stains other polyphenols thereby helping to detect colorless polyphenols including proanthocyanidins in the testa of barley grains as confirmed by *p*-dimethylaminocinnamaldehyde (DMACA) staining. The applicability of this assay to polyphenol profiling was demonstrated not only in the barley grain but also in wheat and common vetch grains. The proposed histochemical assay allows rapid polyphenol screening using a single grain, making it a practical and efficient alternative to time-consuming chromatographic methods for preliminary selection from large sample sets prior to detailed quantitative and qualitative chemical analysis.

Key words: anthocyanin; analytical technique; DMACA; proanthocyanidins; melanins; Fontana–Masson

For citation: Mursalimov S.R., Shoeva O.Yu. A histochemical assay for polyphenolic profiling in cereal grains. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):53-60. doi 10.18699/vjgb-26-05

Funding. The study was supported by RSF grant No. 25-16-20101, https://rscf.ru/project/25-16-20101/ and the Ministry of Science and Innovation Policy of the Novosibirsk region (the agreement No. 30-2025-000848).

Acknowledgements. We are very grateful for Dr. A. Zukin (SPBGU) for providing us with the DMACA reagent. Plants for gathering developing seeds were grown using resources of the Greenhouse Core Facility supported by ICG project number FWNR-2026-0029.

Гистохимический тест для определения профиля полифенольных соединений в зерне злаков

С.Р. Мурсалимов, О.Ю. Шоева

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
mursalimovsr@gmail.com

Аннотация. Зерна злаков в некоторых типах клеток могут накапливать разнообразные экономически важные полифенольные соединения, такие как окрашенные антоцианы и меланины, а также бесцветные проантоцианидины. Для эффективного создания новых сортов с различными комбинациями этих соединений в зернах необходим простой, быстрый и точный метод определения их полифенольного профиля. В данной статье для идентификации указанных веществ предложено использовать гистохимический анализ, включающий комбинацию обработок криосрезов зерен горячим этанолом, кислотой, щелочью и аммиачным раствором серебра с последующей микроскопией. В качестве модели использовались линии ячменя, зерна которых ранее были охарактеризованы с помощью аналитических методов на наличие антоцианов, проантоцианидинов и меланинов. В черных зернах ячменя данный подход позволил отличить нерастворимые меланины, не реагирующие на изменение pH, от антоцианов, которые могут быть растворимыми или нерастворимыми, но всегда меняют окраску при изменении pH. Впервые обработка аммиачным серебром, широко применяемая для идентификации меланина в тканях человека и животных, была адаптирована для использования в растительных тканях. Наряду с меланинами, этот реагент окрашивает и другие полифенолы, что позволяет выявлять в том числе бесцветные соединения. С помощью этого подхода были выявлены проантоцианидины в оболочке зерна ячменя, что подтверждено окрашиванием 4-(диметиламино)циннамальдегидом (DMACA). Разработанный протокол успешно применен для определения полифенольного профиля зерен ячменя, пшеницы и вики посевной. Метод позволяет проводить быструю оценку полифенольного профиля с использованием единичных зерен, что является эффективной альтернативой трудоемким хроматографическим методам на этапе предварительного отбора коллекционного материала перед его детальным химическим анализом.

Ключевые слова: антоцианы; аналитическая техника; DMACA; проантоцианидины; меланины; реакция Фонтана–Массона

© Mursalimov S.R., Shoeva O.Yu., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Introduction

In different cell layers, cereal grains may accumulate different polyphenols such as colored anthocyanins and melanins and colorless proanthocyanidins. Melanins and anthocyanins are known to participate in protection of grain against severe environments and threats by predators and pathogens (Winkel-Shirley, 1998; Glagoleva et al., 2020), while proanthocyanidins are conducive to dormancy and longevity of seeds (Debeaujon et al., 2000).

The aforementioned polyphenols are industrially important; they affect final application of grains as a raw material. For example, due to their protein-binding capacity leading to chill haze (reducing beer quality), proanthocyanidins synthesized in the testa (i. e., the seed coat) of barley under the control of the *Ant28* gene are undesirable in malt cultivars (von Wettstein, 2007). Anthocyanins that accumulate either in the aleurone layer or in the pericarp under the control of genes *Blx1–5* or *Ant1* and *Ant2* cause the grain color to be blue and purple, respectively, and melanins that accumulate in husks and the pericarp under the control of *Blp1* cause the grain color to be black (Shoeva et al., 2018). These anthocyanins and melanins are promising functional food ingredients and are desirable in cultivars for human consumption (Matseychik et al., 2020; Loskutov, Khlestkina, 2021). Currently, the generation of new cultivars with distinct compositions of polyphenolic pigments in grain is an advanced task for plant breeders. To reach this goal, analysis of many specimens for certain compounds to choose adequate donors carrying the desired genes and screen the resulting hybrids is required.

Currently available methods of chemical profiling of polyphenols such as high-performance liquid chromatography (HPLC) or thin layer chromatography (TLC) allow to separate a large number of individual compounds (Vermerris, Nicholson, 2008), but they are unnecessarily time-consuming for preliminary screening of many specimens. In addition, they do not show histological patterns of polyphenol accumulation, which are known to be unevenly distributed within cereal grain and should be taken into account while processing grain for food (Barron et al., 2017). As an alternative to these methods, microscopy techniques can be used (Panato et al., 2017).

Among polyphenols, anthocyanins can be easily detected by microscopy owing to their color, which may vary from red to blue and their shades and depends on the chemical structure of the molecules, the presence of copigments, and metal ions (Davies, 2009). Change of color at different pH levels is a well-known characteristic feature of anthocyanins. In a strongly acidic medium (pH < 1), anthocyanins are in the cationic form, having a red color, whereas between pH 4 and 6, the cation loses two protons and turns blue (Davies, 2009). This pH-dependent color change of anthocyanins has been adapted to the detection and quantification of these compounds (Lee et al., 2005; Wrolstad et al., 2005).

Melanins are dark brown to black pigments formed in plants by oxidation of diverse phenolic precursors among which catechol, caffeic, chlorogenic, protocatechuic, *p*-coumaric and gallic acid have been reported (Bell, Wheeler, 1986; Solano, 2014; Varga et al., 2016). The polymeric nature and poor solubility of melanins constrains research on their chemical structure. In plants, melanins are the least studied group of pigments unlike melanins of the other kingdoms of living organisms, such as animals, bacteria, and fungi melanins, which were surveyed in details at the biochemical and molecular genetic levels (Britton, 1983; Solano, 2014; Glagoleva et al., 2020). It has been shown that melanins in barley grains accumulate in specific plastids called melanoplasts (Shoeva et al., 2020; Mursalimov et al., 2022).

Proanthocyanidins are colorless and require histochemical staining to be detected (Gardner, 1975; Treutter, 1989). An aromatic aldehyde *p*-dimethylaminocinnamaldehyde (DMACA) that binds to meta-oriented dihydroxy- or trihydroxy-substituted benzene rings is routinely used to detect proanthocyanidins and their immediate precursor molecules, namely, flavan-3,4-diols and flavan-3-ols, in plant tissues, including grains of cereals (Aastrup et al., 1984; Abeynayake et al., 2011; Kohyama et al., 2017; Zykin et al., 2020).

In barley, it can often be challenging to distinguish visually dark melanins from anthocyanins owing to a brownish hue they acquire during grain maturation (Glagoleva et al., 2022b). Furthermore, anthocyanins and melanins can accumulate in one grain simultaneously (Glagoleva et al., 2022a). Melanins are usually identified by a series of solubility tests, and spectroscopy techniques such as Fourier transform infrared (FT–IR), nuclear magnetic resonance (NMR), and the electron paramagnetic resonance (EPR) spectroscopy (Glagoleva et al., 2020). They dissolve in alkali solutions, discolor under the influence of strong oxidizing agents, react with $FeCl_3$ and are stable in common organic solvents that have been reported as hallmarks of melanin (Sava et al., 2001). Histochemical analysis of the melanin pigments in plant tissues is not common. For detection of melanins in human and animal tissues, the Fontana–Masson (FM) protocol with ammoniacal silver staining is commonly used. The protocol is based on the reduction of ammoniacal silver to metallic silver by phenolic substances. The product of the reaction is an insoluble black precipitate, which can be identified on tissue sections by light microscopy (Wildi, 1951; Bancroft, Gamble, 2008).

Here, we use the barley lines, previously characterized chemically for the presence of anthocyanins, proanthocyanins, and melanins, to demonstrate the ability of a new histochemical assay to differentiate these substances in grain tissue. Among the lines used in the study are Bowman (hereafter: Bw) backcross-derived near-isogenic lines (NILs) developed for the presence of anthocyanin (i:*BwAnt1Ant2*, hereafter: PLP) and melanin (i:*BwBlp1*, hereafter: BLP) compounds chemically identified by HPLC, a series of solubility tests, and Fourier transform infrared spectroscopy (Shoeva et al., 2020; Glagoleva et al., 2022c). The lines have been created by introgression of recombinant genetic fragments from colored donor lines into the genetic background of the parental cultivar Bowman, which lacks pigmentation in the grain (Druka et al., 2011). Another line used in the study is the mutant line *ant25.264*, which lacks proanthocyanidins in the grain. This line was induced by chemical mutagenesis of the parental cultivar Secobra 18193, which retained its ability to accumulate proanthocyanidins in the grain (Jende-Strid, 1993). NILs and mutant lines, which have relatively few genetic differences from their parental

cultivars, and are used extensively in genetic studies aiming to discover the association of phenotypic differences between the lines with specific chromosome regions or nucleotide polymorphisms. The use of these lines in studies of the genetic control of the synthesis of different classes of polyphenolic compounds in barley grains has been accompanied by their chemical profiling (Jende-Strid, 1993; Shoeva et al., 2020; Glagoleva et al., 2022c).

Here, taking advantage of the detailed characterization of these lines, a histochemical assay was developed to identify the different classes of polyphenolic compounds in barley grains. The described protocol may be employed for fast screening of a large number of samples and requires only a single grain for every sample. The method could be adapted to the grains/seeds/fruits of a wide variety of plant species. Here, as an example, besides barley, we successfully applied the assay to polyphenolic pigment profiling of black grains of wheat line s:S29*Ba1*4Th(4D)*Pp3*^{P}*Pp-D1*^{PF} (hereafter: S29BLACK) accumulating anthocyanins in the pericarp and aleurone layers simultaneously (Gordeeva et al., 2022), and common vetch cultivar Obskaya 16, having a black color of grain due to accumulation of anthocyaninis in the macrosclereids (Goncharova, 2020; Mursalimov et al., 2021).

Materials and methods

Plant material and growing conditions. The plant genetic resources utilized in the current study and their origin are listed in Table S1[1]. To obtain seeds for developing the protocol for histochemical polyphenolic profiling, cereal plants were grown in a greenhouse of the ICG SB RAS (Novosibirsk, Russia) under a 16 h photoperiod at a temperature range of 20–25 °C, whereas vetch was grown in an experimental field of the ICG SB RAS (Novosibirsk, Russia) in 2021.

The histochemical assay. The grains were sampled at the hard dough developmental stage (BBCH-87) from each of the aforementioned barley and wheat lines and at the brown pod stage for vetch grains. All of them were snap-frozen in liquid nitrogen and stored at –70 °C until analysis. Prior to sectioning, the frozen grains were kept at –20 °C for 30 min, mounted, and embedded in the Tissue-Tek O.C.T.TM compound medium (Sakura Finetek Europe B.V., Netherlands). The sectioning was carried out at –20 °C using an HM 505N cryostat microtome (Microm, Germany). Sections of 20 µm thickness were mounted on a poly-L-lysine slide (Thermo Fisher Scientific, Germany) and fixed by the addition of 1 mL of a precooled fixative (4 °C) onto the slide. Then, the slides were air dried at room temperature. For sections prepared for hot ethanol treatment, 80 % ethanol in water was utilized as a fixative, and 8 % formaldehyde (Sigma-Aldrich, USA) solution in phosphate-buffered saline (pH 7.4) was used as a fixative for the rest of the samples. All the slides were next rinsed for 15 min in distilled water to remove the fixative and embedding medium. After that, for hot ethanol treatment, the slides were transferred into 80 % (v/v) ethanol in water at 80 °C for 60 min incubation. The slides were then rinsed in distilled water and mounted in glycerol. For the pH test, the slides were immersed for a few seconds either in 1 M NaOH or 1 M HCl, quickly rinsed in distilled water and mounted in glycerol.

For ammoniacal silver staining, the FM protocol (Bancroft, Gamble, 2008; Kiernan, 2008) was performed. A stock solution was prepared by adding ammonium hydroxide solution drop by drop to a 10 % (w/v) silver nitrate water solution, until the solution yielded a precipitate and cleared again. After that, a 0.01 % working solution was prepared by dilution with distilled water. Next, the slides with the sections were incubated in the working solution for 15 min at 60 °C. The slides were then rinsed in distilled water and incubated in a 5 % (w/v) sodium thiosulfate water solution for 15 min, rinsed in distilled water again, and mounted in glycerol. The sections were analyzed under an Axio Observer Z1 microscope equipped with an AxioCam HRc camera (Zeiss, Germany). The ZEN software was used for image analysis and processing (Zeiss, Germany). The microscopic analysis was carried out at the Multi-Access Center for Microscopy of Biological Objects at the Institute of Cytology and Genetics, SB RAS (Novosibirsk, Russia). For each variety, at least three independent grains were analyzed, and a minimum of ten sections were examined for each grain.

Dimethylaminocinnamaldehyde (DMACA) staining. The sections fixed as described above were incubated overnight in a fresh solution of 0.5 % (w/v) DMACA (Sigma-Aldrich, Germany, kindly provided by Dr. Pavel Zykin from St. Petersburg State University, Russia) in 6 N HCl. The microscopic analysis was carried out as described above. For each variety, at least three independent grains were analyzed, and a minimum of ten sections were examined for each grain.

Results

Grain sections of three barley lines were analyzed by the proposed assay: the control unpigmented Bowman line, the PLP line with purple grains, and the BLP line with black grains. Five groups of sections were obtained for every grain: untreated or treated with hot ethanol, HCl, NaOH, or ammoniacal silver. All the resulting specimens were examined by brightfield microscopy (Fig. 1).

Grain tissues on sections of the control Bowman line did not have any visible pigmentation at the hard dough stage (Fig. 1*A*). There was no visible staining after hot ethanol treatment as well as after the pH change. The only noticeable dark staining appeared after ammoniacal silver treatment of the testa.

Untreated sections of the PLP line are characterized by accumulation of purple pigments in cells of the pericarp (Fig. 1*B*). The other cell types of this line do not have any pigmentation. The purple color disappeared totally after hot ethanol treatment, and the whole slice became colorless. At pH 1, the purple color persisted, whereas at pH 13, it turned bluish-green and notably leached into surrounding tissues. Intensive dark staining was observed in the pericarp, testa, and husk after ammoniacal silver treatment.

In the BLP line, brownish-black pigments were observed in untreated sections in the pericarp and husk (Fig. 1*C*). This pigmentation did not change after hot ethanol treatment or after the pH changes. After ammoniacal silver treatment, the

[1] Supplementary Table S1 is available at:
https://vavilovj-icg.ru/download/pict-2026-30/appx4.pdf

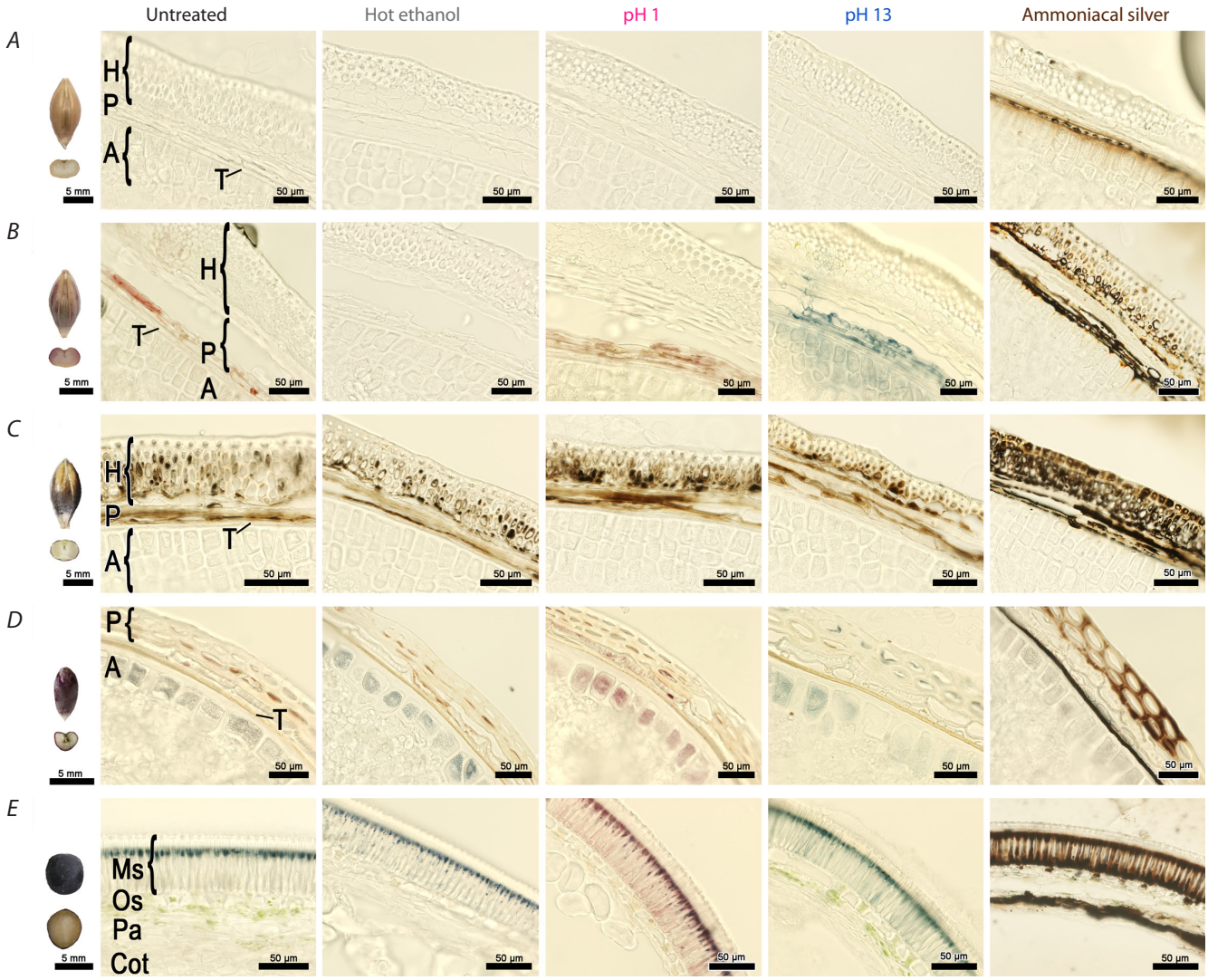


Fig. 1. Seeds and cross-sections of the seeds of barley, wheat and vetch after corresponding treatment (brightfield microscopy).
Horizontal lines: *A* – Bowman, *B* – PLP, *C* – BLP, *D* – S29BLACK, *E* – vetch. A: aleurone layer, H: husk, P: pericarp, T: testa, Cot: cotyledon, Ms: macrosclereids, Os: osteosclereids, Pa: parenchyma.

brownish-black pigmentation became even more intense in the pericarp and husk, where it was present before, and it emerged in the previously colorless testa and upper epidermal cells of the husk.

The sections prepared from wheat and vetch grains were assayed in the same manner. In black grains of wheat line S29BLACK, pale bluish-gray and purple pigments were revealed in aleurone and pericarp cells, respectively (Fig. 1*D*). The pigmentation did not disappear after hot ethanol treatment, and the bluish-gray color turned blue. At pH 1, blue aleurone cells became purple, while the purple pericarp cells did not change their color. At pH 13, the purple color of pericarp cells became blue, while the bluish-gray color of aleurone cells changed to bluish-green. Ammoniacal silver treatment gave rise to gray pigmentation in the aleurone, a brownish-black color in pericarp cells, and an intensive black color in the testa. It should be pointed out that only in wheat grains was the brownish-black color detected in cell walls of the pericarp.

In black vetch grains, blue pigmentation was observed in macrosclereids (Fig. 1*E*), becoming more intense after hot ethanol treatment and failing to be washed out. The blue pigments turned purple at pH 1 and became bluish-green at pH 13. In macrosclereids, ammoniacal silver stained the blue pigments brownish-black. In addition, the brownish-black pigmentation emerged in previously colorless parts of macrosclereids as well as in osteosclereids and cells of the parenchyma.

A comparison between ammoniacal silver staining and DMACA staining for revealing colorless polyphenols was performed on proanthocyanidin-less barley mutant line *ant25.264* and its parental cultivar Secobra 18193 containing proanthocyanidins in grains (Fig. 2).

There was no noticeable staining after the DMACA treatment of grains of *ant25.264*, whereas pale brown staining appeared in testa after ammoniacal silver treatment (Fig. 2). The brownish-purple staining was observed in the testa of the parental Secobra 18193 grains after DMACA treatment, and

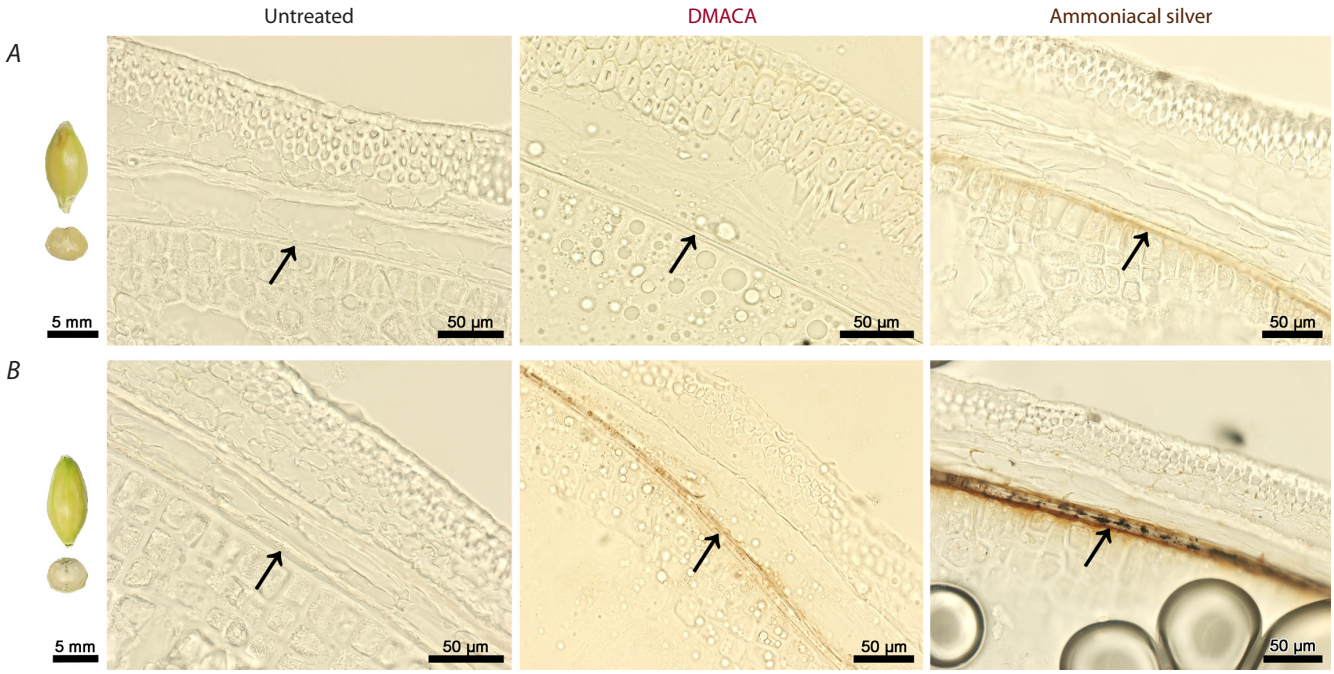


Fig. 2. Seeds and cross-sections of the seeds of anthocyanin-free (*A*) and anthocyanin-rich (*B*) grains of barley after DMACA and ammoniacal silver staining (brightfield microscopy).

Horizontal lines: *A* – *ant25.264*, *B* – Secobra 18193. Arrows indicate testa.

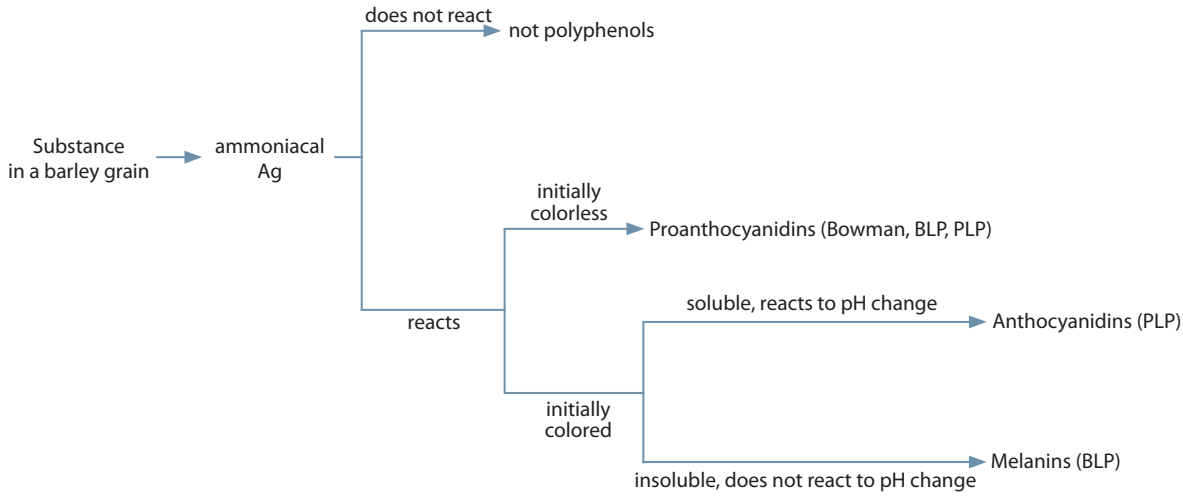


Fig. 3. The scheme for identification of polyphenols in barley grains.

in the same tissue, intense black staining was registered after ammoniacal silver treatment (Fig. 2).

After summarizing the results of all the above-mentioned tests, we built a scheme for identifying anthocyanins, melanins, and proanthocyanidins in barley grains (Fig. 3).

Discussion

The analysis of polyphenolic profiles of crops is quite important for both the elucidation of fundamental aspects of polyphenol biosynthesis and for industrial applications of the raw material. Some attempts have been made to find a fast and simple way to perform polyphenolic profiling of a large number of samples, such as high-throughput phenotyping approaches based on image analysis of color characteristics (ElMasry et al., 2019; Komyshev et al., 2023). Nevertheless, histochemical techniques combined with microscopy seem to be the only approach nowadays allowing a fast and precise polyphenolic profiling on a limited amount of the experimental material and providing information about the cell type where these polyphenols accumulate. Our assay involving a series of standard chemical tests applied to cryosections followed by microscopic examination successfully showed the possibility of identifying polyphenolic pigments in barley, wheat, and vetch grains.

Hot ethanol treatment, which is a common test for solubility (Takahashi, Hara, 2014), allowed us to easily distinguish soluble anthocyanins in PLP grains from insoluble melanins in BLP grains. It is known that brown-pigment melanins necessitate a special harsh treatment to become soluble and are insoluble under the hot ethanol conditions we used (Sava et al., 2001). Nevertheless, not all anthocyanins are soluble. Melanins can be distinguished from insoluble anthocyanins by the absence of a color transition at changing pH, which is a well-known characteristic feature of anthocyanins (Davies, 2009).

Thus, in the PLP line, the solubility and pH tests enabled us to identify the purple pigments – that completely disappeared during hot ethanol treatment and changed color to blue in the presence of sodium hydroxide – as anthocyanins, and to identify the BLP line's black pigments – that did not react to hot ethanol treatment and changes in pH – as melanins. The same results have been previously obtained by HPLC, a series of solubility tests, and Fourier transform infrared spectroscopy, when extracts from PLP and BLP grains have been assayed (Shoeva et al., 2020; Glagoleva et al., 2022c).

As a first step, a FM protocol was included in our histochemical assay to confirm the presence of melanins, which do not react to hot ethanol treatment and a pH change. The FM protocol is commonly used for the detection of melanins in human and animal tissues (Bancroft, Gamble, 2008; Kiernan, 2008) and has not been used previously for the detection of melanins in plant tissues. It turned out that the FM protocol allows the detection of not only melanins by staining them deep black but also of other polyphenols in a grain, including colorless ones, owing to their common ability to reduce ammoniacal silver to metallic one. It is known that silver nitrate reacts with aromatic compounds that contain two or more phenolic hydroxyl groups (Wildi, 1951).

In the barley grain, black melanins, purple anthocyanins, and colorless proanthocyanidins can be detected by ammoniacal silver staining. In grains of the BLP line, the melanin-containing tissues (pericarp and husk) were expectedly stained by ammoniacal silver; however, metallic silver was also noted in the testa, implying the presence of hidden polyphenols. Ammoniacal silver staining in the BLP line was the strongest in comparison to lines Bowman and PLP, thus pointing to the presence of a much greater amount of polyphenols in BLP. The higher accumulation of phenolic compounds in the BLP line has also been observed elsewhere by HPLC analysis (Shoeva et al., 2016). Moreover, for the black-grained BLP line, the highest antioxidant activity has been documented (Glagoleva et al., 2017). The ability of melanins and colorless polyphenols to reduce ammoniacal silver may be explained by their well-known antioxidant properties (Quideau et al., 2011; Panzella et al., 2012). In the PLP line, ammoniacal silver strongly stained both visible anthocyanin pigments and previously colorless tissues, but less strongly than in the BLP line. In all three barley lines, ammoniacal silver stained colorless polyphenols in the testa. It was especially noticeable in noncolored control Bowman grains, which do not contain any visible pigments and do not react to hot ethanol and to changes in pH but show prominent black granules after ammoniacal silver staining in only one tissue: the testa. It is in this tissue of the barley grain that colorless proanthocyanidins are synthesized (Jende-Strid, 1993).

To evaluate the efficiency of ammoniacal silver staining at detecting proanthocyanidins, it was compared with the DMACA staining. Both methods were applied to barley grain sections enriched with and low in proanthocyanidins. We observed the better sensitivity of ammoniacal silver staining: DMACA did not detect any proanthocyanidins in proanthocyanidin-less *ant25.264*, while the ammoniacal silver reaction in its turn was weak but positive. It is also important to state that DMACA staining is performed in 6 M HCl, and HCl negatively affects tissue preservation. Moreover, owing to the low pH of the DMACA solution, this reagent must not be used in the presence of anthocyanins because they could change their color under such conditions and yield a false positive result. Due to the aforementioned problem, we could not apply DMACA staining to the detection of proanthocyanidins in grains of PLP barley, S29BLACK wheat, and vetch. The results allowed us to conclude that ammoniacal silver staining has some advantages against DMACA staining and could replace it to locate proanthocyanidins in grains.

The observations made on the barley grain sections treated with hot ethanol, HCl, NaOH, and ammoniacal silver helped us to develop a protocol for the identification of polyphenolic pigments in barley grains; it is presented as a scheme in Figure 3. The relevance of the protocol was confirmed by chromatography and independent chemical tests reported previously for the same barley lines (Shoeva et al., 2020; Glagoleva et al., 2022c). It is worth noting that this protocol was devised for barley. For polyphenolic profiling of grains of other plant species in the same manner, it is desirable to verify the reliability of the proposed histochemical techniques by precise analytical chemistry approaches. Nevertheless, for a preliminary analysis, the proposed assay could be used in its current state even for different species. Here we extended the application of this assay to polyphenol profiling of darkly pigmented grains of wheat and vetch. Wheat line S29BLACK was constructed by marker-assisted selection for the presence of genes *ThMyc4E* and *Pp*, which control the synthesis of blue and purple anthocyanins in the aleurone and pericarp (Gordeeva et al., 2022). The black color of grains of the vetch line we analyzed in the current work has been ascribed to the blue pigments accumulating in macrosclereids; however, the precise chemical composition has not been determined (Mursalimov et al., 2021).

Our assay revealed a sufficient amount of polyphenols in testa and pericarps of the tested wheat grains, which yielded positive staining with ammoniacal silver. This approach allowed us to detect deposition of the polyphenols in cell walls of the wheat grains pericarp with the bound phenolic acids among which dehydrodiferulates have been reported as predominant (Parker et al., 2005). Abundant colorless polyphenols were also detected by ammoniacal silver staining in seed envelopes of vetch; this feature may affect feeding properties of this crop and deserves deeper research too. Unlike purple pigments in the barley pericarp, the purple pigments in the wheat pericarp are insoluble and are not washed out by hot ethanol; however,

they alter their color after a pH change, thereby enabling us to identify these purple pigments as anthocyanins. Previously, cyanidin-based compounds have been identified in the wheat pericarp as major pigments, among which cyanidin-3-glucoside, cyanidin-3-rutinoside, and peonidin-3-glucoside are predominant (Abdel-Aal et al., 2006; Ficco et al., 2014). Insoluble blue pigments have also been observed in the aleurone layer of wheat and in macrosclereids of vetch. They have been identified as anthocyanins owing to their color transition from blue (original color) to red in an acidic medium. In blue-grained wheat, delphinidin-derived anthocyanins – delphinidin-3-glucoside and delphinidin-3-rutinoside – have been identified as predominant anthocyanins (Knievel et al., 2009; Abdel-Aal et al., 2016), whereas in vetch, the precise chemical structure of the blue pigments has not been determined yet. Judging by the results of our test, we can predict that vetch grains accumulate similar blue delphinidin-derived anthocyanins.

It also should be noted that differential behavior of anthocyanins during hot ethanol treatment may be attributed not only to their distinct chemical structure but also to the type of cells and organelles accumulating these pigments. For example, in wheat, blue anthocyanins accumulate in aleurone cells' vacuoles. There are spherical particles called aleurone granules in aleurone cells, which are either phytate inclusions (type 1: composed of phytic acid minerals) or niacin inclusions (type 2: composed of niacin and proteins), each granule being surrounded by a fine layer of lipidic droplets (Morrison et al., 1975; Yu et al., 2021). The anthocyanins accumulated in aleurone cells may form complexes with metal ions or interact with the proteins; both these properties are attributed to anthocyanins and reportedly improve stability of the molecules (Li et al., 2021).

Conclusion

In the current work, the histochemical assay consisting of a series of standard chemical tests applied to cryosections followed by microscopic examination successfully showed the possibility of distinguishing polyphenolic pigments and uncolored polyphenols in barley, wheat, and vetch grains. However, the extrapolation of results obtained on barley (and confirmed by analytical chemistry approaches) to other species requires caution, and further chemical analyses are necessary for exact identification of polyphenols. Nevertheless, the proposed histochemical assay has a high potential for polyphenolic profiling as part of preliminary screening of grains of barley and many other species where only a single grain is required. After the proposed preliminary screening, the selected lines/varieties could be propagated and analyzed in more detail for the presence of certain chemicals by more precise quantitative and qualitative methods.

References

Aastrup S., Outtrup H., Erdal K. Location of the proanthocyanidins in the barley grain. *Carlsberg Res Commun.* 1984;49:105-109. doi 10.1007/BF02913969

Abdel-Aal el-S.M., Young J.C., Rabalski I. Anthocyanin composition in black, blue, pink, purple, and red cereal grains. *J Agric Food Chem.* 2006;54(13):4696-4704. doi 10.1021/jf0606609

Abdel-Aal el-S.M., Hucl P., Shipp J., Rabalski I. Compositional differences in anthocyanins from blue- and purple-grained spring wheat grown in four environments in central Saskatchewan. *Cereal Chem J.* 2016;93(1):32-38. doi 10.1094/CCHEM-03-15-0058-R

Abeynayake S.W., Panter S., Mouradov A., Spangenberg G. A high-resolution method for the localization of proanthocyanidins in plant tissues. *Plant Methods.* 2011;7(1):13. doi 10.1186/1746-4811-7-13

Bancroft J.D., Gamble M. (Eds) Theory and Practice of Histological Techniques. Edinburgh: Churchill Livingstone, 2008. doi 10.1016/B978-0-443-10279-0.50001-4

Barron C., Holopainen-Mantila U., Sahlstrom S., Hotekjolen A.K., Lullien-Pellerin V. Assessment of biochemical markers identified in wheat for monitoring barley grain tissue. *J Cereal Sci.* 2017;74:11-18. doi 10.1016/j.jcs.2017.01.004

Bell A.A., Wheeler M.H. Biosynthesis and functions of fungal melanins. *Annu Rev Phytopathol.* 1986;24:411-451. doi 10.1146/annurev.py.24.090186.002211

Britton G. The Biochemistry of Natural Pigments. Cambridge Univ. Press, 1983

Davies K.M. Modifying anthocyanin production in flowers. In: Winefield C., Davies K., Gould K. (Eds) Anthocyanins: Biosynthesis, Functions, and Applications. Springer, 2009;49-80. doi 10.1007/978-0-387-77335-3_3

Debeaujon I., Léon-Kloosterziel K.M., Koornneef M. Influence of the testa on seed dormancy, germination, and longevity in *Arabidopsis*. *Plant Physiol.* 2000;122(2):403-414. doi 10.1104/pp.122.2.403

Druka A., Franckowiak J., Lundqvist U., Bonar N., Alexander J., Houston K., Radovic S., Shahinnia F., Vendramin V., Morgante M., Stein N., Waugh R. Genetic dissection of barley morphology and development. *Plant Physiol.* 2011;155(2):617-627. doi 10.1104/pp.110.166249

ElMasry G., Mandour N., Al-Rejaie S., Belin E., Rousseau D. Recent applications of multispectral imaging in seed phenotyping and quality monitoring – an overview. *Sensors.* 2019;19(5):1090. doi 10.3390/s19051090

Ficco D.B.M., De Simone V., Colecchia S.A., Pecorella I., Platani C., Nigro F., Finocchiaro F., Papa R., De Vita P. Genetic variability in anthocyanin composition and nutritional properties of blue, purple, and red bread (*Triticum aestivum* L.) and durum (*Triticum turgidum* L. ssp. *turgidum* convar. *durum*) wheats. *J Agric Food Chem.* 2014;62(34):8686-8695. doi 10.1021/jf5003683

Gardner R.O. Vanillin-hydrochloric acid as a histochemical test for tannin. *Stain Technol.* 1975;50(5):315-317. doi 10.3109/10520297509117081

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

Glagoleva A.Y., Shoeva O.Y., Khlestkina E.K. Melanin pigment in plants: current knowledge and future perspectives. *Front Plant Sci.* 2020;11:770. doi 10.3389/fpls.2020.00770

Glagoleva A., Kukoeva T., Mursalimov S., Khlestkina E., Shoeva O. Effects of combining the genes controlling anthocyanin and melanin synthesis in the barley grain on pigment accumulation and plant development. *Agronomy.* 2022a;12(1):112. doi 10.3390/agronomy12010112

Glagoleva A.Y., Novokreschenov L.A., Shoeva O.Y., Kovaleva O.N., Khlestkina E.K. Studying grain color diversity in the barley collection of VIR. *Proc Appl Bot Genet Breed*. 2022b;183(3):76-84. doi 10.30901/2227-8834-2022-3-76-84 (in Russian)

Glagoleva A.Y., Vikhorev A.V., Shmakov N.A., Morozov S.V., Chernyak E.I., Vasiliev G.V., Shatskaya N.V., Khlestkina E.K., Shoeva O.Y. Features of activity of the phenylpropanoid biosynthesis pathway in melanin-accumulating barley grains. *Front Plant Sci.* 2022c;13:923717. doi 10.3389/fpls.2022.923717

Goncharova A.V. Spring common vetch sowing cultivar Obskaya 16. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Lett Vavilov J Genet Breed*. 2020;6(1):15-17. doi 10.18699/Letters2020-6-03 (in Russian)

Gordeeva E., Shoeva O., Mursalimov S., Adonina I., Khlestkina E. Fine points of marker-assisted pyramiding of anthocyanin biosynthesis regulatory genes for the creation of black-grained bread wheat (*Triticum aestivum* L.) lines. *Agronomy.* 2022;12:2934. doi 10.3390/agronomy12122934

Jende-Strid B. Genetic control of flavonoid biosynthesis in barley. *Hereditas.* 1993;119(2):187-204. doi 10.1111/j.1601-5223.1993.00187.x

Kiernan J. Histological and Histochemical Methods: Theory and Practice. Cold Spring Harbor, 2008

Knievel D.C., Abdel-Aal el-S.M., Rabalski I., Nakamura T., Hucl P. Grain color development and the inheritance of high anthocyanin blue aleurone and purple pericarp in spring wheat (*Triticum aestivum* L.). *J Cereal Sci.* 2009;50(1):113-120. doi 10.1016/j.jcs.2009.03.007

Kohyama N., Chono M., Nakagawa H., Matsuo Y., Ono H., Matsunaka H. Flavonoid compounds related to seed coat color of wheat. *Biosci Biotechnol Biochem.* 2017;81(11):2112-2118. doi 10.1080/09168451.2017.1373589

Komyshev E.G., Genaev M.A., Busov I.D., Kozhekin M.V., Artemenko N.V., Glagoleva A.Y., Koval V.S., Afonnikov D.A. Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(7):859-868. doi 10.18699/VJGB-23-99

Lee J., Durst R.W., Wrolstad R.E. Determination of total monomeric anthocyanin pigment content of fruit juices, beverages, natural colorants, and wines by the pH differential method. *J AOAC Int.* 2005;88(5):1269-1278. doi 10.1093/jaoac/88.5.1269

Li J., Wang B., He Y., Wen L., Nan H., Zheng F., Liu H., Lu S., Wu M., Zhang H. A review of the interaction between anthocyanins and proteins. *Food Sci Technol Int.* 2021;27(5):470-482. doi 10.1177/1082013220962613

Loskutov I.G., Khlestkina E.K. Wheat, barley, and oat breeding for health benefit components in grain. *Plants.* 2021;10(1):86. doi 10.3390/plants10010086

Matseychik I.V., Korpacheva S.M., Lomovsky I.O., Serasutdinova K.R. Prospects for the use of products of processing of buckwheat as functional ingredients. *Technol Merchandising Innovative Foodstuff.* 2020;61(2):53-57. doi 10.33979/2219-8466-2020-61-2-53-57 (in Russian)

Morrison I.N., Kuo J., O'Brien T.P. Histochemistry and fine structure of developing wheat aleurone cells. *Planta.* 1975;123(2):105-116. doi 10.1007/BF00383859

Mursalimov S.R., Goncharova A.V., Glagoleva A.Yu., Shoeva O.Yu. Black seed color of the spring common vetch (*Vicia sativa* L.) cultivar Obskaya 16 is caused by blue anthocyanins accumulating in macrosclereids. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Lett Vavilov J Genet Breed*. 2021;7(2):91-95. doi 10.18699/Letters-VJ2021-7-10

Mursalimov S., Glagoleva A., Khlestkina E., Shoeva O. Chlorophyll deficiency delays but does not prevent melanogenesis in barley seed melanoplasts. *Protoplasma.* 2022;259(2):317-326. doi 10.1007/s00709-021-01669-3

Panato A., Antonini E., Bortolotti F., Ninfali P. The histology of grain caryopses for nutrient location: a comparative study of six cereals. *Int J Food Sci Technol.* 2017;52(5):1238-1245. doi 10.1111/ijfs.13390

Panzella L., Eidenberger T., Napolitano A., d'Ischia M. Black sesame pigment: DPPH assay-guided purification, antioxidant/antinitrosating properties, and identification of a degradative structural marker. *J Agric Food Chem.* 2012;60(36):8895-8901. doi 10.1021/jf2053096

Parker M.L., Ng A., Waldron K.W. The phenolic acid and polysaccharide composition of cell walls of bran layers of mature wheat (*Triticum aestivum* L. cv. Avalon) grains. *J Sci Food Agric.* 2005;85(15):2539-2547. doi 10.1002/jsfa.2304

Quideau S., Deffieux D., Douat-Casassus C., Pouységu L. Plant polyphenols: chemical properties, biological activities, and synthesis. *Angew Chem Int.* 2011;50(3):586-621. doi 10.1002/anie.201000044

Sava V.M., Yang S.-M., Hong M.-Y., Yang P.-C., Huang G.S. Isolation and characterization of melanic pigments derived from tea and tea polyphenols. *Food Chem.* 2001;73(2):177-184. doi 10.1016/S0308-8146(00)00258-2

Shoeva O.Y., Mock H.-P., Kukoeva T.V., Börner A., Khlestkina E.K. Regulation of the flavonoid biosynthesis pathway genes in purple and black grains of *Hordeum vulgare*. *PLoS One.* 2016;11(10):e0163782. doi 10.1371/journal.pone.0163782

Shoeva O.Yu., Strygina K.V., Khlestkina E.K. Genes determining the synthesis of flavonoid and melanin pigments in barley. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2018;22(3):333-342. doi 10.18699/VJ18.369 (in Russian)

Shoeva O.Y., Mursalimov S.R., Gracheva N.V., Glagoleva A.Y., Börner A., Khlestkina E.K. Melanin formation in barley grain occurs within plastids of pericarp and husk cells. *Sci Rep.* 2020;10:179. doi 10.1038/s41598-019-56982-y

Solano F. Melanins: skin pigments and much more – types, structural models, biological functions, and formation routes. *New J Sci.* 2014;5:1-28. doi 10.1155/2014/498276

Takahashi I., Hara M. Enhancement of starch accumulation in plants by exogenously applied methyl jasmonate. *Plant Biotechnol Rep.* 2014;8:143-149. doi 10.1007/s11816-013-0304-1

Treutter D. Chemical reaction detection of catechins and proanthocyanidins with 4-dimethylaminocinnamaldehyde. *J Chromatogr.* 1989;467(1):185-193. doi 10.1016/S0021-9673(01)93963-9

Varga M., Berkesi O., Darula Z., May N.V., Palágyi A. Structural characterization of allomelanin from black oat. *Phytochemistry.* 2016;130:313-320. doi 10.1016/j.phytochem.2016.07.002

Vermerris W., Nicholson R. Phenolic compounds and their effects on human health. In: Phenolic Compound Biochemistry. Springer, 2008;235-255. doi 10.1007/978-1-4020-5164-7_7

von Wettstein D. From analysis of mutants to genetic engineering. *Annu Rev Plant Biol.* 2007;58:1-19. doi 10.1146/annurev.arplant.58.032806.104003

Wildi B.S. Silver nitrate as a test for ortho and para dihydric phenols. *Science.* 1951;113(2929):188. doi 10.1126/science.113.2929.188

Winkel-Shirley B. Flavonoids in seeds and grains: physiological function, agronomic importance and the genetics of biosynthesis. *Seed Sci Res.* 1998;8(4):415-422. doi 10.1017/S0960258500004372

Wrolstad R.E., Durst R.W., Lee J. Tracking color and pigment changes in anthocyanin products. *Trends Food Sci Technol.* 2005;16(9):423-428. doi 10.1016/j.tifs.2005.03.019

Yu R., Wu X., Liu J., Howitt C.A., Bird A.R., Liu C.M., Larkin P.J. Rice with multilayer aleurone: a larger sink for multiple micronutrients. *Rice.* 2021;14(1):102. doi 10.1186/s12284-021-00543-3

Zykin P.A., Andreeva E.A., Tsvetkova N.V., Voylokov A.V. Anatomical and image analysis of grain coloration in rye. *Preprints.* 2020. doi 10.20944/preprints202011.0530.v1

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

Received May 25, 2025. Revised September 1, 2025. Accepted September 24, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):61-71

doi 10.18699/vjgb-26-03

Modification of the BphP1-QPAS1 optogenetic system for gene expression regulation in *Nicotiana benthamiana* tobacco leaves using near-infrared light

E.S. Surkova [1, 2 $], Y.A. Galimova [1 $], N.V. Battulina [1], D.M. Motorina [1], E.S. Omelina [1]

[1] Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
omelina@mcb.nsc.ru

Abstract. In plants, the regulation of transgene transcription is typically achieved using chemical agents. A safe alternative to chemically induced systems may be optogenetic systems. The BphP1-QPAS1 system has distinct advantages over other optogenetic systems, as it is activated by near-infrared (NIR, 780 nm) light, which is beyond the spectrum of plant photoreceptors. This system is based on the use of a split transcription factor (TF), consisting of the DNA-binding and dimerization domains of the yeast TF Gal4, fused to the QPAS1 component, along with the transactivation domain VP16 fused to BphP1. Under NIR light, BphP1 interacts with QPAS1, leading to the formation of the functional TF Gal4-VP16. A primary obstacle to using optogenetic systems in plants is their undesired activation under white light, which is vital for normal plant growth. A potential solution to this issue is temporarily removing one component of the split TF from the nucleus under white light. We modified the BphP1-QPAS1 system to activate reporter gene expression in *Nicotiana benthamiana* leaves using NIR light. We combined BphP1-QPAS1 with several variants of LOV domain-containing proteins activated by blue light (460–480 nm). The best results were achieved by combining the BphP1-QPAS1 system with the *As*LOV2 domain, which carries the degron sequence RRRG at the C-terminal Jα helix and initiates the degradation of the chimeric protein NES-Gal4-QPAS1-*As*LOV2-RRRG under white light. This modification induced the BphP1-QPAS1 system in tobacco leaves only under NIR light, but not in the dark or under white light. We believe that, in the future, the BphP1-QPAS1 system could be applied to enhance plant resistance to adverse environmental conditions, pests, and viral diseases.

Key words: optogenetic system; BphP1-QPAS1; VVD; *As*LOV2; *Nicotiana benthamiana*; gene expression regulation; Gal4/UAS

For citation: Surkova E.S., Galimova Y.A., Battulina N.V., Motorina D.M., Omelina E.S. Modification of the BphP1-QPAS1 optogenetic system for gene expression regulation in *Nicotiana benthamiana* tobacco leaves using near-infrared light. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):61-71. doi 10.18699/vjgb-26-03

Funding. This research was funded by the Russian Science Foundation, grant No. 22-74-10118.

Модификация оптогенетической системы BphP1-QPAS1 для регуляции экспрессии генов в листьях табака *Nicotiana benthamiana* с помощью ближнего инфракрасного света

Э.С. Суркова [1, 2 $], Ю.А. Галимова [1 $], Н.В. Баттулина [1], Д.М. Моторина [1], Е.С. Омелина [1]

[1] Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
omelina@mcb.nsc.ru

Аннотация. В растениях регуляция транскрипции трансгенов обычно осуществляется с помощью химических агентов. Безопасной альтернативой химически индуцируемым системам могут служить оптогенетические (т. е. светоиндуцируемые) системы. Бактериальная система BphP1-QPAS1 выгодно отличается от других оптогенетических систем, поскольку активируется ближним инфракрасным светом (БИК, 780 нм), выходящим за пределы спектра восприятия растительных фоторецепторов. Система BphP1-QPAS1 основана на использовании «расщепленного» транскрипционного фактора (ТФ), состоящего из ДНК-связывающего и димеризующего доменов дрожжевого ТФ Gal4, слитого с компонентом QPAS1, и трансактивационного домена VP16, слитого с BphP1. Под действием БИК света BphP1 взаимодействует с QPAS1, что приводит к сближению Gal4 с VP16 и запуску экспрессии репортерного гена. Основным препятствием для широкого использования оптогенетических систем в растениях является нежелательная активация таких систем при воздействии дневного (белого) света, который включает широкий спектр длин волн и жизненно важен для нормального роста растений. Решением проблемы нежелательной активации может стать временное удаление из ядра клетки одного из компонентов

© Surkova E.S., Galimova Y.A., Battulina N.V., Motorina D.M., Omelina E.S., 2026

$ These authors contributed equally.

This work is licensed under a Creative Commons Attribution 4.0 License

«расщепленного» ТФ при воздействии белого света. В данной работе мы модифицировали систему BphP1-QPAS1 для активации экспрессии репортерного гена в листьях табака *Nicotiana benthamiana* с помощью БИК света. Для этого мы комбинировали систему BphP1-QPAS1 с несколькими вариантами LOV домен-содержащих белков, активируемых синим светом (460–480 нм). Наилучшие результаты были достигнуты при комбинации компонентов системы BphP1-QPAS1 с доменом *As*LOV2 из белка фототропин 1 *Avena sativa*, несущим на С концевой Jα спирали последовательность дегрона RRRG и запускающим деградацию химерного белка NES-Gal4-QPAS1-*As*LOV2-RRRG при воздействии белого света. Такая модификация активировала систему BphP1-QPAS1 в листьях табака только при воздействии БИК света, но не в темноте либо при белом свете. Мы полагаем, что в будущем система BphP1-QPAS1 может быть применена для повышения устойчивости растений к неблагоприятным условиям среды, вредителям, вирусным заболеваниям.

Ключевые слова: оптогенетическая система; BphP1-QPAS1; VVD; *As*LOV2; *Nicotiana benthamiana*; регуляция экспрессии генов; Gal4/UAS

Introduction

The use of non-channel light-sensitive proteins, which can change conformation or form homo- or heterodimeric complexes under specific wavelengths of light, has enabled the application of optogenetic systems for precise temporal transcription regulation (Shimizu-Sato et al., 2002; Wang et al., 2012; Konermann et al., 2013; Ochoa-Fernandez et al., 2016; Gligorovski et al., 2023). Light control may serve as a valuable alternative to chemically inducible gene expression systems. The advantages of optogenetic systems over chemical ones include the following: (1) the ability to focus light on a specific small area of an organism or cell; (2) control over the intensity and duration of light exposure; (3) activation or deactivation of the system by switching the light On or Off, respectively; (4) independence from the diffusion rate of a chemical agent; and (5) absence of Off-target effects caused by chemical inducers (Shimizu-Sato et al., 2002; Omelina et al., 2022).

One of the strategies for regulating expression by light is based on photoreversible conformational changes (Fig. 1*A*, *B*). For instance, light illumination may induce the unfolding of one of the α-helices in the LOV domain-containing proteins (Crosson, Moffat, 2001; Zayner et al., 2012), which can be used to expose an engineered peptide that controls intracellular protein localization or protein-protein interactions (Krueger et al., 2019). Another strategy is the use of split transcription factors (TFs), where one component is a DNA-binding domain of a TF fused to a light-sensitive protein, while its partner is fused to an activator or repressor domain. When exposed to light, heterodimers are formed, resulting in the generation of an active transcription factor (Fig. 1*C*, *D*) (Pathak et al., 2017; Noda, Ozawa, 2018; Omelina, Pindyurin, 2019; Hernández-Candia, Tucker, 2020).

One of the most popular systems for the activation and suppression of gene transcription is the LOV domain-containing proteins. LOV domains belong to the Per-Arnt-Sim (PAS) protein family (Herrou, Crosson, 2011). The small size (~11–15 kDa) and the presence of flavin chromophores in most, if not all, cell types are key advantages of these domains for the design of optogenetic tools (Wu Y.I. et al., 2009; Losi, Gärtner, 2011; Christie et al., 2012; Renicke et al., 2013; Niopek et al., 2016; Redchuk et al., 2017). One of the strategies for developing LOV-derived optogenetic tools is based on light-induced homodimerization. This approach utilises the LOV domain-containing protein Vivid (VVD) from the filamentous fungus *Neurospora crassa* (Wang et al., 2012). VVD has an N-terminal helix that is docked on the core in the dark state. Blue light (460–480 nm) leads to the dissociation of this helix from the VVD core and interaction with the N-terminal helix of another VVD, resulting in the formation of a VVD homodimer (Fig. 1*A*) (Zoltowski et al., 2007; Zoltowski, Crane, 2008). The VVD-based strategy was used to develop the LightON system, providing blue-light-induced transcription of target transgenes in mammalian cells and in mice (Wang et al., 2012). VVD was fused with the DNA-binding domain of the yeast TF Gal4 and the p65 transcription activation domain. It formed homodimers under blue light, resulting in the generation of the active dimer Gal4-p65 and the transcriptional activation of the UAS reporter genes. Dark relaxation caused dimer dissociation, meaning the system returned to the non-active state.

Another LOV domain is derived from the LOV2 domain of *Avena sativa* phototropin 1 (*As*LOV2). In LOV2, illumination with blue light leads to the unwinding of the caged C-terminal Jα helix and the exposure of the fused C-terminal peptide, thus making it available for protein-protein interactions (Peter et al., 2010; Lungu et al., 2012; Strickland et al., 2012; Yumerefendi et al., 2015; Kim et al., 2024; Kaya et al., 2025). This property of the LOV2 domain has been utilized to create a light-inducible nuclear export system (LEXY) (Niopek et al., 2016). In this system, blue light releases the photocaged nuclear export signal (NES) from the *As*LOV2 core, thereby inducing the export of the protein of interest from the nucleus (Fig. 1*B*). To create this system, a library of 33 different NESs was generated, providing varying efficiencies of nuclear export for the chimeric protein NLS-mCherry-LEXY from the nucleus to the cytoplasm under blue light. This system has been tested in various mammalian cell cultures, and it has been demonstrated that during the dark relaxation period, the protein NLS-mCherry-LEXY returns to the nucleus, while the export of the protein from the nucleus can be repeatedly induced under blue light. An interesting application of LEXY is the spatiotemporal control of transcriptional repressors by sequestering them into the cytosol, which could be of particular value for controlling reporter gene activity (Fig. 1*B*) (Niopek et al., 2016).

Phytochromes are also a widely used optogenetic tool for the activation and suppression of gene transcription. These photoreceptors are found in plants, bacteria, cyanobacteria, and fungi, playing essential roles in light-adaptive processes. They share common domains in a photosensory core module and phytochrome-specific domains (Wagner et al., 2007; Yang et al., 2008; Chernov et al., 2017). Bacterial and fungal phytochromes incorporate a tetrapyrrole biliverdin IXα (BV)

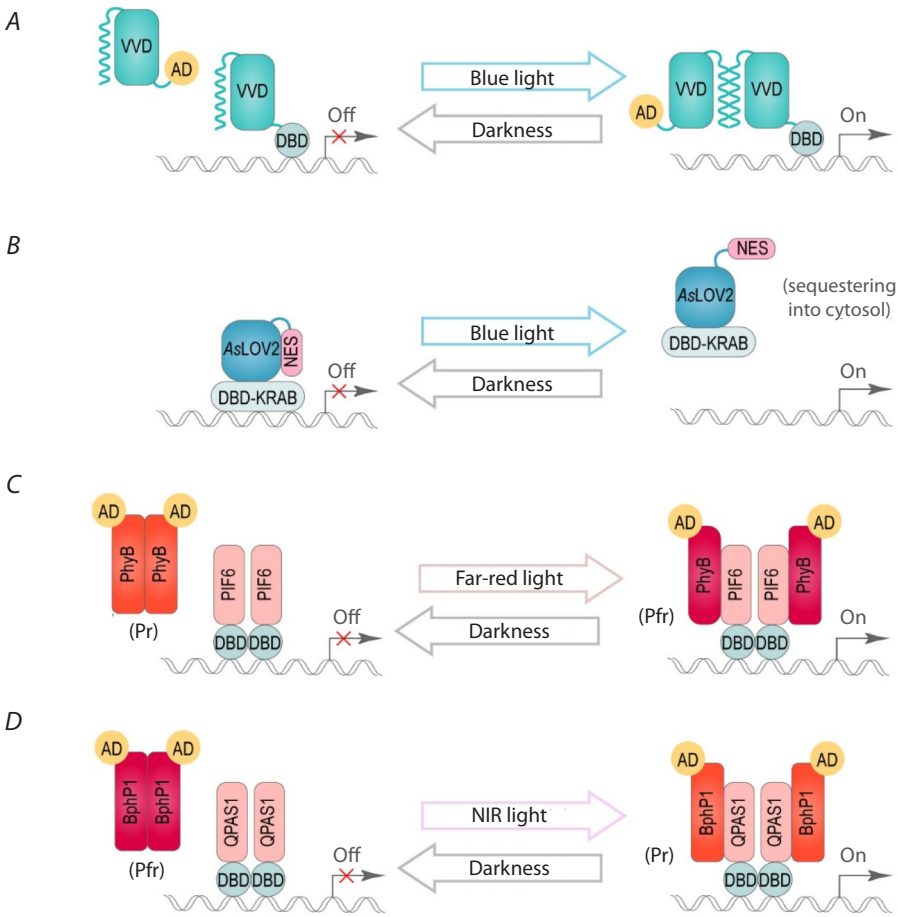


Fig. 1. Different strategies for regulating gene expression using light.

A, A schematic representation of the formation of homodimers due to conformational changes in the LOV domain of the VVD protein induced by blue light illumination. The N-terminal helix dissociates from the VVD core and interacts with the N-terminal helix of another VVD under blue light. As a result, the DBD and AD of the split TF associate into a functional TF, leading to the transcription of the reporter gene. *B*, A schematic representation of the unfolding of the C-terminal Jα helix from the *As*LOV2 domain and the exposure of the fused NES signal, leading to the relocalization of the transcriptional repressor (DBD-KRAB) from the nucleus into the cytoplasm to control gene activity. *C*, *D*, A schematic representation of the formation of heterodimers PhyB-PIF6 under far-red light (*C*) or BphP1-QPAS1 under NIR light (*D*), and their dissociation in darkness. Upon light induction, the DBD interacts with the AD, resulting in the transcription of the reporter gene. The Pr and Pfr states refer to two interconvertible forms of phytochromes. AD, activating domain; DBD, DNA-binding domain; KRAB, Kruppel associated box; NES, nuclear export signal.

as a chromophore. In plants and cyanobacteria, BV is further enzymatically reduced to phytochromobilin (PΦB) or phycocyanobilin (PCB), which bind to plant and cyanobacterial phytochromes, respectively. Phytochromes exist in one of two interconvertible states: either the Pr state or the Pfr state (Fig. 1*C*, *D*). The Pr state absorbs light at 660–700 nm, whereas the Pfr state absorbs light at 740–780 nm (Li J. et al., 2011; Chernov et al., 2017). Typically, the ground (inactive) state of phytochromes is Pr, which converts to the active Pfr state upon illumination with 660–680 nm light (Fig. 1*C*) (Shcherbakova et al., 2015). A specific group of phytochromes, termed bathy, adopts Pfr as the ground state (Fig. 1*D*). Many photoreversible dimerization tools have been developed based on the light-regulated interaction of plant phytochrome with basic helix–loop–helix proteins known as phytochrome-interacting factors (PIFs). The plant phytochromes PhyA and PhyB from *Arabidopsis thaliana* are readily activated by far-red (640–680 nm) light and deactivated by near-infrared (NIR) (740–760 nm) light or in darkness. In the active Pfr state, PhyA and PhyB interact with PIF3 or PIF6 (Fig. 1*C*) (Ni et al., 1999; Zhu et al., 2000). A far-red light-inducible gene expression system in yeast was developed by fusing the components of the PhyB-PIF3 system to the DNA-binding and transactivation domains of the Gal4 TF (Shimizu-Sato et al., 2002). The interaction of PhyA and PhyB with PIF3 was exploited to develop a light-switchable gene expression system in mammalian cells (Müller et al., 2013). The PhyB-PIF6 system for activating transcription in plants has also been reported (Ochoa-Fernandez et al., 2020). PhyB is known to be very light-sensitive and can be easily activated even in dim white light. To prevent this, the authors combined the PhyB-PIF6 protein pair with a LOV-based optogenetic system. The system was successfully tested in tobacco and *Arabidopsis*.

Among phytochromes, a subclass of bacterial phytochrome photoreceptors (BphPs) that incorporate BV should be emphasized (Giraud, Verméglio, 2008; Auldridge, Forest, 2011;

Piatkevich et al., 2013). BV has the most red-shifted absorbance relative to other tetrapyrroles, including PΦB and PCB. The first BphP1-PpsR2 optogenetic system for transcriptional activation in eukaryotic cells was developed based on the bacterial bathy phytochrome BphP1 from the nonsulphur purple bacterium *Rhodopseudomonas palustris*. Under NIR light, BphP1 interacts with its natural partner protein PpsR2. This property was utilized to create a transcription activation system in mammalian cell cultures and mice (Kaberniuk et al., 2016). The original BphP1-PpsR2 system was optimized by reducing PpsR2 to a single-domain binding partner, QPAS1, which is three times smaller than full-length PpsR2, includes only the α-helical Q-linker and PAS1 domain and lacks oligomerization (Redchuk et al., 2017). The pair of proteins BphP1 and QPAS1 was used to create NIR light-induced systems for the activation (Fig. 1*D*) and suppression of gene expression, as well as for the control of protein localization. These systems were tested in various mammalian cell cultures (Redchuk et al., 2017, 2018).

Previously, optogenetic systems for the activation of expression in plants utilized the plant phytochrome PhyB (Müller et al., 2014; Ochoa-Fernandez et al., 2016, 2020) or the bacterial photoreceptor CarH (Chatelle et al., 2018) activated by far-red (640–680 nm) or green light (525 nm), respectively (Omelina et al., 2022). A significant drawback of these systems is the potential effect on endogenous plant photoreceptors (Müller et al., 2014; Battle, Jones, 2020; Christie, Zurbriggen, 2021). In this study, for the regulation of the reporter gene transcription in plants, we applied the bacterial BphP1-QPAS1 system activated by NIR light (780 nm), which is outside the spectra of plant photoreceptors' excitation and differs from the wavelength required for their reversion. BphP1 is known to incorporate BV as a chromophore, which is synthesized in chloroplasts as a precursor of PΦB, suggesting the usability of BV-required tools in plants (Shikata, Denninger, 2022). However, the application of optogenetic systems in plant research is known to have a significant obstacle. Plants are phototrophic organisms and require exposure to white light, which may nonspecifically activate optogenetic systems (Andres et al., 2019). To block the activity of the BphP1-QPAS1 system under white light, we combined the BphP1-QPAS1 protein pair with blue light-controlled systems (the LOV domain-containing VVD protein or *As*LOV2 domain fused to several variants of the NES and degron sequences). The best results were observed for the combination of the BphP1-QPAS1 system with the *As*LOV2 domain fused to a degron sequence. In the leaves of these transiently transformed tobacco plants, we observed strong fluorescence of the plant EGFP (pEGFP) reporter under NIR light and found no pEGFP signal when the plants were incubated in standard day/night conditions.

We believe that in the future, the optogenetic system BphP1-QPAS1 will enable the development of several approaches for creating safe and highly productive agricultural plants that are resistant to stress conditions, pests, and viral diseases, which are common among agricultural crops. Unlike methods that use various chemicals to combat pests and infectious diseases, optogenetic biotechnologies are non-toxic and facilitate safe and high-quality food production.

Materials and methods

Design of plasmids. All plasmids generated in this study are presented in the Table.

Control reporter plasmids without the activator (*R1*, *R2*). Two variants of reporter plasmids R1 and R2 (Fig. 2*A*, *B*; Supplementary Figs. 1, 2)[1] were obtained using the Gibson assembly method based on the pBI121 vector (Chen et al., 2003), which was hydrolyzed at the SbfI/ClaI restriction sites. For that, the UAS sequence was amplified from the pUASTattB vector (GenBank EF362409.1) (Bischof et al., 2007). The pEGFP coding sequence, fused with the NOS terminator, and the sequence of the minimal promoter from the Cauliflower mosaic virus (mini35S) were amplified from the pCAMBIA_pGFP vector (Gayatri, Basu, 2020).

Control reporter plasmids with the activator (*R1A*, *R2A*). The sequences of the full-length promoter of the Cauliflower mosaic virus (CaMV35S) and light-independent activator Gal4-VP16 were added to the R1 and R2 plasmids to obtain the R1A and R2A plasmids. The CaMV35S promoter sequence was amplified from the pBI121 vector. The sequence of the activator Gal4-VP16 (Fig. 2*C*; Suppl. Fig. 3) was amplified from the plasmid pGal4-VP16 (Redchuk et al., 2017). The NOS terminator sequence was amplified from the pCAMBIA_pGFP vector. All fragments were inserted into the reporter plasmids R1 or R2 using the Gibson assembly method after being hydrolyzed at the PmeI/ClaI restriction sites.

Control reporter plasmids without the full-length CaMV35S promoter from the pBI121 plasmid. To eliminate the influence of the CaMV35S promoter from the pBI121 plasmid on the expression of pEGFP, this fragment was excised from R1, R2, R1A and R2A plasmids at the SbfI/BamHI sites. Subsequently, the vectors were treated with a Klenow fragment (NEB), followed by self-ligation. These constructs were named R1d, R2d, R1Ad and R2Ad, respectively.

Plasmid carrying the coding sequences of the BphP1-VP16, NLS-Gal4-QPAS1-VVD and VVD-CAAX chimeric proteins (*BQV*). To generate the BQV plasmid (Suppl. Fig. 4), the components of the BphP1-QPAS1 system were amplified from the plasmid pQP-T2A (Addgene #102583 (Redchuk et al., 2017)). The coding sequence of the VVD protein was amplified from the GAVPO construct (Wang et al., 2012). The NLS and CAAX sequences were included in the primers. These fragments were inserted into the reporter plasmid R2d using the Gibson assembly method after being hydrolyzed at the PmeI/ClaI restriction sites.

Plasmids carrying the coding sequences of the BphP1-VP16 and NES-Gal4-QPAS1-AsLOV2-NES21 or NES-Gal4-QPAS1-AsLOV2-NES27 or NES-Gal4-QPAS1-AsLOV2-RRRG chimeric proteins (*BQL1*, *BQL2*, *BQL3*, *respectively*). To generate the BQL1, BQL2 and BQL3 plasmids (Suppl. Fig. 5), the sequences of the NOS and CaMV35S promoters, the NOS terminator, and the NeoR/KanR sequence were amplified from the pBI121 vector. The components of the BphP1-QPAS1 system were amplified from the plasmid pQP-T2A. The Cauliflower mosaic virus polyadenylation signal sequence was amplified from the pCAMBIA_pGFP vector. The *As*LOV2

[1] Supplementary Figures 1–5 are available at:
https://vavilovj-icg.ru/download/pict-2026-30/appx5.zip

sequence was amplified from the plasmid pKM546 (Müller et al., 2015). The N-terminal NES and C-terminal NES21, NES27, and degron sequences were included in the primers. These fragments were inserted into the BQV plasmid using the Gibson assembly method after being hydrolyzed at the PmeI/HindIII restriction sites.

Preparation of chemically competent *Agrobacterium tumefaciens* cells. An overnight culture (3 ml) of *A. tumefaciens* strain GV3101 was grown in sterile YEP medium (10 g/l yeast extract, 10 g/l Bacto Peptone, 5 g/l NaCl) for 16–20 hours at 28 °C. Subsequently, 2 ml of this culture was transferred to a flask containing 50 ml of sterile YEP medium and incubated at 28 °C until an optical density at 600 nm (OD_{600}) of 0.5–1.0 was reached (typically several hours to overnight). The culture was then cooled on ice for 30 minutes. Cells were harvested by centrifugation at 5,000g at 4 °C for 8 minutes. The supernatant was discarded, and the cell pellet was gently resuspended in 2 ml of ice-cold 20 mM $CaCl_2$. The suspension was incubated on ice for an additional 30 minutes. Aliquots of 100 µl were dispensed into pre-chilled microcentrifuge tubes, flash-frozen in liquid nitrogen, and stored at –70 °C.

***A. tumefaciens* transformation.** Chemically competent *A. tumefaciens* strain GV3101 was transformed with the plasmids of interest (see the Table). Specifically, 1 µg of plasmid DNA was added to 100 µl of competent cells. These cells were thawed for 5 minutes at 37 °C, after which 1 ml of YEP medium supplemented with 50 µg/ml kanamycin and 100 µg/ml rifampicin was added. The mixture was incubated with shaking for 3 to 5.5 hours at 28 °C. The cells were then centrifuged at 1,000g for 5 minutes, after which the pellet was resuspended in 100 µl of YEP medium and plated on YEP agar plates containing kanamycin and rifampicin. Plates were incubated in an inverted position at 28 °C for 72 hours until individual colonies appeared. Colonies were confirmed by colony PCR using Taq DNA polymerase (Biolabmix).

Transient transformation of *Nicotiana benthamiana* plants. *A. tumefaciens* cultures were adjusted to an optical density measured at a wavelength of 600 nm (OD_{600} = 0.5) in infiltration medium (10 mM $MgCl_2$, 10 mM MES in H_2O). The cultures were incubated for 1–3 hours at 28 °C prior to infiltration through the adaxial surface of the leaves of 4- to 5-week-old *N. benthamiana*.

Illumination conditions and microscopy. To test the Gal4/UAS system (plasmids #1–8 in the Table) in tobacco leaves, the plants were kept in a plant incubator with fluorescent tubes (cool white light, OSRAM), employing alternating cycles of light and darkness (16 hours of light and 8 hours of darkness) to simulate standard day/night conditions for 48 hours.

To analyze the modified BphP1-QPAS1 system variants, *N. benthamiana* plants transiently transformed with the plasmids BQV, BQL1, BQL2, or BQL3 (see the Table) were grown for 48 hours in simulated day/night conditions (16 hours of white light and 8 hours of darkness). Thereafter, half of the plants were incubated under NIR light for 24 hours to induce pEGFP expression, while the other half continued to grow under the same day/night conditions.

NIR illumination was performed using custom-assembled LED arrays with a wavelength of 780/20 nm. For microscopic analysis, we conducted at least three independent experiments, taking 1–3 samples from different areas of the tobacco leaves and using a Zeiss Axio Imager M2 (Carl Zeiss).

Results

Testing of light-independent expression of the control reporter constructs in tobacco leaves

To test the functioning of the optogenetic system BphP1-QPAS1 in plants, we first selected an optimal structure for the regulatory region of the UAS enhancer-based reporter construct. This structure should provide minimal reporter activity in the absence of the activator and a high level of reporter gene expression in the presence of the activator. The creation of different variants of the reporter construct for use in *N. benthamiana* is necessary due to the lack of a single version of the UAS regulatory regions of reporter genes in plants, as their structure depends on the plant species (Wu C. et al., 2003; Johnson A.A.T. et al., 2005; Sakvarelidze et al., 2007; Yun et al., 2023). Additionally, the DNA-binding activity of Gal4 is known to be sensitive to the methylation of its binding site in plant chromatin, which can result in weak expression or even the loss of Gal4-mediated expression (Gälweiler et al., 2000). To develop these control reporter constructs, we employed the widely used binary *Agrobacterium* vector pBI121, which carries a GUS (*β-glucuronidase*) reporter gene for plant transformation (Chen et al., 2003). Based on the pBI121 vector backbone, we designed two variants of the control reporter plasmids, R1 (Fig. 2*A*; the Table; Suppl. Fig. 1) and R2 (Fig. 2*B*; the Table; Suppl. Fig. 2). Both plasmids contain five UAS repeats for the binding of the TF Gal4-VP16 and the reporter gene *pEGFP*. The R2 plasmid additionally contains the sequence of the minimal promoter from the Cauliflower mosaic virus (CaMV35S) – mini35S.

The sequence of the light-insensitive Gal4-VP16 activator was added to the reporter plasmids R1 and R2, resulting in the creation of the R1A and R2A plasmids (see the Table). The Gal4-VP16 activator consists of the DNA-binding and dimerization domains of the yeast TF Gal4 (148 a. a.) fused with the transactivation domain VP16 (Fig. 2*C*; Suppl. Fig. 3). For testing in plants, the plasmids R1, R2, R1A, and R2A were used to transform *A. tumefaciens* cells, which were subsequently used to infiltrate the leaves of the tobacco *N. benthamiana*. A microscopic analysis of pEGFP protein expression in the transiently transformed tobacco leaves was conducted 48 hours after infiltration. We anticipated observing pEGFP expression in the tobacco leaves transformed with one of the variants of the reporter plasmids in the presence of the Gal4-VP16 activator (R1A or R2A), but not in its absence (R1 or R2) (see the Table). However, all the analyzed constructs exhibited strong pEGFP expression (Fig. 2*D*–*E*′).

As we mentioned above, we cloned the reporter plasmids based on the pBI121 vector backbone. This vector carries the GUS coding sequence under the control of the full-length CaMV35S promoter located 494 bp from the UAS repeats in the reverse orientation (Fig. 2*F*, *G*). We did not initially remove the GUS sequence and its promoter from the vector, assuming that its expression would not interfere with the analysis of

Plasmids for testing the Gal4/UAS (#1–8) and optogenetic (#9–12) systems in tobacco leaves

#	Name	Key element(s) (reporter, activator, components of the optogenetic system)	Bb	Cloning method, restriction sites	Presence of the CaMV35S promoter*
1	R1	5×UAS-pEGFP	pBI121	GA, SbfI/ClaI	+
2	R2	5×UAS-mini35S-pEGFP	pBI121	GA, SbfI/ClaI	+
3	R1A	5×UAS-pEGFP, CaMV35S:Gal4-VP16	R1	GA, PmeI/ClaI	+
4	R2A	5×UAS-mini35S-pEGFP, CaMV35S:Gal4-VP16	R2	GA, PmeI/ClaI	+
5	R1d	5×UAS-pEGFP	R1	Restriction and self-ligation, SbfI/BamHI	–
6	R2d	5×UAS-mini35S-pEGFP	R2	Restriction and self-ligation, SbfI/BamHI	–
7	R1Ad	5×UAS-pEGFP, CaMV35S:Gal4-VP16	R1A	Restriction and self-ligation, SbfI/BamHI	–
8	R2Ad	5×UAS-mini35S-pEGFP, CaMV35S:Gal4-VP16	R2A	Restriction and self-ligation, SbfI/BamHI	–
9	BQV	5×UAS-mini35S-pEGFP, CaMV35S:BphP1-VP16-T2A-VVD-CAAX, NOS:NLS-Gal4-QPAS1-VVD	R2d	GA, PmeI/ClaI	–
10, 11, 12	BQL1, BQL2, BQL3	5×UAS-mini35S-pEGFP, CaMV35S:BphP1-VP16-T2A-NeoR/KanR, NOS:NES-Gal4-QPAS1-AsLOV2-NES21/NES27/degron	BQV	GA, PmeI/HindIII	–

Note. Bb, backbone; GA, Gibson assembly method. * The full-length promoter CaMV35S from the pBI121 vector backbone, which controls expression of the GUS reporter gene.

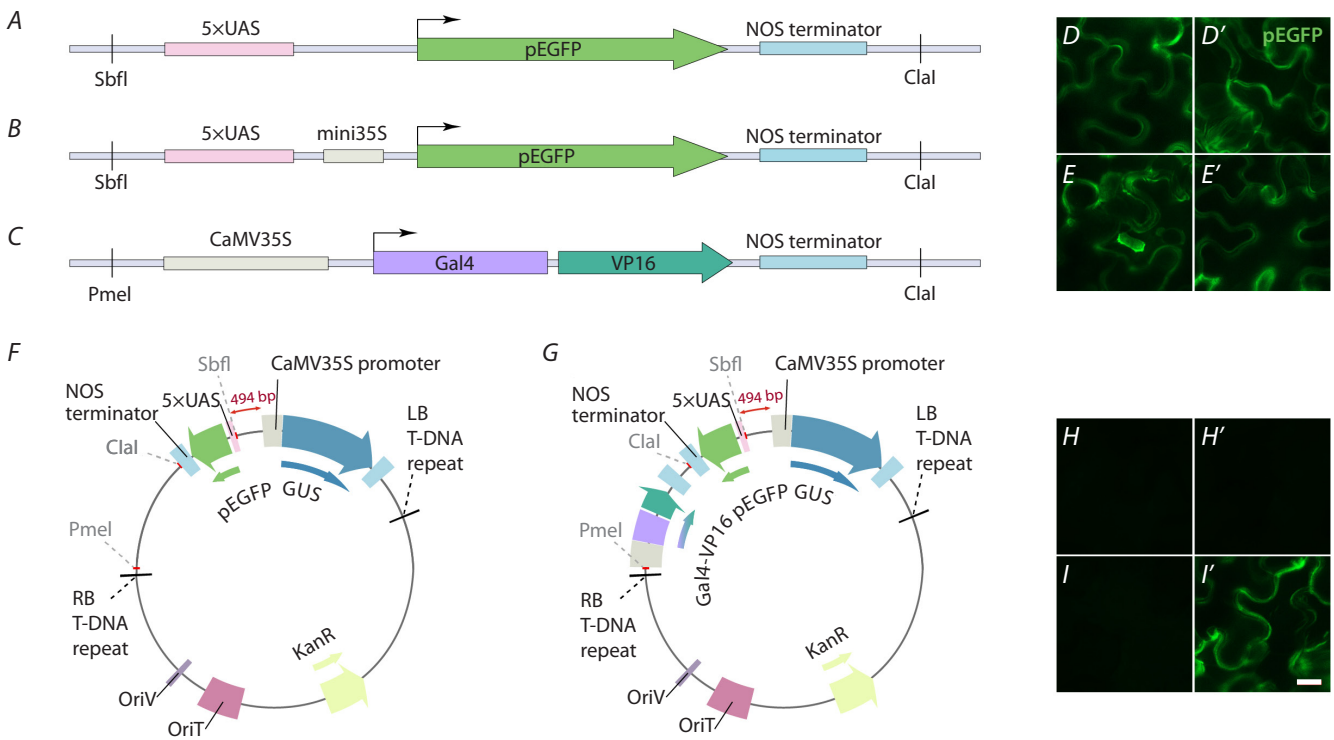


Fig. 2. Structure of the control reporter plasmids for testing the Gal4/UAS system in transiently transformed *N. benthamiana* leaves.

A, Reporter construct R1 without a minimal promoter. *B*, Reporter construct R2, containing the sequence of the minimal CaMV35S promoter – mini35S. *C*, Scheme of the light-independent Gal4-VP16 activator, incorporated into the reporter plasmids R1 and R2 for the activation of the reporter gene. It consists of the DNA-binding and dimerization domains of the yeast TF Gal4 (148 a.a.) fused with the transactivation domain VP16. The plasmid elements are not shown to scale. *D–E′*, Microscopy images of *N. benthamiana* leaves transiently transformed with the constructs R1 (*D*), R1A (*D′*), R2 (*E*), R2A (*E′*). *F*, A scheme of the reporter construct R1 (analogous to R2) based on the pBI121 vector backbone. *G*, A scheme of the construct R1A (analogous to R2A) with the Gal4-VP16 activator based on the pBI121 vector backbone. *F*, *G*, The full-length CaMV35S promoter, which regulates the activity of the GUS gene, is located 494 bp from the UAS repeats in the reverse orientation. *H–I′*, Microscopy images of *N. benthamiana* leaves transiently transformed with the constructs R1d (*H*), R1Ad (*H′*), R2d (*I*), R2Ad (*I′*). Plants were incubated in simulated day/night conditions for 48 hours (*D–E′*, *H–I′*). Scale, 20 µm.

the pEGFP fluorescence. However, a thorough review of the literature indicated that the full-length CaMV35S promoter could alter the activity level of adjacent genes located approximately 2 kb downstream of the promoter (Zheng et al., 2007). Furthermore, this promoter can act as an enhancer in either orientation and lead to an increase in the expression level of neighboring genes (Zheng et al., 2007). To eliminate the presumed influence of the CaMV35S promoter on the expression of the pEGFP reporter, we removed its sequence from plasmids R1, R2, R1A and R2A. As a result, we obtained two new variants of the reporter constructs (R1d and R2d) and two constructs that carry both the reporter gene pEGFP and the activator Gal4-VP16 (R1Ad and R2Ad) (see the Table). Analysis of the activity of these constructs in tobacco leaves showed that pEGFP protein expression was only observed in the leaves transformed with the R2Ad construct with the reporter containing the mini35S sequence in the presence of the Gal4-VP16 activator (Fig. 2*I′*). In the case of the other constructs, the level of the pEGFP signal corresponded to autofluorescence (Fig. 2*H–I*). Therefore, we subsequently used the R2d reporter plasmid for the application of the optogenetic system BphP1-QPAS1 in plants.

Optogenetic control of gene expression with the BphP1-QPAS1 system in transiently transformed *N. benthamiana* leaves

The application of optogenetic systems in plant research is known to face a significant obstacle in the form of undesired activity under ambient white light, which is vital for normal plant growth and development (Braguy, Zurbriggen, 2016). The original BphP1-QPAS1 system is based on the split TF Gal4-VP16, which is assembled by exposure to NIR light. The DNA-binding domain of Gal4 is fused to QPAS1 (Gal4-QPAS1) and BphP1 is fused with the VP16 transactivation domain (BphP1-VP16) (Fig. 1*D*) (Redchuk et al., 2017). To address the problem of undesired activation, we fused the Gal4-QPAS1 chimeric protein with several variants of blue light-sensitive LOV domains. This approach might provide the removal of the Gal4-QPAS1 component from the cell nucleus under white light.

We combined the components of the BphP1-QPAS1 system with the VVD protein (see the BQV plasmid in the Table), which forms homodimers under blue light (or white light covering a broad spectrum of different wavelengths, including blue light at 460–480 nm). To achieve this, we fused the DNA-binding protein NLS-Gal4-QPAS1 with VVD, resulting in the chimeric protein NLS-Gal4-QPAS1-VVD (Fig. 3*A*). We also added to the BQV plasmid an additional VVD-containing construct, VVD-CAAX, which facilitates the attachment of the VVD protein to the plasma membrane via the CAAX peptide (Fig. 3*A*; Suppl. Fig. 4) (Redchuk et al., 2017; Willumsen et al., 1984a, b). We hypothesized that in this modified variant of the optogenetic system, BphP1-VP16 would predominantly reside in the nucleus of the cells (Fig. 3*A*). In darkness, the system exists in a state of relaxation. Under white light, the chimeric proteins NLS-Gal4-QPAS1-VVD and VVD-CAAX form dimers due to VVD homodimerization, resulting in the relocalization of the chimeric protein NLS-Gal4-QPAS1-VVD to the membrane. Under NIR light, BphP1 interacts with QPAS1, leading to the activation of the reporter gene (Fig. 3*A*).

To test the functionality of the BphP1-QPAS1 system combined with VVD, we transformed *Agrobacterium* with the BQV plasmid (see the Table), followed by infiltration of the leaves of tobacco plants. The plants were incubated in simulated day/night conditions (16 hours of light and 8 hours of darkness) for 48 hours. Thereafter, half of the plants were incubated under NIR light for 24 hours, while the other plants continued to grow in simulated day/night conditions. As a result, we detected pEGFP fluorescence in the transiently transformed tobacco leaves incubated both under NIR light and in white light/dark conditions (Fig. 3*B*, *B′*).

It should be noted that a standard practice for testing optogenetic systems involves incubating samples under different light conditions: in darkness and under pulsed light (e. g.,

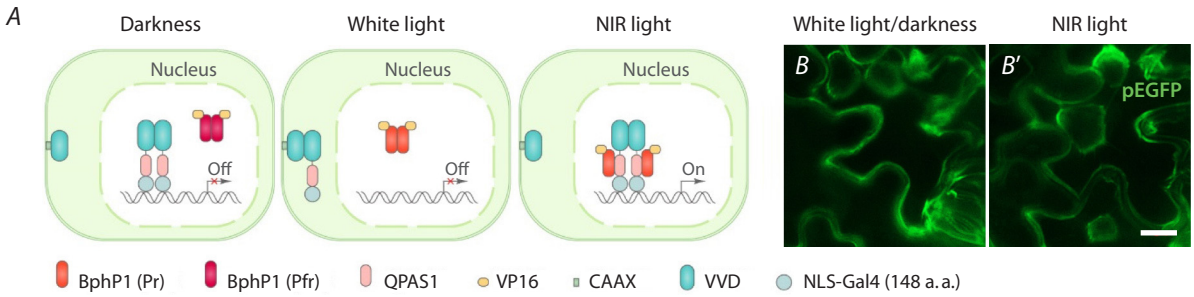


Fig. 3. NIR light-controlled transcription activation of the reporter gene using the BphP1-QPAS1 system in combination with the VVD protein.

A, A scheme of the proposed functioning of the BphP1-QPAS1 system under different light conditions. In darkness, BphP1 exists in the inactive Pfr state and cannot interact with QPAS1. Under white light, BphP1 converts to the active Pr state. However, it cannot interact with QPAS1, as the NLS-Gal4-QPAS1-VVD chimeric protein relocalizes to the plasma membrane due to dimer formation with VVD-CAAX. Exposure to NIR light leads to the interaction of the BphP1 and QPAS1 components in the nucleus and activation of the reporter gene expression. *B*, *B′*, Microscopy images of *N. benthamiana* leaves transiently transformed with the plasmid BQV. The plants were incubated in simulated day/night conditions for 72 hours (*B*) or in simulated day/night conditions for 48 hours, followed by incubation under NIR light for 24 hours (*B′*). Scale, 20 µm.

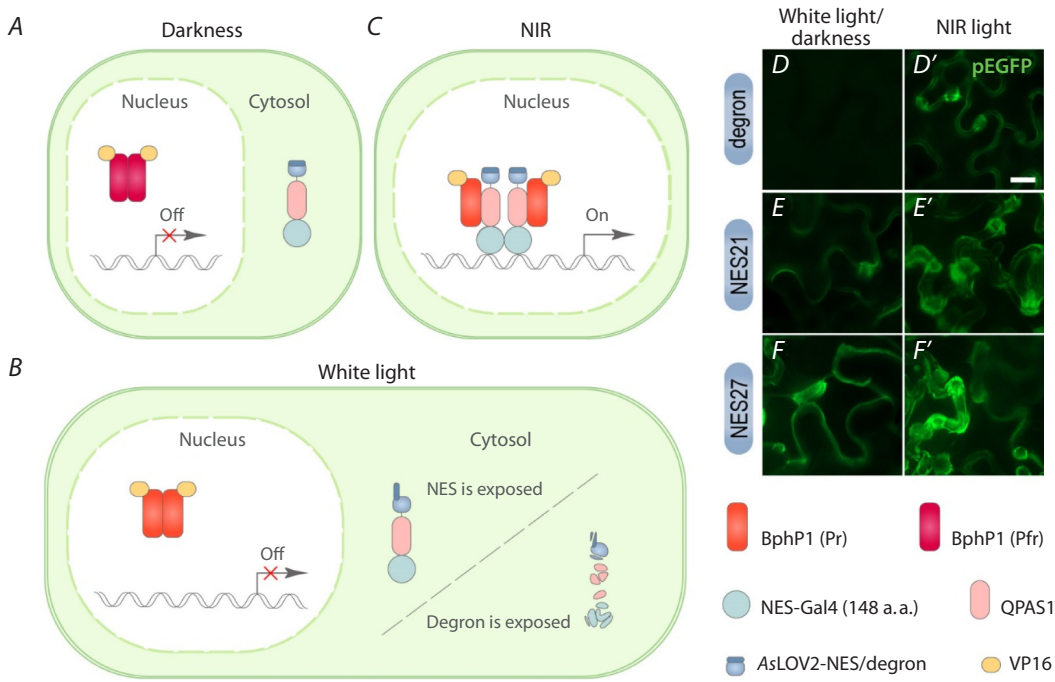


Fig. 4. Two strategies of the BphP1-QPAS1 system functioning to avoid undesired activation in simulated day/night conditions.

A, A schematic representation of the proposed functioning of the modified BphP1-QPAS1 system in darkness. BphP1-VP16 exists in the ground Pfr state and predominantly resides in the nucleus of the cells. The NES-Gal4-QPAS1-*As*LOV2-NES/degron localizes in the cytoplasm due to a weak N-terminal NES signal. *B*, A scheme of the proposed functioning of the modified BphP1-QPAS1 system under white light. BphP1-VP16 possibly exists in the active Pr state. However, it cannot interact with QPAS1, as under white light, the NES or degron signal unwinds from the *As*LOV2 core. This results in the relocalization from the nucleus to the cytoplasm or degradation of the NES-Gal4-QPAS1-*As*LOV2-NES/degron chimeric protein. *C*, A scheme of the proposed functioning of the modified BphP1-QPAS1 system under NIR light. BphP1 exists in the active Pr state and interacts with QPAS1, leading to the formation of the functional TF Gal4-VP16 and the activation of the reporter gene expression. *D–F′*, Microscopy images of *N. benthamiana* leaves transiently transformed with plasmids carrying the components of the BphP1-QPAS1 system in combination with *As*LOV2 fused to different peptides (degron in *D*, *D′*; NES21 in *E*, *E′*; NES27 in *F*, *F′*). The plants were incubated in simulated day/night conditions for 72 hours (*D*, *E*, *F*) or in simulated day/night conditions for 48 hours, followed by incubation under NIR light for 24 hours (*D′*, *E′*, *F′*). Scale bar, 20 μm.

30 seconds On and 180 seconds Off). Initially, after the transient transformation of the tobacco leaves with the BQV plasmid, we employed various light conditions. We kept the plants either in darkness or under continuous blue (480 nm), NIR (780 nm), or white light for 48 hours. However, these conditions were not optimal for the normal growth of the plants, and they exhibited wilting. Therefore, in all subsequent experiments, we decided to simulate the most natural lighting conditions with a day/night cycle during the first 48 hours after transformation. Thereafter, half of the plants were incubated under NIR light for 24 hours to induce pEGFP expression, while the remaining plants continued to grow in simulated day/night conditions.

Thus, the combination of BphP1-QPAS1 with the used variant of the VVD protein appears to be ineffective in plants, as the addition of VVD did not eliminate the undesired activity of the BphP1-QPAS1 system under white light (and possibly in darkness). To address this issue, we first replaced the N-terminal nuclear localization signal (NLS) with a weak NES signal in the Gal4-QPAS1 component in the BQV plasmid. It may relocalize part of this chimeric protein from the cell nucleus, which could potentially reduce the level of undesired activation in darkness. Secondly, we replaced VVD with the blue-light-sensitive *As*LOV2 domain, which carries a specific peptide on its C-terminal Jα helix. Utilizing the different peptides allowed for two strategies: (1) white light-induced nuclear export of the chimeric protein NES-Gal4-QPAS1-*As*LOV2-NES; (2) white light-induced degradation of the chimeric protein NES-Gal4-QPAS1-*As*LOV2-degron (Fig. 4*A–C*). To relocalize the NES-Gal4-QPAS1-*As*LOV2-NES chimeric protein from the nucleus to the cytosol, we used two different NES signals, presented in plasmids BQL1 and BQL2 (see the Table; Suppl. Fig. 5). These signals (NES21 and NES27, respectively) were taken from a library of NES signals, which, in fusion with *As*LOV2, have shown different efficiencies of nucleocytoplasmic translocation under blue light (Niopek et al., 2016). For light-inducible degradation of the NES-Gal4-QPAS1-*As*LOV2-degron protein, the plasmid BQL3 was generated (see the Table; Suppl. Fig. 5). For this strategy, we utilized the previously described B-LID system carrying the degron variant RRRG (Bonger et al., 2014).

We transformed *Agrobacterium* with the plasmids BQL1, BQL2, and BQL3, followed by the infiltration of tobacco leaves. The plants were incubated for 48 hours in simulated day/night conditions. Thereafter, half the plants were incubated under NIR light for 24 hours, while the others continued to

grow in simulated day/night conditions for a further 24 hours. The strategy based on white light-sensitive degradation enabled the avoidance of undesired activation of the BphP1-QPAS1 system under white light and in darkness. We observed a strong pEGFP signal under NIR light and did not detect any pEGFP fluorescence in plants grown under standard day/night conditions (Fig. 4*D*, *D′*). The strategy based on white light-induced nucleocytoplasmic translocation was much less effective. Both variants of the NES signals displayed pEGFP fluorescence under NIR light as well as in alternating cycles of white light and darkness (Fig. 4*E–F′*).

Discussion

The undesired activation of optogenetic systems under ambient white light is known to be a significant obstacle to using these systems in plant research, as white light is vital for normal plant growth. This explains the relatively low number of articles on the application of optogenetic systems in plants compared to studies employing chemically inducible systems (Omelina et al., 2022). However, optogenetic systems offer several undeniable advantages over chemical systems, making their application in plants very appealing. To address the problem of undesired activation, optogenetic systems may be combined with other light-sensitive proteins that have higher light sensitivity to block undesired activity under white light.

In this study, we first applied the BphP1-QPAS1 system in the leaves of *N. benthamiana* to induce the expression of a reporter gene using NIR light. To avoid undesired activity of the BphP1-QPAS1 system under white light in plants, we combined it with the blue light-sensitive VVD protein (Wang et al., 2012). Previous studies in mice have demonstrated that VVD is more light-sensitive compared to the BphP1-QPAS1 protein pair (Kaberniuk et al., 2016). However, in plants, the combination of these systems did not yield the desired result. This may be related to the previously described dependence of the photoinducible dimerization properties of VVD on temperature (Qian et al., 2023). In this study, we used a variant of the VVD protein bearing the mutations N56K and C71V, which have previously shown a high On/Off ratio at 37 °C and a low On/Off ratio at 25 °C in HEK293 cells (Qian et al., 2023). The combination of the BphP1-QPAS1 system with the VVD protein is very appealing for use in plants, as all components are orthogonal to plant cells. Using another mutant version of the VVD protein or other localization signals in the future might enable avoiding undesired activation of the BphP1-QPAS1 system in combination with VVD under white light. We replaced the VVD protein with the plant-derived *As*LOV2 domain. As a result, a variant of the *As*LOV2 domain fused with the degron sequence has shown very good results. We observed a high level of pEGFP fluorescence after NIR light illumination and did not detect the pEGFP signal in plants grown in simulated day/night conditions.

It is important to note that light-sensitive proteins of plant origin have been applied in plants before. For instance, the PULSE system, which was previously successfully tested in plants, is based on the use of the plant phytochrome PhyB (Müller et al., 2014; Ochoa-Fernandez et al., 2020). In this system, the problem of undesired activation under white light was solved by the addition of the bacterial LOV-domain protein EL222 fused with the repressor domain SRDX and binding to the same promoter of the reporter gene as the PhyB-PIF6-based split TF. We propose that the BphP1-QPAS1 system has several advantages compared to the PhyB-based system, as its activation occurs at NIR light (780 nm) rather than at 640–680 nm. NIR light is preferable not only because it is not toxic to eukaryotic cells, including plant cells, but also because it does not affect the excitation or relaxation of endogenous phytochromes (Johnson M.P., 2016; Guidi et al., 2017). Additionally, passive activation of degradation by white light (but not blue light) does not affect the activation of endogenous phototropins and cryptochromes. The chromophores required for photoactivation of BphP1 and *As*LOV2 (BV and flavin chromophore, respectively) are plant metabolites and are present in sufficient quantities in plant cells. In contrast to BphP1-QPAS1, the functioning of other optogenetic systems may lead to the potential activation of endogenous photoreceptors, which play a crucial role in plant development. Moreover, it should also be noted that the BphP1-based system is significantly less sensitive to ambient light compared to PhyB-derived optogenetic tools (Kaberniuk et al., 2016). Furthermore, the PhyB protein is present in all seed plants (Mathews, 2010; Li F.W. et al., 2015), so the interaction of the components of PhyB-based systems with endogenous proteins cannot be ruled out. These points favourably distinguish the BphP1-QPAS1 system from previously used systems for light-inducible regulation of gene expression in plants. Thus, in the near future, the BphP1-QPAS1 system has all the potential for wide application in plant research and biotechnology, for example, for plant protection from adverse climatic and chemical conditions (such as drought, toxins, soil salinity, and radiation), plant diseases (including moulds, bacteria, and viruses), and pests (such as insects and rodents).

References

Andres J., Blomeier T., Zurbriggen M.D. Synthetic switches and regulatory circuits in plants. *Plant Physiol.* 2019;179(3):862-884. doi 10.1104/pp.18.01362

Auldridge M.E., Forest K.T. Bacterial phytochromes: more than meets the light. *Crit Rev Biochem Mol Biol*. 2011;46(1):67-88. doi 10.3109/10409238.2010.546389

Battle M.W., Jones M.A. Cryptochromes integrate green light signals into the circadian system. *Plant Cell Environ*. 2020;43(1):16-27. doi 10.1111/pce.13643

Bischof J., Maeda R.K., Hediger M., Karch F., Basler K. An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc Natl Acad Sci USA*. 2007;104(9):3312-3317. doi 10.1073/pnas.0611511104

Bonger K.M., Rakhit R., Payumo A.Y., Chen J.K., Wandless T.J. General method for regulating protein stability with light. *ACS Chem Biol*. 2014;9(1):111-115. doi 10.1021/cb400755b

Braguy J., Zurbriggen M.D. Synthetic strategies for plant signalling studies: molecular toolbox and orthogonal platforms. *Plant J*. 2016; 87(1):118-138. doi 10.1111/tpj.13218

Chatelle C., Ochoa-Fernandez R., Engesser R., Schneider N., Beyer H.M., Jones A.R., Timmer J., Zurbriggen M.D., Weber W. A green-light-responsive system for the control of transgene expression in mammalian and plant cells. *ACS Synth Biol*. 2018;7(5): 1349-1358. doi 10.1021/acssynbio.7b00450

Chen P.-Y., Wang C.-K., Soong S.-C., To K.-Y. Complete sequence of the binary vector pBI121 and its application in cloning T-DNA in-

sertion from transgenic plants. *Mol Breed*. 2003;11(4):287-293. doi 10.1023/A:1023475710642

Chernov K.G., Redchuk T.A., Omelina E.S., Verkhusha V.V. Near-infrared fluorescent proteins, biosensors, and optogenetic tools engineered from phytochromes. *Chem Rev.* 2017;117(9):6423-6446. doi 10.1021/acs.chemrev.6b00700

Christie J.M., Zurbriggen M.D. Optogenetics in plants. *New Phytol*. 2021;229(6):3108-3115. doi 10.1111/nph.17008

Christie J.M., Gawthorne J., Young G., Fraser N.J., Roe A.J. LOV to BLUF: flavoprotein contributions to the optogenetic toolkit. *Mol Plant.* 2012;5(3):533-544. doi 10.1093/mp/sss020

Crosson S., Moffat K. Structure of a flavin-binding plant photoreceptor domain: insights into light-mediated signal transduction. *Proc Natl Acad Sci USA*. 2001;98(6):2995-3000. doi 10.1073/pnas.051520298

Gälweiler L., Conlan R.S., Mader P., Palme K., Moore I. Technical advance: the DNA-binding activity of Gal4 is inhibited by methylation of the Gal4 binding site in plant chromatin. *Plant J*. 2000;23(1):143-157. doi 10.1046/j.1365-313x.2000.00805.x

Gayatri T., Basu A. Development of reproducible regeneration and transformation system for *Sesamum indicum*. *Plant Cell Tiss Organ Cult*. 2020;143(2):441-456. doi 10.1007/s11240-020-01931-1

Giraud E., Verméglio A. Bacteriophytochromes in anoxygenic photosynthetic bacteria. *Photosynth Res*. 2008;97(2):141-153. doi 10.1007/s11120-008-9323-0

Gligorovski V., Sadeghi A., Rahi S.J. Multidimensional characterization of inducible promoters and a highly light-sensitive LOV-transcription factor. *Nat Commun*. 2023;14(1):3810. doi 10.1038/s41467-023-38959-8

Guidi L., Tattini M., Landi M. How does chloroplast protect chlorophyll against excessive light? In: Chlorophyll. InTech, 2017. doi 10.5772/67887

Hernández-Candia C.N., Tucker C.L. Optogenetic control of gene expression using cryptochrome 2 and a light-activated degron. *Methods Mol Biol*. 2020;2173:151-158. doi 10.1007/978-1-0716-0755-8_10

Herrou J., Crosson S. Function, structure and mechanism of bacterial photosensory LOV proteins. *Nat Rev Microbiol*. 2011;9(10):713-723. doi 10.1038/nrmicro2622

Johnson A.A.T., Hibberd J.M., Gay C., Essah P.A., Haseloff J., Tester M., Guiderdoni E. Spatial control of transgene expression in rice (*Oryza sativa* L.) using the GAL4 enhancer trapping system. *Plant J*. 2005;41(5):779-789. doi 10.1111/j.1365-313X.2005.02339.x

Johnson M.P. Photosynthesis. *Essays Biochem*. 2016;60(3):255-273. doi 10.1042/EBC20160016

Kaberniuk A.A., Shemetov A.A., Verkhusha V.V. A bacterial phytochrome-based optogenetic system controllable with near-infrared light. *Nat Methods*. 2016;13(7):591-597. doi 10.1038/nmeth.3864

Kaya S.G., Hovan A., Fraaije M.W. Engineering of LOV-domains for their use as protein tags. *Arch Biochem Biophys*. 2025;763:110228. doi 10.1016/j.abb.2024.110228

Kim C., Yun S.R., Lee S.J., Kim S.O., Lee H., Choi J., Kim J.G., Kim T.W., You S., Kosheleva I., Noh T., Baek J., Ihee H. Structural dynamics of protein-protein association involved in the light-induced transition of *Avena sativa* LOV2 protein. *Nat Commun.* 2024;15(1):6991. doi 10.1038/s41467-024-51461-z

Konermann S., Brigham M.D., Trevino A.E., Hsu P.D., Heidenreich M., Cong L., Platt R.J., Scott D.A., Church G.M., Zhang F. Optical control of mammalian endogenous transcription and epigenetic states. *Nature.* 2013;500(7463):472-476. doi 10.1038/nature12466

Krueger D., Izquierdo E., Viswanathan R., Hartmann J., Pallares Cartes C., De Renzis S. Principles and applications of optogenetics in developmental biology. *Development*. 2019;146(20):dev175067. doi 10.1242/dev.175067

Li F.W., Melkonian M., Rothfels C.J., Villarreal J.C., Stevenson D.W., Graham S.W., Wong G.K., Pryer K.M., Mathews S. Phytochrome diversity in green plants and the origin of canonical plant phytochromes. *Nat Commun*. 2015;6(1):7852. doi 10.1038/ncomms8852

Li J., Li G., Wang H., Wang Deng X. Phytochrome signaling mechanisms. *Arabidopsis Book*. 2011;9:e0148. doi 10.1199/tab.0148

Losi A., Gärtner W. Old chromophores, new photoactivation paradigms, trendy applications: flavins in blue light-sensing photoreceptors. *Photochem Photobiol*. 2011;87(3):491-510. doi 10.1111/j.1751-1097.2011.00913.x

Lungu O.I., Hallett R.A., Choi E.J., Aiken M.J., Hahn K.M., Kuhlman B. Designing photoswitchable peptides using the AsLOV2 domain. *Chem Biol*. 2012;19(4):507-517. doi 10.1016/j.chembiol.2012.02.006

Mathews S. Evolutionary studies illuminate the structural-functional model of plant phytochromes. *Plant Cell*. 2010;22(1):4-16. doi 10.1105/tpc.109.072280

Müller K., Engesser R., Metzger S., Schulz S., Kämpf M.M., Busacker M., Steinberg T., Tomakidi P., Ehrbar M., Nagy F., Timmer J., Zubriggen M.D., Weber W. A red/far-red light-responsive bi-stable toggle switch to control gene expression in mammalian cells. *Nucleic Acids Res*. 2013;41(7):e77. doi 10.1093/nar/gkt002

Müller K., Siegel D., Rodriguez Jahnke F., Gerrer K., Wend S., Decker E.L., Reski R., Weber W., Zurbriggen M.D. A red light-controlled synthetic gene expression switch for plant systems. *Mol Biosyst*. 2014;10(7):1679-1688. doi 10.1039/c3mb70579j

Müller K., Zurbriggen M.D., Weber W. An optogenetic upgrade for the Tet-OFF system. *Biotechnol Bioeng.* 2015;112(7):1483-1487. doi 10.1002/bit.25562

Ni M., Tepperman J.M., Quail P.H. Binding of phytochrome B to its nuclear signalling partner PIF3 is reversibly induced by light. *Nature.* 1999;400(6746):781-784. doi 10.1038/23500

Niopek D., Wehler P., Roensch J., Eils R., Di Ventura B. Optogenetic control of nuclear protein export. *Nat Commun*. 2016;7:10624. doi 10.1038/ncomms10624

Noda N., Ozawa T. Light-controllable transcription system by nucleocytoplasmic shuttling of a truncated phytochrome B. *Photochem Photobiol*. 2018;94(5):1071-1076. doi 10.1111/php.12955

Ochoa-Fernandez R., Samodelov S.L., Brandl S.M., Wehinger E., Müller K., Weber W., Zurbriggen M.D. Optogenetics in plants: red/far-red light control of gene expression. *Methods Mol Biol.* 2016;1408:125-139. doi 10.1007/978-1-4939-3512-3_9

Ochoa-Fernandez R., Abel N.B., Wieland F.-G., Schlegel J., Koch L.A., Miller J.B., Engesser R., Giuriani G., Brandl S.M., Timmer J., Weber W., Ott T., Simon R., Zurbriggen M.D. Optogenetic control of gene expression in plants in the presence of ambient white light. *Nat Methods*. 2020;17(7):717-725. doi 10.1038/s41592-020-0868-y

Omelina E.S., Pindyurin A.V. Optogenetic regulation of endogenous gene transcription in mammals. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2019;23(2):219-225. doi 10.18699/VJ19.485

Omelina E.S., Yushkova A.A., Motorina D.M., Volegov G.A., Kozhevnikova E.N., Pindyurin A.V. Optogenetic and chemical induction systems for regulation of transgene expression in plants: use in basic and applied research. *Int J Mol Sci.* 2022;23(3):1737. doi 10.3390/ijms23031737

Pathak G.P., Spiltoir J.I., Höglund C., Polstein L.R., Heine-Koskinen S., Gersbach C.A., Rossi J., Tucker C.L. Bidirectional approaches for optogenetic regulation of gene expression in mammalian cells using *Arabidopsis* cryptochrome 2. *Nucleic Acids Res*. 2017;45(20):e167. doi 10.1093/nar/gkx260

Peter E., Dick B., Baeurle S.A. Mechanism of signal transduction of the LOV2-Jα photosensor from *Avena sativa*. *Nat Commun*. 2010;1:122. doi 10.1038/ncomms1121

Piatkevich K.D., Subach F.V., Verkhusha V.V. Engineering of bacterial phytochromes for near-infrared imaging, sensing, and light-control in mammals. *Chem Soc Rev*. 2013;42(8):3441-3452. doi 10.1039/c3cs35458j

Qian Y., Li T., Zhou S., Chen X., Yang Y. A single-component optogenetic Gal4-UAS system allows stringent control of gene expression in zebrafish and Drosophila. *ACS Synth Biol*. 2023;12(3):664-671. doi 10.1021/acssynbio.2c00410

Redchuk T.A., Omelina E.S., Chernov K.G., Verkhusha V.V. Near-infrared optogenetic pair for protein regulation and spectral multiplexing. *Nat Chem Biol*. 2017;13(6):633-639. doi 10.1038/nchembio.2343

Redchuk T.A., Karasev M.M., Omelina E.S., Verkhusha V.V. Near-infrared light-controlled gene expression and protein targeting in neurons and non-neuronal cells. *Chembiochem.* 2018;19(12):1334-1340. doi 10.1002/cbic.201700642

Renicke C., Schuster D., Usherenko S., Essen L.O., Taxis C. A LOV2 domain-based optogenetic tool to control protein degradation and cellular function. *Chem Biol*. 2013;20(4):619-626. doi 10.1016/j.chembiol.2013.03.005

Sakvarelidze L., Tao Z., Bush M., Roberts G.R., Leader D.J., Doonan J.H., Rawsthorne S. Coupling the GAL4 UAS system with *alcR* for versatile cell type-specific chemically inducible gene expression in *Arabidopsis*. *Plant Biotechnol J*. 2007;5(4):465-476. doi 10.1111/j.1467-7652.2007.00254.x

Shcherbakova D.M., Shemetov A.A., Kaberniuk A.A., Verkhusha V.V. Natural photoreceptors as a source of fluorescent proteins, biosensors, and optogenetic tools. *Annu Rev Biochem*. 2015;84(1):519-550. doi 10.1146/annurev-biochem-060614-034411

Shikata H., Denninger P. Plant optogenetics: applications and perspectives. *Curr Opin Plant Biol*. 2022;68:102256. doi 10.1016/j.pbi.2022.102256

Shimizu-Sato S., Huq E., Tepperman J.M., Quail P.H. A light-switchable gene promoter system. *Nat Biotechnol.* 2002;20(10):1041-1044. doi 10.1038/nbt734

Strickland D., Lin Y., Wagner E., Hope C.M., Zayner J., Antoniou C., Sosnick T.R., Weiss E.L., Glotzer M. TULIPs: tunable, light-controlled interacting protein tags for cell biology. *Nat Methods*. 2012;9(4):379-384. doi 10.1038/nmeth.1904

Wagner J.R., Zhang J., Brunzelle J.S., Vierstra R.D., Forest K.T. High resolution structure of *Deinococcus* bacteriophytochrome yields new insights into phytochrome architecture and evolution. *J Biol Chem*. 2007;282(16):12298-12309. doi 10.1074/jbc.M611824200

Wang X., Chen X., Yang Y. Spatiotemporal control of gene expression by a light-switchable transgene system. *Nat Methods*. 2012;9(3):266-269. doi 10.1038/nmeth.1892

Willumsen B.M., Christensen A., Hubbert N.L., Papageorge A.G., Lowy D.R. The p21 *ras* C-terminus is required for transformation and membrane association. *Nature.* 1984a;310(5978):583-586. doi 10.1038/310583a0

Willumsen B.M., Norris K., Papageorge A.G., Hubbert N.L., Lowy D.R. Harvey murine sarcoma virus p21 *ras* protein: biological and biochemical significance of the cysteine nearest the carboxy terminus. *EMBO J*. 1984b;3(11):2581-2585. doi 10.1002/j.1460-2075.1984.tb02177.x

Wu C., Li X., Yuan W., Chen G., Kilian A., Li J., Xu C., Li X., Zhou D.X., Wang S., Zhang Q. Development of enhancer trap lines for functional analysis of the rice genome. *Plant J*. 2003;35(3):418-427. doi 10.1046/j.1365-313x.2003.01808.x

Wu Y.I., Frey D., Lungu O.I., Jaehrig A., Schlichting I., Kuhlman B., Hahn K.M. A genetically encoded photoactivatable Rac controls the motility of living cells. *Nature.* 2009;461(7260):104-108. doi 10.1038/nature08241

Yang X., Kuk J., Moffat K. Crystal structure of *Pseudomonas aeruginosa* bacteriophytochrome: photoconversion and signal transduction. *Proc Natl Acad Sci USA*. 2008;105(38):14715-14720. doi 10.1073/pnas.0806718105

Yumerefendi H., Dickinson D.J., Wang H., Zimmerman S.P., Bear J.E., Goldstein B., Hahn K., Kuhlman B. Control of protein activity and cell fate specification via light-mediated nuclear translocation. *PLoS One*. 2015;10(6):e0128443. doi 10.1371/journal.pone.0128443

Yun A., Kang J., Lee J., Song S.J., Hwang I. Design of an artificial transcriptional system for production of high levels of recombinant proteins in tobacco (*Nicotiana benthamiana*). *Front Plant Sci.* 2023;14:1138089. doi 10.3389/fpls.2023.1138089

Zayner J.P., Antoniou C., Sosnick T.R. The amino-terminal helix modulates light-activated conformational changes in *As*LOV2. *J Mol Biol*. 2012;419(1-2):61-74. doi 10.1016/j.jmb.2012.02.037

Zheng X., Deng W., Luo K., Duan H., Chen Y., McAvoy R., Song S., Pei Y., Li Y. The cauliflower mosaic virus (CaMV) 35S promoter sequence alters the level and patterns of activity of adjacent tissue- and organ-specific gene promoters. *Plant Cell Rep*. 2007;26(8):1195-1203. doi 10.1007/s00299-007-0307-x

Zhu Y., Tepperman J.M., Fairchild C.D., Quail P.H. Phytochrome B binds with greater apparent affinity than phytochrome A to the basic helix-loop-helix factor PIF3 in a reaction requiring the PAS domain of PIF3. *Proc Natl Acad Sci USA*. 2000;97(24):13419-13424. doi 10.1073/pnas.230433797

Zoltowski B.D., Crane B.R. Light activation of the LOV protein vivid generates a rapidly exchanging dimer. *Biochemistry.* 2008;47(27):7012-7019. doi 10.1021/bi8007017

Zoltowski B.D., Schwerdtfeger C., Widom J., Loros J.J., Bilwes A.M., Dunlap J.C., Crane B.R. Conformational switching in the fungal light sensor Vivid. *Science.* 2007;316(5827):1054-1057. doi 10.1126/science.1137128

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

Received April 8, 2025. Revised April 30, 2025. Accepted May 28, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):72-84

doi 10.18699/vjgb-26-04

Base editing in the *AUTS2* gene and high-throughput NGS genotyping of clones: a strategy for generating a cellular model

A.P. Yan [1, 2, 3], P.A. Salnikov [1, 2], A.A. Buzdin [4, 5, 6], V.A. Kovalskaia [7], E.V. Musatova [8], P.S. Orlov [1, 9], O.P. Ryzhkova [7], A.I. Subbotovskaia [9], M.V. Suntsova [4, 5], A.U. Khristichenko [5], A.A. Khabarova [1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia
[4] I.M. Sechenov First Moscow State Medical University of the Ministry of Healthcare of the Russian Federation, Moscow, Russia
[5] National Medical Research Center for Endocrinology named after Academician I.I. Dedov, Moscow, Russia
[6] M.M. Shemyakin–Yu.A. Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, Russia
[7] Research Centre for Medical Genetics, Moscow, Russia
[8] Center of Genetics and Reproductive Medicine "Genetico", Moscow, Russia
[9] Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, Russia
a.yan@alumni.nsu.ru, khabarova@bionet.nsc.ru

Abstract. Studying the molecular mechanisms underlying autism spectrum disorders (ASD) requires cellular models capable of capturing *cis*-regulatory effects and allele-specific gene expression. In this study, we present an approach for generating induced pluripotent stem cells (iPSCs) modified using an adenine base editor (ABE) to introduce synonymous single-nucleotide substitutions in the *AUTS2* gene – a candidate involved in ASD pathogenesis. These substitutions serve as allele-specific markers, enabling the tracking of expression differences between normal and rearranged alleles in a *cis*-regulatory context. We developed a high-efficiency strategy for genotyping clones using amplicon-based next-generation sequencing (NGS). Analysis of over 100 subclones demonstrated that this approach surpasses Sanger sequencing in scalability, sensitivity, and cost-effectiveness. We selected clones with targeted heterozygous substitutions, assessed mosaicism levels, and performed phasing with germline heterozygous variants to confirm monoclonal origin and identify the allele carrying the chromosomal rearrangement. The resulting iPSC lines mark distinct *AUTS2* alleles, providing a foundation for analyzing the impact of *cis*-regulatory elements on gene expression across different cell types. Our findings highlight the practical value of base editors and targeted NGS genotyping in creating cellular models with single-nucleotide substitutions for both basic and applied research.

Key words: induced pluripotent stem cells (iPSCs); CRISPR/Cas9; single-nucleotide substitutions; allele-specific expression analysis; chromosomal rearrangement

For citation: Yan A.P., Salnikov P.A., Buzdin A.A., Kovalskaia V.A., Musatova E.V., Orlov P.S., Ryzhkova O.P., Subbotovskaia A.I., Suntsova M.V., Khristichenko A.U., Khabarova A.A. Base editing in the *AUTS2* gene and high-throughput NGS genotyping of clones: a strategy for generating a cellular model. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):72-84. doi 10.18699/vjgb-26-04

Funding. This work was supported by the RSF grant 24-25-00152.

Acknowledgements. DNA library sequencing was supported by the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2022-310 dated April 20, 2022). Cell culture was performed at the Core Facility "Collection of Pluripotent Human and Mammalian Cell Cultures for General Biological and Biomedical Research" of ICG SB RAS. Bioinformatic analysis of data was carried out with the financial support from the budget project No. FWNR-2025-0017. The work of A.A. Buzdin, M.V. Suntsova, and A.U. Khristichenko was supported by the project of the Ministry of Health of the Russian Federation "Development of a liquid biopsy system for monitoring therapy response and progression of lung cancer".

©Yan A.P., Salnikov P.A., Buzdin A.A., Kovalskaia V.A., Musatova E.V., Orlov P.S., Ryzhkova O.P., Subbotovskaia A.I., Suntsova M.V., Khristichenko A.U., Khabarova A.A. , 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Редактирование оснований в гене *AUTS2* и высокопроизводительное NGS-генотипирование клонов: стратегия создания клеточной модели

А.П. Ян [1, 2, 3], П.А. Сальников [1, 2], А.А. Буздин [4, 5, 6], В.А. Ковальская [7], Е.В. Мусатова [8], П.С. Орлов [1, 9], О.П. Рыжкова [7], А.И. Субботовская [9], М.В. Сунцова [4, 5], А.Ю. Христиченко [5], А.А. Хабарова [1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия
[4] Первый Московский государственный медицинский университет им. И.М. Сеченова Министерства здравоохранения Российской Федерации, Москва, Россия
[5] Национальный медицинский исследовательский центр эндокринологии им. академика И.И. Дедова, Москва, Россия
[6] Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук, Москва, Россия
[7] Медико-генетический научный центр им. академика Н.П. Бочкова, Москва, Россия
[8] Центр генетики и репродуктивной медицины "Genetico", Москва, Россия
[9] Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
a.yan@alumni.nsu.ru, khabarova@bionet.nsc.ru

Аннотация. Изучение молекулярных механизмов, лежащих в основе расстройств аутистического спектра (РАС), требует создания клеточных моделей, способных отражать цис-регуляторные эффекты и аллель-специфичную экспрессию генов. В настоящем исследовании мы представляем подход к получению индуцированных плюрипотентных стволовых клеток (ИПСК), модифицированных с использованием аденинового редактора оснований (ABE), для введения синонимичных однонуклеотидных замен в ген *AUTS2* – кандидата на участие в патогенезе РАС. Эти замены позволяют маркировать аллели и отслеживать различия в экспрессии нормального и реорганизованного аллелей в цис-контексте. Мы разработали стратегию высокоэффективного генотипирования клонов с использованием секвенирования продуктов ПЦР (ампликонов) на платформе нового поколения (NGS). Анализ более 100 субклонов показал, что предложенный подход превосходит секвенирование по Сэнгеру по масштабируемости, чувствительности и экономичности. Мы отобрали клоны с целевыми гетерозиготными заменами, оценили уровень мозаицизма и провели фазирование с герминальными гетерозиготными вариантами, позволяющее убедиться в моноклональном происхождении клеточной линии или идентифицировать аллель, ассоциированный с мутацией. Полученные линии ИПСК маркируют разные аллели гена *AUTS2*, что открывает перспективу анализа влияния цис-регуляторных элементов на экспрессию гена в различных типах клеток. Результаты работы подчеркивают практическую значимость редакторов оснований и целевого NGS-генотипирования при создании клеточных моделей с однонуклеотидными заменами для фундаментальных и прикладных исследований.

Ключевые слова: индуцированные плюрипотентные стволовые клетки (ИПСК); CRISPR/Cas9; однонуклеотидные замены; аллель-специфичная оценка экспрессии; хромосомная перестройка

Introduction

According to Global Burden of Disease 2021 data, approximately 61.8 million people worldwide lived with autism spectrum disorders, with cumulative disability-adjusted life years lost totaling 11.5 million – equivalent to 147 years per 100,000 people (Global Burden of Disease Study 2021…, 2025). Investigating the genetic causes of autism, like other congenital neurological disorders, often requires studying pathological processes occurring in brain tissue. Given the limited availability of this organ for direct cell sampling, establishing cellular models to monitor changes in target gene expression is especially important. iPSCs derived from various patient cell types serve as indispensable tools, as they can be directed to differentiate into multiple lineages, providing flexibility for modeling processes across different tissues (Rowe, Daley, 2019; De Masi et al., 2020).

CRISPR/Cas9 is widely used to recreate pathogenic mutations in cellular models. However, the system has notable limitations, including off-target activity and the imprecise repair of double-strand breaks, both of which can introduce unintended mutations (Smirnov et al., 2016; Uddin et al., 2020). To overcome these drawbacks, more precise CRISPR/Cas9-based technologies are currently being developed. One such approach is base editing, which enables specific nucleotide substitutions (A→G or C→T) without generating double-strand breaks (Gaudelli et al., 2017).

Introducing modifications with single-nucleotide precision enables not only modeling genetic diseases

caused by point mutations (Lu, Huang, 2018; Geurts et al., 2023), but also allele labeling, providing a unique opportunity to trace the impact of *cis*-regulatory variants on gene expression and functional activity at the monoallelic level. Comparing allele expression levels allows the assessment of *cis*-regulatory effects while neutralizing any trans-influences (Salnikov et al., 2024). Additionally, base editors are actively used for developing gene therapy strategies for diseases associated with variants in nuclear and mitochondrial DNA (Billon et al., 2017; Liang et al., 2023).

The base editing system uses a hybrid protein consisting of nuclease-inactive Cas9 and a nucleotide deaminase enzyme. This chimeric protein enables A→G or C→T substitutions without creating double-strand DNA breaks. For the adenine base editor (ABE), deamination occurs on one DNA strand within a narrow deaminase activity window (four nucleotides) while a single-strand nick on the other strand stimulates the repair system to restore the sequence according to the modified strand (Rees, Liu, 2018). The ABE system induces A→G conversion by deaminating adenine to hypoxanthine, which DNA polymerase then recognizes as guanine (Gaudelli et al., 2017; Chen et al., 2023).

Thus, base editors catalyze direct chemical nucleotide conversion without double-strand breaks, minimizing chromosomal aberration risk and eliminating the need for exogenous DNA delivery as a homology-directed repair template. This substantially reduces off-target insertions, deletions, and chromosomal rearrangements compared to classical Cas9.

Despite the obvious advantages of this editing system, base editors have several limitations in application. A key drawback is the catalytic specificity of deaminase enzymes, enabling only transitions (C→T or A→G; purine-to-purine or pyrimidine-to-pyrimidine substitutions), but not transversions (purine-to-pyrimidine or vice versa), significantly narrowing the spectrum of correctable mutations (Komor et al., 2016; Gaudelli et al., 2017; Rees, Liu, 2018). Another serious issue is off-target activity, manifesting both as editing of non-fully complementary DNA sites within the "editing window" and nonspecific cellular RNA modification. Deaminases can randomly deaminate cytosines in RNA, potentially disrupting the transcriptome and causing toxic effects (Grünewald et al., 2019; Jin et al., 2019; Zuo et al., 2019; Yu et al., 2020).

This article focuses on using the ABE base editing system and an innovative cellular clone genotyping strategy for creating iPSC-based cellular models. Karyotype analysis of an autism spectrum disorder patient revealed a chromosomal rearrangement on one allele, with its boundary in an intergenic region near the *AUTS2* gene (Gridina et al., 2025). We hypothesize that this rearrangement may disrupt *AUTS2* expression regulation and potentially be linked to the patient's phenotype. To experimentally test this hypothesis, a model is needed to assess expression differences between the normal allele and the one in *cis* with the chromosomal rearrangement across cell types. However, no single-nucleotide substitutions were found in this patient's *AUTS2* coding region to discriminate transcripts from different alleles. From previously obtained iPSCs of this patient, using the ABE system, we generated model lines carrying heterozygous single-nucleotide substitutions in exon 10 of the *AUTS2* gene. These lines enable obtaining neurons/neural precursors *in vitro* and quantitatively assessing allele-specific *AUTS2* transcription from rearranged *vs*. intact chromosomes.

Clone genotyping and editing efficiency assessment are routinely performed by Sanger sequencing of target locus amplicons. The most convenient and accurate approach is sequencing the amplicon containing the target substitution. For amplicon sequencing, two methods can be applied: Sanger and NGS (Fig. 1*c*, *d*). The choice between Sanger and NGS depends on specific research tasks. Sanger remains optimal for analyzing individual PCR products with few samples. Its limited throughput makes it economically unfeasible for large sample sets. In contrast, NGS technologies enable simultaneous sequencing of numerous samples, significantly reducing per-sample analysis cost and making the method indispensable for high-throughput analysis. However, NGS requires bioinformatics. Thus, for routine single amplicon analysis Sanger retains advantages, while for scaling analysis to large sample numbers NGS demonstrates clear superiority.

In this study, NGS analysis of 117 subclones identified cell lines suitable for differentiation and *AUTS2* expression analysis. We provide a detailed protocol for NGS genotyping sample preparation and bioinformatics analysis of obtained data, adaptable for other genotyping tasks.

Materials and methods

Creation of the MLM3636-BstV2I vector. The base vector was plasmid MLM3636 (Addgene #43860), containing a guide RNA (gRNA) sequence with a spacer

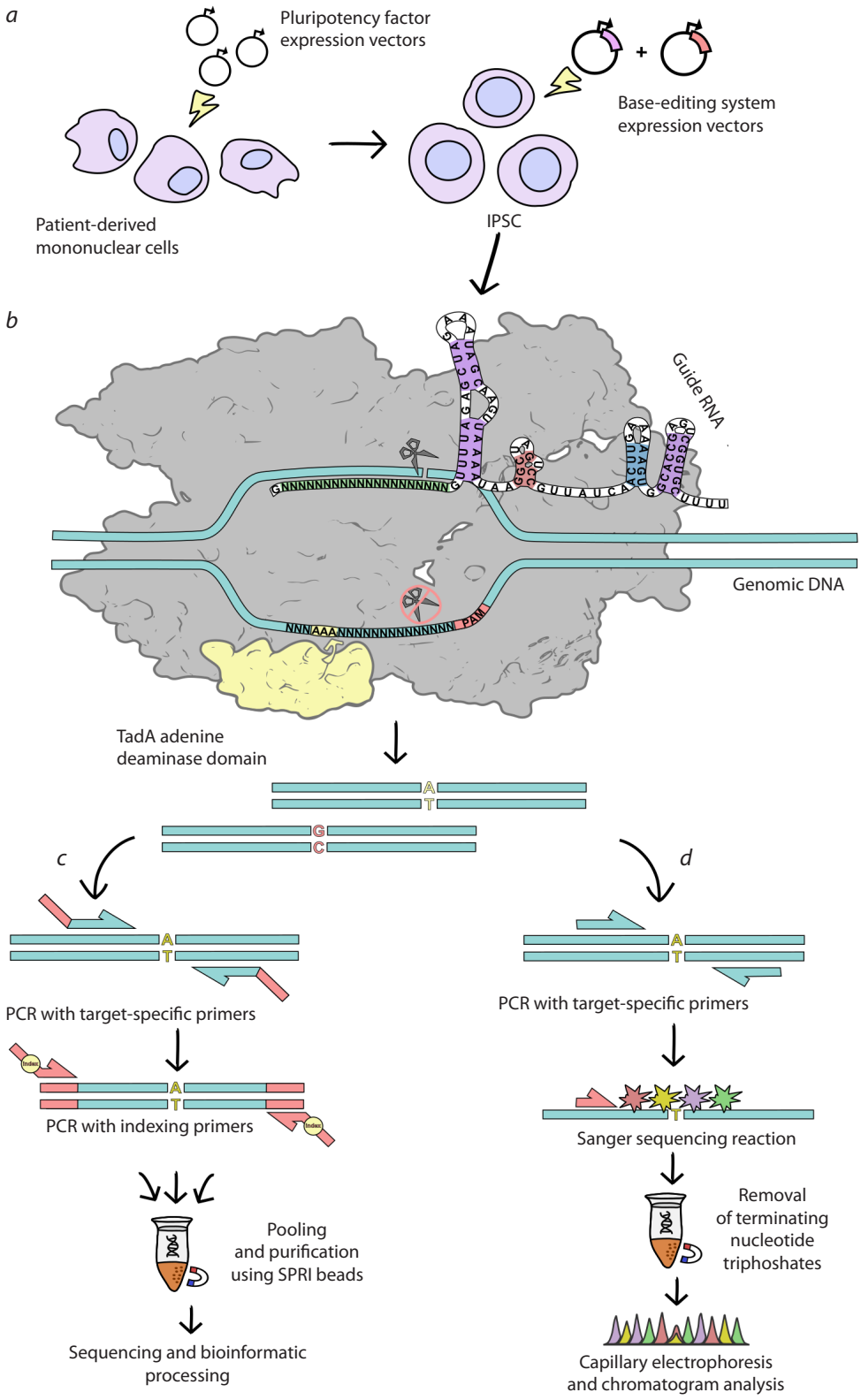


Fig. 1. General workflow for generating single-nucleotide substitutions in human iPSCs and comparison of clone genotyping protocols using NGS and Sanger sequencing.

a – electroporation of patient-derived blood monocytes with reprogramming factor vectors to obtain iPSCs, followed by electroporation of iPSCs with plasmids encoding the base editing system: the target sgRNA and a modified Cas9 fused to the TadA adenine deaminase; *b* – schematic representation of the complex of modified Cas9 (gray), fused to TadA (light yellow), with an sgRNA and protospacer sequence on genomic DNA. Adenines within the editing window are highlighted in yellow; *c* – NGS library preparation scheme using addition of adapter sequences and indexing via two rounds of PCR; *d* – standard Sanger sequencing workflow.

cloning site designed for BsmBI restriction enzyme (New England Biolabs, USA). The following protospacer sequences were used: 5′-CAAAAGTTGACCCATTC TAC-3′ for N044 and 5′-TCACACTGTGCCGGTA GAAT-3′ for N068. Since this enzyme is not produced in Russia, to simplify subsequent manipulations, we replaced the original cloning site with an analogous site from plasmid pSpCas9(BB)-2A-GFP (PX458) (Addgene #48138), recognized by BstV2I enzyme produced by the Russian manufacturer SibEnzyme LLC.

For this, we amplified the fragment containing the pSpCas9(BB)-2A-GFP (PX458) cloning site using primers pSpCas9_clonF and pSpCas9_clonR, and nearly the entire MLM3636 vector sequence using primers MLM3636_clonF and MLM3636_clonR. The product was treated with 1 U DpnI enzyme (New England Biolabs, USA) and purified using SPRI bead suspension VAHTS DNA Clean Beads (Vazyme, China). Assembly was performed using NEBuilder HiFi (New England Biolabs, USA) according to the manufacturer's protocol.

To obtain vectors expressing gRNAs, 1 μg of MLM3636-BstV2I plasmid was linearized with 25 U BstV2I enzyme in a 50 μl reaction mixture at 55 °C overnight. After restriction, DNA was purified using 0.8× volume of SPRI bead suspension VAHTS DNA Clean Beads (Vazyme, China). Single-stranded oligonucleotides N044_F and N044_R (2 μM each) were phosphorylated in 1× T4 DNA ligase buffer with 100 nM ATP and 10 U T4 polynucleotide kinase (SibEnzyme, Russia). The same procedure was applied for the N068 pair. Oligonucleotides were then heated to 95 °C and hybridized by slow cooling at 0.1 °C/s to 4 °C.

The resulting double-stranded oligonucleotide N044 (N068) F+R solution was ligated into the linearized MLM3636-BstV2I vector backbone. Components mixed on ice were: 100 nM ds-oligo solution, up to 20 ng previously linearized plasmid, 100 nM ATP, 1× T4 DNA ligase SE buffer, 20 U T4 DNA ligase. The reaction mixture was incubated overnight at 4 °C. The product was purified using 0.8× volume of SPRI bead suspension VAHTS DNA Clean Beads (Vazyme, China) and eluted in 5 μl H_2O. Half the product volume was used for subsequent bacterial cell transformation.

Escherichia coli transformation was performed by electroporation, plasmid DNA was isolated using Midi-prep kit (Evrogen, Russia), and additionally purified using 0.8× volume of VAHTS DNA Clean Beads (Vazyme, China). Elution was performed in a minimal water volume to achieve DNA concentration ≥1 μg/μl. For puromycin selection, plasmid pURC_puro (provided by A. Nurislamov) was used. This construct expresses a puromycin resistance gene under a CMV promoter integrated into the plasmid, enabling selection and enrichment of the cell population for editing events.

iPSC culture and neon electroporation. iPSCs were cultured on plastic plates pre-coated with Matrigel cultural matrix solution (Corning, USA) using mTeSR1 medium (StemCell, Canada) supplemented with penicillin/streptomycin antibiotic mixture (PanEco, Russia). Cells were dissociated using TripLE reagent (ThermoFisher, USA) and passaged at a 1:3–1:5 ratio with the addition of the Y-27632 inhibitor (10 μM, Rho-kinase inhibitor) (StemCell, Canada). The day before electroporation, iPSC lines were passaged to achieve ≤70 % confluency. Before transferring transfected cells, culture plate bottoms were coated with Matrigel matrix (1 ml per 10 cm^2) and incubated at 37 °C for 20–30 minutes. 1 hour before transformation, 10 μM Y-27632 was added to cells.

Cell culture was dissociated using 700 μl TripLE (ThermoFisher, USA). 1 ml wash medium was added per well, cells were gently resuspended, and the suspension was collected into a centrifuge tube. Cell counting was performed using a hemocytometer. Suspension was centrifuged at 300 g for 5 min. After counting, the required cell suspension volume (150–200 thousand cells per 1 electroporation reaction) was washed once in phosphate buffer. In a 1,500 μl tube, three plasmids were mixed (0.5 μg each): gRNA expression plasmid (MLM3636-BstV2I), base editor (Addgene #108382), and pURC_puro (puromycin resistance). 150–250 thousand cells, resuspended after washing in phosphate buffer, were added into Neon electroporation buffer R. Total cell suspension and plasmid mixture volume was not supposed to exceed 10 μl.

Electroporation was performed using Neon system at 1,100 V – 30 ms – 1 pulse, and cells were transferred to a 6-well culture plate well in 3 ml antibiotic-free mTeSR1 medium. After 24 hours, antibiotics (1× penicillin/streptomycin mixture) and selective antibiotic puromycin (1 μg/ml for 48 hours) (Gibco, USA) were added to the medium. The medium was fully changed after 1 day. At 7–10 days, individual cell colonies were manually transferred to separate wells of a 24-well plate (~30 colonies picked). When colonies reached 1/4 well area, they were dissociated using TrypLE, and after centrifugation, part of the cell suspension was used for DNA extraction and

Table 1. Primers and oligonucleotides

Name	Sequence 5′–3′
MLM_genF	TCGGGCAGGAAGAGGGCCTATTTC
MLM_genR	CCTCGAGCGGCCCAAGCTTAAAAA
H044_F	CACCGCAAAAGTTGACCCATTCTAC
H044_R	AAACGTAGAATGGGTCAACTTTTGC
H068_F	CACCGTCACACTGTGCCGGTAGAAT
H068_R	AAACATTCTACCGGCACAGTGTGAC
AUTS_gen_F	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTGACAGCTTAATGACAGGGAAGC
AUTS_gen_R	TCTTTCCCTACACGACGCTCTTCCGATCTCTGATCTGTGGCAAGAGCACAA
P5_IP{index_ID}	AATGATACGGCGACCACCGAGATCTACAC{8bp_Index}ACACTCTTTCCCTACACGAC
P7_IP{index_ID}	CAAGCAGAAGACGGCATACGAGAT{8bp_Index}GTGACTGGAGTTCAGACGTGT
pSpCas9_clonF	GGCCTATTTCCCATGATTCCT
pSpCas9_clonR	CGACTCGGTGCCACTTTT
MLM3636_clonF	AGGAATCATGGGAAATAGGCC
MLM3636_clonR	AAAAGTGGCACCGAGTCG
AUTS2_phase_F1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTACTAGCTTTTGCTTTGATCCC
AUTS2_phase_R1	TCTTTCCCTACACGACGCTCTTCCGATCTGCCTGACTGCCACTAAAGAG
AUTS2_phase_F2	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCACTGTAGCAGTGAACAAA
AUTS2_phase_R2	TCTTTCCCTACACGACGCTCTTCCGATCTATGCCTGACTGCCACTAAAG

subsequent genotyping, while the remainder was frozen in KSR (Gibco, USA) with 10 % DMSO.

Clone genotyping using amplicon NGS sequencing. PCR setup used reagents from MBS-Technology (Russia). Total 50 µl volume contained: 50 ng DNA, 1× MBUsion buffer, 0.2 mM of each dNTP, 0.2 µM AUTS_gen_F primer, 0.2 µM AUTS_gen_R primer (Table 1), 0.4 µl MBUsion polymerase (MBS, Russia), H_2O to 50 µl. PCR conditions were as follows: 95 °C for 3 min, 20 cycles: 95 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s, then 72 °C for 1 min, 4 °C. PCR products were diluted to 1:100, and 2 µl was used as a template for indexed PCR including full sequencing-required sequences P5_IP{index_ID} and P7_IP{index_ID} (Table 1).

Amplification conditions were: 95 °C for 3 min, 8 cycles: 95 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s, then 72 °C for 1 min, 4 °C. PCR product concentrations were approximately assessed by agarose gel electrophoresis visualization, samples were pooled in equal amounts, purified using 1 volume VAHTS DNA Clean Beads (Vazyme, China), and eluted in 20 µl H_2O.

Sequencing was performed in paired-end 150 bp mode (~10,000 reads per sample). Adapter sequences were removed (cutadapt) and aligned to indexed target region reference sequence using bowtie2 (default parameters). BAM files were visualized in Integrative Genomics-Viewer (IGV, USA).

Variant identification used bcftools mpileup with allele depth recording (AD flag). A custom Python script calculated mutant allele frequency per position, aggregated results into matrix, and generated line graphs for sample groups to assess allele frequency distribution across all possible substitution positions. All scripts with examples are available on GitHub: https://github.com/Somatich/NGS-genotyping-Yan-et-al.-2025.

Phasing introduced substitutions with a germline variant. Phasing of single-nucleotide substitutions determines whether two variants are in *cis* or *trans* configuration. We designed primer pairs flanking both induced mutagenesis site and germline variant (chr7:70768282 hg38, rs3829006 G/A), yielding 363 bp of PCR product; 5′-ends included NGS technical sequences. This germline variant is 238 bp from the mutagenesis site, selected from patient exome sequencing data (Gridina et al., 2025). Distance ≤200–300 nucleotides yields optimal PCR product length for Illumina NGS sequencing

(DNA fragments >500 bp significantly reduce efficiency). Two primer pairs were used (AUTS2_phase_F1/R1 and AUTS2_phase_F2/R2) (Table 1).

PCR setup used MBS-Technology reagents (Russia) as described above. Indexed PCR and cleanup were performed identically. Libraries were sequenced paired-end (150 bp), yielding ~5,000 read pairs per sample. Reads were mapped using Bowtie2 (default settings). SNP phasing was performed manually by analyzing alignments in the IGV genome browser.

Results

Design of sgRNAs and evaluation of editing efficiency in exon 10 of the *AUTS2* gene

The aim of this study was to generate an iPSC-based cellular model carrying a heterozygous synonymous SNV in exon 10 of the *AUTS2* gene using adenine base editing (ABE) (Fig. 1*a*). Exon 10 was selected because it is present in most of the described and predicted transcripts; moreover, a germline single-nucleotide variant located near exon 10 enables phasing of the introduced substitution with respect to alleles.

We identified adenine positions located at the third codon positions and falling within the active ABE editing window (nucleotides 4–7 of the protospacer). Thus, A→G substitutions at these positions are synonymous. Two such adenines were identified: chr7:70768044 A→G and chr7:70768068 T→C (hg38). The chr7:70768044 A→G substitution yields a synonymous AAA→AAG change (p.Lys570Lys). This variant occurs in the population with a frequency of 0.000005596 and is not predicted to affect splicing or create a novel splice site (SpliceAI, MobiDetails). The chr7:70768068 T→C substitution is likewise synonymous (AGT→AGC), is absent from gnomAD, and is also not predicted to influence splicing (SpliceAI, MobiDetails) (Reese et al., 1997; Jaganathan et al., 2019; Baux et al., 2021).

For both candidate sites, two guide RNAs – H044 and H068 – were designed (Table 1). Their *in vivo* activity was tested in iPSC culture using a plasmid expressing Cas9 (Addgene #62988). Both sgRNAs demonstrated comparable editing efficiencies: 9.2 % for H044 at chr7:70768044 and 7.6 % for H068 at chr7:70768068.

Editing outcomes were analyzed using Cas-Analyser (Crisper Rgen) (Hwang et al., 2018). It is important to emphasize that genome editing efficiency is determined by multiple poorly controllable factors, such as the local chromatin environment in a given cell line. Therefore, computational predictions of sgRNA activity and specificity cannot be fully relied upon. To increase the likelihood of obtaining the desired mutation, it is reasonable to use multiple sgRNAs directed at the same region. Based on this rationale, both sgRNAs were carried forward into subsequent experiments.

As a pilot experiment, we assessed the efficiency of each guide in combination with the adenine base editor Cas9(ABE7.10) in a bulk cell population (i. e., without selecting for transfected cells). Three days after electroporation with plasmids encoding Cas9 and the sgRNA, genomic DNA was extracted and libraries were prepared for NGS. BE-Analyser (Crisper Rgen, South Korea) (Hwang et al., 2018) was used for analysis. The ABE-mediated editing efficiency for H044 at chr7:70768044 was 5.1 %.

Four adenines lie adjacent to the target nucleotide (chr7:70768041–70768043). Their editing frequencies were 3 % at chr7:70768043; 0.4 and 0.7 % at chr7:70768041 and chr7:70768042, respectively. For H068, the efficiency at chr7:70768068 was 3 %. These results indicate that, despite the presence of a cluster of four adenines near the target site, the H044 protospacer yields higher editing efficiency at the intended position compared to H068. Although the relatively high off-target editing frequency at chr7:70768043 was not critical for further use of H044, it necessitated screening a larger number of clones to identify lines carrying only the desired substitution.

To increase the proportion of successfully edited cells in the population, we additionally co-transfected a plasmid containing a selectable marker (puromycin resistance), enabling enrichment of transfected cells and substantially improving the yield of edited clones. Taken together, the pilot data confirm that ABE editing is functional at the AUTS2 locus. The H044 sgRNA demonstrates superior on-target efficiency relative to H068, despite potential issues associated with the adenine cluster.

Generation and selection of iPSC clones carrying the introduced nucleotide substitution

Electroporation using the Neon™ system was performed to introduce the target SNV into exon 10 of *AUTS2*. Two patient-derived iPSC lines were electroporated with a mixture of three plasmids encoding the sgRNA, the adenine base editor and a puromycin resistance marker. After electroporation, cells were plated at low density onto matrix-coated plates and cultured in mTesR1 me-

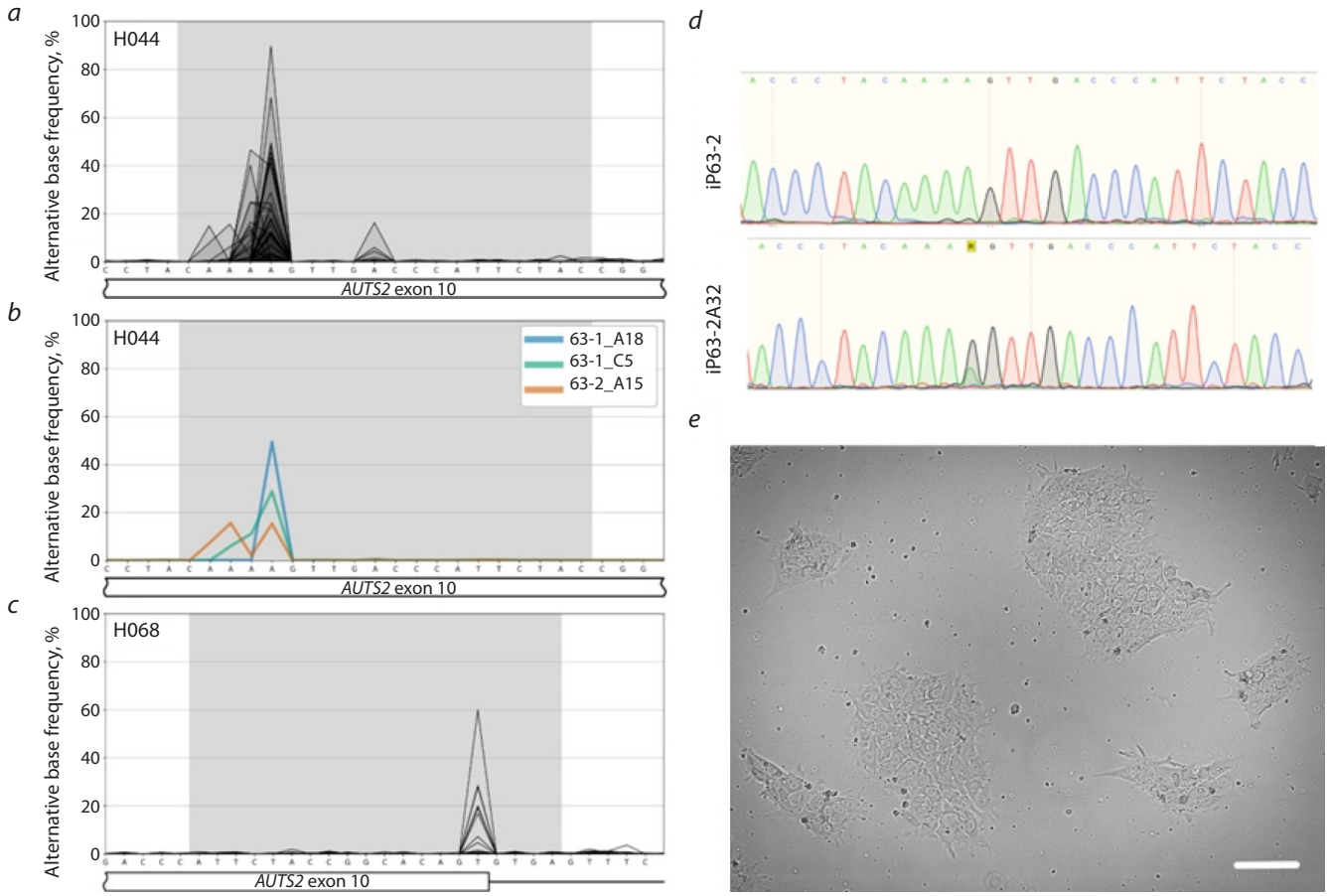


Fig. 2. Examples of the results obtained using the presented protocol.

a – visualization of the aggregated genotyping results for 69 cell clones modified using guide RNA H044; *b* – example of the visualization of genotyping results for three selected samples. The X axis shows the genomic sequence chr7:70768037–70768062, hg38 (in nucleotides), and the Y axis shows the percentage of sequencing reads containing a substitution; *c* – visualization of the aggregated genotyping results for 48 cell clones modified using guide RNA H068. The X axis shows the genomic sequence chr7:70768048–70768077, hg38. The grey area corresponds to the guide RNA sequence used in the experiment; *d* – sequencing chromatograms of samples iP-63-2 (the original iPSC line) and iP-63-2A32 (the iPSC line with a synonymous AAA/AAG triplet substitution) at the mutagenesis site; *e* – morphology of the iPSC culture in transmitted light; the white bar represents 100 µm.

dium supplemented with puromycin for three days. On days 7–10, individual colonies exhibiting characteristic iPSC morphology were visually identified and manually picked (Fig. 2*e*). A total of 117 clones displaying stable growth and typical pluripotent morphology were retained for further analysis.

A key challenge at this stage was genotyping over 100 samples. Conventional assessment of genome editing outcomes relies on Sanger sequencing of PCR amplicons spanning the edited region (Fig. 1*c*, *d*). However, this approach becomes labor-intensive and economically inefficient when analyzing >80 samples (Fig. 3). We therefore implemented high-throughput genotyping via next-generation sequencing of amplicons. This method provides excellent scalability and dramatically accelerates analysis, allowing simultaneous determination of both allele composition and allele frequencies within each clone.

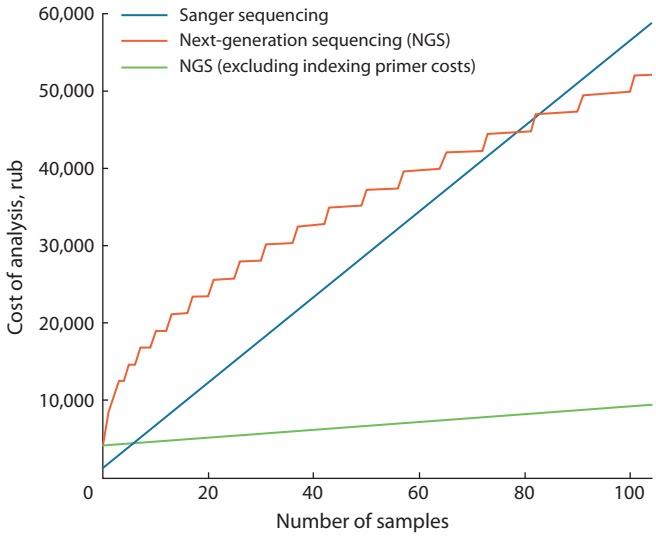


Fig. 3. Dependence of genotyping cost by Sanger sequencing (blue) and by NGS, with (orange) and without taking into account (green) the cost of synthesizing indexing primers.

For NGS analysis, the target region harboring the substitution was amplified with site-specific primers flanking the editing site. For paired-end 150 bp reads, optimal primer placement is 50–130 bp from the mutation site (Fig. 1*c*), ensuring that both reads cover the edited base. Technical sequences corresponding to Illumina 3′ adapter regions were added to the 5′ ends of site-specific primers. A second PCR was performed using universal primers containing index sequences for sample pooling and demultiplexing.

Assessment of editing efficiency and mosaicism in clones

In total, 117 clones were analyzed: 69 edited with H044 (Fig. 2*a*), and 48, with H068 (Fig. 2*c*). Overall editing efficiency for H044 – defined as the proportion of clones with substitutions occurring at ≥2 % allele frequency – was 54 % (37/69). For downstream applications, we specifically required heterozygous substitutions; so, we filtered clones with edited allele frequencies ≥30 %. At this threshold, H044 editing efficiency was 13 % (9/69). Two clones exhibited substitution frequencies of 68 and 89 %, suggesting potential homozygosity. For H068, the overall efficiency was 31 % (15/48) at the 2 % threshold, and 4% (2/48) at the 30 % threshold.

Observed allele frequencies often deviated from the theoretically expected values (0, 50, 100 %), even in ostensibly clonal lines. Some variability likely arises from library preparation-related biases (multiple PCR cycles, cross-contamination risk). However, a major contribution likely stems from cellular cross-contamination during subcloning, as imperfect dissociation, reaggregation, and migration of cells during early culture can occur. Additional variability is introduced by asynchronous editing events: between transfection and onset of editor expression, some cells may divide, giving rise to subpopulations with different mutations or no mutations, producing mosaic colonies. Therefore, regardless of initial editing efficiency, screening a large number of clones is essential to exclude mosaic lines.

Within the editing window (protospacer positions 4–6), multiple adenines were present. Substitutions occurred at all "editable" positions but with varying efficiency: the highest frequency was observed at position 5 (chr7:70768044; ~45 % positive clones at ≥2 % threshold), approximately twice the frequency of position 4 (chr7:70768043; ~17 %) (Fig. 2*a*). Position 6 harbors a guanine and therefore cannot be assessed, although it falls within the theoretical window. Editing frequencies at other adenines in the region did not exceed 5 %. Notably, no cases of simultaneous double substitutions within the same allele were detected, nor were T→C conversions observed on the non-target strand.

Figure 2 illustrates the visualization strategy used for presenting these results (Fig. 2*a*–*c*) – a convenient method for rapid comparison of samples, noise assessment, and identification of candidate clones for further work.

We selected five colonies carrying appropriate substitutions and displaying minimal mosaicism (<5 % mosaic alleles). To determine which allele carried the introduced substitution, as well as to assess clone monoclonality, phasing was performed with a germline heterozygous variant in the surrounding region (SNV chr7:70768282 hg38, rs3829006 G/A), previously identified by whole-genome sequencing and marking the allele in *cis* with the structural variant. Three non-mosaic clones were selected; two of them carried a synonymous SNV in a heterozygous state, each on a different allele (Table 2).

After confirming the absence of mosaicism, these iPSC lines were expanded for subsequent analysis of *AUTS2* allele-specific expression. To prevent errors during large-scale culturing, heterozygous substitutions

Table 2. Phasing and assessment of mosaicism in AUTS2-edited colonies

Clone	SNV position, hg38	Frequency, %	Allele	Mosaicism
63-2_A32	chr7:70768044	41.5	SNV	–
63-1_C16	chr7:70768043	40.2	WT	–
63-1_C21	chr7:70768044	48.4	WT	–
63-2_A30	chr7:70768043	46.5	WT	+
63-2_A30	chr7:70768044	39.5	SNV	+
63-2_C19	chr7:70768064	28.60	SNV	+

Note. WT – wild type, SNV – single nucleotide variant.

were re-validated by Sanger sequencing (Fig. 2*d*). Thus, we successfully generated the required iPSC lines and obtained a tool suitable for analyzing the impact of the patient's chromosomal rearrangement on AUTS2 expression.

Discussion

In this study, we generated three genetically modified patient-derived iPSC lines carrying single-nucleotide substitutions in AUTS2 exon 10 and demonstrated the effectiveness of an NGS-based strategy for high-throughput genotyping of edited clones. The resulting lines are genetically homogeneous, derived from two independent iPSC lines from the same individual; two of them carry synonymous SNV marking different alleles. These lines are fully suitable for subsequent analysis of how the patient's chromosomal rearrangement influences AUTS2 expression and its functional consequences.

Allelic marking enables discrimination of *cis*-regulatory differences within the same cellular environment. This experimental design minimizes the influence of poorly controlled external factors: because both alleles are exposed to identical trans-acting regulators, any detected expression differences reflect only *cis*-regulatory variation. For this reason, even non-synonymous substitutions can be used, since any feedback-mediated effect on transcription would affect both alleles equally and thus would not distort *cis*-regulation assessment.

The application of NGS for genotyping significantly reduces laboratory handling time and simplifies the analysis process compared to first-generation sequencing. In contrast to Sanger sequencing of individual samples, NGS enables process scaling and allows the analysis of more than 100 samples in a single run (Table 3). Preparation of an NGS library requires routine DNA extraction and two rounds of PCR using primers specific to the target genomic region of interest, which also introduce the necessary technical sequences required for sequencing.

The integration of NGS with bioinformatics tools enables automated identification of nucleotide variants

Table 3. Comparison of genotyping by Sanger sequencing and by NGS

Criterion	Sanger sequencing	NGS genotyping
Scalability	Can sequence up to eight samples at once; results analyzed individually	Enables sequencing and analysis of dozens to hundreds of samples simultaneously
Sample preparation workload	One PCR round, sequencing reaction, and cleanup; individual sample prep required	Two PCR rounds with target-specific and indexed primers; after indexing, samples can be pooled and SPRI-cleaned
Speed of data analysis	Manual chromatogram interpretation	Fast and automated analysis
Sensitivity	Averaged signal; difficult to interpret with multiple alleles; detection threshold ~15–20 %	Allele-specific quantification; detects variants down to ~0.1 %, including low-frequency variants and mosaicism
Sequence quality	Terminal regions (~100 bp) are often low-quality and unsuitable for analysis	High-quality terminal read regions
Primer requirements	Standard primers	Long primers; error-sensitive; require PAGE purification
Clone selection	Cannot reliably detect mosaic or subclonal events	Enables selection of clones with defined allele frequencies; detects subclones; excludes mosaicism
Method applicability	Suited for verifying high-frequency mutations; cannot detect low editing efficiency (<~20 %)	Applicable to many tasks: quantitative editing assessment, subpopulation tracking; detects as little as 5–9 % editing
Cost	~700 RUB per sample	~100 RUB per sample

Table 4. Cost estimation of sample genotyping using Sanger sequencing and NGS

Stage		Item	Quantity	Price per unit	Total	
		Sanger sequencing				
Per site	Site-specific primers	Synthesis of 20-nt oligonucleotides	2*	540	1,080	
Per sample	PCR	MBUsion polymerase	1	20	20	555***
	Sanger reaction	GenSeq-1000	2	159.5	319	
	Cleanup	Ethanol precipitation	2	0	0	
		Sephadex G-50	2 for 20 mg	4	8	
		SeqMag-1000	2	33	66	
		VAHTS DNA Clean Beads	2	55	110	
	Sequencing	Genetic Analyzer "NANOFOR 05"	2	104**	208	
		NGS				
Per site	Site-specific primers	Synthesis of 55-nt oligonucleotides	2	2,035	4,070	
Depends on the number of samples in pool	Indexing primers*	Synthesis of 55-nt oligonucleotides	$2\sqrt{N}$, where N is the number of samples	2,035	$4{,}060\sqrt{N}$, where N is the number of samples	
Per sample	PCR	MBUsion polymerase	2	20	40	49.744
	Sequencing	GenoLab M V2.0 (FCM Cell, 300 cycles, 250M reads)	0.01M reads	243,600	9,744	
Per pool	Cleanup	VAHTS DNA Clean Beads	1	110	110	

* Ideally, a system of four primers should be used: one pair for product generation via PCR, and two more for sequencing in both directions.
** The cost of analyzing one sample recalculated per polymer consumption in a 35-cm capillary.
*** Including Sephadex G-50 cleanup.

and eliminates errors associated with chromatogram quality in the Sanger method. NGS-based genotyping significantly reduces labor and simplifies analysis compared to first-generation sequencing. Unlike Sanger sequencing, which measures an averaged signal and can detect variants only at ≥20 % abundance (based on chromatogram inspection) (Davidson et al., 2012; Bennett et al., 2020), NGS provides allele-level information with detection sensitivity down to ~2 %, eliminating interpretative ambiguity.

Using this approach, we assessed sgRNA performance in bulk cell populations and identified a low modification rate (~5 %), prompting a shift to constructs with selectable markers to enrich for successfully transfected cells.

Subsequent NGS analysis enabled identification of non-mosaic clones with the desired edits – critical for establishing stable genetic lines. High sequencing depth (typically ~10,000 reads per 300-bp amplicon) allows reliable detection of rare variants that are missed by less sensitive methods.

Although NGS is often perceived as prohibitively expensive, targeted amplicon sequencing reduces per-sample sequencing costs to <10 RUB. The main expenses are index primers, with total cost scaling as $\sqrt{N}$ (where N is the number of samples). These primers are universal and compatible with both the library preparation scheme described here and standard protocols using ligated adapters. Commercial index kits are available but generally more expensive. Accounting for primer costs, NGS becomes more cost-effective than Sanger for >80 samples; without primer synthesis costs (i. e., using existing primer sets), the cost advantage appears already at >7 samples (Fig. 3; Table 4). Thus, the proposed method is highly suitable for widespread use.

Amplicon NGS can also be applied to assess outcomes of other CRISPR/Cas9-based editing systems. Even when induced mutations (e. g., large deletions or inversions) exceed the read length limit (~300 bp paired-end), sequencing only the mutation boundaries is sufficient to verify adjacent sequence integrity and identify undesired events. The method remains applicable as long as primer sites are intact – a limitation shared with all PCR-based genotyping.

Nonetheless, several limitations must be considered. Multiple PCR cycles can introduce amplification artifacts. Because amplicons undergo many manipulations prior to indexing and the method is highly sensitive, the risk of contamination is significant; strict post-PCR handling practices must be followed, and contact with pre-PCR areas must be avoided. Additionally, primers containing long non-genomic 5′ extensions often reduce amplification efficiency and require optimization of PCR conditions for each primer pair.

Conclusion

This article presents a method for obtaining model lines based on iPSCs with a synonymously mutated allele of the *AUTS2* gene marked using base editor technology. The obtained iPSC lines will be used to study the influence of chromosomal aberrations on *AUTS2* expression and to investigate the functional consequences of this impairment.

The proposed approach to genotyping a large number of clones using NGS amplicon sequencing demonstrates high efficiency and scalability while reducing the cost of analysis compared to traditional methods. The high throughput of the method allowed the analysis of a large number of samples, as well as the assessment of mosaicism, composition, and allele frequencies in the cell population.

Thus, the presented strategy allows highly efficient generation of cellular models for studying the influence of *cis*-regulatory variants on gene transcriptional activity, while controlling the genetic homogeneity of the modified cellular models, which can be in demand both in basic research and in solving applied problems.

References

Baux D., Van Goethem C., Ardouin O., Guignard T., Bergougnoux A., Koenig M., Roux A.-F. MobiDetails: online DNA variants interpretation. *Eur J Hum Genet.* 2021;29(2):356-360. doi 10.1038/s41431-020-00755-z

Bennett E.P., Petersen B.L., Johansen I.E., Niu Y., Yang Z., Chamberlain C.A., Met Ö., Wandall H.H., Frödin M. INDEL detection, the 'Achilles heel' of precise genome editing: a survey of methods for accurate profiling of gene editing induced indels. *Nucleic Acids Res*. 2020;48(21):11958-11981. doi 10.1093/nar/gkaa975

Billon P., Bryant E.E., Joseph S.A., Nambiar T.S., Hayward S.B., Rothstein R., Ciccia A. CRISPR-mediated base editing enables efficient disruption of eukaryotic genes through induction of STOP codons. *Mol Cell*. 2017;67(6):1068-1079.e4. doi 10.1016/j.molcel.2017.08.008

Chen X., McAndrew M.J., Lapinaite A. Unlocking the secrets of ABEs: the molecular mechanism behind their specificity. *Biochem Soc Trans*. 2023;51(4):1635-1646. doi 10.1042/BST20221508

Davidson C.J., Zeringer E., Champion K.J., Gauthier M.-P., Wang F., Boonyaratanakornkit J., Jones J.R., Schreiber E. Improving the limit of detection for Sanger sequencing: a comparison of methodologies for *KRAS* variant detection. *BioTechniques*. 2012;53(3):182-188. doi 10.2144/000113913

De Masi C., Spitalieri P., Murdocca M., Novelli G., Sangiuolo F. Application of CRISPR/Cas9 to human-induced pluripotent stem cells: from gene editing to drug discovery. *Hum Genomics*. 2020;14(1):25. doi 10.1186/s40246-020-00276-2

Gaudelli N.M., Komor A.C., Rees H.A., Packer M.S., Badran A.H., Bryson D.I., Liu D.R. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*. 2017;551(7681):464-471. doi 10.1038/nature24644

Geurts M.H., Gandhi S., Boretto M.G., Akkerman N., Derks L.L.M., Van Son G., Celotti M., … Andersson-Rolf A., Chuva De Sousa Lopes S.M., Van Es J.H., Van Boxtel R., Clevers H. One-step generation of tumor models by base editor multiplexing in adult stem cell-derived organoids. *Nat Commun*. 2023;14(1):4998. doi 10.1038/s41467-023-40701-3

Global Burden of Disease Study 2021 Autism Spectrum Collaborators. The global epidemiology and health burden of the autism spectrum: findings from the Global Burden of Disease Study 2021. *Lancet Psychiatry*. 2025;12(2):111-121. doi 10.1016/S2215-0366(24)00363-8

Gridina M., Lagunov T., Belokopytova P., Torgunakov N., Nuriddinov M., Nurislamov A., Nazarenko L.P., … Filipenko M., Rogaev E., Shilova N.V., Lebedev I.N., Fishman V. Combining chromosome conformation capture and exome sequencing for simultaneous detection of structural and single-nucleotide variants. *Genome Med*. 2025;17(1):47. doi 10.1186/s13073-025-01471-3

Grünewald J., Zhou R., Garcia S.P., Iyer S., Lareau C.A., Aryee M.J., Joung J.K. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature.* 2019;569(7756):433-437. doi 10.1038/s41586-019-1161-z

Hwang G.-H., Park J., Lim K., Kim S., Yu J., Yu E., Kim S.-T., Eils R., Kim J.-S., Bae S. Web-based design and analysis tools for CRISPR base editing. *BMC Bioinformatics*. 2018;19(1):542. doi 10.1186/s12859-018-2585-4

Jaganathan K., Kyriazopoulou Panagiotopoulou S., McRae J.F., Fazel Darbandi S., Knowles D., Li Y.I., Kosmicki J.A., … Gao H., Kia A., Batzoglou S., Sanders S.J., Farh K.K.-H. Predicting splicing from primary sequence with deep learning. *Cell.* 2019;176(3):535-548.e24. doi 10.1016/j.cell.2018.12.015

Jin S., Zong Y., Gao Q., Zhu Z., Wang Y., Qin P., Liang C., Wang D., Qiu J.-L., Zhang F., Gao C. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science.* 2019;364(6437):292-295. doi 10.1126/science.aaw7166

Komor A.C., Kim Y.B., Packer M.S., Zuris J.A., Liu D.R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature.* 2016;533(7603):420-424. doi 10.1038/nature17946

Liang Y., Chen F., Wang K., Lai L. Base editors: development and applications in biomedicine. *Front Med.* 2023;17(3):359-387. doi 10.1007/s11684-023-1013-y

Lu Z., Huang X. Base editors: a powerful tool for generating animal models of human diseases. *Cell Stress*. 2018;2(10):242-245. doi 10.15698/cst2018.10.156

Rees H.A., Liu D.R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat Rev Genet*. 2018;19(12): 770-788. doi 10.1038/s41576-018-0059-1

Reese M.G., Eeckman F.H., Kulp D., Haussler D. Improved splice site detection in Genie. In: Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB '97). 1997;232-240. doi 10.1145/267521.267766

Rowe R.G., Daley G.Q. Induced pluripotent stem cells in disease modelling and drug discovery. *Nat Rev Genet*. 2019;20(7):377-388. doi 10.1038/s41576-019-0100-z

Salnikov P., Belokopytova P., Yan A., Viesná E., Korablev A., Serova I., Lukyanchikova V., Stepanchuk Y., Torgunakov N., Tikhomirov S., Fishman V. Direction and modality of transcription changes caused by TAD boundary disruption in *Slc29a3/Unc5b* locus depends on tissue-specific epigenetic context. *Epigenetics Chromatin*. 2025; 18(1):55. doi 10.1186/s13072-025-00618-1

Smirnov A.V., Yunusova A.M., Lukyanchikova V.A., Battulin N.R. CRISPR/Cas9, a universal tool for genomic engineering. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2016;20(4): 493-510. doi 10.18699/VJ16.175 (in Russian)

Uddin F., Rudin C.M., Sen T. CRISPR gene therapy: applications, limitations, and implications for the future. *Front Oncol.* 2020;10:1387. doi 10.3389/fonc.2020.01387

Yu Y., Leete T.C., Born D.A., Young L., Barrera L.A., Lee S.-J., Rees H.A., Ciaramella G., Gaudelli N.M. Cytosine base editors with minimized unguided DNA and RNA off-target events and high on-target activity. *Nat Commun.* 2020;11(1):2052. doi 10.1038/s41467-020-15887-5

Zuo E., Sun Y., Wei W., Yuan T., Ying W., Sun H., Yuan L., Steinmetz L.M., Li Y., Yang H. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science*. 2019;364(6437):289-292. doi 10.1126/science.aav9973

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

Received June 14, 2025. Revised November 5, 2025. Accepted November 11, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):85-93

doi 10.18699/vjgb-26-02

The computational analysis of tumor cell sensitivity to supertarget deletion

D.A. Chetverina [1], N.Y. Kozelchuk [1], D.V. Lomaev [1], A.A. Shtil [2], M.M. Erokhin [1]

[1] Institute of Gene Biology Russian Academy of Sciences, Moscow, Russia
[2] N.N. Blokhin National Medical Research Center of Oncology, Moscow, Russia
yermaxbio@yandex.ru

Abstract. Gene mutations and altered epigenetic regulation of gene expression are characteristic features of malignant neoplasms. Combinations of these abnormalities form molecular features of individual tumors. In the large-scale Dependency Map (DepMap) project, the broad panels of human tumor cell lines are being tested for sensitivity to single gene inactivation. Using DepMap data, we have previously identified a set of genes termed supertargets, the deletion of which significantly reduced the survival of cells of a particular tissue origin while minimally impairing the unrelated cell lines. In the present study, we determined the factors of viability (inhibition of proliferation or death) of cell lines in which the supertarget genes have been deleted. We found that, in 79 % of cases, the reduced survival may be caused by epigenetic changes of gene expression. In the remaining 21 % of cases, it is associated with altered gene structure. Three groups containing different types of gene expression alterations can be distinguished. In the first group, the reduced cell survival correlated with a higher expression of the supertarget gene (e.g., *SOX10* and *HNF1B*). In the second group, a gene different from the deleted supertarget was overexpressed (gene pairs: *FOXA1* and *SPDEF*, *TP63* and *SERPINB13*, etc.). The third group was characterized by correlations between low expression of a certain gene and tumor cell sensitivity (e.g., *FAM126A* and *FAM126B*, *SMARCA2* and *SMARCA4*). The genetic changes included GOF mutations (*KRAS*, *BRAF* genes, etc.), LOF mutations (*STAG1*, *SMARCA2* genes, etc.), gene fusions (*BCR-ABL1*, *PAX3-FOXO1*, etc.), and amplification (*CPM*, *BEST3*, etc.). Therefore, many different molecular mechanisms act as predictors of tumor cell response to inhibition of supertarget genes.

Key words: tumors; cancer; oncomarkers; Dependency Map; DepMap; transcription; mutations; *SMARCA2*; *SMARCA4*; *APP*; *FOXA1*; *ATP6V0A2*; *ATP6V0A1*; bioinformatics; database analysis; personalized medicine

For citation: Chetverina D.A., Kozelchuk N.Y., Lomaev D.V., Shtil A.A., Erokhin M.M. The computational analysis of tumor cell sensitivity to supertarget deletion. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):85-93. doi 10.18699/vjgb-26-02

Funding. The study was supported by Russian Science Foundation grant No. 20-74-10099.

Acknowledgements. This study was performed using the equipment of the Institute of Gene Biology RAS facilities supported by the Ministry of Science and Higher Education of the Russian Federation.

Биоинформатический анализ механизмов жизнеспособности линий опухолевых клеток при делеции генов-супермишеней

Д.А. Четверина [1], Н.Я. Козельчук [1], Д.В. Ломаев [1], А.А. Штиль [2], М.М. Ерохин [1]

[1] Институт биологии гена Российской академии наук, Москва, Россия
[2] Национальный медицинский исследовательский центр онкологии им. Н.Н. Блохина, Москва, Россия
yermaxbio@yandex.ru

Аннотация. Мутации генов и изменения эпигенетической регуляции экспрессии генов являются характерными признаками злокачественных новообразований. Сочетания данных нарушений формируют биологические особенности индивидуальных опухолей на молекулярном уровне. Разработка стратегий персонализированного лечения требует глубокого понимания молекулярных «портретов» отдельных опухолей. В рамках крупномасштабного проекта Dependency Map (DepMap) обширные панели линий опухолевых клеток человека тестируются на чувствительность к инактивации отдельных генов. Ранее на основе данных DepMap нами охарактеризованы гены, получившие название «супермишени», при делеции которых существенно снижена жизнеспособность клеток конкретного тканевого происхождения при минимальном нарушении жизнеспособности других линий. В настоящем исследовании определены факторы низкой жизнеспособности (ингибирования пролиферации или гибели) клеточных линий при инактивации генов-супермишеней. В результате установлено, что в 79 % случаев низкая жизнеспособность может быть вызвана эпигенетическими изменениями в экспрессии генов. В остальных случаях (21 %) она вызвана нарушениями структуры генов. Исходя из полученных данных, можно выделить три группы, содержащие разного типа нарушения экспрессии генов.

© Chetverina D.A., Kozelchuk N.Y., Lomaev D.V., Shtil A.A., Erokhin M.M., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

В первой группе низкая жизнеспособность клеток коррелирует с повышением экспрессии гена-супермишени (например, *SOX10* и *HNF1B*). Во второй группе детектируется гиперэкспрессия гена, отличного от делетируемой супермишени (пары генов *FOXA1* и *SPDEF*, *TP63* и *SERPINB13* и др.). Третья группа характеризуется корреляциями между пониженной экспрессией определенных генов и чувствительностью опухолевых клеток (пары генов *FAM126A* и *FAM126B*, *SMARCA2* и *SMARCA4* и др.). Наблюдаемые генетические изменения включали GOF-мутации (*KRAS*, *BRAF* и др.), LOF-мутации (*STAG1*, *SMARCA2* и др.), слияние генов (*BCR-ABL1*, *PAX3-FOXO1* и др.) и амплификации (*CPM*, *BEST3* и др.). Таким образом, разные молекулярные механизмы выступают предикторами ответа опухолевых клеток на ингибирование генов-супермишеней.
Ключевые слова: опухоли; рак; онкомаркеры; Dependency Map; DepMap; транскрипция; мутации; *SMARCA2*; *SMARCA4*; *APP*; *FOXA1*; *ATP6V0A2*; *ATP6V0A1*; биоинформатика; анализ баз данных; персонализированная медицина

Introduction

The current approach to antitumor therapy involves identification of molecular mechanisms specific to a particular tumor type. The principle of targeted therapy is the inactivation of these factors to achieve an antiproliferative effect and/or cell death with minimal damage to non-malignant counterparts (Verma, 2012; Pfohl et al., 2021). The search for tumor-specific targets includes the screening of cell lines of various tissue origin on broad panels. The most large-scale project is the Dependency Map (DepMap, https://depmap.org/), which analyzes inactivation of an individual gene via RNAi and CRISPR/Cas9 technologies (Tsherniak et al., 2017; Arafeh et al., 2025).

Previously, using the DepMap resource, we studied 27 histological types of tumors; for each type, five genes were identified, the CRISPR/Cas9-mediated knockout of which reduced the viability (i. e. inhibition of proliferation or death) of cells of a particular type (Chetverina et al., 2023). These genes, termed "supertargets", can be considered promising candidates for personalized therapy.

In this study, we used the DepMap resource to dissect the genetic and epigenetic changes that correlate with reduced cell viability upon supertarget deletion. The identified factors can be used to predict the sensitivity of cells of an individual tumor type to inactivation of a particular target.

Materials and methods

Analysis of tumor cell viability was performed using the DepMap database (https://depmap.org/portal/, Tsherniak et al., 2017), Public 22Q4+Score, Chronos release (https://doi.org/10.25452/figshare.plus.27993248.v1). Viability was analyzed using the "Gene effect" value. To attribute low viability to a factor, two criteria were used: the value of the significance parameter (Importance) >10 %, and the degree of reliability (Student's *t*-test) <0.01. Gene expression levels were analyzed using DepMap Expression Public Release 22Q4 (https://depmap.org/portal/data_page/?tab=allData&releasename=DepMap%20Public%2024Q4&filename=OmicsExpressionProteinCodingGenesTPMLogp1.csv) The gene copy number was determined using the DepMap Copy Number Public version 22Q4 (https://depmap.org/portal/data_page/?tab=allData&releasename=DepMap%20Public%2024Q4&filename=OmicsCNGene.csv).

The STRING database (https://string-db.org/, version 12.0) was used to analyze protein-protein interactions. Only interactions confirmed by biochemical methods were taken into account.

Results and discussion

Low viability of tumor cell lines upon deletion of genes encoding supertargets

Previously, we analyzed 27 histological tumor types to identify the supertarget genes, that is, the ones the knockout of which most specifically affected the viability of a particular tumor compared with other cancers. For each tumor type, the top five genes critical for survival were identified, yielding a total of 124 unique supertargets (nine genes were repeated twice in different tumors and one was found thrice) (Chetverina et al., 2023). To identify genetic and/or epigenetic changes that correlate with low cell viability upon knockout of an individual supertarget, we used the Importance parameter set by the DepMap project. The search for correlation factors included data on all cell lines in the database without grouping. The RNAseq data and genetic alterations in intact cell lines were compared with gene effect in cell lines with inactivated supertarget genes. The search was performed for each supertarget individually.

A total of 167 associated correlations were found to link low viability of tumor cells (i. e. inhibited proliferation and/or death) with the knockout of supertarget genes. For a number of genes, two or more predictors of low viability were discovered whereas no reliable correlations were detectable for 23 out of 124 supertargets. Among the factors of low viability, the majority (132 cases) were associated with changes in gene expression, and 35 cases with altered gene structure.

Gene expression abnormalities that correlate with low cell viability upon knockout of supertarget genes

Overexpression of the supertarget. Detailed analysis revealed three groups of gene expression abnormalities that correlate with low cell viability upon supertarget knockout. In the first group (50 cases), low cell viability correlated with high expression of the supertarget (Table 1).

Figure 1 shows two typical examples. The *SOX10* gene (SRY-box transcription factor 10, Fig. 1*a*) has been previously identified as a supertarget in melanoma-derived cell lines. It encodes a transcription factor important for cell differentiation in the central and peripheral nervous system, melanocytes, and chondrocytes. Dysregulation of *SOX10* is known to be associated with carcinogenesis (Bahmad et al., 2023).

The *HNF1B* gene (HNF1 homeobox B, Fig. 1*b*) has been identified as a supertarget for renal cancer cell lines. This gene encodes a homeobox-containing transcription factor

Table 1. Low cell viability upon supertarget knockout correlating with a high level of expression of the same supertarget

No.	Supertarget gene	Importance, %	Pearson	*p*-value	No.	Supertarget gene	Importance, %	Pearson	*p*-value
1	*SOX10*	81.8	–0.84	2.99E-268	26	*SOX17*	39.6	–0.446	2.90E-50
2	*IRF4*	78.9	–0.813	7.16E-238	27	*BATF3*	38.7	–0.347	9.19E-30
3	*POU2AF1*	73.3	–0.67	7.79E-132	28	*DUSP4*	38.3	–0.416	2.64E-43
4	*HNF1B*	72.7	–0.751	1.76E-182	29	*PHOX2A*	37.3	–0.455	2.36E-52
5	*PAX8*	71.9	–0.677	1.38E-135	30	*MYOG*	37.0	–0.554	1.03E-81
6	*MYB*	71.4	–0.671	5.92E-132	31	*TP63*	36.4	–0.702	1.19E-149
7	*PAX5*	67.6	–0.645	3.21E-119	32	*LEF1*	35.7	–0.278	3.05E-19
8	*SPDEF*	66.7	–0.446	4.28E-50	33	*MECOM*	34.5	–0.363	1.38E-32
9	*MYCN*	62.9	–0.647	5.57E-120	34	*FLI1*	33.9	–0.312	4.86E-24
10	*MEF2C*	62.1	–0.464	1.27E-54	35	*SPI1*	31.8	–0.523	2.14E-71
11	*FOXA1*	61.0	–0.433	3.48E-47	36	*CCNE1*	31.5	–0.451	2.29E-51
12	*ESR1*	59.0	–0.481	3.55E-59	37	*PPARG*	31.3	–0.385	1.05E-36
13	*SATB2*	58.5	–0.278	3.06E-19	38	*PRDM1*	29.4	–0.286	2.27E-20
14	*WT1*	52.4	–0.209	2.34E-11	39	*ISL1*	27.5	–0.456	1.42E-52
15	*GATA3*	52.1	–0.522	3.84E-71	40	*HAND2*	27.4	–0.469	6.63E-56
16	*KLF5*	51.8	–0.609	8.22E-103	41	*ZBTB18*	24.5	–0.247	2.16E-15
17	*MITF*	51.4	–0.472	9.45E-57	42	*KLB*	20.8	–0.383	1.80E-36
18	*MYOD1*	51.3	–0.711	3.27E-155	43	*TEAD3*	20.2	–0.355	3.67E-31
19	*GFI1*	49.2	–0.423	8.75E-45	44	*NFIA*	15.1	–0.155	8.37E-07
20	*TRPS1*	48.0	–0.458	4.83E-53	45	*DOCK5*	14.9	–0.312	4.96E-24
21	*BCL6*	44.7	–0.383	2.03E-36	46	*PAX3*	13.2	–0.367	2.26E-33
22	*NKX2-1*	41.9	–0.391	5.39E-38	47	*PARD3*	12.3	–0.243	6.06E-15
23	*PHOX2B*	41.4	–0.458	4.49E-53	48	*SOX4*	12.3	–0.163	2.09E-07
24	*IGF2BP1*	40.3	–0.348	5.62E-30	49	*CDX2*	11.8	–0.399	1.09E-39
25	*IKZF1*	40.2	–0.583	1.58E-92	50	*IRS1*	10.0	–0.292	3.81E-21

important for the liver, kidney, and pancreatic development during embryogenesis (Chandra et al., 2021). Dysfunctions of *HNF1B*, including germline and somatic mutations, have been identified in a variety of solid tumors (Bártů et al., 2018).

Overexpression of a gene different from the inactivated supertarget. The second group of correlations contains 71 cases in which low cell viability upon supertarget deletion correlates with increased expression of a gene unrelated to the knocked-out supertarget (Table 2).

In 65 cases, correlations were found between low cell viability upon knockout of a supertarget gene and elevated expression of an unrelated gene. An example of such dependence is the *TP63* and *SERPINB13* pair (Fig. 2*a*). The *TP63* gene, which we assigned as a supertarget for upper aerodigestive cancer cell lines (Chetverina et al., 2023), encodes the p63 protein, a transcriptional regulator, like its homologs p53 and p73, involved in numerous processes in tumor biology (Sadu Murari et al., 2025). The *SERPINB13* gene encodes a serine protease antagonist that inhibits the activity of cathepsins K and L (Jayakumar et al., 2003; Welss et al., 2003). Recent studies have shown that *SERPINB13* may act as an oncogene in squamous cell lung cancer (Zhang et al., 2025). However, to our knowledge, there are no data on functional interactions between p63 and SERPINB13.

In six cases, we observed a correlation between low cell viability upon supertarget knockout and overexpression of a gene classified as a supertarget in the previous screening: the *FOXA1–SPDEF*, *MYOD1–MYOG*, *TCF7L2–HNF4A*, *RXRA–PPARG*, *HAND2–PHOX2A*, *SMARCAL1–MYOD1* gene pairs. For example, the *FOXA1* (forkhead box A1) and *SPDEF* (SAM pointed domain containing the ETS transcription factor, Fig. 2*b*) genes have been identified as supertargets in breast cancer cell lines (Chetverina et al., 2023). The FOXA1 protein activates the *SPDEF* gene, while SPDEF activates *FOXA1* transcription (Buchwalter et al., 2013; Paranjapye et al., 2020). Although there are no data on direct interactions between FOXA1 and SPDEF proteins, their genes are located

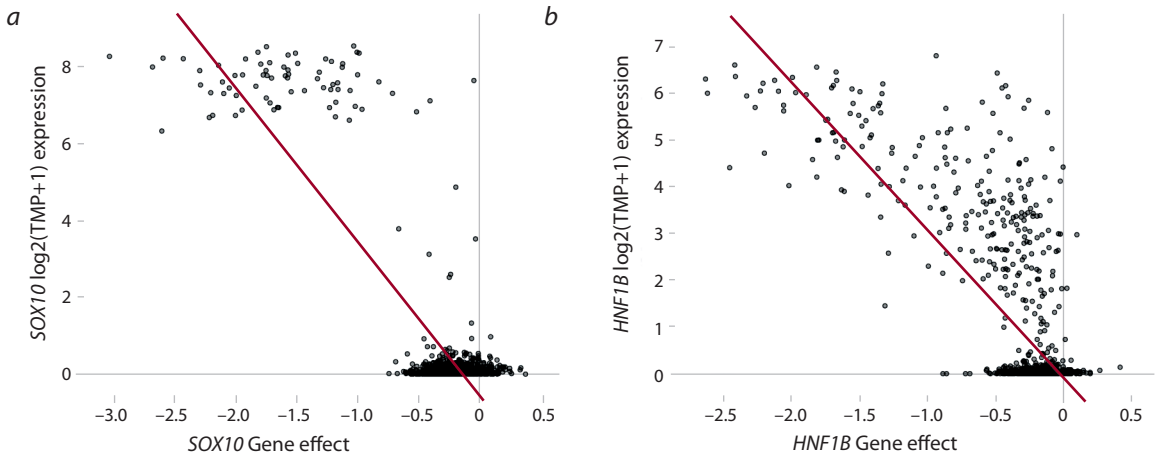


Fig. 1. Overexpression of the supertarget gene and low viability of tumor cells.

Correlations between gene expression and low cell viability upon knockout of the *SOX10* (*a*) and *HNF1B* (*b*) genes. The Y axis shows the expression level of the corresponding gene (log2(TPM+1)), the X axis shows the estimated Gene effect value reflecting cell viability upon knockout of the supertarget gene. The smaller the Gene effect value, the lower the viability and the higher the inhibition of proliferation and/or cell death of a given cell line upon inactivation of the corresponding gene. Here and in Tables 2–4: black dots show values for individual cell lines. The diagonal line is the regression.

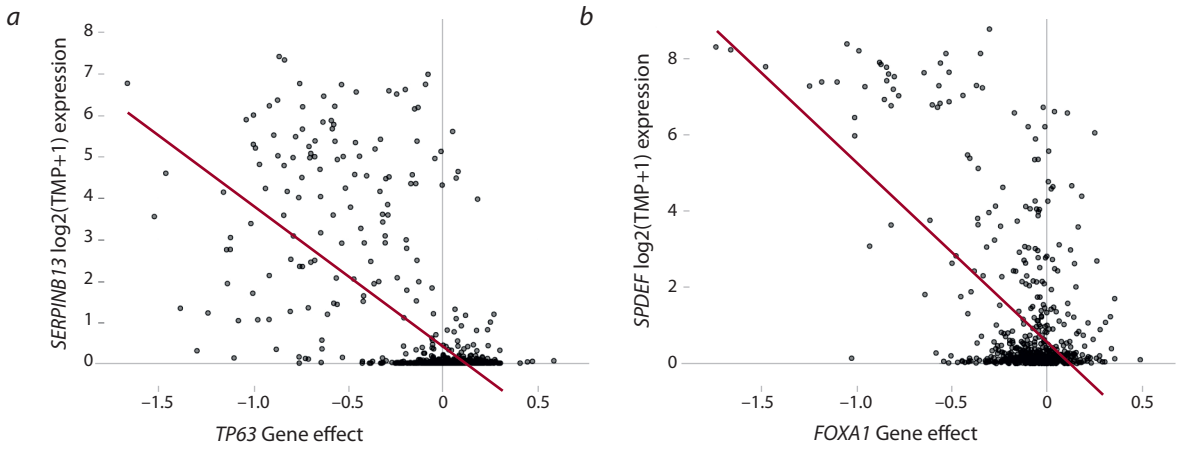


Fig. 2. Cell viability correlates with expression of genes unrelated to the inactivated supertarget.

a – correlation between *SERPINB13* expression and low cell viability upon *TP63* deletion; *b* – correlation between *SPDEF* expression and low cell viability upon *FOXA1* deletion.

in the same regulatory cluster, which also includes the genes for estrogen receptor (*ER*) and *GATA3* transcription factor, both important in mammary gland carcinogenesis (Lemieux et al., 2017).

The detected dependencies may be due to physical and functional interactions between proteins encoded by identified gene pairs. To address this possibility, we tested the presence of physical contacts between the proteins encoded by the identified genes using the STRING database. We found that 12 gene pairs encoded direct protein–protein interactors (Table 2). It can be assumed that, for other gene pairs, there could be an indirect effect, or else physical interactions between proteins have not yet been established.

Low expression of a non-supertarget gene. The third group of correlations contained 11 cases in which low cell viability upon supertarget knockout correlated with low expression of a non-supertarget gene. Most frequently (nine cases), the low-expressed non-supertarget gene was a close homologue of the supertarget (Table 3).

Typical examples are shown in Figure 3*a*, *b*. Low cell viability upon *FAM126B* knockout correlated with the low expression of its homologue gene *FAM126A* (Fig. 3*a*). The *FAM126A* and *FAM126B* genes (also known as *HYCC1* and *HYCC2*, stands for hyccin PI4KA lipid kinase complex subunit) encode the members of the PI4KIIIα/PI4KA protein kinase complex which regulates lipid composition and asymmetry of the plasma membrane (Suresh et al., 2024). Figure 3*b* demonstrates the correlation of low gene effect upon *SMARCA2* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 2) knockout with low expression of its homologue *SMARCA4*. Proteins encoded by *SMARCA2* and *SMARCA4* are the compo-

Table 2. Correlation of low cell viability upon supertarget knockout with overexpression of an unrelated gene

No.	Supertarget gene	Sensivity factor gene	Importance, %	Pearson	*p*-value	No.	Supertarget gene	Sensivity factor gene	Importance, %	Pearson	*p*-value
1	*EBF1*	*CD79A*	83.5	–0.76	1.57E-189	37	*SOS1*	*ITGA2B*	20.6	–0.333	2.37E-27
2	*EBF1*	*VPREB3*	39.8	–0.758	4.90E-188	38	*TRIM8*	*TPTE2*	20.6	–0.536	8.63E-76
3	*MEF2B*	*BORCS8-MEF2B*	73.8	–0.575	2.46E-89	39	*TRIM8*	*CYP4F22*	15.3	–0.501	7.02E-65
4	*MEF2B*	*ELL3*	39.2	–0.375	7.15E-35	40	*MECOM*	*WNT7A*	18.6	–0.355	3.26E-31
5	*TP63*	*SERPINB13*	61.1	–0.659	8.31E-126	41	*CRTC2*	*AVPI1*	18.5	–0.284	4.89E-20
6	*FOXA1*	*SPDEF*	57.6	–0.596	1.86E-97	42	*CRTC2*	*SIK1*	16.3	–0.194	5.41E-10
7	*CTNNB1**	*AXIN2**	55.9	–0.619	2.43E-107	43	*BCL6*	*ELL3*	18.3	–0.221	1.31E-12
8	*PAX5*	*CD19*	45.4	–0.701	5.36E-149	44	*EPAS1*	*SEC14L6*	18.3	–0.27	3.30E-18
9	*PAX5*	*VPREB3*	14.9	–0.706	2.08E-152	45	*EPAS1**	*EGLN3**	11.8	–0.123	8.80E-05
10	*PAX5*	*CD79A*	11.8	–0.682	6.22E-138	46	*RUNX1**	*TAL1**	17.5	–0.435	1.87E-47
11	*TCF7L2**	*RNF43**	42.6	–0.5	1.40E-64	47	*RUNX1*	*GMFG*	10.9	–0.497	9.38E-64
12	*TCF7L2*	*HNF4A*	14.6	–0.604	8.80E-101	48	*HAND2*	*PHOX2A*	15.7	–0.509	3.09E-67
13	*PAX3*	*SOX8*	40.4	–0.382	3.72E-36	49	*POU2AF1*	*JCHAIN*	14.7	–0.676	9.14E-135
14	*NFE2L2*	*AKR1C1*	38.9	–0.545	1.27E-78	50	*ISL1*	*STMN2*	14.1	–0.539	8.01E-77
15	*NFE2L2**	*MAFG**	16.2	–0.267	7.06E-18	51	*EGFR*	*ITGB4*	13.9	–0.491	5.21E-62
16	*NFE2L2*	*NQO1*	16.1	–0.304	7.02E-23	52	*EGFR**	*AREG**	12.3	–0.504	1.01E-65
17	*LMO2*	*LYL1*	36.1	–0.496	2.06E-63	53	*EGFR*	*C1orf116*	10.5	–0.53	9.47E-74
18	*LMO2*	*NFE2*	30.0	–0.525	5.43E-72	54	*EGFR*	*GJB3*	10.5	–0.53	8.17E-74
19	*LMO2*	*SIGLEC8*	11.1	–0.599	1.14E-98	55	*HS2ST1**	*CKAP4**	13.7	–0.355	4.24E-31
20	*LMO2*	*GATA1*	10.7	–0.612	4.03E-104	56	*PARD3**	*PARD6B**	13.7	–0.347	9.26E-30
21	*PRDM1*	*TXNDC5*	35.2	–0.281	1.12E-19	57	*PARD3*	*KCNJ16*	13.0	–0.44	9.86E-49
22	*CDX2*	*CDX1*	34.3	–0.495	2.98E-63	58	*DUSP4*	*MIA*	13.0	–0.509	3.01E-67
23	*MAPK1*	*S100B*	31.5	–0.489	2.32E-61	59	*STAT5B*	*PRSS57*	12.9	–0.453	5.73E-52
24	*MAPK1*	*CPN1*	16.8	–0.548	1.52E-79	60	*HMGA2*	*OS9*	12.8	–0.282	9.55E-20
25	*IL13RA1*	*IL5RA*	28.0	–0.345	2.30E-29	61	*EDF1*	*HCAR2*	12.4	–0.347	8.23E-30
26	*JAK2**	*CSF2RB**	26.8	–0.434	2.75E-47	62	*FZD5*	*FER1L6*	12.2	–0.32	3.01E-25
27	*JAK2**	*MPL**	20.8	–0.404	1.42E-40	63	*PRKAR1A**	*PRKACA**	12.2	–0.333	1.95E-27
28	*JUN*	*SERPINE1*	24.8	–0.502	5.32E-65	64	*PRKAR1A*	*TICAM2*	11.5	–0.45	2.94E-51
29	*JUN*	*PDCD1LG2*	12.8	–0.544	2.63E-78	65	*PRKAR1A*	*RTL3*	10.4	–0.351	2.13E-30
30	*MYOD1*	*MYOG*	23.7	–0.724	1.95E-163	66	*PHOX2A*	*SLC6A2*	11.7	–0.412	2.60E-42
31	*SPI1*	*TYROBP*	22.9	–0.509	2.52E-67	67	*MITF*	*MLANA*	11.6	–0.579	8.32E-91
32	*FGFR1*	*FGF2*	22.3	–0.432	8.11E-47	68	*SMARCAL1*	*MYOD1*	11.1	–0.298	5.12E-22
33	*ITGB1*	*GJB5*	22.3	–0.488	2.82E-61	69	*BATF3*	*MYBPC2*	10.8	–0.439	1.37E-48
34	*ITGB1*	*LAMA3*	11.1	–0.39	6.99E-38	70	*TRPS1*	*CREB3L4*	10.2	–0.303	8.35E-23
35	*RXRA**	*PPARG**	20.8	–0.371	3.87E-34	71	*RAB10*	*BICDL2*	10.0	–0.387	2.86E-37
36	*ETV6*	*PTGER3*	20.7	–0.411	4.64E-42						

Note. The "*" symbol marks pairs of proteins for which direct interactions have been shown by biochemical methods (STRING database). Groups corresponding to the same gene are shown in grey.

Table 3. Low viability upon supertarget knockout correlates with low expression of an unrelated gene

No.	Supertarget gene	Sensivity factor gene	Importance, %	Pearson	*p*-value	Homologues
1	*RPP25L*	*RPP25*	75.0	0.743	1.05E-176	Yes
2	*FAM126B*	*FAM126A*	71.6	0.408	1.64E-41	Yes
3	*ATP6V0E1*	*ATP6V0E2*	60.3	0.564	2.13E-85	Yes
4	*SMARCA2*	*SMARCA4*	55.0	0.39	9.62E-38	Yes
5	*ATP6V0A2*	*ATP6V0A1*	51.6	0.412	2.61E-42	Yes
6	*STAG1*	*STAG2*	28.9	0.253	3.61E-16	Yes
7	*SH3GL1*	*APP*	26.0	0.535	1.86E-75	No
8	*TIMM17A*	*TIMM17B*	24.7	0.399	1.04E-39	Yes
9	*ARHGEF7*	*ARHGEF6*	23.0	0.457	6.96E-53	Yes
10	*VRK1*	*VRK2*	12.2	0.357	1.66E-31	Yes
11	*STAT5B*	*NPHP1*	10.8	0.198	2.33E-10	No

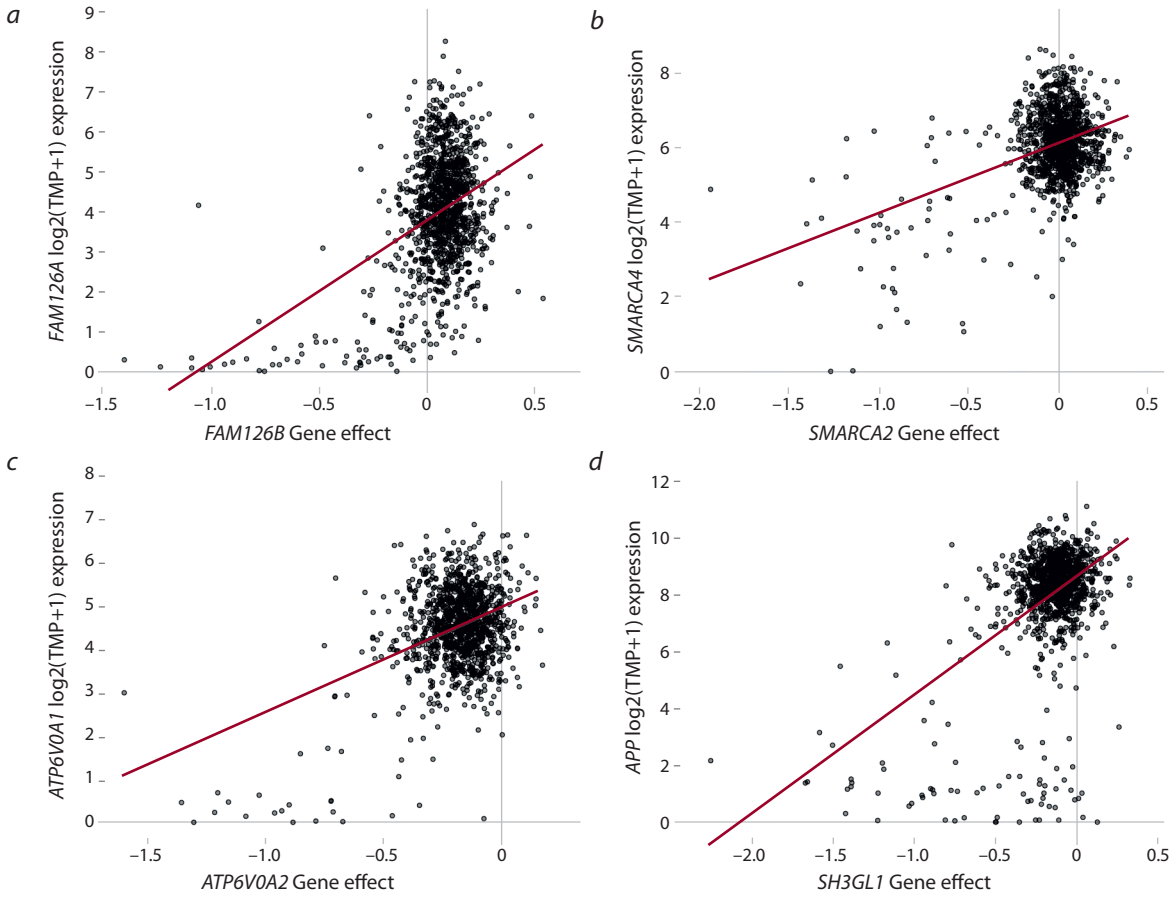


Fig. 3. Low expression of a non-supertarget gene correlates with poor viability upon supertarget knockout.

a – correlation of *FAM126A* expression and cell sensitivity to *FAM126B* knockout; *b* – correlation of *SMARCA4* expression and low viability upon *SMARCA2* knockout; *c* – correlation of *ATPV0A1* expression and sensitivity to *ATPV0A2* knockout; *d* – correlation of *APP* expression and low cell sensitivity to *SH3GL1* deletion.

nents of the SWI/SNF chromatin remodeling complex. Altered structure and function of this complex have been often found in tumors (Nguyen et al., 2023; Reddy et al., 2023).

In the case of the *ATP6V0A2* (ATPase H+ transporting V0 subunit a2) gene knockout, low cell viability correlated with low expression of its homologue *ATP6V0A1* (Fig. 3*c*). The *ATP6V0A2* and *ATP6V0A1* genes encode a component of the V-ATPase proton channel, which maintains an electrochemical proton gradient across the plasma membrane. In addition to its main function, V-ATPase is involved in the Notch/Wnt pathway (Eaton et al., 2021). Detailed functions of ATP6V0A1/2 remain to be established.

For *SH3GL1* (SH3 domain containing GRB2 like 1) knockout, low viability correlated with low expression of the *APP* (amyloid precursor protein) gene (Fig. 4*d*). The *SH3GL1* gene encodes endophilin A2 important for the dynamics of intracellular membranes, in particular, for endocytosis (Yang et al., 2023). The role of APP has been investigated in Alzheimer's disease: an APP fragment (β-amyloid) forms characteristic plaques in the brain (Chen et al., 2024). There are indications of the involvement of APP in carcinogenesis, but its role has not been established conclusively (Lee et al., 2021). No interactions between SH3GL1 and APP have been reported, nor was their mutual role in tumor biology studied. Thus, our analysis allows to predict tumor cell response to deletion of a particular gene as well as to reveal previously unknown functional interactions between the gene products.

Gene structure alterations. In 35 cases, correlations were observed between low cell viability upon supertarget knockout and altered gene structure. The types of abnormalities included point mutations (12 cases), gene fusions (13), and amplifications (10 cases) (Table 4).

Figure 4*a* shows the correlation for the *KRAS* (Kirsten rat sarcoma virus) gene, which encodes a small GTPase, a key component of the signaling pathway activated by the interaction of epidermal growth factor with its receptor (Seres et al., 2025). KRAS-activating mutations (gain-on-function, GOF) correlated with the sensitivity of cells to *KRAS* knockout.

The example in Figure 4*b* demonstrates the correlation between the low gene effect upon *SMARCA2* knockout and the number of inactivating (loss-of-function, LOF) mutations in its close homologue *SMARCA4*. As indicated above, *SMARCA2* and *SMARCA4* are the components of the SWI/SNF chromatin-remodeling complex. Increased cell sensitivity to *SMARCA2* knockout correlated with LOF mutations in the *SMARCA4* gene (Fig. 4*b*) and with low expression of *SMARCA4* (Fig. 3*b*).

Table 4. Changes in the structure of supertarget and other genes that correlate with low cell viability upon deletion of the supertarget

No.	Supertarget gene	Sensivity factor gene	Importance, %	Pearson	*p*-value	No.	Supertarget gene	Sensivity factor gene	Importance, %	Pearson	*p*-value
GOF mutation						LOF mutation					
1	*KRAS*	*KRAS*	69.3	–0.680	3.91E-146	1	*STAG1*	*STAG2*	32.8	–0.425	2.65E-48
2	*BRAF*	*BRAF*	68.2	–0.689	7.68E-152	2	*SMARCA2*	*SMARCA4*	31.9	–0.399	3.33E-42
3	*MAPK1*	*BRAF*	15.7	–0.438	2.06E-51	3	*CTNNB1*	*APC*	24.2	–0.528	6.34E-78
4	*CTNNB1*	*CTNNB1*	13.8	–0.288	6.57E-22	4	*WRN*	*KMT2B*	19.4	–0.517	2.50E-74
5	*DOCK5*	*KRAS*	13.1	–0.331	9.41E-29	5	*WRN*	*RPL22*	12.2	–0.479	2.02E-62
Fusion						6	*FAM126B*	*RNF43*	14.8	–0.231	1.78E-14
1	*BCR*	*BCR_ABL1*	44.8	–0.703	2.50E-150	7	*EPAS1*	*VHL*	10.5	–0.292	1.45E-22
2	*ABL1*	*BCR_ABL1*	42.6	–0.626	2.73E-110	Amplification					
3	*TRIM8*	*EWSR1_FLI1*	23.4	–0.546	5.04E-79	Negative correlation					
4	*NFE2L2*	*AKR1C1_AKR1C6P*	18.7	–0.328	1.59E-26	1	*CPM*	*CPM*	32.0	–0.376	1.30E-37
5	*PAX3*	*PAX3_FOXO1*	18.6	–0.662	1.99E-127	2	*BEST3*	*BEST3*	20.8	–0.175	6.97E-09
6	*MYOG*	*PAX3_FOXO1*	16.5	–0.638	1.00E-115	3	*MYCN*	*NBAS*	16.2	–0.42	2.28E-47
7	*FLI1*	*EWSR1_FLI1*	13.5	–0.548	1.24E-79	4	*NFIA*	*NFIA*	11.9	–0.213	1.74E-12
8	*MECOM*	*AC093821.1_LINC01091*	13.3	–0.278	2.86E-19	5	*CCNE1*	*CCNE1*	11.0	–0.299	9.61E-24
9	*DLL1*	*PAX3_FOXO1*	12.8	–0.41	4.85E-42	6	*KLB*	*UPF1*	10.9	–0.093	2.26E-03
10	*ZBTB18*	*PAX3_FOXO1*	11.2	–0.482	1.60E-59	7	*MEF2B*	*KIR2DS4*	10.8	–0.089	3.60E-03
11	*CCNE1*	*AC093821.1_LINC01091*	10.8	–0.298	4.77E-22	8	*CAPS2*	*CAPS2*	10.7	–0.176	5.99E-09
12	*MYOD1*	*PAX3_FOXO1*	10.5	–0.659	8.23E-126	9	*HMGA2*	*HMGA2*	10.0	–0.229	3.01E-14
13	*JAK2*	*HBB_AC104389.5*	10.3	–0.365	6.35E-33	Positive correlation					
						1	*DDX5*	*DDX17*	13.3	0.338	3.28E-30

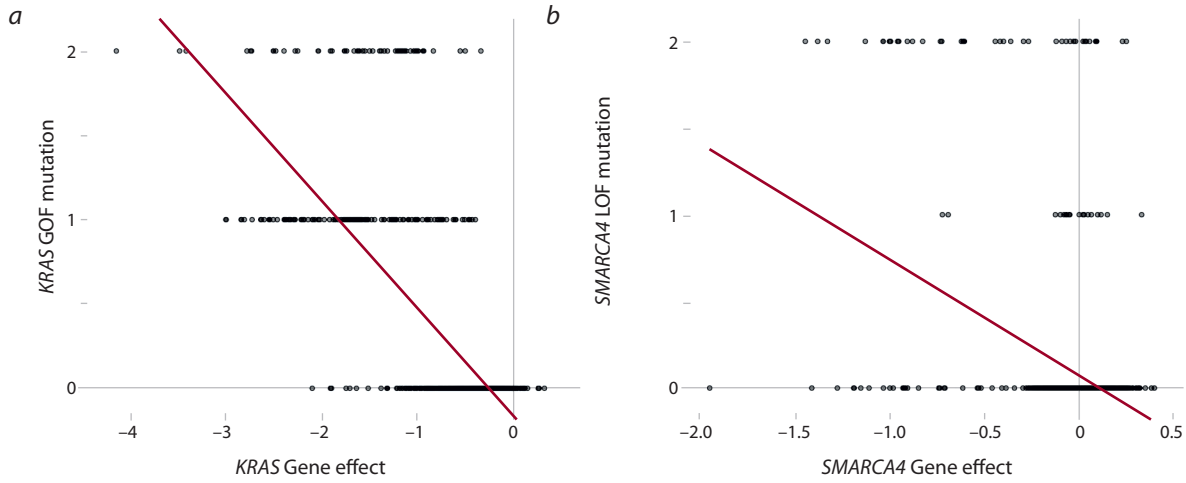


Fig. 4. Genetic mutations correlate with low tumor cell viability upon deletion of supertarget genes.

a – correlation between the number of the *KRAS* mutations and cell sensitivity to *KRAS* knockout; *b* – correlation between the number of *SMARCA4* mutations and cell sensitivity to *SMARCA2* knockout.

Conclusion

A detailed analysis of sensitivity of human tumor cell lines to knockout of supertarget genes was performed. We use the term "supertarget" for genes, the inactivation of which, according to the DepMap database, significantly reduces the viability of tumor cells of a particular tissue origin. Most frequently, low cell viability correlated with the expression of particular genes, i. e. supertargets as well as unrelated genes. Also, cell viability can be affected by genetic mutations such as GOF, LOF, gene fusion and amplification. Data on functional interdependences can be used to test the sensitivity of tumor cells of different origin to inactivation of supertarget genes by conventional and investigational drugs.

The established correlations are relevant to the development of personalized treatment strategies based on biological characteristics of the patient's tumor, that is, its molecular "portrait". Interpretation of tumor sensitivity to a specific drug presumes the identification of genetic as well as epigenetic mechanisms.

The therapeutic outcome (i. e. tumor eradication or growth retardation) is determined not as much by one factor but by their combinations. For a drug to be efficient, a coincidence of conditions must take place: cell death would be especially pronounced if inactivation of one gene is accompanied by the second mechanism ("synthetic lethality"). Increasing evidence points to the mutational "burden" of individual tumors, meaning the pairs of synthetic lethal genes (Du et al., 2023; Previtali et al., 2024). For tumors with a transcriptional "burden" (in particular, pediatric malignanices), genetic mutations are less important than epigenetic dysregulation (Comitani et al., 2023). Using *Drosophila* as a model, it was demonstrated that even temporary transcriptional disturbances can be carcinogenic without genetic alterations (Parreno et al., 2024).

Molecular correlations established in the present study determine cell fate upon inactivation of supertarget genes, thereby providing the mechanistic basis for rational drug combinations to treat "mutational" and "transcriptional" tumors.

References

Arafeh R., Shibue T., Dempster J.M., Hahn W.C., Vazquez F. The present and future of the Cancer Dependency Map. *Nat Rev Cancer*. 2025;25(1):59-73. doi 10.1038/s41568-024-00763-x

Bahmad H.F., Thiravialingam A., Sriganeshan K., Gonzalez J., Alvarez V., Ocejo S., Abreu A.R., Avellan R., Arzola A.H., Hachem S., Poppiti R. Clinical significance of *SOX10* expression in human pathology. *Curr Issues Mol Biol*. 2023;45(12):10131-10158. doi 10.3390/cimb45120633

Bártů M., Dundr P., Němejcová K., Tichá I., Hojný H., Hájková N. The role of HNF1B in tumorigenesis of solid tumours: a review of current knowledge. *Folia Biol* (*Praha*). 2018;64(3):71-83. doi 10.14712/fb2018064030071

Buchwalter G., Hickey M.M., Cromer A., Selfors L.M., Gunawardane R.N., Frishman J., Jeselsohn R., Lim E., Chi D., Fu X., Schiff R., Brown M., Brugge J.S. PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. *Cancer Cell*. 2013;23(6):753-767. doi 10.1016/j.ccr.2013.04.026

Chandra S., Srinivasan S., Batra J. Hepatocyte nuclear factor 1 beta: a perspective in cancer. *Cancer Med*. 2021;10(5):1791-1804. doi 10.1002/cam4.3676

Chen J., Chen J.-S., Li S., Zhang F., Deng J., Zeng L.-H., Tan J. Amyloid precursor protein: a regulatory hub in Alzheimer's disease. *Aging Dis*. 2024;15(1):201-225. doi 10.14336/AD.2023.0308

Chetverina D., Vorobyeva N.E., Gyorffy B., Shtil A.A., Erokhin M. Analyses of genes critical to tumor survival reveal potential 'supertargets': focus on transcription. *Cancers* (*Basel*). 2023;15(11):3042. doi 10.3390/cancers15113042

Comitani F., Nash J.O., Cohen-Gogo S., Chang A.I., Wen T.T., Maheshwari A., Goyal B., … Behjati S., Malkin D., Villani A., Irwin M.S., Shlien A. Diagnostic classification of childhood cancer using multiscale transcriptomics. *Nat Med*. 2023;29(3):656-666. doi 10.1038/s41591-023-02221-x

Du Y., Luo L., Xu X., Yang X., Yang X., Xiong S., Yu J., Liang T., Guo L. Unleashing the power of synthetic lethality: augmenting treatment efficacy through synergistic integration with chemotherapy drugs. *Pharmaceutics*. 2023;15(10):2433. doi 10.3390/pharmaceutics15102433

Eaton A.F., Merkulova M., Brown D. The H^+-ATPase (V-ATPase): from proton pump to signaling complex in health and disease. *Am J Physiol Cell Physiol*. 2021;320(3):C392-C414. doi 10.1152/ajpcell.00442.2020

Jayakumar A., Kang Y., Frederick M.J., Pak S.C., Henderson Y., Holton P.R., Mitsudo K., Silverman G.A., EL-Naggar A.K., Brömme D., Clayman G.L. Inhibition of the cysteine proteinases cathepsins K and L by the serpin headpin (SERPINB13): a kinetic analysis. *Arch Biochem Biophys*. 2003;409(2):367-374. doi 10.1016/s0003-9861(02)00635-5

Lee H.N., Jeong M.S., Jang S.B. Molecular characteristics of amyloid precursor protein (APP) and its effects in cancer. *Int J Mol Sci*. 2021;22(9):4999. doi 10.3390/ijms22094999

Lemieux S., Sargeant T., Laperrière D., Ismail H., Boucher G., Rozendaal M., Lavallée V.-P., Ashton-Beaucage D., Wilhelm B., Hébert J., Hilton D.J., Mader S., Sauvageau G. MiSTIC, an integrated platform for the analysis of heterogeneity in large tumour transcriptome datasets. *Nucleic Acids Res*. 2017;45(13):e122. doi 10.1093/nar/gkx338

Nguyen V.T., Tessema M., Weissman B.E. The SWI/SNF complex: a frequently mutated chromatin remodeling complex in cancer. In: Chen J., Wang G.G., Lu J. (Eds) Epigenetics in Oncology. *Cancer Treat Res*. Springer, Cham, 2023;190:211-244. doi 10.1007/978-3-031-45654-1_7

Paranjapye A., Mutolo M.J., Ebron J.S., Leir S.-H., Harris A. The FOXA1 transcriptional network coordinates key functions of primary human airway epithelial cells. *Am J Physiol Lung Cell Mol Physiol*. 2020;319(1):L126-L136. doi 10.1152/ajplung.00023.2020

Parreno V., Loubiere V., Schuettengruber B., Fritsch L., Rawal C.C., Erokhin M., Győrffy B., … Butova N.L., Chiolo I., Chetverina D., Martinez A.-M., Cavalli G. Transient loss of Polycomb components induces an epigenetic cancer fate. *Nature*. 2024;629(8012):688-696. doi 10.1038/s41586-024-07328-w

Pfohl U., Pflaume A., Regenbrecht M., Finkler S., Graf Adelmann Q., Reinhard C., Regenbrecht C.R.A., Wedeken L. Precision oncology beyond genomics: the future is here – it is just not evenly distributed. *Cells*. 2021;10(4):928. doi 10.3390/cells10040928

Previtali V., Bagnolini G., Ciamarone A., Ferrandi G., Rinaldi F., Myers S.H., Roberti M., Cavalli A. New horizons of synthetic lethality in cancer: current development and future perspectives. *J Med Chem*. 2024;67(14):11488-11521. doi 10.1021/acs.jmedchem.4c00113

Reddy D., Bhattacharya S., Workman J.L. (mis)-Targeting of SWI/SNF complex(es) in cancer. *Cancer Metastasis Rev*. 2023;42(2):455-470. doi 10.1007/s10555-023-10102-5

Sadu Murari L.S., Kunkel S., Shetty A., Bents A., Bhandary A., Rivera-Mulia J.C. p63: a master regulator at the crossroads between development, senescence, aging, and cancer. *Cells*. 2025;14(1):43. doi 10.3390/cells14010043

Seres M., Spacayova K., Sulova Z., Spaldova J., Breier A., Pavlikova L. Dynamic multilevel regulation of EGFR, KRAS, and MYC oncogenes: driving cancer cell proliferation through (epi)genetic and post-transcriptional/translational pathways. *Cancers* (*Basel*). 2025;17(2):248. doi 10.3390/cancers17020248

Suresh S., Shaw A.L., Pemberton J.G., Scott M.K., Harris N.J., Parson M.A.H., Jenkins M.L., Rohilla P., Alvarez-Prats A., Balla T., Yip C.K., Burke J.E. Molecular basis for plasma membrane recruitment of PI4KA by EFR3. *Sci Adv*. 2024;10(51):eadp6660. doi 10.1126/sciadv.adp6660

Tsherniak A., Vazquez F., Montgomery P.G., Weir B.A., Kryukov G., Cowley G.S., Gill S., … Garraway L.A., Root D.E., Golub T.R., Boehm J.S., Hahn W.C. Defining a cancer dependency map. *Cell*. 2017;170(3):564-576.e16. doi 10.1016/j.cell.2017.06.010

Verma M. Personalized medicine and cancer. *J Pers Med*. 2012;2(1):1-14. doi 10.3390/jpm2010001

Welss T., Sun J., Irving J.A., Blum R., Smith A.I., Whisstock J.C., Pike R.N., von Mikecz A., Ruzicka T., Bird P.I., Abts H.F. Hurpin is a selective inhibitor of lysosomal cathepsin L and protects keratinocytes from ultraviolet-induced apoptosis. *Biochemistry*. 2003;42(24):7381-7389. doi 10.1021/bi027307q

Yang L.-Q., Huang A.-F., Xu W.-D. Biology of endophilin and it's role in disease. *Front Immunol*. 2023;14:1297506. doi 10.3389/fimmu.2023.1297506

Zhang X., Zhang P., Ren Q., Li J., Lin H., Huang Y., Wang W. Integrative multi-omic and machine learning approach for prognostic stratification and therapeutic targeting in lung squamous cell carcinoma. *BioFactors*. 2025;51(1):e2128. doi 10.1002/biof.2128

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

Received February 27, 2025. Revised June 8, 2025. Accepted June 10, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):94-100

doi 10.18699/vjgb-26-07

The specific features of the thyroid hormone receptor gene *THRB* polymorphism in indigenous populations of Siberia

B.A. Malyarchuk, N.V. Pokhilyuk, G.A. Denisova, A.N. Litvinov

Institute of Biological Problems of the North of the Far Eastern Branch of the Russian Academy of Sciences, Magadan, Russia
malbor@mail.ru

Abstract. In the process of adaptation to cold in humans, genes belonging to the thyroid system signaling pathways that regulate thermogenesis, energy expenditure, and metabolic rearrangements are implicated. One such gene is the *THRB* gene, which encodes the nuclear receptor TRβ, with which the thyroid hormone triiodothyronine (T3) interacts. The activity of thermogenin UCP1 is influenced by the concentration of TRβ-T3 complexes, which serve to uncouple oxidative phosphorylation in mitochondria, thereby enhancing heat production. Consequently, thyroid hormone receptors have been demonstrated to play a significant role in adaptive thermogenesis. In the present study, we conducted a comprehensive analysis of published data on the *THRB* gene polymorphism in Siberian indigenous populations, with the objective of identifying potential associations between polymorphism variants and adaptation to cold. The analysis of exon and adjacent noncoding regions of the *THRB* gene revealed a single nucleotide substitution in the protein-coding region (synonymous substitution in the locus rs3752874). All other nucleotide substitutions were detected primarily in 3′-untranslated regions and introns. Analysis of the *THRB* haplotype distribution revealed two Koryak-specific haplotypes characterized by the rs762175401-A substitution. The results of population screening demonstrated that this substitution is prevalent among the Koryak population, with a frequency of 13.8 %, and is also present in the Siberian Eskimo population. However, in other global populations, the frequency of the rs762175401-A substitution does not exceed 0.05 % (in the Japanese and Koreans) or has even lower values (less than 0.02 %). The analysis of the nucleotide sequence of the *THRB* gene indicates that the rs762175401 locus is situated in the 3′-untranslated region at position +2 from the terminating codon. It is plausible that this substitution may have led to alterations in translation termination efficiency. In the case of enhanced termination efficiency, it is conceivable that it contributed to an elevated rate of protein synthesis, thereby resulting in an increase in the concentration of TRβ-T3 complexes. The higher frequency of the rs762175401-A variant in the Koryak and Eskimo populations, representing the oldest populations of Northeastern Siberia, is assumed to be due to long-term adaptation of these populations to cold.
Key words: *THRB* gene; human populations; Siberia; cold adaptation; thyroid system; adaptive thermogenesis

For citation: Malyarchuk B.A., Pokhilyuk N.V., Denisova G.A., Litvinov A.N. The specific features of the thyroid hormone receptor gene *THRB* polymorphism in indigenous populations of Siberia. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):94-100. doi 10.18699/vjgb-26-07

Особенности полиморфизма гена рецептора тиреоидных гормонов *THRB* у коренного населения Сибири

Б.А. Малярчук, Н.В. Похилюк, Г.А. Денисова, А.Н. Литвинов

Институт биологических проблем Севера Дальневосточного отделения Российской академии наук, Магадан, Россия
malbor@mail.ru

Аннотация. В процессах адаптации к холоду у человека задействованы гены, входящие в сигнальные пути тиреоидной системы, которыми регулируются термогенез, энергетические затраты и метаболические перестройки. Один из таких генов – *THRB*, кодирующий ядерный рецептор TRβ, с которым взаимодействует гормон щитовидной железы – трийодтиронин (T3). От концентрации комплексов TRβ-T3 зависит активность термогенина UCP1, с помощью которого происходит разобщение окислительного фосфорилирования в митохондриях и усиливается производство тепла. Таким образом, рецепторы тиреоидных гормонов играют важную роль в адаптивном термогенезе. В настоящей работе нами впервые проведен анализ опубликованных данных о полиморфизме гена *THRB* в популяциях коренного населения Сибири с целью поиска вариантов полиморфизма, потенциально связанных с адаптацией к холоду. Анализ полиморфизма экзонов и прилегающих некодирующих участков гена *THRB* показал наличие всего одной нуклеотидной замены в белок-кодирующей области (синонимичная замена в локусе rs3752874), все остальные нуклеотидные замены выявлены преимущественно в 3′-нетранслируемых участках и интронах. Анализ распределения гаплотипов гена *THRB*

© Malyarchuk B.A., Pokhilyuk N.V., Denisova G.A., Litvinov A.N., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

позволил обнаружить два специфичных для коряков гаплотипа, характеризующиеся заменой rs762175401-A. Популяционный скрининг показал, что эта замена распространена среди коряков с частотой 13.8 %, а также присутствует у сибирских эскимосов, хотя в остальных группах населения мира частота замены rs762175401-A не превышает 0.05 % (у японцев и корейцев) или имеет еще более низкие значения (менее 0.02 %). Анализ нуклеотидной последовательности гена *THRB* показывает, что локус rs762175401 находится в 3′-нетранслируемой области в позиции +2 от терминирующего кодона. Вполне вероятно, что эта замена могла привести к изменениям в эффективности терминации трансляции и в случае повышения эффективности терминации способствовала увеличению скорости синтеза белка и, соответственно, концентрации комплексов TRβ-T3. Предполагается, что повышение частоты варианта rs762175401-A у коряков и эскимосов, представляющих древнейшее население Северо-Востока Сибири, обусловлено долговременной адаптацией популяций к холоду.

Ключевые слова: ген *THRB*; популяции человека; Сибирь; адаптация к холоду; тиреоидная система; адаптивный термогенез

Introduction

Physiological studies have shown that the indigenous peoples of Siberia have an increased metabolic rate, especially in winter, and changes in thyroid hormone levels synchronous with the type of metabolism (Leonard, 2024). In Yakuts, the indigenous inhabitants of one of the coldest regions of the world, metabolic heat production increases by an average of 6 % in winter (Leonard et al., 2014). In winter, the body's tissues – mostly brown adipose tissue (BAT) – absorb thyroid hormones more actively. This increases heat production, leading to a significant drop in blood levels of triiodothyronine (T3) and thyroxine (T4) (Levy et al., 2013; Nikanorova et al., 2023). It is well established that thermogenin UCP1 (uncoupling protein-1) is expressed in BAT, where it facilitates uncoupling oxidative phosphorylation in mitochondria and heat release (Bianco, Silva, 1988). UCP1-mediated thermogenesis is activated by the interaction of thyroid hormones with the nuclear receptor TRβ. Higher concentrations of TRβ-T3 complexes result in higher UCP1 activity (Martinez de Mena et al., 2010; Lee et al., 2012; Yau, Yen, 2020; Ma et al., 2023). Thyroid hormone receptors clearly play an important role in the nonshivering thermogenesis associated with adaptation to cold.

It is well established that the distribution of polymorphism variants of the *UCP1*, *UCP2*, and *UCP3* uncoupling protein genes in human populations is associated with a number of natural and climatic factors, including geographic latitude, altitude, and the severity of natural conditions (Hancock et al., 2011; Nikanorova et al., 2021, 2022; Kozlov et al., 2024).

The polymorphism of the *THRB* gene encoding the nuclear receptor TRβ has been characterized in various genetic databases (dbSNP, https://www.ncbi.nlm.nih.gov/snp/). However, we did not find any special publications devoted to the analysis of the distribution of polymorphic variants of this gene in human populations. The main publications related to the polymorphism of the *THRB* gene focus on the search for genetic variants associated with thyroid hormone resistance syndrome (Dumitrescu, Refetoff, 2013), the risk of cancer (González-Sancho et al., 2003), and the regulation of transcription and chromatin remodeling (Grøntved et al., 2015).

The present work aims to characterize the polymorphism of the *THRB* gene in the indigenous populations of Siberia and to search for polymorphism variants associated with adaptation to cold.

Materials and methods

Data on whole exome-wide polymorphisms in indigenous populations of Northeastern Siberia (Eskimos, Chukchi, Koryaks; $N = 25$), Central Siberia (Evens, Evenks, Yakuts; $N = 29$), Southern Siberia (Tuvinians, Shorians, Altaians, Buryats; $N = 28$), and Western Siberia (Kets, Khanty, Mansi, Selkups, Nenets, Nganasans; $N = 20$), with the total number of 102 individuals, were used (Pagani et al., 2016). We analyzed the polymorphism of all exons and adjacent noncoding regions of the *THRB* gene located on chromosome 3 between positions 24158651 and 24536773. We used the ELB algorithm (Excoffier et al., 2003) implemented in the Arlequin 3.5 software package to identify haplotypes from genotypes with unknown gametic phase. Fisher's exact test was applied to assess the statistical significance of differences in the frequencies of polymorphic variants. The median network of *THRB* gene haplotypes was constructed using the Network 10.2 program (www.fluxus-engineering.com). This work used information from genomic databases in human populations for comparative analysis: dbSNP (www.ncbi.nlm.nih.gov/projects/SNP), 1000 Genomes (https://www.internationalgenome.org/), and gnomAD (https://gnomad.broadinstitute.org/).

We performed population screening of polymorphisms in loci rs762175401 (nucleotide position 3:24164373) and rs72619908 (position 3:24164268) of the *THRB* gene. Nucleotide numbers are given according to the human genome reference sequence GRCh37.p13 (hg19). The study used total DNA isolated from whole blood from representatives of the indigenous populations of the Severo-Evensky District of the Magadan Region. The samples include 98 Koryaks and 110 Evens. The questionnaire data show that the surveyed Koryaks and Evens have identified themselves as members of the above ethnic groups for at least two to three generations.

The nucleotide sequence of the *THRB* gene including loci rs762175401 and rs72619908 was amplified using oligonucleotide primers 5′-GCGCCATTTTGCTGACTCAA-3′ and 5′-TCTTCTCTCTTCCCCGCAGA-3′. Primers were designed based on the nucleotide sequence of the *THRB* gene (number NC_000003.12 in the GenBank database) using the Primer3 program (Untergasser et al., 2012).

Amplification products were sequenced using the BrilliantDye™ Terminator Cycle DNA Sequencing kit v3.1 (Netherlands) and an ABI Prism 3500xL genetic analyzer (Applied Biosystems, USA). MEGA5 package programs were used for

nucleotide sequence alignment and analysis (Tamura et al., 2011). We calculated the allele frequencies, heterozygosity, and correspondence of the genotype distribution to Hardy–Weinberg equilibrium using the Arlequin 3.5 software package (Excoffier, Lischer, 2010).

Results and discussion

The analysis of the *THRB* gene nucleotide sequences from 102 individuals belonging to different ethnic groups within the indigenous Siberian population revealed polymorphisms at 22 nucleotide positions (Table 1). However, only one polymorphism variant was detected in the exons: a synonymous substitution at locus rs3752874 (amino acid position 245). All other polymorphisms were found in the non-coding region, primarily in the 3′-untranslated regions and introns of the gene. Analysis of the distribution of polymorphic variants revealed that a similar distribution of allele frequencies was observed only in two cases (rs56204436-A and rs13326381-T) in Siberia, East Asia, and Europe ($p > 0.05$).

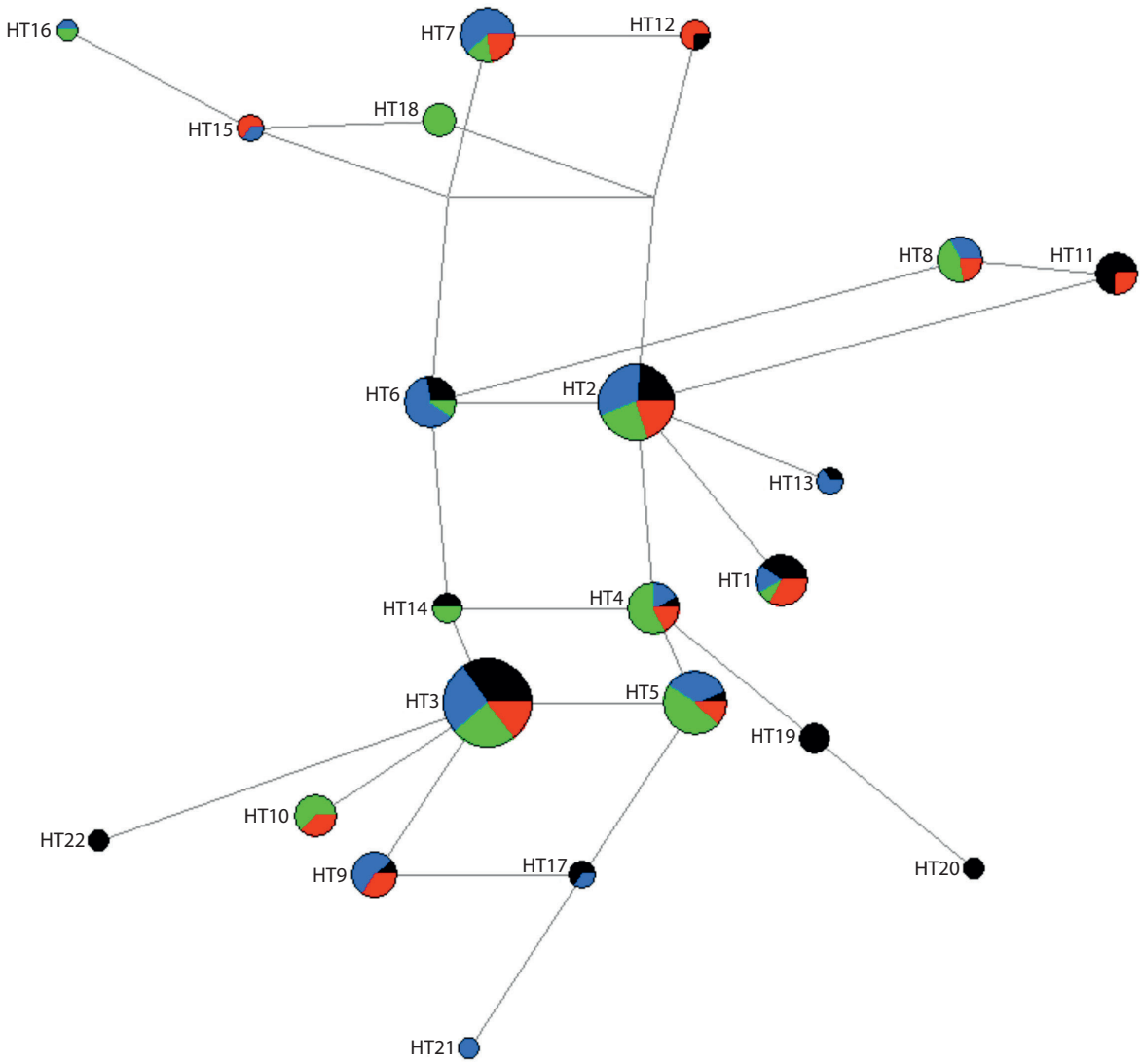
In most cases, there are statistically significant differences in the frequency of polymorphic variants in Siberia compared to Europe and East Asia. For several loci, the allele frequencies of Siberian populations are more similar to European than East Asian values (rs75272640, rs79270057, rs3752874, rs60502621, rs58274299, rs150604595, rs13090120). In three cases, polymorphisms characteristic of Siberian populations were detected in all four regions (rs561918607-T) or in particular populations (rs572372574-C in Chukchi and Evenki, and rs762175401-A in Eskimos and Koryaks) (Table 1).

To analyze the distribution of *THRB* gene haplotypes in Siberian populations, we narrowed down the set of polymorphic loci to 16: rs75272640, rs2167116, rs56204436, rs1349265,

Table 1. Frequencies of *THRB* gene polymorphism variants in indigenous populations of Siberia, East Asia and Europe

Polymorphism variant	Site	Northeastern Siberia (*N* = 25)	Central Siberia (*N* = 29)	Southern Siberia (*N* = 28)	Western Siberia (*N* = 20)	**Siberia** (*N* = 102)	**East Asia** (*N* = 1170)	*p* (Siberia-East Asia)	**Europe** (*N* = 1266)	*p* (Siberia-Europe)
rs3752874-A	syn	0.16	0.052	0.107	0.15	0.113	0.056	0.0031	0.149	0.181
rs34833017-G	intron	0	0.017	0.161	0.075	0.064	0.006	0	0.191	0.000001
rs13326381-T		0.6	0.397	0.5	0.525	0.5	0.546	**0.187**	0.472	**0.511**
rs60502621-C		0	0.017	0.018	0.025	0.015	0.101	0.000002	0.024	0.481
rs58274299-C		0	0.017	0.018	0.025	0.015	0.101	0.000002	0.024	0.481
rs150604595-A		0	0.017	0.018	0.025	0.015	0.101	0.000002	0.024	0.481
rs13090120-C		0.12	0.034	0.054	0.175	0.088	0.012	0	0.113	0.3
rs572372574-C	**5′-UTR**	0.02	0.069	0	0	0.025	nd	nd	nd	nd
rs75272640-T	3′-UTR	0.12	0.052	0.089	0.15	0.098	0.05	0.0085	0.137	0.134
rs2167116-A		0.06	0.19	0.143	0.225	0.152	0.092	0.0089	0.074	0.0004
rs56204436-A		0.04	0.052	0.107	0.075	0.069	0.046	**0.169**	0.057	**0.53**
rs1349265-G		0.54	0.431	0.607	0.375	0.495	0.697	0	0.694	0
rs79270057-T		0.12	0.052	0.089	0.15	0.098	0.05	0.0085	0.139	0.111
rs561918607-T		0.08	0.138	0.018	0.075	0.078	0	0	0.0003	0
rs844107-C		0.52	0.586	0.536	0.65	0.569	0.626	0.114	0.419	0.00005
rs826371-G		0.38	0.397	0.393	0.35	0.382	0.529	0.00007	0.262	0.0004
rs826372-C		0.38	0.397	0.393	0.35	0.382	0.529	0.00007	0.262	0.0004
rs826373-G		0.4	0.534	0.429	0.5	0.466	0.574	0.0032	0.257	0
rs826374-C		0.38	0.397	0.393	0.35	0.382	0.529	0.00007	0.262	0.0004
rs826375-C		0.4	0.534	0.429	0.5	0.466	0.574	0.0032	0.257	0
rs72619908-G		0.02	0.138	0.036	0.15	0.083	0.045	0.024	0	0
rs762175401-A		0.12	0	0	0	0.029	0	0	0.00003	0

Note. Data for East Asian and European populations are from the 1000 Genomes (https://www.internationalgenome.org/) and gnomAD (https://gnomad.broadinstitute.org/) databases. The statistical significance of differences (*p*) between the frequencies of polymorphism variants in the compared regions was evaluated using Fisher's exact test. Designations: 3′-UTR and 5′-UTR – 3′- and 5′-untranslated regions; syn – synonymous substitution in the protein-coding region; nd – no data. The polymorphism variants for which additional population screening was performed in the present study, as well as *p*-values greater than 0.05 in cases with a similar allele frequency distribution in Siberia, East Asia, and Europe (for rs56204436-A and rs13326381-T), are highlighted in bold.



Median network of *THRB* gene haplotypes in the indigenous Siberian populations.
Black indicates Northeastern Siberia, blue – Central Siberia, green – Southern Siberia, red – Western Siberia.

rs79270057, rs561918607, rs844107, rs826371, rs72619908, rs762175401, rs3752874, rs572372574, rs34833017, rs13326381, rs60502621, rs13090120. This was done because several loci located close to each other showed similar frequencies in Siberian, East Asian, and European samples and were linked. The ELB algorithm identified 34 16-locus haplotypes (Table S1 of the Supplementary Material)[1]. For the subsequent phylogenetic analysis of the haplotypes, we used the 22 haplotypes that were recorded more than once in the populations.

The median network of haplotypes obtained demonstrates rather complex phylogenetic relationships, likely due to recurrent mutations in the non-coding regions of the *THRB* gene or sequencing errors at the rs13326381 locus (see the Figure). The haplotypes under study were detected in different regional groups in Siberia; therefore, geographic clustering of haplotype groups is not apparent. Of particular interest are the HT19 and HT20 haplotypes found in samples from Northeastern Siberia. These haplotypes are characterized by the presence of a G→A substitution at locus rs762175401, which is located in the 3′-untranslated region of the gene. This variant polymorphism was only found in Eskimos (25 %) and Koryaks (12 %). According to the dbSNP and gnomAD data, the rs762175401-A allele was detected at very low frequencies in the Japanese (0.042 %) and Koreans (0.04 %), and in large samples from East Asia (0.013 %), the Middle East (0.016 %), South Asia (0.005 %), and Europe (0.002 %).

Since the Siberian sample size was insufficient, we investigated the polymorphism of the rs762175401 locus in a more representative sample of Koryaks and Evens from the Magadan region. We also included the rs72619908 locus, located 104 bp from rs762175401, in the investigated region of the *THRB* gene (Tables 2 and 3). According to historical data, the Koryaks (along with the Chukchi) belong to north-

[1] Supplementary Table S1 is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Mal_Engl_30_1.xlsx

Table 2. Genotype and allele frequencies of the rs762175401 locus of the *THRB* gene in the Koryaks and Evens

Population (*N*)	Genotypes			Alleles		H_e	*p*
	GG	GA	AA	G	A		
Koryaks (98)	0.765	0.194	0.041	0.862	0.138	0.239	0.08
Evens (110)	1.0	0	0	1.0	0	0	1.0

Note. Here and in Table 3: *N* – sample size, H_e – expected heterozygosity, *p* – statistical significance of deviation from Hardy–Weinberg equilibrium (significant at $p < 0.05$).

Table 3. Genotype and allele frequencies of the rs72619908 locus of the *THRB* gene in the Koryaks and Evens

Population (*N*)	Genotypes			Alleles		H_e	*p*
	CC	CG	GG	C	G		
Koryaks (96)	0.948	0.052	0	0.974	0.026	0.051	1.0
Evens (110)	0.927	0.064	0.009	0.959	0.041	0.079	0.156

eastern Paleoasians, the oldest population in Northeastern Siberia. The Tungus-speaking Even people began settling in Koryak areas around the 17th century (Khakhovskaya, 2024). However, despite living in the same neighborhood for a long time and intermarrying, these ethnic groups have retained features of their gene pools that are adapted to varying degrees to the extreme conditions of their natural environment (Cardona et al., 2014; Derenko et al., 2023; Malyarchuk, Derenko, 2024).

Analysis of the rs762175401 polymorphism showed that, of the studied samples, only Koryaks have the rs762175401-A allele, which occurs at a frequency of 13.8 % (Table 2). The polymorphic variants of the rs72619908 locus are distributed similarly in both ethnic groups (Table 3). The high frequency of the rs762175401-A variant in Koryaks, as well as its presence in Siberian Eskimos, suggests that this polymorphism has adaptive significance with respect to adaptive thermogenesis. The rs762175401 locus is located in the 3′-untranslated region of the *THRB* gene, at position +2 relative to the UAG termination codon. A substitution at this nucleotide position could potentially alter translation termination efficiency, as the 3′ context of stop codons has been shown to influence translation termination in eukaryotes (Cridge et al., 2018; Sokolova et al., 2020). If the rs762175401-A substitution contributes to increased termination efficiency, it could lead to a higher rate of protein synthesis, thus optimizing translation (increasing mRNA stability, ribosome recycling, and translation fidelity), as was previously found for *Escherichia coli* and yeast (Baggett et al., 2017; Wu et al., 2020).

Improved thyroid hormone receptor synthesis presumably contributes to an increased concentration of TRβ-T3 complexes and may therefore have adaptive significance in cold conditions. Thus, the increased frequency of the rs762175401-A variant in the Koryak and Eskimo populations, which are among the oldest in Northeastern Siberia, can be explained by the long-term influence of climatic factors on *THRB* gene function.

It is known that several genes associated with cold adaptation in humans, including *DIO2*, *UCP1*, *UCP3*, *THRB*, *PPARGC1A*, and *RXRA*, are involved in thyroid signaling pathways that regulate thermogenesis as well as energy expenditure and metabolic rearrangements (Laurberg et al., 2005; Bianco et al., 2019; Tsibulnikov et al., 2020). Genetic studies have revealed population genetic effects with respect to certain genes, such as *UCP1* and *UCP3*, which are manifested by an increased frequency of specific polymorphism variants in Northeast Asia (Nikanorova et al., 2021, 2022; Kozlov et al., 2024). Chronic exposure to cold in Arctic populations likely increases the activity of the DIO2 enzyme, which regulates T3 levels in cells. This increases the production of T3 to compensate for the high metabolic demands of thermogenesis (Noahsen et al., 2021).

Natural selection has likely affected the *ANGPTL8* and *PLA2G2A* genes in indigenous Siberian peoples, such as the Koryaks, Yukaghirs, and Nenets (Hallmark et al., 2019). These genes are involved in the adaptive response to cold; they are activated by hormone T3 and participate in regulating lipid metabolism (Sharma et al., 2014; Tseng et al., 2014). The effect of selection on the *THRB* gene has also been reported in the Yakut sample using the PBS test (Cardona et al., 2014). However, an analysis of the effect of selection in regional groups worldwide found statistical associations with thyroid system function, BAT, and thermoregulation only in Central Siberians and Africans using the EHH test (Pagani et al., 2016).

However, a recent genomic study of Greenland Eskimos did not reveal selection acting on thyroid system genes (Stæger et al., 2025). Polymorphism variants specific to Greenlandic and North American Eskimos as well as Siberian populations (*FADS1/2*, *SI*, *CPT1A*, *TBC1D*), or only to Greenlandic and American Eskimos (*LDLR*, *HNF1A*, *ADCY3*, *ATP8B1*, *PCCA/PCCB*) were found mainly in lipid and carbohydrate metabolism genes (Stæger et al., 2025).

These data suggest the necessity of further investigating thyroid system gene polymorphisms in indigenous populations of the Far North of various ethnical backgrounds, as adaptive changes in gene pools may exhibit population-specific characteristics.

Conclusion

Thus, our study provided the first estimate of the prevalence of different *THRB* gene polymorphisms that encode the thyroid hormone receptor in ethnic groups of the indigenous Siberian population. Our analysis of *THRB* gene haplotype distribution revealed that the rs762175401-A variant, observed in Koryaks (at a frequency of 13.8 %) and Siberian Eskimos (25 %), is prevalent in Northeastern Siberia. The increase in the frequency of this polymorphism is likely associated with the adaptation of indigenous Far North populations to cold and with rearrangements of the thyroid system, which is directly involved in nonshivering thermogenesis. Further molecular genetic studies will help clarify these mechanisms.

References

Baggett N.E., Zhang Y., Gross C.A. Global analysis of translation termination in *E. coli*. *PLoS Genet.* 2017;13(3):e1006676. doi 10.1371/journal.pgen.1006676

Bianco A.C., Silva J.E. Cold exposure rapidly induces virtual saturation of brown adipose tissue nuclear T3 receptors. *Am J Physiol.* 1988;255(4):E496-E503. doi 10.1152/ajpendo.1988.255.4.E496

Bianco A.C., Dumitrescu A., Gereben B., Ribeiro M.O., Fonseca T.L., Fernandes G.W., Bocco B.M.L.C. Paradigms of dynamic control of thyroid hormone signaling. *Endocr. Rev.* 2019;40(4):1000-1047. doi 10.1210/er.2018-00275

Cardona A., Pagani L., Antao T., Lawson D.J., Eichstaedt C.A., Yngvadottir B., Shwe M.T.T., … Willerslev E., Tyler-Smith C., Malyarchuk B.A., Derenko M.V., Kivisild T. Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS One*. 2014; 9(5):e98076. doi 10.1371/journal.pone.0098076

Cridge A.G., Crowe-McAuliffe C., Mathew S.F., Tate W.P. Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 2018; 46(4):1927-1944. doi 10.1093/nar/gkx1315

Derenko M., Denisova G., Litvinov A., Dambueva I., Malyarchuk B. Mitogenomics of the Koryaks and Evens of the northern coast of the Sea of Okhotsk. *J Hum Genet.* 2023;68(10):705-712. doi 10.1038/s10038-023-01173-x

Dumitrescu A.M., Refetoff S. The syndromes of reduced sensitivity to thyroid hormone. *Biochim Biophys Acta.* 2013;1830(7):3987-4003. doi 10.1016/j.bbagen.2012.08.005

Excoffier L., Lischer H.E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 2010;10(3):564-567. doi 10.1111/j.1755-0998.2010.02847.x

Excoffier L., Laval G., Balding D. Gametic phase estimation over large genomic regions using an adaptive window approach. *Hum Genomics*. 2003;1(1):7-19. doi 10.1186/1479-7364-1-1-7

González-Sancho J.M., García V., Bonilla F., Muñoz A. Thyroid hormone receptors/THR genes in human cancer. *Cancer Lett.* 2003; 192(2):121-132. doi 10.1016/S0304-3835(02)00614-6

Grøntved L., Waterfall J.J., Kim D.W., Baek S., Sung M.H., Zhao L., Park J.W., Nielsen R., Walker R.L., Zhu Y.J., Meltzer P.S., Hager G.L., Cheng S.Y. Transcriptional activation by the thyroid hormone receptor through ligand-dependent receptor recruitment and chromatin remodelling. *Nat Commun.* 2015;6:7048. doi 10.1038/ncomms8048

Hallmark B., Karafet T.M., Hsieh P., Osipova L.P., Watkins J.C., Hammer M.F. Genomic evidence of local adaptation to climate and diet in indigenous Siberians. *Mol Biol Evol.* 2019;36(2):315-327. doi 10.1093/molbev/msy211

Hancock A.M., Clark V.J., Qian Y., Di Rienzo A. Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Biol Evol.* 2011;28(1):601-614. doi 10.1093/molbev/msq228

Khakhovskaya L.N. Nomenclature of indigenous peoples of Magadan oblast in the historical context. *Bulletin of the North-East Scientific Center of FEB RAS*. 2024;2:108-115. doi 10.34078/1814-0998-2024-2-108-115 (in Russian)

Kozlov A.I., Vershubskaya G.G., Malyarchuk B.A., Nagornaya E.G., Parfenteva O.I., Balanovska E.V. Variability of *UCP1* and *UCP3* uncoupling protein genes in relation to climate in indigenous populations of Siberia and the Far East. *Lomonosov Journal of Anthropology* (*Moscow University Anthropology Bulletin*). *Series XXIII. Anthropology.* 2024;3:79-90. doi 10.55959/MSU2074-8132-24-3-7 (in Russian)

Laurberg P., Andersen S., Karmisholt J. Cold adaptation and thyroid hormone metabolism. *Horm Metab Res.* 2005;37(9):545-549. doi 10.1055/s-2005-870420

Lee J.-Y., Takahashi N., Yasubuchi M., Kim Y.-I., Hashizaki H., Kim M.-J., Sakamoto T., Goto T., Kawada T. Triiodothyronine induces UCP-1 expression and mitochondrial biogenesis in human adipocytes. *Am J Physiol Cell Physiol.* 2012;302(2):C463-C472. doi 10.1152/ajpcell.00010.2011

Leonard W.R. *Pearl Memorial Lecture*. Humans at the extremes: exploring human adaptation to ecological and social stressors. *Am J Hum Biol.* 2024;36(3):e24010. doi 10.1002/ajhb.24010

Leonard W.R., Levy S.B., Tarskaia L.A., Klimova T.M., Fedorova V.I., Baltakhinova M.E., Krivoshapkin V.G., Snodgrass J.J. Seasonal variation in basal metabolic rates among the Yakut (Sakha) of northeastern Siberia. *Am J Hum Biol.* 2014;26(4):437-445. doi 10.1002/ajhb.22524

Levy S.B., Leonard W.R., Tarskaia L.A., Klimova T.M., Fedorova V.I., Baltakhinova M.E., Krivoshapkin V.G., Snodgrass J.J. Seasonal and socioeconomic influences on thyroid function among the Yakut (Sakha) of Eastern Siberia. *Am J Hum Biol.* 2013;25(6):814-820. doi 10.1002/ajhb.22457

Ma Y., Shen S., Yan Y., Zhang S., Liu S., Tang Z., Yu J., … Li Y., Hu C., Jiang J., Li Y., Ying H. Adipocyte thyroid hormone receptor-mediated hormone action fine-tunes intracellular glucose and lipid metabolism and systemic homeostasis. *Diabetes.* 2023;72(5):562-574. doi 10.2337/db22-0656

Malyarchuk B.A., Derenko M.V. Genetic history of the Koryaks and Evens of the Magadan region based on Y-chromosome polymorphism data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed.* 2024;28(1):90-97. doi 10.18699/vjgb-24-11

Martinez de Mena R., Scanlan T.S., Obregon M.J. The T_3 receptor beta1 isoform regulates UCP1 and D2 deiodinase in rat brown adipocytes. *Endocrinology*. 2010;151(10):5074-5083. doi 10.1210/en.2010-0533

Nikanorova A.A., Barashkov N.A., Pshennikova V.G., Nakhodkin S.S., Gotovtsev N.N., Romanov G.P., Solovyev A.V., Kuzmina S.S., Sazonov N.N., Fedorova S.A. The role of nonshivering thermogenesis genes on leptin levels regulation in residents of the coldest region of Siberia. *Int J Mol Sci.* 2021;22(9):4657. doi 10.3390/ijms22094657

Nikanorova A.A., Barashkov N.A., Pshennikova V.G., Gotovtsev N.N., Romanov G.P., Solovyev A.V., Kuzmina S.S., Sazonov N.N., Fedorova S.A. Relationships between uncoupling protein genes *UCP1*, *UCP2* and *UCP3* and irisin levels in residents of the coldest region of Siberia. *Genes.* 2022;13(9):1612. doi 10.3390/genes13091612

Nikanorova A.A., Barashkov N.A., Pshennikova V.G., Teryutin F.M., Nakhodkin S.S., Solovyev A.V., Romanov G.P., Burtseva T.E., Fedorova S.A. A systematic review and meta-analysis of free triiodothyronine (FT3) levels in humans depending on seasonal air temperature changes: is the variation in FT3 levels related to nonshivering thermogenesis? *Int J Mol Sci.* 2023;24(18):14052. doi 10.3390/ijms241814052

Noahsen P., Rex K.F., Bülow Pedersen I., Mulvad G., Florian-Sørensen H.C., Pedersen M.L., Andersen S. Adaptation to a high iodine intake in Greenland Inuit suggested by thyroid disease pattern. *Thyroid.* 2021;31(12):1850-1857. doi 10.1089/thy.2021.0342

Pagani L., Lawson D.J., Jagoda E., Mörseburg A., Eriksson A., Mitt M., Clemente F., … Nielsen R., Villems R., Willerslev E., Kivisild T., Metspalu M. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016;538(7624):238-242. doi 10.1038/nature19792

Sharma P., Levesque T., Boilard E., Park E.A. Thyroid hormone status regulates the expression of secretory phospholipases. *Biochem Biophys Res Commun.* 2014;444(1):56-62. doi 10.1016/j.bbrc.2014. 01.003

Sokolova E.E., Vlasov P.K., Egorova T.V., Shuvalov A.V., Alkalaeva E.Z. The influence of A/G composition of 3′ stop codon contexts on translation termination efficiency in eukaryotes. *Mol Biol.* 2020;54(5):739-748. doi 10.1134/S0026893320050088

Stæger F.F., Andersen M.K., Li Z., Hjerresen J.P., He S., Santander C.G., Jensen R.T., … Nielsen R., Jørgensen M.E., Hansen T., Moltke I., Albrechtsen A. Genetic architecture in Greenland is shaped by demography, structure and selection. *Nature.* 2025;639(8054):404-410. doi 10.1038/s41586-024-08516-4

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 2011;28:2731-2739. doi 10.1093/molbev/msr121

Tseng Y.H., Ke P.Y., Liao C.J., Wu S.M., Chi H.C., Tsai C.Y., Chen C.Y., Lin Y.H., Lin K.H. Chromosome 19 open reading frame 80 is upregulated by thyroid hormone and modulates autophagy and lipid metabolism. *Autophagy.* 2014;10(1):20-31. doi 10.4161/auto.26126

Tsibulnikov S., Maslov L., Voronkov N., Oeltgen P. Thyroid hormones and the mechanisms of adaptation to cold. *Hormones.* 2020;19(3):329-339. doi 10.1007/s42000-020-00200-2

Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B.C., Remm M., Rozen S.G. Primer3 – new capabilities and interfaces. *Nucleic Acids Res.* 2012;40(15):e115. doi 10.1093/nar/gks596

Wu C., Roy B., He F., Yan K., Jacobson A. Poly(A)-binding protein regulates the efficiency of translation termination. *Cell Rep.* 2020;33(7):108399. doi 10.1016/j.celrep.2020.108399

Yau W.W., Yen P.M. Thermogenesis in adipose tissue activated by thyroid hormone. *Int J Mol Sci.* 2020;21(8):3020. doi 10.3390/ijms21083020

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

Received February 24, 2025. Revised May 9, 2025. Accepted May 21, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):101-116

doi 10.18699/vjgb-26-08

Transcriptomics of severe COVID-19

A.A. Gusarova, E.A. Trifonova, A.A. Babovskaya, M.M. Gavrilenko, V.A. Stepanov

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
anastasia.gusarova@medgenetics.ru

Abstract. Currently, identifying biomarkers that can reliably predict the risk of developing severe COVID-19, potentially leading to fatal outcomes, remains a critical challenge. Studying the pathogenetic mechanisms underlying the progression from moderate to severe disease through blood transcriptome analysis enables the identification of differentially expressed genes (DEGs), which may serve as potential prognostic biomarkers of disease severity and as novel therapeutic targets for managing COVID-19 complications. In this review, we have summarized and analyzed studies that compared gene expression profiles between moderate and severe COVID-19 cases using bulk RNA sequencing of blood cell samples. Based on the results of five studies, five commonly and significantly differentially expressed genes were identified (*CD177*, *PPARG*, *PCOLCE2*, *SLC51A* and *ADAMTS2*), and their potential roles in the progression to severe COVID-19 are discussed. Functional enrichment analysis was performed, and shared pathways associated with severe COVID-19 were identified, including neutrophil degranulation, interleukin signaling, collagen biosynthesis, and suppression of adaptive and NK cell-mediated immune responses. Additionally, single-cell RNA sequencing (scRNA-seq) studies were reviewed, comparing moderate and severe cases, supporting some of the bulk RNA-seq findings. Due to the limited overlap of data in the reviewed articles, one section of this review focuses on the study designs, including analytical tools, sample collection protocols, and criteria used to define comparison groups. Transcriptomic analysis of the COVID-19 severe form reveals both cellular and molecular mechanisms of the immune response, the dysregulation of which can lead to the development of severe manifestations. RNA-markers seem to be promising predictors of the severity of COVID-19. At the same time, other omics technologies can fill in the gaps in understanding the characteristics of severe COVID-19 and identify mechanisms of disease progression to develop approaches for COVID-19 prevention and treatment.

Key words: severe COVID-19; RNA sequencing; transcriptomics; bulk RNA-seq; single-cell RNA-seq; differentially expressed genes

For citation: Gusarova A.A., Trifonova E.A., Babovskaya A.A., Gavrilenko M.M., Stepanov V.A. Transcriptomics of severe COVID-19. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):101-116. doi 10.18699/vjgb-26-08

Funding. The research was carried out at the expense of the State assignment (fundamental scientific research No. 122020200083-8).

Транскриптомика тяжелой формы COVID-19

А.А. Гусарова, Е.А. Трифонова, А.А. Бабовская, М.М. Гавриленко, В.А. Степанов

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
anastasia.gusarova@medgenetics.ru

Аннотация. В настоящее время выявление биомаркеров, позволяющих эффективно определить пациентов с риском развития тяжелой формы COVID-19, которая может привести к летальному исходу, считается важной задачей. Изучение патогенетических механизмов перехода умеренной формы в тяжелую с помощью анализа транскриптома крови обеспечивает идентификацию дифференциально экспрессирующихся генов (ДЭГ), которые могут стать потенциальными прогностическими биомаркерами тяжести инфекции, а также новыми терапевтическими мишенями в борьбе с осложнениями COVID-19. В данном обзорном исследовании проведены поиск и анализ работ по изучению различий в экспрессии генов при применении подхода секвенирования РНК пула клеток крови (bulk RNA-seq) при сравнении умеренной и тяжелой степени коронавирусной инфекции. По результатам пяти работ были определены пять общих, наиболее значимых дифференциально экспрессирующихся генов (*CD177*, *PPARG*, *PCOLCE2*, *SLC51A* и *ADAMTS2*) и рассмотрена их предполагаемая роль в развитии тяжелой формы COVID-19. Проведен анализ функционального обогащения, который определил общие пути, в которые вовлечены гены, дифференциально экспрессирующиеся при тяжелой форме COVID-19, такие как активация процессов дегрануляции нейтрофилов, путей интерлейкинов, биосинтеза коллагена и подавление путей адаптивного и опосредованного NK-клетками иммунного ответа. Проанализированы также результаты секвенирования РНК единичных клеток (single-cell RNA-seq) в изучении умеренной и тяжелой форм, подтверждающие некоторые результаты bulk RNA-seq. В связи с низкой общностью данных в рассматриваемых

© Gusarova A.A., Trifonova E.A., Babovskaya A.A., Gavrilenko M.M., Stepanov V.A., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

работах один из разделов обзора посвящен анализу дизайнов выбранных исследований, включая инструменты анализа, сбор материала и критерии формирования сравниваемых групп. Транскриптомика тяжелой формы COVID-19 раскрывает как клеточные, так и молекулярные механизмы иммунного ответа, дисрегуляция которого может привести к развитию тяжелых проявлений. При этом другие омиксные технологии смогут дополнить пробелы в изучении особенностей тяжелой формы и раскрыть механизмы прогрессирования заболевания для разработки подходов профилактики и терапии COVID-19.

Ключевые слова: тяжелая форма COVID-19; секвенирование РНК; транскриптомика; bulk RNA-seq; single-cell RNA-seq; дифференциально экспрессирующиеся гены

Introduction

COVID-19 (Coronavirus Disease 2019) is an acute respiratory infection caused by SARS-CoV-2. The clinical picture of COVID-19 has a wide range of manifestations and the course of the disease: it varies from mild to severe and critical (Prevention… COVID-19, 2023). The severe form may be accompanied by the development of acute respiratory distress syndrome (ARDS), which is characterized by an acute onset, severe hypoxemia, bilateral infiltration and pulmonary edema. Most patients with the severe form have lymphopenia, and some have thromboembolic complications. Severe COVID-19 can lead to multiple organ failure and death (Berlin et al., 2020; Huang C. et al., 2020).

Elucidating the pathogenetic basis of the disease, as well as key immune and inflammatory processes that differentiate a severe case from non-severe cases, is required to determine therapeutic strategies and prevention of severe COVID-19 (Jovic et al., 2022). Currently, laboratory blood tests are used for early identification of patients at risk of severe COVID-19, which can also be sensitive predictors of an unfavorable outcome of the disease (Chen et al., 2022; Roessler et al., 2023). In addition, the identification of such markers can contribute to the timely diagnosis of complications, monitoring and determination of optimal treatment (Tabassum et al., 2021). Biomarkers used to predict the severity of COVID-19 include complete blood count (white blood cell levels, levels of lymphocytes, neutrophils, neutrophil/lymphocyte ratio, platelet levels), markers of inflammation and proinflammatory cytokines (C-reactive protein, procalcitonin, ferritin, IL-6, IL-8, TNF-alpha, etc.), markers of damage to target organs (blood clotting factors, D-dimer, cardiomarkers and markers of kidney function), as well as markers of oxidative stress (levels of reactive oxygen species and antioxidants: vitamin C, thiol proteins, serum parameters of the glutathione system, redox status, lipid peroxidation) (Polonikov, 2020; Pincemail et al., 2021; Tabassum et al., 2021; Chen et al., 2022; Karu et al., 2022; Roessler et al., 2023; Liu X. et al., 2024). These biomarkers are significant indicators of severe COVID-19, but they are often identified only during the advanced stages of the disease (Chen et al., 2022). In this regard, studies aimed at identifying potential biomarkers that can be measured at earlier stages of infection and can predict the features of the organism's immune defense against SARS-CoV-2 are becoming relevant (Schultze, Aschenbrenner, 2021).

Blood transcriptomics plays a significant role in determining the mechanisms that reflect the immune responses of the host organism. The study of transcriptomic profiles of patients with varying severity of infection can provide unique information about the biological processes underlying the severity of the course, which can be used to identify prognostic and diagnostic RNA markers, as well as therapeutic targets (Lee H.J. et al., 2018; Huang W. et al., 2021; Schultze, Aschenbrenner, 2021). Today, the most common method of transcriptome analysis is high-throughput RNA sequencing (RNA-seq), which is a quantitative system for genome-wide expression profiling and, therefore, is applicable to characterize events related to dysregulation of gene expression in patients with severe COVID-19 (Hegenbarth et al., 2022).

Differences in gene expression identified in different clinical severity of the disease indicate that there are potential genes associated with the progression of the disease, rather than with the disease as a whole. Identification and functional characterization of differentially expressed genes in severe and moderate forms of the disease seem to be a promising strategy for identifying biomarkers of the severity of COVID-19. This can provide a deeper understanding of the pathogenesis of COVID-19, helping in the choice of treatment methods by deciphering the complexity of the immune response of organism and discovering new therapeutic targets (Arunachalam et al., 2020; Bando et al., 2023). However, reliable RNA markers have not been established in clinical practice to identify patients who may develop severe COVID-19.

Therefore, in this review, we focused on the search for shared genes when analyzing the differential expression of genes obtained in the study of moderate and severe forms of COVID-19 using leukocyte RNA sequencing.

Analysis of whole blood transcriptome in the various COVID-19 severity

Up to 2025, multiple transcriptome COVID-19 studies have been conducted. The number of different datasets in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) is about 300, whereas there are more than 60 original studies in the PubMed database. The design of such studies includes various RNA sequencing approaches, comparison groups, and analyzed materials (Fig. 1).

The main sample type for gene expression profiling is peripheral blood and its components. In connection with the main clinical manifestations of COVID-19, transcriptomes of nasopharyngeal/oropharyngeal swabs (Chua et al., 2020; Hadzega et al., 2024), bronchoalveolar lavage fluid (BALF) samples (Xiong et al., 2020; Nassir et al., 2021), etc. are also being studied. Since the entry receptor for SARS-CoV-2 is found in many tissues (https://www.uniprot.org/uniprotkb/

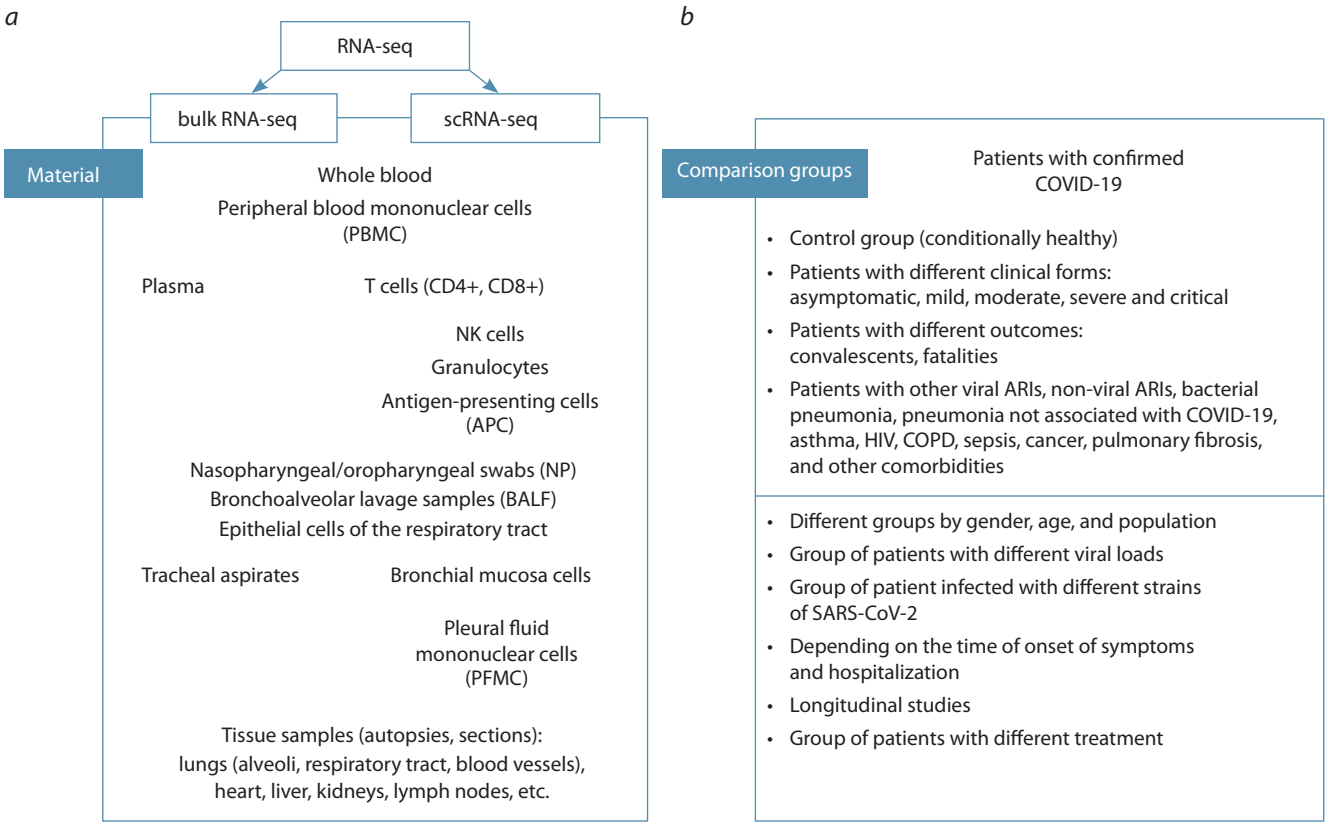


Fig. 1. RNA sequencing approaches, materials (*a*) and comparison groups (*b*) in the study of moderate and severe forms of COVID-19.

Q9BYF1/entry), the effect of coronavirus infection on other organ systems is being investigated, including the analysis of transcriptomes of heart, liver, kidney, and lymph node tissues (Delorey et al., 2021), as well as the main target of SARS-CoV-2: lung tissue (Fig. 1*a*). The main comparison groups include patients with confirmed COVID-19 and a control group consisting of conditionally healthy individuals. Individual studies compare transcriptomic profiles across various clinical forms: asymptomatic, mild, moderate, severe, critical, as well as across different outcomes, including convalescent patients and fatal cases. In addition, the differences between coronavirus infection and other viral and non-viral acute respiratory infections, bacterial pneumonia, asthma and COPD, sepsis, pulmonary fibrosis, and comorbidities are being investigated (Blanco-Melo et al., 2020; COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022). The analysis may also include different groups of COVID-19 patients by gender, age, population, high and low viral load, depending on the time of onset of symptoms and time of sampling, groups of patients without treatment and with specific treatment, as well as different strains of SARS-CoV-2 (Fig. 1*b*).

In most COVID-19 transcriptomics studies, peripheral blood serves as the primary source for identifying biomarkers of the disease. Changes in blood transcriptomic profiles can be caused by effects of immunogenic factors and/or changes in immune cell proportions (Chaussabel et al., 2010). The general dysregulation of certain genes may imply a specific mechanism of immune response. To identify the most reliable biomarkers in this study, a search was conducted for overlapping differentially expressed genes (DEGs) across several independent studies of moderate and severe forms of COVID-19. This may reduce the probability that the observed expression pattern is caused by the heterogeneity of the cell population (Song et al., 2017).

Keywords and expressions such as “severe COVID-19”, “moderate COVID-19”, “mild COVID-19”, “RNA-sequencing”, “bulk RNA-sequencing” were used when searching in the PubMed system for studies that obtained data on differential gene expression in severe and moderate COVID-19 using RNA sequencing. So, the criteria for selecting studies were the analysis of the whole blood transcriptome obtained by sequencing RNA from the pool of blood cells (bulk RNA-seq) in certain comparison groups – severe and moderate/mild COVID-19 (“severe versus moderate/mild”). The analysis included works that exactly meet the criteria; their characteristics are presented in Table 1.

A total of 7,650 differentially expressed genes were identified in these studies when comparing groups of severe and moderate forms (Fig. 2). For each pair of studies, the number of overlapping genes ranges from 23 (Aschenbrenner et al., 2021 and Armignacco et al., 2024) to 1,102 (Tang et al., 2020; Jackson et al., 2022).

When analyzing the generality of the DEGs, low replication of the results was noted. Only five genes were found, which were identified when comparing severe and moderate forms of COVID-19 in each of the selected studies. These observations

Table 1. Studies that conducted a whole transcriptome analysis of severe and mild/moderate forms of COVID-19

No.	Study	Comparison groups and number of patients	Platform	Results
1	Tang et al., 2020	Severe form (*n* = 6) and moderate form (*n* = 6)	HiSeq 4000 (Illumina)	3,082 DEGs (2,267↑, 815↓)
2	Aschenbrenner et al., 2021	Severe form (*n* = 20) and mild form (*n* = 19)	NovaSeq 6000 (Illumina)	1,097 DEGs (623↑, 474↓)
3	Jackson et al., 2022	Severe form (*n* = 10) and mild form (*n* = 19); Severe form and moderate form (*n* = 26)		7,343 DEGs (3,329↑, 4,014↓); 8,971 DEGs(4,380↑, 4,591↓)
4	Wang Y. et al., 2023	Severe form (*n* = 32) and moderate form (*n* = 25); Severe form and mild form (*n* = 31)		1,448 DEGs (1,013↑, 435↓); 4,592 DEGs (2,617↑, 1,975↓)
5	Armignacco et al., 2024	Severe pneumonia (*n* = 11) and mild pneumonia (*n* = 53)		345 DEGs (237↑, 108↓)

Notes. DEGs – differentially expressed genes. ↑ – upregulation. ↓ – downregulation.

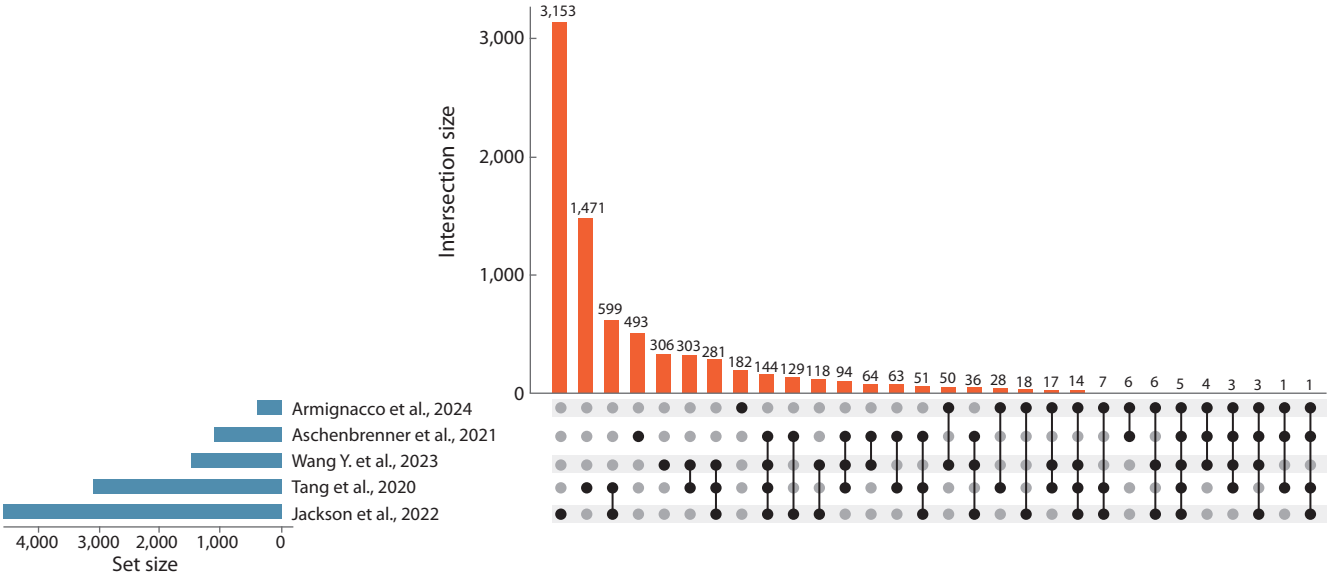


Fig. 2. UpSet plot, showing the number of differentially expressed genes when comparing the results of the selected studies.

The bar graph located on the left shows the number of detected DEGs for each study. The circles that make up the matrix represent sections of the Venn diagram. The connected circles indicate the intersection of genes between certain studies. The bar graph above the matrix shows the number of unique or overlapping DEGs. For example, the first and second columns of the chart reflect the number of unique genes for research (Jackson et al., 2022) (3,153 DEGs) and (Tang et al., 2020) (1,471 DEGs). The third column shows the number of overlapping genes for this pair only (599 DEGs). For the (Jackson et al., 2022; Wang Y. et al., 2023) studies, overlapping DEGs were included, identified when comparing severe and non-severe forms according to the WHO severity classification. The diagram is constructed using https://intervene.shinyapps.io/intervene/ (Khan, Mathelier, 2017).

may indicate the significance of changes in the expression of the identified genes in the pathogenesis of severe COVID-19.

Characterization of the most significant genes

Let us take a closer look at the characteristics of five overlapping differentially expressed genes: *CD177*, *PPARG*, *PCOLCE2*, *SLC51A* (upregulated) and *ADAMTS2* (downregulated in the study (Armignacco et al., 2024) and upregulated in studies (Tang et al., 2020; Aschenbrenner et al., 2021; Jackson et al., 2022; Wang Y. et al. al., 2023).

The ***PPARG*** gene encodes the nuclear receptor PPARγ, activated by the peroxisomal proliferator gamma. PPARγ regulates the peroxisomal pathway of beta-oxidation of fatty acids, as well as macrophage activation by inhibiting the production of inflammatory cytokines by monocytes (https://www.genecards.org/).

According to the DisGeNET database, this gene is significantly associated with adult-onset diabetes mellitus and obesity, for which the gene-disease association score (GDA score) is 1. These pathologies are risk factors for severe COVID-19 (de Seabra Rodrigues Dias et al., 2022). In addition, an association with COVID-19 was found with a GDA score of 0.3 (https://disgenet.com/).

It is assumed that SARS-CoV-2 suppresses the expression of PPAR in the lungs and disrupts the anti-inflammatory mechanism of NF-κB, thereby causing a hyperinflammatory

reaction in patients with severe COVID-19 (Desterke et al., 2020; Hasankhani et al., 2024). In transcriptomic studies, various changes in its expression are observed, including different tissues. For example, in the study (Vlasov et al., 2021), *PPARG* expression was increased in patients with an unfavorable outcome. The authors suggested that elevated PPARγ levels may be a sign of unresolved inflammation in conditions of lipid depletion, characteristic of the severe form of COVID-19 (Pei et al., 2021). According to the results of studies using network approaches, the *PPARG* gene has been proposed as a promising therapeutic target for controlling inflammation in COVID-19 (Auwul et al., 2021; Oh et al., 2021).

SLC51A. The alpha organic solute transporter encoded by the *SLC51A* gene is the main component of the OSTα/OSTβ complex, which acts as a transporter responsible for the export of bile acids from enterocytes.

According to the DisGeNET database (https://disgenet.com/), *SLC51A* is significantly associated with primary biliary cirrhosis (GDA score = 0.65), Byler's syndrome (GDA score = 0.5) and with a disorder associated with diabetes mellitus (GDA score = 0.4).

It has been revealed that *SLC51A* is a target for drugs in the treatment of COVID-19 (Morselli Gysi et al., 2021). In addition, *SLC51A* is included in significant genes in the differentiation of patients with sepsis and the determination of its endotypes, the clinical characteristics of which are often discussed as common with severe COVID-19 (Baghela et al., 2023; Fang, Ma, 2023). In the work (Pei et al., 2021), it was revealed that in SARS-CoV-2 infected respiratory tract and alveolar organoids derived from human embryonic stem cells, the SLC51A transporter was found in reduced amounts. Despite the importance of this gene in differentiating patients and identifying therapeutic targets for COVID-19, the role of the *SLC51A* gene in the mechanisms of severity of coronavirus infection remains poorly understood.

The ***CD177*** gene mediates TNF-alpha-induced neutrophil activation, including degranulation and superoxide production, and promotes neutrophil adhesion (https://www.genecards.org/).

According to the DisGeNET database, this gene is significantly associated with myeloproliferative disorders (GDA score = 0.35), as well as with COVID-19 with a GDA score of 0.25 (https://disgenet.com/).

The relevance of this biomarker in predicting the severe course of coronavirus infection has been revealed in several transcriptomic and proteomic studies (Derakhshani et al., 2021; Lévy et al., 2021; Meizlish et al., 2021; Schimke et al., 2022; Wang Q.S. et al., 2022; Lei, 2024). In the work (Lévy et al., 2021), a high expression of *CD177* and a higher average level of this serum protein in blood were observed in critically ill patients. It is assumed that neutrophil degranulation causes endothelial damage and, consequently, thrombotic complications in COVID-19 (Reusch et al., 2021). Thus, hyperexpression of *CD177* is a sign of the physiopathology of COVID-19 and may act as a possible prognostic factor for the progression of the disease.

PCOLCE2 encodes a procollagen C-endopeptidase enhancer 2 protein, which provides collagen and heparin binding activity.

Increased expression of the *PCOLCE2* gene is detected in patients with severe COVID-19 (Alqutami et al., 2021; Che et al., 2022). The enrichment analysis revealed that this gene is involved in the extracellular matrix organization and regulation of the inflammatory response. It is noted that *PCOLCE2* can enhance the activity of collagen, and SARS-CoV-2 infection causes an increase in the level of collagen 1 in organoids and promotes the activation of fibrosis signaling pathways (Jansen et al., 2022). It is also assumed that *PCOLCE2* is a factor stimulating the production of reactive oxygen species by neutrophils and the formation of neutrophil extracellular traps (NETs) (Yoon et al., 2022).

The ***ADAMTS2*** gene encodes a member of the ADAMTS protein family, which plays a key role in the conversion of procollagen fibrillar precursors into collagen molecules (https://www.genecards.org/).

According to the DisGeNET database, *ADAMTS2* is significantly associated with the development of malignant mesothelioma (GDA score = 0.4), as well as with COVID-19 with a GDA score of 0.25 (https://disgenet.com/).

ADAMTS proteins are involved in extracellular matrix remodeling, which could potentially be important in the development of pulmonary fibrosis seen in patients with severe COVID-19. DEG analysis and analysis of gene ontology (GO) in groups of severe, precritical and critical COVID-19 showed that fibrosis-related genes (*AREG*, *EREG*, the IL-18 cytokine gene) and *ADAMTS2* are highly expressed in monocytes (Zhang Y. et al., 2022).

It is assumed that ADAMTS2 mediates the pathway of TGF-β, transforming growth factor-β (de Seabra Rodrigues Dias et al., 2022). Disruption of TGF-β signaling contributes to excessive extracellular matrix deposition in tissues, which may be caused by infection and inflammation (Togami et al., 2017; Deng et al., 2024). In addition, TGF-β is involved in maintaining immune homeostasis by suppressing the activity of immunocompetent cells: this cytokine prevents the differentiation of naive T cells into classical effector T cells, suppresses the expression of cytotoxic factors and the maturation of NK cells (Deng et al., 2024). It is also assumed that TGF-β is a key cytokine regulating the chronic immune response in severe COVID-19 (Ferreira-Gomes et al., 2021).

Due to the association of *ADAMTS2* with the activation of the TGF-β pathway and the formation of the extracellular matrix, increased expression of this gene in severe COVID-19 may reflect immunopathological processes characteristic of the progression of the disease and the development of pulmonary fibrosis.

So, the proposed mechanisms by which the identified genes are involved in the pathogenesis of severe COVID-19 were considered. They are reflected in the general scheme (Fig. 3), which also includes the most significant associations of these genes with diseases according to the DisGeNET database. A decrease in *PPARG* expression can lead to low PPARGγ expression, which leads to increased proinflammatory reactions; an increase in *CD177* expression leads to neutrophil degranulation and the release of reactive oxygen species, and then, to the development of hyperinflammation characteristic of the severe form. Involvement of neutrophils in the pathogenesis is also

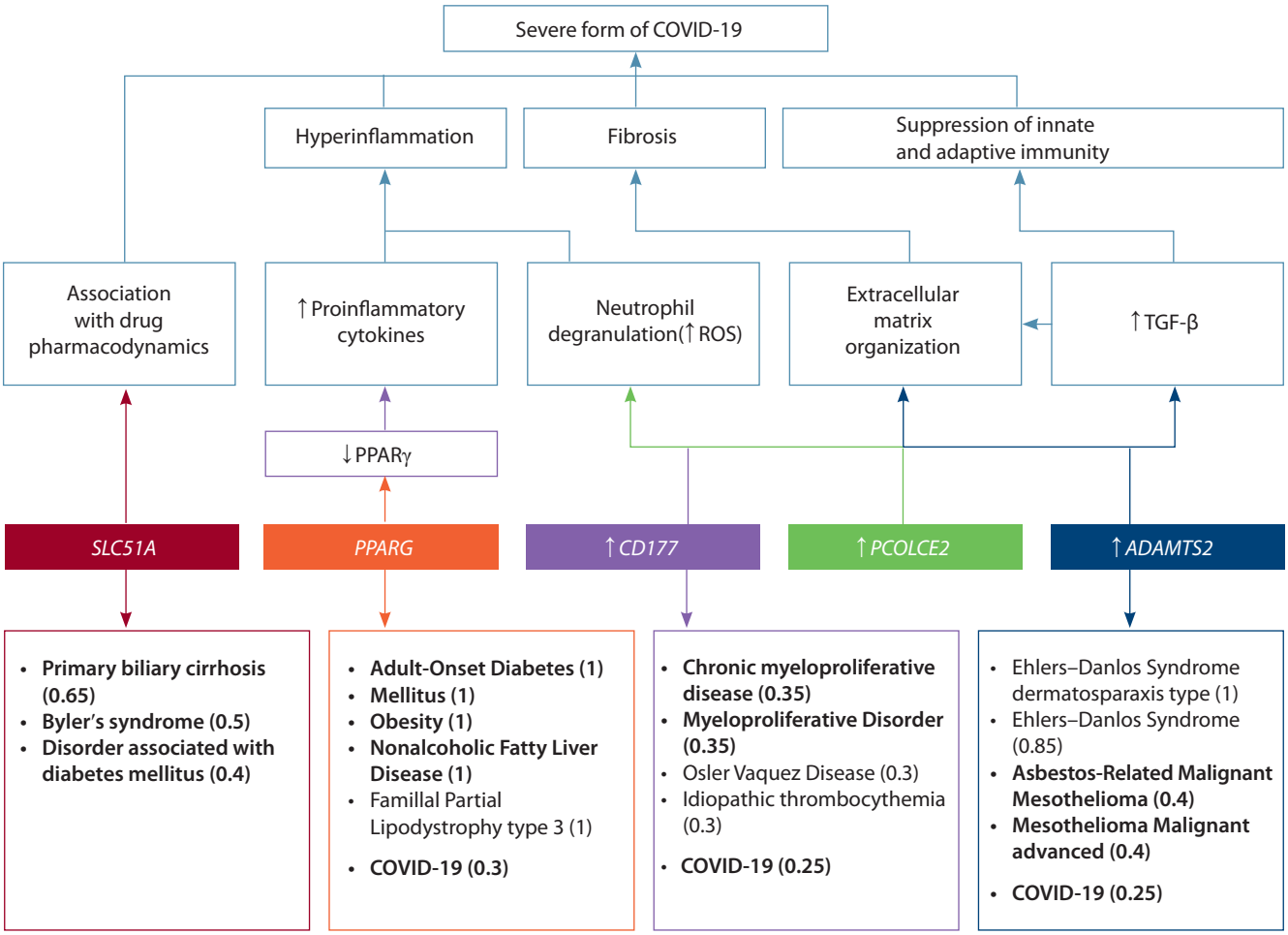


Fig. 3. Pathogenetic aspects of severe COVID-19, including shared genes found in five selected whole transcriptome studies of moderate and severe forms of coronavirus infection and the most significant associations of these genes with diseases according to the DisGeNET database (GDA score ≥ 0.3).

Diseases belonging to pathology groups that are risk factors for severe COVID-19 are highlighted in bold. ROS – reactive oxygen species.

possible through increased expression of *PCOLCE2* as a factor that also stimulates the production of reactive oxygen species. *PCOLCE2* and *ADAMTS2* may be involved in the development of pulmonary fibrosis as a complication of severe pneumonia through various extracellular matrix organization pathways by activating collagen synthesis and the TGF-β pathway. Changes in *ADAMTS2* expression can also lead to dysregulation of TGF-β and, thus, cause suppression of both innate and adaptive immunity. Interestingly, according to DisGeNET, these genes are associated with diseases identified as risk factors for severe COVID-19 (Prevention… COVID-19, 2023): diabetes mellitus and its associated complications (*PPARG*, *SLC51A*), obesity (*PPARG*), cancers (*CD177*, *ADAMTS2*) and chronic liver diseases (*SLC51A*).

The mechanisms driving the progression from moderate to severe COVID-19 are not yet fully elucidated, and the specific roles of these genes in the pathogenesis of severe symptoms require further investigation. Furthermore, discordant expression patterns observed across studies may indicate different regulatory signatures. Analyzing the signaling pathways involving these genes provides a novel perspective on the pathogenesis of severe SARS-CoV-2 infection, accounting for the functional relationships between genes.

Functional enrichment analysis of overlapping differentially expressed genes associated with the severity of COVID-19

Functional enrichment was also analyzed in each of the reviewed studies for DEGs obtained by comparing severe and moderate forms of coronavirus infection. In all studies, significant signaling pathways associated with gene upregulation are associated with the inflammatory immune response of the organism. For example, the processes of degranulation and activation of neutrophils. The lymphocytic immune response pathways have been enriched with both upregulated and downregulated genes in various studies. Significant enrichments have been identified in the pathways associated with type 2 T-helper cells involved in the humoral immune response (Jackson et al., 2022), and in the processes associated with lymphocyte activation, proliferation, and regulation (Tang

et al., 2020). At the same time, the processes enriched with genes with a decreased expression did not overlap between the studies. Thus, questions arise about the mechanisms of disruption of the interaction of innate and adaptive immune responses and the transition from a "viral" response in the mild form of COVID-19 to a severe inflammatory process (Jackson et al., 2022).

Since most of the significant signaling pathways in the analyzed studies are different when comparing severe and moderate forms of coronavirus infection, we performed an enrichment analysis for overlapping differentially expressed genes in order to identify common patterns of development of the severe form. Four papers were selected (Tang et al., 2020; Aschenbrenner et al., 2021; Jackson et al., 2022; Wang Y. et al., 2023), in which the identified differentially expressed genes overlapped the most. They included 149 DEGs: 102 genes with hyperexpression and 47 genes with hypoexpression. The functional enrichment analysis was performed using the WebGestalt resource (https://www.webgestalt.org/), as well as Reactome (Fabregat et al., 2017).

The results of the functional enrichment analysis using Reactome are presented in Table 2.

The upregulated genes in the severe disease group are associated with neutrophil degranulation (R-HSA-6798695) and collagen formation (R-HSA-1474290). Activation of interleukin signaling pathways (R-HSA-449147) – IL-18 (R-HSA-9012546), interleukin-1 family (R-HSA-446652), including IL-33 (R-HSA-9014843), IL-4 and IL-13 (R-HSA-6785807) – was also revealed. IL-4, the key cytokine of the Th2 immune response, alongside IL-13, is mainly associated with fibrotic inflammatory remodeling (Vaz de Paulaet al., 2020; Wang F. et al., 2020). In addition, IL-33, a signaling protein that warns the immune system of damage, may play an important role in all stages of COVID-19 (Zizzo, Cohen, 2020), and may also be associated with the development of fibrosis (Uchasova et al., 2018). It has been reported that IL-18 signaling, involved in the regulation of the Th1, Th2, and Th17 types of immune response, reflects inflammasome activation associated with the severity of the disease in the lungs (Filbin et al., 2021; Nasonov, Avdeeva, 2022).

Downregulated genes in severe COVID-19 were associated with adaptive immunity (R-HSA-1280218), as well as with the expression (R-HSA-9839394) and signaling of TGFBR3 (R-HSA-9839373), a type III TGF-β receptor that inhibits TGF-β signaling (Ahn et al., 2009; Chu et al., 2010). The role of this pathway was considered in the characterization of the *ADAMTS2* gene.

Thus, both the analysis of pathways identified in certain studies and the analysis of pathways associated with shared genes reveal unbalancing changes in the immune response during the transition to severe form, characterized by activation of pathways of neutrophil degranulation, signals of pro-inflammatory (IL-1 β, IL-18) and anti-inflammatory (IL-4/13) cytokines, as well as suppression of the adaptive immune response, which may be a key factor influencing the severity of COVID-19.

At the same time, the results of the Gene Ontology analysis (Fig. 4) included significant signaling pathways associated with downregulated genes that belong to the functional categories of NK cell mediated immunity (GO:0002228), lymphocyte mediated immunity (GO:0002449), and leukocyte mediated cytotoxicity (GO:0001909). These observations indicate the important involvement of immune cells in the response to SARS-CoV-2.

Research designs

In order to determine the causes of variation and low replication of genes and signaling pathways, a comparison of research

Table 2. The most significant Reactome pathways with FDR ≤ 0.05

Pathway	*p*-value	FDR
Pathways associated with upregulated genes		
Neutrophil degranulation (R-HSA-6798695)	7.24e–10	2.83e–07
Interleukin-18 signaling (R-HSA-9012546)	1.22e–04	0.024
Collagen formation (R-HSA-1474290)	2.76e–04	0.035
Signaling by interleukins (R-HSA-449147)	3.86e–04	0.035
Interleukin-4 and interleukin-13 signaling (R-HSA-6785807)	4.58e–04	0.035
Interleukin-33 signaling (R-HSA-9014843)	5.52e–04	0.035
Interleukin-1 family signaling (R-HSA-446652)	6.45e–04	0.035
Pathways associated with downregulated genes		
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933)	1.22e–07	2.53e–05
TGFBR3 expression (R-HSA-9839394)	1.56e–04	0.016
Adaptive immune system (R-HSA-1280218)	7.70e–04	0.041
Signaling by TGFBR3 (R-HSA-9839373)	7.83e–04	0.041

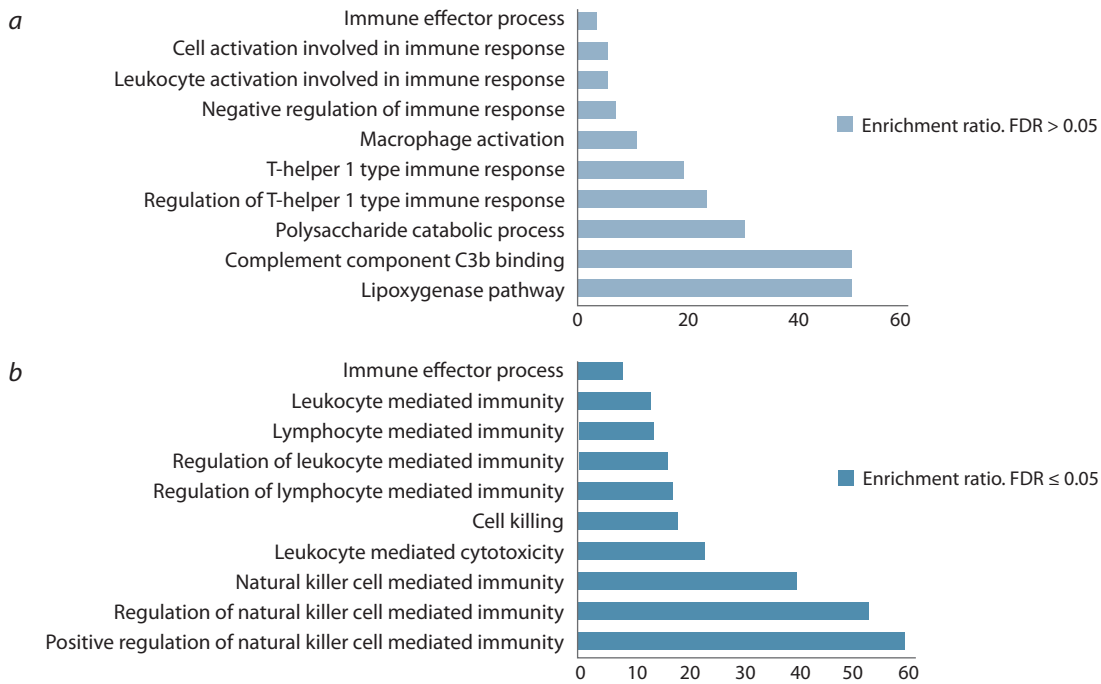


Fig. 4. Gene Ontology pathways enriched in upregulated (*a*) and downregulated (*b*) genes.

designs was carried out, including a comparison of the characteristics of patient cohorts, RNA sequencing platforms, as well as statistical data processing methods. Further, studies (Tang et al., 2020; Aschenbrenner et al., 2021; Jackson et al., 2022; Wang Y. et al., 2023; Armignacco et al., 2024) will correspond to the numbers 1, 2, 3, 4 and 5, respectively.

Analysis tools. Different software packages of the R environment were used to obtain differentially expressed genes: DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and limma (Ritchie et al., 2015). An adjusted *p*-value of less than 0.05 was considered significant for all studies. The threshold for absolute log2 fold change (|log2FC|) was set at ≥1 in analysis (1) and >1.5 in analyses (4, 5).

Severity of COVID-19. The inclusion of patients in the analyzed studies followed certain general criteria: age 18 years and older; positive result of the RT-PCR test for SARS-CoV-2 in respiratory samples (nasal/pharyngeal swab/sputum/bronchoalveolar lavage) and/or serological test (5) and/or diagnosis of COVID-19 based on the presence of typical clinical symptoms and CT results (2, 5). Patients in all studies were classified as having a mild, moderate, or severe form of the disease. The classification of the severity of COVID-19 in four papers was based on the recommendations of the World Health Organization. However, the rules for the formation of severity groups were different. The mild severity corresponded to the WHO 1–2 in the study (3) and the WHO 1–4 in (2, 4). The moderate severity corresponded to the WHO level 3–4 in (3), and to the WHO level 5 in the study (4). Severe COVID-19 corresponded to WHO severity levels 5–7 in (2), WHO 5–8 in (3) and WHO 6–9 in (4). WHO Severity Classification Score is given in (3). In study (5), the severity of patients was determined by the course of pneumonia development. The patients' condition was classified as mild, moderate, or severe pneumonia, depending on the onset and development of COVID-19 complications, including the duration of hospitalization, the need for oxygen, mechanical ventilation, or extracorporeal membrane oxygenation. Due to the variation in the formation of comparison groups, the DEG analysis included groups representing the mild form of the disease.

Collecting material. The total collection period for all works was one year, from February 2020 to February 2021. During this period, such variants of SARS-CoV-2 as Beta, Alpha, Delta, Gamma, etc. circulated (https://www.who.int/ru/activities/tracking-SARS-CoV-2-variants). It is assumed that the virus variant may affect the severity of the disease. Transcriptomic studies were conducted to identify differences in the host response to infection with different SARS-CoV-2 variants: "Pre-VOC" and "VOCs" (variants of concern). "Pre-VOC" infection was characterized by moderate manifestations and persisted much longer than infection during VOCs circulation. Delta infection was severe, leading to high rates of hospitalization and mortality (Hughes et al., 2023; Maurya et al., 2023). It has been demonstrated that this variant is capable of enhanced replication due to sustained suppression of the innate immune response of the host organism, which potentially contributes to severe symptoms and long-term recovery (Laine et al., 2022). Despite the identification of differentiated transcriptomic responses, data on the effect on the severity of a particular variant of SARS-CoV-2 still remain contradictory. In some studies, at a significance level of 0.05, no association was found between viral lines and the severity of the disease at an early stage of the pandemic (Parikh et al., 2022).

Patients' blood samples were collected at different time periods: during the first 24 hours after hospitalization (2, 5) or 5–7 days after admission to the hospital (3). In addition,

blood was collected not only in a hospital environment. For example, in study (3), blood samples were collected from patients with mild severity at home. Blood collection for further RNA isolation was carried out in PAXgene™ tubes and, after the necessary manipulations, stored at a temperature of –80 °C; the type of tube was not disclosed in the study (1). It is important to note that incomplete convergence of the results is observed when comparing the gene expression profiles of blood cells collected in different types of RNA stabilization tubes (Menke et al., 2012).

Criteria for the formation of a patient sample. The criterion for excluding patients for all studies was identification of coinfection (for example, human immunodeficiency virus). Study (2) excluded neutropenia, hematological malignancies and/or active chemotherapy, solid organ transplantation, autoimmune diseases, and any previous use of immunosuppressants (corticosteroids, anti-cytokine biologics, and biological response modifiers), while study (3) accounted for immunomodulatory treatment as a covariate in its analysis.

When comparing the formed groups, in some studies, there were no statistically significant differences in gender (1, 2, 4, 5) and age (1, 2, 4). However, study (3) reported a correlation between severity and gender/age. Additionally, the study populations varied: East Asia, Australia and Europe.

Comorbidities. The analysis also considered such an important factor for the progression of the severity of COVID-19 as comorbidities. In study (2), the Charlson comorbidity index was calculated for each of the comparison groups (Charlson et al., 1987). According to this index, there were no statistically significant differences between the groups of mild, moderate and severe severity. In study (3), the incidence of endocrine comorbidities, smoking, and obesity was highest in the group with severe disease: 70, 50, and 50 %, respectively, compared with 53.8, 15.4, and 46.2 % of patients with moderate COVID and 21, 5.3, and 5.3 % with mild COVID-19. In study (5), significant differences between patients of varying severity among comorbidities were identified only for diabetes mellitus (adjusted $p = 0.048$).

Thus, despite meeting the criteria for selecting studies, noticeable differences were found in the characteristics of the comparison groups, as well as the analysis tools, which may explain the observed variation in the results of the selected studies. Another key factor influencing the variability of the results is the cellular heterogeneity of the analyzed samples. It is possible to study the contribution of cellular heterogeneity and consider in more detail the composition and functional characteristics of cells using the single-cell RNA sequencing approach (scRNA-seq).

Single-cell RNA sequencing in studying severe COVID-19

Numerous studies have recently employed single-cell RNA-seq technology to characterize the immune landscape in COVID-19 patients. Table S1 (Supplementary materials)[1] shows the features of gene expression and cellular composition when comparing severe and moderate forms of COVID-19. According to the results of scRNA-seq of blood cells, the severe form is characterized by certain changes in the composition of immune cells: there was a decrease in plasmocytoid dendritic cells, NK cells, non-classical monocytes and an increase in the proportion of classical monocytes, mature neutrophils and immature subpopulations of monocytes and neutrophils. In the severe form of COVID-19, low-density neutrophils (LDN) associated with dysfunctional immune responses have also been found to occur in conditions of emergency myelopoiesis (Schultze et al., 2019), which was not observed in the mild form. LDN, like mature neutrophils, secreted high levels of alarmins S100A8 and S100A9 (Schulte-Schrepping et al., 2020; Silvin et al., 2020; Ren et al., 2021; Wilk et al., 2021). Alarmins are released under inflammatory conditions and form a stable heterodimer known as “calprotectin” (Wang S. et al., 2018), which is involved in the activation and chemotaxis of neutrophils (Ryckman et al., 2003), and is also suspected to be the cause of cytokine release syndrome (Silvin et al., 2020). However, in severe cases, multidirectional gene expression of certain proinflammatory (*IL1B*, *TNF*) and anti-inflammatory (*IFNG*) cytokines was also detected, with an overall increased production of proinflammatory cytokines by immune cells.

The revealed differences in the composition and functional activity of T cells with varying severity and at different stages of COVID-19 may indicate the complexity of T cell responses to infection (Ren et al., 2021). At the same time, each type of cell involved in the response to coronavirus infection differs in a specific pattern of gene and membrane protein expression, reflecting the complex pathophysiology of severe COVID-19 manifestations. The main differences were found in the response of cells to type I interferon, as well as in the expression of interferon-stimulated genes (*ISG15* and *IFITM1/2*). Such observations can be explained by temporary changes at different stages of disease progression. In severe cases, there is a time shift: from an early but short-term reaction to type I interferon to a proinflammatory reaction at later stages (Arunachalam et al., 2020); therefore, it is crucial to account for the timing of sample collection relative to symptom onsetduring the analysis. However, not all studies specify cell types when analyzing gene expression, which also makes it more difficult to identify common changes in severe COVID-19.

Several studies have also observed shared pathophysiological pathways driving severe disease, reflecting the phenomenon of dysregulation of the immune response. For example, most studies have revealed a disrupted major histocompatibility complex class II antigen presentation by monocytes to T cells (*HLA-DR*), which indicates the development of immunosuppression (Zurochka et al., 2008). In addition, an activated pathway of NF-κB, a transcription factor contributing to the development of a “cytokine storm” and oxidative stress by enhancing the synthesis of proinflammatory cytokines and reactive oxygen species (ROS) by activated macrophages and neutrophils, was identified (Bolevich S.B., Bolevich S.S., 2020; Kesika et al., 2024). Oxidative stress resulting from increased ROS formation and decreased antioxidant protection contributes to the pathogenesis of the severe form of

[1] Supplementary Table S1 is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Gusar_Engl_30_1.pdf

COVID-19, leading to damage to cells and tissues through direct damage, lipid peroxidation and protein oxidation. Oxidative stress, coupled with cytokine release, induces endothelial dysfunction and activates the coagulation cascade, leading to microvascular thrombosis. Proinflammatory cytokines additionally stimulate ROS synthesis, exacerbating ARDS and lung tissue damage, causing a vicious cycle between oxidative stress and a cytokine storm (Polonikov, 2020; Gadotti et al., 2021; Alam, Czajkowsky, 2022; Labarrere, Kassab, 2022). The analysis of transcriptomic data also confirms the importance of a disturbance of the redox balance in the pathogenesis of severe COVID-19. The authors of the study (Saheb Sharif-Askari et al., 2021) conducted an *in silico* analysis of publicly available transcriptomic data from COVID-19 patients to assess the expression levels of 125 genes associated with oxidative stress. Seven genes (*MPO*, *S100A8*, *S100A9*, *SRXN1*, *GCLM*, *SESN2* and *TXN*) were found to have increased expression in whole blood and lung autopsies of patients with severe COVID-19 compared with patients with a non-severe form. In the work (Tavassolifar et al., 2023), increased levels of intracellular ROS, decreased levels of glutathione, and upregulation of oxidant and antioxidant genes were detected in peripheral blood mononuclear cells of patients with COVID-19 (*CAT*, *NFE2L2*, *SOD1*, *SOD2* and *CYBB*).

Commonality of bulk and single-cell RNA-seq studies

Some patterns revealed by analyzing the transcriptomes of bulk blood samples of patients with moderate and severe forms of COVID-19 were also replicated using single-cell transcriptomics.

An increase in the proportion of circulating neutrophils, their hyperactivation and dysregulation, and the appearance of immature or developing neutrophils associated with the severity of COVID-19 have been found in several studies using the scRNA-seq approach. The analysis revealed hyperexpression of genes encoding proinflammatory cytokines associated with phagocytosis and degranulation, as well as genes (for example, *PADI4*, *MPO*, *ELANE*, and *PRTN3*) that are involved in the formation of neutrophil extracellular traps (Barnes et al., 2020; Schulte-Schrepping et al., 2020; Silvin et al., 2020; Wilk et al., 2021). Excessive NET release from neutrophils contributes to the development of oxidative stress, hypercoagulation, impaired alveolar microcirculation and damage to lung tissue. The levels of components of neutrophil extracellular traps – for example, *DEFA1* RNA and neutrophil elastase (ELANE) activity in the blood – are considered as potential biomarkers of the severe form of COVID-19 (Wargodsky et al., 2022).

In addition, activated mature neutrophils in severe COVID-19 showed increased expression of the *CD177* gene, the possible role of which in pathogenesis was previously discussed, as well as hyperexpression of the *CD274* gene and the *ARG1* gene (Schulte-Schrepping et al., 2020; Wilk et al., 2021), which is part of 149 overlapping DEGs. *CD274* (PD-L1) and *ARG1* are associated with suppression of T cell activation, suggesting that neutrophils may perform immunosuppressive functions in severe COVID-19 (Schulte-Schrepping et al., 2020). These observations are consistent with suppression of shared lymphocyte signaling pathways.

The enrichment of both pro- and anti-inflammatory interleukin signaling pathways identified by bulk RNA-seq analysis is consistent with the detected high cytokine expression in various immune cells, mainly classical monocytes and macrophages via single-cell analysis approach. Increased levels of proinflammatory cytokines are thought to play an important role in the severe progression of COVID-19, causing a hyperinflammatory reaction called a "cytokine storm" (Ershov et al., 2020; Guo et al., 2020). It was found that the IFN-I reaction can contribute to the development of the hyperinflammatory response caused by IL-1β in severe progression of COVID-19 (Lee J.S. et al., 2020). The genetic signature of the IL-4/13 and IL-18 signaling pathways in monocytes also increased significantly in severe cases (Lee J.S. et al., 2020; Liu C. et al., 2021; Jeong et al., 2023).

Natural killer (NK) cells play an important role in innate immune responses to viral infections. Decreased functional activity of NK cells is associated with acute and chronic viral infection (Abakushina et al., 2012). Analysis of single-cell RNA sequencing results revealed differences in the transcriptome of NK cells between groups of patients with moderate and severe COVID-19. For example, a significant transcriptional remodeling was driven by an increase in the expression of canonical NK cell activation genes in the severe form, including increased expression of genes encoding cytotoxic effector molecules *GZMB*, *PRF1*, *GZMA*, as well as the proliferation marker *MKI67* (Wilk et al., 2021; Shaymardanov et al., 2022). However, genes associated with NK cell cytotoxicity (*GZMM*, *GZMH*, *GZMA*) were identified among the common low-expression genes in bulk RNA-seq analysis, and the NK cell-mediated immunity signaling pathway (GO:0002228) was enriched in downregulated genes. At the same time, signs of NK cell depletion at the transcriptome level in patients with the severe form were observed in scRNA-seq studies (Lee J.S. et al., 2020; Krämer et al., 2021; Liu C. et al., 2021; Wilk et al., 2021; Witkowski et al., 2021). In the work (Witkowski et al., 2021) NK cells demonstrated dysregulation of cytokine production, cell-mediated cytotoxicity, and response to the virus, despite high expression of cytotoxic effector molecules. At the same time, transcriptional networks responsible for the activation of NK cells were superimposed by the dominant signature of the response to TGF-β. In the study (McClain et al., 2023) differential expression of the *TGFB1* gene encoding transforming growth factor beta-1 was observed in $CD14^+$ monocytes and was associated with a worsening of COVID-19, and in the study (Ren et al., 2021) upregulated expression of *TGFB1* was observed in T cells, B cells, NK and dendritic cells. The TGF-β pathway was identified as one of the key pathways activated in severe cases of SARS-CoV-2 Delta infection (Shaymardanov et al., 2022). Thus, the involvement of TGF-β is confirmed by the analysis of the transcriptome at the level of a single cell. These results suggest that disturbances in the cytotoxicity of NK cells, including through the influence of the TGF-β pathway, may be associated with the mechanisms of severe COVID-19 development (Su et al., 2020; Lee M.L., Blish, 2023).

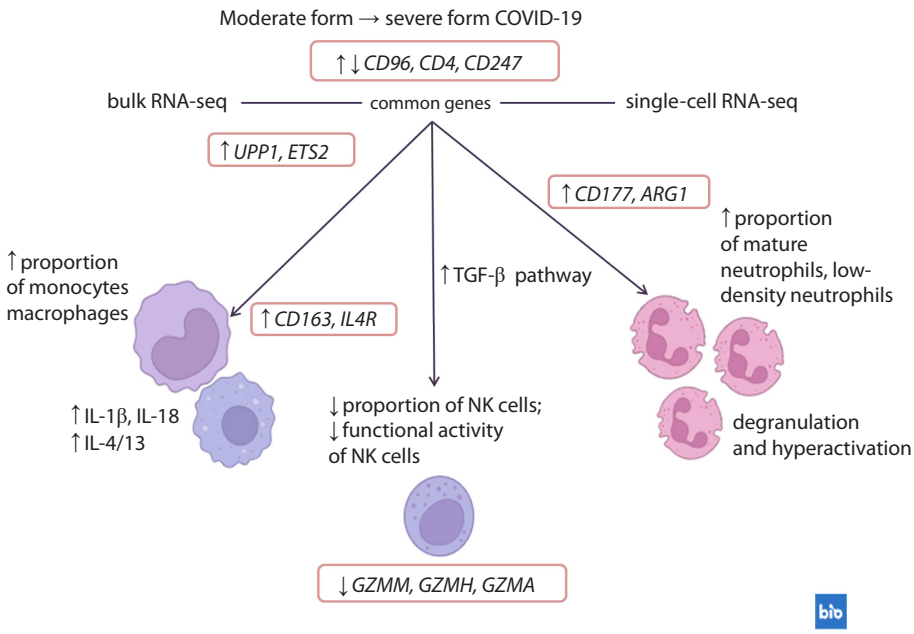


Fig. 5. Shared differentially expressed genes and consistent signatures identified through the comparison of moderate and severe COVID-19 cases using bulk and single-cell RNA-sequencing approaches.

↑ – increase in proportion or expression; ↓ – decrease in proportion or expression; ↑↓ – multidirectional expression. Made with https://BioRender.com.

The features of the development of the severe form of COVID-19 are characterized by a multifaceted immune dysregulation, which is described as a state of imbalance (Yao et al., 2021). Despite the differences in RNA-sequencing methodologies, shared differentially expressed genes and consistent patterns were identified when comparing moderate and severe forms (Fig. 5). There is a dysregulation of innate (hyperinflammatory reactions, decreased cytotoxicity of NK cells, activation and degranulation of neutrophils) and adaptive (suppression of lymphocytic immunity pathways) immune responses. At the same time, the multidirectional expression of some shared DEGs, for example *CD96*, *CD4* and *CD247*, expressed on different immune cells and involved in the immune response, was also revealed, which is consistent with the phenomenon of cellular heterogeneity. In bulk RNA-seq, the overall expression levels are influenced by shifts in cell type proportions within the patients' blood samples, while in single-cell RNA-seq, gene expression levels are obtained at single-cell resolution.

Conclusion

Peripheral blood transcriptomics is a powerful tool for studying the immune system in patients with coronavirus infection. The dynamics of the organism's immune response during an infectious process is reflected in changes in gene expression and can be studied in patients with severe disease development to identify key pathways for disease progression. It is assumed that changes in the expression of certain genes may be predictors of severe COVID-19.

So, this work analyzes the core set of differentially expressed genes and signaling pathways identified across COVID-19 transcriptomic studies. Studies that met the selection criteria were analyzed, including bulk RNA-seq of whole blood and a comparison of moderate and severe cases. A total of 7,605 differentially expressed genes were identified; five overlapping genes were detected: *CD177*, *PPARG*, *PCOLCE2*, *SLC51A* and *ADAMTS2*. Their possible role in the pathogenesis of severe COVID-19 was considered. Pathological processes such as hyperinflammation, fibrosis, and dysregulation of adaptive and innate immunity may be among the possible mechanisms of progression of coronavirus infection. For the four papers closest in design, shared enriched pathways for 149 DEGs were found, which included activation of neutrophil degranulation, interleukin signaling pathways, collagen biosynthesis, and suppression of adaptive and NK cell immune pathways.

The analysis also included transcriptomic studies of various clinical forms of COVID-19 using single-cell technology, which confirm the hypothesis of immune dysregulation identified by the results of bulk RNA-seq in patients with development of the severe form. The severe course is characterized by increased proportions and activity of mature and developing neutrophils, the onset of cytokine release syndrome and oxidative stress, as well as impaired cytotoxicity of NK cells involving the TGF-β pathway.

Research in the field of COVID-19 transcriptomics reveals the features of cellular and molecular processes that can lead to the development of a severe form of coronavirus infection. Identification of RNA biomarkers in patients with COVID-19 can contribute to more effective recognition of patients from risk groups, stratification of disease severity, and prediction of severe course and outcomes of COVID-19. Transcriptomic

biomarkers also show their effectiveness in differentiating patients with SARS-CoV-2 infection from patients with other lung infections. In addition, they play an important role in determining the directions of effective treatment of infection, including long-term COVID-19. RNA markers can be used as therapeutic targets, as well as in evaluating treatment response. These advantages highlight the potential of RNA biomarkers for clinical applications in diagnosis, prognosis, and the development of personalized treatment and rehabilitation strategies for patients.

However, despite the high sensitivity of RNA biomarkers, their integration into clinical practice faces significant difficulties. Methodological diversity of studies often makes it difficult to generalize the results and limits the possibility of identifying reproducible RNA biomarkers of COVID-19 severity. The variability of RNA sequencing data may also stem from individual physiological differences in patients and environmental factors. The results may not be representative for patients from different populations because of the population-specific immune response (Nédélec et al., 2016; Randolph et al., 2024). It is also important to consider the specifics of obtaining the material in a clinical setting: accurate measurement of RNA is hampered by the instability of RNA in the blood and the difficulty of purification. In the future, integrating new technologies such as single-cell RNA sequencing, validation of the effectiveness of the identified predictors in larger multicenter trials, and the development of standardized research protocols will help overcome some of the limitations of the clinical use of RNA markers (Schultze, Aschenbrenner, 2021; Chen et al., 2022; Wargodsky et al., 2022; Eldien et al., 2025; Shimansky et al., 2025). Furthermore, to fully elucidate the complex pathogenesis of severe COVID-19, it is also important to apply an integrated "omics" approach, which, in addition to transcriptomics, will include other methods for studying the immune response to SARS-CoV-2: proteomics, metabolomics, epigenetics, multiplex measurements of cytokines/chemokines, etc. Based on multi-omic technologies, new approaches can be developed for outcome prediction, prevention and treatment of severe COVID-19.

References

Abakushina E.V., Kuzmina E.G., Kovalenko E.I. The main characteristics of human natural killer cells. *Immunologiya = Immunology.* 2012;33(4):220-224 (in Russian)

Ahn J.Y., Park S., Yun Y.S., Song J.Y. Inhibition of type III TGF-β receptor aggravates lung fibrotic process. *Biomed Pharmacother*. 2010;64(7):472-476. doi 10.1016/j.biopha.2010.01.006

Alam M.S., Czajkowsky D.M. SARS-CoV-2 infection and oxidative stress: pathophysiological insight into thrombosis and therapeutic opportunities. *Cytokine Growth Factor Rev.* 2022;63:44-57. doi 10.1016/j.cytogfr.2021.11.001

Alqutami F., Senok A., Hachim M. COVID-19 transcriptomic atlas: a comprehensive analysis of COVID-19 related transcriptomics datasets. *Front Genet.* 2021;12:755222. doi 10.3389/fgene.2021.755222

Armignacco R., Carlier N., Jouinot A., Birtolo M.F., de Murat D., Tubach F., Hausfater P., … Beurton A., Goulet H., Manivet P., Bertherat J., Assié G.; COVIDeF group. Whole blood transcriptome signature predicts severe forms of COVID-19: results from the COVIDeF cohort study. *Funct Integr Genomics.* 2024;24(3):107. doi 10.1007/s10142-024-01359-2

Arunachalam P.S., Wimmers F., Mok C.K.P., Perera R.A.P.M., Scott M., Hagan T., Sigal N., … Maecker H.T., Khatri P., Rouphael N., Peiris M., Pulendran B. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science.* 2020; 369(6508):1210-1220. doi 10.1126/science.abc6261

Aschenbrenner A.C., Mouktaroudi M., Krämer B., Oestreich M., Antonakos N., Nuesch-Germano M., Gkizeli K., … Nattermann J., Koutsoukou A., Giamarellos-Bourboulis E.J., Ulas T. German COVID-19 Omics Initiative (DeCOI). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med.* 2021;13(1):7. doi 10.1186/s13073-020-00823-5

Auwul M.R., Rahman M.R., Gov E., Shahjaman M., Moni M.A. Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19. *Brief Bioinform*. 2021;22(5): bbab120. doi 10.1093/bib/bbab120

Baghela A., An A., Zhang P., Acton E., Gauthier J., Brunet-Ratnasingham E., Blimkie T., Freue G.C., Kaufmann D., Lee A.H.Y., Levesque R.C., Hancock R.E.W. Predicting severity in COVID-19 disease using sepsis blood gene expression signatures. *Sci Rep.* 2023;13(1):1247. doi 10.1038/s41598-023-28259-y

Bando S.Y., Bertonha F.B., Vieira S.E., de Oliveira D.B.L., Chalup V.N., Durigon E.L., Palmeira P., … Antonangelo L., Lauterbach G.D.P., Regalio F.A., Cesar R.M., Jr, Moreira-Filho C.A. Blood leukocyte transcriptional modules and differentially expressed genes associated with disease severity and age in COVID-19 patients. *Sci Rep.* 2023; 13(1):898. doi 10.1038/s41598-023-28227-6

Barnes B.J., Adrover J.M., Baxter-Stoltzfus A., Borczuk A., Cools-Lartigue J., Crawford J.M., Daßler-Plenker J., Guerci P., Huynh C., Knight J.S. Targeting potential drivers of COVID-19: neutrophil extracellular traps. *J Exp Med.* 2020;217:e20200652. doi 10.1084/jem.20200652

Berlin D.A., Gulick R.M., Martinez F.J. Severe Covid-19. *N Engl J Med.* 2020;383(25):2451-2460. doi 10.1056/NEJMcp2009575

Blanco-Melo D., Nilsson-Payant B.E., Liu W.C., Uhl S., Hoagland D., Møller R., Jordan T.X., … Wang T.T., Schwartz R.E., Lim J.K., Albrecht R.A., ten Oever B.R. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*. 2020;181(5):1036-1045.e9. doi 10.1016/j.cell.2020.04.026

Bolevich S.B., Bolevich S.S. Complex mechanism of COVID-19 development. *Sechenov Med J*. 2020;11(2):50-61. doi 10.47093/2218-7332.2020.11.2.50-61 (in Russian)

Charlson M.E., Pompei P., Ales K.L., MacKenzie C.R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis*. 1987;40(5):373-383. doi 10.1016/0021-9681(87)90171-8

Chaussabel D., Pascual V., Banchereau J. Assessing the human immune system through blood transcriptomics. *BMC Biol.* 2010;8:84. doi 10.1186/1741-7007-8-84

Che Y., Jiang D., Zhang Y., Zhang J., Xu T., Sun Y., Fan J., … Ding J., Hu C., Huang Y., Zhang J., Yang K. Elevated ubiquitination contributes to protective immunity against severe SARS-CoV-2 infection. *Clin Transl Med*. 2022;12(12):e1103. doi 10.1002/ctm2.1103

Chen C.H., Lin S.W., Shen C.F., Hsieh K.S., Cheng C.M. Biomarkers during COVID-19: mechanisms of change and implications for patient outcomes. *Diagnostics* (*Basel*). 2022;12(2):509. doi 10.3390/diagnostics12020509

Chu W., Li X., Li C., Wan L., Shi H., Song X., Liu X., Chen X., Zhang C., Shan H., Lu Y., Yang B. TGFBR3, a potential negative regulator of TGF-β signaling, protects cardiac fibroblasts from hypoxia-induced apoptosis. *J Cell Physiol*. 2011;226(10):2586-2594. doi 10.1002/jcp.22604

Chua R.L., Lukassen S., Trump S., Hennig B.P., Wendisch D., Pott F., Debnath O., … Laudi S., Lehmann I., Conrad C., Sander L.E., Eils R. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat Biotechnol.* 2020; 38(8):970-979. doi 10.1038/s41587-020-0602-4

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*. 2022;185(5):916-938.e58. doi 10.1016/j.cell.2022.01.012

Delorey T.M., Ziegler C.G.K., Heimberg G., Normand R., Yang Y., Segerstolpe Å., Abbondanza D., … Li B., Shalek A.K., Villani A.C., Rozenblatt-Rosen O., Regev A. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature*. 2021;595(7865):107-113. doi 10.1038/s41586-021-03570-8

Deng Z., Fan T., Xiao C., Tian H., Zheng Y., Li C., He J. TGF-β signaling in health, disease, and therapeutics. *Signal Transduct Target Ther*. 2024;9(1):61. doi 10.1038/s41392-024-01764-w

Derakhshani A., Hemmat N., Asadzadeh Z., Ghaseminia M., Shadbad M.A., Jadideslam G., Silvestris N., Racanelli V., Baradaran B. Arginase 1 (*Arg1*) as an up-regulated gene in COVID-19 patients: a promising marker in COVID-19 immunopathy. *J Clin Med*. 2021;10(5):1051. doi 10.3390/jcm10051051

de Seabra Rodrigues Dias I.R., Cao Z., Kwok H.F. Adamalysins in COVID-19 – potential mechanisms behind exacerbating the disease. *Biomed Pharmacother*. 2022;150:112970. doi 10.1016/j.biopha.2022.112970

Desterke C., Turhan A.G., Bennaceur-Griscelli A., Griscelli F. PPARγ cistrome repression during activation of lung monocyte-macrophages in severe COVID-19. *iScience*. 2020;23(10):101611. doi 10.1016/j.isci.2020.101611

Eldien H.M.S., Almaeen A.H., El Fath A.A., Taha A.E., Ahmed R., Elfadil H., Hetta H.F. Unlocking the potential of RNA sequencing in COVID-19: toward accurate diagnosis and personalized medicine. *Diagnostics* (*Basel*). 2025;15(2):229. doi 10.3390/diagnostics15020229

Ershov A.V., Surova V.D., Dolgikh V.T., Dolgikh T.I. Cytokine storm in the novel coronavirus infection and methods of its correction. *Antibiotiki i Khimioterapiya = Antibiotics and Chemotherapy*. 2020;65(11-12):27-37. doi 10.37489/0235-2990-2020-65-11-12-27-37 (in Russian)

Fabregat A., Sidiropoulos K., Viteri G., Forner O., Marin-Garcia P., Arnau V., D'Eustachio P., Stein L., Hermjakob H. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*. 2017;18(1):142. doi 10.1186/s12859-017-1559-2

Fang C., Ma Y. Peripheral blood genes crosstalk between COVID-19 and sepsis. *Int J Mol Sci*. 2023;24(3):2591. doi 10.3390/ijms24032591

Ferreira-Gomes M., Kruglov A., Durek P., Heinrich F., Tizian C., Heinz G.A., Pascual-Reguant A., … Radbruch H., Witkowski M., Melchers F., Radbruch A., Mashreghi M.F. SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself. *Nat Commun*. 2021;12(1):1961. doi 10.1038/s41467-021-22210-3

Filbin M.R., Mehta A., Schneider A.M., Kays K.R., Guess J.R., Gentili M., Fenyves B.G., … Parry B.A., Villani A.C., Sade-Feldman M., Hacohen N., Goldberg M.B. Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell Rep Med.* 2021;2(5):100287. doi 10.1016/j.xcrm.2021.100287

Gadotti A.C., Lipinski A.L., Vasconcellos F.T., Marqueze L.F., Cunha E.B., Campos A.C., Oliveira C.F., Amaral A.N., Baena C.P., Telles J.P., Tuon F.F., Pinho R.A. Susceptibility of the patients infected with Sars-Cov2 to oxidative stress and possible interplay with severity of the disease. *Free Radic Biol Med*. 2021;165:184-190. doi 10.1016/j.freeradbiomed.2021.01.044

Guo C., Li B., Ma H., Wang X., Cai P., Yu Q., Zhu L., … Weng J., Wei H., Jin T., Lin J., Qu K. Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat Commun*. 2020;11(1):3924. doi 10.1038/s41467-020-17834-w

Hadjadj J., Yatim N., Barnabei L., Corneau A., Boussier J., Smith N., Péré H., … Fischer A., Duffy D., Rieux-Laucat F., Kernéis S., Terrier B. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science*. 2020;369(6504):718-724. doi 10.1126/science.abc6027

Hadzega D., Babisova K., Hyblova M., Janostiakova N., Sabaka P., Janega P., Minarik G. Analysis of transcriptomics data from COVID-19 patients: a pilot research. *Folia Microbiol* (*Praha*). 2024;69(1):155-164. doi 10.1007/s12223-024-01130-x

Hasankhani A., Bahrami A., Tavakoli-Far B., Iranshahi S., Ghaemi F., Akbarizadeh M.R., Amin A.H., Abedi Kiasari B., Mohammadzadeh Shabestari A. Corrigendum: the role of peroxisome proliferator-activated receptors in the modulation of hyperinflammation induced by SARS-CoV-2 infection: a perspective for COVID-19 therapy. *Front Immunol*. 2024;15:1422261. doi 10.3389/fimmu.2024.1422261

Hegenbarth J.C., Lezzoche G., De Windt L.J., Stoll M. Perspectives on bulk-tissue RNA sequencing and single-cell RNA sequencing for cardiac transcriptomics. *Front Mol Med*. 2022;2:839338. doi 10.3389/fmmed.2022.839338

Huang C., Wang Y., Li X., Ren L., Zhao J., Hu Y., Zhang L., … Jiang R., Gao Z., Jin Q., Wang J., Cao B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. doi 10.1016/S0140-6736(20)30183-5

Huang W., Wang D., Yao Y.F. Understanding the pathogenesis of infectious diseases by single-cell RNA sequencing. *Microb Cell*. 2021;8(9):208-222. doi 10.15698/mic2021.09.759

Hughes T.D., Subramanian A., Chakraborty R., Cotton S.A., Del Pilar Giraldo Herrera M., Huang Y., Lambert N., Pinto M.D., Rahmani A.M., Sierra C.J., Downs C.A. The effect of SARS-CoV-2 variant on respiratory features and mortality. *Sci Rep.* 2023;13(1):4503. doi 10.1038/s41598-023-31761-y

Jackson H., Rivero Calle I., Broderick C., Habgood-Coote D., D'Souza G., Nichols S., Vito O., … Levin M., Martinon-Torres F., Kaforou M.; PERFORM consortium; GEN-COVID (www.gencovid.eu) study group. Characterisation of the blood RNA host response underpinning severity in COVID-19 patients. *Sci Rep*. 2022;12(1):12216. doi 10.1038/s41598-022-15547-2

Jansen J., Reimer K.C., Nagai J.S., Varghese F.S., Overheul G.J., de Beer M., Roverts R., … COVID Moonshot consortium; van Rij R.P., Costa I.G., Schneider R.K., Smeets B., Kramann R. SARS-CoV-2 infects the human kidney and drives fibrosis in kidney organoids. *Cell Stem Cell*. 2022;29(2):217-231.e8. doi 10.1016/j.stem.2021.12.010

Jeong K., Kim Y., Jeon J., Kim K. Subtyping of COVID-19 samples based on cell-cell interaction in single cell transcriptomes. *Sci Rep.* 2023;13(1):19629. doi 10.1038/s41598-023-46350-2

Jovic D., Liang X., Zeng H., Lin L., Xu F., Luo Y. Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med*. 2022;12(3):e694. doi 10.1002/ctm2.694

Karu N., Kindt A., van Gammeren A.J., Ermens A.A.M., Harms A.C., Portengen L., Vermeulen R.C.H., Dik W.A., Langerak A.W., van der Velden V.H.J., Hankemeier T. Severe COVID-19 is characterised by perturbations in plasma amines correlated with immune response markers, and linked to inflammation and oxidative stress. *Metabolites*. 2022;12(7):618. doi 10.3390/metabo12070618

Kesika P., Thangaleela S., Sisubalan N., Radha A., Sivamaruthi B.S., Chaiyasut C. The role of the nuclear factor-kappa B (NF-κB) pathway in SARS-CoV-2 infection. *Pathogens*. 2024;13(2):164. doi 10.3390/pathogens13020164

Khan A., Mathelier A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*. 2017;18(1):287. doi 10.1186/s12859-017-1708-7

Krämer B., Knoll R., Bonaguro L., ToVinh M., Raabe J., Astaburuaga-García R., Schulte-Schrepping J., … Sawitzki B.; Deutsche COVID-19 OMICS Initiative (DeCOI); Aschenbrenner A.C., Schultze J.L., Nattermann J. Early IFN-α signatures and persistent dysfunction are distinguishing features of NK cells in severe COVID-19. *Immunity*. 2021;54(11):2650-2669.e14. doi 10.1016/j.immuni.2021.09.002

Labarrere C.A., Kassab G.S. Glutathione deficiency in the pathogenesis of SARS-CoV-2 infection and its effects upon the host immune response in severe COVID-19 disease. *Front Microbiol*. 2022;13: 979719. doi 10.3389/fmicb.2022.979719

Laine L., Skön M., Väisänen E., Julkunen I., Österlund P. SARS-CoV-2 variants Alpha, Beta, Delta and Omicron show a slower host cell interferon response compared to an early pandemic variant. *Front Immunol*. 2022;13:1016108. doi 10.3389/fimmu.2022.1016108

Lee H.J., Georgiadou A., Walther M., Nwakanma D., Stewart L.B., Levin M., Otto T.D., Conway D.J., Coin L.J., Cunnington A.J. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Sci Transl Med*. 2018; 10(447):eaar3619. doi 10.1126/scitranslmed.aar3619

Lee J.S., Park S., Jeong H.W., Ahn J.Y., Choi S.J., Lee H., Choi B., … Park S.H., Choi J.Y., Kim S.H., Jung I., Shin E.C. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol*. 2020; 5(49):eabd1554. doi 10.1126/sciimmunol.abd1554

Lee M.J., Blish C.A. Defining the role of natural killer cells in COVID-19. *Nat Immunol*. 2023;24(10):1628-1638. doi 10.1038/s41590-023-01560-8

Lei H. Hypoxia and activation of neutrophil degranulation-related genes in the peripheral blood of COVID-19 patients. *Viruses*. 2024;16(2): 201. doi 10.3390/v16020201

Lévy Y., Wiedemann A., Hejblum B.P., Durand M., Lefebvre C., Surénaud M., Lacabaratz C., … Yazdanpanah Y., Pantaleo G., Hocini H., Thiébaut R.; French COVID cohort study group. CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. *iScience*. 2021;24(7):102711. doi 10.1016/j.isci.2021.102711

Liu C., Martins A.J., Lau W.W., Rachmaninoff N., Chen J., Imberti L., Mostaghimi D., … COVID Clinicians; Rossi C., Su H.C., Kuhns D.B., Cohen J.I., Notarangelo L.D., Tsang J.S. Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell.* 2021;184(7):1836-1857.e22. doi 10.1016/j.cell.2021.02.018

Liu X., Chen R., Li B., Zhang J., Liu P., Li B., Li F., Zhang W., Lyu X., Hu M. Oxidative stress indexes as biomarkers of the severity in COVID-19 patients. *Int J Med Sci*. 2024;21(15):3034-3045. doi 10.7150/ijms.102879

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

Maurya R., Swaminathan A., Shamim U., Arora S., Mishra P., Raina A., Ravi V., Tarai B., Budhiraja S., Pandey R. Co-evolution of SARS-CoV-2 variants and host immune response trajectories underlie COVID-19 pandemic to epidemic transition. *iScience.* 2023;26(12): 108336. doi 10.1016/j.isci.2023.108336

McClain M.T., Zhbannikov I., Satterwhite L.L., Henao R., Giroux N.S., Ding S., Burke T.W., Tsalik E.L., Nix C., Balcazar J.P., Petzold E.A., Shen X., Woods C.W. Epigenetic and transcriptional responses in circulating leukocytes are associated with future decompensation during SARS-CoV-2 infection. *iScience*. 2023;27(1):108288. doi 10.1016/j.isci.2023.108288

Meizlish M.L., Pine A.B., Bishai J.D., Goshua G., Nadelmann E.R., Simonov M., Chang C.H., … Halene S., Damsky W., van Dijk D., Lee A.I., Chun H.J. A neutrophil activation signature predicts critical illness and mortality in COVID-19. *Blood Adv*. 2021;5(5):1164-1177. doi 10.1182/bloodadvances.2020003568

Menke A., Rex-Haffner M., Klengel T., Binder E.B., Mehta D. Peripheral blood gene expression: it all boils down to the RNA collection tubes. *BMC Res Notes*. 2012;5:1. doi 10.1186/1756-0500-5-1

Morselli Gysi D., do Valle Í., Zitnik M., Ameli A., Gan X., Varol O., Ghiassian S.D., Patten J.J., Davey R.A., Loscalzo J., Barabási A.L. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc Natl Acad Sci USA*. 2021;118(19): e2025581118. doi 10.1073/pnas.2025581118

Nasonov E.L., Avdeeva A.S. Interleukin 18 in immune-mediated rheumatic diseases and COVID-19. *Rheumatology Science and Practice*. 2022;60(2):195-204. doi 10.47360/1995-4484-2022-195-204 (in Russian)

Nassir N., Tambi R., Bankapur A., Al Heialy S., Karuvantevida N., Khansaheb H.H., Zehra B., … Yaseen Hachim M., Casanova J.L., Berdiev B.K., Alsheikh-Ali A., Uddin M. Single-cell transcriptome identifies *FCGR3B* upregulated subtype of alveolar macrophages in patients with critical COVID-19. *iScience*. 2021;24(9):103030. doi 10.1016/j.isci.2021.103030

Nédélec Y., Sanz J., Baharian G., Szpiech Z.A., Pacis A., Dumaine A., Grenier J.C., … Hernandez R.D., Pique-Regi R., Tung J., Yotova V., Barreiro L.B. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*. 2016; 167(3):657-669.e21. doi 10.1016/j.cell.2016.09.025

Oh K.K., Adnan M., Cho D.H. Network pharmacology approach to decipher signaling pathways associated with target proteins of NSAIDs against COVID-19. *Sci Rep*. 2021;11(1):9606. doi 10.1038/s41598-021-88313-5

Parikh V.N., Ioannidis A.G., Jimenez-Morales D., Gorzynski J.E., De Jong H.N., Liu X., Roque J., … Palacios J.A., Pinsky B.A., Bustamante C.D., Rivas M.A., Ashley E.A. Deconvoluting complex correlates of COVID-19 severity with a multi-omic pandemic tracking strategy. *Nat Commun*. 2022;13(1):5107. doi 10.1038/s41467-022-32397-8

Pei R., Feng J., Zhang Y., Sun H., Li L., Yang X., He J., … Wen K., Zhou H., Chen J., Rong Z., Chen X. Host metabolism dysregulation and cell tropism identification in human airway and alveolar organoids upon SARS-CoV-2 infection. *Protein Cell*. 2021;12(9):717-733. doi 10.1007/s13238-020-00811-w

Pincemail J., Cavalier E., Charlier C., Cheramy-Bien J.P., Brevers E., Courtois A., Fadeur M., Meziane S., Goff C.L., Misset B., Albert A., Defraigne J.O., Rousseau A.F. Oxidative stress status in COVID-19 patients hospitalized in intensive care unit for severe pneumonia. A pilot study. *Antioxidants* (*Basel*). 2021;10(2):257. doi 10.3390/antiox10020257

Polonikov A. Endogenous deficiency of glutathione as the most likely cause of serious manifestations and death in COVID-19 patients. *ACS Infect Dis*. 2020;6(7):1558-1562. doi 10.1021/acsinfecdis.0c00288

Prevention, Diagnosis and Treatment of Novel Coronavirus Infection (COVID-19). Temporary guidelines. Version 18. The Ministry of Health of the Russian Federation, 2023 (in Russian)

Randolph H.E., Aracena K.A., Lin Y.L., Mu Z., Barreiro L.B. Shaping immunity: the influence of natural selection on population immune diversity. *Immunol Rev*. 2024;323(1):227-240. doi 10.1111/imr.13329

Ren X., Wen W., Fan X., Hou W., Su B., Cai P., Li J., … Qu K., Wang X., Chen J., Jin R., Zhang Z. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell.* 2021;184(7):1895-1913.e19. doi 10.1016/j.cell.2021.01.053

Reusch N., De Domenico E., Bonaguro L., Schulte-Schrepping J., Baßler K., Schultze J.L., Aschenbrenner A.C. Neutrophils in COVID-19. *Front Immunol.* 2021;12:652470. doi 10.3389/fimmu.2021.652470

Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., Smyth G.K. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi 10.1093/nar/gkv007

Robinson M.D., McCarthy D.J., Smyth G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi 10.1093/bioinformatics/btp616

Roessler C., de Oliveira K.C.S., de Oliveira Portella A.X., Fortes P.C.N., Machado F.R., Araujo S.M., Prigol M., Lucio L.C., Benvegnú D.M., Ferreto L.E.D. Evaluation of oxidative stress level: reactive oxy-

gen species, reduced glutathione, and D-dimer in patients hospitalized due to COVID-19. *Redox Rep*. 2023;28(1):1-6. doi 10.1080/13510002.2023.2272384

Ryckman C., Vandal K., Rouleau P., Talbot M., Tessier P.A. Proinflammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and adhesion. *J Immunol*. 2003;170(6):3233-3242. doi 10.4049/jimmunol.170.6.3233

Saheb Sharif-Askari N., Saheb Sharif-Askari F., Mdkhana B., Hussain Alsayed H.A., Alsafar H., Alrais Z.F., Hamid Q., Halwani R. Upregulation of oxidative stress gene markers during SARS-COV-2 viral infection. *Free Radic Biol Med*. 2021;172:688-698. doi 10.1016/j.freeradbiomed.2021.06.018

Saichi M., Ladjemi M.Z., Korniotis S., Rousseau C., Ait Hamou Z., Massenet-Regad L., Amblard E., Noel F., Marie Y., Bouteiller D., Medvedovic J., Pène F., Soumelis V. Single-cell RNA sequencing of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity. *Nat Cell Biol*. 2021;23(5):538-551. doi 10.1038/s41556-021-00681-2

Schimke L.F., Marques A.H.C., Baiocchi G.C., de Souza Prado C.A., Fonseca D.L.M., Freire P.P., Rodrigues Plaça D., … Ulas T., Schultze J.L., Nakaya H.I., Jurisica I., Cabral-Marques O. Severe COVID-19 shares a common neutrophil activation signature with other acute inflammatory states. *Cells*. 2022;11(5):847. doi 10.3390/cells11050847

Schulte-Schrepping J., Reusch N., Paclik D., Baßler K., Schlickeiser S., Zhang B., Krämer B., … Li Y., Nattermann J., Sawitzki B., Saliba A.E., Sander L.E.; Deutsche COVID-19 OMICS Initiative (DeCOI). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*. 2020;182(6):1419-1440.e23. doi 10.1016/j.cell.2020.08.001

Schultze J.L., Aschenbrenner A.C. COVID-19 and the human innate immune system. *Cell*. 2021;184(7):1671-1692. doi 10.1016/j.cell.2021.02.029

Schultze J.L., Mass E., Schlitzer A. Emerging principles in myelopoiesis at homeostasis and during infection and inflammation. *Immunity.* 2019;50(2):288-301. doi 10.1016/j.immuni.2019.01.019

Shaymardanov A.M., Antonova O.A., Sokol A.D., Deinichenko K.A., Kazakova P.G., Milovanov M.M., Zakubansky A.V., … Keskinov A.A., Kraevoy S.A., Snigir E.A., Svetlichnyy D.V., Skvortsova V.I. Single-cell gene expression analysis revealed immune cell signatures of Delta COVID-19. *Cells*. 2022;11(19):2950. doi 10.3390/cells11192950

Shimansky V., Popov O., Kel A., Goryanin I., Klochkova T., Apalko S., Sushentseva N., Anisenkova A., Mosenko S., Shcherbak S. Analysis of the expression profile in COVID-19 patients in the Russian population considering disease severity, mortality, and cytokine storm. *Biomedicines*. 2025;13(4):863. doi 10.3390/biomedicines13040863

Silvin A., Chapuis N., Dunsmore G., Goubet A.G., Dubuisson A., Derosa L., Almire C., … André F., Zitvogel L., Ginhoux F., Fontenay M., Solary E. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Cell.* 2020;182(6):1401-1418.e18. doi 10.1016/j.cell.2020.08.002

Song F., Qian Y., Peng X., Li X., Xing P., Ye D., Lei H. The frontline of immune response in peripheral blood. *PLoS One*. 2017;12(8):e0182294. doi 10.1371/journal.pone.0182294

Su Y., Chen D., Yuan D., Lausted C., Choi J., Dai C.L., Voillet V., … Greenberg P.D., Gottardo R., Davis M.M., Goldman J.D., Heath J.R. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell*. 2020;183(6):1479-1495.e20. doi 10.1016/j.cell.2020.10.037

Tabassum T., Rahman A., Araf Y., Ullah M.A., Hosen M.J. Prospective selected biomarkers in COVID-19 diagnosis and treatment. *Biomark Med*. 2021;15(15):1435-1449. doi 10.2217/bmm-2021-0038

Tang H., Gao Y., Li Z., Miao Y., Huang Z., Liu X., Xie L., Li H., Wen W., Zheng Y., Su W. The noncoding and coding transcriptional landscape of the peripheral immune response in patients with COVID-19. *Clin Transl Med*. 2020;10(6):e200. doi 10.1002/ctm2.200

Tavassolifar M.J., Aghdaei H.A., Sadatpour O., Maleknia S., Fayazzadeh S., Mohebbi S.R., Montazer F., Rabbani A., Zali M.R., Izad M., Meyfour A. New insights into extracellular and intracellular redox status in COVID-19 patients. *Redox Biol*. 2023;59:102563. doi 10.1016/j.redox.2022.102563

Togami K., Yamaguchi K., Chono S., Tada H. Evaluation of permeability alteration and epithelial-mesenchymal transition induced by transforming growth factor-β_1 in A549, NCI-H441, and Calu-3 cells: development of an *in vitro* model of respiratory epithelial cells in idiopathic pulmonary fibrosis. *J Pharmacol Toxicol Methods*. 2017;86:19-27. doi 10.1016/j.vascn.2017.02.023

Uchasova E.G., Gruzdeva O.V., Dileva Yu.A., Karetnikova V.N. Interleukin 33 and fibrosis: pathogenesis updated. *Meditsinskaya Immunologiya = Medical Immunology*. 2018;20(4):477-484. doi 10.15789/1563-0625-2018-4-477-484 (in Russian)

Vaz de Paula C.B., de Azevedo M.L.V., Nagashima S., Martins A.P.C., Malaquias M.A.S., Miggiolaro A.F.R.D.S., da Silva Motta Júnior J., Avelino G., do Carmo L.A.P., Carstens L.B., de Noronha L. IL-4/IL-13 remodeling pathway of COVID-19 lung injury. *Sci Rep*. 2020;10(1):18689. doi 10.1038/s41598-020-75659-5

Vlasov I., Panteleeva A., Usenko T., Nikolaev M., Izumchenko A., Gavrilova E., Shlyk I., Miroshnikova V., Shadrina M., Polushin Y., Pchelina S., Slonimsky P. Transcriptomic profiles reveal downregulation of low-density lipoprotein particle receptor pathway activity in patients surviving severe COVID-19. *Cells*. 2021;10(12):3495. doi 10.3390/cells10123495

Wang F., Xia H., Yao S. Regulatory T cells are a double-edged sword in pulmonary fibrosis. *Int Immunopharmacol*. 2020;84:106443. doi 10.1016/j.intimp.2020.106443

Wang Q.S., Edahiro R., Namkoong H., Hasegawa T., Shirai Y., Sonehara K., Tanaka H., … Miyano S., Ogawa S., Kanai T., Fukunaga K., Okada Y. The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 task force. *Nat Commun*. 2022;13(1):4830. doi 10.1038/s41467-022-32276-2

Wang S., Song R., Wang Z., Jing Z., Wang S., Ma J. S100A8/A9 in inflammation. *Front Immunol.* 2018;9:1298. doi 10.3389/fimmu.2018.01298

Wang Y., Schughart K., Pelaia T.M., Chew T., Kim K., Karvunidis T., Knippenberg B., … Burcham J., McLean A.; PREDICT-19 consortium; Tang B., Shojaei M. Blood transcriptome responses in patients correlate with severity of COVID-19 disease. *Front Immunol*. 2023;13:1043219. doi 10.3389/fimmu.2022.1043219

Wargodsky R., Dela Cruz P., LaFleur J., Yamane D., Kim J.S., Benjenk I., Heinz E., … Farrar K., Toma I., Jordan T., Goldman J., McCaffrey T.A. RNA sequencing in COVID-19 patients identifies neutrophil activation biomarkers as a promising diagnostic platform for infections. *PLoS One*. 2022;17(1):e0261679. doi 10.1371/journal.pone.0261679

Wilk A.J., Lee M.J., Wei B., Parks B., Pi R., Martínez-Colón G.J., Ranganath T., … Holmes S., Rabinovitch M., Rogers A.J., Greenleaf W.J., Blish C.A. Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. *J Exp Med*. 2021;218(8):e20210582. doi 10.1084/jem.20210582

Witkowski M., Tizian C., Ferreira-Gomes M., Niemeyer D., Jones T.C., Heinrich F., Frischbutter S., … Durek P., Kruglov A., Radbruch A., Mashreghi M.F., Diefenbach A. Untimely TGFβ responses in COVID-19 limit antiviral functions of NK cells. *Nature*. 2021;600(7888):295-301. doi 10.1038/s41586-021-04142-6

Xiong Y., Liu Y., Cao L., Wang D., Guo M., Jiang A., Guo D., … Shi M., Liu Y., Zhou Y., Lan K., Chen Y. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. *Emerg Microbes Infect*. 2020;9(1):761-770. doi 10.1080/22221751.2020.1747363

Yao C., Bora S.A., Parimon T., Zaman T., Friedman O.A., Palatinus J.A., Surapaneni N.S., … Martins G.A., Stripp B.R., Gharib S.A., Goodridge H.S., Chen P. Cell-type-specific immune dysregulation in severely ill COVID-19 patients. *Cell Rep*. 2021;34(1):108590. doi 10.1016/j.celrep.2020.108590

Yoon H.S., Kim H.Y., Cho K.A., Kim Y.H., Woo S.Y., Kim H.S., Kang J.L., Ryu K.H., Park J.W. Procollagen C-endopeptidase enhancer 2 secreted by tonsil-derived mesenchymal stem cells increases the oxidative burst of promyelocytic HL-60 cells. *Biology* (*Basel*). 2022;11(2):255. doi 10.3390/biology11020255

Zhang J.Y., Wang X.M., Xing X., Xu Z., Zhang C., Song J.W., Fan X., … Zhang Y., Zumla A., Maeurer M., Bai F., Wang F.S. Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol*. 2020;21(9):1107-1118. doi 10.1038/s41590-020-0762-x

Zhang Y., Wang S., Xia H., Guo J., He K., Huang C., Luo R., Chen Y., Xu K., Gao H., Sheng J., Li L. Identification of monocytes associated with severe COVID-19 in the PBMCs of severely infected patients through single-cell transcriptome sequencing. *Engineering* (*Beijing*). 2022;17:161-169. doi 10.1016/j.eng.2021.05.009

Zizzo G., Cohen P.L. Imperfect storm: is interleukin-33 the Achilles heel of COVID-19? *Lancet Rheumatol*. 2020;2(12):e779-e790. doi 10.1016/S2665-9913(20)30340-4

Zurochka A.V., Kotlyarov A.N., Kuvaytsev M.V., Kvyatkovskaya S.V., Zurochka V.A., Ryabova L.V., Khaidukov S.V. Changes of HLA-DR antigen expression on monocytes in children and their clinical significance in sepsis. *Meditsinskaya Immunologiya = Medical Immunology*. 2008;10(4-5):379-388. doi 10.15789/1563-0625-2008-4-5-379-388 (in Russian)

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

Received July 16, 2025. Revised December 5, 2025. Accepted December 5, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):117-125

doi 10.18699/vjgb-26-12

Single nucleotide polymorphisms are typical for tick-borne encephalitis and West Nile viruses during triple natural mixed infections in Blyth's reed warbler (*Acrocephalus dumetorum*)

E.P. Ponomareva [1], V.A. Ternovoi [1], E.V. Protopopova [1], N.L. Tupota [1], V.B. Loktev [1, 2]

[1] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk Region, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
ponomareva_ep@vector.nsc.ru

Abstract. Wild bird species contribute significantly to the rapid geographic dissemination of tick-borne encephalitis viruses (TBEV) and West Nile virus (WNV), facilitating the establishment of new natural foci of these *orthoflaviviruses*. However, the impact of TBEV and WNV population variability on shaping these foci, as well as the potential emergence of new human-pathogenic viral variants, remain underexplored. This study aimed to assess the genetic heterogeneity of TBEV (Siberian and Far Eastern genotypes) and WNV, isolated simultaneously from the tissues of a single garden reed warbler (*Acrocephalus dumetorum*) collected in the suburbs of Tomsk. The methods of viral strain isolation on various cell cultures were used in combination with a whole-genome analysis of isolates through traditional and high-throughput sequencing (NGS) methods. Consensus full-genome nucleotide sequences of the viruses were obtained by Sanger sequencing and compared with those obtained by NGS, with single nucleotide substitutions (single nucleotide variants, SNVs) accounting for 2 % or higher within the population under study. Our findings revealed single nucleotide polymorphisms (SNPs) associated with both synonymous and non-synonymous nucleotide substitutions, primarily located within the non-structural protein genes of TBEV and WNV. Notably, recombination events were not detected in the genomes of isolated *orthoflaviviruses*. The WNV isolate, Tomsk/bird/2006/A4, and the TBEV isolates, PT12 and PT122, obtained from *A. dumetorum*, exhibited heterogeneous viral populations, with SNVs ranging in frequency from 1.75 to 19.88 % for WNV and from 2.08 to 23.73 % for TBEV. Most identified SNPs shared similar nucleotide substitutions in the genomes of already known strains of TBEV and WNV, suggesting that these SNVs could play a crucial role in viral adaptation and underscore the genetic and phenotypic diversity of these viruses in nature.

Key words: tick-borne encephalitis virus; West Nile virus; single nucleotide polymorphism; viral genome; high-throughput sequencing; garden warbler; *orthoflaviviruses*

For citation: Ponomareva E.P., Ternovoi V.A., Protopopova E.V., Tupota N.L., Loktev V.B. Single nucleotide polymorphisms are typical for tick-borne encephalitis and West Nile viruses during triple natural mixed infections in Blyth's reed warbler (*Acrocephalus dumetorum*). *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):117-125. doi 10.18699/vjgb-26-12

Funding. The study was conducted within the framework of the State assignments No. 7/21 and No. 9/21 of the Federal Budgetary Institution of Science – State Research Center of Virology and Biotechnology "Vector" of Rospotrebnadzor.

Acknowledgements. The authors express their sincere gratitude to Dr. Alexander N. Shvalov and Dr. Tamara P. Mikryukova for their assistance in conducting experiments, sequencing, and analyzing viral genomes.

Однонуклеотидные полиморфизмы в геномах вирусов клещевого энцефалита и Западного Нила при тройной природной инфекции у садовой камышовки (*Acrocephalus dumetorum*)

Е.П. Пономарева [1], В.А. Терновой [1], Е.В. Протопопова [1], Н.Л. Тупота [1], В.Б. Локтев [1, 2]

[1] Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р. п. Кольцово, Новосибирская область, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
ponomareva_ep@vector.nsc.ru

© Ponomareva E.P., Ternovoi V.A., Protopopova E.V., Tupota N.L., Loktev V.B., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Аннотация. Вовлечение различных видов диких птиц в формирование природных очагов вирусов клещевого энцефалита (ВКЭ) и Западного Нила (ВЗН) обеспечивает быстрое распространение этих ортофлавивирусов в различных географических районах и формирование новых природных очагов данных инфекций. Однако роль популяционной вариабельности ВКЭ и ВЗН в формировании природных очагов, возможность появления (селекции) новых вирусных вариантов, патогенных для человека, в этих природных очагах остаются еще недостаточно изученными. Цель настоящего исследования – оценить популяционную гетерогенность геномов ВКЭ сибирского и дальневосточного генотипов и ВЗН, которые были одновременно выделены из тканей одной особи садовой камышовки (*Acrocephalus dumetorum*), отловленной в природных ландшафтах пригорода Томска. С этой целью были использованы методы выделения вирусных штаммов на различных культурах клеток в сочетании с анализом полных вирусных геномов полученных изолятов, определенных методами традиционного и высокопроизводительного секвенирования (NGS). Консенсусные полногеномные нуклеотидные последовательности вирусов получали секвенированием по Сэнгеру и сравнивали с последовательностями, полученными методом NGS, с однонуклеотидными заменами (single nucleotide variant, SNV) 2 % и выше, в пределах изучаемой популяции. Обнаруженный однонуклеотидный полиморфизм (SNP) ассоциировался как с синонимичными, так и несинонимичными нуклеотидными заменами преимущественно локализующихся в генах неструктурных белков ВКЭ и ВЗН. При этом возможных рекомбинационных событий в геномах изолированных ортофлавивирусов обнаружить не удалось. Результаты показали, что изоляты ВЗН Tomsk/bird/2006/A4, ВКЭ PT12 и PT122, выделенные из *A. dumetorum*, представлены гетерогенными вирусными популяциями, в которых обнаруживаются множественные SNV, возникающие с частотой от 1.75 до 19.88 % для ВЗН и от 2.08 до 23.73 % для ВКЭ. Для большинства выявленных SNP найдены аналогичные нуклеотидные замены в геномах уже известных штаммов ВКЭ и ВЗН, что может свидетельствовать о важной роли этих SNV-спектров в вирусной адаптации и обеспечении генетического и фенотипического разнообразия этих вирусов в природе.

Ключевые слова: вирус клещевого энцефалита; вирус Западного Нила; однонуклеотидный полиморфизм; вирусный геном; высокопроизводительное секвенирование; садовая камышовка; ортофлавивирусы

Introduction

Tick-borne encephalitis virus (*Orthoflavivirus encephalitidis*) and West Nile virus (*Orthoflavivirus nilense*) belong to the *Flaviviridae* family, genus *Orthoflavivirus* (Current ICTV Taxonomy Release; Postler et al., 2023). Tick-borne encephalitis virus (TBEV) and West Nile virus (WNV) are capable of inducing severe human diseases, potentially resulting in significant damage to the central nervous system (Worku, 2023; Singh et al., 2024). These viruses establish natural foci, with human infection commonly resulting from mosquito (WNV) or tick (TBEV) bites. The natural foci of TBEV are characteristic of northern Eurasia, whereas WNV exhibits a near-global distribution (Pustijanac et al., 2023; Simonin, 2024). Co-circulation of these two *orthoflaviviruses* has been observed in southern Western Siberia (Ternovoi et al., 2004; Kononova et al., 2006).

TBEV and WNV are characterized by genetic diversity, with at least five main genotypes described for TBEV and at least nine genotypes for WNV (Dai et al., 2018; Kozlova et al., 2018; Simonin, 2024). The level of nucleotide differences between different genotypes can reach 18–20 %. Siberian and Far Eastern TBEV genotypes, as well as WNV genotype 1a, are prevalent within the Siberian region.

Modern high-throughput sequencing methods enable the detection of SNPs in relatively small viral populations. A study examining the presence of SNVs in populations of the EK-328 TBEV strain, which had been adapted to various cultivation conditions, and its cloned variants revealed that SNVs occur at a frequency of approximately 1 % in populations of TBEV laboratory strains (Litov et al., 2018). This finding allowed us to assume that minor SNVs of TBEV could facilitate microevolution and rapid adaptation of the viral population to environmental changes, even under laboratory conditions. This hypothesis was confirmed by the studies on the variability of the genome of the C11-13 TBEV strain of the Siberian genotype when cultivated in the laboratory (Ternovoi et al., 2024). The presence of single nucleotide polymorphism has also been described in the genomes of Zika, dengue, Japanese encephalitis viruses, and WNV (Kaiser et al., 2019; Zaráte et al., 2019; Borda et al., 2021). However, attempts to attenuate West Nile virus using SNPs specific to the Japanese encephalitis virus E protein gene yielded no positive results. A suggestion was made that each *orthoflavivirus* species possesses an unique SNP profile.

A previous study (Mikryukova et al., 2014; Moskvitina et al., 2014; Korobitsyn et al., 2021) showed the involvement of 42 out of 60 bird species studied in the circulation of TBEV and WNV in the Tomsk and Novosibirsk regions. Additionally, a significant percentage of both birds (up to 39 %) and ticks removed from the birds (2.13 %) exhibited evidence of mixed infections. Given the diversity of bird species that circulate TBEV and WNV, we should consider how these *orthoflaviviruses* adapt to different hosts. The involvement of various wild bird species in the formation of natural foci of TBEV and WNV was assumed to facilitate the rapid spread of these *orthoflaviviruses* across different geographic areas, leading to the establishment of new natural foci of these infections. Further research is needed to elucidate the impact of TBEV and WNV population variability on the formation of natural foci and the potential emergence (selection) of new human-pathogenic viral variants within these natural foci.

This study aimed to assess the population heterogeneity of the WNV and TBEV genomes found in the tissues of one garden reed warbler captured in the suburbs of Tomsk. To achieve this, we employed methods for isolating viral strains from various cell cultures, along with an analysis of complete

viral genomes from the obtained isolates. This analysis was conducted using both traditional and high-throughput sequencing techniques.

Materials and methods

Samples studied. A 10 % homogenate, combining spleen and liver tissue, was prepared from a garden reed warbler (*Acrocephalus dumetorum*) captured in the Tomsk region in 2006 (Mikryukova et al., 2014). *Orthoflavivirus* isolation was achieved using cells culture porcine embryo kidney (PEK) and the *Aedes albopictus* mosquito (C6/36) obtained from the cell culture collection of the State Research Center of Virology and Biotechnology "Vector". The cell cultures were grown in DMEM/F12 medium ("Vector", Russia) supplemented with 10 % fetal bovine serum (Gipco, USA) and 80 µg/ml gentamicin sulfate at 37 °C for PEK cells and 28 °C for C6/36 cells.

The infectious activity of the viral isolates was determined by the development of the cytopathogenic effect (CPE). For this purpose, 10^5 cells of PEK were seeded into 96-well culture microplates in a volume of 50 µl per well and infected with virus-containing material. The viruses were titrated in DMEM/F12 medium containing 2 % fetal bovine serum in a volume of 100 µl per well. Following a five-day period, the viral infectious titer was calculated as described previously (Svyatchenko et al., 2021).

Enzyme immunoassay. TBEV and WNV antigens were detected in the culture medium using mouse monoclonal antibodies 13F6 (against TBEV) and 9E2 (against WNV) as at antigen capture (Razumov et al., 2005; Shanshin et al., 2024). Bound antigens were detected using monoclonal antibodies 10H10 (TBEV) and 5H6 (WNV) labeled with biotin and streptavidin-peroxidase conjugate (ICN, USA), as described previously (Korobitsyn et al., 2021).

Sample preparation. Post-treatment with benzonase (Law et al., 2013), the total RNA was extracted using the Extract RNA reagent (Eurogen, Russia), according to the manufacturer's protocol. The first DNA strand was constructed using the MMLV RT kit (Eurogen, Russia), according to the manufacturer's instructions. PCR was performed using "BioMaster LR HS-PCR" (BioLabMix, Russia), with primers for the detection of TBEV and WNV RNA, respectively (Supplementary Materials 1 and 2)[1]. PCR mode (C1000 amplifier, Bio-Rad, USA) was as follows: 94 °C for 10 s, 58 °C for 20 s, and 72 °C for 30 s (40 cycles), followed by 72 °C for 7 minutes.

Electrophoretic analysis and isolation of viral DNA fragments from gel. Amplification products were analyzed in a 2 % agarose gel in TAE x1 buffer (40 mM Tris, 1 mM Na_2EDTA). The diaGene kit (Dia-M, Russia) was used to isolate the amplification products from the agarose gel.

Determination of nucleotide sequences of viral cDNA. Nucleotide sequences of amplification products were determined using an ABI 3130*xl* automated genetic analyzer (Applied Biosystems, USA) with the BigDye reagent kit. We also used the Terminator v 3.1 Cycle Sequencing Kit (Applied Biosystems, USA), according to the manufacturer's instructions. The alignment of nucleotide sequences was performed using the Lasergene 7 (DNASTAR) application.

First-strand cDNA synthesis for NGS was performed using the NEBNext® Ultra Directional (NEB) module. Second-strand synthesis was performed using the UMI Second Strand Synthesis Module for QuantSeq FWD (Illumina, Lexogen). Cutadapt (version 1.18) and SAMtools (version 0.1.18) were used to remove Illumina adapters and reformat the reads. The contigs were assembled *de novo* using the MIRA assembler (version 4.9.6).

The nucleotide sequences for comparison were taken from the GenBank database. Multiple nucleotide sequence alignments were performed using the AlignX application within the Vector NTI 11 software package (InforMax). The analysis of the obtained nucleotide sequences was performed using the Unipro UGENE v.1.30 and MEGA 7/10 programs (Kumar et al., 2018). The phylogenetic trees were calculated using the maximum likelihood method using 1,000 bootstrap replicates.

Results

This study utilized a combined liver and spleen sample from the garden warbler (*A. dumetorum*) caught in the Tomsk region in 2006 (Mikryukova et al., 2014). Figure 1 illustrates the process of isolating viral isolates from this sample using C6/36 and PEK cell cultures.

The first instance of a cytopathogenic effect, attributable to TBEV replication, appeared after just two passages. However, after seven additional passages, the antigen and genetic material of WNV and TBEV were detected in the sample. Passaging the infectious material in mosquito C6/36 cells allowed us to obtain a pure culture of WNV. Supplementary passages of identical PEK cell material resulted in the eradication of the WNV population and the establishment of a stable TBEV population.

The genomic sequencing of the samples allowed us to determine the nucleotide sequences for two TBEV isolates (Tomsk-PT12 and Tomsk-PT122) and one WNV isolate (Tomsk/bird/2006/A4). The phylogenetic analysis of whole-genome sequences revealed that isolate Tomsk-PT12 belonged to the Far Eastern genotype of TBEV, while isolate Tomsk-PT122 represented the Siberian genotype of TBEV. The Tomsk/bird/2006/A4 isolate was genotyped as a virus belonging to genotype Ia of WNV (Supplementary Material 3).

High-throughput sequencing analysis of nucleotide sequences identified SNVs within viral populations. Figure 2 presents the single nucleotide substitutions identified in the three viral populations, with a frequency threshold of 1.75 % or greater.

Our findings indicate the presence of heterogeneous viral populations in the WNV strains (Tomsk/bird/2006/A4) and the TBEV strains (Tomsk-PT12 and Tomsk-PT122) isolated from *A. dumetorum*. These populations contain multiple SNVs, with frequencies ranging from 1.75 to 19.88 % for WNV and from 2.08 to 23.73 % for TBEV (see the Table). The identified SNPs are mapped throughout the viral genome, both in the genes of structural viral proteins and in the genes of non-structural proteins. At the same time, isolate Tomsk-PT12 of the Far

[1] Supplementary Materials 1–4 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Pon_Engl_30_1.pdf

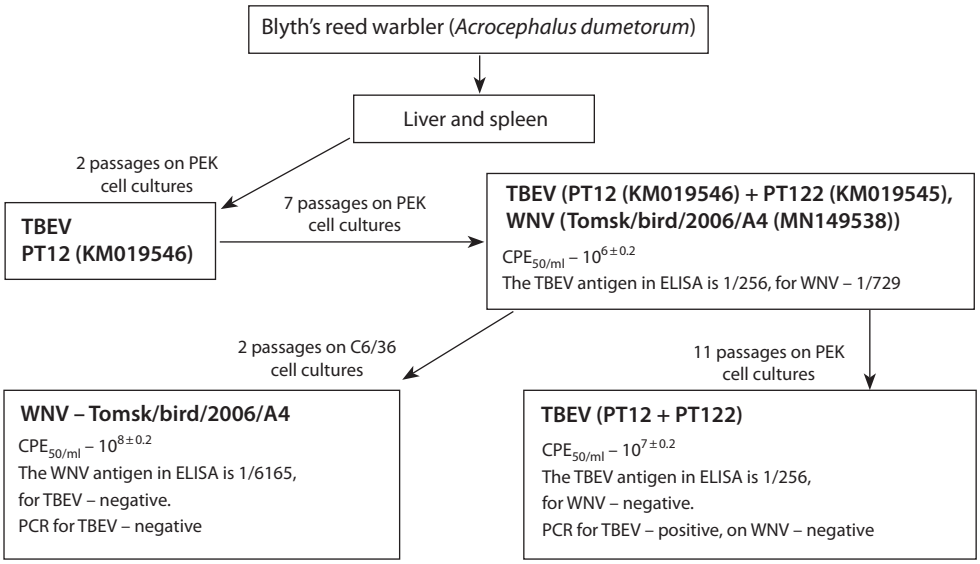


Fig. 1. Scheme for isolating TBEV and WNV isolates from the combined homogenate of the spleen and liver of the garden warbler on cell cultures.

As cell culture passages were performed, each block marked in the diagram was subjected to determination of infectious activity, determination of the presence of viral antigen in ELISA using monoclonal antibodies, PCR, and sequencing of the sample using Sanger and high-throughput sequencing methods, as described in the "Materials and methods" section.

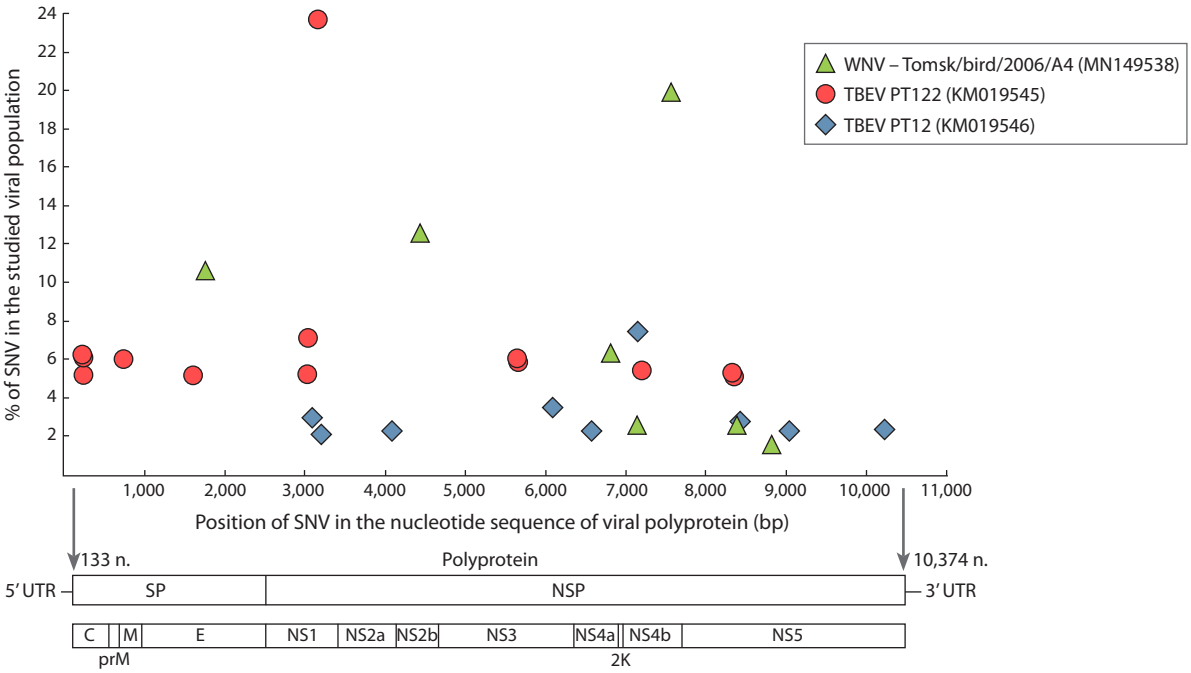


Fig. 2. Mapping of single nucleotide substitutions in the TBEV and WNV genomes using high-throughput sequencing.

Eastern genotype of TBEV was completely devoid of SNVs in the genomic region encoding the structural proteins. The most significant number of SNPs was detected in this region of the genome for the Siberian genotype of TBEV.

Of further interest is the observation that all 29 SNPs identified in the three *orthoflavivirus* isolates proved to be original. These results indicate that the detected SNV spectrum is characteristic and specific for the studied TBEV and WNV viral populations within a single infected host. This finding is indirectly confirmed by previously obtained data on the mismatch of SNVs for Japanese encephalitis virus and WNV (Kaiser et al., 2019). Whole-genome sequencing of these viruses did not detect evidence of recombination across extensive fragments of the isolated TBEV and WNV strains

Single nucleotide substitutions in the genomic RNA of WNV and TBEV during associated infection in *A. dumetorum*

No.	Position in the nucleotide sequence viral polyprotein activity (bp)	Predominant nucleotide/ SNV in the studied viral population	SNP frequency, %	Prototype strains of tick-borne encephalitis and West Nile viruses (GenBank accession numbers) with identified similar nucleotide substitutions in the genomes	Gene	Amino acid substitutions	
TBEV Tomsk-PT12 (KM019546), Far Eastern genotype							
1	3,093	C/T	2.95	JF819648 JN229223 JN003205 and others	*NS1*	Y_{1031}	**Y**
2	3,201	T/C	2.08	KF880803 KU761572 EU816451 and others		D_{1067}	**D**
3	4,083	T/C	2.23	JQ825154 EF469661 KY069125 and others	*NS2a*	S_{1361}	**S**
4	6,087	C/ T	3.48	JN003205 AF069066 KP716978 and others	*NS3*	T_{2029}	**T**
5	6,572	T/C	2.23	JF819648 JN003205 JN003206	*NS4a*	M_{2191}	**T**
6	7,150	G/T	7.41	No	*NS4b*	V_{2383}	**L**
7	8,430	C/T	2.76	JQ825147 JQ650523 JX534167 and others	*NS5*	H_{2810}	**H**
8	9,038	G/A	2.27	KJ633033 DQ862460 KF880803 and others		G_{3013}	**E**
9	10,217	A/C	2.33	No		E_{3406}	**A**
TBEV Tomsk-PT122 (KM019545), Siberian genotype							
1	218	T/C	6.32	EU816451 KT321430 JN003208 and others	*C*	V_{73}	**A**
2	231	A/G	6.10	JQ825155 DQ862460 OP902895 and others		K_{77}	**K**
3	234	C/T	5.18	KJ626343 JX498939 JN003205 and others		I_{78}	**I**
4	732	C/T	6.01	JQ825147 FJ402886 JF819648 and others	*preM*	N_{244}	**N**
5	1,605	T/C	5.17	JN003207 EF469661 AF069066 JQ650523 and others	*E*	Y_{534}	**Y**
6	3,027	C/T	5.22	KJ000002 KJ739729 KC414090 FJ572210 and others	*NS1*	D_{1009}	**D**
7	3,031	G/C	7.09	FJ968751		A_{1011}	**P**
8	3,159	A/C	23.73	U27495 KP716978 KJ922516 KY069125 and others		T_{1053}	**I**

Table (end)

No.	Position in the nucleotide sequence viral polyprotein activity (bp)	Predominant nucleotide/ SNV in the studied viral population	SNP frequency, %	Prototype strains of tick-borne encephalitis and West Nile viruses (GenBank accession numbers) with identified similar nucleotide substitutions in the genomes	Gene	Amino acid substitutions	
TBEV Tomsk-PT122 (KM019545), Siberian genotype							
9	5,643	C/T	6.05	KJ633033 EU816451 JQ825147 and others	*N* *NS3*	N_{1881}	N
10	5,658	G/A	5.86	KJ922516 KC414090 MF774565 and others		E_{1886}	E
11	7,194	C/T	5.42	MF774565 KJ633033	*NS4b*	V_{2498}	V
12	8,220	C/T	5.20	JN003208 KF826914 JQ654701 and others	*NS5*	N_{2740}	N
13	8,238	C/T	5.15	KU761572 JQ825155 KP716978 and others		N_{2746}	N
WNV Tomsk/bird/2006/A4 (MN149538), genotype Ia							
1	1,748	A/G	10.55	KX547363 KX547219 MH507756 and others	*E*	E_{583}	G
2	4,434	C/T	12.58	AF196835 EF657887 AF404754 and others	*NS2b*	C_{1478}	C
3	6,807	C/T	6.29	AY262283 DQ786572	*2k*	S_{2269}	S
	7,149	G/T	2.52	AF196835 EF657887 AF404754 and others	*NS4b*	V_{2383}	V
5	7,566	C/A	19.88	KJ958922 GQ851607 GQ851608 and others		D_{2522}	E
6	8,388	G/A	2.55	AY701413 JN858069 GQ851607 and others	*NS5*	G_{2796}	G
7	8,907	C/T	1.75	FJ766332		R_{2969}	R

Note. Non-synonymous nucleotide substitutions are highlighted with a dark gray background; a light gray background indicates the absence of prototype strains with identified similar nucleotide substitutions in the genomes. The amino acid designations are given under the generally accepted international single-letter code.

found in a single wild bird specimen. The unique SNV pattern for all three isolates confirms the absence of recombination during the simultaneous circulation of three viruses in a single host.

The Siberian genotype TBEV isolate Tomsk-PT122 exhibited the highest number of detected SNPs (13), while the lowest number (7) was observed in WNV. More than 60 % of SNPs were mapped in genes encoding non-structural viral proteins. The mean SNP frequency, relative to the consensus genome, was 3.1 % for the Tomsk-PT12 isolate of TBEV (Far Eastern genotype), 7.1 % for the Tomsk-PT122 isolate of TBEV (Siberian genotype), and 8.0 % for the WNV genotype 1a isolate (see the Table).

Notably, the identified SNPs exhibited analogous confirmed substitutions within the genomes of the established TBEV and WNV strains. Only theSNP_{7150}NS4b and SNP_{10217}NS5

proteins mapped for the Far Eastern genotype of TBEV had no analogs among the known TBEV genomes. Hence, the presence of identical nucleotide substitutions across the genomes of previously described viral isolates indicates a non-random nature of SNV occurrence, potentially contributing to the genetic variability of these flaviviruses.

Discussion

It is known that isolates of various RNA-containing viruses typically represent a heterogeneous population of closely related variants, often referred to as quasispecies (Eigen et al., 1988; Holland et al., 1992; Domingo, Holland, 1997; Domingo et al., 2012; Karbowiak et al., 2016). Viral populations are characterized by a high number of viral particles, with sizes potentially reaching 10^{10}–10^{12} virions or higher per infected macroorganism (Marí Saéz et al., 2015; Diallo et al., 2016; Thorson et al., 2016; Domingo et al., 2021). Moreover, a minimal quantity of viral particles is sufficient for infection of a susceptible organism. Further replication of RNA-containing viruses in the host organism ensures the formation of a heterogeneous viral population.

Predominantly, the extensive variability of viral RNA genomes arises from the high level of RNA polymerase errors and nucleotide substitutions, deficiencies in error correction mechanisms, and the action of selective forces exerted during replication within the host.

Previous findings reported the circulation of WNV and TBEV in the urban and suburban biotopes of Tomsk (Moskvitina et al., 2008; Chausov et al., 2009). Notably, the natural foci of these infections in this region involved over 42 diverse wild bird species (Mikryukova et al., 2014; Moskvitina et al., 2014; Korobitsyn et al., 2021). At the same time, the genetic markers of both TBEV and WNV were identified in 1.7 % of the examined samples. The literature also reports cases of mixed infections caused by different TBEV genotypes (Bezrukova et al., 2008; Kovalev et al., 2008; Kozlova et al., 2010; Pogodina et al., 2012; Bezrukov et al., 2015). Alternatively, the prevalence of diverse TBEV subtypes identified in ixodid ticks ranges from 4.4 to 15 %.

The level of genetic differences between the European, Siberian, and Far Eastern genotypes of TBEV can reach 18–20 % (Ternovoi et al., 2007). TBEV and WNV belonging to different types of *orthoflaviviruses* demonstrate significant genomic heterogeneity, with sequence divergence frequently exceeding 28–32 %. The distinct genetic profiles of these two orthoflavivirus types enabled the identification of at least five main genotypes for TBEV and nine for WNV (Dai et al., 2018; Kozlova et al., 2018; Simonin, 2024). The various genotypes of these viruses tend to be associated with distinct natural foci situated in different parts of the world.

Recently accumulated data suggest a possible joint circulation and widespread distribution of various *orthoflaviviruses* in new regions. This trend is corroborated by the circulation of TBEV and WNV within the urban and suburban biotopes of Tomsk. The discovery of the joint circulation of several *orthofaviviruses* in an individual garden reed warbler clearly demonstrates the distinctive epidemiological features of these viruses in southern West Siberia.

The identification of a whole set of SNVs in populations of the isolates (Tomsk/bird/2006/A4 WNV, Tomsk-PT12, and Tomsk-PT122 TBEV) in the tissues of one wild bird indicates a pronounced heterogeneity of these viral populations. The possibility of random nucleotide substitutions in genomic RNA cannot be excluded. However, most of the identified SNPs share similar nucleotide substitutions with those found in the genomes of already-known WNV and TBEV. The genetic diversity of these flaviviruses could be limited by the number of SNPs, which predetermines the scope of genetic variation of these *orthoflaviviruses* within natural foci. Mapping of the original SNV patterns characteristic of each of the three different isolates of *orthoflaviviruses* under study, which were circulating simultaneously in a single infected host, was performed. The results demonstrate that the viral population heterogeneity is maintained within the Siberian and Far Eastern genotypes of TBEV and the genotype Ia of WNV via independent mechanisms. It is worth noting that we failed to identify the signs of new recombination events during the assembly and analysis of the consensus genomes of these viruses (Supplementary Material 4).

Our findings reveal pronounced heterogeneity of the flavivirus genomic RNA population, a characteristic preserved during triple flavivirus coinfection of a single host. The presence of multiple SNVs within even a limited viral population of *orthoflaviviruses* is likely to serve as a significant mechanism for the formation of new TBEV and WNV genovariants, even during the replication of these viruses in the tissues of the infected host.

Conclusion

The evaluation of potential population heterogeneity of tick-borne encephalitis and West Nile viruses using metagenomic analysis methods has revealed multiple SNVs in the viral populations of all three *orthoflavivirus* isolates studied, which were isolated from tissues of a single garden reed warbler. In the TBEV and WNV isolates under investigation, SNV detection rates were observed to be from 1.75 to 23.73 %.

The identified SNPs were associated with synonymous and non-synonymous single-nucleotide substitutions, both predominantly localized in the genes of viral non-structural proteins. Tomsk-PT122 TBEV (Siberian genotype) exhibited the highest SNP count (13), with a notable prevalence (23.73 % for SNP_{3159}NS1) within the viral population. In the Tomsk-PT12 strain of TBEV of the Far Eastern genotype, SNP_{7150}NS4b was found to be the most prevalent, with the amino acid substitution $V_{2383}L$ occurring at a frequency of 7.41 %. This has not been previously described for other known strains of TBEV, as well as SNP_{10271}NS5. The smallest number of SNVs (7) was found in WNV. However, SNP_{1748}E, SNP_{4434}NS3b, and SVP_{7566}NS4b had a high frequency of occurrence, from 10.55 to 19.88 % in the studied viral population.

Twenty-nine SNPs were identified in the TBEV and WNV strains under study, leading to nine viral variants with amino acid substitutions. Taken together, the results obtained highlight the significance of SNPs in ensuring the genetic diversity of *orthoflaviviruses.*

References

Bezrukova (Gamova) E.G., Pogodina V.V., Levina L.S., Karan L.S., Malenko G.V. The research of different genotypes of TBE strain virus isolated from patients with chronic course of the disease. *Medicina v Kuzbasse = Medicine in Kuzbass*. 2008;S5:21-28 (in Russian)

Borda V., da Silva Francisco Junior R., Carvalho J.B., Morais G.L., Duque Rossi Á., Pezzuto P., Azevedo G.S., … Tanuri A., Stratakis C.A., Aguiar R.S., Cardoso C.C., Vasconcelos A.T.R. Whole-exome sequencing reveals insights into genetic susceptibility to Congenital Zika Syndrome. *PLoS Negl Trop Dis*. 2021;15(6): e0009507. doi 10.1371/journal.pntd.0009507

Chausov E.V., Ternovoi V.A., Protopopova E.V., Konovalova S.N., Kononova J.V., Pershikova N.L., Moskovitina N.S., Romanenko V.N., Ivanova N.V., Bol'shakova N.P., Moskvitin S.S., Korobitsyn I.G., Gashkov S.I., Tiutenkov O.I., Kuranova V.N., Kravchenko L.B., Suchkova N.G., Agulova L.P., Loktev V.B. Genetic diversity of ixodid tick-borne pathogens in Tomsk City and suburbs. *Parazitologiya*. 2009;43(5):374-388 (in Russian)

Current ICTV Taxonomy Release. Taxonomy Browser. Ch. Family: Flaviviridae. Available at: https://ictv.global/report/chapter/flaviviridae/flaviviridae/orthoflavivirus

Dai X., Shang G., Lu S., Yang J., Xu J. A new subtype of eastern tick-borne encephalitis virus discovered in Qinghai-Tibet Plateau, China. *Emerg Microbes Infect*. 2018;7:74. doi 10.1038/s41426-018-0081-6

Diallo B., Sissoko D., Loman N.J., Bah H.A., Bah H., Worrell M.C., Conde L.S., … Formenty P., Keïta S., Günther S., Rambaut A., Duraffour S. Resurgence of Ebola virus disease in Guinea linked to a survivor with virus persistence in seminal fluid for more than 500 days. *Clin Infect Dis*. 2016;63(10):1353-1356. doi 10.1093/cid/ciw601

Domingo E., Holland J.J. RNA virus mutations and fitness for survival. *Annu Rev Microbiol.* 1997;51:151-178. doi 10.1146/annurev.micro.51.1.151

Domingo E., Sheldon J., Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev.* 2012;76(2):159-216. doi 10.1128/mmbr.05023-11

Domingo E., García-Crespo C., Perales C. Historical perspective on the discovery of the quasispecies concept. *Annu Rev Virol*. 2021;8(1): 51-72. doi 10.1146/annurev-virology-091919-105900

Eigen M., McCaskill J., Schuster P. Molecular quasi-species. *J Phys Chem*. 1988;92(24):6881-6891. doi 10.1021/j100335a010

Holland J.J., De La Torre J.C., Steinhauer D.A. RNA-virus populations as quasispecies. In: Holland J.J. (Ed.) Genetic Diversity of RNA Viruses. Current Topics in Microbiology and Immunology. Vol. 176. Berlin; Heidelberg: Springer, 1992;176:1-20. doi 10.1007/978-3-642-77011-1_1

Kaiser J.A., Luo H., Widen S.G., Wood T.G., Huang C.Y., Wang T., Barrett A.D.T. Japanese encephalitis vaccine-specific envelope protein E138K mutation does not attenuate virulence of West Nile virus. *NPJ Vaccines*. 2019;4:50. doi 10.1038/s41541-019-0146-0

Karbowiak G., Biernat B., Werszko J., Rychlik L. The transstadial persistence of tick-borne encephalitis virus in *Dermacentor reticulatus* ticks in natural conditions. *Acta Parasitol*. 2016;61(1):201-203. doi 10.1515/ap-2016-0028

Kononova Yu.V., Ternovoi V.A., Shchelkanov M.Yu., Protopopova E.V., Zolotykh S.I., Yurlov A.K., Druziaka A.V., Slavskii A.A., Shestopalov A.M., L'vov D.K., Loktev V.B. West Nile virus genotyping among wild birds belonging to ground and tree-brush bird populations on the territories of the Baraba forest-steppe and Kulunda steppe (2003-2004). *Voprosy Virusologii = Problems of Virology*. 2006;51(4):19-23 (in Russian)

Korobitsyn I.G., Moskvitina N.S., Tyutenkov O.Y., Gashkov S.I., Kononova Y.V., Moskvitin S.S., Romanenko V.N., … Konovalova S.N., Tupota N.L., Sementsova A.O., Ternovoi V.A., Loktev V.B. Detection of tick-borne pathogens in wild birds and their ticks in Western Siberia and high level of their mismatch. *Folia Parasitol* (*Praha*). 2021;68:024. doi 10.14411/fp.2021.024

Kovalev S.Yu., Umpelova T.V., Snitkovskaya T.E., Kilyachina A.S., Romanenko V.V., Kokorev V.S., Glinskikh N.P. Molecular and epidemiological characteristics of tick-borne encephalitis virus in the Sverdlovsk region on the basis of genotype-specific RT-PCR. *Voprosy Virusologii = Problems of Virology.* 2008;53(2):27-31 (in Russian)

Kozlova I.V., Verkhozina M.M., Demina T.V., Dzhioev Yu.P., Doroshchenko E.K., Lisak O.V., Karan L.S., Koliasnikova N.M., Rar V.A., Fomenko N.V., Tkachev S.E., Bogomazova O.L., Borisov V.A., Tuvakov M.K., Zlobin V.I. Combined natural foci of tick-borne infections in Baikal region. *Epidemiologiya i Vakcinoprofilaktika = Epidemiology and Vaccinal Prevention*. 2010;4(53):40-46 (in Russian)

Kozlova I.V., Demina T.V., Tkachev S.E., Doroshchenko E.K., Lisak O.V., Verkhozina M.M., Karan L.S., … Savinova Y.S., Chernoivanova O.O., Ruzek D., Tikunova N.V., Zlobin V.I. Characteristics of the Baikal subtype of tick-borne encephalitis virus circulating in Eastern Siberia. *Acta Biomed Scientifica*. 2018;3(4):53-60. doi 10.29413/ABS.2018-3.4.9

Kumar S., Stecher G., Li M., Knyaz C., Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547-1549. doi 10.1093/molbev/msy096

Law J., Jovel J., Patterson J., Ford G., O'Keefe S., Wang W., Meng B., … Mitchell T., Jordan T., Carpenter E., Mason A.L., Wong G.K. Identification of hepatotropic viruses from plasma using deep sequencing: a next generation diagnostic tool. *PLoS One.* 2013;8(4):e60595. doi 10.1371/journal.pone.0060595

Litov A.G., Deviatkin A.A., Goptar I.A., Dedkov V.G., Gmyl A.P., Markelov M.L., Shipulin G.A., Karganova G.G. Evaluation of the population heterogeneity of TBEV laboratory variants using high-throughput sequencing. *J Gen Virol*. 2018;99(2):240-245. doi 10.1099/jgv.0.001003

Marí Saéz A., Weiss S., Nowak K., Lapeyre V., Zimmermann F., Düx A., Kühl H.S., … Fahr J., Borcher M., Gogarten J.F., Calvignac-Spencer S., Leendertz F.H. Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO Mol Med.* 2015;7(1):17-23. doi 10.15252/emmm.201404792

Mikryukova T.P., Moskvitina N.S., Kononova Y.V., Korobitsyn I.G., Kartashov M.Y., Tyuten Kov O.Y., Protopopova E.V., … Moskvitin S.S., Tupota N.L., Sementsova A.O., Ternovoi V.A., Loktev V.B. Surveillance of tick-borne encephalitis virus in wild birds and ticks in Tomsk city and its suburbs (Western Siberia). *Ticks Tick Borne Dis.* 2014;5(2):145-151. doi 10.1016/j.ttbdis.2013.10.004

Moskvitina N.S., Romanenko V.N., Ternovoi V.A., Ivanova N.V., Protopopova E.V., Kravchenko L.B., Kononova Yu.V., Kuranova V.N., Chausov E.V., Moskvitin S.S., Pershikova N.L., Gashkov S.I., Konovalova S.N., Bolshakova N.P., Loktev V.B. Detection of the West Nile virus and its genetic typing in ixodid ticks (Parasitiformes: Ixodidae) in Tomsk and its suburbs. *Parazitologiya*. 2008; 42(3):210-225 (in Russian)

Moskvitina N.S., Korobitsyn I.G., Tyuten'kov O.Yu., Gashkov S.I., Kononova Yu.V., Moskvitin S.S., Romanenko V.N., … Konovalova S.N., Tupota N.L., Sementsova A.O., Ternovoi V.A., Loktev V.B. The role of birds in the maintenance of tick borne infections in the Tomsk anthropurgic foci. *Biol Bull Russ Acad Sci*. 2014;41(4): 387-393. doi 10.1134/S1062359014040086

Pogodina V.V., Karan L.S., Kolyasnikova N.M., Gerasimov S.G., Levina L.S., Bochkova N.G., Andayev E.I., Trukhina A.G., Borisova T.I., Sidorova E.A., Nagibina O.A., Malenko G.V., Bezrukova E.G. Polytypic strains in the genofund of tick-borne encephalitis virus. *Voprosy Virusologii = Problems of Virology*. 2012;57(3):30-37 (in Russian)

Postler T.S., Beer M., Blitvich B.J., Bukh J., de Lamballerie X., Drexler J.F., Imrie A., Kapoor A., Karganova G.G., Lemey P., Lohmann V., Simmonds P., Smith D.B., Stapleton J.T., Kuhn J.H. Renaming of the genus *Flavivirus* to *Orthoflavivirus* and extension of binomial species names within the family *Flaviviridae. Arch Virol.* 2023;168(9):224. doi 10.1007/s00705-023-05835-1

Pustijanac E., Buršić M., Talapko J., Škrlec I., Meštrović T., Lišnjić D. Tick-borne encephalitis virus: a comprehensive review of transmission, pathogenesis, epidemiology, clinical manifestations, diagnosis, and prevention. *Microorganisms.* 2023;11(7):1634. doi 10.3390/microorganisms11071634

Razumov I.A., Kazachinskaia E.I., Ternovoi V.A., Protopopova E.V., Galkina I.V., Gromashevskii V.L., Prilipov A.G., Kachko A.V., Ivanova A.V., L'vov D.K., Loktev V.B. Neutralizing monoclonal antibodies against Russian strain of the West Nile virus. *Viral Immunol*. 2005;18(3):558-568. doi 10.1089/vim.2005.18.558

Shanshin D.V., Borisevich S.S., Shaprova O.N., Nesmeyanova V.S., Bondar A.A., Porozov Y.B., Khamitov E.M., Kolosova E.A., Shelemba A.A., Ushkalenko N.D., Protopopova E.V., Sergeev A.A., Loktev V.B., Shcherbakov D.N. Phage display revealed the complex structure of the epitope of the monoclonal antibody 10H10. *Int J Mol Sci.* 2024;25(19):10311. doi 10.3390/ijms251910311

Simonin Y. Circulation of West Nile virus and Usutu virus in Europe: overview and challenges. *Viruses.* 2024;16(4):599. doi 10.3390/v16040599

Singh P., Khatib M.N., Ballal S., Kaur M., Nathiya D., Sharma S., Siva Prasad G.V., … Lakhanpal S., Shabil M., Bushi G., Sah S., Serhan H.A. West Nile virus in a changing climate: epidemiology, pathology, advances in diagnosis and treatment, vaccine designing and control strategies, emerging public health challenges – a comprehensive review. *Emerg Microbes Infect.* 2024;30:2437244. doi 10.1080/22221751.2024.2437244

Svyatchenko V.A., Nikonov S.D., Mayorov A.P., Gelfond M.L., Loktev V.B. Antiviral photodynamic therapy: inactivation and inhibition of SARS-CoV-2 *in vitro* using methylene blue and Radachlorin. *Photodiagnosis Photodyn Ther.* 2021;33:102-112. doi 10.1016/j.pdpdt.2020.102112

Ternovoi V.A., Shchelkanov M.Yu., Shestopalov A.M., Aristova V.A., Protopopova E.V., Gromashevsky V.L., Druzyaka A.V., Slavsky A.A., Zolotykh S.I., Loktev V.B., Lvov D.K. Detection of West Nile virus in birds in the territories of Baraba and Kulunda lowlands (West Siberian migration way) during summer-autumn of 2002. *Voprosy Virusologii = Problems of Virology*. 2004;49(3):52-56 (in Russian)

Ternovoi V.A., Protopopova E.V., Chausov E.V., Novikov D.V., Leonova G.N., Netesov S.V., Loktev V.B. Novel variant of tickborne encephalitis virus, Russia. *Emer Infec Dis*. 2007;13(10):1574-1578. doi 10.3201/eid1310.070158

Ternovoi V.A., Ponomareva E.P., Protopopova E.V., Tupota N.L., Mikryukova T.P., Loktev V.B. Changes in the genome of the Tick-Borne encephalitis virus during cultivation. *Mol Biol.* 2024;58(2):266-278. doi 10.1134/S0026893324020146

Thorson A., Formenty P., Lofthouse C., Broutet N. Systematic review of the literature on viral persistence and sexual transmission from recovered Ebola survivors: evidence and recommendations. *BMJ Open*. 2016;6:e008859. doi 10.1136/bmjopen-2015-008859

Worku D.A. Tick-borne encephalitis (TBE): from tick to pathology. *J Clin Med.* 2023;12(21):6859. doi 10.3390/jcm12216859

Zárate S., Hernández-Perez F., Taboada B., Martínez N.E., Alcaráz-Estrada S.L., Del Moral O., Yocupicio-Monroy M. Complete genome of DENV2 isolated from mosquitoes in Mexico. *Infect Genet Evol.* 2019;71:98-107. doi 10.1016/j.meegid.2019.03.018

Conflict of interest. The authors declare no conflict of interest.
Received January 28, 2025. Revised May 23, 2025. Accepted May 30, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):126-135

doi 10.18699/vjgb-26-13

New arguments in the discussion about the nature of picobirnaviruses

A.Yu. Kashnikov, N.V. Epifanova, N.A. Novikova

I.N. Blokhina Nizhny Novgorod Research Institute of Epidemiology and Microbiology, Nizhny Novgorod, Russia
a.kashn@yandex.ru

Abstract. Picobirnaviruses (PBVs), members of the *Picobirnaviridae* family, are found in a wide range of hosts, including eukaryotes (both higher and lower), fungi, and bacteria. However, scientists are unsure about their "true master" or primary host. While often found in animals, including cases of gastroenteritis, they are also detected in environmental samples and have shown genetic links to bacterial and fungal viruses. The lack of a reliable cell culture or animal model for PBV propagation further complicates determining their host specificity. Due to the discovery of prokaryotic regions (motifs) in segments of the PBV genome, it was suggested that their hosts are prokaryotic. However, even this discovery did not pin one specific host to PBVs; since then PBV-like genomes not characteristic of the studied PBV strains, with a mitochondrial genetic code characteristic of lower eukaryotes (molds and invertebrates), were discovered. And recently, a new version of the origin of PBVs from vertebrate viruses and fungi has appeared, denying their phage nature. To understand the nature of genetically diverse PBV strains detected in different organisms, researchers were guided by information about the presence of motifs specific to the viral family in the genome, the genetic code used, and the method of distribution. Recent research suggests that PBVs, previously thought to have a vertebrate origin, may have also evolved from fungal sources denying their phage nature. Some PBV-like sequences have been found to utilize the fungal mitochondrial genetic code, indicating a possible fungal origin or a close relationship with fungal viruses like mitoviruses. This discovery challenges the previously held view of PBVs as exclusively vertebrate viruses and suggests a more complex evolutionary history. The information available today inspires confidence in the imminent conclusion of the ongoing discussion about the possible PBV hosts. In particular, a hypothesis has recently emerged demonstrating a possible mechanism for the replacement of the genetic code in RNA viruses, which makes it possible to explain the origin of PBV forms with the mitochondrial genetic code capable of reproduction in cells of lower eukaryotes using the example of phages. However, an evolutionarily deterministic model demonstrating the path of PBV formation with the genetic code of mold and invertebrate cells has not yet been presented. According to the authors of this review, this evolutionary path is due to the endosymbiotic relationships between the putative PBV hosts, contributing to the horizontal virus spread. The purpose of this review article is to attempt to describe a possible path of formation from the ancestral PBV form and its derived evolutionary forms, some of which inherited a genome with a prokaryotic motif and a standard genetic code, while others acquired a non-standard form of the genome with the code of lower eukaryotes. This review article focuses on the leading role of horizontal transmission in the formation of non-standard intermediate PBV forms.

Key words: picobirnavirus; genome segment; host cell; mitochondrial genetic code; reassortment

For citation: Kashnikov A.Yu., Epifanova N.V., Novikova N.A. New arguments in the discussion about the nature of picobirnaviruses. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):126-135. doi 10.18699/vjgb-26-13

Новые аргументы в дискуссии о природе пикобирнавирусов

А.Ю. Кашников, Н.В. Епифанова, Н.А. Новикова

Нижегородский научно-исследовательский институт эпидемиологии и микробиологии им. академика И.Н. Блохиной Роспотребнадзора, Нижний Новгород, Россия
a.kashn@yandex.ru

Аннотация. Пикобирнавирусы (ПБВ) из семейства *Picobirnaviridae* находят у самых разных хозяев – высших и низших эукариот, а также в грибах и бактериях. Однако среди ученых в отношении ПБВ на сегодняшний день нет однозначного понимания, кто является их истинным хозяином. Принадлежность ПБВ к вирусам высших эукариот не доказана, поскольку не подобраны ни культура животных клеток для их размножения, ни животное-гнотобионт. В связи с обнаружением прокариотических участков (мотивов) в сегментах генома ПБВ было высказано предположение о прокариотической природе их хозяев. Однако и это открытие не закрепило одного конкретного хозяина за ПБВ, так как затем были обнаружены ПБВ-подобные геномы, не характерные для изученных штаммов ПБВ, – с митохондриальным генетическим кодом, свойственным низшим

© Kashnikov A.Yu., Epifanova N.V., Novikova N.A., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

эукариотам (плесени и беспозвоночным). А недавно появилась новая версия происхождения ПБВ от вирусов позвоночных и грибов, отрицающая их фаговую природу. Для понимания природы генетически разнородных штаммов ПБВ, обнаруженных у разных организмов, исследователи руководствовались информацией о присутствии специфических для вирусного семейства мотивов в геноме, используемом генетическом коде и способе распространения. Существующая в настоящее время информация вселяет уверенность в скором завершении продолжающейся дискуссии о возможных хозяевах ПБВ. В частности, недавно появилась гипотеза, демонстрирующая вероятный механизм замены генетического кода у РНК-вирусов, которая позволяет объяснить происхождение форм ПБВ с митохондриальным генетическим кодом, способных к репродукции в клетках низших эукариот на примере фагов. Однако еще не представлена эволюционно детерминированная модель, демонстрирующая путь формирования ПБВ с генетическим кодом клеток плесени и беспозвоночных. Этот эволюционный путь в представлении авторов данного обзора обусловлен эндосимбиотическими отношениями между предполагаемыми хозяевами ПБВ, способствующими горизонтальному распространению вируса. Цель нашей статьи – попытка описания возможного пути формирования из предковой формы ПБВ ее производных эволюционных форм, одни из которых унаследовали геном с прокариотическим мотивом и стандартным генетическим кодом, а другие обрели нестандартную форму генома с кодом низших эукариот. В статье делается акцент на ведущей роли горизонтальной передачи в формировании нестандартных промежуточных форм пикобирнавирусов.

Ключевые слова: пикобирнавирус; сегмент генома; клетка-хозяин; митохондриальный генетический код; реассортация

Discussion about the nature of picobirnaviruses

Picobirnavirus (PBV) is the only genus in the *Picobirnaviridae* family placed under the order "*Diplornavirales"* (Reddy et al., 2023). PBV particles with a diameter of 33–37 nm with a single-layer protein shell have icosahedral symmetry. There is no lipoprotein membrane. PBV genome consists of two dsRNA segments (Delmas et al., 2019). The larger segment (2.4–2.6 kb) (segment 1) encodes a capsid protein and a protein of unknown function, and the smaller segment (1.5–1.9 kb) (segment 2) encodes RNA-dependent RNA polymerase (RdRp of the Pfam family RdRp_1). Based on the RdRp sequence, PBVs are grouped into genogroups, with genogroup I (GI) and genogroup II (GII) being the most common (Malik et al., 2014; Reddy et al., 2023).

Initially, PBVs were detected in the intestinal contents and in the respiratory tract of higher eukaryotes (Pereira et al., 1988; Novikova et al., 2003; Delmas et al., 2019; Kumar et al., 2020; Ghosh, Malik, 2021). In symptomatic infections, PBVs were often detected in the intestines of animals and humans in association with viruses, the pathogenicity of which has been established. However, these viruses are detected both in symptomatic and asymptomatic cases. For this reason, PBVs have traditionally been considered opportunistic intestinal viruses of mammals and birds (Shi et al., 2016; Delmas et al., 2019; Kumar et al., 2020). Since PBVs could not be successfully propagated in mammalian or gnotobiont cell cultures, researchers had doubts about the association of PBVs present in the intestine with animal disease (Ghosh, Malik, 2021; Sadiq et al., 2024). PBVs have been found in invertebrates (mollusks, arthropods, insects) (Shi et al., 2016), and more recent studies have shown that these viruses are likely to infect prokaryotic or fungal host cells (Adriaenssens et al., 2018; Boros et al., 2018; Krishnamurthy, Wang, 2018; Yinda et al., 2018; Kleymann et al., 2020).

Currently, three alternative versions about the nature of PBV hosts are being discussed. According to the first version, the PBV hosts are cells of higher eukaryotes (animals). According to the second version, PBVs may be prokaryotic viruses. This assumption was associated with the discovery of regions (motifs) characteristic of prokaryotic viruses in the PBV genome. According to this version, since PBV-like nucleotide sequences are often found in the animal stool samples, they may be related not to the cells of these animals, but to the presence of the corresponding bacteria in the intestinal microbiome. However, atypical PBV-like genomes with prokaryotic motifs, but with a mitochondrial genetic code characteristic of lower eukaryotes, such as molds and invertebrates, were subsequently isolated from the animal intestinal microbiome (Yinda et al., 2018; Kleymann et al., 2020). So, a third version appeared according to which in addition to prokaryotic cells, some PBVs isolated from the animal intestinal contents can infect the mitochondria of fungi or invertebrates (Shi et al., 2016; Ghosh, Malik, 2021).

There is still no clear answer on the debated issue – which organism(s) are the PBV hosts – although over the years alternative points of view about the nature of PBVs have been supplemented with new arguments in their support. In particular, recently, in favor of the hypothesis denying the phage nature of PBVs, a new idea has emerged about the parallel evolution of PBVs from two different ancestors that were parasites of fungi and vertebrates (Perez et al., 2023).

The version according to which the PBV ancestors are vertebrate and fungal viruses

This version describes an immune response in a vertebrate host based on the recognition of a pathogen, specifically a virus called PBV. The immune system reacts when it detects PBVs containing the standard genetic code, but it does not mount a response when it encounters PBVs with an alternative genetic code, specifically one used by a fungus. The version provides two different evolutionary development models of the PBV amily, which was formed from three different ancestral

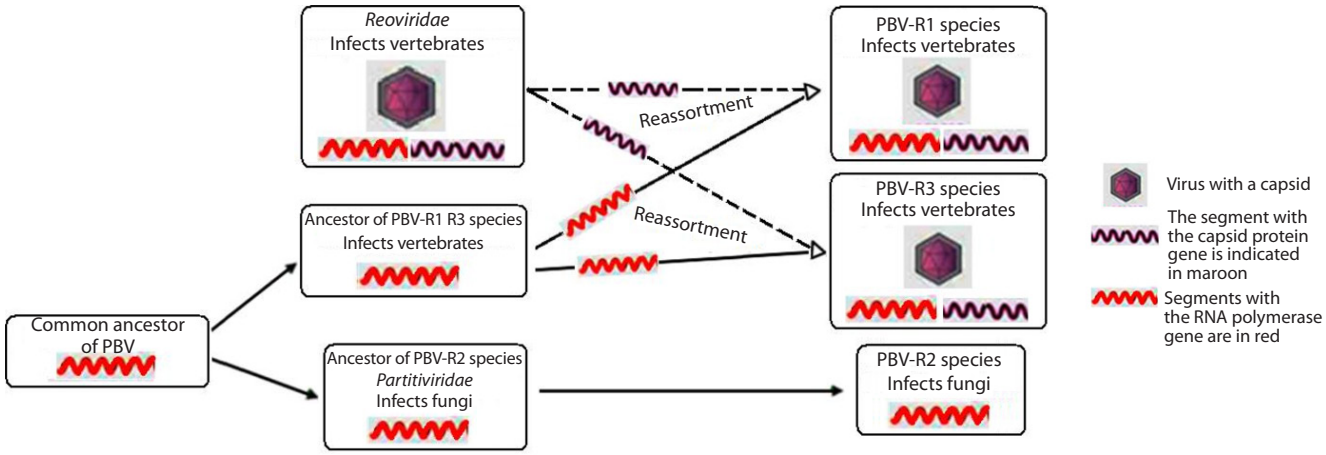


Fig. 1. The first evolutionary development model of the PBV family according to L.J. Perez et al. (2023).

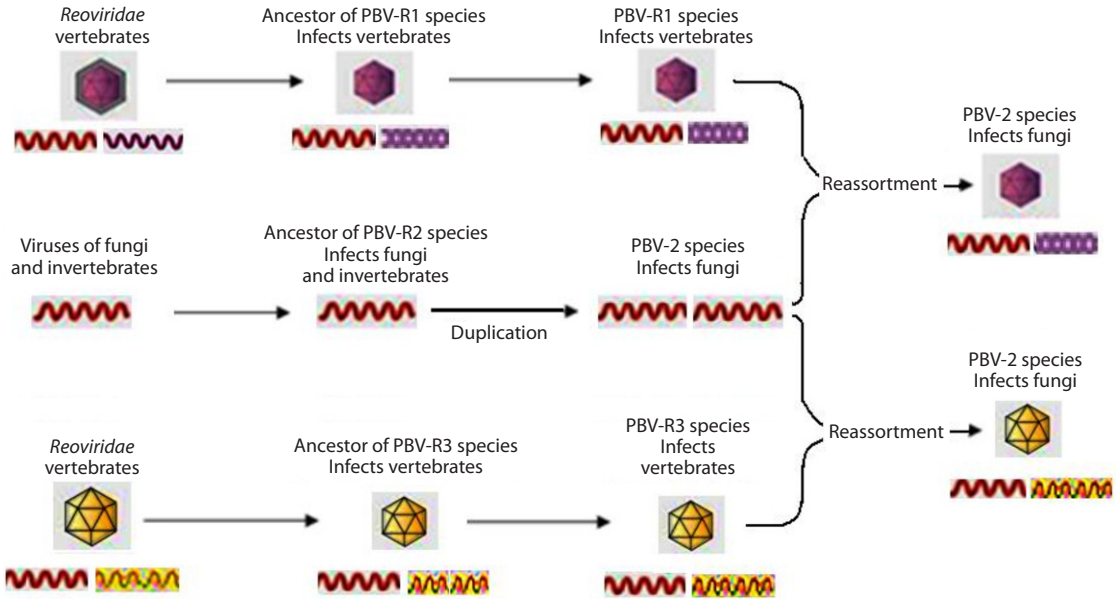


Fig. 2. The second evolutionary development model of the PBV family according to L.J. Perez et al. (2023).
Segments with capsid protein genes are indicated in maroon and yellow; segments with RNA polymerase genes are indicated in red.

lines. According to L.J. Perez et al. (2023), the ancestral PBV lineages (PBV-R1 and PBV-R3) are descended from vertebrate reoviruses (because they provide immunity in the vertebrate in the intestine of which they are detected), and the ancestral PBV-R2 line is descended from fungal partiviruses (immunity in the animal in which they were detected not observed).

According to the first evolutionary development model, at first the common PBV ancestor, containing only the RdRp gene, split into two species. One of them later split into two more species with the simultaneous evolutionary acquisition of a capsid by both species as a result of the reassortment of segments of their genomes with segments of the genome of their common ancestor belonging to the family *Reoviridae*. As a result, two encapsidated ancestral PBV species were located on the same branch of the phylogenetic tree. Together with the previously separated PBV-R2 species, three PBV ancestral species (lines) evolving in parallel were formed during further diversification (Fig. 1).

According to the second evolutionary development model, three PBV-like ancestral species containing the RdRp gene descended from different ancestors, while two species possessed a capsid from the moment of their appearance and, as in the first model, formed a common branch presumably evolving from vertebrate reoviruses (Fig. 2).

The third fungal species (PBV-R2), initially devoid of a capsid, acquired a segment with the capsid gene later as a result of duplication and reassortment of genome segments with one of the two PBV species with capsid infecting vertebrates, according to the mechanism described (Luo et al., 2018). This ancestral species initially resembled mitoviruses

with a mitochondrial code, similar to PBV-like strains found by C.K. Yindaet al. (2018) and A. Kleymann et al. (2020), and later it was located on the same branch as the *Partitiviridae* family, which infects non-chordate eukaryotes (fungi and invertebrates).

The parallel evolution of these three PBV ancestors under similar living conditions could lead to the formation of similar and at the same time genetically diversified representatives of the *Picobirnaviridae* family (Perez et al., 2023). Like the version about the phage nature of PBVs, this version has the right to exist. But the proponents of the version that defends the phage nature of PBVs also present new arguments in favor of their point of view.

The point of view according to which PBVs can infect prokaryotic cells

Proponents of the hypothesis suggesting that prokaryotic cells may be the PBV hosts cite the following arguments in support of the phage nature of PBVs.

1. The PBV genome is enriched (compared to eukaryotic viruses) with prokaryotic motifs – Shine–Dalgarno sequences (SD, "GGAGG" hexamer) similar to the genomes of bacteriophages of the *Cystoviridae* family (Ghosh, Malik, 2021), in which this motif occurs even less frequently (Krishnamurthy, Wang, 2018). For example, a recent study by S. Sadiq et al. (2024) showed that the PBV genome was enriched with the SD motif in 83–85 % of ORF (open reading frames) in the *Picobirnavirus* genomes, which exceeds the enrichment of the genomes of the RNA bacteriophages *Leviviricetes* and *Cystoviridae* with this motif.
2. PBVs are detected almost exclusively in animal stool samples and still cannot be cultured in any eukaryotic cell lines. This may also mean that PBVs are actually bacterial viruses that are part of the animal intestinal microflora or components of their food (Adriaenssens et al., 2018; Delmas et al., 2019; Bell et al., 2020; Guajardo-Leiva et al., 2020; Ghosh, Malik, 2021; Knox et al., 2023; Sadiq et al., 2024). This assumption is consistent with the fact that their closest relatives, representatives of the *Partitiviridae* family, infecting plants, fungi and protozoa (Vainio et al., 2018), are also found in the animal intestinal microflora (Chen et al., 2021) and in the microbiome of invertebrates (Shi et al., 2016; Le Lay et al., 2020).
3. The arrangement of genome segments in typical PBVs is observed in a single particle, as in bacterial viruses, while in most fungal viruses with a segmented genome, the segments are encapsulated separately (Luque et al., 2018).
4. The assumption of opponents of the phage nature of PBVs (Perez et al., 2023) that viruses forming immunity in infected animals (or humans) should belong to eukaryotic viruses cannot unequivocally support the version of the eukaryotic nature of PBV hosts, since it has been established that host immune responses can also occur against bacterial viruses (Dabrowska et al., 2005; Górski et al., 2006). Consequently, PBVs can elicit an immune response not to infection of the animal's own cells, but of the bacterial cells that make up its microbiome (Ghosh, Malik, 2021).
5. Identification of a protein unique to a virus is a key step in determining its specific host, since the presence of bacteriolytic properties in its owner can convincingly indicate that this virus is a bacteriophage (Gan, Wang, 2023; Kashnikov et al., 2023). As it turned out, a protein with a lysing function, which lyses *Escherichia coli*, is present in the PBV capsid, and such proteins are encoded only by the genes of two known families of RNA phages, *Leviviridae* and *Cystoviridae* (Cai et al., 2021). However, perhaps not all, but only some PBVs encode bacteriolysins (Gan, Wang, 2023).
6. A recent study by S. Sadiq et al. (2024) can also confirm the assumption about the phage nature of PBVs. In this study, based on clustering, seven PBV clusters were identified in the microbiome, which could only be hosted by cells of different microbes, since there was no phylogenetic grouping by PBVs belonging to host animals with high genetic variability of PBVs. The lack of grouping between the virus and the host indicates a phylogenetic discrepancy between them (Duraisamy et al., 2018; Woo et al., 2019; Mahar et al., 2020). Based on this, S. Sadiq et al. (2024) suggested that PBVs infect various microbial organisms associated with vertebrates and invertebrates through their diet and habitat.

Based on these observations, proponents of the phage nature of PBVs believe that there could not have been a wide variety of PBVs if their hosts were only mammals, birds, and invertebrates (Sadiq et al., 2024). They suggest that clustering associated with bacteria is confirmed by the high rate of spread of PBVs among animals (even higher than in any other family of RNA viruses). This rate of spread is more consistent with PBVs belonging to prokaryotic hosts, with their extensive interspecific transmission. In addition, proponents of the phage nature of PBVs guess that the preservation of prokaryotic SD sequences is impossible in viruses infecting fungi (Ullah et al., 2022; Wang, 2022). This argument is supported by the probable ability of PBVs, like some phages, to change their genetic code.

Replacement of the genetic code as a trend in phage evolution

In the genomes of some microorganisms, it is possible to replace the standard genetic code with an alternative (mitochondrial) one. This replacement, according to Y. Shulgina and S.R. Eddy (2021), is associated with a decrease in the GC base content during evolution due to genome reduction in microorganisms, which may be due to their parasitic mode of existence – endosymbiosis (McCutcheon, Moran, 2011). In particular, a decrease in the number of GC bases in the genome of organisms leads to a decrease in the proportion of the TGA stop codon (with an increased probability of reassignment of this codon) (Korkmaz et al., 2014).

The discovery of alternative genetic codes in the mitochondrial genome of mold fungi and invertebrates, in bacteria and

archaea demonstrates the genetic code's ability to evolve (Shackelton, Holmes, 2008; Kollmar, Mühlhausen, 2017; Shulgina, Eddy, 2021). The replacement of the genetic code with an alternative one in the mitochondria of eukaryotic cells, bacteria, and archaea could presumably be related to their drive to escape from bacteriophages, which initially used the same code as bacteria (Bender et al., 2008).

Bacteriophages are also capable of changing their genetic code when it becomes necessary to deceive the protective antiviral systems of the host bacterium. For example, it was found that the genetic code in the genome of some phages isolated from baboon stool samples corresponds to an "alternative" genetic code of the genomes of *Bacilli* bacteria (Al-Shayeb et al., 2020). A study by N. Yutin et al. (2021) reports on the replacement (recoding) of stop codons (TAGS) in crAssphage DNA bacteriophages during translation to a codon encoding the amino acid glutamine. The reassignment of the standard TGA stop codon to an alternative one encoding the amino acid tryptophan, corresponding to the code of fungal and invertebrate mitochondria, was observed in some PBV-like genomes by C.K. Yinda et al. (2018), who first discovered this phenomenon in PBVs.

According to D. Wang (2022), the tendency to change the PBV genetic code, as in bacteriophages, is probably related to the ability to read or recode the TGA stop codon during translation, regardless of the taxonomic affiliation of the cells they infect (since phages probably target mitochondria in eukaryotic cells, given their bacterial origin). It is possible that some PBV strains, like bacteriophages, once in a fungal cell, are able to capture the suppressor transfer RNA (tRNA) encoded by the host, or somehow disrupt the host's translation mechanism (Wang, 2022).

According to S.L. Peters et al. (2022), some phages probably use bacterial ribosomes using both standard and alternative codes to translate their proteins. At the same time, the reassignment of normal TAG and TGA stop codons to translation to glutamine and tryptophan is especially common in phages infecting gram-positive bacteria such as *Firmicutes* and *Bacteroidetes* (Peters et al., 2022). *Firmicutes*, known for their low GC content in the genome (less than 50 %), probably related to their parasitic mode of existence, "endosymbiosis", often experience a change of the standard code to an alternative one. It was previously noted that *Firmicutes* are hypothetically most suitable as PBV hosts, which presumably may be phages (Krishnamurthy, Wang, 2018). It is possible that atypical PBV strains with an "alternative" fungal genetic code (Yinda et al., 2018; Kleymann et al., 2020), like phages, were transcoded during translation (Peters et al., 2022) following their host *Clostridial Firmicutes*, which parasitizes the fungal cell.

Thus, it remains likely that PBVs using the fungal mitochondrial code are bacteriophages in which the code replacement is due to evolution associated with the prokaryotic host. Moreover, E.V. Koonin et al. (2020) suggest that PBVs, which were previously considered animal viruses, are exclusively bacterial viruses, in contrast to the phylogenetically close family *Partitiviridae*, which includes eukaryotic and, according to recent data, bacterial viruses, since they, like PBVs, have bacteriolysins. According to U. Neri et al. (2022), the *Picobirnaviridae* family may represent the third clade of RNA bacteriophages together with the *Leviviridae* and *Cystoviridae* families.

Can cells of mold fungi and invertebrates be PBV hosts?

Some arguments by proponents of the phage nature of PBVs cast doubt on the existence of forms of PBVs capable of infecting cells of lower eukaryotes. However, the fact that atypical PBV forms have been identified, the genetic code of which corresponds to viruses of mold fungi and invertebrates, suggests that not all PBVs encode bacteriolysins and, probably, in some PBV forms, like *Mitoviridae* or *Partitiviridae*, cells of lower eukaryotes can be hosts instead of bacterial cells (Luo et al., 2018; Shi et al., 2018; Yinda et al., 2018; Ghosh, Malik, 2021; Ullah et al., 2022; Reddy et al., 2023). The prerequisites for this possibility are: the origin of eukaryotic RNA viruses from +RNA phages (Wolf et al., 2018) and their predisposition to horizontal transmission (Son et al., 2015; Dolja, Koonin, 2018).

The origin of eukaryotic RNA viruses from +RNA phages as evidence of the ability of viruses to change their taxonomic nature

The currently available results of phylogenetic studies indicate the existence of a relationship between the families of eukaryotic and prokaryotic RNA viruses and their common origin (Fig. 3) (Dolja, Koonin, 2018; Wolf et al., 2018). In particular, according to E.V. Koonin et al. (2015), bacteriophages of the *Cystoviridae* family are evolutionarily related to the *Reoviridae* family (eukaryotic viruses) and may be direct ancestors of the *Picobirnaviridae* family.

Based on the analysis of RNA-dependent RNA polymerase, the relationship between single-stranded RNA bacteriophages of the *Leviviridae* family from the *Leviricetes* class and mitoviruses has been proven (Wang, 2022). According to researchers, the families of *Lenarviricota*-type RNA viruses (*Mitoviridae*, *Narnaviridae*, *Urmiaviridae*) are evolutionarily transitional forms from RNA bacteriophages to early RNA viruses infecting lower eukaryotes (Wolf et al., 2018; Sadiq et al., 2022). According to the researchers, the evolutionary transition of prokaryotic viruses to forms of RNA viruses infecting eukaryotes occurred approximately 1.45 billion years ago due to the transition of α-proteobacteria to an endosymbiotic mode of existence. At the same time, forms that lost their capsid due to this evolutionary transition, such as *Mitoviridae*, switched to reproduction in mitochondria, and later, like *Narnaviridae*, entered the cytosol of the cell (Dolja, Koonin, 2018; Wolf et al., 2018).

Y.I. Wolf et al. (2018) believe that different groups of RNA viruses could have evolved independently from prokaryotic +RNA viruses, since their RdRp genes are more similar to

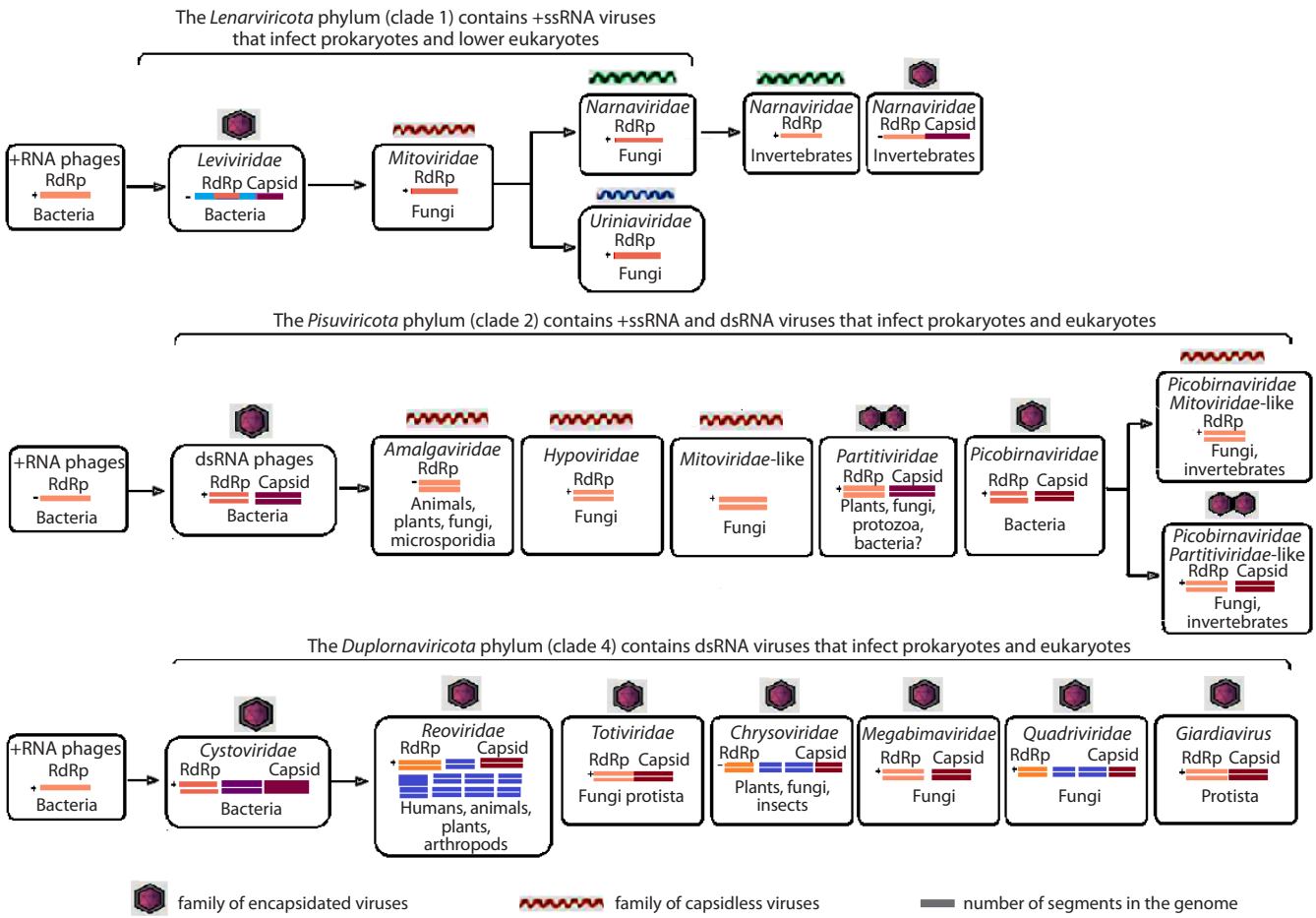


Fig. 3. The evolutionary development model of viral families originating from prokaryotic RNA viruses, according to a version supported by researchers (Koonin et al., 2015; Wolf et al., 2018; Ghosh, Malik, 2021; Sadiq et al., 2022).

the RdRp genes of ancestral +RNA viruses than to the genes of other RNA viruses from parallel branches of the phylogenetic tree.

Role of the method of RNA virus spread in the formation of their taxonomic affiliation

Phylogenetic studies have shown that RNA viruses belonging to the same family can infect hosts from different taxa, including fungi, plants, animals, and protozoa (Son et al., 2015). Moreover, the vast majority of RNA virus families infect unicellular eukaryotes using alternative genetic codes corresponding to these organisms (Neri et al., 2022). These studies are consistent with the "Ancient Coevolution Hypothesis" (Pearson et al., 2009), which states that viruses can migrate from a host belonging to one taxonomic category to a host from another taxonomic category. A similar relationship with organisms from different taxa is observed, for example, in RNA viruses of the families *Partitiviridae*, *Reoviridae*, *Amalgaviridae*, *Totiviridae*, and *Chrysoviridae*. The members of the *Lenarviricota*, *Duplornaviricota*, and *Pisuviricota* RNA viruses (PBVs belong to *Pisuviricota*) are very diverse, distributed in almost any environment, and associated with a wide range of hosts, encompassing bacteria, protozoa, fungi, and plants (Dolja, Koonin, 2018; Sadiq et al., 2022).

According to the "Hypothesis of ancient coevolution", capsidless mitovirus-like PBV forms, like mitoviruses, may represent a transitional evolutionary form from the ancestral +RNA prokaryotic viruses to the simplest eukaryotic viruses (Wolf et al., 2018). Having inherited from its ancestor, the prokaryotic virus, the RdRp gene with a prokaryotic motif and specific motifs (DFXKFD, SGSGGT, and GDD), this form, upon transition to a new host, the fungus, could acquire its genetic code, which its original prokaryotic host could borrow while in an endosymbiotic relationship with the fungus.

Endosymbiosis (symbiogenesis) is a type of symbiosis when one of the partners lives inside the other's cell. In endosymbiosis, the larger of the partners is usually called the host, and the partner is an organism living inside its host, a parasite if it competitively affects its host, suppressing its reproduction. In many associations, parasitic microorganisms move to a permanent intracellular existence and are inherited. The loss of the capsid by the transitional PBV form (similar to *Mitoviridae*) required reproduction inside mitochondria or vacuoles (as in *Narnaviridae*) in order to avoid selective

destruction by the cell protection system of its ds genome corresponding to the replicative form of the *Mitoviridae* genome formed at the intermediate stage of replication. In this case, the PBV-like form P16-366 characterized by the presence of a capsid protein (Yinda et al., 2018) can be considered as an evolutionarily more advanced transitional form, since the presence of a capsid protein in this PBV form does not require hiding its ds genome in the mitochondria of the fungus (Wolf et al., 2018).

A recent large-scale metagenomic study of eukaryotic +RNA cells may confirm the possibility of such an evolutionary scenario. +RNA viruses of the *Narnaviridae* family, which is one of the two known descendant families of +RNA phages hosted by fungi. This study revealed numerous narna-like genovariants among the representatives of *Narnaviridae*, both without the capsid protein gene and with the capsid gene, which were hosted by various invertebrates (Shi et al., 2016). In modern taxonomy, mito- and narna-like RNA viruses descending from bacteriophages with the fungal mitochondrial genetic code have received the status of a family and are defined as eukaryotic RNA viruses. Moreover, while at the family level the spectrum of RNA virus hosts can be wide, at the genus level this spectrum is usually limited and there are clear phylogenetic differences between genera infecting hosts from different taxa (Sadiq et al., 2024).

According to Y.I. Wolf et al. (2018), the origin of PBVs or partiviruses is explained by reassortment between the genome segment of the RdRp virus with +ssRNA from the partivirus-picobirnavirus clade (with a mitovirus-like reproduction method) and the segment with the capsid gene of the virus with the dsRNA genome from the clade of cystoviruses, totiviruses and reoviruses.

Reassortment is a form of exchange of genetic information between viruses with a segmented genome. In viruses with a segmented genome, segment exchange is possible when two (or more) genetically distinct viruses simultaneously infect the same cell. The viruses between which reassortment occurs form a new species or strain of virus with different qualities and, in most cases, increased pathogenicity. Reassortment in nature most often occurs within a single species, but it can also occur within a single genus.

The possibility of reassortment between segments of viral genomes formed by dsRNA and +ssRNA can be explained by the participation in reassortment by +RNA viruses of segments at the stage of replicative intermediate forms of RNA (dsRNA). This variant of the origin of PBVs explains the fact that C.K. Yinda et al. (2018) identified PBV forms with and without capsid with an alternative genetic code, resembling fungal viruses by the method of reproduction.

However, in accordance with the "theory of endosymbiosis", capsidless PBV-like replicons with the fungal mitochondrial genetic code (P11-300, P11-378, P14-90, P15-218, WGML128211, M17A) discovered by (Yinda et al., 2018; Kleymann et al., 2020) can be considered not only as products of the exchange of homologous segments between families of related RNA viruses, but also as a result of symbiotic relationships.

According to the currently available data, PBV genomes with prokaryotic motifs and motifs of mold fungi and invertebrates have been found in the intestines of animals, as part of a single microbiota consisting of bacterial cells and protozoan eukaryotes. The detection of PBVs in the same microbiome with the genetic code of bacterial and mold fungal cells suggests the presence of endosymbiotic relationships between their hosts (Bruto et al., 2014). Recent studies examining the composition of the intestinal microbiota, including bacteria and fungi, indicate that there is a significant interdependence between fungi and bacteria (Li et al., 2022).

The existence of symbiotic relationships between taxonomically different RNA virus hosts allows the virus to overcome the barrier between symbiont cells by horizontal transfer with the possibility of further reproduction in the cells of a new host. An extended phylogenetic analysis proves the widespread horizontal transfer of RNA viruses between different hosts and its leading role in the evolution of these viruses (Dolja, Koonin, 2018). It has been established that bacteria are able to penetrate into the cytoplasm of fungi and stay there for a long time (https://www.pravda.ru/news/science/2024301-simbioz/). Moreover, horizontal transfer of genes or viruses, as a rule, is carried out from bacteria to fungi. In the opposite direction, the transfer of genes (viruses) is hardly possible (mycoviruses are transmitted by fungal cells exclusively vertically – from parents to offspring) (Bruto et al., 2014).

Firmicutes, the alleged PBV hosts (Krishnamurthy, Wang, 2018), being the most common microorganism in the intestines of animals and humans, could end up there along with fungal cells. There is evidence that some *Firmicutes* species, in particular *Clostridial Firmicutes*, competitively interact in the intestine with cells of the fungus *Candida albicans*, preventing the colonization of cells of higher eukaryotes by this fungus (Shulgina. Eddy, 2021). The parasitic mode of existence (endosymbiosis), which binds *Firmicutes* to the unicellular fungi *C. albicans*, allows horizontal transfer of viruses and genes between them (Taggart et al., 2023).

The discovery of PBV genomes with the genetic code of taxonomically different hosts also explains some general patterns in the evolution of dsRNA viruses, which the researchers noticed. This pattern is related to the ability of small dsRNA viruses with minimalistic genomes formed by a minimum number of segments encoding one RdRp protein (*Narnaviridae*) or two RdRPs and a capsid protein (*Totiviridae*, *Partitiviridae*, *Picobirnaviridae*) to perform non-specific horizontal distribution among taxonomically diverse hosts (Dolja, Koonin, 2018). For this reason, PBVs could probably also enter invertebrate cells, since invertebrates are particularly indiscriminate viral hosts.

Often, distantly related invertebrates can be hosts of the same group of viruses (Wolf et al., 2018). On the other hand, dsRNA viruses with the largest possible genome size, such as the *Reoviridae* family, are characterized by a much higher

degree of host specificity, probably due to greater adaptation to it through the acquisition of genes involved in virus–host interactions (Dolja, Koonin, 2018).

It is believed that bacterial viruses cannot directly infect cells of higher eukaryotic organs, and they can enter these organs only by non-specific translocation with the help of bacteria in which they multiply (Dabrowska et al., 2005). It has only recently been reported that mammalian cells are able to directly internalize bacteriophages (Bichet et al., 2023) through macropinocytosis (non-specific internalization) and in rare cases through receptor-mediated endocytosis (specific internalization) (Bichet et al., 2023). However, to date, the ability of PBVs to penetrate animal cells has not been proven.

Conclusion

Considering the sources at our disposal, which provide the molecular and genetic characteristics of the PBV strains discovered so far, it can be assumed that prokaryotes are most likely the hosts of naturally occurring PBV strains, and their transitional evolutionary forms may be present in unicellular eukaryotes (molds and invertebrates).

In favor of this assumption, we present the following arguments, which, as it seems to us, are the most significant.

Arguments supporting the possibility of PBV reproduction in prokaryotic cells:

- The PBV genome is enriched with SD sequences, which are inherent in prokaryotes (Krishnamurthy, Wang, 2018).
- The presence of a bacteriolytic protein in PBVs can serve as a convincing argument that PBVs are bacteriophages (Gan, Wang, 2023).
- The propensity of PBVs to genetic changes (characteristic of viruses with a segmented genome) is more characteristic of prokaryotic viruses.
- Ubiquitous detection of PBVs in the intestines of animals of various levels of organization (vertebrates and invertebrates) and in wastewater (Ghosh, Malik, 2021);
- Inability to cultivate PBVs in the laboratory or isolate them from animal tissue samples (Sadiq et al., 2024).
- More frequent interspecific transmission of PBVs than animal RNA viruses from any other family (Sadiq et al., 2024).
- The phenomenon of reassignment during translation of TAG and TGA stop codons to alternative ones (encoding amino acids glutamine and tryptophan) is especially common in phages infecting Firmicutes and Bacteroidetes (Peters et al., 2022).

Arguments pointing to the possibility of the existence of some transitional evolutionary PBV forms in molds and invertebrates:

- The common origin of the families of prokaryotic and eukaryotic RNA viruses allows for the existence of a wide range of RNA virus hosts related to different taxa at the family level (Wolf et al., 2018).
- The endosymbiotic relationship between *Firmicutes* (putative PBV hosts) and *C. albicans* fungi (Peters et al., 2022), occupying the same ecological niche (animal or human intestines), suggests the possibility of horizontal transfer of genes and viruses between them (Shulgina, Eddy, 2021).
- The endosymbiotic relationship between *Firmicutes* and *C. albicansis* being the reason for the PBV code replacement following its host *Firmicutes* (Bender et al., 2008).
- The mechanism by which PBV genetic code changes (Wang, 2022) may demonstrate not only the ability of phages to reproduce in cells regardless of their taxonomic affiliation, but also a possible pathway for the formation of transitional evolutionary PBV forms found in the cells of some lower eukaryotes.
- Only some PBVs have proteins with bacteriolytic functionin the capsid (Wang, 2022), as well as in the closely related family *Partitiviridae*, the representatives of which belong to fungal viruses (Neri et al., 2022).

These observations confirm the possibility of the existence of separate PBV forms, taxonomically related to both bacteria and unicellular eukaryotes – mold fungi and invertebrates. Therefore, as D. Wang et al. (2022) rightly noted, the final conclusions regarding the true host(s) of the identified PBVs cannot be generalized at the family level, but require studies of the whole diversity of PBVs in order to determine the taxonomic affiliation of the entire spectrum of their hosts.

References

Adriaenssens E.M., Farkas K., Harrison C., Jones D.L., Allison H.E., McCarthy A.J. Viromic analysis of wastewater input to a river catchment reveals a diverse assemblage of RNA viruses. *mSystems.* 2018;3(3):e00025-18. doi 10.1128/mSystems.00025-18

Al-Shayeb B., Sachdeva R., Chen L.X., Ward F., Munk P., Devoto A., Castelle C.J., … Lehours A.C., Warren L., Cate J.H.D., Santini J.M., Banfield J.F. Clades of huge phages from across Earth's ecosystems. *Nature.* 2020;578(7795):425-431. doi 10.1038/s41586-020-2007-4

Bell N., Khamrin P., Kumthip K., Rojjanadumrongkul K., Nantachit N., Maneekarn N. Molecular detection and characterization of picobirnavirus in environmental water in Thailand. *Clin Lab*. 2020;66(5): 855-858. doi 10.7754/Clin.Lab.2019.191013

Bender A., Hajieva P., Moosmann B. Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. *Proc Natl Acad Sci USA*. 2008;105(43):16496-16501. doi 10.1073/pnas.0802779105

Bichet M.C., Adderley J., Avellaneda-Franco L., Magnin-Bougma I., Torriero-Smith N., Gearing L.J., Deffrasnes C., … Lin R.C., Patwa R., Moseley G.W., Doerig C., Barr J.J. Mammalian cells internalize bacteriophages and use them as a resource to enhance cellular growth and survival. *PLoS Biol*. 2023;21(10):e3002341. doi 10.1371/journal.pbio.3002341

Boros Á., Polgár B., Pankovics P., Fenyvesi H., Engelmann P., Phan T.G., Delwart E., Reuter G. Multiple divergent picobirnaviruses with functional prokaryotic Shine–Dalgarno ribosome binding sites present in cloacal sample of a diarrheic chicken. *Virology*. 2018; 525:62-72. doi 10.1016/j.virol.2018.09.008

Bruto M., Prigent-Combaret C., Luis P., Moënne-Loccoz Y., Muller D. Frequent, independent transfers of a catabolic gene from bacteria to contrasted filamentous eukaryotes. *Proc Biol Sci*. 2014;281(1789): 20140848. doi 10.1098/rspb.2014.0848

Cai X., Tian F., Teng L., Liu H., Tong Y., Le S., Zhang T. Cultivation of a lytic double-stranded RNA bacteriophage infecting *Microvirgula aerodenitrificans* reveals a mutualistic parasitic lifestyle. *J Virol*. 2021;95(17):e0039921. doi 10.1128/jvi.00399-21

Chen Y.-M., Sadiq S., Tian J.-H., Chen X., Lin X.-D., Shen J.-J., Chen H., … Xu Q.-Y., Wang W., Gao W.-H., Holmes E.C., Zhang Y.-Z. RNA virome composition is shaped by sampling ecotype. *SSRN Electron J.* 2021. doi 10.2139/ssrn.3934022

Dabrowska K., Switała-Jelen K., Opolski A., Weber-Dabrowska B., Gorski A. Bacteriophage penetration in vertebrates. *J Appl Microbiol*. 2005;98(1):7-13. doi 10.1111/j.1365-2672.2004.02422.x

Delmas B., Attoui H., Ghosh S., Malik Y.S., Mundt E., Vakharia V.N. ICTV virus taxonomy profile: *Picobirnaviridae*. *J Gen Virol*. 2019; 100(2):133-134. doi 10.1099/jgv.0.001186

Dolja V.V., Koonin E.V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* 2018;244:36-52. doi 10.1016/j.virusres.2017.10.020

Duraisamy R., Akiana J., Davoust B., Mediannikov O., Michelle C., Robert C., Parra H.-J., Raoult D., Biagini P., Desnues C. Detection of novel RNA viruses from free-living gorillas, Republic of the Congo: genetic diversity of picobirnaviruses. *Virus Genes*. 2018;54(2):256-271. doi 10.1007/s11262-018-1543-6

Gan T., Wang D. Picobirnaviruses encode proteins that are functional bacterial lysins. *Proc Natl Acad Sci USA*. 2023;120(37): e2309647120. doi 10.1073/pnas.2309647120

Ghosh S., Malik Y.S. The true host/s of picobirnaviruses. *Front Vet Sci*. 2021;7:615293. doi 10.3389/fvets.2020.615293

Górski A., Kniotek M., Perkowska-Ptasińska A., Mróz A., Przerwa A., Gorczyca W., Dabrowska K., Weber-Dabrowska B., Nowaczyk M. Bacteriophages and transplantation tolerance. *Transplant Proc*. 2006;38(1):331-333. doi 10.1016/j.transproceed.2005.12.073

Guajardo-Leiva S., Chnaiderman J., Gaggero A., Díez B. Metagenomic insights into the sewage RNA virosphere of a large city. *Viruses*. 2020;12(9):1050. doi 10.3390/v12091050

Kashnikov A.Yu., Epifanova N.V., Novikova N.A. On the nature of picobirnaviruses. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(3):264-275. doi 10.18699/VJGB-23-32

Kleymann A., Becker A.A.M.J., Malik Y.S., Kobayashi N., Ghosh S. Detection and molecular characterization of picobirnaviruses (PBVs) in the mongoose: identification of a novel PBV using an alternative genetic code. *Viruses*. 2020;12(1):99. doi 10.3390/v12010099

Knox M.A., Wierenga J., Biggs P.J., Gedye K., Almeida V., Hall R., Kalema-Zikusoka G., Rubanga S., Ngabirano A., Valdivia-Granda W., Hayman D.T.S. Abundant dsRNA picobirnaviruses show little geographic or host association in terrestrial systems. *Infect Genet Evol*. 2023;112:105456. doi 10.1016/j.meegid.2023.105456

Kollmar M., Mühlhausen S. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. *BioEssays*. 2017;39(5):1600221. doi 10.1002/bies.201600221

Koonin E.V., Dolja V.V., Krupovic M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology*. 2015;479-480: 2-25. doi 10.1016/j.virol.2015.02.039

Koonin E.V., Dolja V.V., Krupovic M., Varsani A., Wolf Y.I., Yutin N., Zerbini F.M., Kuhn J.H. Global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev*. 2020;84(2): e00061-19. doi 10.1128/MMBR.00061-19

Korkmaz G., Holm M., Wiens T., Sanyal S. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem*. 2014;289(44):30334-30342. doi 10.1074/jbc.M114.606632

Krishnamurthy S.R., Wang D. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology*. 2018;516:108-114. doi 10.1016/j.virol.2018.01.006

Kumar N., Mascarenhas J.D.A.P., Ghosh S., Masachessi G., da Silva Bandeira R., Nates S.V., Dhama K., Singh R.K., Malik Y.S. Picobirnavirus. In: Malik Y.S., Singh R.K., Dhama K. (Eds) Animal-Origin Viral Zoonoses. Singapore: Springer, 2020;291-312. doi 10.1007/978-981-15-2651-0_13

Le Lay C., Shi M., Buček A., Bourguignon T., Lo N., Holmes E.C. Unmapped RNA virus diversity in termites and their symbionts. *Viruses*. 2020;12(10):1145. doi 10.3390/v12101145

Li H., Miao M.X., Jia C.L., Cao Y.B., Yan T.H., Jiang Y.Y., Yang F. Interactions between *Candida albicans* and the resident microbiota. *Front Microbiol*. 2022;13:930495. doi 10.3389/fmicb.2022.930495

Luo X.L., Lu S., Jin D., Yang J., Wu S.S., Xu J. *Marmota himalayana* in the Qinghai-Tibetan plateau as a special host for bi-segmented and unsegmented picobirnaviruses. *Emerg Microbes Infect*. 2018; 7(1):20. doi 10.1038/s41426-018-0020-6

Luque D., Mata C.P., Suzuki N., Ghabrial S.A., Castón J.R. Capsid structure of dsRNA fungal viruses. *Viruses*. 2018;10(9):481. doi 10.3390/v10090481

Mahar J.E., Shi M., Hall R.N., Strive T., Holmes E.C. Comparative analysis of RNA virome composition in rabbits and associated ectoparasites. *J Virol*. 2020;94(11):e02119-19. doi 10.1128/JVI.02119-19

Malik Y.S., Kumar N., Sharma K., Dhama K., Shabbir M.Z., Ganesh B., Kobayashi N., Banyai K. Epidemiology, phylogeny and evolution of emerging enteric Picobirnaviruses of animal origin and their relationship to human strains. *Biomed Res Int.* 2014;2014:780752. doi 10.1155/2014/780752

McCutcheon J.P., Moran N.A. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 2011;10(1):13-26. doi 10.1038/nrmicro 2670

Neri U., Wolf Y.I., Roux S., Camargo A.P., Lee B., Kazlauskas D., Chen I.M., … Krupovic M., Dolja V.V., Kyrpides N.C., Koonin E.V., Gophna U. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*. 2022;185(21):4023-4037.e18. doi 10.1016/j.cell.2022.08.023

Novikova N.A., Epifanova N.V., Fedorova O.F., Golitsyna L.N., Kuprianova N.V. The detection of picobirnavirus by RNA electrophoresis in polyacrylamide gel. *Voprosy Virusologii = Problems of Virology*. 2003;48(6):41-43 (in Russian)

Pearson M.N., Beever R.E., Boine B., Arthur K. Mycoviruses of filamentous fungi and their relevance to plant pathology. *Mol Plant Pathol*. 2009;10(1):115-128. doi 10.1111/j.1364-3703.2008.00503.x

Pereira H.G., Fialho A.M., Flewett T.H., Teixeira J.M.S., Andrade Z.P. Novel viruses in human faeces. *Lancet*. 1988;332(8602):103-104. doi 10.1016/S0140-6736(88)90032-3

Perez L.J., Cloherty G.A., Berg M.G. Parallel evolution of picobirnaviruses from distinct ancestral origins. *Microbiol Spectr*. 2023;11(6): e02693-23. doi 10.1128/spectrum.02693-23

Peters S.L., Borges A.L., Giannone R.J., Morowitz M.J., Banfield J.F., Hettich R.L. Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun.* 2022;13(1): 5710. doi 10.1038/s41467-022-32979-6

Reddy M.V., Gupta V., Nayak A., Tiwari S.P. Picobirnaviruses in animals: a review. *Mol Biol Rep*. 2023;50(2):1785-1797. doi 10.1007/s11033-022-08133-2

Sadiq S., Chen Y.M., Zhang Y.Z., Holmes E.C. Resolving deep evolutionary relationships within the RNA virus phylum *Lenarviricota*. *Virus Evol*. 2022;8(1):veac055. doi 10.1093/ve/veac055

Sadiq S., Holmes E.C., Mahar J.E. Genomic and phylogenetic features of the *Picobirnaviridae* suggest microbial rather than animal hosts. *Virus Evol*. 2024;10(1):veae033. doi 10.1093/ve/veae033

Shackelton L.A., Holmes E.C. The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol*. 2008;254(1):128-134. doi 10.1016/j.jtbi.2008.05.024

Shi M., Lin X.D., Tian J.H., Chen L.J., Chen X., Li C.X., Qin X.C., … Buchmann J., Wang W., Xu J., Holmes E.C., Zhang Y.-Z. Redefining the invertebrate RNA virosphere. *Nature*. 2016;540(7634):539-543. doi 10.1038/nature 20167

Shi M., Lin X.D., Chen X., Tian J.H., Chen L.J., Li K., Wang W., Eden J.S., Shen J.J., Liu L., Holmes E.C., Zhang Y.Z. The evolutionary history of vertebrate RNA viruses. *Nature*. 2018;556(7700): 197-202. doi 10.1038/s41586-018-0012-7

Shulgina Y., Eddy S.R. A computational screen for alternative genetic codes in over 250,000 genomes. *eLife*. 2021;10:e71402. doi 10.7554/eLife.71402

Son M., Yu J., Kim K.-H. Five questions about Mycoviruses. *PLoS Pathog*. 2015;11(11):e1005172. doi 10.1371/journal.ppat.1005172

Taggart N.T., Crabtree A.M., Creagh J.W., Bizarria R. Jr., Li S., de la Higuera I., Barnes J.E., Shipley M.A., Boyer J.M., Stedman K.M., Ytreberg F.M., Rowley P.A. Novel viruses of the family *Partitiviridae* discovered in *Saccharomyces cerevisiae*. *PLoS Pathog*. 2023; 19(6):e1011418. doi 10.1371/journal.ppat.1011418

Ullah K., Mehmood A., Chen X., Dar M.A., Yang S., Zhang W. Detection and molecular characterization of picobirnaviruses in the wild birds: identification of a novel picobirnavirus possessing yeast mitochondrial genetic code. *Virus Res*. 2022;308:198624. doi 10.1016/j.virusres.2021.198624

Vainio E.J., Chiba S., Ghabrial S.A., Maiss E., Roossinck M., Sabanadzovic S., Suzuki N., Xie J., Nibert M. ICTV virus taxonomy profile: *Partitiviridae*. *J Gen Virol.* 2018;99(1):17-18. doi 10.1099/jgv.0.000985

Wang D. The enigma of picobirnaviruses: viruses of animals, fungi, or bacteria? *Curr Opin Virol*. 2022;54:101232. doi 10.1016/j.coviro.2022.101232

Wolf Y.I., Kazlauskas D., Iranzo J., Lucía-Sanz A., Kuhn J.H., Krupovic M., Dolja V.V., Koonin E.V. Origins and evolution of the global RNA virome. *MBio*. 2018;9(6):e02329-18. doi 10.1128/mBio.02329-18

Woo P.C.Y., Teng J.L.L., Bai R., Tang Y., Wong A.Y.P., Li K.S.M., Lam C.S.F., Fan R.Y.Y., Lau S.K.P., Yuen K.-Y. Novel picobirnaviruses in respiratory and alimentary tracts of cattle and monkeys with large intra- and inter-host diversity. *Viruses*. 2019;11(6):574. doi 10.3390/v11060574

Yinda C.K., Ghogomu S.M., Conceição-Neto N., Beller L., Deboutte W., Vanhulle E., Maes P., Van Ranst M., Matthijnssens J. Cameroonian fruit bats harbor divergent viruses, including rotavirus H, bastroviruses, and picobirnaviruses using an alternative genetic code. *Virus Evol*. 2018;4(1):vey008. doi 10.1093/ve/vey008

Yutin N., Benler S., Shmakov S.A., Wolf Y.I., Tolstoy I., Rayko M., Antipov D., Pevzner P.A., Koonin E.V. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun*. 2021;12(1):1044. doi 10.1038/s41467-021-21350-w

Conflict of interest. The authors declare no conflict of interest.
Received February 17, 2025. Revised April 16, 2025. Accepted June 25, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):136-145

doi 10.18699/vjgb-26-15

Evolution of gene order in mtDNA of Baikal endemic amphipods and its possible mechanisms

E.A. Sirotinina, D.Yu. Sherbakov, E.V. Romanova

Limnological Institute of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
haleo.inc@gmail.com

Abstract. Significant gene order diversity of mitochondrial (mt) genomes of invertebrates is peculiar to subphylum Crustacea, and to order Amphipoda in particular. Amphipods from Lake Baikal are also known as a group with unique gene orders in their mt genomes. To estimate the diversity of protein-coding gene orders (GOs) in amphipods, a comparative analysis of gene rearrangements in the mt genomes of Baikal and non-Baikal species was performed. In some cases, gene rearrangement data and the history of gene relocation in different taxonomic groups can also supplement the results of phylogenetic inferences. Among the thirteen mt genomes of Baikal species sequenced in previous studies, four gene order patterns were identified, and fourteen gene order patterns for 114 mt genomes of non-Baikal species were observed. The type and number of rearrangement steps (from 1 to 3) required to transition from one order to another and the number of mt genes rearranged in each GO (from 1 to 5) were also defined. Baikalian amphipods belong to two lineages (I and II) according to molecular data which reveal their origin from two independent introductions of ancestral species into the lake. All cases of mt gene order rearrangements have been detected in species from the first lineage, whereas the mt gene order in the second lineage is conserved in all species studied and corresponds to the Pancrustacean pattern (PanGO). PanGO has been determined as the ancestral gene order for both Baikalian amphipod lineages. The possible mechanisms of mt gene order rearrangements such as a complete or partial duplication of mt genome and subsequent random deletions are discussed in our study. It is supposed that increased mutation rate, weakening of stabilizing selection and other specific factors may influence the probability of emergence and fixation of different GOs in mt genomes of Baikalian amphipods.
Key words: amphipods; Lake Baikal; mitochondrial genome; gene rearrangement

For citation: Sirotinina E.A., Sherbakov D.Yu., Romanova E.V. Evolution of gene order in mtDNA of Baikal endemic amphipods and its possible mechanisms. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):136-145. doi 10.18699/vjgb-26-15

Эволюция порядка генов в мтДНК байкальских эндемичных амфипод и ее возможные механизмы

Е.А. Сиротинина, Д.Ю. Щербаков, Е.В. Романова

Лимнологический институт Сибирского отделения Российской академии наук, Иркутск, Россия
haleo.inc@gmail.com

Аннотация. Значительное разнообразие в порядке генов митохондриальных (мт) геномов беспозвоночных отмечается в подтипе Crustacea, и, в частности, в отряде Amphipoda. Амфиподы озера Байкал также известны как группа с уникальными порядками генов в мт геномах. Чтобы оценить разнообразие порядков белок-кодирующих генов (ПГ) амфипод, был проведен сравнительный анализ генных перестроек в мт геномах байкальских и небайкальских видов. В некоторых случаях данные о перестройках генов и истории их перемещений у различных таксономических групп также могут информативно дополнять данные филогенетического анализа. Для 13 ранее полученных нуклеотидных последовательностей мт геномов байкальских видов мы определили 4 варианта ПГ, для 114 мт геномов небайкальских видов – 14 вариантов ПГ. Были также рассчитаны типы и число шагов перестроек (от 1 до 3), требующихся для перехода от одного порядка генов к другому, и число мт генов, подвергшихся перестройкам в каждом ПГ (от 1 до 5). Байкальские амфиподы принадлежат к двум линиям (I и II) в соответствии с молекулярными данными, указывающими на их происхождение от двух независимых вселений предковых видов в озеро. Все случаи перестроек порядка мт генов обнаружены у видов из линии I, тогда как порядок мт генов в линии II консервативен у всех изученных видов и соответствует модели Pancrustacean pattern (ПанПГ). ПанПГ был определен как предковый порядок генов для обеих линий байкальских амфипод. В исследовании обсуждаются возможные механизмы перестроек порядка мт генов, такие, как полная или частичная дупликация мт генома и последующие случайные делеции. Высказывается предположение, что повышенная скорость мутаций, ослабление стабилизирующего отбора и другие особые факторы могут влиять на вероятность появления и фиксации различных ПГ в мт геномах байкальских амфипод.
Ключевые слова: амфиподы; озеро Байкал; митохондриальный геном; перестройки генов

© Sirotinina E.A., Sherbakov D.Yu., Romanova E.V., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

Introduction

The diversity of mitochondrial (mt) genome gene order in eukaryotes is remarkably high. Depending on evolutionary pathways in different lineages, the types of mt genome organization (linear, circular, or assembled into multiple chromosomes), gene content (involving gene loss and acquisition), and gene arrangement are variable (Sterling-Montealegre, Prada, 2024). The broad range in mt genome sizes (from 11–50 kb in animals to 66 kb–11.3 Mb in plants) is observed largely due to variations in gene number, as well as the length of non-coding mtDNA. The diversity of gene orders (GOs) is determined by both changes in the relative positions of genes and the presence of deletions and duplications. In some major taxa, the GO is conserved (e. g., in vertebrates), while in others it varies widely (e. g., in some groups of invertebrates) (Zardoya, 2020). The gene orders in the mt genomes of crustaceans are among the most diverse found in invertebrates (subphylum Crustacea), which manifested through changes in gene positions relative to the basal order, which is also characteristic of insects (the Pancrustacean pattern, PanGO) (Boore, 1999; Kilpert, Podsiadlowski, 2006; Sterling-Montealegre, Prada, 2024).

The types of gene order alterations can be different. A recent study R.A. Sterling-Montealegre and C.F. Prada (2024) analyzed 299 genomes from the subphylum Crustacea among 464 mt genomes from 47 arthropod orders and identified 87 distinct gene orders, including four gene orders most frequently found in crustaceans.

According to this data, rearrangements in mt genomes of Crustacea that involve tRNA genes accounted for 70.1 % of all structural changes, while alterations in the positions of other genes comprised 29.9 %. Changes in the positions of protein-coding and rRNA genes are less common (Castellucci et al., 2022) than those of tRNA genes, the positions of which can differ even among species within the same family (Jühling et al., 2012). Previous research has demonstrated a positive association between a high rate of gene rearrangement and an elevated nucleotide substitution rate in protein-coding genes of the subclass Caenogastropoda Cox, 1959 (gastropod mollusks) (Fourdrilis et al., 2018). Baikalian amphipods, a group characterized by diverse mt gene order, exhibit significantly accelerated evolution in several protein-coding genes, compared to related freshwater species of the genus *Gammarus* (Romanova, Sherbakov, 2019). Nevertheless, this correlation is not observed in all groups of organisms (Shao et al., 2003; Xu et al., 2006). Remarkably, the order of genes in mt genomes is highly conserved in some crustacean lineages over long evolutionary timescales, while being highly variable in other groups (Kilpert et al., 2012; Tan et al., 2019; Zardoya, 2020).

Phylogenetic and statistical analyses allow to reconstruct the history of GO changes in different taxa. In some instances, GO can represent informative synapomorphies for particular groups, providing additional support for resolving phylogenetic relationships at the level of infraorders, superfamilies, and families (Tan et al., 2019).

Mt genomes with rearranged gene orders are frequently found in Baikal amphipods (Romanova et al., 2016, 2020; Drozdova et al., 2024). This group of invertebrates originated in the lake through adaptive radiation and has persisted over a long period (Sherbakov, 1999; Mats et al., 2011). The causes leading to GO rearrangements in this group of endemic invertebrates, as well as their putative mechanisms, remain unknown (Mueller, Boore, 2005).

The article investigates the diversity of mt gene order in amphipods through a comparative analysis of gene rearrangements, as well as suggestions on the mechanisms for their origin within an evolutionary framework. The results supported PanGO as the ancestral state for Baikalian species and revealed certain patterns in the emergence of rearrangements within this group. Clades exhibiting divergent GO patterns were also identified among the studied species.

Materials and methods

Amphipod gene order dataset. Mt genome sequences for this study were retrieved from the GenBank database (Supplementary Material 1)[1]. The mt genomes of Baikalian species were published in earlier studies (Rivarola-Duarte et al., 2014; Romanova et al., 2016, 2021; Mamos et al., 2021). Specimen collection localities, along with detailed protocols for DNA extraction, PCR amplification, sequencing, genome assembly, and annotation are described in previous publications (Romanova et al., 2016, 2021) and in Supplementary Material 2.

The dataset for GO comparison comprised only mt genomes from 127 species which had complete coding sequences of 13 protein-coding genes. Ribosomal and tRNA genes, as well as the control region (CR), were excluded from the dataset because in some species they were either missing or present as duplicated copies, which would have compromised comparative analysis. The mt genome of the Baikalian species *Linevichella vortex* (Dybowsky, 1874) was also excluded from the GO analysis due to its incomplete genome but was retained for the phylogenetic analysis. For pairwise comparisons, we employed the PanGO scheme. It is important to note that the method of defining gene orders based only on the 13 protein-coding genes (as opposed to all 37 mt genes) more often results in orders being classified as identical, whether they are compared to PanGO or to one another. In contrast, their complete gene orders would not be identical due to the differences in the positions of the remaining genes.

Analysis of gene order using the CREx program. The types of mitochondrial gene rearrangements in amphipods were assessed using the program CREx (Common Interval Rearrangement Explorer), which is part of the CREx2 suite (Bernt et al., 2007; Hartmann et al., 2019), on the Galaxy server (The Galaxy Community, 2024) (https://usegalaxy.eu/). The CREx algorithm employs the concept of common intervals – groups of genes that are consecutively arranged in the compared GOs. The number of common intervals (NSCI) is used to assess the similarity of GOs. A higher NSCI value indicates a greater degree of identity between them. Using event models, CREx identified possible rearrangement scenarios, including tandem duplication of a segment of adjacent genes followed by random

[1] Supplementary Materials 1–4 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Sirot_Engl_30_1.zip

loss of some gene copies (TDRL), transposition (T), defined as the movement of a gene within the same coding strand of the mtDNA, gene inversion (I) to the other coding strand of the mtDNA, and inverse transposition (iT) (Bernt et al., 2007; Basso et al., 2017).

Within the CREx algorithm, the aforementioned types of rearrangements are assessed exclusively using mathematical models. TDRL is more frequently postulated in scenarios involving larger-scale mt genome rearrangements and the movement of a greater number of genes in a single step compared to transposition. It is also an asymmetric rearrangement that can only occur in one direction for any two compared gene orders, thereby enabling the determination of the ancestral state (Bernt et al., 2007). When two or more alternative scenarios were predicted, the one requiring the smallest number of steps was selected.

Manual counting of position changes in specific genes. In pairwise comparisons of PanGO with the GO of each species, the gene order within gene blocks on both mtDNA strands was examined, along with whether the order of blocks was conserved in the genome. We performed a visual analysis, inspecting the sequence from *cox1* to *nad1* to identify the boundaries of each conserved block, between which gene rearrangements occurred. Genes occupying the same position as in PanGO were designated as "0", while individual genes, the position of which was altered relative to the conserved gene blocks in PanGO, were designated as "1" (Supplementary Materials 3 and 4). The quantification results were visualized as histograms, and the corresponding GO schemes were aligned with phylogenetic trees.

Phylogenetic reconstructions. For phylogenetic tree construction, amino acid sequences of 13 protein-coding genes from the mt genomes of 128 amphipod species (available in the GenBank database as of December 10, 2023), including 14 Baikalian species (among them the species *L. vortex* with an incomplete mt genome), and three species from the order Isopoda (*Ligia oceanica* (Linnaeus, 1767), *Eophreatoicus karrkkanj* Wilson & Humphrey, 2020, *Neomysis japonica* Nakazawa, 1910), used as an outgroup, were employed.

Amino acid sequences of each protein-coding gene were aligned at the codon level using TranslatorX (Abascal et al., 2010) with the ClustalW algorithm and were subsequently concatenated in SeaView 4.5.4 (Gouy et al., 2010). The optimal amino acid substitution model, JTT+F+I+R9, was selected using ModelFinder (Kalyaanamoorthy et al., 2017). A maximum likelihood phylogenetic tree was constructed with IQ-TREE 2 (Minh et al., 2020) and was subsequently visualized in FigTree 1.4.3 (Rambaut, 2010).

The phylogenetic tree of amphipods, integrated with gene order data, was used to identify regularities of gene rearrangements that emerged during the evolution of the studied taxa. Branches and nodes of the tree where changes in the positions of individual genes occurred, as well as the variants of GOs in different species, were annotated. The correspondence between changes in gene order and species phylogeny was also assessed.

Results

Analysis of gene order rearrangements

Among the 127 complete mt genomes of amphipods analyzed, 18 distinct protein-coding gene orders (GO 1–GO 18) were found. Using CREx, it was determined that a minimum of 1 to 3 rearrangement events are required to transition between different gene orders. Amphipod species, their mt gene orders, and the types of rearrangements are presented in Supplementary Material 3. Further analysis of gene rearrangements was performed considering only unique gene orders. The total number of mt genes that had changed position relative to PanGO was manually quantified and annotated for two groups of amphipods: Baikalian and non-Baikalian species (Fig. 1), as well as separately for each GO (Fig. 2, Supplementary Material 2). Such rearrangements were found to be more frequent in the larger and ecologically more diverse sample of non-Baikalian species.

The *nad1* gene was found to be the most frequently repositioned relative to the PanGO (occurring in eight GOs) in both Baikalian and non-Baikalian groups (Fig. 1, Supplementary Material 4). Eight of the thirteen protein-coding genes (*nad1*, *cytb*, *nad6*, *nad5*, *cox2*, *nad3*, *nad4*, *nad4l*) underwent rearrangements, particularly inversions, more frequently than others (Fig. 1). The GOs with the highest number of repositioned genes across all amphipods were found in *Macrohectopus branickii* (Dybowsky, 1874) (GO 4, five genes), in two species of the genus *Caprella* Lamarck, 1801 and in *Cyamus boopis* Lütken, 1870 (GO 5, three genes), and *Crypturopus tuberculatus* with *C. inflatus* (Dybowsky, 1874) (GO 2, three genes) (Fig. 2, Supplementary Material 3).

Complex rearrangements that involve 2 to 5 repositioned genes were identified in 20 species, corresponding to nine gene orders (GO 2, GO 4, GO 5, GO 6, GO 7, GO 11, GO 15, GO 16, GO 17) (Supplementary Material 4). The remaining GOs had only a single repositioned gene. Gene inversions (I) and inverse transpositions (iT) were frequently observed, occurring in seven of the 17 altered GOs (Supplementary Material 3). Transposition was the most prevalent type of gene rearrangement in amphipod mt genomes, found in 11 GOs.

We classified rearrangement events as complex if they met one of the following criteria: required two or more steps according to CREx calculations; involved a combination of different rearrangement types (inversions, transpositions); or included at least one TDRL event (Bernt et al., 2007; Castellucci et al., 2022). The complex TDRL rearrangement type was found exclusively in Baikalian species *M. branickii*, *C. tuberculatus*, and *C. inflatus*. Non-Baikalian species exhibited combinations of different rearrangement types, but no TDRL events were detected. Only the Baikalian species from lineage I (*M. branickii*) showed both a high number of rearrangement steps and high event complexity, involving inversion, TDRL, and transposition events. In contrast, the mt genomes of species from lineage II maintain the ancestral PanGO for protein-coding gene order (Supplementary Material 3).

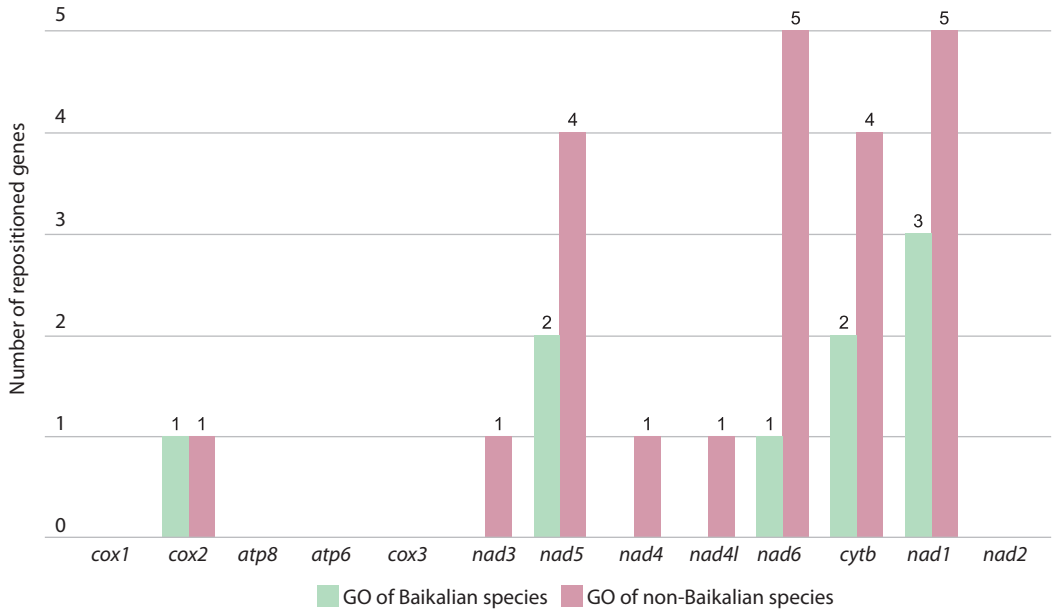


Fig. 1. The number of rearrangement events for each mitochondrial protein-coding gene relative to the PanGO, quantified for Baikalian and non-Baikalian amphipod groups.

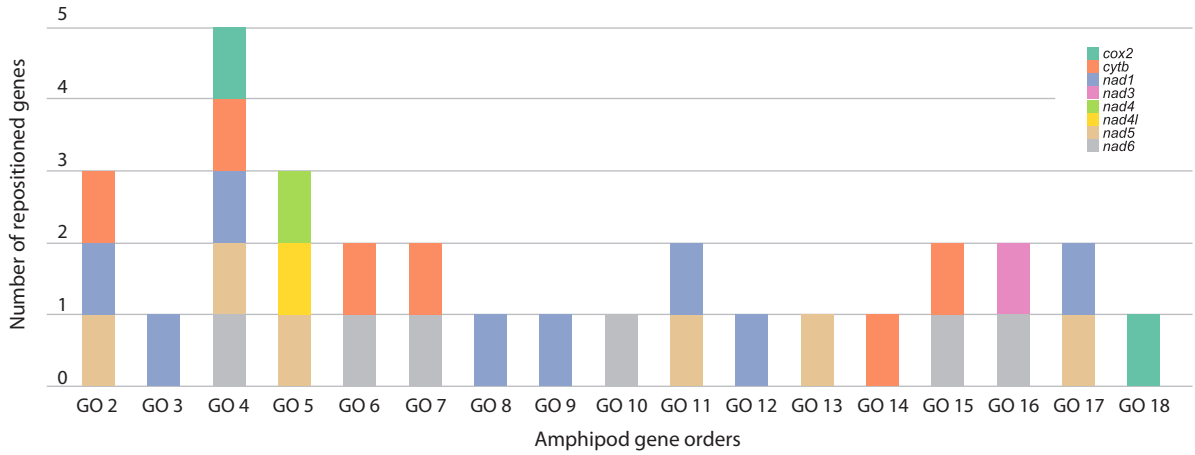


Fig. 2. Gene orders of amphipods (GO 2–GO 18) with the number of mitochondrial protein-coding genes repositioned relative to PanGO. GO1, which corresponds to PanGO, is not shown.

Phylogenetic analysis

Phylogenetic reconstruction was performed using amino acid sequences of 13 protein-coding genes from 128 amphipod species. The outgroup is represented by a clade of isopods (Fig. 3). The group of Baikalian amphipods includes 13 species. The Baikalian species split into two clades, corresponding to the previously identified lineages I and II.

Each Baikalian lineage clusters with different species from the genus *Gammarus* Fabricius, 1775: lineage I clusters with *G. pisinnus* Hou, Li & Li, 2014, *G. fossarum* Koch, 1836, *Baikalogammarus pullus* (Dybowsky, 1874) and *G. roeselii* Gervais, 1835, while lineage II clusters with *Gammarus lacustris* G.O. Sars, 1863. Therefore, the genus *Gammarus* is inferred to be paraphyletic. The clustering pattern of the phylogenetic tree indicates that both Baikalian lineages share a common ancestor with species of the genus *Gammarus*, as *G. duebeni* Lilljeborg, 1852 and *G. chevreuxi* Sexton, 1913 occupy a basal position relative to these lineages. Lineage I comprises species from the families Micruropodidae Kamaltynov, 1999, Macrohectopodidae Sowinsky, 1915 and Crypturopodidae Kamaltynov, 2002, while lineage II consists of species from the families Eulimnogammaridae Kamaltynov, 1999, Acanthogammaridae Garjajeff, 1901, and Pallaseidae Tachteew, 2001.

Species of the genera *Metacrangonyx* Chevreux, 1909 (inhabiting marine brackish waters, rarely freshwaters) and *Pseudoniphargus* Chevreux, 1901 (stygobionts, inhabiting environments from brackish wells to mountain rivers) form monophyletic groups. The species *Echinogammarus berilloni*

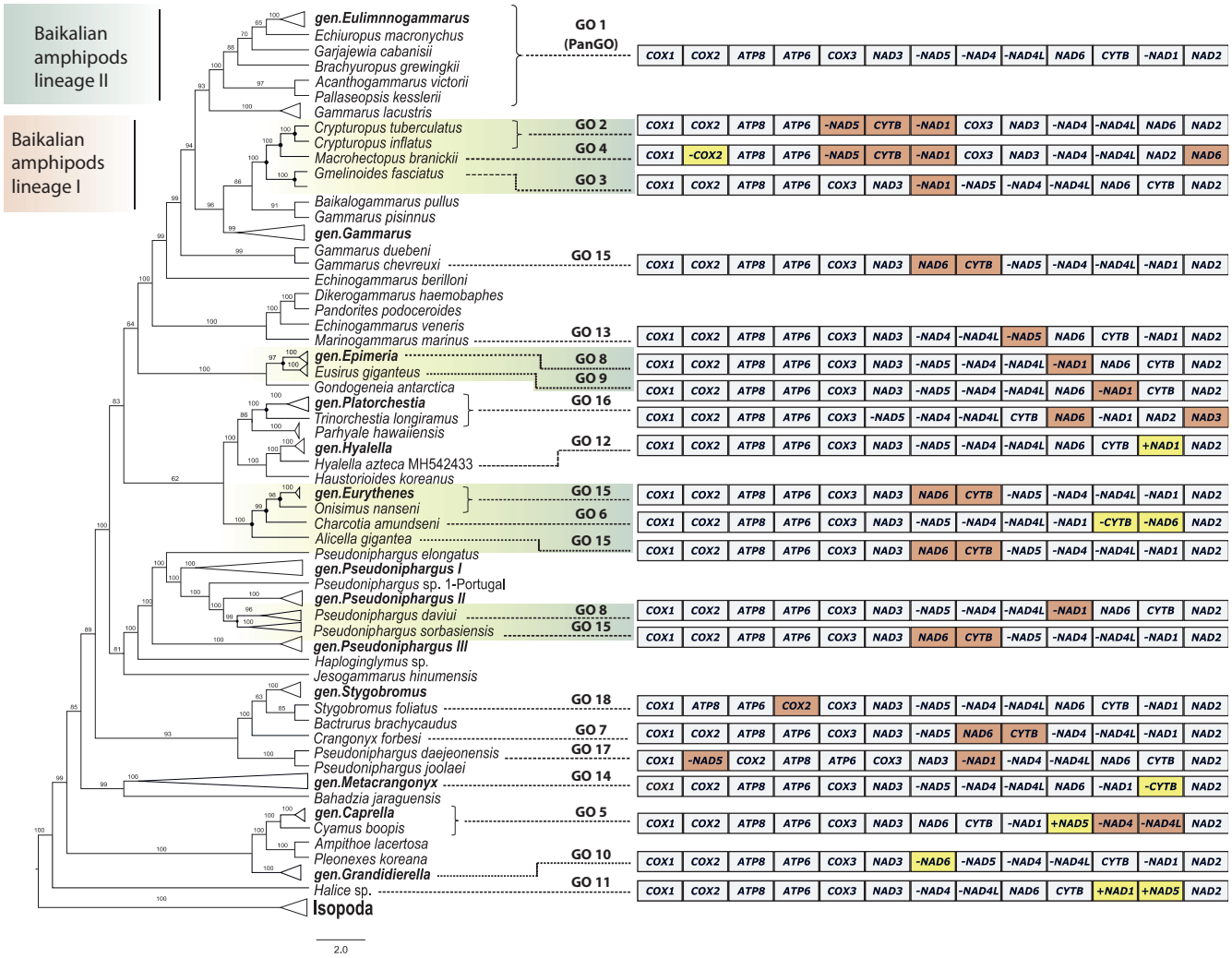


Fig. 3. Phylogenetic tree reconstructed from amino acid sequences of 13 mt protein-coding genes from 128 amphipod species, with a scheme of corresponding mt genome gene orders.

The tree was manually adjusted for proportions, and branch lengths are scaled. Green-shaded areas indicate species groups with dissimilar GOs. The *L. vortex* species is not labeled as it was excluded from the GO analysis. Genes that changed their coding strand (I, iT) are highlighted in yellow, while genes subjected to T and TDRL rearrangements are marked in red.

(Catta, 1878), found in rivers and streams of Western Europe, occupies a basal position to the clade comprising the Baikalian species and species of the genus *Gammarus*. The mt gene order of *E. berilloni* is identical to the GO of the latter species (except for *G. chevreuxi*).

Integration of phylogenetic and manual gene rearrangement analyses data enabled the identification of the amphipod taxa with the most highly rearranged mt gene orders. Figures 2 and 3 and Supplementary Material 3 demonstrate that the same protein-coding genes have been repositioned in the mt genomes of numerous distantly related species. In some instances, such as in *Eurythenes maldoror* d'Udekem d'Acoz & Havermans, 2015 and *G. chevreuxi* (both possessing GO 15), these rearrangements have convergently resulted in identical GOs, representing a case of homoplasy. Analysis of mt genomes from species of the genus *Gammarus* allowed us to conclude that the PanGO represents a symplesiomorphy for both Baikalian lineages.

Discussion

The results of the gene order analysis in amphipod mt genomes characterize this group as one with frequent and diverse gene rearrangements (18 distinct protein-coding GOs identified across 127 species). Four out of 13 Baikalian species correspond to three distinct GOs, while 53 out of 114 non-Baikalian species correspond to fourteen distinct GOs (Fig. 1 and 2). Thus, GO changes are observed in 57 species within the dataset, which represent 17 distinct GOs. The previously described PanGO is the most common protein-coding gene order among amphipods, although whether it represents the ancestral state for this group remains unconfirmed. The GO 11 found in the basal amphipod species *Halice* sp. (Li et al., 2019) is significantly different from PanGO.

Amphipods, being an evolutionarily ancient invertebrate group, have presumably accumulated numerous gene order variants (Hou et al., 2014; Mamos et al., 2021). In a study of mt genome diversity across Decapoda infraorders, M.H. Tan

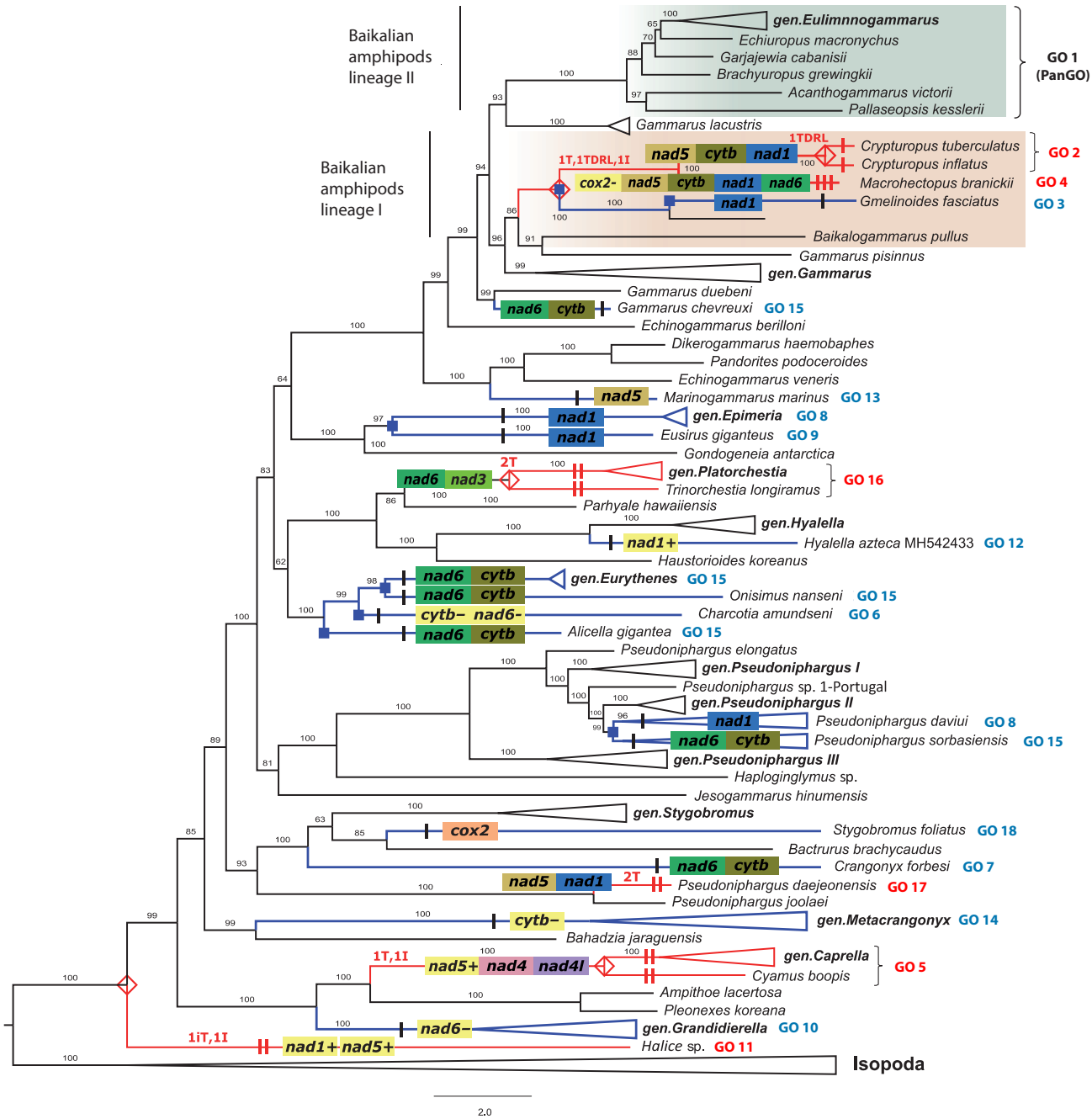


Fig. 4. Phylogenetic tree reconstructed from amino acid sequences of 13 mt protein-coding genes from 128 amphipod species.

Branches are annotated with the names of genes that have changed position relative to PanGO. Vertical bars on branches show the number of rearrangement steps (CREx). Genes that changed their coding strand (I, iT) are highlighted in yellow. Branches and gene orders marked in red indicate complex rearrangement types within the respective species groups, and those in blue denote simple types of changes. Species *L. vortex* is not labeled as it was excluded from the GO analysis.

et al. determined rearrangement "hotspots" on phylogenetic trees annotated with GO schemes (Tan et al., 2019). These hotspots denote clades of closely related species with two or more unique gene orders that are distinct from one another and from the ancestral state. In total, four groups of amphipods containing dissimilar GOs were identified. Two gene orders (GO 6 and GO 15) were found in the superfamily Lysianassoidea Dana, 1849, with GO 15 involving identical gene rearrangements in six species. Two gene orders (GO 8 and GO 9) were found in the genus *Epimeria* A. Costa in Hope, 1851, and two (GO 8 and GO 15) in the genus *Pseudoniphargus*. Three gene orders (GO 2, GO 3, and GO 4) were observed in Baikalian lineage I, with repositioned genes present in four out of its five species (Fig. 3).

Based on CREx calculations, Baikalian lineage I (1–3 steps) shows similarity both to the species group comprising the genus *Caprella* and the species *C. boopis* (two steps), and to the basal species *Halice* sp. (two steps). Gene orders involving more complex rearrangement patterns are concentrated in these groups (Fig. 4). While the formation of GOs in species of

the genus *Platorchestia* Bousfield, 1982 (2T) and the species *P. daejeonensis* (2T) suggests relatively simpler rearrangement scenarios, which still require two steps. Consequently, Baikalian lineage I represents the group with the highest evolutionary dynamics, demonstrating both the greatest number of complex gene rearrangements and the highest diversity of unique GOs.

Amphipod species with altered GOs, which possess diverse biological and ecological traits, are widely distributed across various taxonomic groups in the phylogenetic tree (Fig. 4), demonstrating the instability of GO as a taxonomic character. Novel gene orders, as well as the ancestral PanGO, are present in both ancient lineages and more recent evolutionary radiations. As expected, there are frequent instances where species from the same family forming a single clade share identical GOs – for example, *Caprella mutica* Schurin, 1935 and *C. boopis* (GO 5).

There are also phylogenetically distant amphipod species that show identical GOs when analyzing the protein-coding gene dataset. In particular, homoplasy is observed in three distantly related species groups that share the same rearranged positions of *nad6* and *cytb* genes, corresponding to GO 15 (Supplementary Material 3, Fig. 3 and 4). When tRNA, rRNA, and control regions are included in the analysis, the gene orders of such species are generally no longer identical.

Homoplasies in mt genome gene order are often observed across diverse invertebrate species (Kilpert et al., 2012; Tan et al., 2019; Castellucci et al., 2022). We observed this phenomenon noting both the lower occurrence frequency of protein-coding gene rearrangements compared to tRNA repositioning (Jühling et al., 2012) and its potential role as a convergent trait in distantly related amphipod groups. Similar rearrangement patterns in the same gene cluster *nad5-nad4---nad6---nad1* are observed in many freshwater and marine amphipod species. Rearrangements in four Baikalian species also affect this particular gene cluster.

GO in the mt genomes of invertebrates does not always align with their classification into species, genera or other taxa established through morphological characteristics. However, by combining phylogenetic analysis with gene orders comparison, it is possible to identify clades with the highest GO diversity and nodes on the tree where this feature underwent the most substantial modifications (groups with dissimilar GOs). For example, such discordance is observed within species of the genus *Pseudoniphargus* and among representatives of Baikalian lineage I, where several GOs deviate from the pattern typical for these groups.

Studies report that the uneven distribution of gene rearrangements across the phylogenetic tree is characteristic of Hexapoda and crustaceans, which may be explained by parallel evolution of this trait (Tan et al., 2017; Moreno-Carmona et al., 2021). A representative example includes gene rearrangements in certain cladoceran taxa (e. g., *Daphnia*, *Bosmina*) where PanGO is retained in the mt genomes of most but not all the species (17 out of 32 species conform to PanGO, while 15 species exhibit nine other GOs) (Castellucci et al., 2022). While most *Gammarus* species conform to PanGO, there is a single species, *G. chevreuxi*, which shows altered *cytb* and *nad6* positions. Even greater gene order divergence is observed in the *Hyalella azteca* (Saussure, 1858) mt genome with accession MH542433, which contains a *nad1* inversion absent in genome MT672041. A previous study reports additional inversions and inverse transpositions within the genus *Hyalella* S.I. Smith, 1874, which may be associated with distinct southern and northern populations (Zapelloni et al., 2021).

Since PanGO is ancestral for both Baikalian lineages, all alterations in GO likely emerged in various Baikalian amphipod species during adaptive radiation within the lake. Mt genome rearrangements appear to have occurred in multiple steps toward the more diverse GOs of lineage I Baikalian species. The least altered gene orders (GO 3, GO 8–10, GO 12–14) represent plesiomorphic states. Lineage I Baikalian amphipods from the families Crypturopodidae, Micruropodidae, and Macrohectopodidae (Kamaltynov, 1999) have highly altered GOs not found in other amphipods. It is also noteworthy that the partial mt genome of *L. vortex* (Micruropodidae) retains the ancestral PanGO. The species within this group (excluding *M. branickii*) are shallow-water, thermotolerant amphipods (*Gmelinoides fasciatus* (Stebbing, 1899), *C. tuberculatus*, *L. vortex*) (Kamaltynov, 2001).

Weakened stabilizing selection, resulting from a low effective size and reduced genetic variation in ancestral populations (Charlesworth, 2009; Lavrov, Pett, 2016), may be the factor explaining the increased frequency of GO rearrangements in certain species (Shao et al., 2003). Shallow-water Baikalian amphipods appear particularly prone to rapid population decline due to environmental changes. Specifically, glacial periods during the geological history of Lake Baikal led to cycles of species extinction (Goldberg et al., 2010; Mats et al., 2011), as demonstrated for the southwestern population of *G. fasciatus* (Bukin et al., 2018) and for *M. branickii* (Petunina et al., 2023). Furthermore, high microsporidian infection rates observed in *G. fasciatus* and *M. branickii* (Petunina et al., 2023) may lead to a decreased proportion of males in the population. Since parasites lack mitochondria and utilize host ATP, it is plausible that adaptation to their negative impact has contributed to mtDNA rearrangements in several amphipod species, which may have improved ATP production efficiency (Bukin et al., 2018).

Amphipod species with completely sequenced mt genomes currently represent diverse taxonomic groups inhabiting ecologically distinct biotopes in the lake. A number of authors have attempted to define whether associations exist between GO variations and ecological characteristics, divergence times, features of life cycle or habitat, concluding that GO changes are multifactorial and may be linked to evolution in a number of biochemical and metabolic traits (Romanova, Sherbakov, 2019; Tan et al., 2019; Castellucci et al., 2022; Benito et al., 2024). In a study of the isopod species *Janira maculosa* Leach, 1814, the authors suggested that the frequently observed gene rearrangements in the mt genome could be partially explained by its low structural complexity compared to the nuclear genome (Kilpert et al., 2012). Gene relocations, particularly to the opposite coding strand, affect the mitochondrial transcription process and may influence its efficiency.

Lineage II of Baikalian amphipods is a group with greater taxonomic and ecological diversity that maintains the ances-

tral protein-coding GO. Nevertheless, interspecific differences exist in the number and organization of tRNA genes (Romanova et al., 2016; Romanova, Sherbakov, 2019). It has been proposed in some studies that changes in tRNA gene number and position occur more often than ones affecting rRNA or protein-coding genes in mt genomes (Pääbo et al., 1991; Jühling et al., 2012).

One hypothesized mechanism for mt gene changes is partial or complete genome duplication, followed by subsequent loss of duplicated regions (Jühling et al., 2012). Existing models for calculating scenarios of converting one GO into another are based on the assumption that an equal number of duplication and deletion events occurs at each step of the TDRL process. Although the distance calculation between two gene orders under the TDRL model is known to be asymmetric, requiring a different number of steps for forward and reverse scenario reconstructions (Bernt et al., 2007), the model does not account for the possibility that the ratio of transformations during TDRL (e. g., one duplication to several deletions, or vice versa) could be more complex. Consequently, the inferred minimal number of steps for a rearrangement may overestimate those actually required by the true evolutionary scenario.

The replication of mtDNA via the rolling circle mechanism, which produces a dimeric mtDNA molecule (Fučíková et al., 2016; Xia et al., 2016; Wang et al., 2022), along with deletion events affecting either copy, may result in the duplication of tRNA genes and protein-coding gene clusters. Various mechanisms for the formation of such deletions involving either recombination or errors in replication have been described in diverse taxonomic groups (Nissanka et al., 2019; Oliveira et al., 2020). Furthermore, if a promoter mutation occurs within one copy of mtDNA dimer, it may lead to its pseudogenization (Lavrov et al., 2002).

Indirect evidence for the presence of such a mechanism in Baikalian species includes a duplicated *cox2* gene and three fragmented *atp8* copies in *M. branickii*, along with truncated gene copies within extensive non-coding regions of *G. fasciatus* (*atp6*, *nad4l*), *Brachyuropus grewingkii* (Dybowsky, 1874) (*atp8*, *cox2*, *nad2*), *Garjajewia cabanisii* (Dybowsky, 1874) (*cytb*), including duplicated tRNA genes (Romanova et al., 2020). These regions, typically uncommon in animal mt genomes, may indicate historical duplication events followed by subsequent sequence degeneration (Boore, 1999), while potentially serving as a factor facilitating further gene rearrangements.

In certain cases, mt gene order may provide supplementary diagnostic characters for taxonomic classification (Lavrov, Lang, 2005; Tan et al., 2019). However, explaining the underlying causes of accelerated mtDNA evolutionary rate and the high variability of GOs observed in both Baikalian and other amphipod species will require further investigation.

Conclusion

Analysis of protein-coding gene order in the mitochondrial genomes of Baikalian amphipods allowed us to confirm PanGO as the ancestral state for both Baikalian lineages. Gene order alterations in several Baikalian species emerged during their evolutionary history within the lake. The rearrangement pathways inferred by CREx likely represent simplified models, as the underlying duplication and deletion events may occur with unequal probabilities.

Analysis of gene order rearrangements and phylogenetic analysis of protein-coding sequences of available amphipod mt genomes revealed four distinct groups with complex gene rearrangement patterns. These comprise clades that include species from the superfamily Lysianassoidea, the genus *Epimeria*, the genus *Pseudoniphargus*, and Baikalian amphipods of lineage I. We propose that the more ancient origin of lineage I Baikalian species relative to lineage II may partially explain their greater diversity in gene order arrangements. Additionally, it is hypothesized that the low effective population size in lineage I Baikalian amphipods could be one of the factors that is weakening the effect of stabilizing selection, thereby enabling the fixation of mt genomic rearrangements.

References

Abascal F., Zardoya R., Telford M.J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res*. 2010;38(Suppl. 2):W7-W13. doi 10.1093/nar/gkq291

Basso A., Babbucci M., Pauletto M., Riginella E., Patarnello T., Negrisolo E. The highly rearranged mitochondrial genomes of the crabs *Maja crispata* and *Maja squinado* (Majidae) and gene order evolution in Brachyura. *Sci Rep*. 2017;7(1):4096. doi 10.1038/s41598-017-04168-9

Benito J.B., Porter M.L., Niemiller M.L. Comparative mitogenomic analysis of subterranean and surface amphipods (Crustacea, Amphipoda) with special reference to the family Crangonyctidae. *BMC Genomics*. 2024;25(1):298. doi 10.1186/s12864-024-10111-w

Bernt M., Merkle D., Ramsch K., Fritzsch G., Perseke M., Bernhard D., Schlegel M., Stadler P.F., Middendorf M. CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics*. 2007;23(21):2957-2958. doi 10.1093/bioinformatics/btm468

Boore J.L. Animal mitochondrial genomes. *Nucleic Acids Res*. 1999;27(8):1767-1780. doi 10.1093/nar/27.8.1767

Bukin Yu.S., Petunina J.V., Sherbakov D.Yu. The mechanisms for genetic diversity of Baikal endemic amphipod *Gmelinoides fasciatus*: relationships between the population processes and paleoclimatic history of the lake. *Russ J Genet*. 2018;54(9):1059-1068. doi 10.1134/S1022795418090053

Castellucci F., Luchetti A., Mantovani B. Exploring mitogenome evolution in Branchiopoda (Crustacea) lineages reveals gene order rearrangements in Cladocera. *Sci Rep*. 2022;12(1):4931. doi 10.1038/s41598-022-08873-y

Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 2009;10(3):195-205.

Drozdova P.B., Madyarova E.V., Gurkov A.N., Saranchina A.E., Romanova E.V., Petunina J.V., Peretolchina T.E., Sherbakov D.Y., Timofeyev M.A. Lake Baikal amphipods and their genomes, great and small. *Vavilov J Genet Breed*. 2024;28(3):317-325. doi 10.18699/vjgb-24-36

Fourdrilis S., de Frias Martins A.M., Backeljau T. Relation between mitochondrial DNA hyperdiversity, mutation rate and mitochondrial genome evolution in *Melarhaphe neritoides* (Gastropoda: Littorinidae) and other Caenogastropoda. *Sci Rep*. 2018;8(1):17964. doi 10.1038/s41598-018-36428-7

Fučíková K., Lewis P.O., González-Halphen D., Lewis L.A. Gene arrangement convergence, diverse intron content, and genetic code modifications in mitochondrial genomes of Sphaeropleales (Chlorophyta). *Genome Biol Evol*. 2016;6(8):2170-2180. doi 10.1093/gbe/evu172

Goldberg E.L., Chebykin E.P., Zhuchenko N.A., Vorobyeva S.S., Stepanova O.G., Khlystov O.M., Ivanov E.V., Weinberg E., Gvozdkov A.N. Uranium isotopes as proxies of the environmental history of the Lake Baikal watershed (East Siberia) during the past 150ka. *Palaeogeogr Palaeoclimatol Palaeoecol*. 2010;294(1-2):16-29. doi 10.1016/j.palaeo.2009.08.030

Gouy M., Guindon S., Gascuel O. SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 2010;27(2):221-224. doi 10.1093/molbev/msp259

Hartmann T., Bernt M., Middendorf M. An exact algorithm for sorting by weighted preserving genome rearrangements. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;16(1):52-62. doi 10.1109/TCBB.2018.2831661

Hou Z., Sket B., Li S. Phylogenetic analyses of Gammaridae crustacean reveal different diversification patterns among sister lineages in the Tethyan region. *Cladistics*. 2014;30(4):352-365. doi 10.1111/cla.12055

Jühling F., Pütz J., Bernt M., Donath A., Middendorf M., Florentz C., Stadler P.F. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Res*. 2012;40(7):2833-2845. doi 10.1093/nar/gkr1131

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., Von Haeseler A., Jermiin L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587-589. doi 10.1038/nmeth.4285

Kamaltynov R.M. On the higher classification of Lake Baikal amphipods. *Crustaceana*. 1999;72(8):933-944. doi 10.1163/156854099503834

Kamaltynov R.M. Amphipods (Amphipoda: Gammaridea). In: Index of Animal Species Inhabiting Lake Baikal and its Catchment Area. Novosibirsk: Nauka Publ., 2001;572-831 (in Russian)

Kilpert F., Podsiadlowski L. The complete mitochondrial genome of the common sea slater, *Ligia oceanica* (Crustacea, Isopoda) bears a novel gene order and unusual control region features. *BMC Genomics*. 2006;7:241. doi 10.1186/1471-2164-7-241

Kilpert F., Held C., Podsiadlowski L. Multiple rearrangements in mitochondrial genomes of Isopoda and phylogenetic implications. *Mol Phylogenet Evol*. 2012;64(1):106-117. doi 10.1016/j.ympev.2012.03.013

Lavrov D.V., Lang B.F. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements. *Syst Biol*. 2005;54(4):651-659. doi 10.1080/10635150500221044

Lavrov D.V., Pett W. Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. *Genome Biol Evol*. 2016;8(9):2896-2913. doi 10.1093/gbe/evw195

Lavrov D.V., Boore J.L., Brown W.M. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. *Mol Biol Evol*. 2002;19(2):163-169. doi 10.1093/oxfordjournals.molbev.a004068

Li J.Y., Zeng C., Yan G.Y., He L.S. Characterization of the mitochondrial genome of an ancient amphipod *Halice* sp. MT-2017 (Pardaliscidae) from 10,908 m in the Mariana Trench. *Sci Rep*. 2019;9(1):2610. doi 10.1038/s41598-019-38735-z

Mamos T., Grabowski M., Rewicz T., Bojko J., Strapagiel D., Burzyński A. Mitochondrial genomes, phylogenetic associations, and SNP recovery for the key invasive Ponto-Caspian amphipods in Europe. *Int J Mol Sci*. 2021;22(19):10300. doi 10.3390/ijms221910300

Mats V.D., Shcherbakov D.Y., Efimova I.M. Late Cretaceous-Cenozoic history of the Lake Baikal depression and formation of its unique biodiversity. *Stratigr Geol Correl*. 2011;19(4):404-423. doi 10.1134/S0869593811040058

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R., Teeling E. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530-1534. doi 10.1093/molbev/msaa015

Moreno-Carmona M., Cameron S.L., Prada Quiroga C.F. How are the mitochondrial genomes reorganized in Hexapoda? Differential evolution and the first report of convergences within Hexapoda. *Gene*. 2021;791:145719. doi 10.1016/j.gene.2021.145719

Mueller R.L., Boore J.L. Molecular mechanisms of extensive mitochondrial gene rearrangement in plethodontid salamanders. *Mol Biol Evol*. 2005;22(10):2104-2112. doi 10.1093/molbev/msi204

Nissanka N., Minczuk M., Moraes C.T. Mechanisms of mitochondrial DNA deletion formation. *Trends Genet*. 2019;35(3):235-244. doi 10.1016/j.tig.2019.01.001

Oliveira M.T., Pontes C.B., Ciesielski G.L. Roles of the mitochondrial replisome in mitochondrial DNA deletion formation. *Genet Mol Biol*. 2020;43(Suppl. 1):e20190069. doi 10.1590/1678-4685-GMB-2019-0069

Pääbo S., Thomas W.K., Whitfield K.M., Kumazawa Y., Wilson A.C. Rearrangements of mitochondrial transfer RNA genes in marsupials. *J Mol Evol*. 1991;33(5):426-430. doi 10.1007/BF02103134

Petunina J.V., Vavrischuk N.V., Bukin Yu.S., Romanova E.V. Variability of morphological and genetic characteristics of *Macrohectopus branickii* (Dyb., 1874) (Amphipoda, Macrohectopidae). *Izvestiya Irkutskogo Gosudarstvennogo Universiteta*. *Seriya*: *Biologiya. Ecologiya = The Bulletin of Irkutsk State University. Series*: *Biology. Ecology*. 2023;46:18-28. doi 10.26516/2073-3372.2023.46.18 (in Russian)

Rambaut A. FigTree. Latest Version – v1.4.4. 2018. Available at: http://tree.bio.ed.ac.uk/software/figtree/

Rivarola-Duarte L., Otto C., Jühling F., Schreiber S., Bedulina D., Jakob L., Gurkov A., … Sartoris F., Pörtner H.O., Timofeyev M., Luckenbach T., Stadler P.F. A first Glimpse at the genome of the Baikalian amphipod *Eulimnogammarus verrucosus*. *J Exp Zool B Mol Dev Evol*. 2014;322(3):177-189. doi 10.1002/jez.b.22560

Romanova E.V., Sherbakov D.Y. Different rates of molecular evolution of mitochondrial genes in Baikalian and non-Baikalian amphipods. *Limnol Freshwater Biol*. 2019;(6):339-344. doi 10.31951/2658-3518-2019-A-6-339

Romanova E.V., Aleoshin V.V., Kamaltynov R.M., Mikhailov K.V., Logacheva M.D., Sirotinina E.A., Gornov A.Y., Anikin A.S., Sherbakov D.Y. Evolution of mitochondrial genomes in Baikalian amphipods. *BMC Genomics*. 2016;17(Suppl. 14):1016. doi 10.1186/s12864-016-3357-z

Romanova E.V., Bukin Y.S., Mikhailov K.V., Logacheva M.D., Aleoshin V.V., Sherbakov D.Y. Hidden cases of tRNA gene duplication and remolding in mitochondrial genomes of amphipods. *Mol Phylogenet Evol*. 2020;144:106710. doi 10.1016/j.ympev.2019.106710

Romanova E.V., Bukin Y.S., Mikhailov K.V., Logacheva M.D., Aleoshin V.V., Sherbakov D.Y. The mitochondrial genome of a freshwater pelagic amphipod *Macrohectopus branickii* is among the longest in Metazoa. *Genes* (*Basel*). 2021;12(12):2030. doi 10.3390/genes12122030

Shao R., Dowton M., Murrell A., Barker S.C. Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. *Mol Biol Evol*. 2003;20(10):1612-1619. doi 10.1093/molbev/msg176

Sherbakov D.Y. Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal. *Trends Ecol Evol*. 1999;14(3):92-95. doi 10.1016/s0169-5347(98)01543-2

Sterling-Montealegre R.A., Prada C.F. Variability and evolution of gene order rearrangement in mitochondrial genomes of arthropods (except Hexapoda). *Gene*. 2024;892:147906. doi 10.1016/j.gene.2023.147906

Tan M.H., Gan H.M., Lee Y.P., Poore G.C.B., Austin C.M. Digging deeper: new gene order rearrangements and distinct patterns of codons usage in mitochondrial genomes among shrimps from the Axi-

idea, Gebiidea and Caridea (Crustacea: Decapoda). *PeerJ*. 2017;5: e2982. doi 10.7717/peerj.2982

Tan M.H., Gan H.M., Lee Y.P., Bracken-Grissom H., Chan T.Y., Miller A.D., Austin C.M. Comparative mitogenomics of the Decapoda reveals evolutionary heterogeneity in architecture and composition. *Sci Rep*. 2019;9(1):10756. doi 10.1038/s41598-019-47145-0

The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res*. 2024;52(W1):83-94. https://doi.org/10.1093/nar/gkae410

Wang R., Li X., Qi J. The complete paternally inherited mitochondrial genomes of three clam species in genus *Macridiscus* (Bivalvia: Veneridae): a TDRL model of dimer-mitogenome rearrangement of doubly uniparental inheritance. *Front Mar Sci*. 2022;9:1016779. doi 10.3389/fmars.2022.1016779

Xia Y., Zheng Y., Murphy R.W., Zeng X. Intraspecific rearrangement of mitochondrial genome suggests the prevalence of the tandem duplication-random loss (TDLR) mechanism in *Quasipaa boulengeri*. *BMC Genomics*. 2016;17(1):965. doi 10.1186/s12864-016-3309-7

Xu W., Jameson D., Tang B., Higgs P.G. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. *J Mol Evol*. 2006;63(3):375-392. doi 10.1007/s00239-005-0246-5

Zapelloni F., Jurado-Rivera J.A., Jaume D., Juan C., Pons J. Comparative mitogenomics in *Hyalella* (Amphipoda: Crustacea). *Genes* (*Basel*). 2021;12(2):292. doi 10.3390/genes12020292

Zardoya R. Recent advances in understanding mitochondrial genome diversity. *F1000Res*. 2020;9:270. doi 10.12688/f1000research.21490.1

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

Received December 23, 2024. Revised July 9, 2025. Accepted August 14, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2026;30(1):146-156

doi 10.18699/vjgb-26-16

Morphological and molecular genetic analysis of the genus *Iris* L. polymorphism in the Republic of Bashkortostan and the Orenburg Region

E.A. Aukhadieva[1], A.G. Blinov[2], E.I. Vshivtseva[2, 4] ✉, Z.Kh. Shigapov[3], A.R. Ishbirdin[5], Ya.M. Golovanov[3], A.V. Kryukova[3]

[1] Ufa Research Institute of Occupational Health and Human Ecology, Ufa, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] South-Ural Botanical Garden-Institute – Subdivision of the Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
[5] Ufa University of Science and Technology, Ufa, Russia
✉ evsivceva4@gmail.com

Abstract. *Iris* is a cosmopolitan genus comprising 200 to 340 species distributed throughout the Northern Hemisphere. Although *Iris* is the most diverse group in the family Iridaceae, there are many uncertainties regarding its taxonomic composition and systematics. The aim of this study was to search for taxonomically significant morphological characters of the generative and vegetative spheres and molecular markers with subsequent assessment of their informativeness in identifying phylogenetic relationships and compliance with the most relevant modern classification systems of the genus *Iris*. As a result of constructing the structure of variability of morphometric parameters of 11 species, 10 taxonomic indicators were identified that were common to the analyzed taxa and were characterized by relatively low total and coordinated variability: length and width of the outer perianth lobes, length and width of the inner perianth lobes, length of the filament, anther and pistil, fruit width, as well as seed length and width. Nucleotide sequences of *trnL-trnF* fragments of chloroplast DNA were established for 13 samples of four species of wild flora of the Republic of Bashkortostan and the Orenburg Region: *Iris pumila* L., *I. scariosa* Willd. ex Link., *I. pseudacorus* L., *I. sibirica* L. The obtained sequences were used to construct a phylogenetic tree together with *trnL-trnF* sequences of seven more iris species extracted from the database. The tree contained five clusters: (1) *I. pumila*, *I. scariosa*; (2) *I. pseudacorus*, *I. setosa* Pall. ex Link; (3) *I. lactea* Pall.; (4) *I. sibirica*, *I. sanguinea* Hornem.; (5) *I. spuria* L., *I. xanthospuria* Mathew & Baytop., *I. foetidissima* L., *I. sintenisii* Janka. By the composition of their species, the identified clusters almost completely coincided with the clusters found during the morphological analysis. To confirm the obtained results, a phylogenetic analysis of the species of interest was performed on two more chloroplast sequences available in the database: *matK* and *trnS-trnG*. Clustering of the studied species on *trnS-trnG* and *matK* completely coincided with clustering on *trnL-trnF*. Thus, we can state that the morphological features identified for the *Iris* generic complex work in the taxonomic direction. The analysis also showed that *I. scariosa* from natural populations of the Republic of Bashkortostan and the Orenburg Region were identified correctly.

Key words: *Iris*; taxonomy; taxonomic indicators; nucleotide sequences; *trnL-trnF* chloroplast DNA sequence; phylogenetic tree

For citation: Aukhadieva E.A., Blinov A.G., Vshivtseva E.I., Shigapov Z.Kh., Ishbirdin A.R., Golovanov Ya.M., Kryukova A.V. Morphological and molecular genetic analysis of the genus *Iris* L. polymorphism in the Republic of Bashkortostan and the Orenburg Region. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(1):146-156. doi 10.18699/vjgb-26-16

Funding. The work was carried out with the financial support of the state budget project FWNR-2022-0016.

Морфологический и молекулярно-генетический анализ полиморфизма рода *Iris* L. в Республике Башкортостан и Оренбургской области

Э.А. Аухадиева[1], А.Г. Блинов[2], Е.И. Вшивцева[2, 4] ✉, З.Х. Шигапов[3], А.Р. Ишбирдин[5], Я.М. Голованов[3], А.В. Крюкова[3]

[1] Уфимский научно-исследовательский институт медицины труда и экологии человека, Уфа, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Южно-Уральский ботанический сад-институт – обособленное структурное подразделение Уфимского федерального исследовательского центра Российской академии наук, Уфа, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[5] Уфимский университет науки и технологий, Уфа, Россия
✉ evsivceva4@gmail.com

Аннотация. *Iris* L. (Iridaceae Juss.) – это космополитный род, включающий от 200 до 340 видов, распространенных по всему Северному полушарию. Хотя *Iris* является самой разнообразной группой в семействе Iridaceae, существует множество неопределенностей относительно его таксономического состава и систематики. Цель

© Aukhadieva E.A., Blinov A.G., Vshivtseva E.I., Shigapov Z.Kh., Ishbirdin A.R., Golovanov Ya.M., Kryukova A.V., 2026
This work is licensed under a Creative Commons Attribution 4.0 License

настоящего исследования – поиск молекулярных маркеров и таксономически значимых морфологических признаков генеративной и вегетативной сферы и оценка их информативности для выявления филогенетического родства представителей рода *Iris*. Образцы ирисов взяты из природных популяций Республики Башкортостан и Оренбургской области, а также получены из ботанических садов Санкт-Петербурга, Бонна, Байройта и Брно. В результате построения структуры изменчивости морфометрических показателей 11 видов найдено 10 таксономических индикаторов, общих для анализируемых таксонов и характеризующихся относительно низкими общей и согласованной изменчивостью: длина и ширина наружных долей околоцветника, длина и ширина внутренних долей околоцветника, длина тычиночной нити, пыльника и пестика, ширина плода, а также длина и ширина семени. Установлены нуклеотидные последовательности *trnL-trnF* фрагментов хлоропластной ДНК для 13 образцов четырех видов дикорастущей флоры Республики Башкортостан и Оренбургской области: *Iris pumila* L., *I. scariosa* Willd. ex Link., *I. pseudacorus* L., *I. sibirica* L. Полученные последовательности были использованы для построения филогенетического дерева совместно с *trnL-trnF* последовательностями еще семи видов ирисов, извлеченных из базы данных. На филогенетическом древе формируется пять групп (кластеров): 1) *I. pumila* L., *I. scariosa* Willd. ex Link; 2) *I. pseudacorus* L., *I. setosa* Pall. ex Link; 3) *I. lactea* Pall.; 4) *I. sibirica* L., *I. sanguinea* Hornem.; 5) *I. spuria* L., *I. xanthospuria* Mathew & Baytop., *I. foetidissima* L., *I. sintenisii* Janka. Выявленные кластеры по составу входящих в них видов практически полностью совпадают с кластерами, обнаруженными при морфологическом анализе. Для подтверждения полученных результатов проведен филогенетический анализ интересующих видов еще на двух хлоропластных последовательностях, доступных в базе данных: *matK* и *trnS-trnG*. Кластеризация исследованных видов на *trnS-trnG* и *matK* полностью совпадает с кластеризацией на *trnL-trnF*. Таким образом, можно констатировать, что выявленные для родового комплекса *Iris* морфологические признаки демонстрируют значимую диагностическую ценность при проведении таксономических исследований.
Ключевые слова: *Iris*; таксономия; таксономические индикаторы; нуклеотидные последовательности; *trnL-trnF* последовательность хлоропластной ДНК; филогенетическое древо

Introduction

The genus *Iris* L. is the largest and most cosmopolitan in the family Iridaceae, distributed mainly in the temperate zones of the Northern Hemisphere, and includes from 200 to 340 species (Dorofeeva, Zhurbenko, 2020). Despite significant progress in the study of the genus *Iris*, there are still many uncertainties regarding its taxonomic composition and systematics. The genera boundaries of irises are controversial, and recent data appear to favor much narrower boundaries (Crespo et al., 2015; Boltenkov et al., 2020). In addition, its composition is periodically supplemented with new described species (Zhao, 1992; Mitic, 2002; Alexeeva, 2013a), which is often associated with morphological variability and, as a consequence, repeated descriptions of species with a wide range (Boltenkov et al., 2018a).

Classifications of the genus *Iris* are most often based on anatomical, morphological and cytogenetic characteristics (Doronkin, Krasnikov, 1984; Mathew, 1989; Makarevich et al., 2001), as well as on the results of molecular biological and biochemical studies (Dorogina et al., 2012; Weber et al., 2020). The difficulty is that systematics have both, broad and narrow understandings of the genus *Iris*. There are no commonly accepted classification, and the most popular classification schemes (Rodionenko, 1961; Mathew, 1989; Doronkin, 2006) have differences not only in position of individual species, but also higher taxonomical units – subgenera and sections. Attempts at resolving those contradictions using modern methods often provide ambiguous results. Molecular RAPD-analysis by I.F. Makarevich and colleagues (2001) showed greater agreement with the system of G.I. Rodionenko (1961) in establishing phylogenetic relationships.

Studies of Siberian species have revealed unexpected groupings: species from *Xyridion* and *I. sibirica* (*Limniris*) formed one group, while *I. lactea* and *I. setosa* (*Limniris*) formed another group with species from subgenus *Iris*. These data are contradicting with the existing schemes, especially in regard to polymorphic subgenus *Limniris*. The lack of consensus on the composition of the genus *Iris* among modern researchers requires further comprehensive studies to clarify phylogenetic relationships of the species included in this genus.

There are publications in the literature on how assessment of the structure of morphological variability can be used in biology for classifying morphological characteristics of some plant and animal species according to the ratio of general and coordinated variability. The authors identify four groups of indicators: ecological-biological systemic, biological, genetic (taxonomic), and ecological indicators (Rostova, 2002; Ishbirdin et al., 2005).

Ecological and biological systemic indicators are characterized by high general and coordinated variability, strong dependence on environmental conditions, and the ability to induce coordinated changes in the entire morphosystem of an organism (number of buds, shoot length in *C. rubra*, leaf blade length in *Triticum aestivum* L.).

Biological indicators have moderate environmental dependence, low total and high coordinated variability. They also determine the morphostructure of the plant (e.g., shoot height, leaf parameters in *Cephalanthera rubra*, the number of metameres in the vegetative part of the annual shoot in *Helianthus annuus*).

Genetic (taxonomic) indicators are characterized by low general and coordinated variability, high autonomy and weak dependence on external conditions (for example, the number of leaves in *C. rubra*), and are the most informative for taxonomic studies.

Ecological indicators exhibit strong, relatively independent variation, and their changes are only mildly correlated with the overall system of the organism, but they are sensitive

even to minor external influences (for example, the number of immature flowers, signs of branching in *Rhinanthus* L.). For species such as *H. annuus* L., *C. rubra* L., *Panicum miliaceum* L., and a number of other flowering plants, these traits have already been identified. For the *Iris* genus complex, a correlation analysis of morphological traits to identify these indicators was conducted for the first time.

DNA polymorphisms in noncoding regions of chloroplast genome are successfully used to establish phylogenetic relationships between species of the genus *Iris* (Pleines et al., 2009). Among the most often used regions of chloroplast DNA are the *trnL* intron and the *trnL-trnF* intergenic spacer, as well as other variable regions of chloroplast DNA including: *atpB-rbcL*, *trnS-trnG*, and *trnH-psbA* intergenic spacers (Kozyrenko et al., 2009; Boltenkov et al., 2016).

The aim of the current study is to search for taxonomically significant morphological features of the generative and vegetative spheres and molecular markers, followed by an assessment of their informativeness in identifying phylogenetic relationships and compliance with the most relevant modern classification systems of the genus *Iris*.

Materials and methods

Plant material. The objects of morphological studies were wild-growing representatives of species of the genus *Iris* (*I. pseudacorus* L., *I. sibirica* L., *I. pumila* L., *I. scariosa* Willd. ex Link), collected from natural populations of Bashkortostan Republic and Orenburg Region, and introduced to the collection area of the South-Ural Botanical Garden-Institute of the Ufa Federal Research Center of the Russian Academy of Sciences in 2019–2021. As well as from seed material grown in obtained from delectuses from the botanical gardens of St. Petersburg (*I. lactea* Pall., *I. setosa* Pall., *I. halophila* Pall.), Bonn (*I. sanguinea* Hornem., *I. spuria* L.), Bayreuth (*I. carthaliniae* Fomin) and Brno (*I. graminea* L.). Details on species used, collection locations and geographical coordinates are listed in Table 1.

Due to the identification of a large number of controversial issues regarding the taxonomy of species of the genus *Iris* L., in this work we focused on the classifications of three authors: B. Mathew (1989), G.I. Rodionenko (1961, 2013), and V.M. Doronkin (2006).

Morphological analysis. Morphometric parameters were recorded for 25 plants of each species. To assess variability,

Table 1. Iris samples used in this work, places and time of their collection

Species	Material collection location	Geographical coordinates
I. pseudacorus	Russia, Republic of Bashkortostan, Kushnarenkovsky district, Ilmurzino village	54.952323 N, 55.811026 E
	Russia, Republic of Bashkortostan, Kushnarenkovsky district, Taraberdino village	55.092243 N, 55.418299 E
	Russia, Republic of Bashkortostan, Demsky district of Ufa, Romanovka village	54.736672 N, 55.834274 E
I. sibirica	Russia, Republic of Bashkortostan, Birsk city	55.454523 N, 55.537686 E
	Russia, Republic of Bashkortostan, Agidel city	55.844276 N, 53.935868 E
I. scariosa	Russia, Republic of Bashkortostan, Khaibullinsky district, Tashtugai mountains	51.910798 N, 58.494056 E
	Russia, Orenburg Region, Kuvandyk district, Ramazanovo village	51.526943 N, 57.446162 E
	Russia, Orenburg Region, Svetlinsky district, Tobolsky settlement	51.430173 N, 61.164358 E
I. pumila (yellow flowers)	Russia, Republic of Bashkortostan, Kuyurgazinsky district, Lena village	52.802508 N, 55.603231 E
I. pumila (blue flowers)	Russia, Republic of Bashkortostan, Kuyurgazinsky district, Lena village	52.802508 N, 55.603231 E
I. pumila (yellow flowers)	Russia, Republic of Bashkortostan, Kuyurgazinsky district, Yakshimbetovo village	52.578021 N, 55.654848 E
I. pumila (blue flowers)	Russia, Republic of Bashkortostan, Kuyurgazinsky district, Yakshimbetovo village	52.578021 N, 55.654848 E
I. pumila (yellow flowers)	Russia, Orenburg Region, Perevolotsky district, village of Rychkovka	51.721765 N, 54.523130 E

the following parameters were analyzed: generative shoot height, leaf length and width, perianth segment length and width, reproductive element length (pistil and stamen) during the flowering phase, and fruit and seed length and width after full ripening. Measurements of morphometric parameters of shoots, flowers and fruits were carried out using a ruler, and those of seeds, using a Levenhuk DTX 90 microscope (Golubev, 1962). Standard statistical processing of the obtained data was carried out using the programs MS Excel and IBM SPSS Statistics 21 (Zaitsev, 1984; Dospekhov, 1985). The arithmetic mean, standard error of the mean, and standard deviation were calculated. To assess intrapopulation variability of morphological traits, the coefficient of variation (CV, %) was calculated (Zaitsev, 1990).

A comparison of biometric indicators was conducted to determine the statistical significance of their differences across 2019–2021 years. The values of the studied indicators were tested for normal distribution using the Kolmogorov–Smirnov test. To compare independent samples with a normal distribution, a one-way analysis of variance was used, and for samples that do not obey the law of normal distribution, the Kruskal–Wallis test was used. The analysis showed that for all the parameters studied, the difference is insignificant at a significance level of W = 5 %, which makes it possible to assess the structure of variability of morphological characteristics and classify them into groups.

The structure of trait variability is constructed using the method of N.S. Rostova (2002), using the programs MS Excel and IBM SPSS Statistics 21. Morphological traits are coordinated in the space of total (the trait's coefficient of variation) and consistent (the squared correlation coefficient averaged over the trait) variability. According to their indicator role, traits are divided into three groups: taxonomic (genetic), biological, and ecological-biological systemic.

Isolation of total DNA, PCR amplification and determination of nucleotide sequences and phylogenetic analysis. Total DNA was isolated using the DNeasy Plant Mini Kit (QIAGEN, Germany) according to the manufacturer's protocol. 50–100 mg of dried leaves from each plant were used for DNA extraction. The quantity and quality of the isolated DNA were determined using a NanoDrop2000 spectrophotometer (Thermo Scientific, USA) and electrophoretic separation in a 1 % agarose gel containing ethidium bromide (0.5 mg/ml) in 1xTAE.

PCR amplification was performed as described in (Makarevich et al., 2003) with primers specific for the *trnL/trnF* chloroplast genes:

trnL (5′-CGAAATCGGTAGACGCTACG-3′);
trnF (5′-ATTTGAACTGGTGACACGAG-3′).

The reaction was performed in a 20 µl volume using BioMaster LR HS-PCR-Color (Biolabmix, Novosibirsk, Russia) with 10 pmol of each primer and 30 ng of total DNA. The resulting PCR fragments were separated on a 1 % agarose gel and isolated from the gel using the QIAquick Gel Extraction Kit (QIAGEN, Germany), followed by sequencing. Sequencing reactions were performed using 200 ng of DNA fragment and the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Scientific, USA) on an ABI 3130XL genetic analyzer (Applied Biosystems, USA) at the Genomics Center of the Siberian Branch of the Russian Academy of Sciences (http://www.niboch.nsc.ru/doku.php/corefacility). The nucleotide sequences of the *trnL-trnF* fragments of chloroplast DNA are presented in GenBank (No. PV335670–PV335682).

For phylogenetic analysis, *trnL-trnF* iris chloroplast DNA sequences retrieved from the GenBank database and obtained in this study were used. Nucleotide sequence alignment was performed using MAFFT v7.312 (Katoh, Standley, 2013). Suitable nucleotide substitution model was selected using the Bayesian Information Criterion. The phylogenetic tree was constructed using the maximum likelihood method with the Jukes–Cantor model in the IQ-tree program (Trifinopoulos et al., 2016). The reliability of the constructed tree was tested using the Bootstrap method with a number of repetitions equal to 1,000.

Results

Analysis of morphometric parameters

Morphological data for three years of research were analyzed. Biometric indicators were compared for statistical significance of differences between the years. The analysis showed that for all studied parameters, the differences were insignificant at a significance level of W = 5 %. This allowed us to assess the structure of morphological trait variability based on the ratio of total and consistent variability and classify them into groups. Generalized morphometric parameters of studied species are listed in the Table 2.

An analysis of the variability of morphometric parameters of species of the genus *Iris* revealed the following patterns:

1. The highest variability level is typical for the length of a leaf blade (CV reaches 20.7 %) and the height of generative shoots (CV up to 16.7 %).
2. The most stable indicators – width of the inner perianth lobes (CV = 2.2–5.3 %) and pistil length (CV = 0.8–3.4 %), demonstrating the smallest spread of the variation coefficient values.
3. Fruit and seed parameters show higher variability in fruit length (CV up to 9.4 %), compared to its width (CV up to 5.8 %), and moderate variability in seed size (CV in the range of 0.9–5.9 %).

Morphological features are coordinated in the space of general (coefficient of variation of the feature) and consistent (squared correlation coefficient averaged for the feature) variability. Graphs were compiled showing the structure of morphological trait variability. As an example, Figure 1 shows the structure of variability for *I. pseudacorus*. Variability structure for *I. sibirica*, *I. sanguinea*, *I. setosa*, *I. halophila*, *I. graminea*, *I. carthaliniae*, *I. spuria*, *I. pumila*, *I. scariosa*, *I. lactea* is given in the Supplementary Materials[1].

[1] Supplementary Materials are available at:
https://vavilovj-icg.ru/download/pict-2026-30/appx10.pdf

Table 2. Generalized morphometric parameters of species from genus *Iris* (2019–2021)

Parameter	Value ranges (min–max)	Coefficients of variation (CV, %)
Aboveground shoots		
Height of generative shoot, cm	9.0–97.5	8.5–16.7
Leaf length, cm	7.1–97.2	6.8–20.7
Lead width, cm	0.5–1.9	4.1–10.9
Flower		
Length of outer lobes, cm	4.0–6.1	1.4–4.2
Width of outer lobes, cm	1.1–3.3	0.8–6.0
Length of inner lobes, cm	1.0–5.1	1.6–6.7
Width of inner lobes, cm	0.2–2.2	2.2–5.3
Length of stamen filament, cm	1.1–1.6	1.4–3.8
Anther length, cm	1.1–2.0	1.2–5.0
Pistil length, cm	2.8–4.4	0.8–3.4
Fruits and seeds		
Fruit length, cm	3.3–6.3	3.1–9.4
Fruit width, cm	1.0–2.1	0.7–5.8
Seed length, cm	0.4–0.8	0.9–4.5
Seed width, cm	0.2–0.8	1.2–5.9

In the current study, according to their indicator role, the features were divided into three groups: taxonomic (genetic), biological and ecological-biological (Fig. 1). As a result of constructing the variability structure, 10 taxonomic indicators were identified that are common to the analyzed taxa and characterized by relatively low total (CV = 0.8–6.0 %) and consistent (r_m^2 = 0.005–0.078) variability: the length and width of the outer perianth lobes, the length and width of the inner perianth lobes, the length of the filament, anther and pistil, the width of the fruit, and the length and width of the seed.

To establish the relationship of the studied representatives of the genus *Iris*, a cluster analysis (hierarchical classification, Ward's method) was carried out and a dendrogram of the differences and similarities between the species was constructed based on the identified diagnostic markers (Fig. 2). The species in the dendrogram can be grouped into six clusters:

1. *I. carthaliniae, I. halophila, I. spuria*, representatives of subgenus *Xyridion* (Rodionenko, 1961) or *Limniris* (Mathew, 1989).
2. *I. lactea*, subgenus *Limniris* (Rodionenko, 1961; Mathew, 1989) or *Eremiris* (Doronkin, 2006).
3. *I. sibirica*, *I. sanguinea*, subgenus *Limniris*.
4. *I. pumila*, *I. scariosa*, subgenus *Iris*.
5. *I. graminea*, subgenus *Xyridion* or *Limniris*.
6. *I. pseudacorus*, *I. setosa*, subgenus *Limniris*.

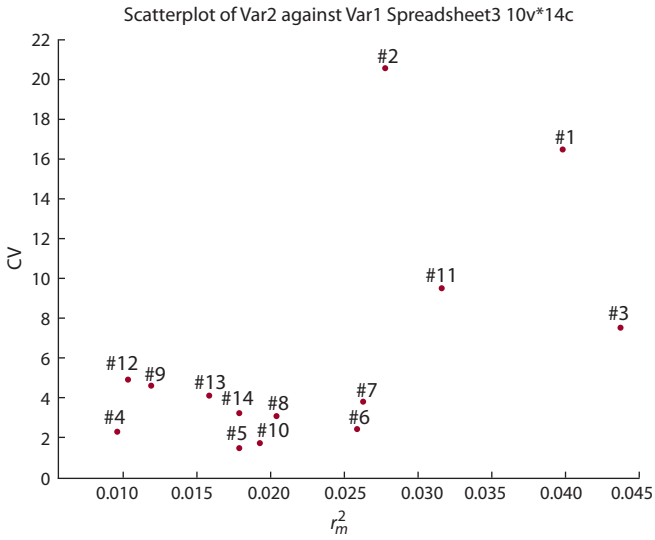


Fig. 1. Structure of variability of morphological characteristics of *I. pseudacorus.*

The *X*-axis shows the consistent variability, the *Y*-axis shows the total variability. 1 – height of generative shoot, 2 – leaf length, 3 – leaf width, 4 – length of outer perianth lobes, 5 – width of outer perianth lobes, 6 – length of inner perianth lobes, 7 – width of inner perianth lobes, 8 – length of filament, 9 – length of anther, 10 – length of pistil, 11 – length of fruit, 12 – width of fruit, 13 – length of seed, 14 – width of seed.

Molecular phylogenetic analysis of species of the genus *Iris*

During the current project, the nucleotide sequences of *trnL-trnF* fragments of chloroplast DNA were established for 13 samples of 4 species of irises: *I. pumila*, *I. pseudacorus*, *I. sibirica*, *I. scariosa* (Table 1). The obtained sequences were used to construct a phylogenetic tree together with the *trnL-trnF* sequences of 7 more iris species retrieved from the database: *I. setosa*, *I. lactea*, *I. sanguinea*, *I. spuria*, *I. xanthospuria* Mathew & Baytop., *I. foetidissima* L., *I. sintenisii* Janka. As a

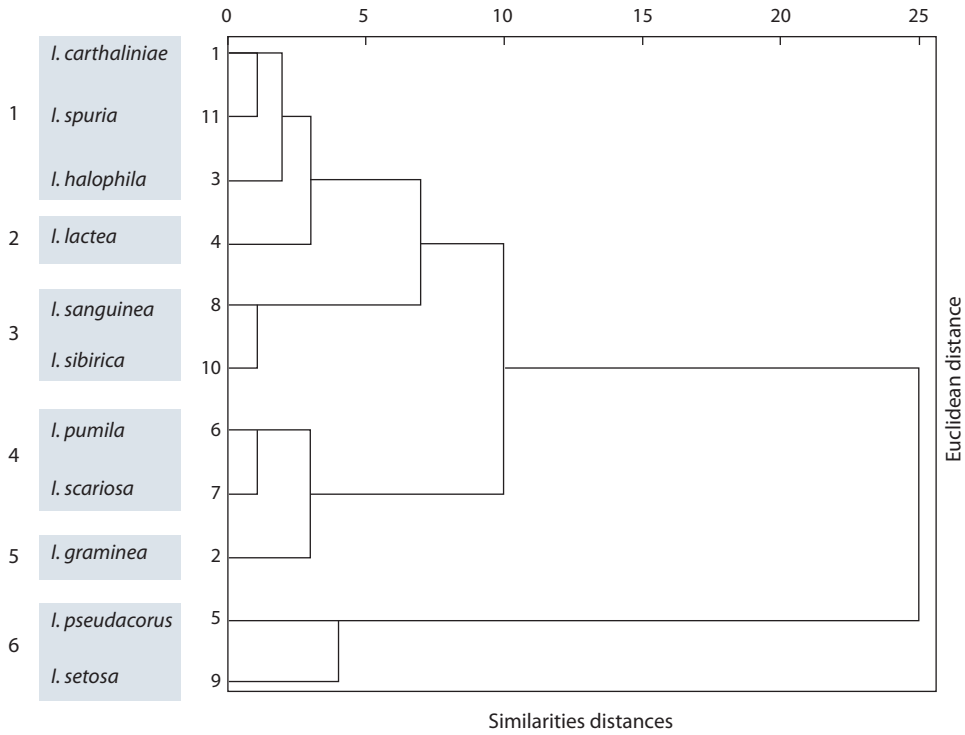


Fig. 2. Dendrogram of differences and similarities of species of the genus *Iris* by taxonomic indicators.

result of this study, *trnL-trnF* sequences from 63 accessions of 11 species were analyzed. The resulting phylogenetic tree is presented in Figure 3. The collection locations of the studied iris accessions are marked on the tree.

There are five clusters on the tree:

1. *I. spuria*, *I. xanthospuria*, *I. sintenisii*, *I. foetidissima*.
2. *I. lactea*.
3. *I. sibirica* and *I. sanguinea*.
4. *I. pumila* and *I. scariosa*.
5. *I. pseudacorus* and *I. setosa*.

These clusters, in terms of the composition of their species, almost completely coincide with the clusters found during morphological analysis (Fig. 2).

To confirm our results, we conducted a phylogenetic analysis of the *Iris* species of interest using two additional chloroplast sequences available in the NCBI database: *matK* and *trnS-trnG*. The dendrogram, obtained using *matK* sequences, showed that the samples split in five groups, similar to the grouping for *trnL-trnF* sequences (Fig. 4):

1. *I. spuria* and *I. halophila*.
2. *I. lactea*.
3. *I. sibirica* and *I. sanguinea*.
4. *I. pumila* and *I. scariosa*.
5. *I. pseudacorus* and *I. setosa*.

However, there are minor differences between the *trnL-trnF* and *matK* trees. In the pairs *I. pumila/I. scariosa*, *I. sibirica/I. sanguinea*, and *I. halophila/I. spuria*, the *matK* sequences do not confirm species differences. While the species *I. pseudacorus* and *I. setosa* are reliably different from each other. If interpopulation differences in the *trnL-trnF* sequences were found in only two species (*I. scariosa* and *I. setosa*), such differences in the *matK* sequences are present in a larger number of species: *I. setosa*, *I. pseudacorus*, *I. sibirica*, *I. sanguinea*, and *I. lactea*. They may also be present in *I. pumila* and *I. scariosa*, but we found only three *matK* gene sequences from these species in the database.

There is less information about the *trnS-trnG* sequences in the database than about the previous two. Figure 5 shows a tree constructed from the *trnS-trnG* sequences. It contains four out of the five clusters obtained in the previous trees: (1) *I. lactea*; (2) *I. sibirica* and *I. sanguinea*; (3) *I. pumila* and *I. scariosa*; (4) *I. pseudacorus* and *I. setosa*. Cluster five is missing because there are no *trnS-trnG* sequences of *I. spuria* and *I. halophile* in the database. We will not go into detail on the analysis of the *trnS-trnG* tree due to the small number of samples for some species. However, the main conclusion is clear: the clustering of the studied species on the *trnS-trnG* tree is completely consistent with that on the *trnL-trnF* and *matK* trees.

Discussion

Morphometric parameters

A comparative analysis of the morphometric parameters of the studied species with published data (Volkova, 2010) revealed a high degree of correspondence in the aboveground traits for most species, with minor deviations observed in some taxa. The detected differences are presumably determined by the specific cultivation conditions, which are supported, in particular, by the influence of temperature amplitude on the variability

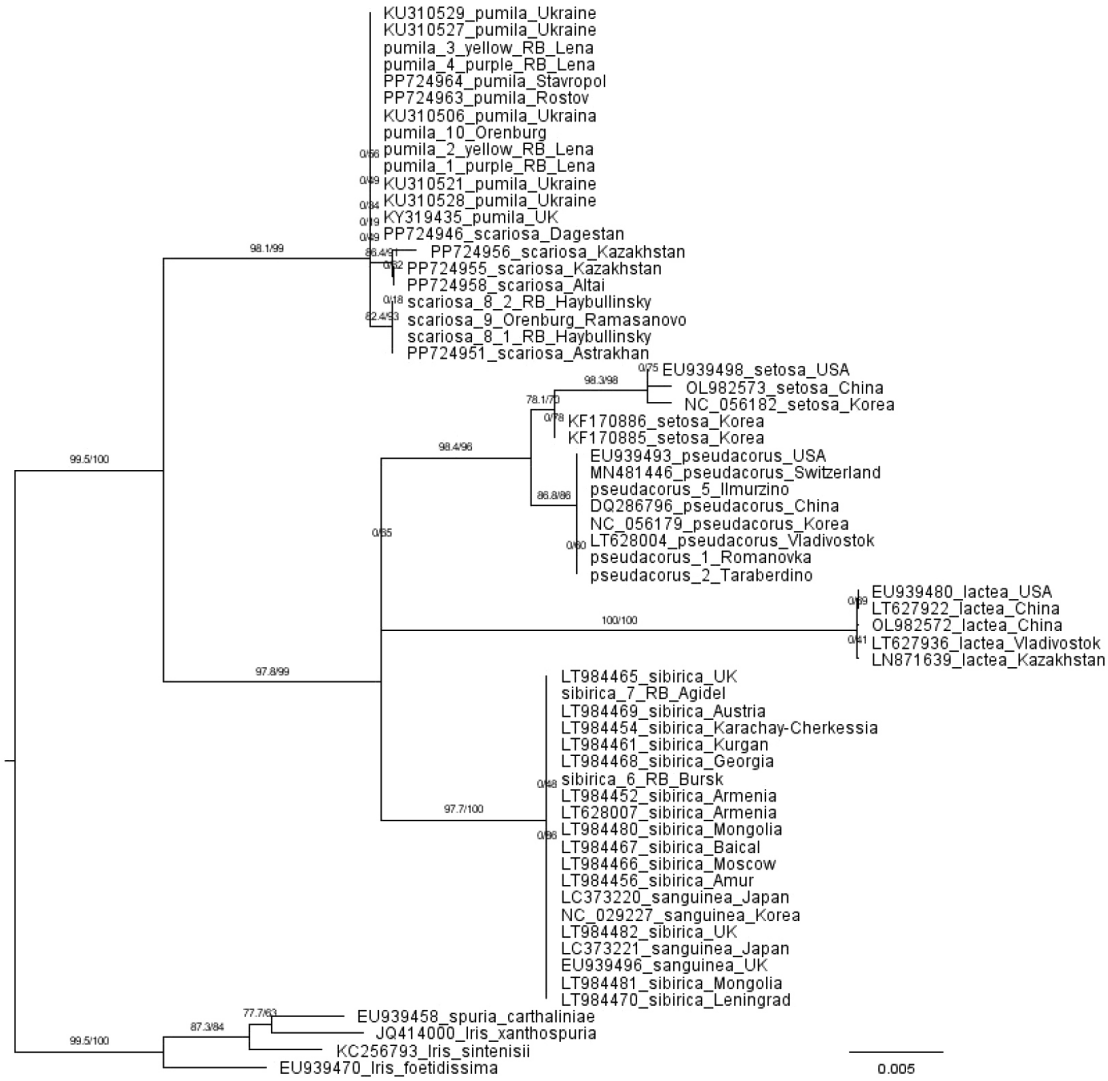


Fig. 3. Phylogenetic tree constructed from iris *trnL-trnF* cpDNA sequences using the ML method.
Collection locations and nucleotide sequence numbers from Genbank database are marked on the tree.

of morphometric parameters of the generative shoot and leaf in *I. pumila* from populations of the Lower Volga region and the Southern Urals (Kashin et al., 2022).

Based on the structure and morphometry of the perianth, the species were divided into conventional groups: (1) with long narrow segments (*I. carthaliniae*, *I. halophila*, *I. spuria*, *I. graminea*, *I. lactea*); (2) with large outer and reduced inner segments (*I. pseudacorus*, *I. setosa*); (3) with large broad segments (*I. sibirica*, *I. sanguinea*); (4) with medium-sized segments (*I. pumila*, *I. scariosa*), which indicates a high degree of reliability and validity of modern classifications of the genus.

A comparative analysis of the morphometric characteristics of fruits and seeds of *Iris* from natural populations of the Republic of Bashkortostan and the Orenburg Region with published data from other regions (Alexeeva, 2020) revealed moderate geographic polymorphism. In terms of seed size, the smallest seeds are characteristic of populations of *I. sibirica* (Saratov Region), *I. halophila* and *I. pumila* (Volgograd Region); a large-seeded form was recorded in *I. pseudacorus* (Primorsky Krai). Regarding fruit parameters, a slight reduction was observed in *I. setosa*, *I. graminea*, *I. spuria*, and *I. pumila* (Belgorod Region). The detected slight variability in size, as well as in the pigmentation of reproductive structures, reflects micropopulation differentiation of the species and local adaptive responses to growing conditions, ensuring their successful reproduction under diverse environmental conditions.

The coefficient of variation of morphological traits serves as a useful tool for the preliminary assessment of the stability of morphological characteristics, however, it is incorrect to judge the taxonomic significance of traits solely on the basis of the coefficient of variation. Taxonomic decisions must be based on a synthesis of morphological, genetic, and ecological data. For the genus *Iris*, this is particularly relevant due to its polymorphism and tendency toward hybridization. The method for identifying taxonomic characters proposed by N.S. Rostova (2002) is based on the analysis of correlation

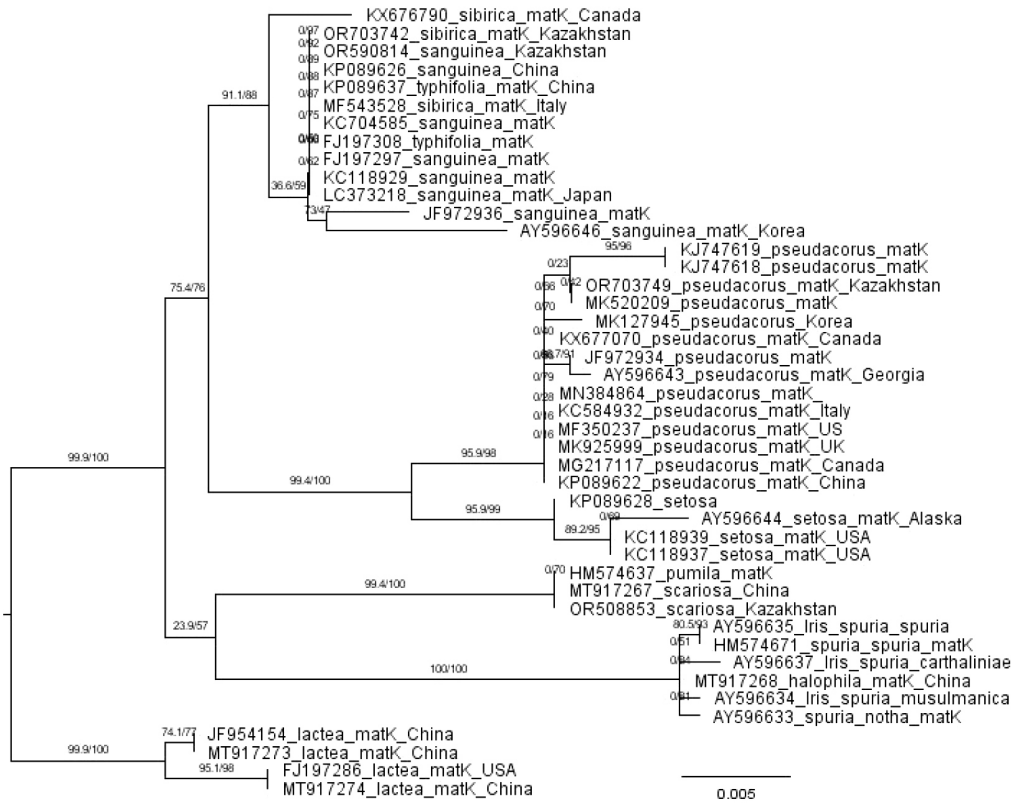


Fig. 4. Phylogenetic tree constructed from the *matK* nucleotide DNA sequences of irises using the ML method.
Collection locations and nucleotide sequence numbers are marked on the tree.

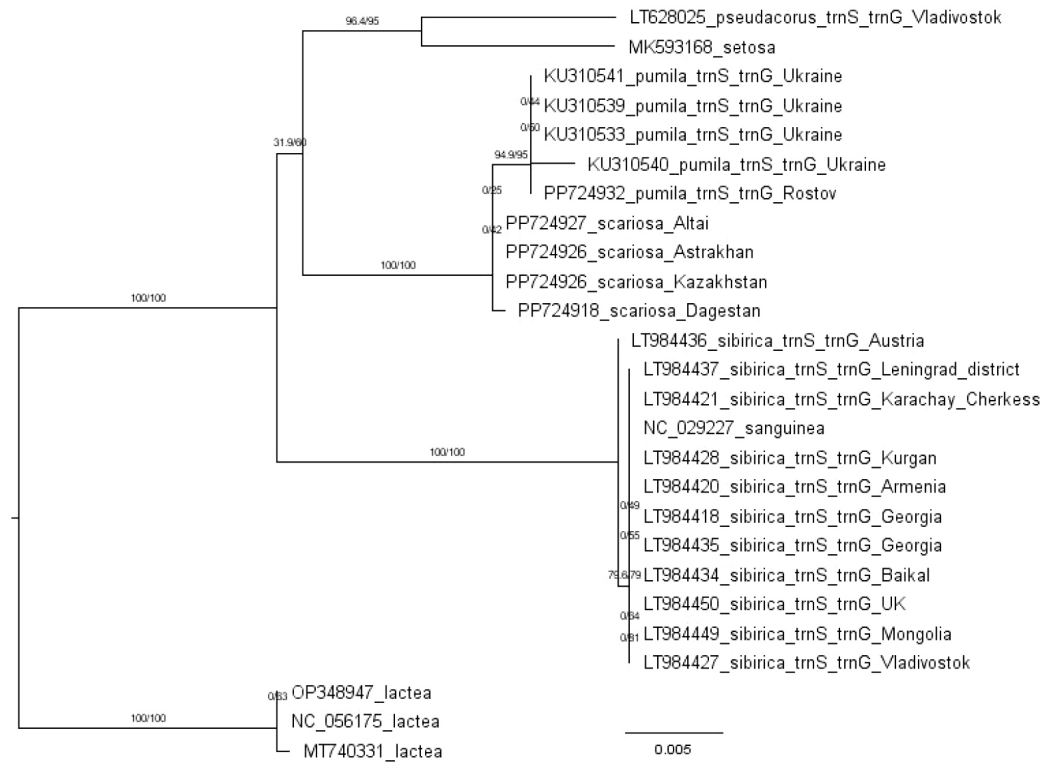


Fig. 5. Phylogenetic tree constructed from *Iris trnS-trnG* cpDNA sequences using the ML method.
Collection locations and nucleotide sequence numbers are marked on the tree.

pleiades – groups of interrelated morphological traits, which detects stable relationships among characters. The method helps distinguish conservative taxonomic traits from those dependent on environmental conditions.

It should be noted that *I. scariosa* is characterized by high variability of morphological traits, especially, like the closely related *I. pumila*, in flower coloration. The species was described from the westernmost margin of its main range based on a specimen collected by Pallas east of the Volga River. It occurs predominantly in the northwestern part of the Caspian Lowland and in the Eastern Ciscaucasia; it is an endemic Caspian European–Caucasian species (Red Book…, 2024).

The taxonomic position of the closely related species *I. scariosa* and *I. glaucescens* Bunge presents certain difficulties. Both species are either synonymized or considered separate (Tsvelev, 1979; Alexeeva, 2020), with *I. scariosa* being the preferred species. *I. glaucescens* was described in Central Asia. The species is represented by small populations found in Russia at the edge of its range in the south of Western Siberia. Outside of Russia, it is known in northeastern Kazakhstan, northwestern Mongolia, and China (Alexeeva, 2020). A similar trend is observed in the Southern Urals: in the Republic of Bashkortostan, *I. scariosa* is included in the regional Red Book (Red Book…, 2021–2024); in the neighboring Chelyabinsk Region, *I. glaucescens* is also listed (Kulikov, 2005). Thus, the taxonomic distinction between these closely related species is challenging. There is evidence in the literature, supported by molecular, morphological, and palynomorphological analysis, that *I. glaucescens*, as well as *I. timofejewii* and *I. curvifolia*, are synonyms of *I. scariosa* (Boltenkov, Artyukova, 2024).

As follows from the dendrogram of species similarities-differences based on the identified diagnostic markers, the closest phylogenetic relationships are observed between the following pairs of species: *I. carthaliniae* and *I. spuria*, *I. sibirica* and *I. sanguinea*, *I. pumila* and *I. scariosa*. Their affinity is also evident in the general habit of the plants. In this way, the pair *I. carthaliniae* and *I. spuria* consists of tall plants characterized by long rhizomes with thickened segments, robust, branched stems, broad sword-shaped leaves, and multi-flowered inflorescences (3–5 flowers). The pair *I. sibirica* and *I. sanguinea* comprises medium-sized irises with short rhizomes bearing narrow segments, hollow, weakly branched stems, narrow linear leaves arranged in basal clumps, and few-flowered inflorescences (2–3 flowers). The pair *I. pumila* and *I. scariosa* includes low-growing plants with thick creeping rhizomes, bearing very short stems, broad-linear or lanceolate leaves in a basal tuft, and large (relative to plant size), few (1–2) flowers positioned close to the ground.

Molecular phylogenetic analysis of species from the genus *Iris*

The results of morphological analysis showed that *I. sibirica* and *I. sanguinea* are very close to each other, and this fact was confirmed by the results of molecular phylogenetic analysis. Moreover, according to the obtained data, we can assume that *I. sibirica* and *I. sanguinea* are the same species, since the *trnL-trnF* sequences of all analyzed samples of these species are identical, regardless of the collection location. Other researchers had previously come to this conclusion by analyzing a number of morphological characteristics and using molecular phylogenetic data (Boltenkov et al., 2020).

A very similar situation is observed in the second pair of closely related species – *I. pumila* and *I. scariosa*. All analyzed *I. pumila* specimens, both those with yellow and those with purple flowers, have identical *trnL-trnF* sequences. The species is polychrome and is characterized by color polymorphism, which is determined by the normal reaction of individuals to environmental conditions (Kashin et al., 2022), and variations in the color of the perianth segments are not associated with molecular genetic polymorphism of populations.

Three sequence variants have been identified in *I. scariosa*. The first variant (from the Republic of Dagestan) is completely identical to the *trnL-trnF* sequence of *I. pumila*, which likely indicates that it is *I. pumila*. The other two variants differ from *I. pumila* by substitutions of several nucleotides. The second variant of *I. scariosa* was found in samples from the Republic of Kazakhstan and the Altai Republic, and the third variant was found in samples from the Republic of Bashkortostan and the Orenburg Region, and the last two variants differ from each other by a single nucleotide substitution. Thus, a comparative analysis of the noncoding sequences of the *trnL-trnF* region of the chloroplast DNA of *I. scariosa* from natural populations, as well as an analysis of literary data, shows that this species is correctly identified in the Republic of Bashkortostan and the Orenburg Region.

The situation is completely different in the third pair of related species – *I. setosa* and *I. pseudacorus*. First of all, these are evidently two different, albeit closely related, species, which is clearly visible on the phylogenetic tree and has been previously noted by other authors (Choi, Lee, 2024). However, these species differ from each other in terms of intraspecific variability. All *I. pseudacorus* samples analyzed showed identical *trnL-trnF* sequences, while five *I. setosa* samples analyzed showed the presence of two types of *trnL-trnF* sequences, with all samples collected either in China or Korea. *I. setosa* is heterogeneous and taxonomically quite clearly delimited from other species, because of this new species and intraspecific taxa have been identified from its composition (Ilyushko et al., 2001). N.B. Alexeeva (2013b) came to the conclusion that the polymorphic complex of *I. setosa* consists of five species that differ in morphology and ecology. Of course, to verify this, more factual material needs to be analyzed.

All five *trnL-trnF* sequences of *I. lactea* are identical to each other, regardless of their habitat. *Iris* Linnaeus ser. *Lacteae* Doronkin (Doronkin, 1990) includes species distributed in the temperate Asian regions of the Northern Hemisphere (Rodionenko, 2006; Boltenkov, 2018). Initially, *I. lactea* was classified in the subgenus *Limniris*, section *Limniris* (Rodionenko, 1961; Mathew, 1989), then in the subgenus *Eremiris*, section *Haloiris* (Doronkin, 2006). Subsequently, it began to be considered as part of a separate genus *Eremiris* (Rodionenko, 2006), with which C.A. Wilson (2011), based on molecular genetic studies, does not agree. For a long time, it was believed that the *Lacteae* Doronkin series was represented by only one

polymorphic species, but molecular studies have confirmed phylogenetic branching in at least three lineages corresponding to the taxa *I. lactea*, *I. oxypetala* and *I. tibetica* (Boltenkov et al., 2018b).

Thus, we can state that morphological and molecular phylogenetic analyses show the same results regarding the phylogeny analysis of the studied species of the genus *Iris*, which means that the identified morphological indicators work in a taxonomic direction.

Conclusion

Based on the conducted assessment of the structure of morphological variability of representatives of the genus *Iris*, 10 taxonomic indicators were identified: the length and width of the outer perianth lobes, the length and width of the inner perianth lobes, the length of the filament, anther and pistil, the width of the fruit, and the length and width of the seed.

The establishment of phylogenetic relationships of the studied species based on the identified taxonomic indicators and the molecular phylogenetic analysis based on *trnL-trnF*, *matK* and *trnS-trnG* chloroplast DNA made it possible to combine the studied *Iris* species into several phylogenetic groups. Thus, it can be concluded that both morphological and molecular phylogenetic analyses yield consistent results regarding the phylogeny of the studied species of the genus *Iris*, which indicates that the identified morphological indicators reliably demonstrate taxonomic validity.

References

Alexeeva N.B. *Iris lokiae*, a new species of genus *Iris* (*Iridaceae*). *Botanicheskii Zhurnal.* 2013a;98(11):1415-1420 (in Russian)

Alexeeva N.B. On the taxa of *Iris setosa* Pall. ex Link affinity (Iridaceae). *Turczaninowia*. 2013b;16(2):13-29 (in Russian)

Alexeeva N.B. *Iris* L. (Iridaceae Juss.) in the Russia. Saint Petersburg, 2020 (in Russian)

Boltenkov E.V. Taxonomic notes on *Iris* ser. *Lacteae* (Iridaceae) with typifications of fifteen names and one new combination. *Phytotaxa*. 2018;383(3):283-292. doi 10.11646/phytotaxa.383.3.5

Boltenkov E.V., Artyukova E.V. Updated taxonomy of *Iris scariosa* (Iridaceae) inferred from morphological and chloroplast DNA sequence data with remarks on classification of *Iris* subg. *Iris*. *Plants.* 2024;13(17):2349. doi 10.3390/plants13172349

Boltenkov E.V., Artyukova E.V., Kozyrenko M.M. Species divergence in *Iris* series *Lacteae* (Iridaceae) in Russia and adjacent countries based on chloroplast DNA sequence data. *Russ J Genet*. 2016;52(5): 507-516. doi 10.1134/S1022795416040037

Boltenkov E.V., Artyukova E.V., Kozyrenko M.M. A taxonomic study of the genus *Iris* (Iridaceae): morphological and molecular data. In: Botany in the Contemporary World. Vol. 1. Mahachkala, 2018a;13-16. Available: https://www.binran.ru/files/publications/Proceedings/Proceedings_RBO/XIV_RBO_Proceedings_T1.pdf (in Russian)

Boltenkov E.V., Artyukova E.V., Kozyrenko M.M., Trias-Blasi A. *Iris tibetica*, a new combination in *I.* ser. *Lacteae* (Iridaceae) from China: evidence from morphological and chloroplast DNA analyses. *Phytotaxa*. 2018b;338(3):223-240. doi 10.11646/phytotaxa.338.3.1

Boltenkov E., Artyukova E., Kozyrenko M., Erst A., Trias-Blasi A. *Iris sanguinea* is conspecific with *I. sibirica* (Iridaceae) according to morphology and plastid DNA sequence data. *Peer J.* 2020;8:e10088. doi 10.7717/peerj.10088

Choi T.-Y., Lee S.-R. Complete plastid genome of *Iris orchioides* and comparative analysis with 19 *Iris* plastomes. *PLoS One.* 2024;19(4): e0301346. doi 10.1371/journal.pone.0301346

Crespo M.B., Martínez-Azorín M., Mavrodiev E.V. Can a rainbow consist of a single colour? A new comprehensive generic arrangement of the '*Iris sensu latissimo*' clade (Iridaceae), congruent with morphology and molecular data. *Phytotaxa*. 2015;232(1):1-78. doi 10.11646/phytotaxa.232.1.1

Dorofeeva M.M., Zhurbenko P.M. Comparative study of ovule structure and development in some species of *Iris* subgenus *Limniris* (Iridaceae). *Botanicheskii Zhurnal.* 2020;105(1):15-31. doi 10.31857/S0006813620010068

Dorogina O.V., Doronkin V.M., Selyutina I.Yu., Konichenko E.S. Structure of the genus *Iris* L. (Iridaceae Juss.) in Asiatic Russia, revealed by SDS-electrophoresis of seed storage proteins. *Turczaninowia*. 2012;15(4):76-81 (in Russian)

Doronkin V.M. The synopsis of Siberian species of the genus *Iris* (Iridaceae). *Botanicheskii Zhurnal.* 1990;75(3):409-416 (in Russian)

Doronkin V.M. Taxonomy of the genus *Iris* L. (Iridaceae Juss.) in Asian Russia. In: Role of Botanical Gardens in the Preservation of Vegetation Diversity in Asian Russia: the Present and the Future. Novosibirsk, 2006;101-103 (in Russian)

Doronkin V.M., Krasnikov A.A. Cytotaxonomic studies in some Siberian species of the genus *Iris* (Iridaceae). *Botanicheskii Zhurnal*. 1984;69(5):683-685 (in Russian)

Doronkin V.M., Baikov K.S. Modeling phylogenetic relationships of Siberian species of the genus *Iris* (Iridaceae) using SYNAP method. *Turczaninowia.* 2004;7(2):45-57 (in Russian)

Dospekhov B.A. Methodology of Field Experiments (with the Basics of Statistical Processing of Research Results). Moscow, 1985 (in Russian)

Golubev V.N. Fundamentals of Biomorphology of Herbaceous Plants in the Central Forest-steppe. Voronezh, 1962 (in Russian)

Ilyushko M.V., Karatavcev Yu.F., Zhuravlev Yu.N. Variability of morphological characters of *I. setosa* (Iridaceae) in the Russian Far East. *Botanicheskii Zhurnal*. 2001;86(3):60-72 (in Russian)

Ishbirdin A.R., Ishmuratova M.M., Zhirnova T.V. Life strategies of the cenopopulation *Cephalanthera rubra* L. (Rich) on the territory of the Bashkir State Nature Reserve. *Vestnik of Lobachevsky University of Nizhni Novgorod*. 2005;1:85-98 (in Russian)

Kashin A.S., Parkhomenko A.S., Abramova L.M., Kondratyeva A.O., Bogoslov A.V., Shilova I.V., Kryukova A.V. Morphological variability of *Iris pumila* (Iridaceae) in the Lower Volga region and the South Urals. *Botanicheskii Zhurnal*. 2022;107(2):180-197. doi 10.31857/S0006813622020065 (in Russian)

Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780. doi 10.1093/molbev/mst010

Kozyrenko M.M., Artyukova E.V., Zhuravlev Y.N. Independent species status of *Iris vorobievii* N.S. Pavlova, *Iris mandshurica* Maxim., and *Iris humilis* Georgi (Iridaceae): evidence from the nuclear and chloroplast genomes. *Russ J Genet*. 2009;45(11):1394-1402. doi 10.1134/S1022795409110143

Kulikov P.V. Synopsis of the Flora of the Chelyabinsk Region (Vascular Plants). Ekaterinburg–Miass, 2005 (in Russian)

Makarevich I.F., Doronkin V.M., Shcherbik S.V., Blinov A.G. Application of RAPD-PCR method in the study of phylogenetic relationships between Siberian species of the genus *Iris* L. (Iridaceae). *Turczaninowia*. 2001;4(4):80-92 (in Russian)

Makarevich I., Golovnina K., Scherbik S., Blinov A. Phylogenetic relationships of the Siberian Iris species inferred from noncoding chloroplast DNA sequences. *Int J Plant Sci*. 2003;164(2):229-237. doi 10.1086/346160

Mathew B. The Iris. Portland, 1989

Mitic B. *Iris adriatica* (Iridaceae), a new species from Dalmatia (Croatia). *Phyton.* 2002;42(2):305-314

Pleines T., Jakob S.S., Blattner F.R. Application of non-coding DNA regions in intraspecific analyses. *Plant Syst Evol*. 2009;282:281-294. doi 10.1007/s00606-008-0036-9

Red Book of the Republic of Bashkortostan. Vols. 1 and 2. Ufa, 2021-2024 (in Russian)

Red Book of the Russian Federation. Plants and Mushrooms. Moscow, 2024 (in Russian)

Rodionenko G.I. Genus Iris. Moscow–Leningrad, 1961 (in Russian)

Rodionenko G.I. *Eremiris*, a new genus of the family Iridaceae. *Botanicheskii Zhurnal*. 2006;91(11):1707-1712 (in Russian)

Rodionenko G.I. Understanding the Secrets of Nature (My Destiny is Irises). St. Petersburg, 2013 (in Russian)

Rostova N.S. Correlations: Structure and Variability. St. Petersburg, 2002 (in Russian)

Trifinopoulos J., Nguyen L.-T., Haeseler A., Minh B.Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 2016;44(W1):W232-W235. doi 10.1093/nar/gkw256

Tsvelev N.N. Genus Iris L. In: Flora of the European Part of the USSR. Vol. 4. 1979;299-307 (in Russian)

Volkova G.A. The results of introduction of natural species of the *Iris* L. genus in northern Europe. In: Botanical Gardens: Issues of Introduction. Tomsk, 2010;274:127-129 (in Russian)

Weber T., Jakše J., Sladonja B., Hruševar D., Landeka N., Brana S., Bohanec B., Milović M., Vladović D., Mitić B., Poljuha D. Molecular study of selected taxonomically critical taxa of the genus *Iris* L. from the broader Alpine-Dinaric area. *Plants*. 2020;9(9):1229. doi 10.3390/plants9091229

Wilson C.A. Subgeneric classification in *Iris* re-examined using chloroplast sequence data. *Taxon.* 2011;60(1):27-35. doi 10.1002/tax.601004

Zaitsev G.N. Mathematical Statistics in Experimental Botany. Moscow, 1984 (in Russian)].

Zaitsev G.N. Mathematics in Experimental Botany. Moscow, 1990 (in Russian)

Zhao Y.T. A new species of Iris from China. *J Syst Evol.* 1992;30(2):181-182

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

Received April 30, 2025. Revised August 10, 2025. Accepted September 9, 2025.

Прием статей через электронную редакцию на сайте http://vavilov.elpub.ru/index.php/jour
Предварительно нужно зарегистрироваться как автору, затем в правом верхнем углу страницы выбрать «Отправить рукопись». После завершения загрузки материалов обязательно выбрать опцию «Отправить письмо», в этом случае редакция автоматически будет уведомлена о получении новой рукописи.

«Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» до 2011 г. выходил под названием «Информационный вестник ВОГиС»/ “The Herald of Vavilov Society for Geneticists and Breeding Scientists”.

Сетевое издание «Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» – реестровая запись СМИ Эл № ФС77-85772, зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 14 августа 2023 г.

Издание включено ВАК Минобрнауки России в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, Russian Science Citation Index, Российский индекс научного цитирования, ВИНИТИ, Web of Science CC, Scopus, PubMed Central, DOAJ, ROAD, Ulrich’s Periodicals Directory, Google Scholar.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте https://vavilovj-icg.ru/
и платформе Научной электронной библиотеки, elibrary.ru/title_about.asp?id=32440
англоязычная версия – на сайте vavilov.elpub.ru/index.php/jour
и платформе PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/journals/3805/

При перепечатке материалов ссылка обязательна.

✉ email: vavilov_journal@bionet.nsc.ru

Издатель: Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук»,
проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Адрес редакции: проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Секретарь по организационным вопросам С.В. Зубова. Тел.: (383)3634977.
Издание подготовлено информационно-издательским отделом ИЦиГ СО РАН. Тел.: (383)3634963*5218.
Начальник отдела: Т.Ф. Чалкова. Редакторы: В.Д. Ахметова, И.Ю. Ануфриева. Дизайн: А.В. Харкевич.
Компьютерная графика и верстка: Т.Б. Коняхина, О.Н. Савватеева.

Дата выхода в свет 27.02.2026. Формат 60 × 84 $^{1}/_{8}$. Уч.-изд. л. 20.9.