

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-26-17***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)

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-26-17

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. Bepalova, 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

Молекулярная и клеточная биология

- 163 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Концепция природной реконструкции генома. Часть 4. Интеграция фрагментов экстраклеточной двуцепочечной ДНК в геном гемопоэтических стволовых клеток и формирование экстра-хромосомальных интермедиатов.
С.Г. Ошихмина, В.С. Рузанова, Г.С. Риттер, Е.В. Долгова, С.С. Кирикович, Е.В. Левитес, Я.Р. Ефремов, Т.В. Карамышева, А.С. Молодцева, Я.В. Райцина, О.С. Таранов, С.В. Сидоров, С.Д. Никонов, О.Ю. Леплина, А.А. Останин, Е.Р. Черных, Н.А. Колчанов, А.С. Проскурина, С.С. Богачев
- 181 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Наивные ИПСК человека, полученные при культивировании линии клеток ICGi022-A с праймированной плюрипотентностью в среде HENSM, способны к эффективной дифференцировке в эндотелиальные производные. М.А. Арссан, А.И. Шевченко, С.М. Закиян, И.С. Захарова (на англ. языке)

Генетика и селекция растений

- 194 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Молекулярно-генетическая изменчивость по ISSR-маркерам у *Caragana jubata* (Pall.) Poir. (Fabaceae) в горах Средней Азии и Южной Сибири. Д.А. Кривенко, О.А. Чернышева, И.Н. Кубан, Е.В. Жмудь, И.В. Горбенко, А.А. Ачимова, О.Ю. Васильева, О.В. Дорогина
- 205 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Изменчивость оргanelльных геномов в коллекции ранних сортов сои.
В.В. Александрович, М.Г. Синявская, О.П. Шатарнов, О.Г. Давыденко
- 212 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Анализ экспрессии генов *PR-1*, *PR-2*, *PR-10*, хитиназ и хитиназа-подобных белков в корнях гороха под влиянием салициловой кислоты и метилжасмоната.
А.М. Егорова
- 222 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Отбор стабильных мутантов риса с помощью линейных моделей со смешанными эффектами (LMM) и индексов стабильности. П. Шарифи, А.А. Эбади, М.Т. Халладжан, Х. Аминпанах (на англ. языке)

Генетическая инженерия

- 233 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Роль типа экспланта и применения селективного агента в первичной эффективности трансформации *Lens culinaris* Medik. Т.В. Дюбенко, К.В. Смирнов, В.Е. Творогова (на англ. языке)
- 241 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Получение и характеристика двойной нокаутной линии *Arabidopsis thaliana*, лишенной экспрессии *AOX1a* и *VTC2*.
Е.В. Гармаш, Е.С. Белых, К.В. Ядрихинский, Р.В. Малышев, И.О. Вележанинов

Медицинская генетика

- 250 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Генетические корреляции между болезнями человека и плейотропные гены.
И.В. Зорколыцева, Н.М. Белоногова, А.В. Кириченко, Я.А. Цепилов, Т.И. Аксенович
- 259 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Новые регуляторные однонуклеотидные полиморфизмы (rSNPs) потенциально участвуют в регуляции глюконеогенеза в печени и вносят вклад в вариативность АМФК-зависимых механизмов индивидуального ответа на метформин.
Е.Е. Корболина, И.С. Дамаров, Т.И. Меркулова (на англ. языке)
- 267 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Особенности коннективности дефолтной сети мозга в зависимости от полиморфизма гена транспортера серотонина (5-HTTLPR).
А.В. Бочаров, А.Н. Савостьянов, С.С. Таможников, А.Е. Сапрыгин, Д.А. Лебедин, Е.А. Меркулова, Г.Г. Князев

Популяционная генетика

- 274 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Филогеография и таксономический статус видового комплекса *Formica picea* (Hymenoptera: Formicidae).
З.А. Жигульская, С.В. Шеховцов, С.В. Чеснокова, А.П. Бурнашева, А.А. Гурина, Р.Ю. Дудко, Т.В. Полубоярова, С.В. Решетников, Ю.Н. Сундуков, Д.И. Берман (на англ. языке)

Биоинформатика и системная биология

- 284 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Функциональная симметрия
и воспроизводимость эволюционного
процесса. С.И. Барцев

- 293 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Alembic: от разрозненных биологических
данных к структурированным ресурсам.
И.В. Бездворных, К.И. Юдыцкий, Н.А. Черкасов,
А.А. Самсонова, А.А. Канапин

Высокопроизводительное секвенирование

- 299 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Разработка и валидация программы
PipeSeq для анализа данных
секвенирования РНК на модели
Chlamydomonas reinhardtii.
А.М. Нерезенко, П.А. Виролайнен, С.А. Тупицына,
Е.М. Чекунова

- 311 **ОБЗОР**
Анализ вирома растений с помощью
высокопроизводительного
секвенирования: принципы и подходы.
Д.Д. Белкина, С.В. Виноградова

Цифровое фенотипирование

- 321 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Применимость инструмента для анализа
временных рядов StatFaRmer
при цифровом фенотипировании сои
(*Glycine max*). Д.С. Ульянов, А.А. Ульянова,
А.А. Кочешкова, А.О. Блинков, А.В. Архипов, Я.С. Меглицкая,
Н.Ю. Свистунова, Г.И. Карлов, М.Г. Дивашук

- 330 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Описание морфологических
характеристик колосьев пшеницы
в базе данных SpikeDroidDB в виде
цифрового паспорта. Е.Г. Кобышев, Ю.В. Кручинина,
В.С. Коваль, А.А. Потешкина, Н.В. Петраш, В.В. Пискарев,
Н.П. Гончаров, Д.А. Афонников

Molecular and cell biology

- 163 **ORIGINAL ARTICLE**
Concept of natural genome reconstruction. Part 4. Integration of extracellular double-stranded DNA fragments into the genome of hematopoietic stem cells and the formation of extrachromosomal intermediates. S.G. Oshikhmina, V.S. Ruzanova, G.S. Ritter, E.V. Dolgova, S.S. Kirikovich, E.V. Levites, Y.R. Efremov, T.V. Karamysheva, A.S. Molodtseva, Y.V. Raitsina, O.S. Taranov, S.V. Sidorov, S.D. Nikonov, O.Y. Leplina, A.A. Ostanin, E.R. Chernykh, N.A. Kolchanov, A.S. Proskurina, S.S. Bogachev
- 181 **ORIGINAL ARTICLE**
Naïve human iPSCs obtained by culturing the ICGi022-A cell line with primed pluripotency in HENSM medium efficiently differentiate into endothelial derivatives. M.A. Arssan, A.I. Shevchenko, S.M. Zakian, I.S. Zakharova

Plant genetics and breeding

- 194 **ORIGINAL ARTICLE**
Molecular genetic variability in coenopopulations of *Caragana jubata* (Pall.) Poir. (Fabaceae) in the mountains of Central Asia and Southern Siberia revealed using ISSR-markers. D.A. Krivenko, O.A. Chernysheva, I.N. Kuban, E.V. Zhmud, I.V. Gorbenko, A.A. Achimova, O.Yu. Vasilyeva, O.V. Dorogina
- 205 **ORIGINAL ARTICLE**
Variability of organelle genomes in a collection of early maturing soybean varieties. V.V. Aleksandrovich, M.G. Siniauskaya, A.P. Shatarnov, O.G. Davydenko
- 212 **ORIGINAL ARTICLE**
Analysis of *PR-1*, *PR-2*, *PR-10*, chitinases and chitinase-like proteins genes expression in pea roots under the action of salicylic acid and methyl jasmonate. A.M. Egorova
- 222 **ORIGINAL ARTICLE**
Selecting stable rice mutants with linear mixed models (LMM) and stability indexes. P. Sharifi, A.A. Ebadi, M.T. Hallajian, H. Aminpanah

Gene engineering

- 233 **ORIGINAL ARTICLE**
The role of explant type and selective agent application in the initial transformation rate of *Lens culinaris* Medik. T.V. Dyubenko, K.V. Smirnov, V.E. Tvorogova
- 241 **ORIGINAL ARTICLE**
Generation and characterization of a double-knockout *Arabidopsis thaliana* line lacking expression of *AOX1a* and *VTC2*. E.V. Garmash, E.S. Belykh, K.V. Yadrikhinskiy, R.V. Malyshev, I.O. Velegzhaninov

Medical genetics

- 250 **ORIGINAL ARTICLE**
Pleiotropic genes underlying genetic correlations across human diseases. I.V. Zorkoltseva, N.M. Belonogova, A.V. Kirichenko, Y.A. Tsepilov, T.I. Axenovich
- 259 **ORIGINAL ARTICLE**
Novel regulatory SNPs that can be activated due to metformin treatment may orchestrate liver gluconeogenesis and add to the variability in AMPK-dependent mechanisms of metformin response. E.E. Korbolina, I.S. Damarov, T.I. Merkulova
- 267 **ORIGINAL ARTICLE**
Features of connectivity of default mode network depending on polymorphism of serotonin transporter gene (5-HTTLPR). A.V. Bocharov, A.N. Savostyanov, S.S. Tamozhnikov, A.E. Saprygin, D.A. Lebedkin, E.A. Merkulova, G.G. Knyazev

Population genetics

- 274 **ORIGINAL ARTICLE**
Phylogeography and taxonomic status of the *Formica picea* complex (Hymenoptera: Formicidae). Z.A. Zhigulskaia, S.V. Shekhovtsov, S.V. Chesnokova, A.P. Burnasheva, A.A. Gurina, R.Yu. Dudko, T.V. Poluboyarova, S.V. Reshetnikov, Yu.N. Sundukov, D.I. Berman

Bioinformatics and systems biology

- 284 **ORIGINAL ARTICLE**
Functional symmetry and reproducibility of the evolutionary process.
S.I. Bartsev

- 293 **ORIGINAL ARTICLE**
Alembic: a framework for converting disparate biological data into structured resources. *I.V. Bezdovnykh, K.I. Yuditskiy, N.A. Cherkasov, A.A. Samsonova, A.A. Kanapin*

High-throughput sequencing

- 299 **ORIGINAL ARTICLE**
Development and validation of the PipeSeq program for RNA-seq data analysis in the *Chlamydomonas reinhardtii* as a model. *A.M. Nerezenko, P.A. Virolainen, S.A. Tupitsyna, E.M. Chekunova*

- 311 **REVIEW**
Plant virome analysis by high-throughput sequencing: concepts and approaches.
D.D. Belkina, S.V. Vinogradova

Digital phenotyping

- 321 **ORIGINAL ARTICLE**
Applicability of the StatFaRmer time series analysis tool in soybean (*Glycine max*) digital phenotyping. *D.S. Ulyanov, A.A. Ulyanova, A.A. Kocheshkova, A.O. Blinkov, A.V. Arkhipov, Ya.S. Meglitskaya, N.Yu. Svistunova, G.I. Karlov, M.G. Divashuk*

- 330 **ORIGINAL ARTICLE**
Description of morphological characteristics of wheat spike as a digital certificate in the SpikeDroidDB database.
E.G. Komyshev, Yu.V. Kruchinina, V.S. Koval, A.A. Poteshkina, N.V. Petrash, V.V. Piskarev, N.P. Goncharov, D.A. Afonnikov

doi 10.18699/vjgb-26-18

Concept of natural genome reconstruction.

Part 4. Integration of extracellular double-stranded DNA fragments into the genome of hematopoietic stem cells and the formation of extrachromosomal intermediates

S.G. Oshikhmina ^{1, 2*}, V.S. Ruzanova ^{1*}, G.S. Ritter ¹, E.V. Dolgova ¹, S.S. Kirikovich ¹, E.V. Levites ¹, Y.R. Efremov ¹, T.V. Karamysheva ¹, A.S. Molodtseva ³, Y.V. Raitsina ^{1, 2}, O.S. Taranov ⁴, S.V. Sidorov ^{2, 5}, S.D. Nikonov ⁶, O.Y. Leplina ⁷, A.A. Ostanin ⁷, E.R. Chernykh ⁷, N.A. Kolchanov ¹, A.S. Proskurina ^{1#}, S.S. Bogachev ^{1#} 

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

³ Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

⁴ State Scientific Center of Virology and Biotechnology "Vector" of Rosпотребнадзор, Koltsovo, Novosibirsk region, Russia

⁵ City Clinical Hospital No. 1, Novosibirsk, Russia

⁶ Novosibirsk Tuberculosis Research Institute, Novosibirsk, Russia

⁷ Research Institute of Fundamental and Clinical Immunology, Novosibirsk, Russia

 labmolbiol@mail.ru

Abstract. To assess the possibility of integrating extracellular double-stranded DNA fragments into the recipient genome of hematopoietic stem cells, a complex substrate was constructed consisting of the entire M13F-*AluI*-M13R fragment and its two restrictive derivatives, appearing after hydrolysis with restriction endonucleases *EcoRI* and *HindIII*: M13F-*AluI*-*EcoRI* and M13R-*AluI*-*HindIII*. The substrate contained a pBlueScript+ plasmid polylinker sequence, absent in the human genome, which framed the human *AluI* fragment cloned at the *EcoRV* site. Human bone marrow cells were treated with the DNA of the constructed complex substrate; taking into account the repair time of pangenomic single-strand breaks, preparations of metaphase plates were obtained. FISH revealed specific fluorescent signals. Simultaneously, DNA isolated from colonies obtained from bone marrow cells treated with a complex substrate was sequenced. Two rounds of sequencing were carried out: whole-genome and selective after targeted hybridization on metal beads. The results obtained indicate that homologous exchange between extrachromosomal and chromosomal DNA is possible. Integration into the genome via the single-strand annealing mechanism, involving microhomologies, is also possible. Intermediates were discovered that suggest the existence of an unusual integration into the genome at the nick of one end of the fragment and the other end of the fragment hanging freely into the interchromosomal space. A direct assessment of the possibility of integrating TAMRA-labeled fragments of fragmented human DNA and *E. coli* DNA into the genome of recipient cells was carried out using a human bone marrow cell model. The results obtained indicate that specific signals of homologous DNA are distributed throughout the chromosome body (human bone marrow cell model). Signals from nonhomologous *E. coli* DNA are predominantly concentrated in the centromeric regions of chromosomes. The ratio of the number of obtained reads with integration elements and FISH signals suggested the existence of a strong interaction between extracellular fragments and chromosomal DNA. Experiments have been conducted showing that linear plasmid DNA, after internalization into hematopoietic stem cells, forms a monomer ring. Internalized into the intracellular space, extracellular plasmid DNA is isolated together with chromosomal DNA after stringent purification and fractionation procedures. This fact suggests the existence of a strong ring associate of plasmid DNA and chromosome DNA formed without the participation of a protein framework in the form of a looped chromosomal strand.

Key words: FISH; whole genome sequencing; integration into the genome; extrachromosomal ring structures

For citation: Oshikhmina S.G., Ruzanova V.S., Ritter G.S., Dolgova E.V., Kirikovich S.S., Levites E.V., Efremov Y.R., Karamysheva T.V., Molodtseva A.S., Raitsina Y.V., Taranov O.S., Sidorov S.V., Nikonov S.D., Leplina O.Y., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Proskurina A.S., Bogachev S.S. Concept of natural genome reconstruction. Part 4. Integration of extracellular double-stranded DNA fragments into the genome of hematopoietic stem cells and the formation of extrachromosomal intermediates. *Vavilovskii Zhurnal Genetiki i Selektcii* = *Vavilov J Genet Breed*. 2026;30(2):163-180. doi 10.18699/vjgb-26-18

Funding. This work was supported by the Ministry of Science and Higher Education of the Russian Federation for the Institute of Cytology and Genetics (state budget-funded project No. FWNR-2026-0025) and by A.A. Purtov, I.N. Zaitseva and LLC "ES.LAB DIAGNOSTIC".

Концепция природной реконструкции генома.

Часть 4. Интеграция фрагментов экстраклеточной двуцепочечной ДНК в геном гемопоэтических стволовых клеток и формирование экстрахромосомальных интермедиатов

С.Г. Ошихмина ^{1, 2*}, В.С. Рузанова ^{1*}, Г.С. Риттер ¹, Е.В. Долгова ¹, С.С. Кирикович ¹,
Е.В. Левитес ¹, Я.Р. Ефремов ¹, Т.В. Карамышева¹, А.С. Молодцева³, Я.В. Райцина^{1, 2},
О.С. Таранов ⁴, С.В. Сидоров ^{2, 5}, С.Д. Никонов⁶, О.Ю. Леплина ⁷, А.А. Останин ⁷,
Е.Р. Черных ⁷, Н.А. Колчанов ¹, А.С. Проскурина ^{1#}, С.С. Богачев ^{1#} 

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный государственный исследовательский университет, Новосибирск, Россия

³ Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия

⁴ Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р. п. Кольцово, Новосибирская область, Россия

⁵ Городская клиническая больница № 1, Новосибирск, Россия

⁶ Новосибирский научно-исследовательский институт туберкулеза, Новосибирск, Россия

⁷ Научно-исследовательский институт фундаментальной и клинической иммунологии, Новосибирск, Россия

 labmolbiol@mail.ru

Аннотация. Для оценки возможности интеграции в реципиентный геном гемопоэтических стволовых клеток экстраклеточных фрагментов двуцепочечной ДНК был сконструирован сложно составленный субстрат, состоящий из целого M13F-*AluI*-M13R фрагмента и двух его производных, появляющихся после гидролиза рестриктазами *EcoRI* и *HindIII*: M13F-*AluI*-*EcoRI* и M13R-*AluI*-*HindIII*. В субстрате была представлена последовательность полилинкера плазмиды pBlueScript+, отсутствующая в геноме человека, которая обрамляла клонированный по сайту *EcoRV* *AluI* фрагмент человека. Клетки костного мозга человека были обработаны ДНК сконструированного сложно составленного субстрата, и с учетом времени репарации пангеномных одноцепочечных разрывов из клеток выросших колоний получены препараты метафазных пластинок. Проведенная FISH выявила специфические сигналы свечения. Одновременно ДНК, выделенная из колоний, полученных из клеток костного мозга, обработанных сложно составленным субстратом, была секвенирована. Проведено два раунда секвенирования: полногеномное и селективное после таргетной гибридизации на металлических бусах. Полученные результаты свидетельствуют, что гомологичный обмен между экстрахромосомальной и хромосомной ДНК возможен. Также возможна интеграция в геном по механизму однокитевого отжига, с участием микрогомологии. Обнаружены интермедиаты с одним концом фрагмента, интегрированным в геном по участку микрогомологии, и другим концом фрагмента, свободно свисающим в межхромосомное пространство. Проведена прямая оценка возможности интеграции TAMRA-меченых фрагментов двуцепочечной ДНК человека и *E. coli* в реципиентный геном на модели клеток костного мозга человека. Обнаружено, что специфические сигналы гомологичной ДНК распределены по телу хромосом (модель клеток костного мозга человека). Сигналы негомологичной ДНК *E. coli* преимущественно сконцентрированы в центромерных районах хромосом. Соотношение количества полученных ридов с элементами интеграции и сигналов FISH предполагало существование прочного взаимодействия экстраклеточных фрагментов и ДНК хромосом. В экспериментах показано, что линейная плазмидная ДНК после интернализации в гемопоэтические стволовые клетки формирует кольцо мономера. Интернализированная во внутриклеточное пространство экстраклеточная плазмидная ДНК выделяется совместно с ДНК хромосом после жестких процедур очистки и фракционирования. Этот факт предполагает существование прочного кольцевого ассоциата ДНК плазмиды и ДНК хромосом, сформированного без участия белкового каркаса в форме закольцованной хромосомной нити.

Ключевые слова: FISH; полногеномное секвенирование; интеграция в геном; экстрахромосомальные кольцевые структуры

Introduction

This article, the fourth in a series, further develops the concept that extracellular double-stranded DNA (dsDNA) fragments – linear DNA molecules of variable origin and size (200–2,000 bp or more) – are naturally delivered to hematopoietic stem cells (HSCs). The internalization process is facilitated by the general positive charge of HSCs, a characteristic determined by specific proteins within the HSC glycocalyx (Dolgorova et al., 2012; Petrova et al., 2022; Ritter et al., 2022). The mobilization and internalization

processes involve proteins containing heparin-binding domains or clusters of positively charged amino acids. The introduction of dsDNA into the cell has been shown to induce genome-wide single-stranded DNA breaks, which subsequently result in terminal hematopoietic stem cell differentiation (Ruzanova et al., 2024).

The process of terminal differentiation has been demonstrated to be associated with the reorganization of higher-order chromatin structure (Ruzanova et al., 2024). This process triggers the activation of a recombinogenic situation

in the cell, with its enzymatic machinery determining the appearance and repair of pangenomic single-chain breaks and leading the cell to the path of development of a specific hematopoietic sprout (Jacobson et al., 1975; Farzaneh et al., 1982; Johnstone, Williams, 1982).

The recombinogenic situation in the cell is defined as the state in which the mechanisms of chromatin DNA metabolic transformations are activated. These processes include the following: activation of supervising kinases, modulation of the cell cycle, induction of replicative stress, and activation of the repair and recombination mechanisms. The structure of the latter includes the enzymatic machinery of the mechanism, as well as DNA substrates that form metabolic intermediates of the DNA core. Cumulatively, in the event of pangenomic single-strand breaks, this process is instrumental in ensuring the restoration of continuity and the integrity of the primary DNA sequence of chromosomes (Likhacheva et al., 2008; Ruzanova et al., 2024).

The introduction of DNA fragments into the cell has been shown to initiate nick formation (Ruzanova et al., 2024), leading to a recombinogenic situation and triggering extensive cellular alterations. The molecular mechanisms that initiate this process, as well as the causal relationships among the initiating factors, remain to be elucidated.

The research on the intracellular behavior of extrachromosomal, non-viral DNA commenced in the 1980s (Smith, Berg, 1984). Nowadays, this subject is not a prevailing area of research. The existing body of research provides an incomplete account of intracellular events triggered by dsDNA fragment entry and the resulting activation of cellular surveillance mechanisms.

Cell cycle progression control systems were reported to be activated in response to the appearance of DNA in the form of synthetic oligonucleotides, SV40 DNA, and apoptotic cells in the nuclear space (Yakubov et al., 1989; Holmgren et al., 1999). Additionally, mechanisms underlying the internalization of plasmid DNA through artificial means were described (Smith, Berg, 1984; Lin J. et al., 1985; Thomas et al., 1986). Several studies demonstrated cellular checkpoint responses to be activated upon the internalization of synthetic (dA/dT)₇₀ oligonucleotides within the nucleus. The induction of cell cycle control systems in the absence of any genomic DNA damage or aberrant replication suggests that the structures of synthetic paired oligonucleotides may mimic intermediates arising from aberrant replication or genomic DNA damage, such as single-stranded DNA, blunt double-stranded ends, single-stranded site-duplex junctions, and cruciform shapes. The process of oligonucleotide pairing is assumed to lead to the creation of DNA molecule ends from paired fragments. These fragments function as inducers of ATM/ATR-dependent checkpoint cellular response through a mechanism that relies on regulators of hierarchical kinase activity such as RPA, RAD1, TopBP1, and claspin (Yoo et al., 2004, 2006; MacDougall et al., 2007; Zou, 2007).

Studies examining molecular processes involved in DNA uptake indicate that single-stranded DNA fragments do not trigger cellular regulatory mechanisms (Kumagai, Dunphy, 2000). The analysis of the minimum number of linear molecules capable of inducing a checkpoint response demonstrated that more than 30 molecules containing ssDNA/duplex structures activated a sufficient checkpoint response required to determine the level of Chk1 phosphorylation (MacDougall et al., 2007). However, with the presence of fewer than 30 linear DNA fragments located in the extrachromosomal space of the nucleus, the cell stress monitoring system remains quiescent. Consequently, this quantity of linear DNA molecules may enter the nucleus without activating the hierarchical kinase system, suggesting that the transient presence of extracellular fragments circulating in the plasma is the norm.

Extracellular linear dsDNA fragments within the cell have been observed to undergo a process of cross-linking, resulting in the formation of concatemeric and ring structures with a size range of up to 10 kbp (Perucho et al., 1980; Lin J. et al., 1985; Lin Y., Waldman, 2001; Rogachev et al., 2006; Potter et al., 2017, 2024). Furthermore, it has been observed that linear plasmid DNA fragments located in HSCs are processed to a depth of up to 1 kbp (Dolgova et al., 2013). Consequently, the collective evidence suggests that the presence of extracellular dsDNA fragments within the cell initiates cellular repair and recombination pathways. This assertion is supported by experimental evidence (Pierandrei et al., 2016). RT2 Profiler PCR Array platform (SABioscience, Qiagen) was utilized to analyze the expression of 84 genes during the transfection of extracellular dsDNA fragments into the cell. The fragments of dsDNA that were delivered to the cell were found to activate genes of various cellular systems, including those involved in DNA repair and homologous recombination, as well as genes regulating the cell cycle and epigenetic modifications. In summary, the principal molecular systems responsible for regulating cellular DNA metabolism are subjected to a stressed condition.

Nicks, as well as the double-stranded ends of extracellular fragments delivered to the nucleus, were found to serve as triggers for repair and recombinogenesis (Vriend, Krawczyk, 2017; Maizels, Davis, 2018). While sharing some overlapping signaling pathways, the mechanisms and factors triggered by these two structures differ (Ruzanova et al., 2024). Consequently, the presence of two DNA metabolites (nicks in the DNA of chromosomes and double-stranded ends of extracellular fragments delivered to the nucleus) in the cell triggers unidirectional repair and recombination processes. Apparently, the factors of both processes are likely to be involved in certain molecular relationships that have yet to be disclosed. A comprehensive analysis is required to assess the viability of such a scenario within the framework of the concept under consideration.

The introduction of fragments into cells using a variety of transfection techniques constitutes a standard procedure. The most common technique is the knockout method. Three independent researchers, Mario R. Capecchi, Oliver Smithies, and Sir Martin J. Evans, won the Nobel Prize in 2007 for this approach. This method relies on the principle that homologous dsDNA fragments within the nucleus will recombine, resulting in a knockout mutation. Consequently, the process of the natural exchange of genetic information between extrachromosomal DNA and chromatin DNA can be objectively confirmed to exist. This exchange represents an intrinsic property of the metabolic activity of the cell, associated with the interaction of chromatin DNA and extrachromosomal DNA (Perucho, Wigler, 1981; Kucherlapati et al., 1984; Smithies et al., 1985; Murnane et al., 1990; Hastings et al., 1993; Leung et al., 1997; Li et al., 2001; Lin Y., Waldman, 2001).

Following their entry into the nucleus, the dsDNA fragments then engage in the processes they trigger. The introduction of recombinogenic nicks and double-stranded ends into the nucleus will inevitably initiate recombination processes, which we hypothesize to result in the emergence of new genetic information within the cell stored in the internalized dsDNA fragments of initially extracellular localization.

The present part of the study evaluates the possibility for extracellular DNA fragments to be integrated into the genome of HSCs and for ring-like associates to be formed between fragment DNA and chromosome DNA.

Materials and methods

Human bone marrow cells. In this study, we used cells from cryopreserved bone marrow separates of Hodgkin's lymphoma patients provided by the cryobank of the Research Institute of Fundamental and Clinical Immunology (RIFCI). The Clinic of Immunopathology of RIFCI (the Department of Hematology equipped with a bone marrow transplantation unit) provides the treatment of patients with hemoblastosis using high-dose chemotherapy and transplantation of autologous or allogeneic peripheral HSCs. When harvesting peripheral stem cells, along with the main apheresis product (which is transplanted to the patient), two or three samples (satellite tubes) of separated cells are prepared for quality control of the apheresis product and scientific research. Such samples were used in the present work along with the main sample. Comprehensive documentation, including patient-signed informed consent and approved protocols for bone marrow research and treatment, is provided for each bone marrow separate sample and satellite samples, adhering to established regulatory guidelines. Subsequent to treatment and use of the primary product, satellite samples undergo disposal conforming to Sanitary Rules and Regulations or are allocated for scientific applications. Documentation pertaining to each bone marrow sample is archived in the RIFCI cryobank and is readily available upon request.

***AluI* double-stranded DNA probe.** The amplification of the human *AluI* repeat (M13F-*AluI*-M13R fragment) was performed by means of PCR. The matrix was the repeat DNA cloned in plasmid pUC19 incorporating the start and the end of the tandemly repeated *AluI* and *AluY* sequences (NCBI: AC002400.1, 53494–53767) (Dolgova et al., 2012). The amplification was performed using either standard M13 primers (M13 for: 5' GTAAAACGACGGCCAGT 3', M13 rev: 5' CAGGAAACAGCTATGAC 3') or *AluI*-specific primers (Dolgova et al., 2012). The PCR fragment was resuspended in 0.1 V NaAc 3 M pH 5.2 and 1 V isopropanol for 10 min at –20 °C, followed by centrifugation. The precipitate was washed in 70 % ethanol and dissolved in sterile water.

DNA preparations. The hDNA^{gr} preparation (a human genome DNA reconstructor) is a human double-stranded DNA in the form of 200–2,000 bp fragments. The hDNA^{gr} preparation was isolated from the placentas of healthy women, fragmented by ultrasonic disintegration to the size of 200–2,000 bp, deproteinized by proteinase K treatment, and isolated by phenol/chloroform extraction.

The human DNA was isolated from the placentas of healthy women using a similar method without fragmentation.

The *E. coli* DNA was isolated in a standard manner using lysozyme, ultrasonic cell disruption, proteinase K deproteinization, and phenol/chloroform extraction.

Design of a probe for bone marrow cell processing. For analyzing potential nonhomologous integration and/or homologous exchange, a probe was constructed consisting of an *AluI* fragment embedded in the pBlueScript+ (pBS) polylinker by the *EcoRV* site and restricted to sequences extending to and including the M13 primer landing sites. The internalization experiments involved using a mixture of the whole M13F-*AluI*-M13R fragment and two of its derivatives, generated through *EcoRI* and *HindIII* restriction enzyme hydrolysis: M13F-*AluI*-*EcoRI* and M13R-*AluI*-*HindIII*.

For analyzing possible direct integration, fragmented human and *E. coli* genomic DNA was tagged with a nucleotide containing the TAMRA fluorochrome using a random sequence primer (stat primer) in an amplification reaction with Taq polymerase (Medigen LLC).

The internalization analysis of the TAMRA-tagged M13F-*AluI*-M13R PCR fragment into Ehrlich carcinoma cells was performed according to the procedure described in (Ruzanova et al., 2022).

FISH was performed according to a standard procedure using TAMRA-labeled probes using the method described in Supplementary Material 1¹. The examination of FISH results was performed with an AXIOSKOP 2 Plus microscope (Zeiss, Germany). Microimage recording and processing were performed using a CCD camera (CoolCube 1, METASystems, Germany), Set 49 filter set

¹ Supplementary Materials 1 and 2 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Oshikh_Engl_30_2.pdf

(G 365, FT 395, BP 445/50) (Zeiss, Germany), ET-orange #1 (ET 546/22 × ET 590/33) (CHROMA, USA), and ISIS5 software (METASystems GmbH, Germany). The number of metaphases analyzed ranged from 14 to 30 metaphase plate images.

Bone marrow cell cultivation within a methylcellulose medium. Brain tissue samples were procured via percutaneous aspiration from the RIFCI cryobank. Standard thawing methods preserve up to 90 % viability of bone marrow cells, confirmed by FACS analysis to include CD34⁺ cells, eliminating the need for additional verification.

Human bone marrow cells contained in the aspirate were washed with RPMI medium and introduced into the experiment. Control cells and probe-treated cell samples were precipitated for 10 min at 400g and resuspended in IMDM + 2 % FBS medium. Myeloid progenitor quantification and analysis were performed using MethoCult H4034 methylcellulose medium (Stem Cell Technologies) for both treated and control cell samples. Cell culturing took 10 to 15 days, on average, depending on the purpose of the experiment. Cells were analyzed and isolated from methylcellulose medium after cultivation according to the manufacturer's instructions.

Targeted sequencing. Sequencing was performed using DNA isolated from cells treated with the compound probe: M13F-*AluI*-M13R, M13F-*AluI*-EcoRI, and M13R-*AluI*-HindIII. Library enrichment was performed by hybridization with biotinylated polylinker DNA of pBS plasmid immobilized on Dynabeads® Streptavidin magnetic particles (Life Technologies, USA) according to the method proposed by T. Maricic et al. (2010). The Illumina MiSeq platform (TruSeq® DNA Sample Preparation Kit (Illumina)) was used for sequencing using a MiSeq Reagent Kit v2 (300 cycles).

Whole-genome sequencing. Sequencing was performed using DNA isolated from cells treated with the compound probe: M13F-*AluI*-M13R, M13F-*AluI*-EcoRI, and M13R-*AluI*-HindIII. A NovaSeq 6000 instrument (Illumina) with a read length of 2 × 150 bp was used. The Q30 value accounted for at least 90 %. TruSeq Nano DNA Library Prep kit (Illumina) was used for library preparation. Ultrasonic DNA fragmentation was performed on an ME220 Focused-ultrasonicator (Covaris). Capillary electrophoresis, performed using a TapeStation 4200 instrument (Agilent), was utilized for the assessment of DNA library quality. The Ugene and Blastn platforms were used to analyze the sequences obtained. Raw sequence data, stored in FASTQ format, are accessible upon request.

Ring formation by extracellular dsDNA fragments internalized by HSCs and their closest progeny. A detailed schematic of the experimental design is described in the "Experimental proof of ring formation by long dsDNA fragments internalized in HSCs and their looping of the chromosome DNA strand" subsection. All the experimental procedures were performed according to

the protocols described in practical manuals (Maniatis et al., 1984; Glover, 1988).

The bone marrow cells were treated with DNA preparation, washed of DNA, and incubated for 2 hours in a CO₂ incubator in a complete RPMI nutrient medium. Then, the cells were harvested and washed with physiological saline. The cytoplasmic membrane was solubilized with 0.2 % Triton X-100, layered on a 10 % sucrose gradient, and centrifuged at 500 g for 20 min. Saline-washed nuclei underwent lysis with 1 % SDS and 50 mM EDTA. Deproteinization was accomplished by treating with 200 µg/ml proteinase K at 58 °C for two hours, followed by phenol-chloroform extraction. The resulting lysate was applied to a 10–30 % NaCl stepwise (5 %) gradient and centrifuged at ~300,000g (L-8 centrifuge, SW 50.1 rotor, 47,000 rpm) for 2 hours at room temperature. The scheme displays DNA molecular weight markers located within the designated percentage gradient zones near the last tube on the right (Glover, 1988; Dolgova et al., 2013). The 7th region of the gradient contains DNA with a size greater than 40 kbp, forming a precipitate at the bottom of the test tube. Precipitation of the DNA supernatant was achieved through the addition of an equal volume of isopropanol, prepared from a 0.3 M sodium acetate solution (pH 4.8), and incubation at –20 °C overnight. Following centrifugation and saline washing, fractions 1–6 were dissolved in 36 µl of H₂O. Following a 70 % ethanol wash, the DNA precipitate from fraction 7 was subsequently dissolved in 36 µl of H₂O. 12 µl of material (1/3 of the total volume) was spread on an agarose gel.

Sample 7 (sediment at the bottom of the gradient) was analyzed, with a Qubit DNA concentration of 1 ng/µL. The total yield was 36 ng. Gradient analysis was performed using one-third (12 ng) of the sample. Quantification of pEGFP-N1 plasmid DNA in the kanamycin resistance gene region was accomplished via Real-Time PCR using another one-third of the sample (12 ng). The remaining third of the sample (12 ng) underwent a transformation into electrocompetent DH-5 cells, yielding a titer of 10⁷–10⁸.

The transformation of 12 ng of fraction 7 produced a count exceeding 50 large colonies and approximately 5,000 small colonies. All large colonies used for DNA isolation (8 colonies) contained the pEGFP-N1 monomer. Small colonies were not found to contain plasmid DNA. The DNA extracted from large-scale bacterial colonies was subjected to enzymatic hydrolysis using the restriction enzymes HindIII, EcoRI, and HaeIII.

Results

Detection of synthetic DNA substrate by FISH on metaphase plates obtained from colony cells

Colony formation, as described above, is a consequence of a single bone marrow hematopoietic stem cell undergoing division. Therefore, cells within a given colony exhibit identical genetic profiles, mirroring the genotype of the

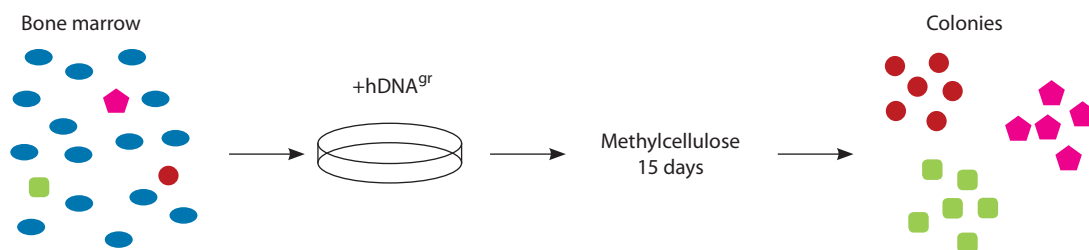


Fig. 1. Experimental design overview.

progenitor CD34⁺ HSCs (Fig. 1). Should the original HSC acquire genomic changes due to inducer treatment, these alterations will be amplified in the resulting colony to a detectable level, unlike the inherent difficulty of such detection within individual CD34⁺ bone marrow HSCs (property of LLC “ES LAB DIAGNOSTIC”).

The experimental protocol necessitated the creation of an artificial substrate for hematopoietic stem cell processing and a fluorescent probe to identify the integration of artificial substrate components. The options selected are detailed below. The amplified polylinker of the pBS plasmid was taken as foreign DNA, which is absent in the human genome. This DNA fragment, labeled with the TAMRA fluorophore, served as a hybridization probe (pBSM13-TAMRA⁺). Simultaneously, we used a complex substrate including an *AluI* repeat cloned at the *EcoRV* site, framed by the pBS polylinker, and *εIII* fragments containing a portion of the pBS polylinker with an adjacent *AluI* repeat and bounded by *EcoRI* or *HindIII* restriction sites (M13F-*AluI*-M13R, M13F-*AluI*-*EcoRI*, and M13R-*AluI*-*HindIII*). The internalization of TAMRA was assessed via spectrophotometry, electrophoresis, and visualization in an Ehrlich carcinoma ascites cell model (a standard procedure for assessing the fact of internalization). This approach to assessing DNA labeling in materials is a routine laboratory technique. Supplementary Material 1 provides a summary of the preliminary findings.

The resultant construct was used to treat human bone marrow cells. On day 15 of methylcellulose cultivation, the colonies were harvested, washed free of the medium, and treated with colchicine. Then, metaphase preparations were prepared. Next, FISH with TAMRA-labeled probe pBSM13 (pBSM13-TAMRA⁺) was performed. Figure 2A presents the hybridization results. The preparations were found to exhibit multiple signals. The results obtained indicated a possible integration of the artificial probe into the recipient genome. Also, it was possible for extracellular DNA to be firmly associated with the chromosome body. When analyzing images of metaphase plates (up to 40), we did not focus on mapping the TAMRA signals. Therefore, we were not to achieve a complete representation of the karyotype in the images. The primary objective of the experiments was to ascertain the specificity of hybridization signals, which was confirmed through analysis of the resulting images.

Part of the cell samples used to obtain the metaphase plates under analysis were also used for targeting and full genomic sequencing (total DNA isolated from cells).

Furthermore, experiments employing fluorochrome-labeled DNA probes were conducted to directly ascertain the presence of labeled material in (in association with) chromosomes. Human bone marrow cells, treated with TAMRA-labeled human and *E. coli* DNA, were cultured in methylcellulose. At 15 days, the cells were washed and treated with colchicine, followed by the metaphase plate analysis (Fig. 2B). The presence of specific signals was observed on a number of metaphase plates. In the case of human DNA, the labeled material was found to be detected at multiple sites on the chromosome body. In the case of bacterial DNA, the specific label proved to be concentrated to a greater extent in centromeric regions.

All the experiments with metaphase plates used metaphase preparations obtained from colony cells grown from a sample of bone marrow cells untreated with DNA substrate as control samples.

Proof of recombination interaction of the artificial probe with the HSC genome using modern sequencing technologies

In order to prove the recombination interaction of the M13F-*AluI*-M13R, M13F-*AluI*-*EcoRI*, and M13R-*AluI*-*HindIII* DNA substrates with the genome of human HSCs, we employed the same cell material as used for metaphase preparation and FISH. We constructed a library and performed two types of sequencing: (1) whole-genome sequencing and (2) sequencing involving increasing the number of target fragments by enriching the library by hybridization with biotinylated polylinker DNA of the pBS plasmid included in the DNA substrate used, immobilized on Dynabeads® Streptavidin magnetic particles (Life Technologies, USA) according to the method of (Maricic et al., 2010). The results obtained were summarized in one section.

Figure 3A schematically depicts the substrates used for cell treatment and those present in the mixture. To sum up, the following derivatives of the original substrate were present in the mixture: M13F-*AluI*-M13R, M13F-*AluI*-*EcoRI*, *EcoRI*-M13R, M13R-*AluI*-*HindIII*, and *HindIII*-M13F. The sequences of nucleotides in the substrates are listed

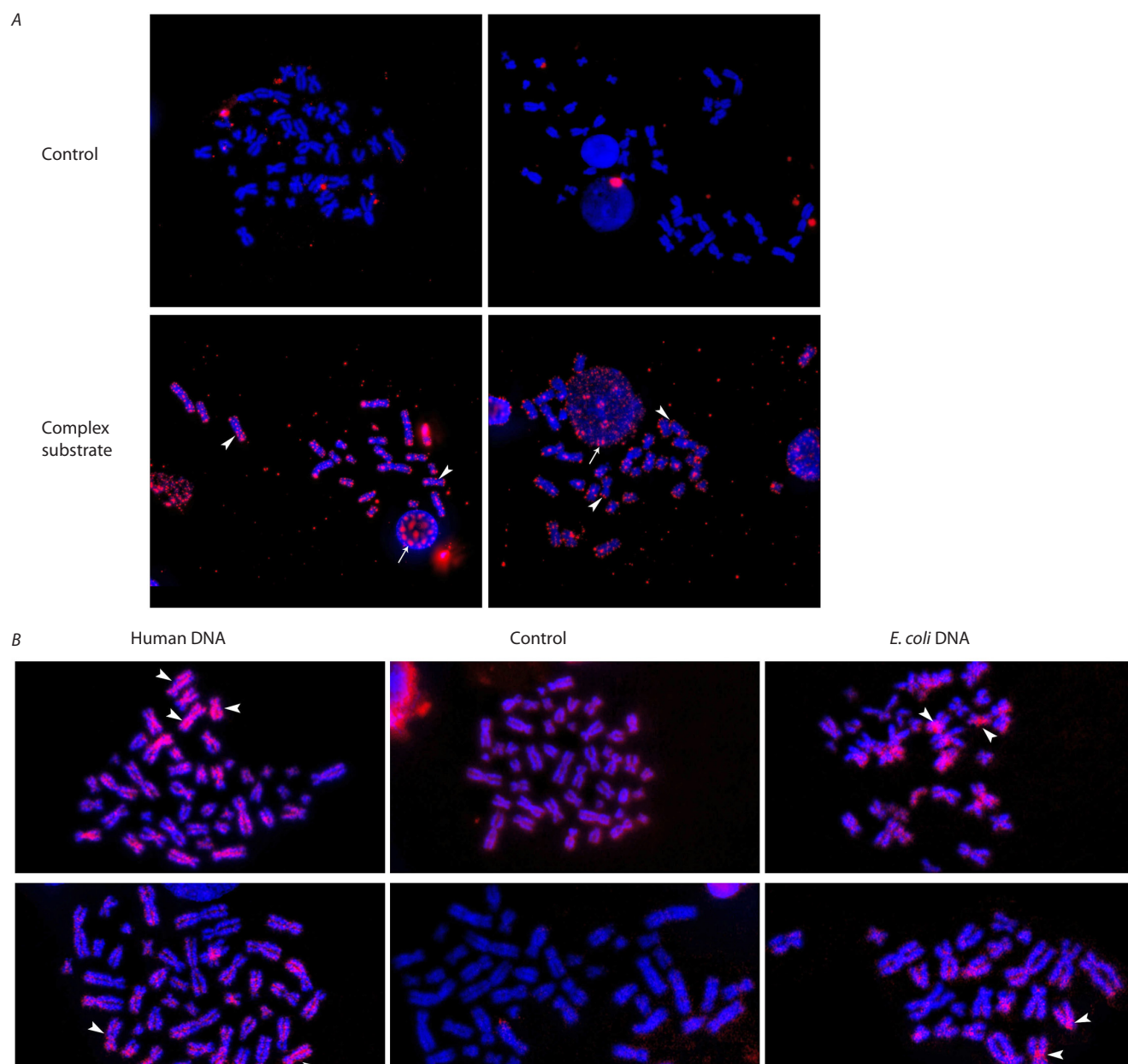


Fig. 2. Analysis of the presence of specific FISH signals after treatment of bone marrow cells with the selected probe and specific signals detected after direct interaction of fluorochrome-labeled dsDNA probes and bone marrow cells on metaphase chromosomes. *A* – FISH with the pBSM13-TAMRA+ probe of metaphase chromosomes obtained after colchicine treatment of colonic cells grown from HSCs without treatment (Control) and treated with a complex substrate (M13F-*Alu*-M13R, M13F-*Alu*-EcoRI, and M13R-*Alu*-HindIII) as part of bone marrow cells. Arrows indicate multiple hybridization sites. *B* – distribution of TAMRA-tagged material on metaphase chromosomes derived from human colony cells after treatment of these cells with TAMRA-tagged human and *E. coli* DNA. The arrows indicate multiple localization sites of the tag in the sample treated with human DNA and centromeric regions in the sample treated with *E. coli* DNA.

in Supplementary Material 2. Both forward and reverse orientations of the reads were analyzed for each chain.

Nucleotide sequence analysis of the selected clones

We identified the reads with distinctly cut-off ends terminating at the HindIII site and the primer sequences for M13 for. This indicates that the fragments from the original substrate are present in the cell in a “free state”. It should be noted that these are short fragments containing only part of the polylinker and the primer sequence for M13 for.

The observed fragments may have survived 15 days of cell cultivation, localized either to a single quiescent HSC or dispersed amongst its progeny. Their size (<100 bp) may have shielded them from enzymatic processing, e. g., ring formation, concatemerization, or nuclease degradation (Supplementary Material 2).

A total of 56 sequences were found to contain a polylinker region from the pBS plasmid extending past the EcoRV site. Most of them were found to contain sequences flanked by the same inverted repeats (TIR) adjacent to both halves of

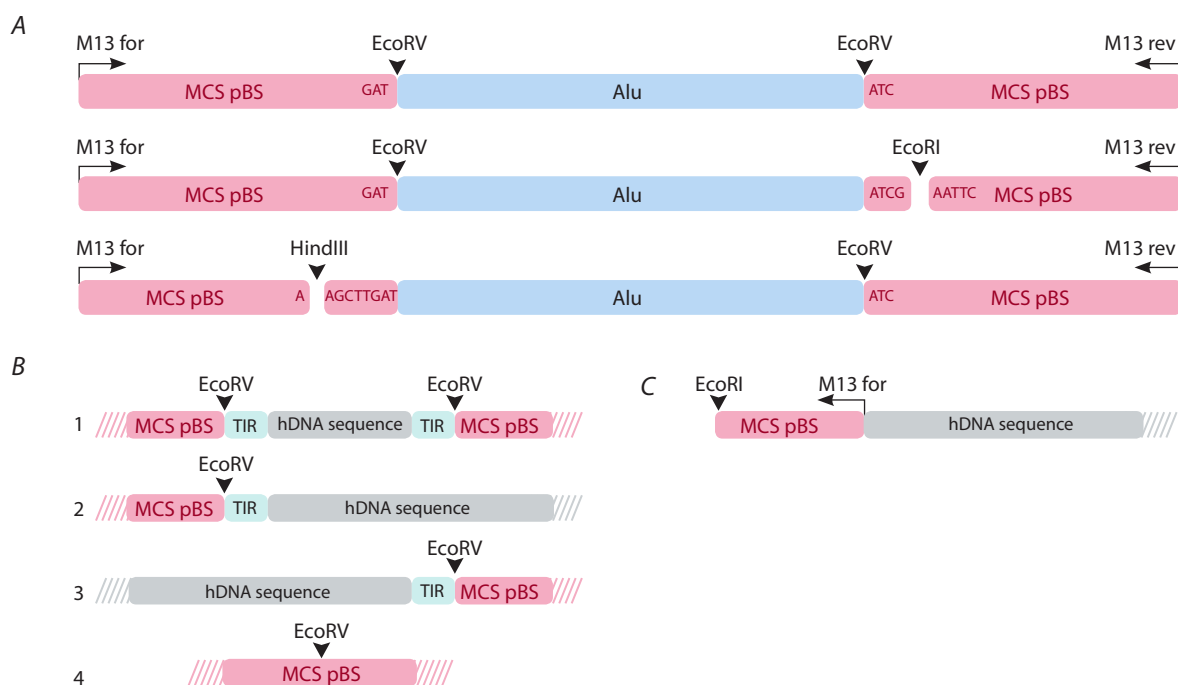


Fig. 3. Schematic representation of the complex substrate used and several of its intermediates resulting from induced recombination processes.

A – schematic representation of the substrate: the M13F-AluI-M13R, M13F-AluI-EcoRI, and M13R-AluI-HindIII fragments used for cell treatment. The original fragment and the same fragment hydrolyzed by EcoRI and HindIII restriction enzymes; B, C – schematic representation of sequencing-derived sequences.

the *AluI* cloning site of the EcoRV fragment (Fig. 3B1). Four reads had the *AluI* repeat replaced by an 84-nucleotide DNA sequence bounded on both sides by 10-bp terminal inverted repeats (TIRs) adjacent to the halves of the EcoRV cloning site (Fig. 3B2). The 21st read had the sequence terminating towards the left half of the conserved *AluI* EcoRV cloning site, failing to reach the opposite half of the EcoRV site, and bounded by the same inverted repeat (TIR) adjacent to the conserved left EcoRV site (Fig. 3B3). This 21st read is represented by three different sequences, one being 37 bp and the other two being approximately 100 bp in length and containing small regions (5–7 bp) of homology with the *AluI* repeat from the substrate.

Eleven reads include an intact full-length EcoRV site with no insertion, indicating the recombination of the *AluI* sequence cloned at the indicated site and the recovery of the original EcoRV (Fig. 3B4).

Fifteen reads match the original substrate containing part of the *AluI* repeat.

Two reads were found with long (54 and 57 bp) regions of genomic DNA joined to the end of the M13 for sequence. These comprise a complete M13 for primer and a polylinker site terminating at a residual HindIII restriction site (Fig. 3C).

Furthermore, we identified reads exhibiting internal recombination at the *AluI*-EcoRV cloning site, those with chaotically shuffled polylinker sites, and one featuring DNA segments of non-substrate origin.

Experimental proof of ring formation by long dsDNA fragments internalized in HSCs and their looping of the chromosome DNA strand

The apparent discrepancies in the number of fluorescent signals in FISH hybridization experiments, the findings of a direct analysis of the integration of extrachromosomal fragments into the genome, and the results of whole-genome sequencing suggested a potential mechanism for the formation of stable associates between DNA fragments localized in the nucleus and DNA strands of the chromosome. Given the lack of detectable large-scale single nonhomologous integration, a different mechanism is proposed for the formation of stable complexes. One type of associates that does not require protein anchoring is the DNA loop formed on the chromosome strand. This loop could have been formed by fragments of extracellular dsDNA delivered into the nuclear space. Experiments, largely replicating the methodology of a prior study (Dolgova et al., 2013), were performed to assess the assumption concerned.

The experiments involved analyzing events and changes at the molecule level occurring with the substrate, plasmid pEGFP-N1, used for prolonged processing of HSCs (Fig. 4A). The analysis of the native plasmid sample revealed the presence of the dimeric form and the absence of the monomeric form. The transformation of 100 pg of linearized pEGFP-N1 (the amount equivalent to that obtained from the experimental sample DNA isolated from the bottom of the salt gradient, see below) into electrocompe-

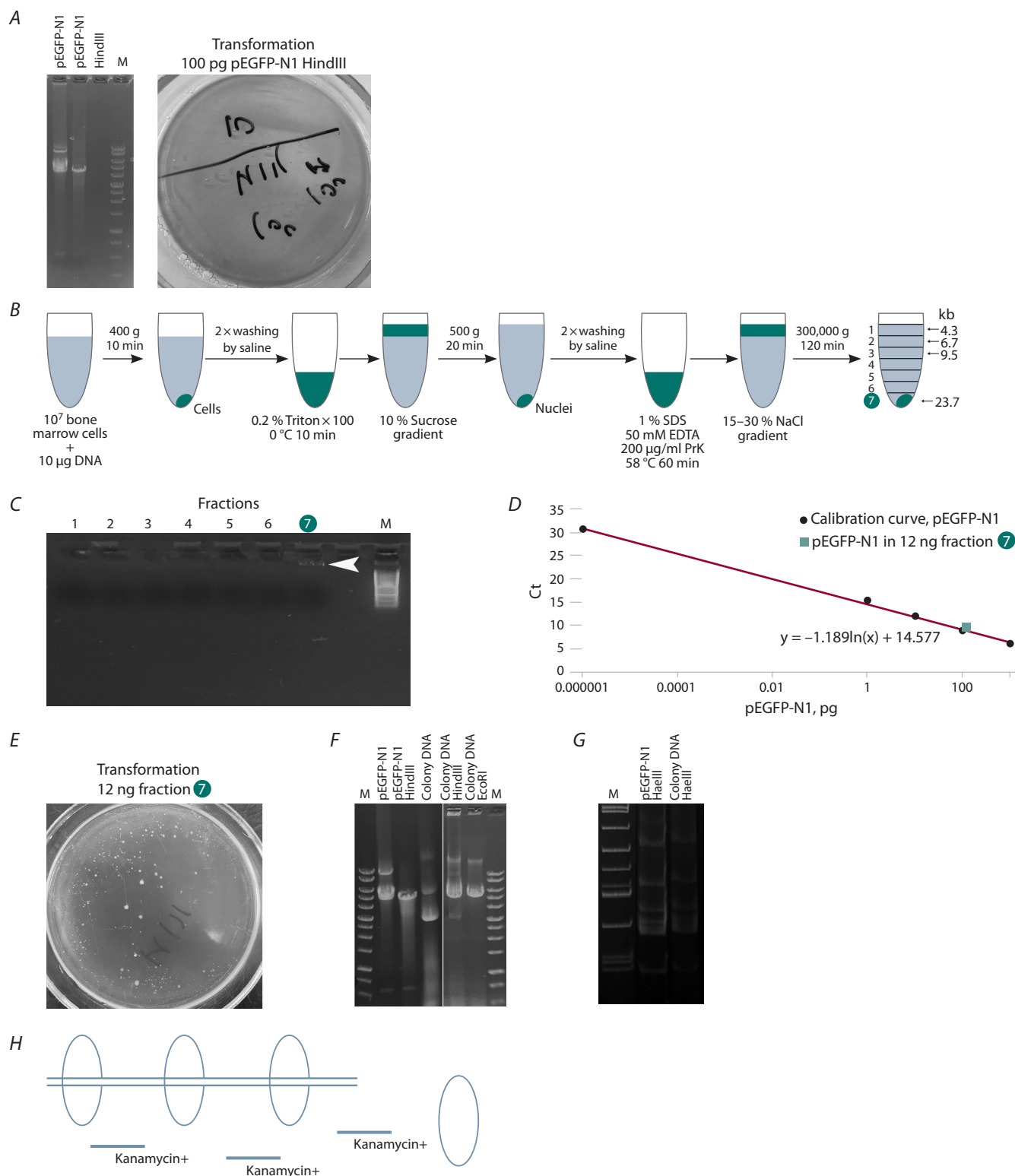


Fig. 4. Experimental evidence of ring formation by long dsDNA fragments internalized in HSCs and their looping of the chromosome DNA strand.

A – substrate of plasmid pEGFP-N1 used for processing of HSCs. Electrophoresis in agarose gel of native pEGFP-N1 plasmid hydrolyzed by HindIII. Transformation of pEGFP-N1 plasmid hydrolyzed by HindIII into *E. coli*, strain DH-5; **B** – the overall design of the experiment; **C** – electrophoresis of samples in 0.7 % agarose; **D** – analysis of the amount of pEGFP-N1 plasmid DNA in the kanamycin resistance gene region; **E** – transfection of 12 ng of fraction 7 resulted in more than 50 large colonies and about 5,000 small colonies; **F**, **G** – agarose and acrylamide gels with different variants of hydrolyzed and non-hydrolyzed plasmid DNA: the original pEGFP-N1 and one of the transformants (large colony). **F** – DNA isolated from large colonies was hydrolyzed by HindIII and EcoRI. Restriction analysis indicates intact HindIII and EcoRI polylinker sites. **G** – restriction analysis indicates an intact HaeIII monomer pattern and, thus, no intraplasmid recombination occurs. **H** – a putative schematic representation of the looped form of the plasmid formed around the chromosome after it enters the cell.

tent DH-5 cells yielded no transformants, thus confirming the absence of the ring form of the plasmid.

Figure 4B presents a schematic overview of the experimental design. A detailed step-by-step description is given in the “Materials and methods” section, specifically in the “Ring formation by extracellular dsDNA fragments internalized by HSCs and their closest progeny” subsection.

An examination of fractions from high-speed salt-gradient centrifugation by electrophoresis revealed the presence of DNA, as detected by ethidium bromide fluorescence, only in the final fraction (fraction 7, sediment) (Fig. 4C).

A quantitative analysis of the pEGFP-N1 plasmid DNA within the kanamycin resistance gene locus revealed that fraction 7 (12 ng) contained approximately 100 pg of homologous DNA (Fig. 4D).

The transformation of 12 ng of fraction 7 yielded over 50 macroscopic colonies and approximately 5,000 microscopic colonies (Fig. 4E). The macroscopic colonies were cultivated and inoculated in a liquid medium supplemented with 50 µg/mL kanamycin. The microscopic colonies exhibited growth in the liquid medium but did not develop further. No plasmid DNA in the form of monomer or dimer was found in the microscopic colonies. All the macroscopic colonies from which DNA was isolated (eight colonies) contained the pEGFP-N1 monomer. DNA isolated from the macroscopic colonies was treated with HindIII, EcoRI (Fig. 4F), and HaeIII (Fig. 4G) restriction enzymes.

Taken together, the results obtained in this series of experiments indicate the following:

- Bone marrow cells were treated with plasmid DNA linearized by the HindIII site. The presence of the original plasmid as a dimer and its monomer form in the resulting transformants suggests the recovery of the ring form linearized by HindIII plasmid within the bone marrow cells (BMCs).
- Restriction analysis with HaeIII, HindIII and EcoRI confirms the absence of intraplasmid recombination (Fig. 4 F, G).
- Stringent fractionation of plasmid and chromosomal DNA via NaCl gradient ultracentrifugation, following exhaustive proteinase K digestion and phenol-chloroform extraction, reveals a stable plasmid-chromosome DNA complex.

The most likely scenario involves genomic DNA looping facilitated by the plasmid ring, which is formed during the linear pEGFP-N1 cyclization upon cellular entry (Fig. 4H).

The disparity between the pEGFP-N1 kanamycin resistance gene copy number, the number of transformants obtained, and the DNA yield, as determined through comparison, suggests the following explanation. Following transformation, multiple degraded plasmids, all possessing a kanamycin resistance gene, may have produced a single kanamycin resistance protein, manifested as small colonies.

It is these fragments that were evaluated by quantitative real-time PCR. Analysis of large transformants indicates fewer than one plasmid DNA molecule per cell with DNA capture capability. Viable transformants may have resulted from plasmid rings released during chromosomal DNA isolation. The remaining portion, associated with chromosomal DNA, failed to yield transformants, thereby compromising the accuracy of quantification.

Our findings corroborate those presented by E.V. Dolgova et al. (2013), confirming the ring closure of internalized dsDNA fragments within HSCs in both experimental contexts. At the moment of closure, some rings seem to stochastically encompass the DNA strand of the chromosome, forming a stable complex.

Discussion

The experimental system demonstrated recombination activity between chromatin DNA and double-stranded extrachromosomal DNA fragments, which functioned as interacting partners in this process. A number of publications characterize experimentally verified and theoretically predicted intermediates resulting from the recombination of chromosomal DNA, single-stranded chromosomal DNA breaks, and extrachromosomal double-stranded DNA fragments (Hastings et al., 1993; Leung et al., 1997; Cromie et al., 2001; Li et al., 2001). Figure 5 offers schematic illustrations of the aforementioned intermediates. The relationship may be categorized into two variants. The first scenario presents an active recombinogenic condition that generates an enzymatic mechanism for recombination. This can occur at any genomic site through the interaction of double-stranded fragment ends participating in a reciprocal exchange with homologous chromosome

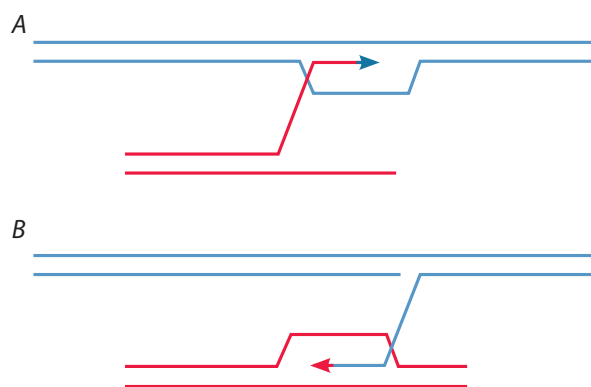


Fig. 5. Schematic representation of recombination processes (single-strand annealing, gene conversion, microhomology pairing) occurring in the cell involving genomic DNA and dsDNA fragments delivered to the cell.

A – recombination due to the open double-stranded ends of the delivered fragments; B – recombination due to single-strand breaks in genomic DNA.

sequences (Fig. 5A). In the second scenario, it is the DNA strand termini exhibiting the nick that are recombinogenic (Fig. 5B).

This work explored the potential for incorporating extracellular DNA into the genome. Two methodologies were employed: FISH and whole-genome sequencing. In pursuit of this objective, a composite probe was developed, consisting of an *AluI* fragment as its core, framed by the left and right pBS polylinker sequences and a number of their derivatives. Hydrolysis with *EcoRI* and *HindIII* restriction enzymes yielded derivative fragments representing portions of the complete substrate. Figure 3A depicts a mixture containing five distinct elements. A hypothesis was formulated suggesting that the successful association of probe and chromosome DNA would result in the observation of specific signals on metaphase chromosomes through the application of FISH.

The structured nuclear hybridization signals revealed through image analysis might demonstrate specific characteristics. The FISH signal indicated potential homologous and nonhomologous substrate integration into the genome. An alternative hypothesis regarding the origin of specific signaling was proposed. This phenomenon was associated with the integration of extrachromosomal elements

into actively transcribed chromatin regions, resulting in the formation of stable chromatin complexes between extrachromosomal and chromosomal DNA, persisting until metaphase (Møller et al., 2018; Wu et al., 2019; Zhu et al., 2021; Pecorino et al., 2022). This issue is considered in the last part of the “Discussion” section.

Whole-genome and selective sequencing methodologies were implemented to test the stated hypotheses. These methods enabled the identification of two intermediate variants, suggesting recombination between extracellular dsDNA and nuclear chromosome DNA fragments.

1. Over 20 genomic DNA sequences with discernible *AluI* fragment substitutions were detected using *EcoRV* cloning sites for identification. The genomic DNA segment extends significantly from the *AluI* cloning site of the *EcoRV* fragment on both its left and right flanks, or it is positioned between the two halves of the *EcoRV* cloning site. Without exception, this genomic segment is situated adjacent to the halved *EcoRV* cloning site, with the boundary defined by specific sequences constituting a 10 bp terminal inverted repeat (TIR). In some reads, this fragment exhibits homology with cloned *AluI*, while others contain small regions (5–7 bp) of homology with the *AluI* repeat of the substrate (Fig. 6).

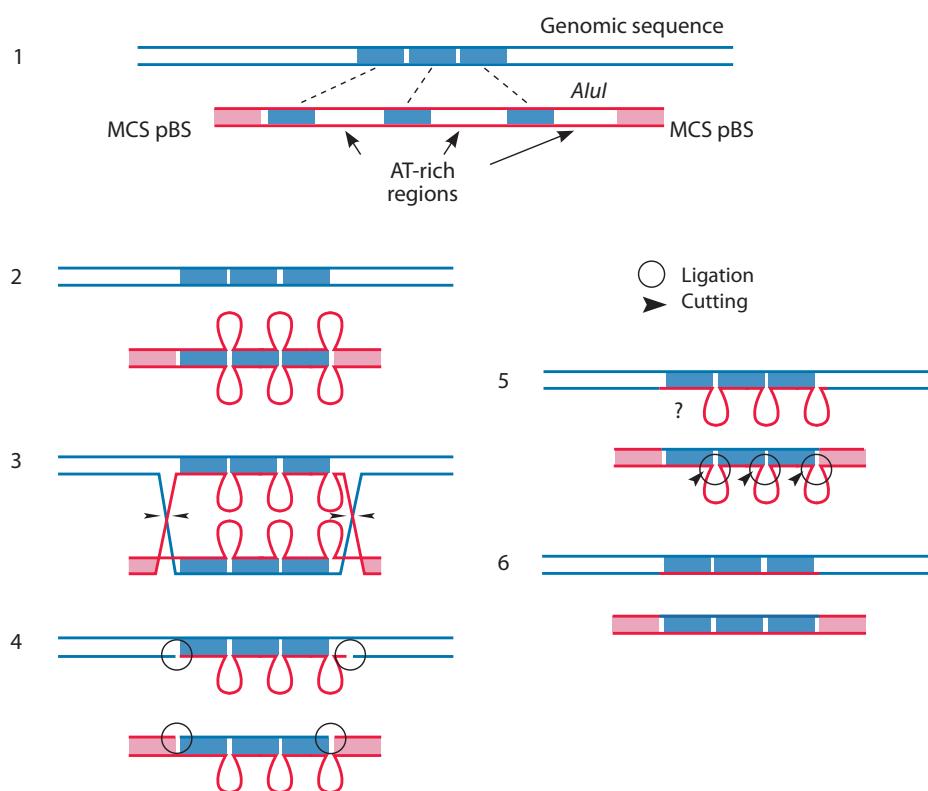


Fig. 6. Schematic representation of possible events during the formation of one of the detected intermediates identified by analyzing the full-genome sequencing reads.

1 – genomic DNA and substrate delivered to the cell. Homologous regions are highlighted in blue; 2–4 – sequential steps of recombination events, with regard to AT looping of AT-rich nonhomologous sites; 5, 6 – predicted recombinant products, with sequencing data included.

The TIR is known to flank the sequence of the mobile genetic element hAT-5_RIr. hAT elements are a superfamily of DNA transposons that are moved around in the genome by the transposase enzyme. Their presence has been confirmed within the genomes of plants, fungi, and animals. In mammals and humans, however, these elements exhibit no activity. They are approximately 2.5–5 kbp in size and contain terminal inverted repeats (TIR). The analysis of the structures of the final intermediates reveals that the complete substitution of the *AluI* repeat of the original substrate for the DNA sequence bounded by terminal inverted repeats occurred exactly on the flanks of the *AluI* repeat, a mobile genetic element of the human genome (Hagan et al., 2003). One intermediate, in which substitution preserved both halves of the *AluI*-EcoRV cloning site, exhibits clear homology with the *AluI* substrate.

Central to the analysis is the observation that, in all instances, the centrally located *AluI* nucleotide sequence has been replaced solely by genomic sequences at the EcoRV cloning site. Recombination was limited to homologous sequences between the genome and the *AluI* substrate, with nonhomologous plasmid DNA excluded. The analysis revealed no reads in which plasmid DNA sequences were enclosed by genomic DNA sequences, thus suggesting nonhomologous insertion events. A logical interpretation of recombinational events in a read, showing complete preservation of substrate structure and complete replacement of the genomic sequence with the *AluI* DNA sequence, is depicted in Figure 6.

2. Two reads were identified, with the right side of the hydrolysate (HindIII-primer M13 for) of the starting fragment being in the same continuous sequence as the human genome segment (54 and 57 bp), joined through the M13 for and genome microhomology site (Fig. 7). Moreover, the second part of both reads terminates with a cut-off strictly at the HindIII restriction site, which is the “terminus” of the substrate. It should be noted that the genomic DNA region of the two reads is 95 % homologous, implying that the integration may have occurred either in different parts of the genome or in the same part of the genome but in different cells where the acceptor sequences are not 100 % homologous.

A hypothesis is proposed to explain the discovered structure. The genomic sequence of the intermediate experienced a single-stranded break at its outset, or four letters before the start of the genomic sequence of the hybrid read. In this region, four nucleotides within the genomic sequence are complementary (homologous) to the four terminal nucleotides of the M13 forward primer. The fusion of extrachromosomal and genomic DNA within the polylinker matrix resulted from single-strand annealing, a process initiated by microhomology-based complementary pairing and D-loop formation (McVey, Lee, 2008; Hastings et al., 2009). The characteristic feature of both read sequences is their strict termination at the HindIII restriction site. The probability of the original DNA undergoing a precise, HindIII-clear

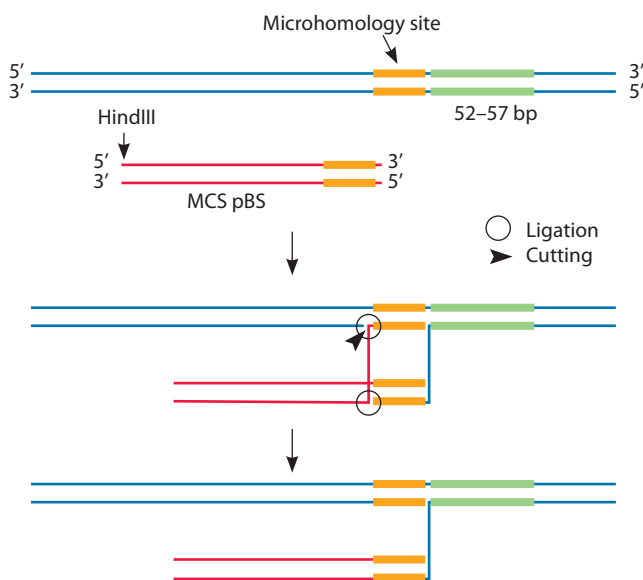


Fig. 7. Schematic representation of possible events during the formation of intermediates detected when analyzing the results of full-genome sequencing, in which a section of genomic DNA is combined with a plasmid polylinker.

It is suggested that recombination occurs at the site of a single-stranded genomic DNA break by a single-stranded annealing mechanism due to complementary microhomology pairing. The genomic DNA sequence is indicated in blue. The green color indicates the section of genomic DNA detected in the two derived reads. The yellow color indicates the microhomology section. The pBS polylinker (MCS pBS) is indicated in red.

double break during simultaneous ultrasonic processing of two fragments is rather low. This finding demonstrates that HindIII-terminated tails are extranuclear and persist freely in the nucleus for 15 days without genomic integration. Figure 7 illustrates potential intermediates from the aforementioned process.

So, intermediates were found that indicate the possibility of homologous exchange between extracellular DNA and chromatin DNA.

Additionally, unusual structures were identified that exhibited covalent attachment at one terminus to the recipient genome, while the opposing terminus was cleaved at the restriction site employed in substrate preparation (Fig. 3C). The data suggest that this part of the sequence is either free within the nucleus, lacking association with the second DNA strand, or is associated with a protein structure required for chromatin compaction into chromosomes and subsequent cell cycle events. We hypothesize that the observed association in these cases resulted from microhomology-driven interactions between genomic and M13 primer DNA at the single-strand break.

The identification of several intermediates is indicative of substrate DNA recombination. The identified reads exhibited multiple, chaotic permutations of both substrate and non-substrate sequences. These findings further support the occurrence of active recombination within the cell.

Schematic representation of the sequence of events and intermediate variants occurring during the act of homologous recombination

An intermediate resulting from recombination of an internal *AluI* repeat cloned at the *EcoRV* site and framed by polylinker regions bounded by primers to M13 with genomic DNA. The genomic DNA sequence located after the first terminal inverted repeat (TIR) is homologous to segments of the *AluI* repeat that are separated by AT-rich regions (spacers) (Fig. 6). A possible explanation for the convergence of the *AluI* spacer-containing substrate sequence to a homologous genomic region involves the looping-out of AT-rich spacer sequences, followed by the linear convergence of *AluI* substrate segments exhibiting homology with the genomic DNA. In this scenario, it is possible that homologous recombination took place between the genomic sequence and the substrate used in the cellular treatment. This case involved the complete replacement of the *AluI* repeat element in the extrachromosomal DNA with a genomic DNA fragment. Such events can arise from reciprocal homologous recombination at either open chromatin sites or arbitrary genomic locations. This result indicates an active process of homologous precise recombination between extracellular DNA and chromatin DNA.

A part of the polylinker that terminates at the *HindIII* restriction site at one end and is joined at the other end to DNA of genomic origin by microhomology. Genomic restructuring in differentiating cells is initiated via pangenomic single-strand breaks, which resolve within one week. The observed intermediates appear to correlate with the presence of single-stranded DNA breaks concurrent with the nuclear entry of extracellular DNA probe fragments (Fig. 7).

The situation allows for two possible interpretations. The integration event may have occurred during the initial committing phase, with intermediates emerging and existing prior to metaphase formation. The intermediary steps may have taken place concurrently with, or shortly prior to, DNA isolation on day 15 of cell culture. This variant involves the prolonged retention of short substrate fragments in the nucleus space. The retention of intermediate products throughout the duplication process is not clearly established. Considering the integration, it is probable that embedding occurred following chromosome doubling at those sites. The significance of such post-replicative disruptions to chromatin condensation and metaphase chromosome formation is seemingly negligible. The polymerase is unlikely to adopt such intermediate conformations during replication, which would require subsequent correction by specialized complexes.

The genomic DNA regions in both strains share 95 % homology, as previously stated. This indicates integration events at either different genomic loci or at the same locus in cells with non-identical acceptor sequences.

Reads with multiple chaotic permutations of substrate sequences and non-substrate sequences. Figure 8 presents a schematic representation of sequences exhibiting multiple chaotic permutations generated through sequencing.

In the last series of experiments, the feasibility of integrating extracellular fragments into the genome by direct detection of the integrated fragment was evaluated. Two types of TAMRA-labeled DNA, homologous human DNA and *E. coli* DNA, were used. The main conclusion of these experiments is that, in the case of human DNA, specific signals are detected throughout the chromosomes. In the case of *E. coli* DNA, these signals were found to be more localized in centromeric heterochromatin. Assuming that integration into the genome has occurred, the mode of this integration will be different for homologous and heterologous DNA, namely homologous incorporation of the human substrate and nonhomologous integration of *E. coli* DNA. Centromeric heterochromatin is a bulk (conglomerate) of repeats corresponding by mass to 1/3 of the chromosome mass, creating its known three-dimensional architecture. Assuming integration into centromeric heterochromatin, the considerable size and specific organization of α -satellites suggest that the addition of DNA does not compromise the functional integrity of the centromeric locus. This observation may clarify the process by which heterologous DNA is integrated.

Discrepancies between the results of FISH, direct detection of the fluorescently labeled probe at metaphases, and the data of whole-genome sequencing

An analysis was conducted on the results of FISH, i. e. direct detection of possible integration of a fluorescence-labeled probe into the genome, and data of full genomic sequencing in the experiments performed using a probe consisting of a mixture of M13F-*AluI*-M13R, M13F-*AluI*-*EcoRI*, and M13R-*AluI*-*HindIII* fragments. This analysis revealed the apparent discrepancy between the number of hybridization signals, or simply specific fluorescent signals, and the number of reads demonstrating the actual covalent association of the substrate sequence and the chromosome sequence. The most apparent initial assumption attributed the result to methodological artifacts. However, analysis of different experimental variants, and in particular, the nuclear signal patterns, indicated atypical and specific signal characteristics. Therefore, supplementary experiments were undertaken in an attempt to clarify the underlying mechanism of metaphase signaling.

The findings have redirected our focus to a currently debated phenomenon: the presence of substantial ring-shaped extracellular DNA (Pecorino et al., 2022).

The study of naturally occurring extrachromosomal circular DNA (eccDNA), abundant in diverse tumor cell types, commenced over six decades ago with the identification of double minutes, circular extrachromosomal structures

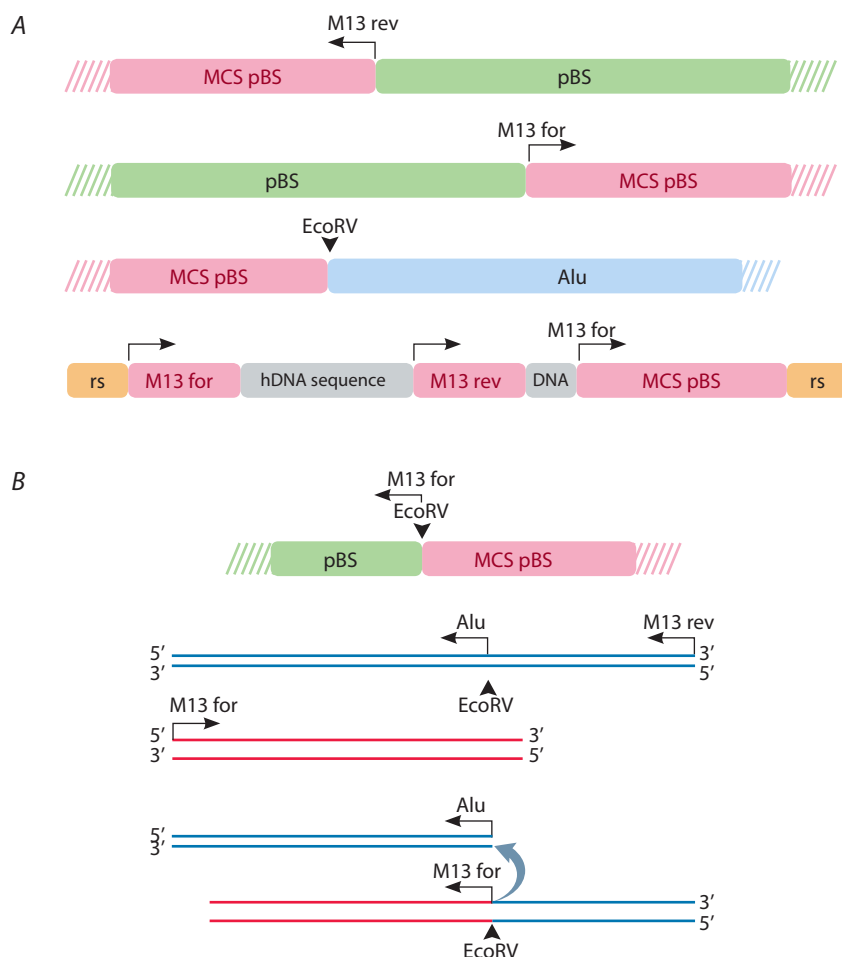


Fig. 8. Schematic representation of sequences obtained by sequencing (A) and a possible variant of the reparative process (B).

lacking centromeres and telomeres (Spriggs et al., 1962). The lack of centromeres leads to a random distribution of extracellular DNA during cell division, with the information encoded in them being transmitted in a non-Mendelian manner. These structures are formed from chromosomal material. Two structural categories are identified, the first characterized by single DNA sequences, and the second, by a chimeric structure composed of multiple chromosomal fragments. This observation indicates that chromosomal DNA fragments, detached through undetermined mechanisms, are randomly joined to form ring structures. Some of these DNA structures are up to 1 million bp in size and contain numerous genes and regulatory regions. Micro extrachromosomal circular DNAs of 100–2,000 bp in length are more prevalent in healthy cells. These DNAs contain ribosomal genes, DNAs of mobile elements, and telomeric ring DNAs. In tumor cells, these extrachromosomal structures can reintegrate into the genome, with “non-native” chromosomal regions with homogeneous staining containing amplified genes (Pecorino et al., 2022). Extrachromosomal circular DNA has been found to bind

closely to the polymerase complex (RNAPII) and is thought to associate with actively transcribed chromosome domains with which it forms chromatin concentration foci (Wu et al., 2019; Zhu et al., 2021). This finding suggests that associations of extrachromosomal ring DNA and chromosome DNA are stable and can persist in various phases of the cell cycle.

Similar DNAs were reported to be present in human myocytes and leukocytes, indicating that these structures appear to be a common extrachromosomal object and represent a general biologic phenomenon (Møller et al., 2018).

Taking into account the above, it can be hypothesized that in our study, the probe containing the plasmid tag formed chromatin concentration foci with chromosomes without integrating into the chromosome DNA molecule. These associates persisted until metaphase and gave rise to multiple specific hybridization signals. Is there consistency between this assumption and the data obtained? Is such a form of intermediates possible?

The findings of whole-genome sequencing indicate that homologous probe integration is possible through either

microhomology pairing or homologous recombination in the analyzed case of the *AluI* sequence. However, the number of reads-integrates obtained does not correspond to the hybridization pattern, with numerous specific signals (as initially assumed hybridization sites) detected at metaphases and, especially, concerning their patterning in interphase nuclei. Additionally, the results of direct integration analysis using fluorescently labeled homologous and heterologous probes, showing multiple specific signals (Fig. 2), require further clarification.

There is a lack of consistency between the quantified sequencing data and the observed hybridization pattern. Sequencing necessitates 40 ng of DNA, representing approximately 4,000 nuclei. Cumulatively, two repeats of the whole-genome reads revealed approximately 50 tag-carrying reads for each sequence. Thus, there is one substrate molecule per 80 cells. This is also inconsistent with the hybridization pattern, as multiple specific signals (hybridization signals) are detected at both metaphase and interphase nuclei (Fig. 2).

The facts and their analysis are more likely to suggest the formation of a large number of strong associates, which, together with the actual integration, give the overall FISH picture. The question remains as to why these multiple associations are not detected by full-genome sequencing. One suggestion is a technical problem.

The internalization of dsDNA fragments (nine telomeric repeats, 54 bp) into human and murine CD34 cells was experimentally confirmed, with results indicating that the dsDNA probe was successfully delivered to both CD34+ and CD34- cell populations (Ruzanova et al., 2024).

In CD34+ cells, the dsDNA substrate treated with the cells was found to be retained in its original form as an intact fragment. At the same time, in a small percentage of CD34- cells, the delivered fragments were concatemerized and possibly cyclized (Dolgova et al., 2013), with the number of monomers (54 bp) reaching seven units (~350 bp). After capturing the fragments, the CD34+ cells do not seem to be able to process them until reaching a committed state. Two potential scenarios can be suggested. The first scenario involves CD34- cells of the first echelon of commitment, which have retained the ability to capture extracellular DNA, process it, and cyclize internalized monomers. In the second scenario, CD34+ cells undergo terminal differentiation upon incubation with extracellular DNA substrate due to intracellular DNA fragments. This leads to CD34 marker loss and the acquisition of the ability to process the internalized fragments, forming detectable ring-like structures characteristic of CD34- cells. These DNA-associated structures successfully reach metaphase stages.

The Illumina platform full-genome sequencing technique involves fragmenting genomic DNA into 100–200 bp fragments. The quantitative likelihood of a 100–300 bp ring remaining intact following sonication is high. However, plasmid tag sequences (approximately 200 bp in this

experiment) will not be detected, although they are fully present on the chromosomes. It is this assumption that we believe to account for the contradictory findings yielded by FISH, whole-genome sequencing, and direct detection of fluorescently labeled probe integration.

A comparative analysis of the size of the ring formed by the oligonucleotide (about 350 bp) (Ruzanova et al., 2024), the diameter of it encompassing a linear molecule of the interphase chromosome, and the size of the polymerase complex suggests the possibility of the replicative enzymatic machinery moving freely through the ring during replication, with dsDNA strand thickness ~2 nm, one 10.5 bp dsDNA strand ~3.4 nm, 300 bp fragment ~100 nm, 300 bp dsDNA ring diameter ~32 nm, and polymerase complex ~300 kDa ~20 nm (<https://www.dynamic-biosensors.com/project/list-of-protein-hydrodynamic-diameters/>). The ring structure remains intact until the metaphase stage, at which point it is detectable via FISH.

Direct evidence for the long fragments of extracellular dsDNA (pEGFP-N1 linearized HindIII) forming a ring spanning the chromosome strand

The objective contradictions and doubts raised prompted a necessary inquiry into the signaling processes occurring on metaphase chromosomes.

Our previous research (Dolgova et al., 2013) using the example of linearized plasmid pEGFP-N1 (4.7 kbp) demonstrated that once a recombinogenic situation is induced in HSCs (mouse model), triggered by double-strand breaks as a result of the cross-linking cytostatic cyclophosphamide (“death window” phenomenon), linear plasmid fragments internalized into these cells undergo metabolic changes. The concatamers arrange themselves into ring formations, with a maximum of four monomers per ring. A subset of fragments undergoes end hydrolysis to a depth of 1 kbp, maintaining this structure within the nucleus. Of key importance is the observation that plasmid ring forms are isolated along with chromosomal DNA at the bottom of a 15–30 % NaCl gradient (~300,000g) subsequent to exhaustive proteinase K digestion and phenolic extraction. Specifically, plasmid DNA interacts with chromosomal DNA within a rigidly fractionated environment, while lacking any protein-mediated structural linkage. Common sense suggests two possibilities: plasmid integration into the chromosome or plasmid circularization around the chromosome, both resulting in cosedimentation with genomic DNA in the salt gradient. The significance of this fact was not initially recognized during the results analysis.

The present study shows that, similar to the results of the cited work, in human HSCs, when extracellular dsDNA fragments enter them, a recombinogenic situation is induced, provoked by the nicks that these fragments initiated. A comparison of the results obtained in the second part of this study (Ruzanova et al., 2024) with the earlier findings (Dolgova et al., 2013) indicates that during the two re-

combinogenic situations with different origins (induced by double-strand breaks and nicks), after entering the cell, the extracellular fragments undergo similar changes at the first stage of metabolic transformations. Their concatamerization (cross-linking/ligation at the ends) possibly implies that all other transformations of the fragments in the nucleus are similar to those shown previously (Dolgova et al., 2013). In other words, the fragments are either concatamerized and/or form a ring, and in the process of cyclization in the stochastic mode, the DNA strand of chromosomes can be looped by the cyclized plasmid. We can thus conclude that all experiments examining metaphase chromosome fluorescence signals have implicitly identified these structures.

To validate the aforementioned hypothesis, a set of experiments was conducted, closely replicating those conducted earlier (Dolgova et al., 2013) (Fig. 4). With the exception of the cell model (murine vs human bone marrow), the key differentiating factor was the method of recombinogenic state induction. Previous studies utilized double-strand breaks (interchain cross-links), whereas the current investigation uses nicks. The results obtained were virtually indistinguishable. Upon entry into hematopoietic stem cells, the linearized plasmid undergoes cyclization via the cohesive HindIII ends, resulting in a ring structure. Cyclization involves plasmid-induced looping of the chromosomal DNA, which co-purifies with the chromosomal DNA fraction at the bottom of a 15–30 % NaCl gradient during density gradient centrifugation. This structure, a long-standing component of the nucleus, is likely detectable “as a constituent” of the metaphase chromosome (Fig. 4). This interpretation of events aligns with whole-genome sequencing data, revealing no evidence of nonhomologous large-scale integration resulting in multiple discernible signals on metaphase chromosomes.

The processes governing the fate of these intermediate structures during metaphase chromosome packaging and division are yet to be elucidated. It can be hypothesized that ring structures undergo either physical fragmentation or enzymatic processing, resulting in linearization. Regardless, the stored genetic information may comprise an informational “reservoir” of currently undetermined function. The loosely arranged rings formed by extracellular dsDNA fragments may be analogs of rings that are involved in the mechanism of alternative telomere elongation.

Conclusion

The occurrence of pangenomic single-strand breaks (nicks) consequently activates cellular repair and recombination processes, thereby inducing recombination between chromatin DNA and exogenous dsDNA. It is worth noting that no nonhomologous end joining (NHEJ) events were detected between extracellular/extrachromosomal substrate DNA and chromosomal DNA. These findings indicate the potential for homologous genome modification during this biological process while significantly reducing the

likelihood of nonhomologous, random, and functionally detrimental integrations. Extracellular DNA fragments trapped in HSCs concatamerize, form rings, and establish long-lasting spatial associations with chromosomal DNA. It is possible for extrachromosomal rings to be the DNA matrix used by the cell in the mechanism of alternative telomere elongation. Finally, extracellular dsDNA fragments in the cell under the pressure of the recombination situation induced by both double-strand breaks and nicks undergo similar metabolic transformations, indicating the commonality and consistency of the two processes, which do not “conflict” and can complement one another.

References

- Cromie G.A., Connelly J.C., Leach D.R.F. Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. *Mol Cell*. 2001;8(6):1163-1174. doi 10.1016/S1097-2765(01)00419-1
- Dolgova E.V., Nikolin V.P., Popova N.A., Proskurina A.S., Orishchenko K.E., Alyamkina E.A., Efremov Y.R., ... Taranov O.S., Rogachev V.A., Sidorov S.V., Bogachev S.S., Shurdov M.A. Internalization of exogenous DNA into internal compartments of murine bone marrow cells. *Russ J Genet Appl Res*. 2012;2(6):440-452. doi 10.1134/S2079059712060056
- Dolgova E.V., Efremov Y.R., Orishchenko K.E., Andrushkevich O.M., Alyamkina E.A., Proskurina A.S., Bayborodin S.I., ... Omigov V.V., Minkevich A.M., Rogachev V.A., Bogachev S.S., Shurdov M.A. Delivery and processing of exogenous double-stranded DNA in mouse CD34+ hematopoietic progenitor cells and their cell cycle changes upon combined treatment with cyclophosphamide and double-stranded DNA. *Gene*. 2013;528(2):74-83. doi 10.1016/j.gene.2013.06.058
- Farzaneh F., Zalin R., Brill D., Shall S. DNA strand breaks and ADP-ribosyl transferase activation during cell differentiation. *Nature*. 1982; 300(5890):362-366. doi 10.1038/300362A0
- Glover D.M. (Ed.) DNA Cloning. A Practical Approach. IRL Press, 1988
- Hagan C.R., Sheffield R.F., Rudin C.M. Human Alu element retrotransposition induced by genotoxic stress. *Nat Genet*. 2003;35(3): 219-220. doi 10.1038/ng1259
- Hastings P.J., McGill C., Shafer B., Strathern J.N. Ends-in vs. ends-out recombination in yeast. *Genetics*. 1993;135(4):973-980. doi 10.1093/genetics/135.4.973
- Hastings P.J., Ira G., Lupski J.R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet*. 2009;5(1):e1000327. doi 10.1371/journal.pgen.1000327
- Holmgren L., Szeles A., Rajnavölgyi E., Folkman J., Klein G., Ernberg I., Falk K.I. Horizontal transfer of DNA by the uptake of apoptotic bodies. *Blood*. 1999;93(11):3956-3963. doi 10.1182/blood.V93.11.3956
- Jacobson G.K., Pinon R., Esposito R.E., Esposito M.S. Single-strand scissions of chromosomal DNA during commitment to recombination at meiosis. *Proc Natl Acad Sci USA*. 1975;72(5):1887-1891. doi 10.1073/pnas.72.5.1887
- Johnstone A.P., Williams G.T. Role of DNA breaks and ADP-ribosyl transferase activity in eukaryotic differentiation demonstrated in human lymphocytes. *Nature*. 1982;300(5890):368-370. doi 10.1038/300368A0
- Kucherlapati R.S., Eves E.M., Song K.Y., Morse B.S., Smithies O. Homologous recombination between plasmids in mammalian cells can be enhanced by treatment of input DNA. *Proc Natl Acad Sci USA*. 1984;81(10):3153-3157. doi 10.1073/pnas.81.10.3153

- Kumagai A., Dunphy W.G. Claspin, a novel protein required for the activation of Chk1 during a DNA replication checkpoint response in *Xenopus* egg extracts. *Mol Cell*. 2000;6(4):839-849. doi 10.1016/S1097-2765(05)00092-4
- Leung W., Malkova A., Haber J.E. Gene targeting by linear duplex DNA frequently occurs by assimilation of a single strand that is subject to preferential mismatch correction. *Proc Natl Acad Sci USA*. 1997;94(13):6851-6856. doi 10.1073/pnas.94.13.6851
- Li J., Read L.R., Baker M.D. The mechanism of mammalian gene replacement is consistent with the formation of long regions of heteroduplex DNA associated with two crossing-over events. *Mol Cell Biol*. 2001;21(2):501-510. doi 10.1128/mcb.21.2.501-510.2001
- Likhacheva A.S., Rogachev V.A., Nikolin V.P., Popova N.A., Shilov A.G., Sebeleva T.E., Strunkin D.N., Chernykh E.R., Gel'fgat E.L., Bogachev S.S., Shurdov M.A. Involvement of exogenous DNA in the molecular processes in somatic cell. *Informatsionny Vestnik VOGiS = Herald Vavilov Soc Genet Breed Sci*. 2008;12(3):426-447 (in Russian)
- Lin J., Krishnaraj R., Kemp R.G. Exogenous ATP enhances calcium influx in intact thymocytes. *J Immunol*. 1985;135(5):3403-3410
- Lin Y., Waldman A.S. Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics*. 2001;158(4):1665-1674. doi 10.1093/genetics/158.4.1665
- MacDougall C.A., Byun T.S., Van C., Yee M.C., Cimprich K.A. The structural determinants of checkpoint activation. *Genes Dev*. 2007;21(8):898-903. doi 10.1101/gad.1522607
- Maizels N., Davis L. Initiation of homologous recombination at DNA nicks. *Nucleic Acids Res*. 2018;46(14):6962-6973. doi 10.1093/nar/gky588
- Maniatis T., Fritsch E., Sambrook D. Methods of Genetic Engineering. Molecular Cloning. Moscow: Mir Publ., 1984 (in Russian)
- Maricic T., Whitten M., Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*. 2010;5(11):e14004. doi 10.1371/journal.pone.0014004
- McVey M., Lee S.E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet*. 2008;24(11):529-538. doi 10.1016/j.tig.2008.08.007
- Møller H.D., Mohiyuddin M., Prada-Luengo I., Sailani M.R., Halling J.F., Plomgaard P., Maretty L., Hansen A.J., Snyder M.P., Pilegaard H., Lam H.Y.K., Regenbreg B. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nat Commun*. 2018;9(1):1069. doi 10.1038/S41467-018-03369-8
- Murnane J.P., Yezzi M.J., Young B.R. Recombination events during integration of transfected DNA into normal human cells. *Nucleic Acids Res*. 1990;18(9):2733-2738. doi 10.1093/nar/18.9.2733
- Pecorino L.T., Verhaak R.G.W., Henssen A., Mischel P.S. Extrachromosomal DNA (ecDNA): an origin of tumor heterogeneity, genomic remodeling, and drug resistance. *Biochem Soc Trans*. 2022;50(6):1911-1920. doi 10.1042/BST20221045
- Perucho M., Hanahan D., Wigler M. Genetic and physical linkage of exogenous sequences in transformed cells. *Cell*. 1980;22(1):309-317. doi 10.1016/0092-8674(80)90178-6
- Perucho M., Wigler M. Linkage and expression of foreign DNA in cultured animal cells. *Cold Spring Harb Symp Quant Biol*. 1981;45(Pt.2):829-838. doi 10.1101/sqb.1981.045.01.101
- Petrova D.D., Dolgova E.V., Proskurina A.S., Ritter G.S., Ruzanova V.S., Efremov Y.R., Potter E.A., Kirikovich S.S., Levites E.V., Taranov O.S., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. The new general biological property of stem-like tumor cells (Part II: surface molecules, which belongs to distinctive groups with particular functions, form a unique pattern characteristic of a certain type of tumor stem-like cells). *Int J Mol Sci*. 2022;23(24):15800. doi 10.3390/ijms232415800
- Pierandrei S., Luchetti A., Sanchez M., Novelli G., Sangiuolo F., Lucarelli M. The gene targeting approach of small fragment homologous replacement (SFHR) alters the expression patterns of DNA repair and cell cycle control genes. *Mol Ther Nucleic Acids*. 2016;5(4):e304. doi 10.1038/mtna.2016.2
- Potter E.A., Dolgova E.V., Proskurina A.S., Efremov Y.R., Minkevich A.M., Rozanov A.S., Peltek S.E., ... Baiborodin S.I., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. Gene expression profiling of tumor-initiating stem cells from mouse Krebs-2 carcinoma using a novel marker of poorly differentiated cells. *Oncotarget*. 2017;8(6):9425-9441. doi 10.18632/oncotarget.14116
- Potter E.A., Dolgova E.V., Proskurina A.S., Ruzanova V.S., Efremov Y.R., Kirikovich S.S., Oshikhmina S.G., ... Grivtsova L.U., Kolchanov N.A., Ostanin A.A., Chernykh E.R., Bogachev S.S. Stimulation of mouse hematopoietic stem cells by angiogenin and DNA preparations. *Braz J Med Biol Res*. 2024;57:e13072. doi 10.1590/1414-431X2024e13072
- Ritter G.S., Dolgova E.V., Petrova D.D., Efremov Y.R., Proskurina A.S., Potter E.A., Ruzanova V.S., Kirikovich S.S., Levites E.V., Taranov O.S., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Bogachev S.S. The new general biological property of stem-like tumor cells. Part I. Peculiarities of the process of the double-stranded DNA fragments internalization into stem-like tumor cells. *Front Genet*. 2022;13:954395. doi 10.3389/fgene.2022.954395
- Rogachev V.A., Likhacheva A., Vratskikh O., Mechetina L.V., Sebeleva T.E., Bogachev S.S., Yakubov L.A., Shurdov M.A. Qualitative and quantitative characteristics of the extracellular DNA delivered to the nucleus of a living cell. *Cancer Cell Int*. 2006;6:23. doi 10.1186/1475-2867-6-23
- Ruzanova V., Proskurina A., Efremov Y., Kirikovich S., Ritter G., Levites E., Dolgova E., Potter E., Babaeva O., Sidorov S., Taranov O., Ostanin A., Chernykh E., Bogachev S. Chronometric administration of cyclophosphamide and a double-stranded DNA-mix at interstrand crosslinks repair timing, called "Karanahan" therapy, is highly efficient in a weakly immunogenic Lewis carcinoma model. *Pathol Oncol Res*. 2022;28:1610180. doi 10.3389/pore.2022.1610180
- Ruzanova V.S., Oshikhmina S.G., Proskurina A.S., Ritter G.S., Kirikovich S.S., Levites E.V., Efremov Y.R., Karamysheva T.V., Meschaninova M.I., Mamaev A.L., Taranov O.S., Bogachev A.S., Sidorov S.V., Nikonov S.D., Leplina O.Y., Ostanin A.A., Chernykh E.R., Kolchanov N.A., Dolgova E.V., Bogachev S.S. A concept of natural genome reconstruction. Part 2. Effect of extracellular double-stranded DNA fragments on hematopoietic stem cells. *Vavilovskii Zhurnal Genetiki i Selektii = Vavilov J Genet Breed*. 2024;28(8):993-1007. doi 10.18699/vjgb-24-106
- Smith A.J.H., Berg P. Homologous recombination between defective neo genes in mouse 3T6 cells. *Cold Spring Harb Symp Quant Biol*. 1984;49:171-181. doi 10.1101/sqb.1984.049.01.020
- Smithies O., Gregg R.G., Boggs S.S., Koralewski M.A., Kucherlapati R.S. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature*. 1985;317(6034):230-234. doi 10.1038/317230A0
- Spriggs A.I., Boddington M.M., Clarke C.M. Chromosomes of human cancer cells. *Br Med J*. 1962;2(5317):1431-1435. doi 10.1136/bmj.2.5317.1431
- Thomas K.R., Folger K.R., Capecchi M.R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell*. 1986;44(3):419-428. doi 10.1016/0092-8674(86)90463-0
- Vriend L.E.M., Krawczyk P.M. Nick-initiated homologous recombination: protecting the genome, one strand at a time. *DNA Repair*. 2017;50:1-13. doi 10.1016/j.dnarep.2016.12.005

- Wu S., Turner K.M., Nguyen N., Raviram R., Erb M., Santini J., Luebeck J., ... Furnari F.B., Chang H.Y., Ren B., Bafna V., Mischel P.S. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*. 2019;575(7784):699-703. doi 10.1038/S41586-019-1763-5
- Yakubov L.A., Deeva E.A., Zarytova V.F., Ivanova E.M., Ryte A.S., Yurchenko L.V., Vlassov V.V. Mechanism of oligonucleotide uptake by cells: involvement of specific receptors? *Proc Natl Acad Sci USA*. 1989;86(17):6454-6458. doi 10.1073/pnas.86.17.6454
- Yoo H.Y., Shevchenko A., Shevchenko A., Dunphy W.G. Mcm2 is a direct substrate of ATM and ATR during DNA damage and DNA replication checkpoint responses. *J Biol Chem*. 2004;279(51):53353-53364. doi 10.1074/jbc.M408026200
- Yoo H.Y., Jeong S.Y., Dunphy W.G. Site-specific phosphorylation of a checkpoint mediator protein controls its responses to different DNA structures. *Genes Dev*. 2006;20(7):772-783. doi 10.1101/GAD.1398806
- Zhu Y., Gujar A.D., Wong C.H., Tjong H., Ngan C.Y., Gong L., Chen Y.A., ... Yi E., deCarvalho A.C., Ruan Y., Verhaak R.G.W., Wei C.L. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell*. 2021;39(5):694-707.e7. doi 10.1016/j.ccell.2021.03.006
- Zou L. Single- and double-stranded DNA: building a trigger of ATR-mediated DNA damage response. *Genes Dev*. 2007;21(8):879-885. doi 10.1101/gad.1550307

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

Received July 4, 2024. Revised April 9, 2025. Accepted May 12, 2025.

doi 10.18699/vjgb-26-19

Naïve human iPSCs obtained by culturing the ICGi022-A cell line with primed pluripotency in HENSM medium efficiently differentiate into endothelial derivatives

M.A. Arssan , A.I. Shevchenko , S.M. Zakian , I.S. Zakharova 

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

 arsanmhdamin@yandex.ru

Abstract. Naïve human pluripotent stem cells (PSCs) are a promising new tool in biomedical research. They provide access to the early embryonic development programmes and offer breakthrough solutions in regenerative medicine. However, the current inability to obtain long-term cultures of genetically and epigenetically stable naïve human PSC lines poses a challenge to their effective application in biomedicine. The recently proposed HENSM culture medium is claimed to enable the obtaining and long-term maintenance of naïve PSC lines. In this study, the potential of the HENSM medium for obtaining stable naïve human PSC lines was investigated. We successfully reset the primed induced pluripotent stem cell (iPSC) line ICGi022-A (K7-4Lf), derived from a healthy donor, to a naïve state using the HENSM medium. Naïve iPSCs grow in the form of dome-shaped colonies, both with and without a feeder layer of cells. The resulting cells retained expression of the key pluripotency factors and activated the naïve PSC transcriptional programme, including expression of endogenous retroviral elements, early epiblast marker genes and genes associated with totipotency. The naïve iPSC line was capable of differentiating into derivatives of the three primary germ layers, as well as producing trophoblast derivatives. Culturing naïve iPSCs in low-adhesion conditions resulted in the spontaneous formation of three-dimensional structures (blastoids) resembling early human blastocysts. The X chromosome, which was in an eroded inactive state in the original cell line, was reactivated in the naïve cells, but returned to its normal inactive state when the naïve cells were re-primed. Notably, naïve iPSCs demonstrated limited ability to directly differentiate into endothelial cells. However, their competence to give rise to mature endothelial derivatives was restored upon returning to the primed state, achieving comparable efficiency to the original primed iPSCs. Thus, the resulting naïve iPSC line has significant potential for studying the early stages of embryogenesis and for other biomedical applications, including disease modelling. However, the naïve ICGi022-A line proved to be karyotypically unstable during long-term cultivation using HENSM medium. As there is a risk of karyotypic aberrations during the maintenance of naïve PSCs, further improvement of the culture conditions is necessary to obtain reliable, karyotypically stable lines of naïve pluripotent cells.

Key words: primed and naïve human pluripotent stem cells; directed endothelial differentiation; hereditary disease cellular models

For citation: Arssan M.A., Shevchenko A.I., Zakian S.M., Zakharova I.S. Naïve human iPSCs obtained by culturing the ICGi022-A cell line with primed pluripotency in HENSM medium efficiently differentiate into endothelial derivatives. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov J Genet Breed.* 2026;30(2):181-193. doi 10.18699/vjgb-26-19

Funding. The work was supported by the Russian Science Foundation grant No. 24-15-00346, <https://rscf.ru/project/24-15-00346/>

Acknowledgements. The immunofluorescent imaging was performed using resources of the Common Facilities Center of Microscopic Analysis of Biological Objects, ICG SB RAS (<https://ckp.icgen.ru/ckpmabo/>), supported by the State project of the Institute of Cytology and Genetics FWNR-2026-0024. M.A. Arssan is a member of Molecular Paleogenetics and Paleogenomics laboratory.

Ethical statement. This study did not involve human or animal subjects. The human iPSC line used in the study was obtained earlier in accordance with ethical considerations.

Наивные ИПСК человека, полученные при культивировании линии клеток ICGi022-A с праймированной плюрипотентностью в среде HENSM, способны к эффективной дифференцировке в эндотелиальные производные

M.A. Арссан , А.И. Шевченко , С.М. Закиан , И.С. Захарова 

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

 arsanmhdamin@yandex.ru

Аннотация. Наивные плюрипотентные стволовые клетки (ПСК) человека – новый многообещающий инструмент биомедицинских исследований, открывающий доступ к ранним программам эмбрионального развития и прорывным решениям задач регенеративной медицины. Препятствием на пути к эффективному внедрению наивных ПСК в биомедицину является отсутствие возможности получать длительно культивируемые генетически и эпигенетически стабильные линии наивных ПСК. Ранее было заявлено, что культуральная среда HENSM позволяет индуцировать и длительно поддерживать линии наивных плюрипотентных стволовых клеток человека. В рамках данной работы мы проверили возможность получать стабильные линии наивных ПСК с использованием среды HENSM. С помощью среды HENSM мы успешно перевели линию ICGi022-A (K7-4Lf) праймированных индуцированных плюрипотентных стволовых клеток (ИПСК) здорового донора в наивное состояние. Наивные ИПСК растут в форме сфероподобных колоний как на питающем слое клеток, так и без него. Полученные клетки сохранили экспрессию ключевых факторов плюрипотентности и одновременно активировали транскрипционную программу наивных ПСК, включающую экспрессию эндогенных ретровирусных элементов, генов-маркеров раннего эпибласта и генов, ассоциированных с тотипотентностью. Наивная линия ИПСК была способна дифференцироваться в производные трех первичных зародышевых листков, а также давать производные трофобласта. При культивировании наивных ИПСК в низкоадгезивных условиях наблюдалось спонтанное формирование трехмерных структур – бластоидов, морфологически напоминающих ранние бластоцисты человека. X-хромосома, имевшая нарушения в неактивном статусе в исходной линии клеток, реактивировалась в наивных клетках и восстанавливала нормальное неактивное состояние при повторном переводе наивных клеток в праймированное состояние. Важно отметить, что наивные ИПСК продемонстрировали низкую способность к прямой направленной дифференцировке в эндотелиоциты, однако их компетентность давать зрелые эндотелиальные производные восстанавливалась после их возврата к праймированному состоянию, достигая эффективности, сопоставимой с таковой у исходных праймированных ИПСК. Таким образом, полученная линия наивных ИПСК обладает значительным потенциалом для исследования ранних стадий эмбриогенеза и других биомедицинских приложений, включая моделирование заболеваний. Тем не менее при длительном культивировании с использованием среды HENSM наивная линия ICGi022-A оказалась кариотипически нестабильной. Поскольку риск кариотипических aberrаций при поддержании наивных ПСК сохраняется, в дальнейшем для получения надежных, кариотипически стабильных линий наивных плюрипотентных клеток необходимо совершенствование условий культивирования.

Ключевые слова: праймированные и наивные плюрипотентные стволовые клетки человека; направленная эндотелиальная дифференцировка; клеточные модели наследственных заболеваний

Introduction

Cultures of human pluripotent stem cells (hPSCs) can reflect various stages of embryonic development. The majority of hPSC lines obtained and used to date are in a primed state, resembling the post-implantation epiblast. These cells are stable in long-term cultivation and are widely used in biomedical research and clinical trials. However, they exhibit limited developmental and differentiation pathways, demonstrating the epigenetic memory phenomenon that biases differentiation towards specific lineages (Stadtfield, Hochedlinger, 2010; Gafni et al., 2013; Valamehr et al., 2014; Lee et al., 2017; Hu et al., 2020; Yu et al., 2021; Dekel et al., 2022). They cannot be used to model the earlier stages of embryonic development, such as X chromosome inactivation, and their chromatin is more closed, making them less accessible to genome editing by programmable nucleases. They also often demonstrate impaired X chromosome inactivation, termed erosion (Mekhoubad et al., 2012). In contrast, naïve hPSCs are in a more primitive developmental state that closely mirrors the pre-implantation epiblast. This gives them broader developmental plasticity than conventional (primed) hPSCs (Theunissen et al., 2016; Collier, Rugg-Gunn, 2018; Rostovskaya, 2022). This unique plasticity includes the ability to contribute to both the embryonic and extraembryonic lineages (Theunissen et al., 2014, 2016; Pham et al., 2022), enhanced proliferation and improved clonal expansion through tolerance of single-cell passaging. Greater amenability to genome editing and erased epigenetic memory through reduction of repressive chromatin marks also render naïve hPSCs a promising platform for advancing regenerative medicine, disease modelling and the study of early human de-

velopment. Recent advances have demonstrated the feasibility of resetting primed hPSCs to the naïve state in defined media composed small molecules and growth factors (Chan et al., 2013; Gafni et al., 2013; Takashima et al., 2014; Theunissen et al., 2014; Ware et al., 2014; Duggal et al., 2015; Zimmerlin et al., 2016; Guo et al., 2017; Lee et al., 2017; Szczerbinska et al., 2019; Bayerl et al., 2021; Khan et al., 2021; Buckberry et al., 2023). Despite these advantages, deriving and maintaining naïve hPSCs in the long term remains challenging. A major concern is the genomic and epigenetic instability reported in naïve cultures, particularly under prolonged or suboptimal conditions. Chromosome aberrations, karyotype abnormalities, and variable X-chromosome reactivation status have all been observed (Theunissen et al., 2014, 2016; Fischer et al., 2025). These vulnerabilities underscore the need for well-defined, optimized culture systems.

The recently developed HENSM medium shows promise in supporting the long-term stable propagation of naïve hPSC lines while preserving the key molecular and functional features of naïve pluripotency (Bayerl et al., 2021). During the medium formulation, it was recognised that WNT signalling activity was the main cause of cell heterogeneity, imprinting loss, and chromosomal rearrangements in naïve hPSC cultures. Thus, the medium lacks WNT activators, such as GSK3 kinase inhibitors, and ensures almost complete downregulation of WNT signalling through tankyrase inhibition. The medium also contains LIF growth factor and MEK kinase inhibitor, two components that are common to all naïve conditions. LIF activates JAK/STAT3 signalling pathway, which is necessary for naïve pluripotency. The MEK inhibitor, together with

the PKC and SRC kinase inhibitors in the HENSM medium, downregulates the MEK/ERK cascade, which is responsible for primed pluripotency. The medium was tested on several ESC and iPSC cell lines. It was claimed that HENSM helps preserve genomic imprints and allows long-term naïve hPSC maintenance, which is crucial for accurate disease modelling and developmental studies. However, there are still no reports of long-term cultured naïve hPSCs obtained in the HENSM medium.

The aim of this study is to obtain stable naïve induced pluripotent stem cells (hiPSCs) in HENSM medium by deriving them from an existing primed hiPSC line, ICGi022-A (K7-4Lf), from a healthy donor (Malakhova et al., 2020). The primed K7-4Lf line is widely used for disease modeling, drug screening, and stem cell research due to its well-characterized genetic background and robust pluripotency (Malakhova et al., 2020; Klepikova et al., 2022; Ustyantseva et al., 2022; Zakharova et al., 2022, 2024; Grigor'eva et al., 2023, 2024; Kopytova et al., 2023; Sheveleva et al., 2023, 2024; Yarkova et al., 2024; Pavlova et al., 2024a, b; Rezvova et al., 2024; Nadtochy et al., 2025). Reprogramming K7-4Lf cells into a naïve state provides an opportunity to evaluate the developmental and differentiation capabilities of naïve pluripotency.

In this study, we reset primed hiPSCs into the naïve state and examined their developmental potential and differentiation abilities. We successfully derived naïve hiPSCs that exhibited key features of early embryonic cells, including spontaneous blastoids formation. Additionally, we investigated the ability of naïve hiPSCs to give rise to differentiated derivatives, with a particular focus on endothelial cells. Furthermore, we analysed the dynamics of X chromosome regulation, focusing on H3K27me3 and *XIST* RNA as indicators of epigenetic fidelity. Our findings emphasise the functional benefits and epigenetic challenges of naïve pluripotency, highlighting the necessity of improved resetting and culture protocols to maintain genomic and epigenetic stability.

Materials and methods

hiPSCs cultivation and differentiation. The human induced pluripotent stem cell line ICGi022-A (hPSCreg identifier: RRID:CVCL_ZE02) (Malakhova et al., 2020) with a karyotype of 46, XX was utilised in this study. The ICGi022-A line was derived and maintained under conditions typical of primed pluripotent cells on a layer of mitotically inactivated mouse embryonic fibroblasts. The medium was composed of DMEM/F12 (1:1), with the addition of 15 % serum replacement KnockOut SR (Thermo Fisher Scientific), 1 mM GlutaMax (Thermo Fisher Scientific), 1 % non-essential amino acid solution (NEAA) (Thermo Fisher Scientific), 50 U/ml penicillin and 50 µg/ml streptomycin (InvivoGen), as well as 0.25 mM 2-mercaptoethanol (Thermo Fisher Scientific) and 10 ng/ml bFGF (Sci-store).

Complete HENSM medium (Bayerl et al., 2021), for induction and maintenance of naïve state, consisted of a 1:1 mixture of Neurobasal and DMEM-F12 media, supplemented with 1 % N2, 1 % B27, 1 mM GlutaMAX, 1 % NEAA (all Thermo Fisher Scientific), 50 U/ml penicillin and 50 µg/ml streptomycin (InvivoGen), 0.2 % Geltrex (Thermo Fisher Scientific),

50 µg/ml L-ascorbic acid 2-phosphate (Sigma-Aldrich), 20 ng/ml recombinant LIF (Sci-store), and kinase inhibitors: 1 µM GSKi PD0325901, 2 µM TNKi XAV939, 2 µM PKCi Gö6983, 1.2 µM SRCi CGP77675 (all R&D) and 5 µM ROCKi thiazovivin (Stemolecule).

In order to expand cell cultures, colonies of both naïve and primed pluripotent cells were dissociated using a TrypLE enzyme (Thermo Fisher Scientific) and subsequently transferred into fresh medium containing 10 µM of the ROCKi thiazovivin (Stemolecule). The seeding ratio was established at 1:10 for primed iPSCs and at 1:3 for naïve iPSCs.

The spontaneous differentiation of naïve hiPSCs into derivatives of three primitive germ layers was performed in a monolayer. To achieve this, the cells were seeded at 30 % confluence on Matrigel (Corning), and the HENSM medium was replaced with a differentiation medium containing a 1:1 DMEM/F12 mixture (Gibco), 10 % fetal bovine serum (FBS) (Cell Technologies), 1 mM GlutaMax (Thermo Fisher Scientific), 1 % NEAA solution (Thermo Fisher Scientific), 50 U/ml penicillin and 50 µg/ml streptomycin (InvivoGen). The medium was changed every other day. The results of the spontaneous differentiation were analysed after 23 days using immunofluorescence.

The directed differentiation protocol for trophoblast derivatives started with the dissociation of confluent naïve iPSC colonies using TrypLE, followed by seeding at a density of 2×10^6 cells on a 10 cm² culture surface treated with collagen IV (Sigma-Aldrich) in HENSM medium. After two days, TSC medium was added to the cells as described in (Okoe et al., 2018). The TSC medium comprises a 1:1 DMEM/F12 mixture (Gibco), 0.2 % FBS (Cell Technologies), 0.3 % BSA (Sigma-Aldrich), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), 1 % ITS-X (Gibco), 1.5 µM ascorbic acid (Sigma-Aldrich), 50 ng/ml EGF (Peprotech), 2 µM CHIR99021 and 0.5 µM A83-01, 1 µM SB431542 (all R&D), 0.8 mM valproic acid (Sigma-Aldrich), 5 µM thiazovivin (Stemolecule), 50 U/ml penicillin and 50 µg/ml streptomycin (InvivoGen). The medium was changed daily. On day 5 after the addition of the TSC medium, once the flattened colonies had reached 50–70 % confluence, they were expanded using TrypLE enzymatic disaggregation at a ratio of 1:4. Homogeneous trophoblast cell cultures were formed by the fifth passage. Thereafter, trophoblast cells were passaged with TrypLE every three to four days at a ratio of 1:4 to 1:6. At the tenth passage, the cells were plated on coverslips and stained with antibodies for specific markers.

Endothelial derivatives of hPSCs were obtained through mesodermal progenitor derivation and subsequent endothelial differentiation, according to the Gu protocol with modifications (Gu, 2018). First, the pluripotent cells were seeded onto a matrigel-treated surface (Corning) at a confluence of 60–70 %, in the original iPSC culture medium. The next day, differentiation in the mesodermal direction was started in RPMI 1640 medium containing 1 % B27 supplementation without insulin (all Thermo Fisher Scientific), 50 U/ml penicillin, 50 µg/ml streptomycin (InvivoGen), and the GSK3 kinase inhibitor CHIR99021 (R&D). The concentration of CHIR99021 in the differentiation medium was 6 µM for two days, after which

it was reduced to 3 μ M for the following two days. The differentiating cells were then transferred to EGM-2 medium (Lonza), supplemented with 50 ng/ml VEGF, 25 ng/ml bFGF (both Sci-Store) and 10 μ M SB431542 (R&D), for endothelial growth. For the next eight days, half of the endothelial medium were changed daily. The proportion of cells containing the CD31 marker of mature endothelial cells was determined using a FACSaria III device (BD, USA) after precipitating the cells with the appropriate antibodies (Supplementary Material 1)¹ and analysing 10⁴ events. Unstained cells and cells incubated with fluorescently labelled mouse IgG1 were used as negative controls. The experiment was repeated three times. The statistical significance of the differences was measured using the Wilcoxon test with Bonferroni correction. Mature endothelial derivatives of hiPSCs were selected using magnetic sorting with antibodies to the human CD31 surface antigen (Miltenyi Biotec) and seeded onto a collagen IV-coated culture surface (Sigma-Aldrich) in EGM2 medium (Lonza).

Generation of human blastoids from naïve human iPSCs maintained in HENSM medium. For the generation of blastoids, naïve hiPSCs with 60–70 % confluence were dissociated into single cells using TrypLE Express at 37 °C for 3 minutes. The cells were then collected by centrifugation at 200g for 5 minutes and resuspended in fresh HENSM medium. The resulting cell suspension was incubated in a gelatinised tissue culture plate for 30 minutes at 37 °C in 5 % CO₂ to remove the feeder cells. The medium containing naïve human iPSCs was collected and passed through a 40 μ m cell strainer. The AggreWell 400 (STEMCELL Technologies) was prepared by rinsing with an anti-adhesion solution (STEMCELL Technologies), followed by centrifugation at 2,000g for 5 minutes and incubation at room temperature for 10 minutes, according to the manufacturer's instructions. After incubation, wells were washed once with HENSM medium, after which 0.5 ml of fresh HENSM containing 5 μ M hiazovivin (Stemolecule) was added.

Approximately 30,000 cells (around 25 cells per micro-well) were resuspended in 1 ml HENSM with 5 μ M thiazovivin (Stemolecule) and seeded into one well of a prepared AggreWell 400 24-well plate. The plate was centrifuged at 200g for 1 minute and placed at 37 °C. Aggregates were formed within 12–16 hours. The HENSM medium was then changed to TDM medium and replaced every two days for six days. The TDM previously described for 5iLA-naïve hiPSCs was slightly modified to adjust the HENSM protocols using the following: 1:1 (v/v) mixture of DMEM/F12 and Neurobasal medium, 0.5 $\times$ N2 supplement, 0.5 % 0.5 $\times$ GlutaMAX, 0.5 $\times$ non-essential amino acids (all Thermo Fisher Scientific), 0.5 $\times$ B27 supplement (Thermo Fisher Scientific), 0.1 mM β -mercaptoethanol (Sigma-Aldrich), 1 % ITS-X (Gibco), 0.5 % KnockOut SR (Thermo Fisher Scientific), 0.1 % FBS (Cell Technologies), 50 mg/ml BSA (Sigma-Aldrich), 50 U/ml penicillin and 50 μ g/ml streptomycin (InvivoGen), 1 μ M Gö6983, 1 μ M PD0325901, 0.5 μ M CGP77675, 0.5 μ M A83-01, 1.25 μ M CHIR99021, 0.5 μ M SB431542 (all R&D), 25 ng/ml recombinant human LIF (Sci-Store), 10 ng/ml EGF

(Peprotech), 0.75 μ g/ml L-ascorbic acid, and 0.4 mM valproic acid (both from Sigma-Aldrich). After six days, the blastoids were isolated and treated with a 4 % formaldehyde fixative solution in PBS for immunostaining with OCT4 antibody.

RNA isolation, cDNA synthesis and semi-quantitative real-time PCR. RNA was isolated from primed and naïve iPSCs using a TRIzol reagent (Thermo Fisher Scientific) and a DNA-free kit (Thermo Fisher Scientific). cDNA was synthesised using a reverse transcriptase (M-MuLV) (Biolabmix) and a random hexamer primer (Thermo Fisher Scientific), according to the manufacturers' instructions. The relative gene expression levels of *NANOG*, *TFCP2L1*, *KLF17*, *TFE3*, *LTR7Y*, *DAZL* and *LEUTX* were determined by semi-quantitative real-time PCR using the $\Delta\Delta$ Ct method, normalised to the levels of the housekeeping genes *ACTB*, *TFRC* and *B2M*. The experiment was performed in three biological and two technical replicates. The results were analysed using the qBase+ software (CellCarta, <https://cellcarta.com/genomic-data-analysis/>) and the generalised $\Delta\Delta$ Ct method was employed, taking into account the reaction efficiency, which was calculated from the calibration curve results. The sequences of the oligonucleotides synthesised at Biosset (<https://www.biosset.com/>) are given in the Supplementary Material 2.

Immunofluorescent staining and RNA FISH. We used indirect immunofluorescence staining to detect surface antigens, transcription factors, and the epigenetic state of the X chromosome in hiPSCs, as well as markers of their differentiated derivatives. During cell preparation and antibody staining, we follow the protocol described previously (Vaskova et al., 2015; Zakharova et al., 2017, 2022). The cells were fixed in a 4 % formaldehyde solution for 10 minutes, permeabilised with a 0.5 % Triton X-100 solution (Sigma-Aldrich) for 30 minutes (this step was omitted for surface antigens), and blocked with a 1 % bovine serum albumin solution in 1x PBS. All procedures were performed at room temperature. Primary antibody incubation was conducted overnight at 4 °C. Precipitation between cells and fluorescently labelled secondary antibodies was performed in the dark at room temperature for one hour. Cell nuclei were stained with DAPI. The preparations were visualized and imaged using a Ti-E inverted fluorescence microscope (Nikon) and NIS Advanced Research software. Antibodies, their manufacturers, species of origin, and dilutions are listed in Supplementary Material 2.

To examine the epigenetic state of the X chromosome, immunofluorescent staining with antibodies to chromatin modification of the inactive X chromosome was combined with RNA FISH with probes to the RNA *XIST* (BAC clone RP11-256P2) and X-linked gene *HUWE1* (BAC clone RP11-975N19), according to the protocol given in (Vaskova et al., 2015). Images were acquired and analysed using a Nikon Ti-E inverted fluorescent microscope (Japan) and NIS Elements AR software. At least 300 nuclei were analysed in each experiment.

STR analysis. The short tandem repeat (STR) profile of the primed and naïve PSC line was determined by polymerase chain reaction with the CORDIS EXPERT 26 reagent kit (Genoanalytica, <https://www.genoanalytica.ru/>) according to the manufacturer's protocol, followed by amplicon separation

¹ Supplementary Materials 1–4 are available at:
<https://vavilovj-icg.ru/download/pict-2026-30/appx12.pdf>

on a 3130 Genetic Analyzer capillary electrophoresis instrument (Applied Biosystems). Electropherograms of the STR profiles obtained by PCR are available upon request from the authors. Table with STR results is provided in Supplementary Material 3.

Mycoplasma and episome detection. The absence of mycoplasma and episome contamination in the cells was detected by PCR. The primer sequences are provided in the Supplementary Material 2. Parameters for episome detection: 95 °C for 5 minutes, followed by 35 cycles of 95 °C for 15 seconds, 58 °C for 15 seconds and 72 °C for 20 seconds, then 72 °C for 5 minutes. Parameters for mycoplasma detection: 95 °C for 3 minutes, followed by 35 cycles of 95 °C for 15 seconds, 67 °C for 15 seconds and 72 °C for 20 seconds, then 72 °C for 5 minutes.

Karyotyping. Naïve hiPSCs were karyotyped at passages 12, 15 and 18. Metaphase chromosome spreads were obtained by methanol-acetic acid fixation, as previously described (Grigor'eva et al., 2024). The karyotype of naïve iPSCs was determined according to the International System of Human Cytogenetic Nomenclature.

Results

Several media have been proposed to date that allow primed human PSCs to be reset to a naïve state (Chan et al., 2013; Gafni et al., 2013; Takashima et al., 2014; Theunissen et al., 2014; Ware et al., 2014; Duggal et al., 2015; Zimmerlin et al., 2016; Guo et al., 2017; Lee et al., 2017; Szczerbinska et al., 2019; Bayerl et al., 2021; Khan et al., 2021; Buckberry et al., 2023). Among these, the HENSM medium has been shown to capture a naïve state that closely resembles the early pre-implantation epiblast (Bayerl et al., 2021). It has been claimed that naïve PSC lines obtained with the HENSM protocol are karyotypically and epigenetically stable during long-term cultivation. In this study, we aimed to return a primed hiPSC line to a naïve state using HENSM conditions. We used K7-4Lf iPSCs that were obtained and described previously (Malakhova et al., 2020). However, direct transferring K7-4Lf cells from a DMEM/F12 medium supplemented with KnockOut SR and FGF2 into a complete HENSM medium resulted in extensive cell death, preventing further progression of the resetting process. To overcome this barrier, we adopted a stepwise approach (Fig. 1a), first culturing the K7-4Lf cells in a primed cell medium supplemented with an SRC kinase inhibitor (CGP77675) and LIF. This pre-conditioning phase was lasted for three passages before the cells were transferred into a DMEM/F12-based HENSM naïve-induction medium that still omitted N2B27 supplements but contained KnockOut SR and FGF2, as well as the complete HENSM small molecule set and growth factor LIF. By the fourth passage in the naïve-inducing medium, the K7-4Lf colonies had begun to exhibit a transition towards the compact, dome-shaped morphology characteristic of naïve-like human pluripotent stem cells (Fig. 1b). At this stage, the cultures were transferred to complete N2B27-based HENSM medium supplemented with Geltrex. Under these conditions, the cells exhibited stable growth and could be propagated enzymatically using TrypLE in the presence of a ROCK inhibitor.

Following the establishment of the naïve culture, we removed FGF2 from the N2B27-based HENSM medium. While FGF2 supports initial cell survival during resetting, its continued presence can hinder the efficient acquisition of the naïve state. Notably, the resulting K7-4Lf cells maintained hallmark features of naïve pluripotency after FGF2 withdrawal at passage 6 of the resetting process, indicating that they had achieved FGF2-independent pluripotency. Furthermore, these cells were capable of forming compact, spherical colonies under feeder-free conditions when cultured on Matrigel-coated plates (Fig. 1c). However, during the early stages of resetting to the naïve state, we observed that K7-4Lf cells failed to maintain under feeder-free conditions, even in complete HENSM medium. This suggests a transient dependency on FGF2 and/or feeder support during the initial phases of resetting to the naïve state.

The resulting naïve K7-4Lf cells demonstrated robust stability under enzymatic passaging and could be cryopreserved and thawed without any apparent loss of viability or naïve morphology. The naïve K7-4Lf line maintained as mycoplasma free culture (Fig. 1d). It showed no signs of the episomes that had been used to derive the original K7-4Lf, indicating that resetting to a naïve state was free from the expression of exogenous reprogramming factors. We also demonstrated that the original cell line and its naïve derivative have identical STR profiles (see Supplementary Material 3).

Naïve K7-4Lf hiPSCs characterisation

The resulting naïve K7-4Lf cells demonstrated characteristics commonly associated with pluripotent cells. A histochemical assay revealed alkaline phosphatase activity (Fig. 1e), and immunostaining confirmed the presence of the transcription factors OCT4 and NANOG, as well as the conventional surface marker TRA-1-60 (Fig. 1e). Additionally, the cells expressed the naïve-specific markers CD75, KLF17, and TFE3 (Fig. 1f). Notably, TFE3 nuclear translocation is considered a pivotal event in the induction of naïve pluripotency and a defining feature of the naïve pluripotent state (Gafni et al., 2013; Mathieu et al., 2019). Functionally, these cells were able to differentiate into derivatives of all three germ layers (Fig. 1g), as well as into trophoblast cells that expressed surface markers typical of early placentation (Fig. 1h). Furthermore, naïve K7-4Lf cultures exhibited spontaneous self-organisation into blastoid-like structures under low-adhesion conditions (Fig. 1i), indicating that the cells possess developmental potency attributable to the naïve pluripotent state.

Furthermore, we analysed the expression of pluripotency-associated genes using RT-PCR. We focused specifically on how gene expression differs in naïve K7-4Lf cultures with and without FGF2. To monitor transcriptional dynamics over time, we analysed gene expression profiles at passages 6 and 11. Generally, we found that the naïve K7-4Lf line exhibited robust upregulation of naïve-specific transcription factors compared to the original primed cells (Fig. 2).

Previous studies have demonstrated that *NANOG* and *TFCP2L1* play a significant role in the acquisition and maintenance of the naïve pluripotent state (Theunissen et al., 2016; An et al., 2020). We found that *NANOG* expression increased

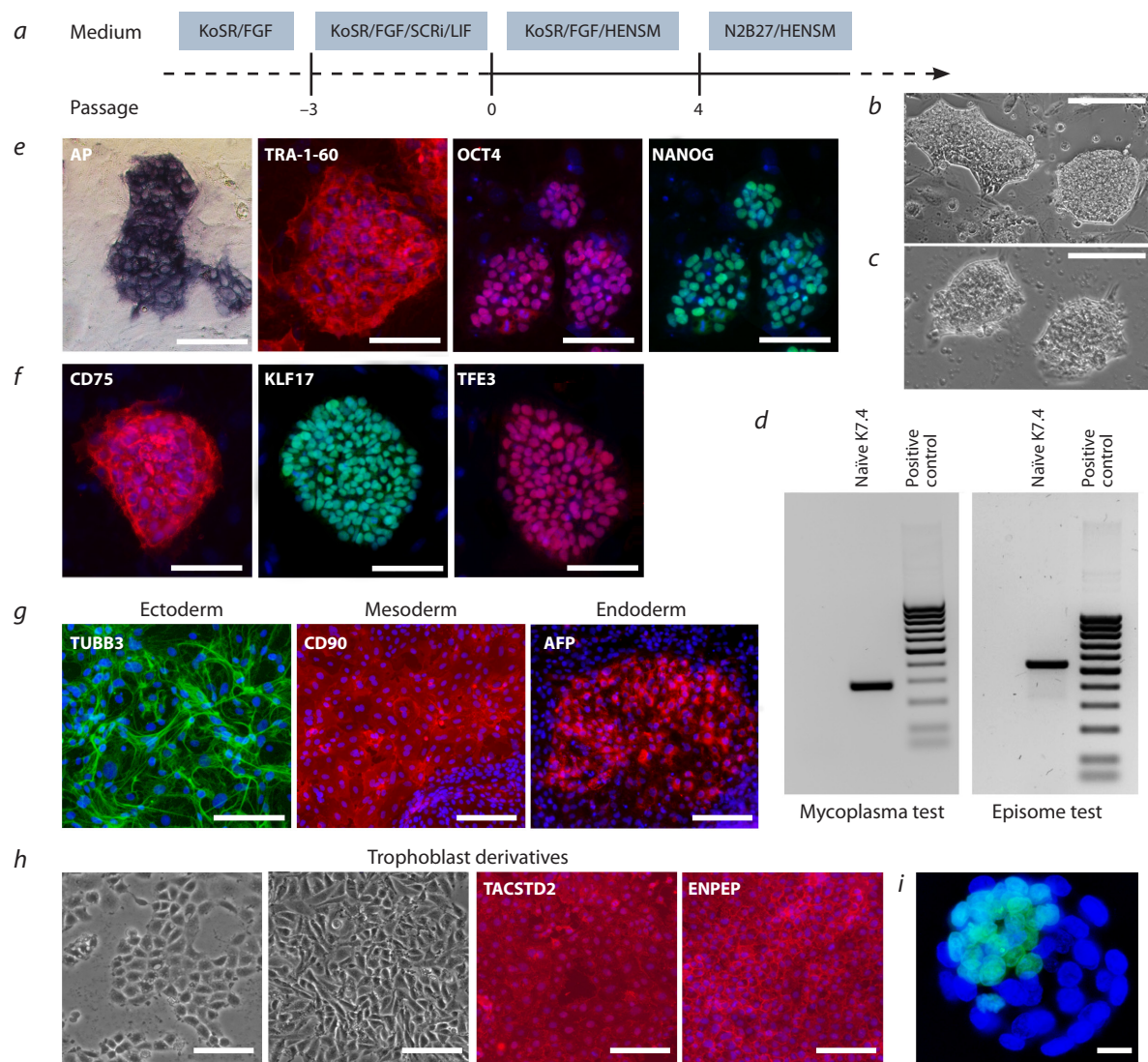


Fig. 1. K7-4Lf hiPSCs after induction of naïve pluripotency in HENSM medium. *a* – schematic representation of the conversion stages from primed to naïve state; *b* – colony morphologies of naïve hiPSCs on a layer of mitotically inactivated mouse embryonic fibroblasts, Scale bar is 100 µm; *c* – colony morphologies of naïve hiPSCs on Matrigel-treated surface; Scale bar is 100 µm; *d* – absence of episomal vectors and mycoplasma contamination in naïve K7-4Lf line; *e* – common pluripotency markers: alkaline phosphatase (AP), TRA-1-60 (red), OCT4 (red), NANOG (green) detected in naïve hiPSCs line. Nuclei are stained with DAPI (blue). Scale bar is 100 µm; *f* – specific markers of naïve pluripotency: surface antigen CD75 (red), transcription factors KLF17 (green) and TFE3 (red). Nuclei are stained with DAPI (blue). Scale bar is 100 µm; *g* – ability of naïve iPSC line to differentiate into derivatives of three germ layers; *h* – morphology of early trophoblast cells derived from naïve hiPSCs and immunostaining with antibodies to early placenta markers TACSTD2 (red) and ENPEP (red). Nuclei are stained with DAPI (blue). Scale bar is 100 µm; *i* – representative image of blastoid obtained by self-aggregation of naïve K7-4Lf iPSCs under low-adhesive conditions. Pluripotent cells are stained green with antibodies to OCT4. Nuclei are stained with DAPI. Scale bar is 10 µm.

substantially in FGF2-independent naïve K7-4Lf cells compared to those maintained with FGF2. Naïve K7-4Lf cultures also demonstrated concomitant *TFCP2L1* upregulation. In line with previous findings that FGF2 impairs the efficient resetting to naïve pluripotency, *TFCP2L1* levels were significantly higher in the FGF2-free condition. *NANOG* expression increased through passage 6, but then declined by passage 11. In contrast, *TFCP2L1* expression remained consistently high across both time points. The expression patterns of *KLF17* and *TFE3* mirrored the dynamic profile of *NANOG*, indicating coordinated regulation during the maturation of naïve identity.

In vivo studies of the early epiblast showed marked activity of transposable elements, particularly *HERVH* and *LTR7Y* (Szczerbinska et al., 2019). Mounting evidence implicates these endogenous retroviral elements as functional regulators of early human development. Consistent with this, we observed robust *LTR7Y* expression in naïve K7-4Lf cultures, with transcription levels progressively increasing over passages. This contrasted with the *NANOG* expression pattern, potentially indicating a transition into a developmental state beyond the *NANOG*-regulated phase of naïve pluripotency.

The expression of a *DAZL* gene, which is also known as a naïve hPSCs marker, was also examined. *DAZL* orchestrates

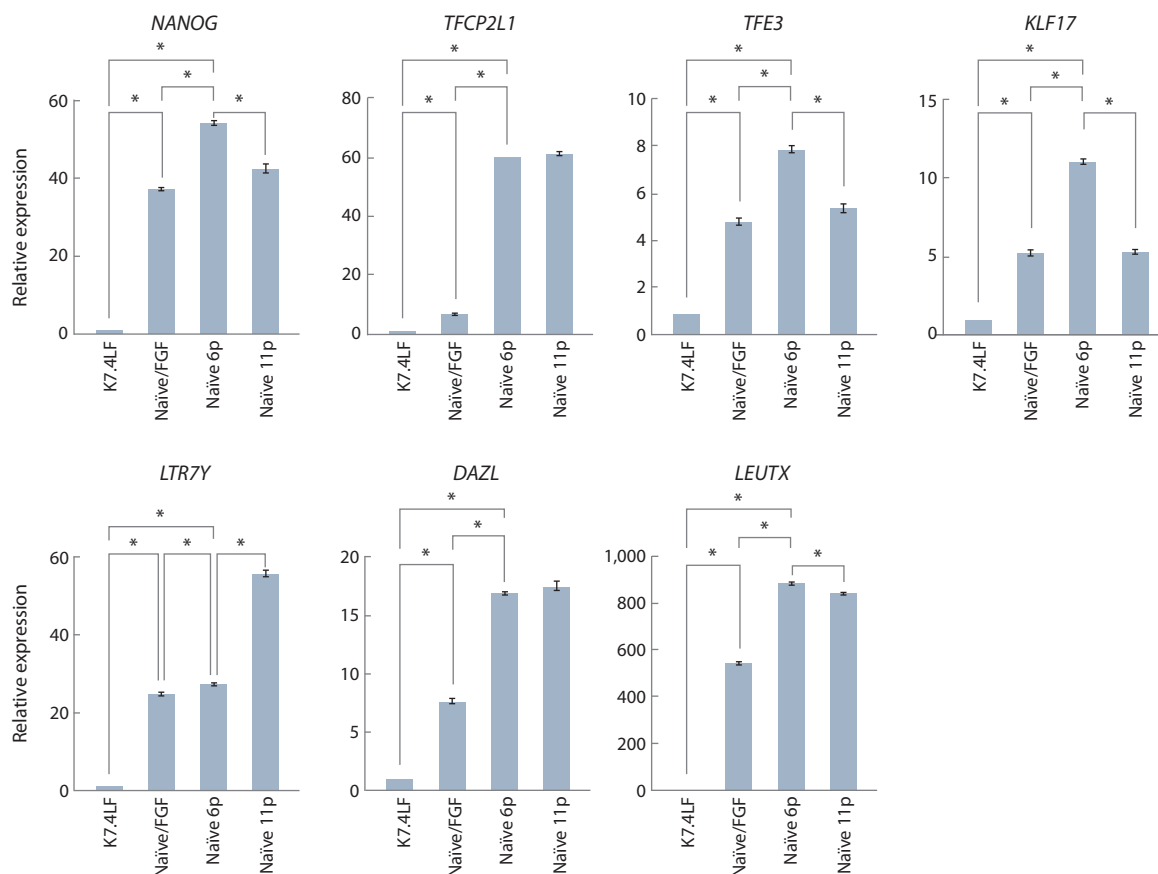


Fig. 2. Semi-quantitative analysis of *NANOG*, *TFCE2L1*, *TFE3*, *KLF17*, *LTR7Y*, *DAZL* and *LEUTX* gene expression in the original primed K7-4Lf line and its naïve derivatives cultured in KoSR/FGF/HENSM (Naïve/FGF) as well as in N2B27/HENSM (Naïve) at 6 and 11 passages.

Statistically significant differences ($p < 0.05$) are marked with an asterisk.

the initial stages of PGC commitment and is involved in the upregulation of *TET1*, which balances the level of 5-hydroxymethylcytosine (Welling et al., 2015). *DAZL* expression was elevated in naïve K7-4Lf hiPSCs and remained consistently higher in FGF2-independent naïve cultures. This finding suggests that the cells have acquired a molecular profile conducive to both germline differentiation competence and stable naïve pluripotency.

Strikingly, the naïve K7-4Lf cells also activated genes associated with a totipotent-like state. Specifically, we observed increased expression of the *LEUTX* gene, which encodes a chromatin protein involved in zygotic genome activation at the eight-cell stage (Mazid et al., 2022), with higher levels detected in FGF2-independent cultures. This is in line with our previous findings and suggests that the removal of FGF2 enables a more advanced reprogramming trajectory towards earlier embryonic-like states (Shevchenko et al., 2025).

In parallel with transcriptional analyses, we monitored chromosomal stability across passages. Karyotype analysis of naïve K7-4Lf cells at passage 12 revealed a normal chromosomal complement (Fig. 3). However, by passage 15, we detected signs of aneuploidy and polyploidy, indicating chromosomal instability during long-term cultivation. The subsequent karyotype analysis at passage 18 revealed a deletion in the short arm

of chromosome 15 (Fig. 3) in more than 70 % of cells. This highlights the importance of ongoing genomic surveillance during long-term culture and suggests that medium composition and culture conditions can affect the karyotype stability in naïve hPSCs obtained and maintained in HENSM.

X chromosome epigenetics in primed, naïve and re-primed K7-4 iPSCs

In primed hPSC lines with two X chromosomes, one is active (X_a) and the other is inactive (X_i) due to X chromosome inactivation (XCI), which is complete by this developmental stage (Vallot et al., 2015; Disteche, 2016; Patel et al., 2017; Sahakyan et al., 2017). However, in most primed hPSC lines, the inactive state on X_i often becomes disrupted and undergoes erosion (X_e). XCI erosion is characterised by the loss of inactive chromatin marks and the partial reactivation of genes, preventing X_e from being re-inactivated during differentiation. This erosion is undesirable for hPSC application in biomedical research. Some naïve media, including HENSM, enable pluripotent cells to restore an inactive state on X_e when the cells are returned to a primed state and differentiated into germinal and somatic cells (Sahakyan et al., 2017; Vallot et al., 2017; An et al., 2020; Raposo et al., 2025). In this study, we evaluated the epigenetic state of the X chromosome in the

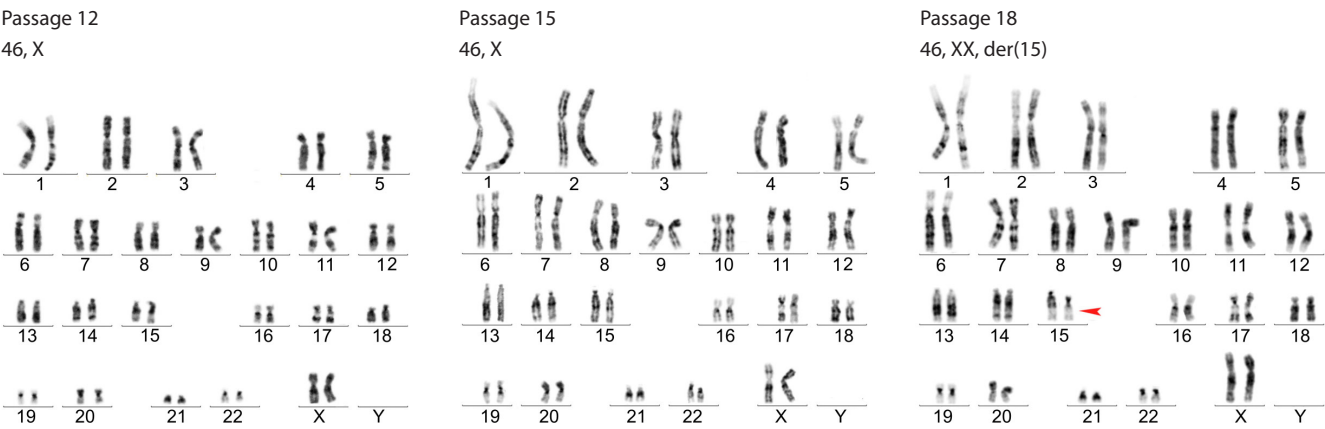


Fig. 3. Naïve hiPSCs K7-4Lf karyotyping at passages 12, 15 and 18.

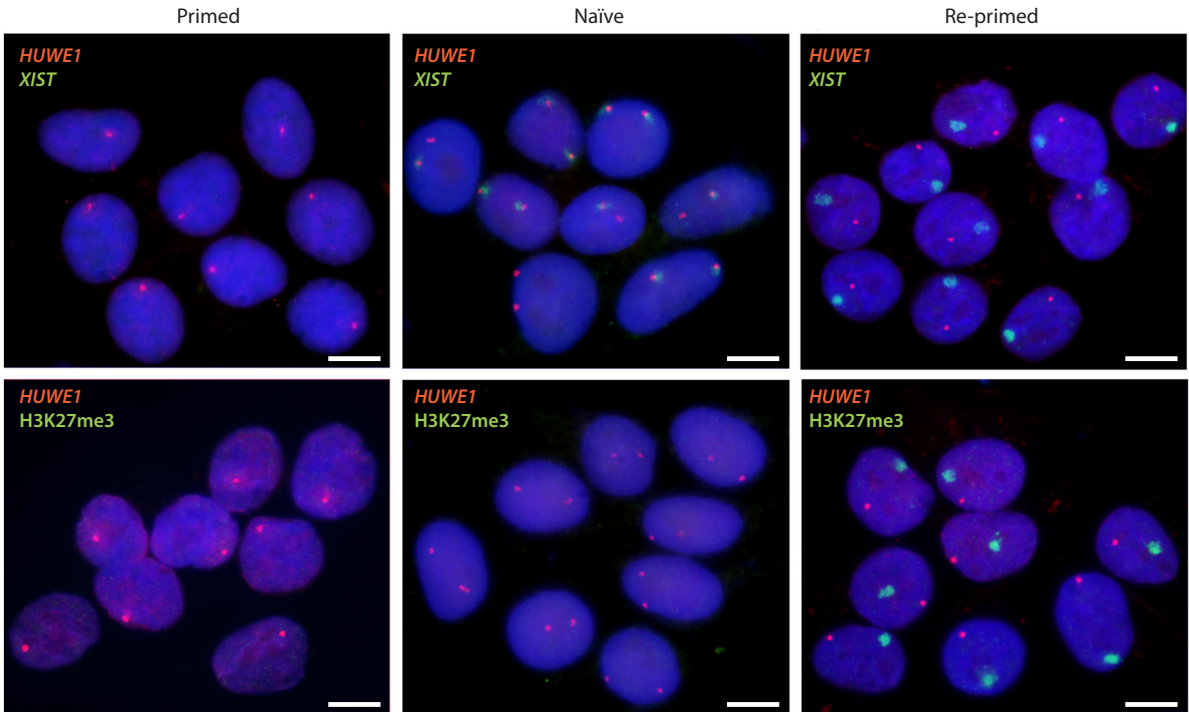


Fig. 4. X chromosome epigenetic state in the original primed, naïve and re-primed K7-4Lf hiPSCs, detected by RNA FISH with probes to RNA *XIST* (green) and *HUWE1* gene (red) as well as immunostaining to H3K27me3 histone mark (green). Nuclei are counterstained with DAPI (blue). Scale bar is 10 μ m.

original primed hiPSC line K7-4Lf, focusing on the changes that occur when the cells are reset to the naïve state and when the naïve cells are returned to the primed state (re-primed). Using RNA FISH and immunofluorescence, we showed that in the original primed K7-4Lf hiPSC line, 96.0 ± 1.2 % of nuclei lacked *XIST* RNA and H3K27me3 histone modification clouds, which are attributable to the Xi territory and are main XCI participants (Fig. 4). Thus, it can be stated that erosion on the inactive X chromosome is observed in the original K7-4Lf line. However, in primed K7-4Lf hiPSCs, the majority of cells (96.5 ± 2.1 %) exhibited just one *HUWE1* signal per nucleus, suggesting that it is transcribed on Xa and remains inactive on Xe. Reactivation of *HUWE1* on Xe is considered a late ero-

sion event, suggesting that XCI erosion in the original K7-4Lf line is at an early stage (Vallot et al., 2015; Patel et al., 2017; Raposo et al., 2025). Naïve K7-4Lf cells cultured in HENSM medium show biallelic expression of the X-linked gene *HUWE1* in 95.0 ± 1.6 % of nuclei. Biallelic expression of *XIST* RNA is also detected in some cells (27.5 ± 2.1 %). Accumulation of the inactive H3K27me3 modification on X chromosomes was not detected. This epigenetic state is typical of naïve hPSCs and reflects the intermediate stages of the X-chromosome reactivation that is characteristic of the early pluripotent state in humans (Vallot et al., 2015; Theunissen et al., 2016; Sahakyan et al., 2017; Bayerl et al., 2021; Khan et al., 2021).

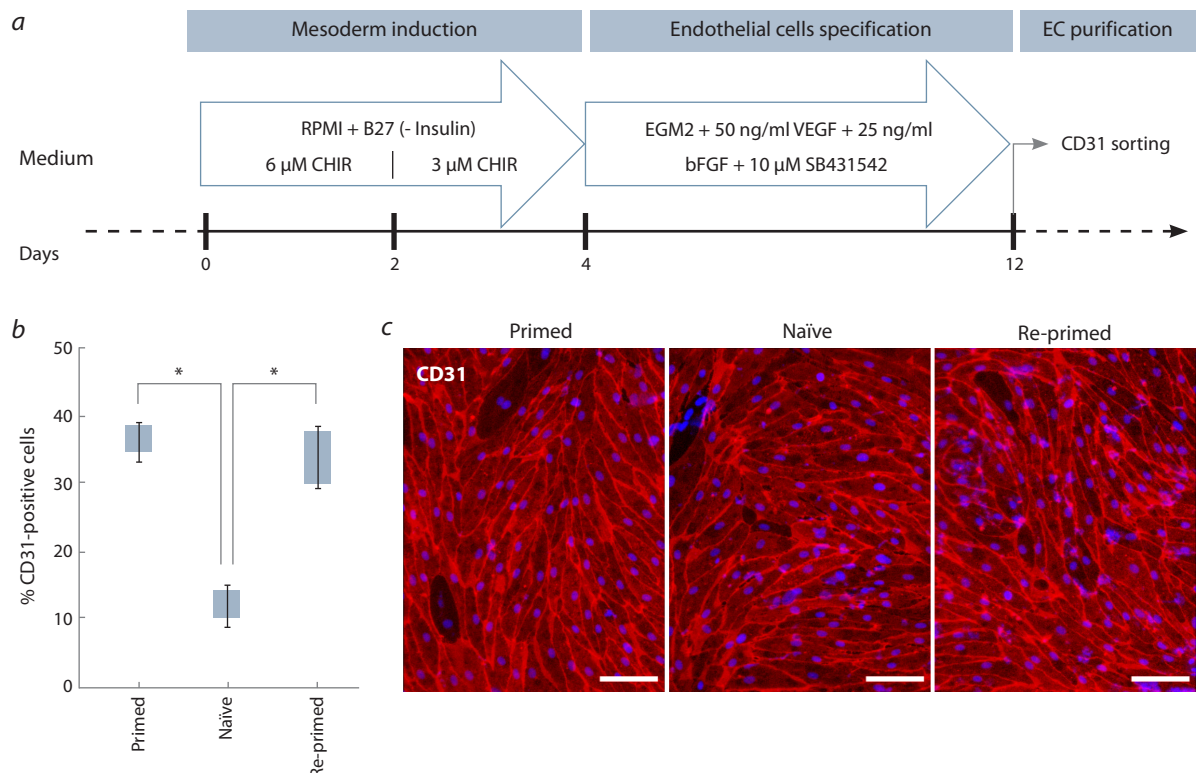


Fig. 5. Directed differentiation of original primed, naïve and re-primed K7-4Lf hiPSCs into endothelial derivatives.

a – schematic representation of the differentiation protocol; *b* – quantification of mature EC using CD31 (PECAM1) surface marker. Statistically significant differences ($p < 0.05$) are marked with an asterisk; *c* – hiPSC derived endothelial cells are stained with antibody to CD31 (red). Nuclei are counterstained with DAPI (blue). Scale bar is 100 μ m.

In K7-4Lf cells derived from naïve hiPSCs after re-priming, we observed pattern of completed XCI (95.5 ± 1.3 % of nuclei). It appears as one distinct RNA *XIST* or H3K27me3 cloud corresponding to Xi and as one signal of X-linked gene *HUWE1* located apart from, corresponding to Xa (Fig. 4).

Thus, obtaining cells in naïve state from primed K7-4Lf hiPSC line in HENSM medium, and re-priming or differentiating them, allows us to get rid of erosion and obtain a culture with normal XCI.

Directed differentiation of naïve hiPSCs into endothelial derivatives

Although naïve pluripotent cells have an expanded differentiation potential, they may not differentiate efficiently into somatic lineages directly (Lee et al., 2017; Rostovskaya et al., 2019; Guo et al., 2021; Buckberry et al., 2023). To achieve efficient differentiation into specific somatic cell types, naïve hPSCs first need to be returned back to the primed state, after which their somatic derivatives can successfully be obtained according to protocols previously established for primed hPSCs.

In this study, we evaluated the effect of naïve pluripotency on the ability of hiPSCs to differentiate into endothelial-like cells (ECs), which had not been investigated previously. We compared the ability of the original primed K7-4Lf line, naïve and re-primed K7-4Lf cells to produce mature endothelial derivatives. Re-priming entails placing naïve K7-4Lf cultures

in the medium used for maintaining primed K7-4Lf hiPSCs for 48 hours prior to directed EC differentiation.

We differentiated primed, naïve and re-primed K7-4Lf cells into ECs using a two-step directed protocol (James et al., 2010; Gu, 2018; Zakharova et al., 2024) (Fig. 5a). It comprises (1) mesoderm induction by WNT pathway activation and subsequent (2) endothelial specification using VEGF and FGF in EGM2 medium along with TGF β pathway inhibition to suppress smooth muscle cell lineage. On day eight of endothelial specification, we used flow cytometry to quantify mature ECs for the endothelial marker CD31 (PECAM1) (Fig. 5b).

We found that attempts to directly differentiate K7-4Lf cells maintained in a naïve state resulted in poor induction of endothelial derivatives. This suggests that naïve hiPSCs have an impaired or incompatible background for endothelial specification. However, re-primed and primed K7-4Lf cells exhibited a significant proportion of mature endothelial cells, which did not differ significantly from each other and were comparable with endothelial specification experiments performed in other studies (Shevchenko et al., 2023; Zakharova et al., 2024). Thus, we found that re-priming renders the naïve K7-4Lf line competent to differentiate into endothelial cells at a level comparable to the original primed K7-4Lf cells (Fig. 5b).

In conclusion, endothelial cells obtained from all three K7-4Lf sources through this differentiation protocol by magnetic sorting were positive for CD31 (Fig. 5c), which is consistent with their identity as vascular endothelial cells.

Discussion

The rigorous selection of HENSM medium composition, taking into account the influence of its components on chromosomal, epigenetic and genetic stability, allowed optimal conditions for long-term cultivation of naïve hPSCs to be identified. This potentially paves the way for the mass production of stable naïve hPSC lines and their broader use in biomedical research.

In this study, we successfully reset the primed K7-4Lf hiPSC line to the naïve state using the HENSM medium. The characteristics of the resulting naïve iPSCs were summarised in the preliminary cell line passport (Supplementary Material 4). However, transferring primed hiPSCs directly into HENSM medium resulted in extensive cell loss, hindering efficient reprogramming and the establishment of stable naïve cultures. To address this issue, we optimised the induction protocol by introducing a transitional phase, culturing the hPSCs in SRC kinase inhibitor and LIF-supplemented medium prior to naïve induction. This intermediate step was crucial, as it conferred the necessary competence for naïve induction, reducing to minimum cell death and improving the efficiency and stability of the resetting process. The effectiveness of the transitional stage likely stems from its ability to reset the cellular signalling and epigenetic landscape, thereby priming the cells for subsequent naïve resetting. By modulating key pathways and reducing lineage priming, the transitional culture creates a more permissive environment for acquiring naïve pluripotency. Our optimised protocol, which included enzymatic passaging with TryPLE, consistently generated stable naïve hPSCs. These cells exhibited robust expression of pluripotency markers and significant upregulation of naïve-specific genes. Notably, the naïve hPSCs produced using our protocol exhibited the unique ability to form trophoblast cells and blastoid-like structures *in vitro*, highlighting their developmental potential and fidelity to the naïve state.

Our naïve K7-4Lf cells showed upregulation of *LEUTX*, which is a key marker of totipotent-like subpopulations that are active during early embryonic development. The increase in *LEUTX* expression following the withdrawal of FGF2 is consistent with our previous study (Shevchenko et al., 2025), which found that removing growth factors such as FGF2 increases the number of *LEUTX*-positive cells. This indicates a shift towards a totipotent-like transcriptional state within naïve hPSC cultures.

Naïve hPSCs demonstrate superior competence in generating neural stem cells, progenitors and primordial germ cell-like cells, compared to their primed counterparts (Irie, Surani, 2016; Kisa et al., 2017; Lee et al., 2017; Ozaki et al., 2022; Buckberry et al., 2023). This enhanced differentiation potential positions naïve hPSCs as a promising platform for various lineage specifications. However, naïve hPSCs appear to be limited in their ability to generate endothelial cells directly *in vitro*. This phenomenon may be explained by the absence of key transcription factors, such as *OTX2* and *ZIC2/3*, which are essential for initiating gastrulation and subsequent somatic lineage specification. These genes are typically expressed in the post-implantation epiblast, a developmental stage that is mimicked by primed hPSCs (Warr et al., 2008; Di Giovan-

antonio et al., 2021; Ee et al., 2024; Hossain et al., 2024), but are absent in the naïve state. Consistent with this, we found that re-priming naïve hPSCs restores their ability to differentiate into endothelial cells. Therefore, the ability of naïve hPSCs to generate mesodermal and endodermal derivatives appears restricted unless the cells undergo a transient re-priming process (Guo et al., 2017; Lee et al., 2017; Liu et al., 2017; Buckberry et al., 2023).

The loss of nuclear *XIST* RNA and H3K27me3 foci in the original K7-4Lf line indicates erosion of the inactive X chromosome. Despite the fact that it occurs frequently, the erosion is often disregarded in stem cell research. XCI erosion leads to the aberrant expression of X-linked genes, which impairs the differentiation potential of hPSCs. This results in abnormal descendants that fail to accurately model physiological conditions (Anguera et al., 2012; Mekhoubad et al., 2012; D'Antonio-Chronowska et al., 2019; Motosugi et al., 2022). Importantly, XCI erosion persists after differentiation, compromising the reliability of disease models and drug screening platforms (Mekhoubad et al., 2012; Vallot et al., 2015; Patel et al., 2017). Differentiated derivatives with erosion may be non-functional and at risk of oncogenic transformation. Therefore, monitoring and preventing XCI erosion is essential for ensuring the fidelity of hPSC-derived models and clinical applications. Resetting to the naïve state enables reactivation of the X chromosome, but this is neither complete nor homogeneous, either between cell lines or between cells within a culture. Contrary to previous findings, a recent study has demonstrated that human pre-implantation epiblasts display two active X chromosomes marked by H3K27me3 (Alfeghaly et al., 2024). These observations suggest that generated naïve hPSCs exist in a metastable, transitional state that does not fully recapitulate the pre-implantation epiblast *in vivo* (Alfeghaly et al., 2024). Nevertheless, re-primed hiPSCs regained *XIST* RNA and H3K27me3 foci, indicating re-establishment of the normal inactive X chromosome state.

Chromosomal instability is a well-recognised challenge in naïve hPSC cultures. Previous reports indicate that naïve hPSCs typically retain a normal karyotype for up to nine passages when maintained in 5iLAF medium before abnormalities arise (Theunissen et al., 2014). In our study, naïve hiPSCs generated using our optimised protocol retained normal karyotypes for up to 15 passages, exceeding earlier benchmarks. However, chromosomal abnormalities were detected by passage 18. These results imply that, although our approach increases the period of genomic stability, the risk of karyotypic aberrations continues with prolonged culture. Ongoing monitoring and further refinement of culture conditions are necessary to ensure the genomic integrity of naïve hPSCs for research and therapeutic applications.

Conclusion

In summary, we successfully reset the primed K7-4Lf hiPSC line to a naïve pluripotent state, using these cells to explore the developmental and functional properties of human naïve pluripotency. Our findings demonstrate the potential of naïve reprogramming in accessing early developmental stages and modelling key aspects of embryogenesis by blastoid formation.

We identified the expression of totipotent-like genes within the naïve hPSC cultures, which reinforces the developmental plasticity of this state. In parallel, we demonstrated that re-priming can effectively recapitulate somatic differentiation competence. This enables the efficient generation of endothelial-like cells, thereby underscoring the developmental relevance of transitioning between pluripotent states. Naïve pluripotency opens up new opportunities to control and understand X chromosome reactivation and inactivation epigenetics in humans. However, this topic remains challenging and requires further study. Moving forward, the field must strive to develop more refined, stable and clinically translatable naïve hPSC systems that preserve the benefits of this state while mitigating its disadvantages. As progress continues, naïve hPSCs will remain pivotal in exploring human development and pioneering next-generation regenerative therapies.

References

- Alfeghaly C., Castel G., Cazottes E., Moscatelli M., Moinard E., Casanova M., Boni J., ... Boers J., Gribnau J., David L., Ouimette J.F., Rougeulle C. XIST dampens X chromosome activity in a SPEN-dependent manner during early human development. *Nat Struct Mol Biol.* 2024;31(10):1589-1600. doi 10.1038/s41594-024-01325-3
- An C., Feng G., Zhang J., Cao S., Wang Y., Wang N., Lu F., Zhou Q., Wang H. Overcoming autocrine FGF signaling-induced heterogeneity in naïve human ESCs enables modeling of random X chromosome inactivation. *Cell Stem Cell.* 2020;27(3):482-497. doi 10.1016/j.stem.2020.06.002
- Anguera M., Sadreyev R., Zhang Z., Szanto A., Payer B., Sheridan S., Kwok S., ... Alvarez J., Gimelbrant A., Mitalipova M., Kirby J.E., Lee J.T. Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes. *Cell Stem Cell.* 2012;11(1):75-90. doi 10.1016/j.stem.2012.03.008
- Bayerl J., Ayyash M., Shani T., Manor Y., Gafni O., Massarwa R., Kalma Y., ... Hanna S., Ben-Yosef D., Novershtern N., Viukov S., Hanna J.H. Principles of signaling pathway modulation for enhancing human naïve pluripotency induction. *Cell Stem Cell.* 2021;28(9):1549-1565.e12. doi 10.1016/j.stem.2021.04.001
- Buckberry S., Liu X., Poppe D., Tan J.P., Sun G., Chen J., Nguyen T.V., ... Breen J., Faulkner G.J., Nefzger C.M., Polo J.M., Lister R. Transient naïve reprogramming corrects hiPS cells functionally and epigenetically. *Nature.* 2023;620(7975):863-872. doi 10.1038/s41586-023-06424-7
- Chan Y.S., Göke J., Ng J.H., Lu X., Gonzales K.A., Tan C.P., Tng W.Q., Hong Z.Z., Lim Y.S., Ng H.H. Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell.* 2013;13(6):663-675. doi 10.1016/j.stem.2013.11.015
- Collier A., Rugg-Gunn P. Identifying human naïve pluripotent stem cells – evaluating state-specific reporter lines and cell-surface markers. *BioEssays.* 2018;40(5):1700239. doi 10.1002/bies.201700239
- D'Antonio-Chronowska A., Donovan M.K.R., Young Greenwald W.W., Nguyen J.P., Fujita K., Hashem S., Matsui H., ... Coulet F., Smith E.N., Adler E., D'Antonio M., Frazer K.A. Association of human iPSC gene signatures and X chromosome dosage with two distinct cardiac differentiation trajectories. *Stem Cell Rep.* 2019;13(5):924-938. doi 10.1016/j.stemcr.2019.09.011
- Dekel C., Morey R., Hanna J., Laurent L., Ben-Yosef D., Amir H. Stabilization of hESCs in two distinct substates along the continuum of pluripotency. *iScience.* 2022;25(12):105469. doi 10.1016/j.isci.2022.105469
- Di Giovannantonio L., Acampora D., Omodei D., Nigro V., Barba P., Barbieri E., Chambers I., Simeone A. Direct repression of *Nanog* and *Otx4* by OTX2 modulates the contribution of epiblast-derived cells to germline and somatic lineage. *Development.* 2021;148(10):dev199166. doi 10.1242/dev.199166
- Disteche C.M. Dosage compensation of the sex chromosomes and autosomes. *Semin Cell Dev Biol.* 2016;56:9-18. doi 10.1016/j.semcdb.2016.04.013
- Duggal G., Warrier S., Ghimire S., Broekaert D., Van der Jeught M., Lierman S., Deroo T., ... Heijmans B.T., Deforce D., De Sutter P., De Sousa Lopes S.C., Heindryckx B. Alternative routes to induce naïve pluripotency in human embryonic stem cells. *Stem Cells.* 2015;33(9):2686-2698. doi 10.1002/stem.2071
- Ee L.S., Medina-Cano D., Uyehara C.M., Schwarz C., Goetzler E., Salataj E., Polyzos A., Madhuranath S., Evans T., Hadjantonakis A.K., Apostolou E., Vierbuchen T., Stadtfeld M. Transcriptional remodeling by OTX2 directs specification and patterning of mammalian definitive endoderm. *bioRxiv.* 2024. doi 10.1101/2024.05.30.596630
- Fischer L.A., Meyer B., Reyes M., Zemke J.E., Harrison J.K., Park K.M., Wang T., Jüppner H., Dietmann S., Theunissen T.W. Tracking and mitigating imprint erasure during induction of naïve human pluripotency at single-cell resolution. *Stem Cell Rep.* 2025;20(3):102419. doi 10.1016/j.stemcr.2025.102419
- Gafni O., Weinberger L., Mansour A.A., Manor Y.S., Chomsky E., Ben-Yosef D., Kalma Y., ... Amann-Zalcenstein D., Benjamin S., Amit I., Tanay A., Massarwa R., Novershtern N., Hanna J.H. Derivation of novel human ground state naïve pluripotent stem cells. *Nature.* 2013;504(7479):282-286. doi 10.1038/nature12745
- Grigor'eva E.V., Kopytova A.E., Yarkova E.S., Pavlova S.V., Sorogina D.A., Malakhova A.A., Malankhanova T.B., Baydakova G.V., Zakharova E.Y., Medvedev S.P., Pchelina S.N., Zakian S.M. Biochemical characteristics of iPSC-derived dopaminergic neurons from N370S *GBA* variant carriers with and without Parkinson's disease. *Int J Mol Sci.* 2023;24(5):4437. doi 10.3390/ijms24054437
- Grigor'eva E.V., Karapetyan L.V., Malakhova A.A., Medvedev S.P., Minina J.M., Hayrapetyan V.H., Vardanyan V.S., Zakian S.M., Arakelyan A., Zakharyan R. Generation of iPSCs from a patient with the M694V mutation in the *MEFV* gene associated with Familial Mediterranean fever and their differentiation into macrophages. *Int J Mol Sci.* 2024;25(11):6102. doi 10.3390/ijms25116102
- Gu M. Efficient differentiation of human pluripotent stem cells to endothelial cells. *Curr Protoc Hum Genet.* 2018;98(1):e64. doi 10.1002/cphg.64
- Guo G., von Meyenn F., Rostovskaya M., Clarke J., Dietmann S., Baker D., Sahakyan A., Myers S., Bertone P., Reik W., Plath K., Smith A. Epigenetic resetting of human pluripotency. *Development.* 2017;144(15):2748-2763. doi 10.1242/dev.146811
- Guo G., Stirparo G.G., Strawbridge S.E., Spindlow D., Yang J., Clarke J., Dattani A., Yanagida A., Li M.A., Myers S., Özel B.N., Nichols J., Smith A. Human naïve epiblast cells possess unrestricted lineage potential. *Cell Stem Cell.* 2021;28(6):1040-1056. doi 10.1016/j.stem.2021.02.025
- Hossain I., Priam P., Reynoso S.C., Sahni S., Zhang X.X., Côté L., Doumat J., Chik C., Fu T., Lessard J.A., Pastor W.A. ZIC2 and ZIC3 promote SWI/SNF recruitment to safeguard progression towards human primed pluripotency. *Nat Commun.* 2024;15(1):8539. doi 10.1038/s41467-024-52431-1
- Hu Z., Li H., Jiang H., Ren Y., Yu X., Qiu J., Stablewski A., Zhang B., Buck M., Feng J. Transient inhibition of mTOR in human pluripotent stem cells enables robust formation of mouse-human chimeric embryos. *Sci Adv.* 2020;6(20):eaaz0298. doi 10.1126/sciadv.aaz0298
- Irie N., Surani M.A. Efficient induction and isolation of human primordial germ cell-like cells from competent human pluripotent stem cells. In: Buszczak M. (Ed.) *Germline Stem Cells. Methods in Molecular Biology.* Vol. 1463. Humana Press, New York, 2016;217-226. doi 10.1007/978-1-4939-4017-2_16
- James D., Nam H.S., Seandel M., Nolan D., Janovitz T., Tomishima M., Studer L., Lee G., Lyden D., Benezra R., Zaninovic N., Rosenwaks Z., Rabbany S.Y., Rafii S. Expansion and maintenance of human embryonic stem cell-derived endothelial cells by TGFβ inhibition

- bition is Id1 dependent. *Nat Biotechnol.* 2010;28(2):161-166. doi 10.1038/nbt.1605
- Khan S.A., Park K.M., Fischer L.A., Dong C., Lungjangwa T., Jimenez M., Casalela D., Chew B., Dietmann S., Auld D.S., Jaenisch R., Theunissen T.W. Probing the signaling requirements for naïve human pluripotency by high-throughput chemical screening. *Cell Rep.* 2021;35(11):109233. doi 10.1016/j.celrep.2021.109233
- Kisa F., Shiozawa S., Oda K., Yoshimatsu S., Nakamura M., Koya I., Kawai K., Suzuki S., Okano H. Naïve-like ESRRB⁺ iPSCs with the capacity for rapid neural differentiation. *Stem Cell Rep.* 2017;9(6):1825-1838. doi 10.1016/j.stemcr.2017.10.008
- Klepikova A., Nenasheva T., Sheveleva O., Protasova E., Antonov D., Gainullina A., Chikina E., Sakovnich O., Gerasimova T., Nikitina I., Shevalie D., Lyadova I. iPSC-derived macrophages: the differentiation protocol affects cell immune characteristics and differentiation trajectories. *Int J Mol Sci.* 2022;23(24):16087. doi 10.3390/ijms232416087
- Kopytova A.E., Rychkov G.N., Cheblokov A.A., Grigor'eva E.V., Nikolaev M.A., Yarkova E.S., Sorogina D.A., ... Bezrukh V.A., Salogub G.N., Zakharova E.Y., Pchelina S.N., Emelyanov A.K. Potential binding sites of pharmacological chaperone NCGC00241607 on mutant β -glucocerebrosidase and its efficacy on patient-derived cell cultures in Gaucher and Parkinson's disease. *Int J Mol Sci.* 2023;24(10):9105. doi 10.3390/ijms24109105
- Lee J.H., Laronde S., Collins T.J., Shapovalova Z., Tanasijevic B., McNicol J.D., Fiebig-Comyn A., Benoit Y.D., Lee J.B., Mitchell R.R., Bhatia M. Lineage-specific differentiation is influenced by state of human pluripotency. *Cell Rep.* 2017;19(1):20-35. doi 10.1016/j.celrep.2017.03.036
- Liu X., Nefzger C.M., Rossello F.J., Chen J., Knaupp A.S., Firas J., Ford E., ... Nilsson S.K., Schittenhelm R.B., Laslett A.L., Lister R., Polo J.M. Comprehensive characterization of distinct states of human naïve pluripotency generated by reprogramming. *Nat Methods.* 2017;14(11):1055-1062. doi 10.1038/nmeth.4436
- Malakhova A.A., Grigor'eva E.V., Pavlova S.V., Malankhanova T.B., Valetdinova K.R., Vyatkin Y.V., Khabarova E.A., Rzaev J.A., Zakian S.M., Medvedev S.P. Generation of induced pluripotent stem cell lines ICGi021-A and ICGi022-A from peripheral blood mononuclear cells of two healthy individuals from Siberian population. *Stem Cell Res.* 2020;48:101952. doi 10.1016/j.scr.2020.101952
- Mathieu J., Detraux D., Kupperts D., Wang Y., Cavanaugh C., Sidhu S., Levy S., ... Hawkins R.D., Moon R.T., Ware C.B., Paddison P.J., Ruohola-Baker H. Folliculin regulates mTORC1/2 and WNT pathways in early human pluripotency. *Nat Commun.* 2019;10(1):632. doi 10.1038/s41467-018-08020-0
- Mazid M.A., Ward C., Luo Z., Liu C., Li Y., Lai Y., Wu L., ... Maxwell P.H., Xu X., Liu L., Li W., Esteban M.A. Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature.* 2022;605(7909):315-324. doi 10.1038/s41586-022-04625-0
- Mekhoubad S., Bock C., De Boer A., Kiskinis E., Meissner A., Eggan K. Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell.* 2012;10(5):595-609. doi 10.1016/j.stem.2012.02.014
- Motosugi N., Sugiyama A., Okada C., Otomo A., Umezawa A., Akutsu H., Hadano S., Fukuda A. De-erosion of X chromosome dosage compensation by the editing of *XIST* regulatory regions restores the differentiation potential in hPSCs. *Cell Rep Methods.* 2022;2(12):100352. doi 10.1016/j.crmeth.2022.100352
- Nadtochy J.A., Medvedev S.P., Grigor'eva E.V., Pavlova S.V., Minina J.M., Chechushkov A.V., Malakhova A.A., Kovalenko L.V., Zakian S.M. Transgenic iPSC lines with genetically encoded MitoTimer to study mitochondrial biogenesis in dopaminergic neurons with tauopathy. *Biomedicines.* 2025;13(3):550. doi 10.3390/biomedicines13030550
- Okae H., Toh H., Sato T., Hiura H., Takahashi S., Shirane K., Kabayama Y., Suyama M., Sasaki H., Arima T. Derivation of human trophoblast stem cells. *Cell Stem Cell.* 2018;22(1):50-63. doi 10.1016/j.stem.2017.11.004
- Ozaki H., Suga H., Sakakibara M., Soen M., Miyake N., Miwata T., Taga S., ... Iguchi G., Takahashi Y., Muguruma K., Inoue H., Arima H. Differentiation of human induced pluripotent stem cells into hypothalamic vasopressin neurons with minimal exogenous signals and partial conversion to the naïve state. *Sci Rep.* 2022;12(1):17381. doi 10.1038/s41598-022-22405-8
- Patel S., Bonora G., Sahakyan A., Kim R., Chronis C., Langerman J., Fitz-Gibbon S., ... Ardehali R., Pellegrini M., Lowry W.E., Clark A.T., Plath K. Human embryonic stem cells do not change their X inactivation status during differentiation. *Cell Rep.* 2017;18(1):54-67. doi 10.1016/j.celrep.2016.11.054
- Pavlova S.V., Shulgina A.E., Minina J.M., Zakian S.M., Dementyeva E.V. Generation of isogenic iPSC lines for studying the effect of the p.N515del (c.1543_1545delAAC) variant on MYBPC3 function and hypertrophic cardiomyopathy pathogenesis. *Int J Mol Sci.* 2024a;25(23):12900. doi 10.3390/ijms252312900
- Pavlova S.V., Shulgina A.E., Zakian S.M., Dementyeva E.V. Studying pathogenetic contribution of a variant of unknown significance, p.M659I (c.1977G>A) in MYH7, to the development of hypertrophic cardiomyopathy using CRISPR/Cas9-engineered isogenic induced pluripotent stem cells. *Int J Mol Sci.* 2024b;25(16):8695. doi 10.3390/ijms25168695
- Pham T.X.A., Panda A., Kagawa H., To S.K., Ertekin C., Georgopoulos G., van Knippenberg S.S.F.A., ... Lluís F., David L., Rivron N., Balaton B.P., Pasque V. Modeling human extraembryonic mesoderm cells using naïve pluripotent stem cells. *Cell Stem Cell.* 2022;29(9):1346-1365.e10. doi 10.1016/j.stem.2022.08.001
- Raposo A.C., Caldas P., Jeremias J., Arez M., Cazaux Mateus F., Barbosa P., Sousa-Luis R., ... Mupo A., Eckersley-Maslin M., Casanova M., Grosso A.R., da Rocha S.T. Gene reactivation upon erosion of X chromosome inactivation in female hiPSCs is predictable yet variable and persists through differentiation. *Stem Cell Rep.* 2025;20(5):102472. doi 10.1016/j.stemcr.2025.102472
- Rezvoova M.A., Ovcharenko E.A., Klyshnikov K.Y., Glushkova T.V., Kostyunin A.E., Shishkova D.K., Matveeva V.G., Velikanova E.A., Shabaev A.R., Kudryavtseva Y.A. Electrospun bioresorbable polymer membranes for coronary artery stents. *Front Bioeng Biotechnol.* 2024;12:1440181. doi 10.3389/fbioe.2024.1440181
- Rostovskaya M. Capacitation of human naïve pluripotent stem cells. In: Rugg-Gunn P. (Ed.) Human Naïve Pluripotent Stem Cells. Methods in Molecular Biology. Vol. 2416. Humana, New York, 2022;117-131. doi 10.1007/978-1-0716-1908-7_9
- Rostovskaya M., Stirparo G., Smith A. Capacitation of human naïve pluripotent stem cells for multi-lineage differentiation. *Development.* 2019;146(7):dev172916. doi 10.1242/dev.172916
- Sahakyan A., Kim R., Chronis C., Sabri S., Bonora G., Theunissen T., Kuoy E., Langerman J., Clark A., Jaenisch R., Plath K. Human naïve pluripotent stem cells model X chromosome dampening and X inactivation. *Cell Stem Cell.* 2017;20(1):87-101. doi 10.1016/j.stem.2016.10.006
- Shevchenko A.I., Arssan A.M., Zakian S.M., Zakharova I.S. Chemokine CCL2 activates hypoxia response factors regulating pluripotency and directed endothelial differentiation of human pluripotent stem cells. *Russ J Dev Biol.* 2023;54(2):134-146. doi 10.1134/s1062360423020054
- Shevchenko A.I., Arssan A.M., Zakharova I.S. Towards the generation of safe naïve human pluripotent cell lines. *Tomsk State Univ J Biol.* 2025;69:184-193. doi 10.17223/19988591/69/21 (in Russian)
- Sheveleva O., Protasova E., Nenasheva T., Butorina N., Melnikova V., Gerasimova T., Sakovnich O., Kurinov A., Grigor'eva E., Medvedev S., Lyadova I. A model of iPSC-derived macrophages with *TNFAIP3* overexpression reveals the peculiarities of TNFAIP3 protein expression and function in human macrophages. *Int J Mol Sci.* 2023;24(16):12868. doi 10.3390/ijms241612868
- Sheveleva O., Protasova E., Grigor'eva E., Butorina N., Kuziaeva V., Antonov D., Melnikova V., Medvedev S., Lyadova I. The generation of genetically engineered human induced pluripotent stem cells overexpressing IFN- β for future experimental and clinically

- oriented studies. *Int J Mol Sci.* 2024;25(22):12456. doi 10.3390/ijms252212456
- Stadtfield M., Hochedlinger K. Induced pluripotency: history, mechanisms, and applications. *Genes Dev.* 2010;24(20):2239-2263. doi 10.1101/gad.1963910
- Szczerbinska I., Gonzales K.A.U., Cukuroglu E., Ramli M.N.B., Lee B.P.G., Tan C.P., Wong C.K., Rancati G.I., Liang H., Göke J., Ng H.H., Chan Y.S. A chemically defined feeder-free system for the establishment and maintenance of the human naïve pluripotent state. *Stem Cell Rep.* 2019;13(4):612-626. doi 10.1016/j.stemcr.2019.08.005
- Takashima Y., Guo G., Loos R., Nichols J., Ficiz G., Krueger F., Oxley D., Santos F., Clarke J., Mansfield W., Reik W., Bertone P., Smith A. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell.* 2014;158(6):1254-1269. doi 10.1016/j.cell.2014.08.029
- Theunissen T.W., Powell B.E., Wang H., Mitalipova M., Faddah D.A., Reddy J., Fan Z.P., ... Gao Q., Dawlaty M.M., Young R.A., Gray N.S., Jaenisch R. Systematic identification of culture conditions for induction and maintenance of naïve human pluripotency. *Cell Stem Cell.* 2014;15(4):471-487. doi 10.1016/j.stem.2014.07.002
- Theunissen T.W., Friedli M., He Y., Planet E., O'Neil R.C., Markoulaki S., Pontis J., ... Drotar J., Lungjangwa T., Trono D., Ecker J.R., Jaenisch R. Molecular criteria for defining the naïve human pluripotent state. *Cell Stem Cell.* 2016;19(4):502-515. doi 10.1016/j.stem.2016.06.011
- Ustyantseva E., Pavlova S.V., Malakhova A.A., Ustyantsev K., Zakian S.M., Medvedev S.P. Oxidative stress monitoring in iPSC-derived motor neurons using genetically encoded biosensors of H₂O₂. *Sci Rep.* 2022;12(1):8928. doi 10.1038/s41598-022-12807-z
- Valamehr B., Robinson M., Abujarour R., Rezner B., Vranceanu F., Le T., Medcalf A., Lee T., Fitch M., Robbins D., Flynn P. Platform for induction and maintenance of transgene-free hiPSCs resembling ground state pluripotent stem cells. *Stem Cell Rep.* 2014;2(3):366-381. doi 10.1016/j.stemcr.2014.01.014
- Vallot C., Ouimette J.F., Makhlof M., Féraud O., Pontis J., Côme J., Martinat C., Bannaceur-Griscelli A., Lalande M., Rougeulle C. Erosion of X chromosome inactivation in human pluripotent cells initiates with *XACT* coating and depends on a specific heterochromatin landscape. *Cell Stem Cell.* 2015;16(5):533-546. doi 10.1016/j.stem.2015.03.016
- Vallot C., Patrat C., Collier A.J., Huret C., Casanova M., Liyakat Ali T.M., Tosolini M., Frydman N., Heard E., Rugg-Gunn P.J., Rougeulle C. *XACT* noncoding RNA competes with *XIST* in the control of X chromosome activity during human early development. *Cell Stem Cell.* 2017;20(1):102-111. doi 10.1016/j.stem.2016.10.014
- Vaskova E.A., Medvedev S.P., Sorokina A.E., Nemudryy A.A., Elisaphenko E.A., Zakharova I.S., Shevchenko A.I., Kizilova E.A., Zhelezova A.I., Evshin I.S., Sharipov R.N. Transcriptome characteristics and X-chromosome inactivation status in cultured rat pluripotent stem cells. *Stem Cells Dev.* 2015;24(24):2912-2924. doi 10.1089/scd.2015.0204
- Ware C.B., Nelson A.M., Mecham B., Hesson J., Zhou W., Jonlin E.C., Jimenez-Caliani A.J., ... Blau C.A., Treuting P.M., Hawkins R.D., Cirulli V., Ruohola-Baker H. Derivation of naïve human embryonic stem cells. *Proc Natl Acad Sci USA.* 2014;111(12):4484-4489. doi 10.1073/pnas.1319738111
- Warr N., Powles-Glover N., Chappell A., Robson J., Norris D., Arkell R. Zic2-associated holoprosencephaly is caused by a transient defect in the organizer region during gastrulation. *Hum Mol Gen.* 2008;17(19):2986-2996. doi 10.1093/hmg/ddn197
- Welling M., Chen H.H., Muñoz J., Musheev M.U., Kester L., Junker J.P., Mischerikow N., Arbab M., Kuijk E., Silberstein L., Kharchenko P.V. DAZL regulates Tet1 translation in murine embryonic stem cells. *EMBO Rep.* 2015;16(7):791-802. doi 10.15252/embr.201540538
- Yarkova E.S., Grigor'eva E.V., Medvedev S.P., Tarasevich D.A., Pavlova S.V., Valetdinova K.R., Minina J.M., Zakian S.M., Malakhova A.A. Detection of ER stress in iPSC-derived neurons carrying the p. N370S mutation in the *GBA1* gene. *Biomedicine.* 2024;12(4):744. doi 10.3390/biomedicine12040744
- Yu L., Wei Y., Sun H.X., Mahdi A.K., Pinzon Arteaga C.A., Sakurai M., Schmitz D.A., ... Okamura D., Muto A.A., Gu Y., Ross P.J., Wu J. Derivation of intermediate pluripotent stem cells amenable to primordial germ cell specification. *Cell Stem Cell.* 2021;28(3):550-567. doi 10.1016/j.stem.2020.11.003
- Zakharova I.S., Zhiven' M.K., Saaya S.B., Shevchenko A.I., Smirnova A.M., Strunov A., Karpenko A.A., Pokushalov E.A., Ivanova L.N., Makarevich P.I., Parfyonova Y.V. Endothelial and smooth muscle cells derived from human cardiac explants demonstrate angiogenic potential and suitable for design of cell-containing vascular grafts. *J Transl Med.* 2017;15(1):54. doi 10.1186/s12967-017-1156-1
- Zakharova I., Saaya S., Shevchenko A., Stupnikova A., Zhiven' M., Laktionov P., Stepanova A., ... Zavjalov E., Chernyavsky A., Romanov A., Karpenko A., Zakian S. Mitomycin-treated endothelial and smooth muscle cells suitable for safe tissue engineering approaches. *Front Bioengin Biotechnol.* 2022;10:772981. doi 10.3389/fbioe.2022.772981
- Zakharova I.S., Shevchenko A.I., Arssan M.A., Sleptcov A.A., Nazarenko M.S., Zarubin A.A., Zheltysheva N.V., ... Saaya S.B., Ezhov M.V., Kukharchuk V.V., Parfyonova Y.V., Zakian S.M. iPSC-derived endothelial cells reveal LDLR dysfunction and dysregulated gene expression profiles in familial hypercholesterolemia. *Int J Mol Sci.* 2024;25(2):689. doi 10.3390/ijms25020689
- Zimmerlin L., Park T.S., Huo J.S., Verma K., Pather S.R., Talbot C.C., Jr., Agarwal J., ... Cope L., Canto-Soler M.V., Friedman A.D., Baylin S.B., Zambidis E.T. Tankyrase inhibition promotes a stable human naïve pluripotent state with improved functionality. *Development.* 2016;143(23):4368-4380. doi 10.1242/dev.138982

Conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received August 12, 2025. Revised October 20, 2025. Accepted November 6, 2025.

doi 10.18699/vjgb-26-22

Molecular genetic variability in coenopopulations of *Caragana jubata* (Pall.) Poir. (Fabaceae) in the mountains of Central Asia and Southern Siberia revealed using ISSR-markers

D.A. Krivenko ¹ , O.A. Chernysheva¹, I.N. Kuban ², E.V. Zhmud ², I.V. Gorbenko¹, A.A. Achimova ³,
O.Yu. Vasilyeva², O.V. Dorogina ^{2, 4}

¹ Siberian Institute of Plant Physiology and Biochemistry of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

² Central Siberian Botanical Garden of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Altai Branch of the Central Siberian Botanical Garden of the Siberian Branch of the Russian Academy of Sciences "Gorno-Altai Botanical Garden", Kamlak, Republic of Altai, Russia

⁴ Novosibirsk State University, Novosibirsk, Russia

 krivenko.irk@gmail.com

Abstract. The genetic structure was studied using five ISSR markers in 44 individuals of samples obtained from eight coenopopulations (CPs) of *Caragana jubata* (Pall.) Poir. in the mountainous conditions of Central Asia – in the Western Pamirs and Inner Tien Shan (Kyrgyzstan) and in Southern Siberia – in Altai (Republic of Altai), Western (Republic of Tyva) and Eastern (Republic of Buryatia) Sayan. We studied the species in a range of geographical distances of more than 2,500 km and in the range of absolute altitudes of 1,570–3,680 m. *Caragana jubata* is listed in the Red Data Books of eight subjects of the Russian Federation. The species population is declining, including due to anthropogenic impact. The aim of the current work is to identify genetic diversity and heterogeneity in *C. jubata* coenopopulations depending on their geographic and altitudinal confinement in the mountains of Central Asia and Southern Siberia. It was shown that in undisturbed locations, the studied CPs of this species were characterized by a high number of individuals and by the occupied area of more than 100 m². Almost every sample from the *C. jubata* CP studied by us (except for representatives from CP7) contained genotypes possessing unique DNA fragments. The highest proportion of such genotypes (75 %) was found in the sample from CP3 (Inner Tien Shan, Teskey-Ala-Too Ridge, absolute altitude 2,550 m). We did not find unique fragments in the genotypes of individuals from the studied sample of CP7 (Western Sayan, Republic of Tuva). Anthropogenic impact on plants at this location is a possible reason for that. The revealed predominance of intrapopulation genetic variability over interpopulation genetic variability in samples from eight *C. jubata* CPs studied by us may indicate the stability of representatives of this species within the parts of the range studied by us.

Key words: camel's tail caragana; vulnerable species; coenopopulations; genetic polymorphism; genetic diversity parameters

For citation: Krivenko D.A., Chernysheva O.A., Kuban I.N., Zhmud E.V., Gorbenko I.V., Achimova A.A., Vasilyeva O.Yu., Dorogina O.V. Molecular genetic variability in coenopopulations of *Caragana jubata* (Pall.) Poir. (Fabaceae) in the mountains of Central Asia and Southern Siberia revealed using ISSR-markers. *Vavilovskii Zhurnal Genetiki i Selektzii* = *Vavilov J Genet Breed*. 2026;30(2):194-204. doi 10.18699/vjgb-26-22

Funding. The work was conducted with funds received from the state assignments of the Central Siberian Botanical Garden of the Siberian Branch of the Russian Academy of Sciences "Analysis of biodiversity, conservation and restoration of rare and resource plant species using experimental methods", state registration number: AAAA-A21-121011290025-2 and the Siberian Institute of Plant Physiology and Biochemistry of the Siberian Branch of the Russian Academy of Sciences "Study of the composition and dynamics of phytodiversity of Baikal Siberia in the original information and analytical environment", state registration number: 125020905982-9, as well as with the financial support of the project of the Ministry of Science and Higher Education of the Russian Federation No. FSUS-2024-002.

Молекулярно-генетическая изменчивость по ISSR-маркерам у *Caragana jubata* (Pall.) Poir. (Fabaceae) в горах Средней Азии и Южной Сибири

Д.А. Кривенко ¹ , О.А. Чернышева¹, И.Н. Кубан ², Е.В. Жмуд ², И.В. Горбенко¹, А.А. Ачимова ³,
О.Ю. Васильева², О.В. Дорогина ^{2, 4}

¹ Сибирский институт физиологии и биохимии растений Сибирского отделения Российской академии наук, Иркутск, Россия

² Центральный сибирский ботанический сад Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Алтайский филиал Центрального сибирского ботанического сада Сибирского отделения Российской академии наук «Горно-Алтайский ботанический сад», Камлак, Республика Алтай, Россия

⁴ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

 krivenko.irk@gmail.com

Аннотация. Исследована генетическая структура по пяти ISSR-маркерам у 44 особей в выборках из восьми ценопопуляций (ЦП) *Caragana jubata* (Pall.) Poir. в горных условиях Центральной Азии – на Западном Памире и Внутреннем Тянь-Шане (Кыргызстан) и в Южной Сибири – на Алтае (Республика Алтай), Западном (Республика Тыва) и Восточном (Республика Бурятия) Саяне. Вид изучен в диапазоне географических расстояний, составляющих более 2.5 тыс. км и в диапазоне абсолютных высот 1570–3680 м. *Caragana jubata* занесена в Красные книги восьми субъектов Российской Федерации. Цель настоящей работы – выявление генетического разнообразия и гетерогенности в ЦП *C. jubata* в зависимости от их географической и высотной приуроченности в горах Центральной Азии и Южной Сибири. Исследования показали, что в ненарушенных местообитаниях изученные ЦП этого вида характеризовались высокими параметрами численности особей и занимали площадь более 100 м². Практически в каждой изученной нами выборке из ЦП *C. jubata* (кроме представителей из ЦП7) обнаружено наличие генотипов, содержащих уникальные фрагменты ДНК. Наибольшая доля (75 %) таких генотипов установлена в выборке из ЦП3 (Внутренний Тянь-Шань, хр. Тескей-Ала-Тоо, абсолютная высота 2550 м). Нами не найдено уникальных фрагментов в генотипах у особей из исследованной выборки в ЦП7 (Западный Саян, Республика Тыва). Возможная причина – антропогенное воздействие на растения в данном местообитании. Выявленное преобладание внутривидовой генетической изменчивости над межпопуляционной в выборках из восьми изученных ЦП *C. jubata* может свидетельствовать об устойчивости представителей этого вида в пределах исследованных нами частей ареала.

Ключевые слова: карагана гривастая; уязвимый вид; ценопопуляции; генетический полиморфизм; параметры генетического разнообразия

Introduction

Caragana jubata (Pall.) Poir. (camel's tail caragana) is a Neogene (Miocene-Pliocene) relict that diverged from its closest relatives more than 4.8 million years ago (Zhang M.-L. et al., 2016). It is a thorny monoecious shrub with bisexual flowers, 0.3–2.5 m tall. It is a heliophilous, predominantly high-mountain species, a mesopsychrophyte that grows in the forest (from 1,100 to 3,000 m above sea level – on the Stanovoy Highlands and Tien Shan, respectively) and high-mountain belts (up to 4,700 m above sea level – on the Tibetan Plateau), in the Arctic tundra (up to 500 m above sea level – in the Kharaulakh Mountains) on rocky placers and cliffs, along the banks of rivers, under the canopy of sparse forests (Koropachinsky, Vstovskaya, 2012; Churyulina, 2021; Oorzhak et al., 2024). The modern range of this relict species is Pan-Asian and disjunctive: from the Tibetan Plateau, the Himalayas, Tien Shan, Altai, Sayan, and Stanovoy Highlands to the Kharaulakh Mountains and the Kolyma Highlands (Churyulina, Bocharnikov, 2019). Optimal conditions are found in the upper part of the spruce forest belt in the Central Tien Shan, where the plants are well-foliated, bear fruit abundantly, and reach a height of 2.0–2.5 m (Lysova, 1967).

C. jubata is listed in the Red Data Books of eight constituent entities of the Russian Federation. In the Red Data Books of the Republic of Altai (2017), Magadan Oblast (2019), and Krasnoyarsk Territory (2022), it is assigned rarity status category 3 (R) – a rare species with a relict disjunctive range. In the Republics of Buryatia (2023) and Sakha (Yakutia) (2017), Irkutsk Oblast (2020), Trans-Baikal (2017), and Khabarovsk (2019) Territories, it is assigned rarity status category 2 (VU) – a species declining in numbers, including as a result of overexploitation by humans.

Limiting factors for the spread of *C. jubata* may include the isolation of populations from each other and their low abundance, as well as the demand for the basic composition of rocks due to growth primarily on limestones and carbonate sandstones. Furthermore, exogenous geomorphological processes, such as landslides and floods, pose a threat to populations of this species (Semenova, 2007; Koropachinsky, Vstovskaya, 2012). *C. jubata* is widely used in folk medicine in Eastern Siberia, which leads to unsystematic harvesting of raw materials and the felling of shrubs for medicinal purposes, since harvesting is carried out year-round and plants are harvested whole, including the root system. Livestock grazing also leads to the depletion of natural populations (Semenova, 2007; Parygin, Khudonogova, 2017; Churyulina et al., 2020).

According to the study of the composition of biologically active substances and the pharmacological activity of an aqueous extract of aerial plant parts, the dominant fraction of biologically active substances in *C. jubata* are polyphenolic compounds (flavonoids), primarily mono- and diglycosides, which are derivatives of O-hydroxylated flavonols. The antioxidant activity of *C. jubata* aqueous extracts *in vitro*, their hepatoprotective, antiviral, anti-inflammatory and wound-healing activity were revealed, while the absence of toxicity was noted. Therefore, aqueous extracts are recommended for the treatment of skin and mucous membrane diseases associated with bacterial inflammatory processes (Kakorin et al., 2020).

Decoctions and alcoholic extracts of the aerial parts of *C. jubata* were found to possess hepatoprotective, antitumor, anti-inflammatory, and antiviral properties (Kakorin et al., 2018a, b). An ethanol extract of *C. jubata* is used in traditional Chinese medicine to treat hypertension and,

in an experiment on rats, reduced infarct size, reduced cerebral edema, and improved neurological parameters in induced ischemic stroke (Wang et al., 2020; Zhao et al., 2023). Methanol extracts of *C. jubata* roots exhibited cytotoxic activity against hepatocellular carcinoma cell lines (Yang et al., 2015).

Studies of the genetic structure of *C. jubata* populations are extremely limited. For example, intraspecific differentiation and phylogeography of this species were studied using sequencing studies of nuclear and chloroplast markers in samples from Tibet, Tien Shan, Eastern Sayan, and the northeastern coast of the Sea of Okhotsk (Zhang et al., 2016; Hantemirova et al., 2024). Using SSR markers, the effect of altitudinal gradients on the genetic structure of alpine *C. jubata* plants was studied (Yao et al., 2025). We found no literature sources on the study of patterns of variability for ISSR markers in CPs from geographically distant locations at different absolute altitudes, which affect the genetic differentiation of *C. jubata* CPs.

In a review by T. Ohsawa and I. Yuji (2008), they summarized data on the influence of absolute elevations of herbaceous perennials and dendroflora on intrapopulation polymorphism using neutral molecular markers at regional scales (approximately 1–10 km). Larger scales (>102 km) were not considered, since, according to the authors, large horizontal distances influence genetic variability no less than limited vertical distances. Isoenzymes, RAPD, or SSR were used to assess genetic polymorphism. According to the authors' conclusions, a specific locus or allele at a given locus can exhibit clear differentiation between populations at different elevations as a result of natural selection.

It was shown that in studies of the genetic heterogeneity of coenopopulations (CPs) of different plant species along altitudinal gradients, the genetic diversity within a CP varied according to one of four patterns of intrapopulation genetic variability, usually represented by expected heterozygosity: $L < M > H$, $L > M > H$, $L < M < H$, or $L = M = H$, where L, M, and H represent low, medium, and high altitudes, respectively. In particular, a number of studies (Oyama et al., 1993; Yifru et al., 2006) have shown that CPs at intermediate altitudes had greater genetic diversity than those in lower and higher habitats, since geographically central populations, according to the authors, are in optimal environmental conditions, while peripheral populations are in suboptimal conditions. For example, in *Arabis serrata* Franch. & Sav. (cressweed) on an absolute altitude gradient of 1,440–2,400 m and *Triticum*

turgidum subsp. *durum* (Desf.) Husn. (durum wheat) – 1,500–3,300 m, the highest intrapopulation variability was observed in middle parts of their altitudinal ranges, respectively, examined with enzyme loci and 29 nSSR loci (Oyama et al., 1993; Yifru et al., 2006).

X. Zhang et al. (2006) found that highly heterogeneous environmental conditions imposed by elevational gradients can influence neutral regions closely associated with the genomic region under selection. Even in cases where differentiation is not detected, intrapopulation genetic variability often changes along elevational gradients.

Identification of the status of genetic variation depending on absolute altitude will help to determine future research directions. In particular, the effectiveness of selection of the most resistant forms of *C. jubata* during introduction to cultivation will be significantly increased by using propagation material from individuals from the most genetically heterogeneous CPs. Therefore, it is crucial to study the genetic diversity and population structure of wild vulnerable plant species to ensure their most rational use (Liu et al., 2024; Yao et al., 2025).

The aim of the current study was to identify genetic diversity and heterogeneity in *C. jubata* coenopopulations depending on their geographic and altitudinal distribution in the mountains of Central Asia and Southern Siberia.

Materials and methods

We collected the research material in 2015 and 2023–2024 in the central areas of five regions located in mountainous conditions: in Kyrgyzstan (Western Pamir – CP1, CP2 and Inner Tien Shan – CP3) and in three constituent entities of the Russian Federation (Southern Siberia): in the Republics of Altai (CP4), Tyva (CP5–CP7) and Buryatia (CP8) at altitudes of 1,575–3,680 m above sea level (Fig. 1, Supplementary Material 1)¹.

For the genetic analysis, leaves from 5–6 individuals in each CP were collected. DNA was isolated from the leaves using a standard CTAB-based method (Nabieva et al., 2020). PCR was performed on a C1000 amplifier (Bio-Rad, USA) in a volume of 25 µl. The reaction mixture contained: 1.5 units of Taq DNA polymerase (Medigen, Russia); 2.7 mM of MgCl₂; 0.8 mM of ISSR primer (Medigen, Russia); 2 µl of DNA solution; 2 µl of mQ H₂O. The amplification program consisted of a DNA denaturation step for 90 s at 94 °C and 35 cycles, each of which included 40 s at 94 °C, 45 s of primer an-

¹ Supplementary Materials 1 and 2 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Kriv_Engl_30_2.pdf



Fig. 1. Locations of the studied specimens of *C. jubata* in Kyrgyzstan (CP1–CP3) and the Russian Federation: in the Republics of Altai (CP4), Tyva (CP5–CP7) and Buryatia (CP8).

nealing, and 90 s at 72 °C. The final step of nucleotide chain elongation lasted 5 min at 72 °C.

Electrophoretic separation of the amplification products was performed in 1.5 % agarose gel in 1x TAE buffer at a voltage of 4 V/cm. To estimate the length of the PCR products, the DNA marker 100 bp DNA Ladder (no stain) (BIORON GmbH, Germany) was used.

In total, we tested five ISSR primers that showed the highest efficiency in working with rare plant species (Dorogina et al., 2023; Zhmud et al., 2023, 2024): HB12 ((CAC)3GC), UBC834 ((AG)8YT), UBC856 ((AC)8YA), UBC857 ((AC)8YG) and UBC881 ((GGTG)₃). Of these, UBC857 was selected, as it turned out to be the most informative (Fig. 2, Supplementary Material 2).

The primary analysis of the study results was performed via GenAlEx6.1, a software package and specialized macro for MS-Excel (Smouse et al., 2017). Further analysis and results visualization were performed using the R programming environment (R Core Team, 2024). Data preparation was conducted using the ‘tidyverse’ package (Wickham et al., 2019). The Hamming distances and the UPGMA method were used to construct the dendrogram, and the tree was visualized using the ‘ggtree’ package (Xu et al., 2022). Correspondence analysis was

performed using the ‘adeigenet’ package (Jombart et al., 2008) and the result was used for additional visualization of the genetic structure of the studied CPs. The correspondence analysis, similar to principal component analysis, allows for data dimensionality reduction and visualization; however, it is designed for categorical data and uses chi-squared distance, which is more suitable for genetic data, where relative allele frequencies are important.

Isolation by distance analysis was performed according to the method described by L.S. Nørgaard et al. (2017), and isolation by elevation analysis was performed according to the method of M.-M. Shi et al. (2011). Both analyses were performed using the ‘adeigenet’ package (Jombart et al., 2008), with Jaccard distances used as genetic distances; the resulting plots were generated using the ‘ggpubr’ package (Kassambara, 2023). Isolation by distance is a phenomenon that describes patterns of genetic variation in populations arising due to spatially restricted gene flow and is defined as a decrease in the genetic similarity between populations with increasing geographic distances between them (Wright, 1974). Isolation by elevation is a similar phenomenon but with elevation differences between populations used as a spatial component.

Table 1. Alleles of genotypes with repetitive and unique fragments in individuals from the studied CPs of *C. jubata*

No.	Individual No.	Alleles	No.	Individual No.	Alleles
Genotype variants with unique fragments			22	3.2	–
1	1.4	011111011	23	3.3	–
2	1.6	010010011	24	1.3	010111011 (2)
3	2.2	010001011	25	1.5	–
4	2.3	000111011	26	1.1	010110011 (2)
5	2.4	001111111	27	2.1	–
6	3.4	110110010	28	5.3	010000100 (3)
7	3.5	110000010	29	5.4	–
8	3.6	000100111	30	5.5	–
9	4.1	010110100	31	4.2	010111100 (2)
10	4.4	000001100	32	5.6	–
11	4.5	011100100	33	7.1	110000000 (5)
12	5.1	110111000	34	7.2	–
13	5.2	110000100	35	7.3	–
14	6.1	001111000	36	7.5	–
15	6.6	010111100	37	8.2	–
16	8.3	010110000	38	6.2	010000100 (2)
17	8.4	110110000	39	6.3	–
18	8.5	110010000	40	6.4	010111000 (2)
Genotype variants with repeating fragments (number)			41	8.1	–
19	1.2	010000011 (5)	42	4.3	010000000 (3)
20	2.5	–	43	6.5	–
21	3.1	–	44	7.4	–

Note. 1 – presence of an allele, 0 – absence.

Results

Our studies have shown that the *C. jubata* CPs we studied are characterized by high individual abundances and occupy areas exceeding 100 m² (Supplementary Material 1). More than 200 individuals were identified in CP1 and CP2, whereas in CP3–CP6, 20 to 50 individuals were found. The individuals inhabited in general undisturbed sites with high grass cover that exceed 50 %. CP7 was an exception, as its few individuals (approximately 10 specimens) were subject to anthropogenic impact due to the construction of a road through its location. In this CP, the aboveground parts of the individuals appeared damaged and lacked developed reproductive organs. The grass cover in this location was also low (Supplementary Material 1).

Molecular genetic studies in samples from the CPs of *C. jubata* showed that the number of genotypes containing unique fragments in the studied sample from all CPs was 18, and the number of repeated variants was nine (Table 1). In each of the seven studied CPs (CP1–CP6, CP8), except CP7, genotypes with unique fragments were identified that were not observed in individuals from the other studied CPs (Table 1). In the sample of individuals from CP7, only two genotype variants identical to the genotype variants of individuals 4.3 and 8.2 (respectively, from CP4 and CP8) were identified (Table 1, Fig. 2). The number of genotype variants in *C. jubata* individuals from the studied CPs varied from 2 to 5 (Table 2). The proportion of genotypes with unique fragments in the studied samples of the CPs varied from none (in CP7)

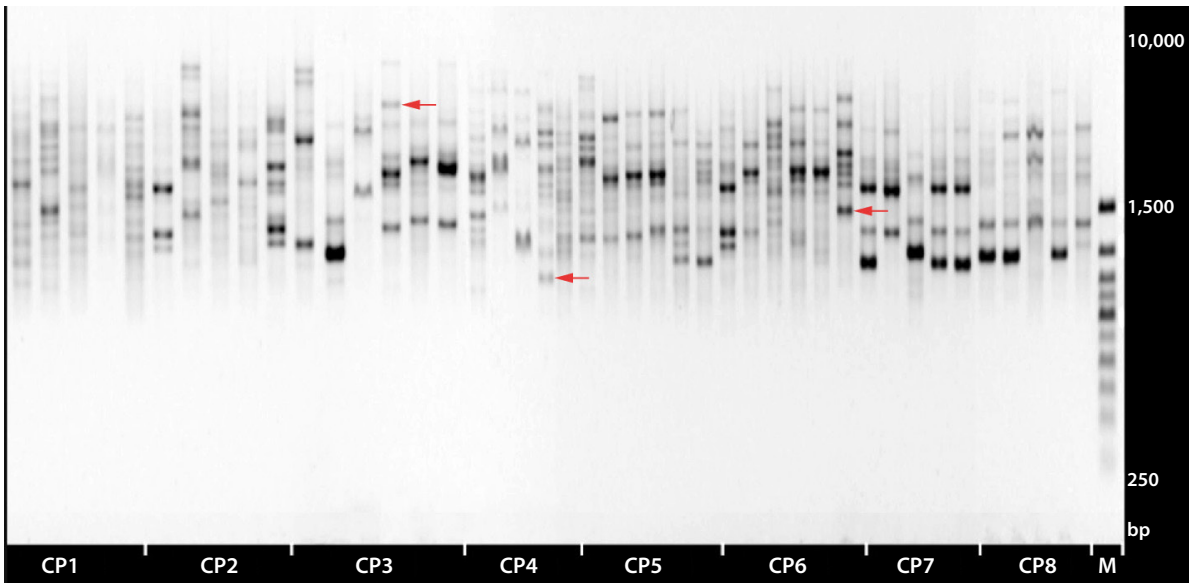


Fig. 2. Electropherogram of DNA fragment amplification products with the UBC857 primer. Samples taken from *C. jubata* CPs.
CP1–CP8 – CP numbers; M – the molecular weight marker; arrows indicate unique fragments of genotypes.

Table 2. Parameters of genetic diversity in samples from *C. jubata* CPs

CP	<i>N</i>	<i>Na</i>	<i>Ne</i>	<i>He</i>	<i>uHe</i>	<i>P</i> , %	<i>Nu/Ng</i>	<i>Nr</i>	Ratio <i>Nu</i> , %
CP1	6	1.22 ± 0.28*	1.31 ± 0.14	0.17 ± 0.07	0.19 ± 0.08	44.4	2/5**	3	40
CP2	5	1.56 ± 0.24	1.44 ± 0.14	0.25 ± 0.07	0.28 ± 0.08	66.7	3/5	2	60
CP3	6	1.44 ± 0.29	1.35 ± 0.13	0.21 ± 0.07	0.23 ± 0.07	66.7	3/4	1	75
CP4	5	1.33 ± 0.33	1.46 ± 0.14	0.26 ± 0.07	0.29 ± 0.08	66.7	3/5	2	60
CP5	6	1.22 ± 0.32	1.29 ± 0.11	0.19 ± 0.06	0.20 ± 0.07	55.6	2/4	2	50
CP6	6	1.33 ± 0.33	1.44 ± 0.13	0.26 ± 0.07	0.28 ± 0.08	66.7	2/5	3	40
CP7	5	0.33 ± 0.24	1.11 ± 0.11	0.06 ± 0.05	0.06 ± 0.06	11.1	0/2	0	0
CP8	5	1.00 ± 0.33	1.33 ± 0.15	0.18 ± 0.08	0.20 ± 0.09	44.4	3/5	2	60

* Stands for mean value ± standard error; *N* – number of samples; *Na* – number of alleles; *Ne* – number of effective alleles; *He* – expected heterozygosity; *uHe* – objective expected heterozygosity; *P* – proportion of polymorphic loci; *Nu/Ng* – ratio of the number of genotypes with unique fragments to the number of genotype variants in the CP; *Nr* – number of repeating genotype variants in each CP.

** The denominator indicates the total number of genotype variants in the samples from the CP.

to 75 % (in CP3) (Table 2, Fig. 2). It was calculated as the ratio of the number of unique genotypes containing unique fragments to the total number of genotype variants in each CP. Representatives from CP3 grew in high-mountain conditions at an absolute altitude of 2,550 m, i. e. they occupied an intermediate geographic and altitudinal position between the Pamir and South Siberian CPs (Supplementary Material 1).

According to the AMOVA analysis, in samples from the eight studied CPs of *C. jubata*, the interpopulation genetic variability was more than 1.5 times lower than the intrapopulation one: 38 and 62 %, respectively.

According to cluster analysis, the presence of two particularly genetically distant groups of individuals was revealed. These are samples from CP1–CP2 and CP5–CP8. They are characterized by low genetic similarity, as shown in the distribution diagram of *C. jubata* individuals according to their genotypic characteristics (Fig. 3), and are located primarily in different clades. Based on genotypic similarity, the studied *C. jubata* individuals were divided into two main clusters that are in general in agreement with their geographic distribution (Fig. 3a).

According to cluster analysis, the upper part of the tree is predominantly composed of individuals from the

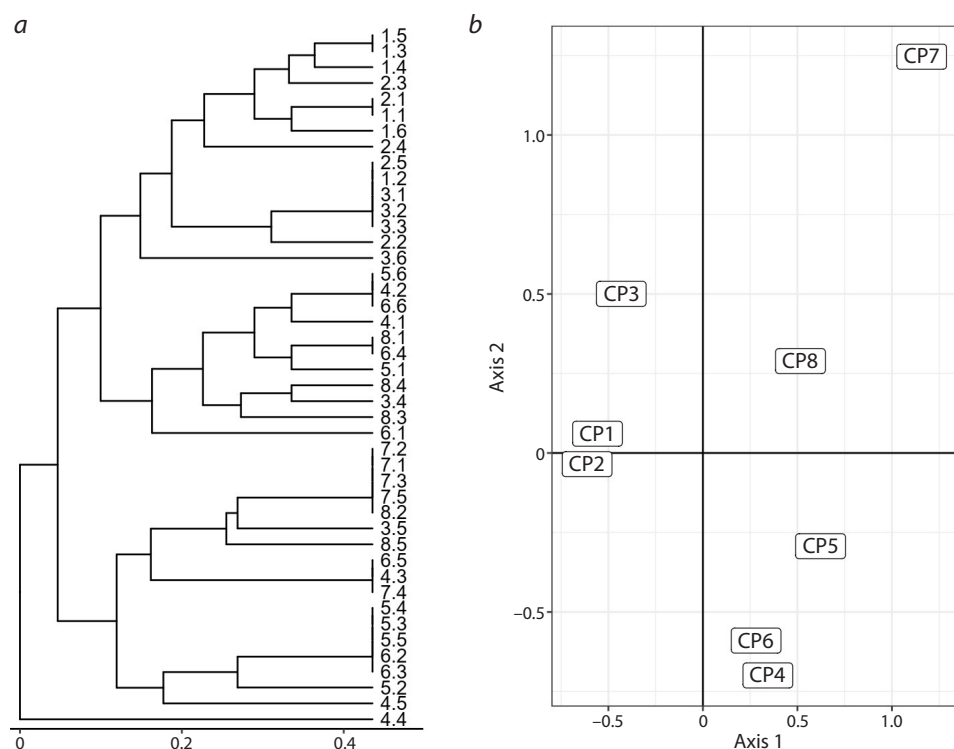


Fig. 3. The clustering of *C. jubata* CPs into groups based on the similarity of genetic structure.

a – UPGMA dendrogram constructed using the ISSR spectra data obtained with the UBC857 primer; *b* – the result of correspondence analysis. CP1 – individuals 1.1–1.6; CP2 – individuals 2.1–2.5; CP3 – 3.1–3.6; CP4 – individuals 4.1–4.5; CP5 – individuals 5.1–5.6; CP6 – individuals 6.1–6.6; CP7 – individuals 7.1–7.5; CP8 – individuals 8.1–8.5.

mountainous regions of the Western Pamirs and Inner Tien Shan (CP1–CP3). This group also includes some southern Siberian individuals from samples of CP4–CP6 and CP8. The lower part of the tree is generally characterized by a distinct distribution of southern Siberian individuals (CP4–CP8). Individuals from CP7 are located separately, and the closest groups to them are those from southern Siberian CPs. The results of the correspondence analysis also support this distribution (Fig. 3b).

According to the obtained data, individuals in the geographically extreme CPs grew at distances of over 2,500 km from each other. Statistical analysis of the data in the samples from the studied CPs revealed significant correlations between genetic distances and both geographic and elevational distances, indicating isolation by distance and isolation by elevation. That is, as the geographic and elevational distances in the samples of individuals from these CPs increase, so do their genetic distances (Fig. 4).

Discussion

Genome sequencing data for representatives of the genus *Caragana* Fabr. revealed that its center of origin is located on the Tibetan Plateau (Zhang M.-L. et al., 2016).

During the xerothermic period of the Pleistocene, *C. jubata* penetrated through mountain systems to the west and north. This is supported, in particular, by data on its modern distribution, which is associated with mountainous regions: the eastern part of the Tibetan Plateau, the northeastern Himalayas, the Central and Southern Tien Shan, the Sayan Mountains, and the Stanovoye Highlands (Churiulina, Bocharnikov, 2019).

We studied *C. jubata* in samples from the Central Plain, also growing over a wide geographic range. Their locations are situated in five mountainous regions of Central Asia – the Western Pamirs, the Inner Tien Shan, and in southern Siberia – the Altai, Western and Eastern Sayan. *C. jubata* CPs were studied across an elevation gradient of more than 2,000 m (Supplementary Material 1) and geographic distances of more than 2,500 km. This wide geographic and altitudinal range of *C. jubata* distribution is consistent with the current understanding that *C. jubata* species exhibits high ecological heterogeneity (Talovina, 2019).

Within the studied altitudinal range of over 2,000 meters and distances of 2,500 kilometers, we identified 18 genotype variants with unique fragments in samples from 44 representatives of this species, a relatively high

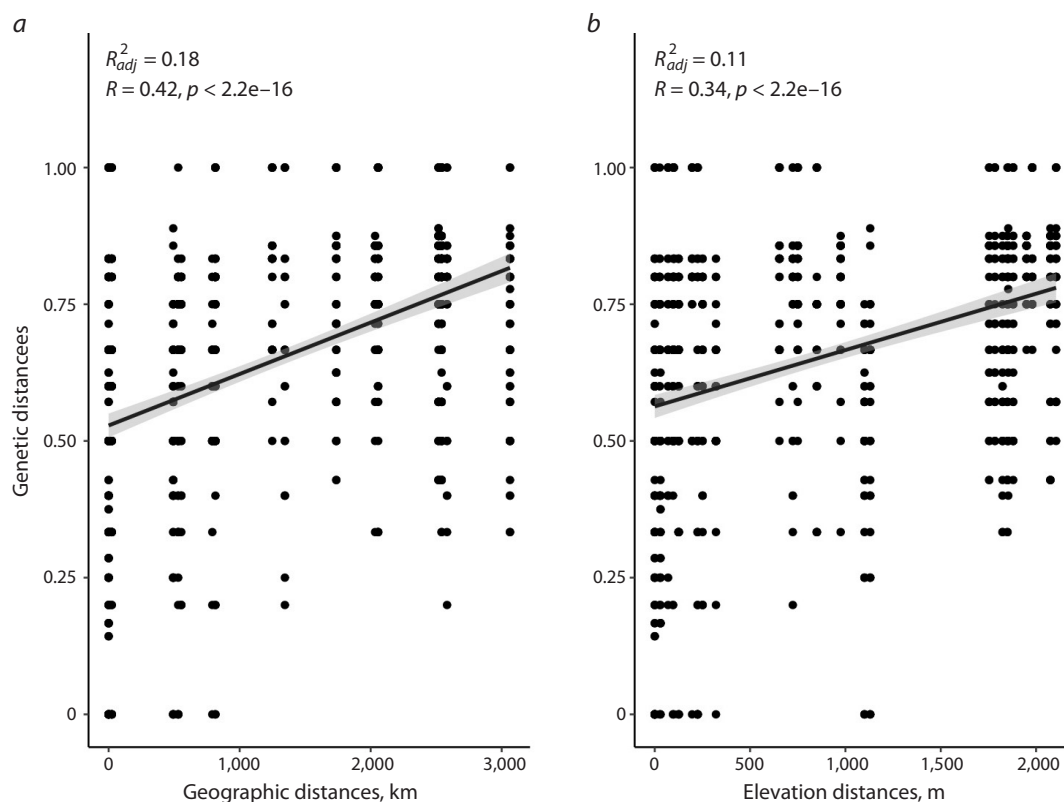


Fig. 4. Correlations of genetic distances of *C. jubata* CPs with geographic distances (a) and relative elevations (b) of their locations.

Each point represents a comparison of two individuals, characterized by genetic distance (Y) and geographic/elevation distance (X).

number for a rare plant species. The value of genetic diversity lies in the fact that a gene variation that is neutral today can become adaptive later, when environmental conditions change (de Lafontaine et al., 2018). When environmental conditions change, a population with higher genetic diversity is more likely to adapt to the new environment than a population with less diversity. Thus, maintaining a species' genetic diversity can significantly increase its chances of long-term survival (Chung et al., 2023).

In mountainous environments, plant populations growing at high altitudes demonstrate later flowering and have a shorter flowering period compared to populations of the same species growing at lower altitudes (Reisch et al., 2005). A simple explanation for differences in genetic variation among populations located at the same altitude but separated by mountainous terrain is the difficulty of their diaspore dispersal (Taberlet et al., 1998). A study using paleobotanical and genetic data for *Fagus sylvatica* L. (European beech) showed that mountain ranges in Europe were not a geographic barrier for this species, but rather facilitated its dispersal. On the contrary, its dispersal was hindered by the presence of large plains or valleys (Magri et al., 2006). From this,

the authors concluded that flat landscapes can act as a barrier to the dispersal of diaspores, including those of high-altitude species.

However, rare and endemic plant species are typically characterized by significantly lower genetic diversity and, consequently, higher homozygosity compared to non-endemic species (Hamabata et al., 2019). Genetic homogenization of populations, especially in rare plant species, results in reduced heterozygosity among their individuals. Populations lacking genetic variability are unable to evolve in response to changing environmental conditions and, therefore, may face an increased risk of extinction (Dorogina, Zhmud, 2020).

In addition to genotypes with unique fragments, we identified nine recurring genotype variants in *C. jubata* specimens taken from the studied CPs. Specimens from CP7 exhibited genotype variants similar to those in specimens from other CPs. The absence of unique genotype fragments in the sample from this CP may be explained by anthropogenic impacts on its specimens in the area where a road is being constructed. We observed a similar situation, for example, in specimens of rare Siberian species *Adonis villosa* Ledeb. (downy adonis), where only four genotype variants were identified in 40 specimens

studied in six CPs in the Republic of Altai. Road construction was also carried out in its habitats, and damage to plants and soil was observed (Zhud et al., 2024).

We also found genetic homogenization in the habitat of one of the CPs of the rare species *Rhodiola rosea* L. (rose rhodiola) in the Republic of Altai. Individuals of this species were damaged by trampling at a summer camp and by grazing by cattle (yaks) (Dorogina et al., 2023). It is possible that the *C. jubata* individuals remaining in CP7 are more resilient to soil damage, in this case related to construction work. Genetic homogenization of individuals in this CP raises concerns about its sustainability and may eventually lead to its complete extinction.

Cluster analysis of the distribution of the most polymorphic components detected via ISSR labeling showed that the studied CPs were divided into different groups according to their genetic structure. We previously noted that the locations of the *C. jubata* CPs we studied differ by more than 2,500 km at their extreme points. A statistically confirmed trend of increasing genetic distances between representatives from different CPs with increasing spatial distance was found.

We also found a tendency toward an increased proportion of unique genotypes (containing unique fragments) in the *C. jubata* CP growing in the Tien Shan mountains at an altitude of 2,550 m – the middle part of the studied altitudinal range – compared to the Pamir and South Siberian CPs. According to T. Ohsawa and I. Yuji (2008), such populations may experience more favorable environmental conditions than peripheral populations growing in suboptimal conditions. The authors believe that mountain isolation plays a crucial role in the diversification and evolution of species.

Adaptation of wild plants to stress with increasing altitude is a complex process achieved through both modification and genetic variability (Safaralikhonov, Aknazarov, 2021). Molecular studies by P.K. Bhardwaj et al. (2013) showed that *C. jubata* exhibits a predominance of genes encoding chaperones and genes involved in growth and development at low temperatures, which are among the main signals acting on plants at high altitudes. According to these studies, such genes are expressed in *C. jubata* in its natural habitat. Their homologs in *Arabidopsis thaliana* (L.) Heynh. (Thal's cress), *Glycine max* (L.) Merr. (soybean), and *Oryza sativa* L. (rice) do not exhibit a similar trend in gene expression at low temperatures.

Constitutive expression and rapid regulation of the mentioned genes suggest the ability of *C. jubata* to tune its cellular machinery to support growth and development in its ecological niche. This molecular and physiological

plasticity allows *C. jubata* to thrive, particularly in the cold, high-altitude desert of the Himalayas (Bhardwaj et al., 2013). The trends in genetic differentiation that we identified in the studied *C. jubata* CPs likely result from the adaptation of this species to various ecological and geographical conditions.

G.P. Semenova (2007) classified *C. jubata* as a medium-potential species when introduced into the forest-steppe zone of Western Siberia. Such species reproduce by seed or vegetatively, require specific soil composition, irrigation, and sites with specific lighting conditions, and flower and fruit annually. However, they are susceptible to frost, and their reproductive capacity may be reduced in unfavorable years. G.P. Semenova (2007) also classified *C. jubata* as a perennial plant, as the lifespan of the introduced population in local conditions reached 17 years.

The extensive range of this species and relatively high genetic diversity suggest that individuals in samples taken from most of the CPs we studied are stable and can serve as source material for the creation of promising introduction populations.

Thus, statistical analysis of genetic variability revealed a significant effect of geographic isolation between genotypes, due to the large distances between the studied *C. jubata* CPs. The influence of altitudinal isolation on genetic distances between individuals from different CPs was also confirmed.

Conclusion

We studied *C. jubata* in samples from eight CPs growing across a wide range of altitudes (over 2,000 m – from 1,575 to 3,680 m) and geographic distances (over 2,500 km) in the mountains of the Pamirs, Tien Shan, and Southern Siberia. ISSR analysis of polymorphic DNA fragments revealed an increasing number of genotypes with unique fragments and the proportion of polymorphic loci in individuals of this species growing in the Tien Shan Mountains at an absolute altitude of 2,550 m, located in the middle of the studied elevation range. We also found a statistically confirmed trend toward increase of genetic distances in samples of representatives from different CPs as geographic distance increase.

The prevalence of intrapopulation genetic variability over interpopulation genetic variability we identified in our studies may indicate the stability of these CPs within the studied parts of this species' range and their good adaptive capacity. The exception was samples of individuals from CP7, which grew in the Republic of Tyva and were subject to anthropogenic impact, where we did not detect any genotypes with unique fragments.

References

- Bhardwaj P.K., Kapoor R., Mala D., Bhagwat G., Acharya V., Singh A.K., Vats S.K., Ahuja P.S., Kumar S. Braving the attitude of altitude: *Caragana jubata* at work in cold desert of Himalaya. *Sci Rep.* 2013;3:1022. doi 10.1038/srep01022
- Chung M.Y., Mmerila J., Li J., Mao K., López-Pujo J., Tsumura Y., Chung M.G. Neutral and adaptive genetic diversity in plants: an overview. *Front Ecol Evol.* 2023;11:1116814. doi 10.3389/fevo.2023.1116814
- Churiulina A.G., Bocharnikov M.V. *Caragana jubata* (Pall.) Poir. (Fabaceae), distribution mapping of rare relict species. *Bot Pacifica.* 2019;8(2):111-114. doi 10.17581/bp.2019.08207
- Churyulina A.G. Geography of rare species (*Caragana jubata* (Pall.) Poir., *Betula lanata* (Regel) V. Vassil.) and communities with their participation in the mountain biomes of the Baikal region: Cand. Sci. (Geogr.) Dissertation. Moscow, 2021 (in Russian)
- Churyulina A.G., Bocharnikov M.V., Ogureeva G.N. Geography of *Caragana jubata* (Pall.) Poir. (Fabaceae) and its phytocenotic role in the mountain vegetation cover. *Vestnik Moskovskogo Universiteta. Seriya 5. Geografiya = Moscow University Bulletin. Series 5. Geography.* 2020;3:108-117 (in Russian)
- de Lafontaine G., Napier J.D., Petit R.J., Hu F.S. Invoking adaptation to decipher the genetic legacy of past climate change. *Ecology.* 2018;99(7):1530-1546. doi 10.1002/ecy.2382
- Dorogina O.V., Zhmud E.V. Molecular-genetic methods in plant ecology. *Contemp Probl Ecol.* 2020;13(4):333-345. doi 10.1134/S1995425520040058
- Dorogina O.V., Kuban I.N., Achimova A.A., Williams N., Lashchinskiy N.N., Zhmud E.V. Morphometric characteristics and genetic ISSR marker variability in *Rhodiola rosea* L. (Crassulaceae) in different ecological and geographic conditions in the Altai Republic. *Int J Molec Sci.* 2023;24(10):15224. doi 10.3390/ijms242015224
- Hamabata T., Kinoshita G., Kurita K., Cao P.-L., Ito M., Murata J., Komaki Y., Isagi Y., Makino T. Endangered island endemic plants have vulnerable genomes. *Commun Biol.* 2019;2:244. doi 10.1038/s42003-019-0490-7
- Hantemirova E.V., Khoreva M.G., Bessonova V.A. Genetic and ecological differentiation of *Caragana jubata* (Pall.) Poir. *Bot Pacifica.* 2024;13(2):15-22. doi 10.17581/bp.2024.13204
- Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics.* 2008;24(11):1403-1405. doi 10.1093/bioinformatics/btn129
- Kakorin P.A., Babenkova I.V., Teselkin Yu.O., Ramenskaya G.V., Demura T.A., Kukes V.G. Hepatoprotective activity of aqueous extract from *Caragana jubata* (Pall.) Poir. shoots in the model of acute hepatitis induced by acetaminophen in rats. *Biomeditsinskaya Khimiya.* 2018a;64(3):241-246. doi 10.18097/PBMC20186403241 (in Russian)
- Kakorin P.A., Tereshkina O.I., Ramenskaya G.V. Potential biological activity and chemical composition of *Caragana jubata* (Pall.) Poir. (review). *Pharm Chem J.* 2018b;52(6):531-535. doi 10.1007/s11094-018-1854-x
- Kakorin P.A., Fateeva T.V., Tereshkina O.I., Perova I.B., Ramenskaya G.V., Sologova S.S., Eller K.I. Antimicrobial activity of lyophilized aqueous extract from *Caragana jubata* (Pall.) Poir. *Pharm Chem J.* 2020;54(3):290-292. doi 10.1007/s11094-020-02193-z
- Kassambara A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.6.0. 2023. <https://rpkgs.datanovia.com/ggpubr/>
- Koropachinsky I.Yu., Vstovskaya T.N. Woody Plants of Asian Russia. Novosibirsk: Geo Publ., 2012 (in Russian)
- Liu C., Wang J., Ko Y.-Z., Shiao M.-S., Wang Y., Sun J., Yuan Q., Wang L., Chiang Y.-C., Guo L. Genetic diversities in wild and cultivated populations of the two closely-related medical plants species, *Tripterygium wilfordii* and *T. hypoglaucum* (Celastraceae). *BMC Plant Biol.* 2024;24(1):195. doi 10.1186/s12870-024-04826-x
- Lysova N.V. Ecology of *Caragana jubata* (Pall.) Poir. in the Central Tien Shan. *Byulleten Glavnogo Botanicheskogo Sada = Bulletin of the Moscow Botanical Garden of the Academy of Sciences.* 1967;64:64-67 (in Russian)
- Magri D., Vendramin G.G., Comps B., Dupanloup I., Geburek T., Gömöry D., Latałowa M., Litt T., Paule L., Roure J.M., Tantau I., Van Der Knaap W.O., Petit R.J., De Beaulieu J.-L. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytol.* 2006;171(1):199-221. doi 10.1111/j.1469-8137.2006.01740.x
- Nabieva A.Yu., Zhmud E.V., Kuban I.N., Dorogina O.V. Morphometric and molecular analysis of *Cypripedium × ventricosum* (Orchidaceae) population in the Novosibirsk Region. *Botanicheskii Zhurnal.* 2020;105(6):597-603. doi 10.31857/S0006813620060058 (in Russian)
- Nørgaard L.S., Mikkelsen D.M.G., Elmeros M., Chriél M., Madsen A.B., Nielsen J.L., Pertoldi C., Randi E., Fickel J., Brygida S., Ruiz-González A. Population genomics of the raccoon dog (*Nyctereutes procyonoides*) in Denmark: insights into invasion history and population development. *Biol Invasions.* 2017;19(5):1637-1652. doi 10.1007/s10530-017-1385-5
- Ohsawa T., Yuji I. Global patterns of genetic variation in plant species along vertical and horizontal gradients on mountains. *Glob Ecol Biogeogr.* 2008;17(2):152-163. doi 10.1111/j.1466-8238.2007.00357.x
- Oorzhak U.S., Lopsan-Endan A.B., Dorzhu-ool Yu.N. Bioecological and chemical features of *Caragana jubata* (Pall.) Poir. *Vestnik Tuvinskogo Gosudarstvennogo Universiteta. Estestvennye i Sel'skokhozyajstvennye Nauki – Vestnik of Tuvan State University. Natural and Agricultural Sciences.* 2024;1(5):16-24. doi 10.24411/2221-0458-2024-01-16-24 (in Russian)
- Oyama K., Ito M., Yahara T., Ono M. Low genetic differentiation among populations of *Arabis serrata* (Brassicaceae) along an altitudinal gradient. *J Plant Res.* 1993;106(2):143-148. doi 10.1007/BF02344418
- Parygin I.A., Khudonogova E.G. Plant communities of *Caragana jubata* (Pallas) Poir. of the southwestern regions of Buryatia. *Vestnik Orenburgskogo Gosudarstvennogo Pedagogicheskogo Universiteta = Vestnik of Orenburg State Pedagogical University.* 2017;4(24):26-33 (in Russian)
- R Core Team. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2024. <https://www.R-project.org/>
- Reisch C., Anke A., Röhl M. Molecular variation within and between ten populations of *Primula farinosa* (Primulaceae) along an altitudinal gradient in the northern Alps. *Basic Appl Ecol.* 2005;6(1):35-45. doi 10.1016/j.baec.2004.09.004
- Safaralikhonov A.B., Aknazarov O.A. The Effect of Environmental Factors of the Pamir Highlands on Plant Life. Dushanbe: Donish Publ., 2021 (in Russian)
- Semenova G.P. Rare and Endangered Species of Siberian Flora: Biology, Conservation. Novosibirsk: Geo Publ., 2007 (in Russian)
- Shi M.-M., Michalski S.G., Chen X.-Y., Durka W. Isolation by elevation: genetic structure at neutral and putatively non-neutral loci in a dominant tree of subtropical forests, *Castanopsis eyrie*. *PLoS One.* 2011;6(6):e21302. doi 10.1371/journal.pone.0021302
- Smouse P.E., Banks S.C., Peakall R. Converting quadratic entropy to diversity: both animals and alleles are diverse, but some are more diverse than others. *PLoS One.* 2017;12(10):e0185499. doi 10.1371/journal.pone.0185499
- Taberlet P., Fumagalli L., Wust-Saucy A.-G., Cosson J.-F. Comparative phylogeography and postglacial colonization route in Europe. *Mol Ecol.* 1998;7(4):453-464. doi 10.1046/j.1365-294x.1998.00289.x
- Talovina G.V. Crop Wild Relatives in the Magadan Province of Russia: Perspectives of Gene Pool Conservation. *Vavilovia.* 2019; 2(3):42-55. doi 10.30901/2658-3860-2019-3-42-55 (in Russian)
- The Red Book of Altay Republic. Plants. Gorno-Altaysk: Gorno-Altayskaya Tipografia Publ., 2017 (in Russian)

- The Red Book of Sakha Republic (Yakutia). Vol. 1. Rare and Endangered Species of Plants and Fungi. Moscow: Reart Publ., 2017 (in Russian)
- The Red Book of Zabaykalskiy Kray. Novosibirsk: Dom Mira Publ., 2017 (in Russian)
- The Red Book of Khabarovsk Kray. Rare and Endangered Species of Plants, Fungi, and Animals. Voronezh: Mir Publ., 2019 (in Russian)
- The Red Book of Magadan Oblast. Rare and Endangered Species of Animals, Plants and Fungi. Magadan: Okhotnik Publ., 2019 (in Russian)
- The Red Book of Irkutsk Oblast. Ulan-Ude, 2020 (in Russian)
- The Red Book of Krasnoyarsk Kray. Vol. 2. Rare and Endangered Species of Plants and Fungi. Krasnoyarsk: Siberian Federal University, 2022 (in Russian)
- The Red Book of Republic of Buryatia. Rare and Endangered Species of Plants and Fungi. Belgorod: Konstanta Publ., 2023 (in Russian)
- Wang Q., Wu X., Yang X., Zhang Y., Wang L., Li X., Qiu Y. Comprehensive quality evaluation of *Lignum Caraganae* and rapid discrimination of *Caragana jubata* and *Caragana changduensis* based on characteristic compound fingerprints by HPLC-UV and HPLC-MS/MS coupled with chemometrics analysis. *Phytochem Anal.* 2020; 31(6):846-860. doi 10.1002/pca.2950
- Wickham H., Averick M., Bryan J., Chang W., McGowan L.D., François R., Golemund G., ... Takahashi K., Vaughan D., Wilke C., Woo K., Yutani H. Welcome to the Tidyverse. *J Open Source Software.* 2019;4(43):1686. doi 10.21105/joss.01686
- Wright S. Breeding structure of populations in relation to speciation. *Am Nat.* 1940;74(752):232-248. doi 10.1086/280891
- Xu S., Li L., Luo X., Chen M., Tang W., Zhan L., Dai Z., Lam T.T., Guan Y., Yu G. *Ggtree*: a serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta.* 2022;1(4):e56. doi 10.1002/imt2.56
- Yang X.-Z., Lyu J.-N., Xu C., Lin Q.-X., Yang J., Wang C., Song P. Cytotoxic chemical constituents from *Caragana jubata* (Pall.) Poir. *J Yunnan Univ Nat Sci Ed.* 2015;37(1):134-139. doi 10.7540/j.ynu.20140364
- Yao N., Gao J., Wang L., Ma Z., Zhang J., Ma F., Xu T. Development of SSR molecular markers and analysis of genetic diversity in *Caragana jubata* (Pall.) Poir. *Genet Resour Crop Evol.* 2025;72(7): 8957-8972. doi 10.1007/s10722-025-02477-x
- Yifru T., Hammer K., Huang X.Q., Röder M.S. Regional patterns of microsatellite diversity in Ethiopian tetraploid wheat accessions. *Plant Breed.* 2006;125(2):125-130. doi 10.1111/j.1439-0523.2006.01147.x
- Zhang M.-L., Xiang X.-G., Xue J.-J., Sanderson S.C., Fritsch P.W. Himalayan uplift shaped biomes in Miocene temperate Asia: evidence from leguminous *Caragana*. *Sci Rep.* 2016;6:36528. doi 10.1038/srep36528
- Zhang X., Korpelainen H., Li C. Microsatellite variation of *Quercus aquifolioides* populations at varying altitudes in the Wolong natural reserve of China. *Silva Fenn.* 2006;40(3):407-415. doi 10.14214/sf.326
- Zhao D., Chen X., Wang R., Pang H., Wang J., Liu L. Determining the chemical profile of *Caragana jubata* (Pall.) Poir. by UPLC-QTOF-MS analysis and evaluating its anti-ischemic stroke effects. *J Ethnopharmacol.* 2023;309:116275. doi 10.1016/j.jep.2023.116275
- Zhmud E.V., Kuban I.N., Achimova A.A., Yamtyrov M.B., Dorogina O.V. Comparative analysis of genetic polymorphism in *Rhaponticum carthamoides* (Asteraceae) populations by ISSR markers in the Altai Republic. *Trudy po Prikladnoy Botanike, Genetike i Selektzii = Proceedings on Applied Botany, Genetics, and Breeding.* 2023;184(2):149-159. doi 10.30901/2227-8834-2023-2-149-159 (in Russian)
- Zhmud E.V., Kuban I.N., Achimova A.A., Papina O.N., Dorogina O.V. Variability of morphometric parameters in *Adonis villosa* Ledeb. (Ranunculaceae) and its genetic differentiation based on ISSR markers in the Altai Republic. *Contemp Probl Ecol.* 2024;17(4):557-565. doi 10.1134/S1995425524700355

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

Received March 12, 2025. Revised October 6, 2025. Accepted November 11, 2025.

doi 10.18699/vjgb-26-23

Variability of organelle genomes in a collection of early maturing soybean varieties

V.V. Aleksandrovich¹ , M.G. Siniauskaya , A.P. Shatarnov¹, O.G. Davydenko²¹ The Institute of Genetics and Cytology of the National Academy of Sciences of Belarus, Minsk, Belarus² Soya-North Co Ltd, Kolodishchi, Minsk oblast, Belarus valeria.alexandrovich@gmail.com

Abstract. Variability of the genomes of cellular organelles (chloroplast and mitochondria) is an important component of the overall variability of the plant genome. A large amount of data has already been obtained on the comparative characteristics of the organization of organelle DNA sequences for different groups of plants. This paper presents new original data on the variability of mitochondrial and chloroplast genomes in soybean (*Glycine max* (L.) Merr.), a crop of great economic importance widely cultivated in Central Europe, including the Republic of Belarus. Initially, we supposed that the peculiarities of soybean organelle DNA sequence or organization promote certain soybean cultivars to be the best maternal and others, alternatively, the best paternal forms. As a result of the study, new complete nucleotide sequences of chloroplast and mitochondrial genomes of 46 soybean samples were obtained by the next generation sequencing method (NGS) on the Illumina platform. A comprehensive bioinformatic comparative study of intraspecific organelle genome variability in 46 soybean varieties of diverse geographical origin was conducted. Polymorphic loci of genomes were discovered. Data on DNA variability were verified by Sanger sequencing. The spectrum of organelle DNA variability of cultivated soybean was represented by three chloroplast DNA haplotypes (C1–C3) and five mitochondrial DNA haplotypes (M1–M5). A comparatively low level of intraspecific variability of organelle genomes in *G. max* was revealed. The soybean chloroplast genome had a lower level of sequence variability than the mitochondrial genome. A set of DNA markers for polymorphic loci of organelle genomes was developed, allowing the differentiation of varieties of the studied group into plasmatypes. Additionally, 90 soybean samples from the collection were studied using PCR followed by Sanger sequencing. The low level of intraspecific variability of organelle genomes in *G. max* was confirmed on the extended group of samples. The majority of cultivars were represented by three plasmatypes – C1/M1, C2/M2 and C1/M3. 46 complete chloroplast DNA sequences have been deposited in NCBI GenBank. The hypothesis that organelle DNA influences the combining ability of different varieties has not yet been confirmed. A more detailed study of the mechanisms of nuclear-cytoplasmic interaction is required, as well as a search for nuclear markers that affect the expression of organelle genes.

Key words: soybean; *Glycine max*; genetic variability; organelles; chloroplasts; mitochondria; NGS

For citation: Aleksandrovich V.V., Siniauskaya M.G., Shatarnov A.P., Davydenko O.G. Variability of organelle genomes in a collection of early maturing soybean varieties. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov J Genet Breed*. 2026;30(2): 205-211. doi 10.18699/vjgb-26-23

Funding. This work was carried out within the framework of the State Program for Scientific Research “Biotechnology 2”, 2021–2025, subprogram “Genomics, Epigenomics, Bioinformatics”, 2.1.3.

Изменчивость оргanelльных геномов в коллекции раннеспелых сортов сои

В.В. Александрович¹ , М.Г. Синявская , О.П. Шатарнов¹, О.Г. Давыденко²¹ Институт генетики и цитологии Национальной академии наук Беларуси, Минск, Республика Беларусь² ООО «Соя-Север Ко», пос. Колодищи, Минская область, Республика Беларусь valeria.alexandrovich@gmail.com

Аннотация. Изменчивость геномов клеточных органелл – хлоропластов и митохондрий – является немаловажной компонентой общей изменчивости генома растений. Получено большое количество данных о сравнительных особенностях организации последовательностей оргanelльных ДНК для различных групп растений. В настоящей работе представлены новые оригинальные данные об изменчивости геномов митохондрий и хлоропластов у сои (*Glycine max* (L.) Merr.), важной хозяйственной и пищевой культуры, широко возделываемой на территории Центральной Европы, в том числе и в Республике Беларусь. Рабочей гипотезой нашего исследования изначально стало предположение: возможно, особенности изменчивости последовательности или структуры ДНК органелл сои определяют способность одних сортов выступать в роли лучших материнских родителей, а других – быть

лучшими отцовскими формами. Получены новые полные нуклеотидные последовательности хлоропластного и митохондриального геномов 46 образцов культурной сои с применением метода секвенирования нового поколения (NGS) на платформе Illumina. Выполнено комплексное биоинформатическое сравнительное исследование внутривидовой изменчивости геномов органелл у 46 сортов сои разнообразного географического происхождения. Выявлены полиморфные локусы геномов. Данные об изменчивости ДНК верифицированы секвенированием по Сэнгеру. Спектр изменчивости оргanelльных ДНК, исследованных методом полногеномного секвенирования сортов сои, представлен тремя гаплотипами хлоропластной ДНК (C1–C3) и пятью гаплотипами митохондриальной ДНК (M1–M5). Обнаружен сравнительно низкий уровень внутривидовой изменчивости геномов органелл у *G. max*. Хлоропластный геном сои обладал меньшим уровнем изменчивости последовательности, чем митохондриальный. Разработан набор ДНК-маркеров к полиморфным локусам геномов органелл, позволяющий дифференцировать сорта сои на плазматипы. Методом ПЦР с последующим секвенированием по Сэнгеру дополнительно изучено 90 образцов сои из коллекции. Низкий уровень внутривидовой изменчивости геномов органелл у *G. max* подтвержден на расширенной группе образцов. Большая часть коллекции была представлена тремя плазматипами – C1/M1, C2/M2 и C1/M3. 46 полных последовательностей хлоропластной ДНК помещено в NCBI GenBank. Гипотеза влияния ДНК органелл на комбинационную способность различных сортов в настоящее время не подтверждена. Требуется более детальное изучение механизмов ядерно-цитоплазматического взаимодействия, поиск ядерных маркеров, влияющих на экспрессию цитоплазматических генов.

Ключевые слова: соя; *Glycine max*; генетическая изменчивость; органеллы; хлоропласты; митохондрии; NGS

Introduction

The study of the organization and function of cellular organelle genomes is a significant area of modern plant genetics. Over the past decades, numerous studies have been dedicated to this topic, addressing issues of taxonomy, phylogeny, nuclear-cytoplasmic coadaptation, and genomic interactions (Gualberto, Newton, 2017; Johnston, 2019; Tsunewaki et al., 2019). The organelle genomes of major agricultural crops are actively studied worldwide (Siniauskaya et al., 2020; Hu et al., 2022; Yue et al., 2023).

Soybean (*Glycine max* (L.) Merr) has been grown in Belarus for many years. In the 1980s, a national school of soybean genetics and breeding originated at the Laboratory of Cytoplasmic Inheritance of the Institute of Genetics and Cytology of the National Academy of Sciences of Belarus. The main focus of soybean breeding in Belarus is the development of early and ultra-early maturing cultivars with sufficiently high yields and seed protein content.

The selection of initial material for hybridization is a key element in the breeding of any crop. Cultivars are known to differ in their ability to produce new hybrids and varieties. Therefore, pedigree analysis and the search for such cultivars are necessary. Based on a study of the pedigrees of hybrids and cultivars developed by other researchers and on our own data, we have identified soybean cultivars that are the most promising as either maternal or paternal parents and are most frequently used in crossbreeding. Given the strictly maternal inheritance of organelles in soybeans, it can be assumed that the observed differentiation of cultivars in their ability to produce successful new gene combinations in hybrids is determined by nuclear-cytoplasmic interactions, possibly due to the organization and expression of organelle genomes.

The availability of collections of original plant breeding material with wide variation in a range of economically valuable traits is a necessary prerequisite for the development of new promising varieties. The study of the phenotypic and genetic variability of collections is one of the fundamental areas of modern breeding, including marker-assisted selection.

The first molecular genetic studies on the variability of the organelle genomes of cultivated and wild soybeans were conducted in the 1980s using restriction fragment length polymorphism (RFLP) analysis. The classification of cultivated and wild soybean lines into three chloroplast DNA (cpDNA) haplotypes and five mitochondrial DNA (mtDNA) haplotypes was the standard at that time (Shoemaker et al., 1986). The RFLP method was used to assess the diversity of organelle genomes in wild soybean (*Glycine soja*) lines (Abe et al., 1999) and soybean varieties cultivated in China (Shimamoto et al., 1998).

The possibility of a more detailed study of the diversity of plant organelle genomes, including those of soybean, using variation in the length of cpDNA microsatellite repeats (SSRs) was demonstrated by W. Powell et al. (1995). This approach enabled the expansion of knowledge on soybean cpDNA variability and the acquisition of qualitatively new data. For example, D. Xu et al. (2002) analyzed six chloroplast microsatellites and identified 52 haplotypes of *G. soja* and eight haplotypes of cultivated soybean, with 75 % of *G. max* lines belonging to a single most common haplotype.

The analysis of the variability of soybean chloroplast genomes began at the Institute of Genetics and Cytology of the National Academy of Sciences of Belarus in the 2000s. As a result of the PCR-RFLP study of the collection of 60 varieties available in the Laboratory of Cytoplasmic Inheritance at that time, three plasmatypes were identified, depending on the absence or presence of *EcoRI* and *ClaI* restriction enzyme recognition sites in cultivars' cpDNA. Most varieties (out of more than 60 studied) had an additional *ClaI* site, three lines had an additional *EcoRI* site, and the cpDNA of two lines lacked *ClaI* and *EcoRI* restriction sites (Siniauskaya et al., 2004). Microsatellite repeats of cpDNA were also studied, and the diversity of soybean varieties collections of various origins was assessed (Aksyonova et al., 2007).

The results of the analysis of the molecular diversity of organelle genomes obtained by a number of author groups have consistently demonstrated that wild soybean has a higher

level of cytoplasmic variability than its cultivated relative: the majority of *G. max* varieties were assigned to one or two major haplotypes (Shimamoto, 2001; Xu et al., 2002). Varieties of Chinese origin were more genetically diverse than those grown in North America and Europe (Yue et al., 2023). The low level of polymorphism of organelle genomes in soybean can be explained by many factors: uniparental inheritance of organelles in soybean through the maternal line, which excludes the possibility of recombination of genomes from two parents; the peculiarities of the propagation of cultivated soybean, for example, strict self-pollination.

The study of the structural features of organelle genomes provides information that can be used to consider (and speculate on) possible microevolutionary changes within the genus *Glycine*. The generally accepted theory is that the cultivated soybean originated monophyletically from the ancestral form *G. soja*. This hypothesis was put forward based on studies of nuclear markers, whole-genome sequencing and nuclear SNP panels (Jeong et al., 2019). However, several studies (Xu et al., 2002; Fang et al., 2016) proposed a theory of repeated processes of soybean domestication. The origin of modern cultivated varieties from multiple primordial maternal lines is indicated by features of the chloroplast and mitochondrial genomes; in particular, in most varieties, organelle genomes belong to one of two main plasmatypes. C. Fang et al. divided all cpDNA haplotypes into two groups and suggested that several maternal lines of wild soybean were involved in the domestication of soybean, giving rise to two groups of plasmatypes in modern varieties (Fang et al., 2016).

The development of next-generation sequencing (NGS) methods has provided fundamentally new information on the organization of plant organelle genomes, including those of soybean, their sequences and variability. In 2023, Y. Yue et al. analyzed NGS data from over 2,000 soybean lines from publicly available databases and found that 69.2 % of all cultivated soybean varieties belong to a single plasmatype, CT1/MT1, while 18.1 % belong to the CT2/MT2 plasmatype, which is consistent with data from earlier studies (Yue et al., 2023).

We previously successfully applied whole-genome sequencing (Illumina platform) to study and assess the diversity of cp and mtDNA in barley (genus *Hordeum*) (Siniauskaya et al., 2020); an approach to analyzing the results of whole-genome sequencing of organelle DNA was developed. In this study, we aimed to use NGS (Illumina) to obtain new data on whole-genome organelle DNA sequences and assess their level of variability using samples from the collection of early and ultra-early maturing soybean varieties of the Laboratory of Cytoplasmic Inheritance of the Institute of Genetics and Cytology of the National Academy of Sciences of Belarus.

Materials and methods

DNA extraction. The study material consisted of organelle and total DNA samples of 46 early maturing soybean varieties from the collection of the Institute of Genetics and Cytology, including varieties of Belarusian breeding (the full list is given in the Supplementary Material)¹. The selected varieties were

successful maternal or paternal forms according to the authors' long-term data (unpublished data).

To isolate organelles using a modification of the method of S.O. Triboush et al. (1998), the first young leaves of 7- to 10-day-old soybean seedlings were used. The isolated organelles were lysed, and the organelle DNA was purified with phenol-chloroform (standard protocol). The quality of the organelle DNA preparation was evaluated by RFLP analysis in 0.8 % agarose gel according to the method of S.O. Triboush et al. (1998).

Total DNA was isolated using phenol-chloroform extraction from the leaves of plants grown in greenhouses or from the first leaves of 7-day etiolated soybean seedlings (standard protocol).

Next-generation sequencing. The organelle DNA was analyzed by paired-end sequencing on an Illumina MiSeq platform using the MiSeq Reagent Kit v3 (600 cycles) (Illumina Inc., USA) following the manufacturer's recommendations. To prepare the DNA library, the Illumina DNA Prep, (M) Tagmentation (24 Samples, IPB) and Nextera XT Index Kit v2 Set A (96 indexes, 384 samples) (Illumina Inc., USA) kits were used following with the manufacturer's recommendations.

NGS data analysis. Whole-genome sequencing data were processed according to a previously developed algorithm (Yermakovich et al., 2020), which included: alignment of reads to reference sequences of the chloroplast and mitochondrial genomes, conversion to bam files and their sorting, generation of VCF files and their filtering. Read alignment files (bam) were visualized using Unipro Ugene and IGV. The chloroplast genome assemblies of the Bragg cultivar (GenBank accession number MW357264) and the mitochondrial genome of the Aiganhuang cultivar (NC020455) were used as reference genomes.

NGS data verification. A set of primers was designed for polymorphic loci of organelle genomes identified after analysis of whole-genome sequencing data to verify sites of variability using Sanger sequencing.

Polymorphic DNA regions were amplified separately in 15 µl of a reaction mixture containing 30–40 ng of sample DNA, 7.5 µl of 2x ArtMix reagent mixture (ArtBioTech LLC, Republic of Belarus), and 1 µl of the corresponding primers (5 pmol/µl) (Primetech ALC, Republic of Belarus). PCR was performed on a C1000 amplifier (Bio-Rad Laboratories, Inc., USA) according to the following protocol: 5 minutes at 95 °C, then 30 cycles, each of which included denaturation at 95 °C for 30 seconds, primer annealing for 30 seconds, and elongation at 72 °C for 25–80 seconds, followed by a final elongation at 72 °C for 5 minutes. The PCR product was identified in 1.5 % agarose gel.

Sanger sequencing of the studied samples was performed on an Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific Inc., USA) using the BrilliantDye Terminator v3.1 Cycle Sequencing Kit (NimaGen B.V., Netherlands), according to the manufacturer's recommendations. Sequencing results were analyzed using Chromas and FinchTV software by comparing the obtained DNA sequences with the *G. max* reference genome (Bragg and Aiganhuang cultivars) from NCBI GenBank.

¹ Supplementary Material is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Alex_Engl_30_2.pdf

Phylogenetic analysis was performed by creating a presence/absence matrix based on cp and mtDNA polymorphisms using Mesquite program (<http://www.mesquiteproject.org>), calculating a distance matrix and constructing a dendrogram in the phylip program (Felsenstein, 1989) based on the neighbor-joining algorithm. Dendrogram visualization was performed using the Iroki online service (Moore, 2020).

Chloroplast genomes assembly. Complete cpDNA sequences of all 46 cultivars used for the study were obtained by manually editing the Bragg chloroplast genome reference sequence to account for the identified polymorphisms.

Results and discussion

NGS of the chloroplast and mitochondrial genomes of 46 cultivated soybean varieties was conducted to study their intra-specific variability. The resulting NGS reads were processed and aligned to the reference sequences (chloroplast genome of the Bragg cultivar and the mitochondrial genome of the Aiganhuang cultivar) to generate VCF files containing information on the differences between the cp and mt genomes of the studied varieties and the reference. As a result of comparative analysis of the VCF files, polymorphic loci in chloroplast and mitochondrial DNA were identified.

Chloroplast genome variability

Nine polymorphic sites were identified in cpDNA. Inter-variety variability in soybean cpDNA was represented by five single-nucleotide polymorphisms (SNPs) and four microsatellite repeat regions. Four out of the five identified SNPs were located in the coding regions of the *atpB*, *rps4*, *accD* and *rps3* genes; all were synonymous. Almost all of the detected SNPs and repeats were located in the large single-copy region (LSC), and only one SNP was in the small single-copy region (SSC).

In a number of studies searching for interspecific and intergeneric variability in the chloroplast genomes of other crops, “hot spots” of cpDNA variability have been identified, such as *ccsA-ndhD*, *trnH-psbA*, *ndhG-ndhI*, *rps18-rpl20*, *rps15-ycf1*, *psbZ-trnG-trnS*, *trnK-rps16*, *trnD-trnY*, *trnW-trnP*, *rpl33-rps18*, *petG-trnW*, *atpB-rbcL* and *rpl32-trnL* (Iram et al., 2019; Mehmood et al., 2020). We did not find any differences in these regions, except for the SSR locus in the intergenic sequence *atpB-rbcL*, apparently due to the taxonomic similarity of the soybean varieties studied.

Of the four polymorphic cpDNA microsatellite loci identified, the SSR locus at position 6,967 (T₁₁–T₁₀) corresponds to the previously described gmcp4 locus (Xu et al., 2002). The remaining microsatellites are newly described and can be used to identify genetic variability in the soybean chloroplast genome.

A single nucleotide substitution at position 82,035 in the coding sequence of the *rps3* gene affects the *Clal* restriction site and was previously described as a polymorphic locus by which soybean varieties are differentiated into two types using RFLP (Kanazawa et al., 1998). This same SNP in the *rps3* gene was included by Y. Yue et al. as a representative locus (position 82,028 in the plastid genome of Zhonghuang 13). This locus, according to the classification of Y. Yue et al., divides

all varieties into two groups: the first group (1,848 lines) is characterized by the presence of an alternative allele at position 82,028 (haplotypes CT1, CT4, CT7, CT22, CT28, CT33 and CT44), the second group has a reference allele (the remaining 732 lines) (Yue et al., 2023).

As a result of our study, nucleotide sequences of chloroplast genomes of 46 soybean cultivars were obtained and published in NCBI GenBank under accession numbers OQ148707–OQ148730 and OR834463–OR834484.

Mitochondrial genome variability

Comparative analysis of mitochondrial genome sequences obtained from whole-genome sequencing of 46 soybean varieties revealed 15 polymorphic loci: three SNPs in noncoding intergenic regions, eight polymorphic SSR loci, three INDELs (insertion/deletion) and one inversion. Only two out of the 15 loci were located in putative coding regions: one mutation in the *orf110c* microsatellite repeat (position 294,729 bp in the reference genome of the Aiganhuang cultivar), and an inversion in *orf160b* (position 205,470).

We identified variability in the intergenic regions *atp6-l-trnK*, *rps3-orf114a*, *orf100c-orf136b*, *trnD-orf114b*, *orf151-orf261*, *atp9-trnM* and in the *nad4* intron of the mitochondrial genome. Similar data were obtained by Y. Yue et al. in 2023, and the polymorphic positions were classified as representative (marker) ones. These are positions 451,199, 284,000, 312,950, 489,992, 256,743, 295,084, and 484,505 in the mtDNA of the Zhonghuang 13 cultivar, which correspond to the regions mentioned above. These polymorphic loci allow differentiation of the MT1 and MT2 haplotypes according to the classification of Y. Yue et al. (2023).

To confirm the detected loci in the cp and mt genomic DNA, a set of primers targeting sites of genomic variation was designed, and a Sanger sequencing study of these polymorphic loci was conducted in 46 accessions. The primers were designed to study the most important marker points of genomic variation used to differentiate individual cp and mtDNA haplotypes in soybean varieties: positions 6,967, 75,657, and 116,598 of the chloroplast genome, and positions 100,092, 197,000, 205,470, and 248,977 of the mitochondrial genome. Primer combinations were also developed to study the most ambiguous or difficult-to-interpret regions of the genome when analyzing bam and VCF files: SSR locus 51,525 of cpDNA and INDELs at positions 158,807 and 321,983 of mtDNA. All differences identified between the varieties were confirmed.

Genetic diversity of soybean organelle genomes

Based on chloroplast genome variability, the studied varieties were divided into three haplotypes: 40 varieties (87 %) were assigned to haplotype I, four varieties (Kitrossa, Lyubasha, Oressa, and Voronezhskaya 31), to haplotype II, and the Legenda and Schara varieties, to haplotype III. The identified haplotypes were designated C1–C3.

Based on 16 polymorphic loci in mitochondrial DNA, five haplotypes were identified, designated M1–M5. Ten INDELs and four SNPs were present in the Lyubasha, Kitrossa, Oressa, and Voronezhskaya 31 varieties (haplotype M2), one INDEL

was found in the related varieties Vasilisa, Amazonka, McCall, Ptsich, and Sahara (haplotype M3). Cultivars Optimus (M5) and Zlata (M4) also differ from other varieties by an inversion and one INDEL, respectively. Most varieties (76 %) belong to the M1 haplotype. Tables 1 and 2 present the differences identified in the organelle genomes and their distribution among the various cp and mtDNA haplotypes.

Based on our data, an analysis of any of the soybean varieties studied by microsatellite repeats No. 1 and 5 allow to distinguish between cpDNA types C1, C2, and C3. These repeats can be recommended for use in studies of genetic diversity in various soybean collections using the chloroplast genome without the use of other cpDNA markers.

Additionally, Sanger sequencing was performed for 90 varieties from the collection using primers differentiating the corresponding haplotypes of cp and mtDNA. Based on the summarized results of whole-genome sequencing and Sanger sequencing, six organelle DNA plasmatypes are identified among the studied varieties of our collection (the so-called northern ecotype): C1/M1, C2/M2, C1/M3, C1/M4, C1/M5 and C3/M1. 65.4 % of soybean varieties (89 out of 136 studied) belong to the C1/M1 plasmatype. 20 (14.7 %) and 23 (16.9 %) varieties have the C2/M2 and C1/M3 plasmatypes, respectively. Three out of 136 varieties (2.2 %) belong to the C1/M4 plasmatype. Plasmatypes C1/M5 and C3/M1 are the rarest, comprising one (0.7 %) and two (1.4 %) varieties, respectively (Table 3).

Based on the phylogenetic analysis, the identified plasmatypes are divided into two clades. One of these clades consists

Table 1. Polymorphic loci of soybean chloroplast DNA and haplotypes identified on their basis

No.	CpDNA polymorphism (relative to MW357264.1)	Haplotype		
		C1	C2	C3
1	6,967 (<i>rbcl-atpB</i>)	(T) ₁₁	(T) ₁₁	(T) ₁₀
2	8,408 (<i>atpB</i>)	G	A	G
3	15,507 (<i>rps4</i>)	T	A	T
4	38,504 (<i>rpoC1</i> intron)	(A) ₄	(A) ₅	(A) ₄
5	51,525 (<i>atpA-trnR</i>)	(A) ₁₉	(A) ₁₃	(A) ₁₉
6	57,873 (<i>accD</i>)	C	A	C
7	75,657 (<i>petD</i> intron)	(T) ₁₄	(T) ₁₅	(T) ₁₄
8	82,035 (<i>rps3</i>)	G	T	G
9	116,598 (<i>ndhA</i> intron)	T	A	T

of plasmatypes C1/M1, C1/M3, C1/M4, C1/M5, and C3/M1, which are most similar to each other in terms of organelle genome sequences. Haplotype C2/M2 constitutes the second clade and has the greatest number of differences from haplotype C1/M1 – eight cpDNA polymorphisms and 13 mtDNA polymorphisms (see the Figure).

Considering that the vast majority of soybean varieties we studied have plasmatype C1/M1, it can be considered the main

Table 2. Polymorphic loci of soybean mitochondrial DNA and haplotypes identified on their basis

No.	MtDNA polymorphism (relative to NC020455.1)	Haplotype				
		M1	M2	M3	M4	M5
1	66,782 (<i>atp6-1-trnK</i>)	(GCTTC) ₂	GCTTC	(GCTTC) ₂	(GCTTC) ₂	(GCTTC) ₂
2	238,088 (<i>orf151-orf261</i>)	(GCTTC) ₂	GCTTC	(GCTTC) ₂	(GCTTC) ₂	(GCTTC) ₂
3	294,604 (<i>orf105b-orf110c</i>)	(GCTTC) ₂	GCTTC	(GCTTC) ₂	(GCTTC) ₂	(GCTTC) ₂
4	321,983 (<i>rps3-orf114a</i>)	(TATAA) ₃	(TATAA) ₂	(TATAA) ₃	(TATAA) ₃	(TATAA) ₃
5	350,927 (<i>orf100c-orf136b</i>)	T	TATAAA	T	T	T
6	100,092 (<i>nad4</i> intron)	GCTCG	(GCTCG) ₂	GCTCG	GCTCG	GCTCG
7	105,562 (<i>trnD-orf114b</i>)	G	T	G	G	G
8	197,000 (<i>orf113-orf160b</i>)	ACCTT	TA	ACCTT	ACCTT	ACCTT
9	221,263 (<i>trnfM-3-orf151</i>)	C	A	C	C	C
10	238,213 (<i>orf151-orf261</i>)	TGTAGT	(TGTAGT) ₂	TGTAGT	TGTAGT	TGTAGT
11	294,729 (<i>orf110c</i>)	TGTAGT	(TGTAGT) ₂	TGTAGT	TGTAGT	TGTAGT
12	333,070 (<i>atp9-trnM</i>)	A	T	A	A	A
13	248,977 (<i>ccmFn-rps4</i>)	T	TTCAC	T	TTCAC	T
14	158,807 (<i>nad4L-1-nad6</i>)	TGCCTA	TGCCTA	(TGCCTA) ₂	TGCCTA	TGCCTA
15	205,470 (<i>orf160b</i>)	GA	GA	GA	GA	TC

Table 3. Differentiation of soybean varieties into plasmatypes based on NGS and Sanger sequencing results

Plasmatype	Soybean variety
C1/M1	Pripyat, Soer 6, Soer 7, Snezhok, Antratsit, Annushka, Mezenka, Yaselda, Adoc, Venus, Gallard, Viliya, Ustyia, Maple Donovan, Lira, Viktoriya, Samer 1, Samer 2 etc.
C2/M2	Oressa, Kitrossa, Luybasha, Lotaro, He Nun 95, Dornburger Stamm, Slavyanka, Volma, Voronezhskaya 31, Violetta, Dobryn', Olvia, Ranitsa, Ros', Selektta 101, Selenia, Pamela, Jing Yuan, Hava, Iney
C1/M3	Sahara, Amazonka, Vasilisa, USHI-6, Tundra, McCall, Ptsich, Viola, AC Colombe, HM 648, Anilin, Barbaro, Belgorodskaya 6, ES Compositor, ES Senator, ES Professor, Lyumaria, OAC Cooper, RGT Ruslana, Siberiya, Ugra, Suedina, Glasier
C1/M4	Zlata, Sharm, Vilana
C1/M5	Optimus
C3/M1	Legenda, Schara

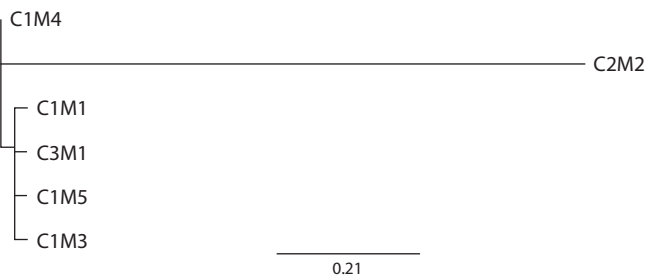
one in the collection. Plasmatypes C1/M3, C1/M4, C1/M5 and C3/M1 are most likely derived from it, since they differ from C1/M1 by only one polymorphic loci in the chloroplast (C3) or mitochondrial genomes (M2–M5).

The C1/M1 and C2/M2 plasmatypes identified by us are consistent with the CT1/MT1 and CT2/MT2 plasmatypes discovered by Y. Yue et al. in 2023. According to their study, these plasmatypes are the most common among cultivated soybean varieties – 69.2 and 18.1 %, respectively (Yue et al., 2023). In our study, we confirmed that the CT1/MT1 plasmatype (C1/M1 according to our classification) is also prevalent among Belarusian varieties. We also identified rare differences not previously described by other researchers. Based on these differences, we identified five additional plasmatypes, of which C1/M3 is more common than C2/M2 in our collection of varieties. This is presumably explained by the geographic and ecological characteristics of Belarus and the selection of varieties that, based on a number of other traits, are most suitable for cultivation in Belarus.

Unfortunately, the fundamental hypothesis of our study regarding the possible organelle genome features that differentiate the best maternal and paternal soybean varieties has not yet been confirmed, as both varieties have similar cp and mtDNA haplotypes. To clarify this issue, a more in-depth study of the subtle mechanisms of interaction between the nucleus and cytoplasm is required.

Conclusion

A comparative analysis of NGS data from 46 early maturing soybean cultivars allowed us to identify polymorphic loci in chloroplast and mitochondrial DNA and assess the intraspecific diversity of organelle DNA in soybean varieties of diverse geographic origin.



Dendrogram of phylogenetic relationships in the studied collection of soybean varieties.

Nine loci in the chloroplast and 15 loci in the mitochondrial genomes of cultivated soybean were identified. The entire spectrum of soybean organelle DNA variability is represented by three chloroplast DNA haplotypes and five mitochondrial haplotypes. Combining variability across both organelle genomes allowed us to identify six organelle DNA plasmatypes and differentiate varieties in the collection accordingly. Using DNA markers for polymorphic loci in organelle genomes developed based on NGS data analysis, we studied plasmon variability in 90 varieties from the collection of Laboratory of Cytoplasmic Inheritance, characterized by a wide range of origins and sensitivity to day length. Forty-six complete soybean chloroplast DNA sequences have been deposited in NCBI GenBank under the accession numbers OQ148707–OQ148730 and OR834463–OR834484.

Our results and analysis of the pedigrees of the varieties confirm the data of other researchers indicating that the C1 and M1 haplotypes are the most common among varieties cultivated in North America, Europe, and the CIS countries. The C1/M1 plasmatype was also detected with the highest frequency in the collection of soybean accessions we studied. The C2 cpDNA and M2 mtDNA haplotypes are characteristic of modern soybean varieties bred in China. Interestingly, the C2/M2 plasmatype was quite common in our collection among a number of accessions with indigenous Chinese maternal forms in their pedigree.

The identification of six plasmatypes indicates a low level of genetic diversity in the studied collection of soybean varieties. The group of early and ultra-early maturing cultivars we studied is an evolutionarily young and very narrow group of varieties (these are varieties of the northern ecotype). Given the targeted selection of these varieties for day length neutrality, a low level of genetic diversity is quite expected in such a limited group of soybean lines (Rosenzweig et al., 2003).

Identifying sources of cytoplasmic variation and varieties with distinct cpDNA and mtDNA haplotypes, as well as introgression of new genetic material into the breeding process are essential for the productive development of new soybean varieties for northern latitudes.

References

Abe J., Hasegawa A., Fukushi H., Mikami T., Ohara M., Shimamoto Y. Introgression between wild and cultivated soybeans of Japan revealed by RFLP analysis for chloroplast DNAs. *Econ Bot.* 1999; 53:285-291. doi 10.1007/BF02866640

- Aksyonova E.A., Rosenzweig V.E., Goloenko D.V., Davydenko O.G. The analysis of belarusian soybean cultivars diversity using nuclear and chloroplast SSR markers. In: Computational Phylogenetics and Molecular Systematics "CPMS 2007". Moscow, 2007
- Fang C., Ma Y., Yuan L., Wang Z., Yang R., Zhou Z., Liu T., Tian Z. Chloroplast DNA underwent independent selection from nuclear genes during soybean domestication and improvement. *J Genet Genomics*. 2016;43(4):217-221. doi 10.1016/j.jgg.2016.01.005
- Felsenstein J. PHYLIP: Phylogeny Inference Package (Version 3.2). *Cladistics*. 1989;5:164-166.
- Gualberto J., Newton K. Plant mitochondrial genomes: dynamics and mechanisms of mutation. *Annu Rev Plant Biol*. 2017;68:224-252. doi 10.1146/annurev-arplant-043015-112232
- Hu Y., Sun Y., Zhu Q., Fan L., Li J. Poaceae chloroplast genome sequencing: great leap forward in recent ten years. *Curr Genomics*. 2022;23(6):369-384. doi 10.2174/1389202924666221201140603
- Iram S., Hayat M.Q., Tahir M., Gul A., Abdullah, Ahmed I. Chloroplast genome sequence of *Artemisia scoparia*: comparative analyses and screening of mutational hotspots. *Plants*. 2019;8(11):476. doi 10.3390/plants8110476
- Jeong S.C., Moon J.K., Park S.K., Kim M.S., Lee K., Lee S.R., Jeong N., Choi M.S., Kim N., Kang S.T., Park E. Genetic diversity patterns and domestication origin of soybean. *Theor Appl Genet*. 2019;132(4):1179-1193. doi 10.1007/s00122-018-3271-7
- Johnston I.G. Tension and resolution: dynamic, evolving populations of organelle genomes within plant cells. *Mol Plant*. 2019;12(6):764-783. doi 10.1016/j.molp.2018.11.002
- Kanazawa A., Tozuka A., Shimamoto Y. Sequence variation of chloroplast DNA that involves *EcoRI* and *ClaI* restriction site polymorphisms in soybean. *Genes Genet Syst*. 1998;73(2):111-119. doi 10.1266/ggs.73.111
- Mehmood F., Shahzadi I., Waseem S., Mirza B., Ahmed I., Waheed M.T. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. *Genomics*. 2020;112(1):581-591. doi 10.1016/j.ygeno.2019.04.010
- Moore R.M., Harrison A.O., McAllister S.M., Polson S.W., Wommack K.E. Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ*. 2020;8:e8584. doi 10.7717/peerj.8584
- Powell W., Morgante M., Andre C., McNicol J.W., Machray G.C., Doyle J.J., Tingey S.V., Rafalski J.A. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Curr Biol*. 1995;5(9):1023-1029. doi 10.1016/s0960-9822(95)00206-5
- Rosenzweig V.E., Goloenko D.V., Davydenko O.G., Shablinskaya O.V., Swiecicki W. Breeding strategies for early soybeans in Belarus. *Plant Breeding*. 2003;122(5):456-458. doi 10.1046/j.1439-0523.2003.00874.x
- Shimamoto Y. Polymorphism and phylogeny of soybean based on chloroplast and mitochondrial DNA analysis. *Jpn Agric Res Q*. 2001;35(2):79-84. doi 10.6090/jarq.35.79
- Shimamoto Y., Fukushi H., Abe J., Kanazawa A., Gai J., Gao Z., Xu D. RFLPs of chloroplast and mitochondrial DNA in wild soybean, *Glycine soja*, growing in China. *Genet Resour Crop Evol*. 1998;45:433-439. doi 10.1023/A:1008693603526
- Shoemaker R.C., Hatfield P.M., Palmer R.G., Atherly A.G. Chloroplast DNA variation in the genus *Glycine* subgenus *Soja*. *J Hered*. 1986;77(1):26-30. doi 10.1093/oxfordjournals.jhered.a110161
- Siniauskaya M.G., Viguro A.V., Rosenzweig V.E., Goloenko D.V. The variability of chloroplast genome of soybean. In: International Scientific Conference on Molecular Genetics, Genomics and Biotechnology (November 24-26, 2004, Minsk, Belarus). Minsk, 2004;346-347 (in Russian)
- Siniauskaya M.G., Makarevich A.M., Goloenko I.M., Pankratov V.S., Liaudanski A.D., Danilenko N.G., Lukhanina N.V., Shimkevich A.M., Davydenko O.G. The study of organelle DNA variability in alloplasmic barley lines in the NGS era. *Vavilov J Genet Breed*. 2020;24(1):12-19. doi 10.18699/VJ19.589
- Triboush S.O., Danilenko N.G., Davydenko O.G. A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. *Plant Mol Biol Rep*. 1998;16:183-189. doi 10.1023/A:1007487806583
- Tsunewaki K., Mori N., Takumi S. Experimental evolutionary studies on the genetic autonomy of the cytoplasmic genome "plasmon" in the *Triticum* (wheat)-*Aegilops* complex. *Proc Natl Acad Sci USA*. 2019;116(8):3082-3090. doi 10.1073/pnas.1817037116
- Xu D., Abe J., Gai J., Shimamoto Y. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. *Theor Appl Genet*. 2002;105:645-653. doi 10.1007/s00122-002-0972-7
- Yermakovich A.E., Siniauskaya M.G., Pankratov V.S., Liaudanski A.D., Halayenka I.M., Shymkevich A.M., Lukhanina N.V., Davydenko O.G. Barley chloroplast and mitochondrial genomes diversity evaluation by NGS of the organelle DNA mixtures. *Vestsi Natsyyanal'nei Akademii Navuk Belarusi. Seriya Biyalagichnykh Navuk = Proceedings of the National Academy of Sciences of Belarus. Biological Series*. 2020;65(3):358-364. doi 10.29235/1029-8940-2020-65-3-358-364 (in Russian)
- Yue Y., Li J., Sun X., Li Z., Jiang B. Polymorphism analysis of the chloroplast and mitochondrial genomes in soybean. *BMC Plant Biol*. 2023;23(1):15. doi 10.1186/s12870-022-04028-3

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

Received June 18, 2025. Revised September 20, 2025. Accepted September 24, 2025.

doi 10.18699/vjgb-26-24

Analysis of *PR-1*, *PR-2*, *PR-10*, chitinases and chitinase-like proteins genes expression in pea roots under the action of salicylic acid and methyl jasmonate

A.M. Egorova  

Kazan Institute of Biochemistry and Biophysics of Kazan Scientific Center of the Russian Academy of Sciences, Kazan, Russia

 egorova@kibb.knc.ru

Abstract. Stress phytohormones – salicylic and jasmonic acids – participate in the plant defense response. The increase in the content of these compounds during plant infection by pathogens leads to the activation of signaling pathways, ultimately resulting in changes in gene expression and protein synthesis, including a group of pathogenesis-related (PR) proteins. Phytohormone-dependent so-called marker genes, are used to assess the activation of these signaling pathways. In this study, *PR-1* genes, which are markers for salicylic acid signaling and expressed in pea roots, were identified and characterized, and the effects of salicylic acid and methyl jasmonate on their expression were analyzed. It was shown that in pea roots, *PR-1*, encoded by the *Psat1g156240* gene, is among the most significantly expressed genes in the control. Salicylic acid did not cause a change in the expression of this gene; however, it was induced by methyl jasmonate after 24 h. Analysis of the expression of other genes encoding *PR-1* proteins showed that salicylate had no effect on their expression after 24 and 72 h. Analysis of the expression of genes encoding chitinases and chitinase-like proteins showed that the former do not exhibit specificity in response to the salicylate and methyl jasmonate, except for the *Psat1g131280*, the expression of which increased after 24 and 72 h of treatment with methyl jasmonate. Genes *Psat1g147600*, *Psat1g147560*, *Psat1g149120* encoding chitinase-like proteins, were barely expressed in pea roots in control, and were specifically induced by salicylic acid. β -1,3-glucanase genes were not induced in roots by the studied phytohormones. The obtained results allowed to reveal specific genes including chitinase-like proteins, the expression of which is salicylate-inducible. These genes can be used for the assessment of the activation of the salicylate-dependent signaling pathway in pea roots.

Key words: marker gene; PR proteins; signaling; salicylic acid; jasmonic acid, chitinase-like proteins; phytoimmunity

For citation: Egorova A.M. Analysis of *PR-1*, *PR-2*, *PR-10*, chitinases and chitinase-like proteins genes expression in pea roots under the action of salicylic acid and methyl jasmonate. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed.* 2026;30(2):212-221. doi 10.18699/vjgb-26-24

Funding. The work was carried out with the financial support of the Russian Foundation for Basic Research and State Institution "Tatarstan Academy of Sciences" grant No. 24-26-20123.

Анализ экспрессии генов *PR-1*, *PR-2*, *PR-10*, хитиназ и хитиназа-подобных белков в корнях гороха под влиянием салициловой кислоты и метилжасмоната

A.M. Egorova  

Казанский институт биохимии и биофизики – структурное подразделение Федерального исследовательского центра «Казанский научный центр Российской академии наук», Казань, Россия

 egorova@kibb.knc.ru

Аннотация. В реализации защитного ответа растений принимают участие стрессовые фитогормоны – салициловая и жасмоновая кислоты. Повышение содержания этих соединений при инфицировании растений патогенами приводит к активации сигнальных путей, в конечном итоге изменяющих экспрессию генов и синтез белков, среди которых выделяется группа индуцируемых, связанных с патогенезом белков (PR). Для оценки активации этих сигнальных путей используются фитогормон-зависимые, так называемые маркерные гены. В настоящей работе выявлены и охарактеризованы экспрессирующиеся в корнях гороха гены маркерных для салицилатного сигналинга белков *PR-1*. Проанализировано влияние салициловой кислоты и метилжасмоната на их экспрессию. Показано, что в корнях гороха ген *Psat1g156240*, кодирующий *PR-1*, входит в число наиболее значительно экспрессирующихся генов. Салициловая кислота не вызывала изменение экспрессии этого гена,

в то же время он индуцировался метилжасмонатом через 24 ч. Анализ экспрессии других генов, кодирующих белки PR-1, показал, что салицилат не влиял на их экспрессию через 24 и 72 ч. При анализе экспрессии генов хитиназ и хитиназа-подобных белков первые не проявляли специфичность ответа на действие салицилата и метилжасмоната, за исключением гена *Psat1g131280*, экспрессия которого повышалась через 24 и 72 ч при действии метилжасмоната. Гены хитиназа-подобных белков, *Psat1g147600*, *Psat1g147560*, *Psat1g149120*, практически не экспрессировались в корнях гороха контрольного варианта и специфично индуцировались салициловой кислотой. Гены β -1,3-глюканаз не индуцировались исследованными фитогормонами в корнях. Полученные результаты позволили определить конкретные гены, кодирующие хитиназа-подобные белки, экспрессия которых индуцируется салициловой кислотой. Эти гены могут быть использованы для оценки активации салицилат-зависимого сигнального пути в корнях гороха.

Ключевые слова: маркерный ген; PR-белки; сигналин; салициловая кислота; жасмоновая кислота; хитиназа-подобные белки; фитоиммунитет

Introduction

Throughout their lifestyle, plants are constantly exposed to pathogens and unfavorable environmental conditions. Phytoimmunity enables plants to cope with pathogen attacks, overcome these unfavorable conditions, and survive. Furthermore, plants have evolved a wide array of defense mechanisms in response to stressors and pathogen attacks. These include pathogen-specific resistance genes, antimicrobial compounds such as phytoalexins, and pathogen-induced defense proteins (de Wit, 2007).

Salicylic acid (SA) and jasmonic acid (JA) are key phytoimmunity factors involved in the realization of plant defense responses (Ding P., Ding Y., 2020; Wang et al., 2021). SA is generally associated with defense against biotrophic and hemibiotrophic pathogens, which feed on living plant tissues, whereas JA is primarily involved in defense against necrotrophic pathogens and herbivorous insects (Glazebrook, 2005). Pathogen attack on plants leads to the production of signaling molecules and the activation of signaling pathways involving one of the immunity factors mentioned above (Ding L.N. et al., 2022). Activation of these signaling cascades alters gene expression and protein synthesis. These proteins either have direct antimicrobial properties or are involved in the synthesis of defense compounds.

When studying plant responses to various factors, it is often necessary to assess the activation of specific signaling pathways to understand the mechanisms underlying the response. This is most commonly achieved by analyzing the activation of expression of so-called marker genes. These genes have primarily been characterized in model plants such as *Arabidopsis*, tobacco, and rice, and mainly in aboveground organs. Nevertheless, results obtained from model plants are not always relevant to other plant species, particularly to significant agricultural crops. For example, research on *Brassica rapa* has shown that widely used marker genes for salicylate- and jasmonate/ethylene-dependent signaling pathways lack specificity in their expression in response to these hormones, and their expression levels differ between shoots and roots (Papadopoulou et al., 2018). In another study on *Capsicum annuum*, the *PR-1* gene, a marker for salicylate signaling, was found to be a more suitable candidate for the JA/ethylene-dependent pathway rather than the SA pathway (Perez-Aranda et al., 2024). In *Brachypodium distachyon*, analysis of *BdPRI-4* and *BdPRI-5* gene expression revealed that they are regulated by both SA and JA (Kouzai et al., 2016).

In the post-genomic era, advances in omics technologies enable the generation and analysis of large datasets, including gene expression and protein synthesis data. The available pool of transcriptomic data allows for the analysis of the full spectrum of genes expressed in a specific organ, both constitutive ones and those altered under various conditions. The identification and characterization of such genes, which play a role in plant adaptation, can be useful for breeding of resistant cultivars (Tyagi et al., 2022).

Pathogen-induced proteins with antipathogenic activity (Pathogenesis-Related or PR proteins) play a crucial role in the realization of plant defense responses. Based on their functional characteristics, they are classified into 17 classes. Many PR proteins are induced not only by pathogens but also by stress-related phytohormones (van Loon et al., 2006). Salicylate-dependent genes are important in mediating plant defense responses (Maleck et al., 2000; Blanco et al., 2009). One of the most well-known marker genes for salicylate-dependent signaling is *PR-1* (Breen et al., 2017). Alongside with *PR-1*, other members of the *PR* gene family – such as β -1,3-glucanases (*PR-2*) (Hennig et al., 1993; Jayaraj et al., 2004) and chitinases (*PR-3*, 4, 8, 11) (Davis et al., 2002; Tarchevsky et al., 2010) – are also regulated by SA.

In recent years, significant attention has been paid to the processes occurring in the rhizosphere and to the interactions between plant roots and pathogens, symbiotic bacteria, and other soil inhabitants. The crop yield also depends on these relationships, since infection of roots with pathogens and the efficiency of symbiosis formation with nodule bacteria (for legumes) significantly impact plant growth and yield (Solomon et al., 2024). Furthermore, roots play a crucial role in plant life not only as “providers” of water and minerals but also as active participants in defense responses. Certain compounds with antipathogenic properties, such as alkaloids and terpenoids, are synthesized in the roots and transported to aboveground organs. In addition to this, roots implement their own specific defense mechanisms (Erb, 2012).

Previously, in our work, we observed a difference in the proteomic response of pea leaves and roots to exogenous SA application, which may also reflect a difference in their defense responses. In pea leaves, SA upregulates the content of several PR proteins – chitinases and β -1,3-glucanases (Tarchevsky et al., 2010) – whereas in the roots, the content of chitinase-like proteins belonging to glycoside hydrolase family 18 (GH18) increased the most significantly (Egorova, 2025). Obviously,

the differing environmental conditions of aerial organs and roots effect the specifics of their responses to various stimuli. Therefore, identifying the characteristics of root responses, including those to pathogen attacks, is important as the data can be utilized in breeding cultivars resistant to soil-borne pathogens and in developing agrobiotechnology approaches to protect specific crops from soil pathogens. When studying the features of root defense responses and assessing the activation of specific signaling pathways, the use of adequate marker genes is essential. In this context, the search for, characterization of, and understanding of the expression patterns and specificity of these genes in response to particular signaling compounds represents an important task.

In this study, we identified and characterized genes encoding *PR-1*, chitinases and chitinase-like proteins, β -1,3-glucanases, and *PR-10* in the *Pisum sativum* roots. We analyzed the expression of these genes in the control and in response to SA and methyl jasmonate (MeJA). Gene expression was assessed using comparative transcriptomic analysis of control, SA-treated, and MeJA-treated pea roots.

Materials and methods

Plant growth and treatment. Pea (*P. sativum* L.) seeds (cultivar Tan) were surface-sterilized in 0.05 % potassium permanganate solution for 20 minutes and then thoroughly rinsed with distilled water. The seeds were germinated on moist filter paper in the dark for 3 days. The seedlings were grown in 1/4-strength Hoagland–Arnon liquid medium under a 16-hour photoperiod at 25 °C. Eight-day-old seedlings were used for subsequent experiments. For the experimental treatments, 50 μ M SA (Sigma, USA) or 20 μ M MeJA (Serva, Germany) was added to the growth medium. The compounds were first dissolved in 500 μ l of ethanol. The control group consisted of seedlings grown on 1/4-strength Hoagland–Arnon solution supplemented with a corresponding volume of ethanol, calculated to account for dilution. Pea roots were harvested and frozen in liquid nitrogen at 24 and 72 h post-treatment.

Transcriptome analysis. Total RNA was isolated from 200 mg of pea roots using the ExtractRNA reagent (Evrogen, Russia) according to the manufacturer's protocol. RNA concentration and quality were assessed using an Implen Nano-Photometer (Implen, GmbH) and by 1 % agarose gel electrophoresis. The preparation of cDNA libraries, high-throughput sequencing, and bioinformatic analysis of the transcriptomic data were performed by Novogene (Beijing, China).

RNA quality was verified using the Agilent Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). mRNA was enriched from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in First Strand Synthesis Reaction Buffer (X5) (Thermo Fisher, USA). First-strand cDNA was synthesized using random hexamer primers and M-MuLV Reverse Transcriptase (Thermo Fisher, USA). Second-strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. After adenylation of the 3' ends of the DNA fragments, Adaptor with hairpin loop structure was ligated to prepare for hybridization.

In order to select cDNA fragments preferentially 370–420 bp in length, the library fragments were purified with the AMPure XP system (Beckman Coulter, USA). PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system (Agilent, USA).

After clusters were generated on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HSv3-cBot-HS (Illumina, USA), the library preparations were sequenced on an Illumina NovaSeq 6000 platform (Illumina, USA), generating 150 bp paired-end reads. Raw data were processed using Perl scripts. At this stage, reads containing adapters, poly-N sequences, and low-quality reads were removed to obtain clean data. Quality metrics Q20, Q30, and GC content were calculated. All downstream analyses were based on these high-quality clean data. The *P. sativum*_v1a genome (Kreplak et al., 2019) was used as the reference (<https://urgi.versailles.inra.fr/Species/Pisum>). The reference genome was indexed, and the clean reads were aligned to it using Hisat2 v2.0.5.

The number of reads mapped to each gene was counted using feature Counts v1.5.0-p3. The FPKM (Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) for each gene was calculated based on the gene length and the mapped read count. FPKM simultaneously accounts for the effects of sequencing depth and gene length on the read count (Trapnell et al., 2010).

Differential gene expression analysis was performed using the DESeq2 R package (1.20.0) (Love et al., 2014). The resulting *p*-values were adjusted using the Benjamini and Hochberg approach for controlling the false discovery rate. Genes with an adjusted *p*-value ≤ 0.05 and an absolute log2 fold change ≥ 1 were considered differentially expressed. A minimum of three biological replicates was used for each condition.

In this study, gene expression levels are presented as FPKM values, as not all analyzed genes were differentially expressed. Figures display FPKM values with standard deviations. Differentially expressed genes are indicated in the figures with an asterisk (*). The sequencing data have been deposited in the NCBI SRA (BioProject PRJNA1358375).

Bioinformatic analysis. BLAST searches for predicted pea genes and protein sequences were performed against the *P. sativum* genome_v.1a (Kreplak et al., 2019) in the Ensemble-Plants database (https://plants.ensembl.org/Pisum_sativum/Info/Index). Multiple alignment of protein sequences was conducted using the MEGA v.10.2.6 software (Kumar et al., 2018). Full-length amino acid sequences for multiple alignments were retrieved from the UniProt and NCBI databases. Protein domain organization was analyzed and predicted using the Conserved Domains Database (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>).

Results

A transcriptomic analysis of pea seedling roots of control plants and those treated with SA (24 and 72 h) and MeJA (24 and 72 h) was performed. Each sample contains between 36 and 57 million reads. The quantity, quality, and characteristics of the obtained data are presented in the Supplementary Material¹.

¹ Supplementary Material is available at: <https://vavilovj-icg.ru/download/pict-2026-30/appx15.xlsx>

Alignment of the reads to the reference pea genome (*P. sativum*_v1a) (Kreplak et al., 2019) identified approximately 20 thousand transcripts. After 24 h, SA altered the expression of 1,476 genes, of which 635 were upregulated and 841 were downregulated. After 72 h of SA treatment, the expression of 2,420 genes was altered, with 977 upregulated and 1,443 downregulated. MeJA treatment for 24 h changed the expression of 3,384 genes (1,546 upregulated, 1,838 downregulated). After 72 h, MeJA altered the expression of 5,462 genes, with 2,218 upregulated and 3,244 downregulated ones.

This study analyzed the expression of *PR-1*, chitinase and chitinase-like protein, β -1,3-glucanase, and *PR-10* genes in pea roots to assess their potential as markers for salicylic acid signaling and to identify specific isoforms induced specifically by SA. These gene groups were selected because *PR-1*, chitinases, β -1,3-glucanases, and *PR-10* have been shown to be induced by SA (Davis et al., 2002; Li et al., 2005; Spoel et al., 2007; Xu et al., 2024) and to possess antipathogenic activity (van Loon et al., 2006), while chitinase-like proteins are significantly induced by SA in pea roots (Egorova, 2025). The expression of the identified genes was evaluated under the influence of both SA and MeJA, as both hormones are involved in plant defense responses and determine the features of the implemented defense strategy.

Identification and expression analysis of *PR-1* genes in pea roots

PR-1 belongs to the family of cysteine-rich secretory proteins (CAPs), and pea genome contains 12 sequences putatively assigned to this family. A BLAST search in the pea genomic database using characterized *PR-1* sequences from tobacco (P09042) and *Arabidopsis* (AT2G14610) revealed similar sequences in pea. Analysis of the expression of the revealed *PR-1* genes in the pea genome showed that nine CAP genes are expressed in the control roots (Fig. 1a). The expression of genes *Psat7g104840*, *Psat6g069320*, and *Psat1g21192* was low (<10 FPKM), and they are not discussed further in this work.

To characterize the pea proteins, a multiple alignment of the amino acid sequences of *PR-1*s, encoded by genes expressed in roots, with annotated proteins from *Arabidopsis* and tobacco, was performed. The protein encoded by the most highly expressed pea roots gene, *Psat1g156240*, shares 60 % amino acid sequence identity with the *PR-1* protein P09042 from *Nicotiana tabacum* and with AtPR1 (AT2G14610) from *Arabidopsis*. The protein encoded by the gene *Psat1g213200* shares 66.4 % identity with PR1A_TOBAC (P08299) and 65 % with AtPR1. CAP proteins are characterized by the presence of conserved CAP motifs (CAP1–CAP 4) and a CAPE peptide (Han et al., 2023). The two *PR-1* proteins encoded by genes *Psat1g211480* and *Psat0s4615g0040* have a truncated CAPE peptide and lack the CAP2 motif (Fig. 1b), while other conserved motifs are present in their sequences.

In the control, the expression of *Psat1g156240* and *Psat1g213200* was the highest compared to other *PR-1* genes. The effect of SA and MeJA on the expression of *PR-1* genes was assessed (Fig. 1a). The analysis showed

that *Psat1g156240* is upregulated by MeJA at 24 h. SA did not cause a significant change in the expression of this gene. The expression of the *Psat1g213200* gene was not changed significantly by SA but decreased under the action of MeJA (Fig. 1c). The expression levels of the *Psat1g211480* and *Psat0s4615g0040* genes in the control were lower compared to the two previous genes and did not change in response to either SA or MeJA at the studied time points. The expression of other CAP genes – *Psat0ss9363g0160*, *Psat4g090120*, *Psat2g163240*, *Psat1g211840*, and *Psat0ss9363g0200* – was low in the control, did not change under the action of SA, but was downregulated by MeJA (Fig. 1a).

Identification and analysis of chitinase and chitinase-like protein gene expression in pea roots

Chitinases (E.C. 3.2.1.14) belong to glycosyl hydrolase families 18 (GH18) and 19 (GH19). The chitinase genes *Psat3g016800*, *Psat1g084840*, *Psat2g145680*, *Psat3g016400*, *Psat5g291720*, *Psat2g128240*, *Psat4g175600*, *Psat4g175720*, *Psat0s794g0040*, *Psat5g256760*, *Psat5g298520*, *Psat1g169040*, *Psat4g171240*, *Psat4g177760*, *Psat5g020240*, *Psat6g166120*, *Psat6g194800* are either not expressed in roots or their expression level is low (<10 FPKM) and are therefore not discussed further in this work (see the Supplementary Material).

Transcriptomic data analysis revealed that six chitinase genes encoding proteins from the GH18 family and two genes from the GH19 family are expressed in pea roots, which in turn belong to different PR protein groups (Fig. 2a). All GH19 chitinases expressed in pea roots contain a lysozyme-like domain, including the most highly expressed gene *Psat5g000800* (Fig. 2a). SA and MeJA did not cause a significant change in the expression of this gene. The expression of the second GH19 chitinase gene, *Psat4g103520*, which was significantly lower than that of the first gene, also did not change under the action of SA or MeJA.

The expression of genes encoding GH18 family proteins in the control was lower compared to the GH19 genes. Chitinases belonging to PR-3 contain a hevein-like domain. The expression of the genes *Psat3g005200* and *Psat6g194680*, belonging to acidic endochitinases, did not change under salicylate treatment, but the expression of *Psat3g005200* decreased under the action of MeJA after 72 h. The expression of *Psat6g194680* did not change under SA or MeJA treatment (Fig. 2a). The expression of *Psat1g131280*, annotated as hevamine A, increased thousands of times under the action of MeJA after 24 and 72 h (Fig. 2). The expression of *Psat0s5923g0040*, also belonging to GH18, was slightly increased by SA and MeJA after 24 h (Fig. 2a).

Another member of the GH18 protein family, the product of the gene *Psat3g007040*, contains a barwin domain in addition to the chitin-binding domain. Proteins with this domain are classified as PR-4 proteins. Neither SA nor MeJA influenced the expression of this pea chitinase representative (Fig. 2a).

In the previous study (Egorova, 2025), significant induction of a group of proteins belonging to the GH18 family in pea roots by salicylate was shown. The identified proteins

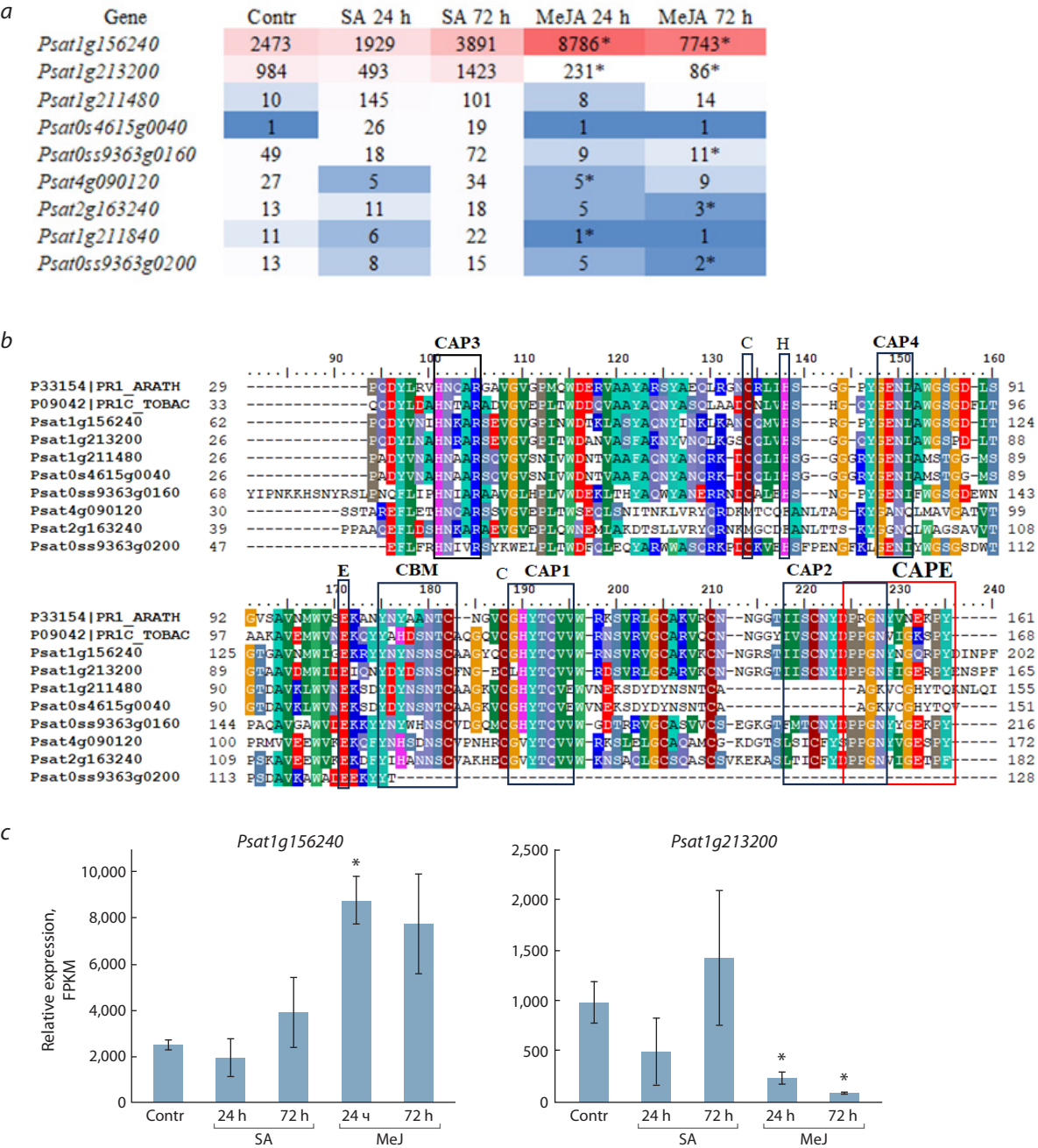


Fig. 1. Cysteine-rich secretory proteins whose genes are expressed in pea roots.
a – list of CAP proteins (PF 00188) expressed in pea roots in the control (Contr) and under the action of SA (50 μ M) and MeJA (20 μ M) for 24 and 72 h. The color scale represents highly expressed (red) and weakly expressed (blue) genes; the expression level is shown in FPKM; *b* – multiple sequence alignment of PR-1 proteins from *Arabidopsis* (P33154), tobacco (P09042), and CAP proteins of pea roots. Black rectangles indicate conserved CAP1–CAP4 domains, conserved cysteine, histidine, glutamic acid residues, and the caveolin-binding motif (CBM) responsible for sterol binding. The CAPE peptide is highlighted in red; *c* – differentially expressed CAP genes in pea roots under SA and MeJA treatment.
Here and in Figures 2–4: data represent $\pm$ SD mean values of three to five biological replicates. * Indicates significant differences in gene expression in SA- and MeJA-treated samples compared to the control ($p < 0.05$ and $|\log_2 FC| \geq 1$).

are encoded by the genes *Psat1g147560*, *Psat1g147600*, *Psat1g149120* and *Psat1g148600*. Analysis of the chitinase activity of these proteins revealed that they do not degrade chitin and are thus classified as chitinase-like proteins (Egorova, 2025). A previous transcriptomic analysis of the effect of SA on these genes showed that the expression of all four genes was significantly increased by SA treatment, while being low in the control (Egorova, 2025). In the present work, we analyzed the influence of MeJA, which had no effect on the expression of any of the four genes studied (Fig. 3).

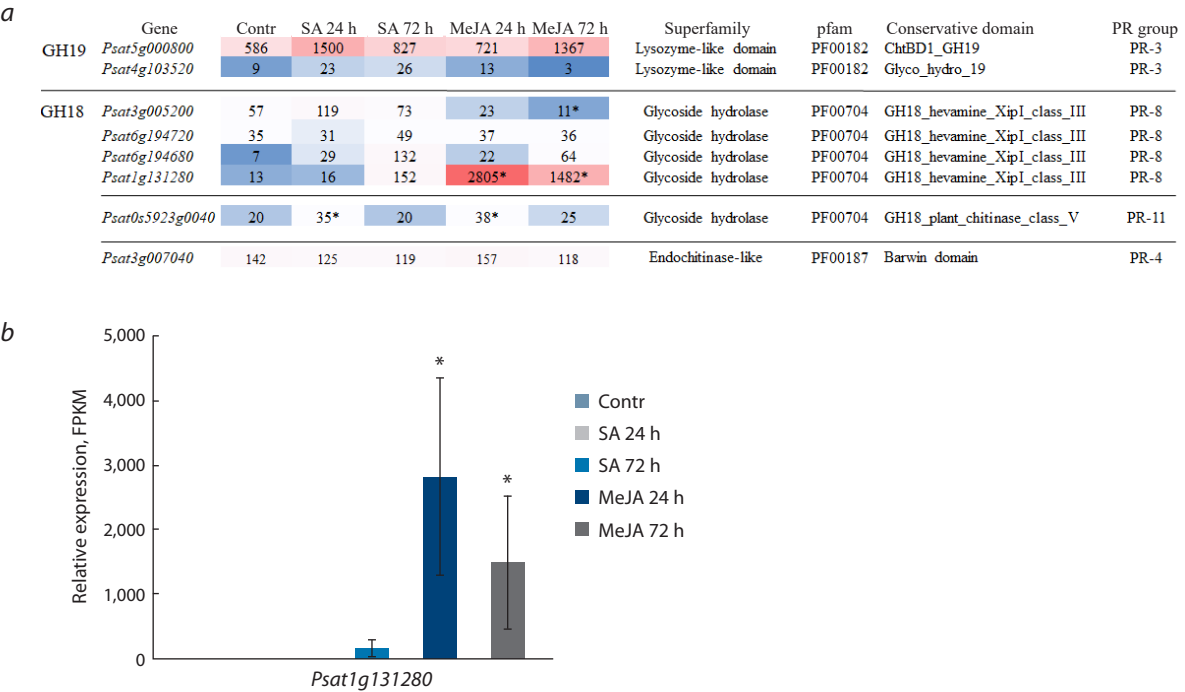


Fig. 2. Effect of salicylic acid and methyl jasmonate on the expression of chitinases in pea roots.
a – List of GH18 and GH19 chitinase genes expressed in pea roots in the control (Contr) and under the action of SA (50 μ M) and MeJA (20 μ M) for 24 and 72 h. The color scale represents highly expressed (red) and weakly expressed (blue) genes; expression level is shown in FPKM; *b* – expression of the *Psat1g131280* gene in pea roots under the action of SA and MeJA.

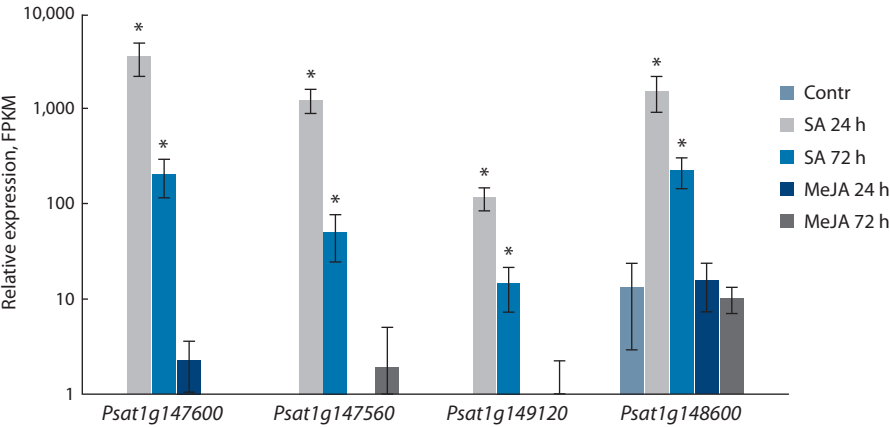


Fig. 3. Expression of chitinase-like protein genes in pea roots under the action of SA (50 μ M) and MeJA (20 μ M) for 24 and 72 h (FPKM).

Identification and analysis
of PR-2 and PR-10 gene expression in pea roots

Our previous work demonstrated the induction of β -1,3-glucanases (PR-2) and a disease resistance response protein (PR-10) in pea leaves by SA (Tarchevsky et al., 2010). A proteomic analysis of SA-treated pea root proteins showed an increased abundance of disease resistance response proteins and ABA-responsive proteins belonging to the PR-10 group, but we did not detect changes in the content of β -1,3-glucanases (Tarchevsky et al., 2010). The pea genome was searched for genes encoding these proteins, and their expression was ana-

lyzed based on transcriptomic data. In this part of the study, we focused on the most highly expressed genes in pea roots (>100 FPKM), as genes with lower expression levels did not show significant changes in response to phytohormone treatment (see the Supplementary Material). More than 50 genes encoding β -1,3-glucanases were identified in the pea genome, of which approximately 20 were expressed in roots at varying levels (see the Supplementary Material). *Psat1g161880* was highly expressed in the control, and its expression did not change in response to SA or MeJA (Fig. 4). The expression of *Psat3g165080* was independent

	Gene	Contr	SA 24 h	SA 72 h	MeJA 24 h	MeJA 72 h	Annotation
PR-2	<i>Psat1g161880</i>	522	310	460	135	174	Glucan endo-1,3-beta-glucosidase
	<i>Psat3g165080</i>	83	40	124	59	39	Glucan endo-1,3-beta-glucosidase 11
	<i>Psat7g179640</i>	3	145*	27	2	9	Glucan endo-1,3-beta-glucosidase
PR-10	<i>Psat4g180800</i>	4654	4696	5324	4000	3280	Pathogenesis-related protein Bet v I family
	<i>Psat1g156920</i>	2405	2555	2784	1744	2675	ABR17
	<i>Psat1g157240</i>	567	1247	1627*	422	579	Disease resistance response protein Pi176
	<i>Psat1g157360</i>	362	1065	841	327	293	Disease resistance response protein Pi49
	<i>Psat1g157200</i>	118	418	252	198	48	Disease resistance response protein Pi49

Fig. 4. List of some most highly expressed (>100 FPKM) *PR-2* (β -1,3-glucanase) and *PR-10* genes in pea roots in the control (Contr) and under the action of SA (50 μ M) and MeJA (20 μ M) for 24 and 72 h. The color scale represents highly expressed (red) and weakly expressed (blue) genes; expression level is shown in FPKM.

of SA and MeJA. The gene *Psat7g179640* was practically not expressed in the control but was activated by SA after 24 h (Fig. 4).

More than 10 genes encoding PR-10 proteins are expressed in pea roots. One of the representatives of this group, encoded by *Psat4g180800*, was the most highly expressed gene in pea roots. Neither SA nor MeJA significantly affected the expression of this gene. The *Psat1g156920*, the product of which is annotated as an ABA-responsive protein 17, was also highly expressed, and its expression was not altered by the studied phytohormones. After 72 h, SA upregulated expression of *Psat1g157240*, the product of which is annotated as a disease resistance response protein (Fig. 4).

Discussion

When working with non-model plants, which include the majority of agricultural species, discrepancies with results obtained on model plants arise. In particular, data on marker genes, the activation of expression of which is associated with the onset of specific phytohormone-dependent signaling pathways, were established using model plants like *Arabidopsis* and tobacco. This is important in studying plant defense mechanisms, as the findings can be applied to develop defense strategies for these species and in the breeding of resistant cultivars.

The main object of our research is pea, and we have previously studied the effect of phytoimmunity factors – salicylic and jasmonic acids – on the pea roots protein spectra, identifying the most significant changes in protein content (Tarchevsky et al., 2010; Yakovleva et al., 2013; Egorova, 2025). This study analyzed the expression of certain classes of *PR* genes and chitinase-like protein genes to assess their response to salicylate and MeJA, since we were faced with the need to choose adequate marker genes to evaluate the activation of the SA-dependent pathway in pea roots. The hormone treatment time points were based on our prior proteomic studies showing changes in the protein spectra at 72 h, and our interest in early events following the activation of phytohormone-dependent defense mechanisms.

First, we analyzed the expression of the widely accepted SA-dependent marker gene *PR-1* and its paralogs in pea

roots. The PR-1 protein was first isolated from TMV-infected tobacco leaves and could constitute up to 2 % of total protein in infected leaves (Alexander et al., 1993). However, not all PR-1 protein isoforms were induced upon infection. Early studies showed that PR-1 proteins participate in plant growth and developmental processes unrelated to pathogenesis – they accumulate in senescing plants leaves (Fraser, 1981) and in sepals of developing flowers (Lotan et al., 1989). Several studies demonstrated that overexpression of *PR-1* enhances resistance to fungi (Sarowar et al., 2005) and bacteria (Kiba et al., 2007), with transgenic plants showing increased pathogen resistance (Ali et al., 2018). The protective functions of PR-1 proteins may be linked to their ability to interact with sterols and the presence of the CAPE peptide in their structure (Breen et al., 2017; Han, 2023). The CAPE peptide was shown to be induced in plants by mechanical wounding and MeJA treatment, cleaved from the N-terminus of PR-1 proteins at a conserved CNYx sequence, and to activate the expression of proteinase inhibitor genes, *PR-1*, *PR-7*, and *ERF-5* (Chen et al., 2014). The most highly expressed *PR-1* genes in pea roots, *Psat1g156240* and *Psat1g213200*, contain the CAPE peptide and the CNYx sequence (Fig. 1b) and the expression of *Psat1g156240* was upregulated by MeJA at 24 h (Fig. 1c). This gene is used by researchers as an SA-dependent marker in pea leaves (Barba-Espín et al., 2011). Our results indicate that, at least in pea roots, this gene cannot be used to assess the activation of salicylate signaling. Conversely, the proteins encoded by *Psat1g211480* and *Psat0s4615g0040* contain a truncated CAPE peptide and lack the CNYx sequence. In pea roots, these genes are expressed at low levels and do not alter their expression in response to SA at the studied time points (Fig. 1a, b). It is possible that activation of these genes occurs earlier than 24 h. Activation of *PR-1* genes by MeJA and other stimuli has been demonstrated previously. Activation of the *AtPRB1* gene by JA/ethylene was shown in *Arabidopsis* roots (Santamaria et al., 2001). In rice, various *PR-1* genes responded to SA and MeJA in a distinct manner, and organ-specific variation in their expression levels was observed (Mitsuhashi et al., 2008).

Chitinases (EC 3.2.1.14) belongs mainly to glycosyl hydrolase families 18 and 19, and degrade chitin, a cell walls

component in many pathogenic microorganisms (Davies, Henrissat, 1995). Since chitinases were among the most significantly induced proteins in pea leaves under SA treatment (Tarchevsky et al., 2010), we analyzed expression of chitinase genes in roots. Furthermore, induction of chitinase transcripts by salicylate was demonstrated in pine (Davis et al., 2002).

According to transcriptomic data, the gene *Psat5g000800* is highly expressed in pea roots, encodes a protein belonging to the GH19 family, and is characterized by the presence of a lysozyme-like domain. The expression of this gene is not sensitive to SA or MeJA (Fig. 2). It is suggested that GH19 chitinases play an important role in defense against soil-borne pathogens and are secreted into the rhizosphere (De-la-Peña et al., 2010). The number of expressed chitinase genes in pea roots encoding GH18 proteins is greater compared to GH19, but their expression levels are lower than those of GH19 chitinase genes. Moreover, they do not show specificity in regulation by the studied phytohormones, except for *Psat1g131280*, the expression of which is significantly upregulated by MeJA (Fig. 2b). These data indicate that the chitinase isoforms expressed in pea roots cannot be used to assess the activation of the SA-dependent pathway.

Alongside genes encoding active chitinases, plant genomes contain genes encoding inactive chitinase-like proteins. During evolution, these proteins lost the ability to hydrolyze chitin due to amino acid substitutions in catalytically important domains and features of tertiary structure formation (Kesari et al., 2015). We identified significant activation of four chitinase-like protein genes in pea roots by SA (Fig. 3). Moreover, these genes were salicylate-dependent and not sensitive to MeJA. This makes them suitable candidates to use as markers of the SA-dependent pathway in pea roots, with the exception of *Psat1g148600*. We have shown that the content of the protein encoded by *Psat1g148600* increases under action of the NO donor – sodium nitroprusside (Egorova, Tarchevsky, 2024).

Plant β -1,3-glucanases work in concert with chitinases, catalyzing the degradation of a microorganism's cell wall components, β -1,3-glucans and chitin (Jongedijk et al., 1995; Dos Santos, Franco, 2023). The expression of β -1,3-glucanases and chitinases in pea roots is comparable and is represented by one highly expressed gene each. While some chitinase genes are induced by SA and MeJA, the β -1,3-glucanase encoded by *Psat7g179640* is induced only by SA. The expression of other β -1,3-glucanase genes did not change under either SA or MeJA treatment.

As mentioned earlier, *Psat4g180800* encoding a PR-10 protein is the most highly expressed gene in pea roots, and its expression was not altered by SA or MeJA. Other genes encoding PR-10 proteins are largely insensitive to these phytohormones, and their expression remains unchanged under their influence.

One of the characteristics of PR proteins is their induction upon pathogen infection or under the action of stress phytohormones. Transcriptomic data analysis and our previous proteomic data indicate that in pea roots, the expression and content of some PR proteins are high – genes encoding certain PR proteins are among the most highly expressed genes

in the control. These include the *PR-1* gene *Psat1g156240*, and the *PR-10* genes *Psat4g180800* and *Psat1g156920*. This may be because roots are in constant contact with rhizosphere inhabitants, including pathogenic organisms. The existing pool of proteins with antipathogenic activity in roots allows them to be prepared for potential infection, and upon activation of the salicylate-dependent defense response, the expression and synthesis of chitinase-like proteins are activated, likely representing an additional defense mechanism in pea roots.

Conclusion

The transcriptomic analysis of basal and inducible expression of certain *PR* genes in pea roots conducted in this work demonstrates that when selecting marker genes to assess the activation of salicylate signaling, attention to their specific expression patterns in the particular organ must be paid. Pea roots are characterized by a high level of expression of genes encoding antipathogenic proteins. Transcriptomic analysis revealed that the most highly expressed *PR-1* genes in pea roots cannot be used to evaluate the activation of salicylate signaling. Chitinases and β -1,3-glucanases, induced by SA in pea leaves, did not show specificity in response to SA in roots. The obtained data indicate a difference in the response of leaves and roots to stress phytohormones. Using pea plants as an example, we showed that genes encoding salicylate-induced antipathogenic proteins in leaves are expressed at high levels in roots under normal conditions, while salicylate-induced genes in roots are those encoding chitinase-like proteins, which also have antipathogenic activity (unpublished data). In addition, the genes of chitinase-like proteins are practically not expressed in the control and are salicylate-inducible, which is an advantage when assessing the expression of these genes in order to evaluate the activation of the salicylate-dependent signaling pathway.

References

- Alexander D., Goodman R.M., Gut-Rella M., Glascock C., Weymann K., Friedrich L., Maddox D., Ahl-Goy P., Luntz T., Ward E. Increased tolerance to two oomycete pathogens in transgenic tobacco expressing pathogenesis-related protein 1a. *Proc Natl Acad Sci USA*. 1993;90(15):7327-7331. doi 10.1073/pnas.90.15.7327
- Ali S., Ganai B.A., Kamili A.N., Bhat A.A., Mir Z.A., Bhat J.A., Tyagi A., Islam S.T., Mushtaq M., Yadav P., Rawat S., Grower A. Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance. *Microbiol Res*. 2018;212-213:29-37. doi 10.1016/j.micres.2018.04.008
- Barba-Espín G., Clemente-Moreno M.J., Alvarez S., García-Legaz M.F., Hernández J.A., Díaz-Vivancos P. Salicylic acid negatively affects the response to salt stress in pea plants. *Plant Biol (Stuttg)*. 2011; 13(6):909-917. doi 10.1111/j.1438-8677.2011.00461.x
- Blanco F., Salinas P., Cecchini N.M., Jordana X., Van Hummelen P., Alvarez M.E., Holuigue L. Early genomic responses to salicylic acid in Arabidopsis. *Plant Mol Biol*. 2009;70:79-102. doi 10.1007/s11103-009-9458-1
- Breen S., Williams S.J., Outram M., Kobe B., Solomon P.S. Emerging insights into the functions of pathogenesis-related protein 1. *Trends Plant Sci*. 2017;22(10):871-879. doi 10.1016/j.tplants.2017.06.013
- Chen Y.L., Lee C.Y., Cheng K.T., Chang W.H., Huang R.N., Nam H.G., Chen Y.R. Quantitative peptidomics study reveals that a wound-induced peptide from PR-1 regulates immune signaling in tomato. *Plant Cell*. 2014;26(10):4135-4148. doi 10.1105/tpc.114.131185

- Davies G., Henrissat B. Structures and mechanisms of glycosyl hydrolases. *Structure*. 1995;3:853-859. doi 10.1016/S0969-2126(01)00220-9
- Davis J.M., Wu H., Cooke J.E., Reed J.M., Luce K.S., Michler C.H. Pathogen challenge, salicylic acid, and jasmonic acid regulate expression of chitinase gene homologs in pine. *Mol Plant Microbe Interact*. 2002;15(4):380-387. doi 10.1094/MPMI.2002.15.4.380
- De-la-Peña C., Badri D.V., Lei Z., Watson B.S., Brandão M.M., Silva-Filho M.C., Sumner L.W., Vivanco J.M. Root secretion of defense-related proteins is development-dependent and correlated with flowering time. *J Biol Chem*. 2010;285(40):30654-30665. doi 10.1074/jbc.M110.119040
- de Wit P.J.G.M. How plants recognize pathogens and defend themselves. *Cell Mol Life Sci*. 2007;64:2726-2732. doi 10.1007/s00018-007-7284-7
- Ding L.N., Li Y.T., Wu Y.Z., Li T., Geng R., Cao J., Zhang W., Tan X.L. Plant disease resistance-related signaling pathways: recent progress and future prospects. *Int J Mol Sci*. 2022;23(24):16200. doi 10.3390/ijms232416200
- Ding P., Ding Y. Stories of salicylic acid: a plant defense hormone. *Trends Plant Sci*. 2020;25(6):549-565. doi 10.1016/j.tplants.2020.01.004
- Dos Santos C., Franco O.L. Pathogenesis-related proteins (PRs) with enzyme activity activating plant defense responses. *Plants (Basel)*. 2023;12(11):2226. doi 10.3390/plants12112226
- Egorova A.M. Unexpected response of pea roots to the salicylic acid: induction of chitinase-like proteins compared to chitinases. *J Plant Growth Regul*. 2025;44:3164-3178. doi 10.1007/s00344-024-11607-1
- Egorova A.M., Tarchevsky I.A. Proteome analysis of the nitric oxide donor effect on *Pisum sativum* L. roots. *Appl Biochem Microbiol*. 2024;60(6):1292-1300. doi 10.1134/S0003683824604694
- Erb M. The role of roots in plant defence. In: Mérillon J., Ramawat K. (Eds) *Plant Defence: Biological Control. Progress in Biological Control*. Vol. 12. Springer, Dordrecht, 2012;291-309. doi 10.1007/978-94-007-1933-0_12
- Fraser R.S.S. Evidence for the occurrence of the “pathogenesis-related” proteins in leaves of healthy tobacco plants during flowering. *Physiol Plant Pathol*. 1981;19(1):69-76. doi 10.1016/S0048-4059(81)80009-4
- Glazebrook J. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu Rev Phytopathol*. 2005;43:205-227. doi 10.1146/annurev.phyto.43.040204
- Han Z., Xiong D., Schneider R., Tian C. The function of plant PR1 and other members of the CAP protein superfamily in plant-pathogen interactions. *Mol Plant Pathol*. 2023;24(6):651-668. doi 10.1111/mpp.13320
- Hennig J., Dewey R.E., Cutt J.R., Klessig D.F. Pathogen, salicylic acid and developmental dependent expression of a β -1,3-glucanase/GUS gene fusion in transgenic tobacco plants. *Plant J*. 1993;4(3):481-493. doi 10.1046/j.1365-313x.1993.04030481.x
- Jayaraj J., Muthukrishnan S., Liang G., Velazhahan R. Jasmonic acid and salicylic acid induce accumulation of β -1,3-glucanase and thaumatin-like proteins in wheat and enhance resistance against *Stagonospora nodorum*. *Biologia Plantarum*. 2004;48:425-430. doi 10.1023/B:BIOP.0000041097.03177.2d
- Jongedijk E., Tigelaar H., van Roekel J.S.C. Bres-Vloemans S.A., Dekker I., van den Elzen P.J.M., Cornelissen B.J.C., Melchers L.S. Synergistic activity of chitinases and β -1,3-glucanases enhances fungal resistance in transgenic tomato plants. *Euphytica*. 1995;85:173-180. doi 10.1007/BF00023946
- Kesari P., Patil D.N., Kumar P., Tomar S., Sharma A.K., Kumar P. Structural and functional evolution of chitinase-like proteins from plants. *Proteomics*. 2015;15:1693-1705. doi 10.1002/pmic.201400421
- Kiba A., Nishihara M., Nakatsuka T., Yamamura S. Pathogenesis-related protein 1 homologue is an antifungal protein in *Wasabia japonica* leaves and confers resistance to *Botrytis cinerea* in transgenic tobacco. *Plant Biotechnol*. 2007;24:247-253. doi 10.5511/plantbiotechnology.24.247
- Kouzai Y., Kimura M., Yamanaka Y., Watanabe M., Matsui H., Yamamoto M., Ichinose Y., Toyoda K., Onda Y., Mochida K., Noutoshi Y. Expression profiling of marker genes responsive to the defence-associated phytohormones salicylic acid, jasmonic acid and ethylene in *Brachypodium distachyon*. *BMC Plant Biol*. 2016;16:59. doi 10.1186/s12870-016-0749-9
- Kreplak J., Madoui M.A., Capal P., Novák P., Labadie K., Aubert G., Bayer P.E., ... Lichtenzweig J., Macas J., Doležel J., Wincker P., Burstin J. A reference genome for pea provides insight into legume genome evolution. *Nat Genet*. 2019;51:1411-1422. doi 10.1038/s41588-019-0480-1
- 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
- Li Y.F., Zhu R., Xu P. Activation of the gene promoter of barley β -1,3-glucanase isoenzyme GIII is salicylic acid (SA)-dependent in transgenic rice plants. *J Plant Res*. 2005;118:215-221. doi 10.1007/s10265-005-0213-7
- Lotan T., Ori N., Fluhr R. Pathogenesis-related proteins are developmentally regulated in tobacco flowers. *Plant Cell*. 1989;1(9):881-887. doi 10.1105/tpc.1.9.881
- Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550. doi 10.1186/s13059-014-0550-8
- Maleck K., Levine A., Eulgem T., Morgan A., Schmid J., Lawton K.A., Dang J.L., Dietrich R.A. The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nat Genet*. 2000;26:403-410. doi 10.1038/82521
- Mitsuhara I., Iwai T., Seo S., Yanagawa Y., Kawahigasi H., Hirose S., Ohkawa Y., Ohashi Y. Characteristic expression of twelve rice *PR1* family genes in response to pathogen infection, wounding, and defense-related signal compounds (121/180). *Mol Genet Genomics*. 2008;279(4):415-427. doi 10.1007/s00438-008-0322-9
- Papadopoulou G.V., Maedick A., Grosser K., van Dam N.M., Martínez-Medina A. Defence signalling marker gene responses to hormonal elicitation differ between roots and shoots. *AoB Plants*. 2018;10(3):ply031. doi 10.1093/aobpla/ply031
- Perez-Aranda A., Loera-Muro A., Caamal-Chan M.G. Expression analysis of defense signaling marker genes in *Capsicum annuum* in response to phytohormones elicitation. *Mol Biol Rep*. 2024;52(1):9. doi 10.1007/s11033-024-10071-0
- Santamaria M., Thomson C.J., Read N.D., Loake G.J. The promoter of a basic *PR1*-like gene, *AtPRB1*, from *Arabidopsis* establishes an organ-specific expression pattern and responsiveness to ethylene and methyl jasmonate. *Plant Mol Biol*. 2001;47:641-652. doi 10.1023/A:1012410009930
- Sarowar S., Kim Y.J., Kim E.N., Kim K.D., Hwang B.K., Islam R., Shin J.S. Overexpression of a pepper basic pathogenesis-related protein 1 gene in tobacco plants enhances resistance to heavy metal and pathogen stresses. *Plant Cell Rep*. 2005;24:216-224. doi 10.1007/s00299-005-0928-x
- Solomon W., Janda T., Molnár Z. Unveiling the significance of rhizosphere: implications for plant growth, stress response, and sustainable agriculture. *Plant Physiol Biochem*. 2024;206:108290. doi 10.1016/j.plaphy.2023.108290
- Spoel S.H., Johnson J.S., Dong X. Regulation of tradeoffs between plant defenses against pathogens with different lifestyles. *Proc Natl Acad Sci USA*. 2007;104(47):18842-18847. doi 10.1073/pnas.0708139104
- Tarchevsky I.A., Yakovleva V.G., Egorova A.M. Salicylate-induced modification of plant proteomes (review). *Appl Biochem Microbiol*. 2010;46(3):241-252. doi 10.1134/S0003683810030026

- Trapnell C., Williams B.A., Pertea G., Davis A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J., Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010; 28:511-515. doi 10.1038/nbt.1621
- Tyagi P., Singh D., Mathur S., Singh A., Ranjan R. Upcoming progress of transcriptomics studies on plants: an overview. *Front Plant Sci.* 2022;13:1030890. doi 10.3389/fpls.2022
- van Loon L.C., Rep M., Pieterse C.M. Significance of inducible defense-related proteins in infected plants. *Annu Rev Phytopathol.* 2006;44:135-162. doi 10.1146/annurev.phyto.44.070505.143425
- Wang Y., Mostafa S., Zeng W., Jin B. Function and mechanism of jasmonic acid in plant responses to abiotic and biotic stresses. *Int J Mol Sci.* 2021;22(16):8568. doi 10.3390/ijms22168568
- Xu L.N., Jiang X.R., Lin J.X., Li K., Javed T., Zhao J.-Y., Gao S.-J. Pathogenesis-related protein 10 family genes involved in sugarcane responses to biotic stressors and salicylic acid. *J Plant Growth Regul.* 2004;43:3907-3919. doi 10.1007/s00344-024-11371-2
- Yakovleva V.G., Egorova A.M., Tarchevsky I.A. Proteomic analysis of the effect of methyl jasmonate on pea seedling roots. *Dokl Biochem Biophys.* 2013;449:90-93. doi 10.1134/S1607672913020105

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

Received March 20, 2025. Revised November 9, 2025. Accepted November 11, 2025.

doi 10.18699/vjgb-26-25

Selecting stable rice mutants with linear mixed models (LMM) and stability indexes

P. Sharifi ¹, A.A. Ebadi², M.T. Hallajian³, H. Aminpanah¹¹ Department of Agronomy, Ra. C., Islamic Azad University, Rasht, Iran² Rice Research Institute of Iran, Agricultural Research, Education and Extension Organization (AREEO), Rasht, Iran³ Nuclear Science and Technology Research Institute of Iran, Tehran, Iran peyman.sharifi@iau.ac.ir

Abstract. Mutation serves as a pivotal source of diversity in plant breeding. This study focused on identifying stable rice mutant lines. Fourteen rice mutant lines, along with four conventional cultivars, were evaluated in a randomized complete block design with three replicates across three Iranian locations (Rasht, ChaparSar, and Fars province) during two growing seasons (2015, 2016). All statistical analyses were performed using the 'metan' (multi-environment trial analysis) R package. Single-environment ANOVA indicated significant genotypic effects for all traits. Likelihood ratio tests (LRTs) confirmed significant environment and genotype-by-environment interaction (GEI) effects for all traits. The first three principal components (PCs) captured 68.13, 14.46, and 9.76 % of the GEI variation, respectively. Heatmap visualization of yield performance and WAASB (weighted average of absolute scores based on best linear unbiased prediction, BLUP) highlighted genotypes G3, G9, G6, G12, and G5 as both high-yielding and stable. Multi-trait stability index (MTSI) analysis, designed to reveal genotypic strengths and weaknesses, selected only genotypes G7, G5, and G1. The top five genotypes based on the harmonic mean of the relative performance of genotypic values (HMRPGV) were G5, G12, G7, G2, and G1. The superior performance of certain mutants demonstrates that mutation has effectively generated significant genetic diversity. Notably, genotypes G12, G5, and G9 exhibited a clear advantage over the other genotypes and warrant consideration for selection or cultivar release; however, only G5 was selected based on all traits in the MTSI index and could therefore undergo selection or cultivar introduction processes.

Key words: genotype-by-environment interaction; MTSI; mutation breeding; simultaneous selection index; WAASB

For citation: Sharifi P, Ebadi A.A., Hallajian M.T., Aminpanah H. Selecting stable rice mutants with linear mixed models (LMM) and stability indexes. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):222-232. doi 10.18699/vjgb-26-25

Funding. The project was supported by the Iranian Rice Research Institute 03-04-0455-9411.

Authors' contribution. MTH treated parental seeds with gamma rays. AAE planned the research, carried the lead role in managing the research and collecting the data. PS analyzed the data and wrote the manuscript with contribution from all the authors.

Отбор стабильных мутантов риса с помощью линейных моделей со смешанными эффектами (LMM) и индексов стабильности

П. Шарифи ¹, А.А. Эбади², М.Т. Халладжян³, Х. Аминпанах¹¹ Департамент агрономии, Исламский университет Азад, Решт, Иран² Иранский исследовательский институт риса, Организация по сельскохозяйственным исследованиям, образованию и распространению знаний (AREEO), Решт, Иран³ Научно-исследовательский институт ядерных наук и технологий Ирана, Тегеран, Иран peyman.sharifi@iau.ac.ir

Аннотация. Мутационные изменения служат основным источником разнообразия в селекции растений. Данное исследование было сосредоточено на выявлении стабильных мутантных линий риса. Четырнадцать мутантных линий и четыре обычных сорта риса были оценены в рандомизированном посеве с тремя повторениями в трех регионах Ирана (Решт, Чапарсар и провинция Фарс) в течение двух вегетационных сезонов 2015 и 2016 гг. Все статистические анализы выполнены с использованием пакета R 'metan' (multi-environment trial analysis). Однофакторный ANOVA показал значимые генотипические эффекты для всех признаков. Тесты отношения правдоподобия (LRT) подтвердили значимые эффекты среды и взаимодействия генотип-среда (GEI) для всех признаков. Первые три главных компонента определяли 68.13, 14.46 и 9.76 % вариаций GEI соответственно. Визуализация тепловой карты для выраженности урожайности и WAASB (взвешенное среднее абсолютных баллов на основе наилучшего линейного несмещенного прогноза, BLUP) позволила выделить генотипы G3, G9, G6, G12 и G5 как высокоурожайные и стабильные. Анализ с помощью индекса мультипризнаковой стабильности

(MTSI), разработанный для выявления сильных и слабых сторон генотипов, отобрал только генотипы G7, G5 и G1. С использованием параметра гармонического среднего относительной производительности генотипических значений (HMRPGV) были отмечены пять лучших генотипов – G5, G12, G7, G2 и G1. Высокая выраженность признаков у некоторых мутантов демонстрирует, что мутационные изменения могут эффективно создавать значительное генетическое разнообразие. В частности, генотипы G12, G5 и G9 имели явное преимущество перед другими генотипами и могут быть использованы для последующей селекции или для создания сорта. Однако только генотип G5 был отобран с учетом индекса MTSI на основе всех признаков и, следовательно, может быть использован в дальнейшем селекционном процессе или для создания сорта.

Ключевые слова: взаимодействие генотип–среда; MTSI; мутационная селекция; индекс одновременного отбора; WAASB

Introduction

Rice stands as a critical cereal crop, serving as a primary food source for over half of the global population. It assumes a vital role in ensuring food and nutritional security, while concurrently contributing to poverty alleviation (Khush, 2005). Crop production is susceptible to abiotic stresses, including drought and salinity, necessitating the evaluation of genotypes across diverse environments. This approach facilitates the identification of adaptable genotypes suitable for broad or specific environmental conditions (Sharifi, 2020).

Mutation is a key source of generating variation in plant breeding programs, allowing for the development of novel traits or the improvement of existing ones. This process relies on alterations within the genetic structure, leading to new heritable traits that can spread within plant populations (Shu, 2012). Mutation breeding is often used to develop cultivars that are more tolerant to biotic and abiotic stresses, enhance quality, or alter other plant characteristics (Ebadi et al., 2019). This strategy is a useful complement to existing germplasm and can help in the development of new cultivars. As plant phenotypes are affected by genotype-by-environment interaction (GEI), stability analysis is essential to understand the performance of genotypes across different environments and guide breeders in selecting superior genotypes (Sharifi et al., 2017). Consequently, yield stability and adaptability analyses in multi-environment trials are a critical component of genotype evaluation programs before a new breeding line is released as a commercial variety (Kang, 1988). This demonstrates the plant's ability to maintain yield potential despite environmental fluctuations (Yan et al., 2000).

Stability analysis has become an important part of research and plant breeding. Various statistical models, including univariate and multivariate methods, are used to assess the stability and adaptability of genotypes. The simultaneous selection index named as grain yield in a single non-parametric index (GSI), which is based on the sum of ranks for yield stability and performance, is often used as a criterion for stability analysis (Farshadfar, 2008). Yan (2016) noted that mean performance and stability are not always equally important when evaluating genotypes. Stability is more of a concern when genotypes are tested in only a few environments. However, when genotypes are adequately tested, information about their stability may already be reflected in their mean performance, because highly unstable genotypes – those that perform very well in some environments but poorly in others – will not be among the top performers across all environments. Newer indices for evaluating stability based on linear mixed mo-

odels (LMMs) include WAASB (weighted average of absolute scores based on best linear unbiased prediction (BLUP)) and WAASBY (WAASB stability index and yield performance, for simultaneous selection based on yield performance and stability) (Olivoto et al., 2019b). This approach combines additive main effects and multiplicative interaction models (AMMI), graphical tools with the predictive precision of BLUP (Olivoto et al., 2019b). The relative importance of mean performance and stability can vary when making selection decisions, which can change how genotypes are ranked (Yan, 2016). This explains why many breeders and variety evaluation organizations base their decisions only on mean performance, yet still achieve significant breeding progress (Yan, 2016). Given the importance of the WAASBY index, it can be used for all stability indices instead of the GSI (Farshadfar, 2008), allowing for the identification of the most superior genotypes.

In addition to high yield, which remains the primary breeding objective, a cultivar must meet minimum requirements for every trait that is significant to growers, processors, and end-users (Yan, 2021). Consequently, when data on multiple traits are available, the multi-trait stability index (MTSI) has several valuable applications for the simultaneous selection of mean performance and stability. Indeed, the greatest challenge in plant breeding is to combine all desirable traits within a single genotype, as key breeding objectives are often negatively correlated due to genetic linkage or pleiotropy (Yan, 2021). These indices are used in the stability analysis of rice (Sharifi et al., 2021), lentil (Karimizadeh et al., 2020), forage (Santos, Marza, 2020), barley (Ahakpaz et al., 2021), and wheat (Verma, Singh, 2020). An alternative analysis for stability studies based on mixed models is the harmonic mean of genotypic values and of the relative performance of genotypic values (HMRPGV), which provides information on stability, adaptability, and yield performance of genotypes in the same unit and scale as the evaluated trait (Resende, 2007). Selecting genotypes with the highest values of the harmonic mean of genotypic values (HMGV), relative performance of genotypic values (RPGV), and HMRPGV allows for simultaneous selection for yield performance and stability. This approach is used to evaluate grain yield stability in rice (Colombari-Filho et al., 2013), wheat (Coan et al., 2018; Verma, Singh, 2020), and corn (Rodvalho et al., 2015).

Several studies have assessed GEI in rice mutant lines. Dushyanthakumar and Shadadashari (2007) found significant effects of genotypes, environment, and their interactions when evaluating the yield stability of 15 rice mutants across nine environments. They used regression coefficients and

deviations from regression to identify stable mutants. Dewi and Dwimahyani (2019) evaluated 12 mutant rice genotypes at 16 locations and estimated yield stability using regression lines. Rahayu (2020) used AMMI analysis to differentiate stable mutant rice lines from 25 genotypes. Relatively few studies have used linear mixed models for rice stability analysis. Donoso-Nanculao et al. (2015) analyzed 10 rice genotypes using BLUP and identified superior genotypes. Balestre et al. (2010) also used BLUP to analyze the stability of rice genotypes and found that phenotypic means had lower predictive potential compared to BLUP.

The objectives of this study were to (1) evaluate GEI in rice genotypes across diverse environments in Iran, and (2) identify stable and high-yielding mutant lines.

Materials and methods

Plant material and experimental design

Fourteen rice mutant lines, along with their parental cultivars (Hashemi, Tarom, and Khazar) and the Gilaneh variety as a control (Table 1), were evaluated in a randomized complete block design (RCBD) with three replications. The experiment was conducted across three geographically distinct research stations in Iran: Rasht, ChaparSar, and Fars province. Data were collected over two years (2015, 2016). In both years, sowing took place in April, and transplanting was carried out in May at the 4–5 leaf stage. Harvesting occurred in August in both years as well.

The rice mutant lines originated from seeds of the parental cultivars subjected to 300 Gray of gamma irradiation from a Cobalt-60 source. Irradiation was performed at the Iranian Agricultural Atomic Research Institute in Tehran in 2009, with

seed moisture content maintained at eight percent. Following irradiation, the M1 generation was grown in the field. Surviving plants were self-pollinated, and seeds from the M2 generation were collected. This self-pollination process was repeated through the M5 generation to achieve homozygosity.

Prior to sowing, seeds of all experimental genotypes (mutant lines and cultivars) were surface-sterilized with a 20 % sodium hypochlorite solution for 10 minutes, followed by rinsing with sterile distilled water. This procedure was carried out in May of each year. Seeds were then placed in Petri dishes containing moistened filter paper to initiate germination. Germination occurred in a controlled environment at 35 °C and 70 % humidity. Seedlings were subsequently transplanted into 120 cm×60 cm boxes within greenhouses, ensuring suitable conditions for transplant production.

Twenty-five-day-old seedlings (approximately 30 cm tall) were transplanted to the main field at a planting density of 25 plants per square meter. Each plot measured two square meters. Data were collected from ten randomly selected plants within each plot. The following traits were measured: plant height, tiller number, panicle length, number of filled grains per panicle, number of unfilled grains per panicle, fertility percentage, and hundred-grain weight. Grain yield was determined on a per-square-meter basis.

Statistical analysis

To assess genotypic stability and adaptability, the harmonic mean of genotypic values (HMGV) and the relative performance of genotypic values (RPGV) were calculated. The harmonic mean of the relative performance of genotypic values (HMRPGV) was employed for the simultaneous evaluation of stability, adaptability, and yield. These indices were computed using the following formulas (Resende, 2007):

$$HMGV_i = \frac{E}{\sum_{j=1}^E \frac{1}{GV_{ij}}}, \tag{1}$$

$$RPGV_i = \frac{1}{E} \sum_{j=1}^E \frac{GV_{ij}}{\mu_j}, \tag{2}$$

$$HMRPGV_i = \frac{E}{\sum_{j=1}^E \frac{1}{GV_{ij}/\mu_j}}, \tag{3}$$

where μ_j represents the general mean for the j -th environment; E denotes the number of environments; and GV_{ij} : $u_j + g_i + ge_{ij}$ indicate the genotypic value of the i -th genotype in the j -th environment. Here, u_j is the mean of the j -th environment, while g_i and ge_{ij} are the BLUP values of the i -th genotype and the interaction between the i -th genotype and j -th environment, respectively.

The weighted average of absolute scores from the singular value decomposition (SVD) of BLUP-estimated interaction effects ($WAASB_i$) was computed using equation (4) (Olivoto et al., 2019a):

$$WAASB_i = \frac{\sum_{k=1}^p |IPCA_{ik} \times EP_k|}{\sum_{k=1}^p EP_k}, \tag{4}$$

where $IPCA_{ik}$ is the score of the i -th genotype (or environment) in the k -th IPCA, and EP_k is the value of the explained variance by the k -th IPCA. The k -th IPCA is the k -th interaction principal

Table 1. The origin/pedigree of rice genotypes used in this experiment

Genotype code	Genotype	Pedigree
G1	TM6-230-VE-7-5-1	Parent: Tarom
G2	TM6-230-VE-8-4-1	
G3	TM6-250-10-7-1	
G4	TM6-B-2-1-E	
G5	TM6-B-7-1	Parent: Hashemi
G6	TM6-B-19-2	
G7	HM5-250-E-1-1	
G8	HM5-250-E-3-2	
G9	HM5-250-6-6	
G10	HM5-250-7-6	
G11	HM5-300-12-1	Parent: Khazar
G12	HM5-300-3-1	
G13	HM5-300-5-1	
G14	KM5-200-4-2-E	
G15	Khazar	IR36/TNAU4756
G16	Hashemi	Iranian landrace
G17	Tarom	Iranian landrace
G18	Gilaneh	Saleh × Abjibuji

Table 2. Evaluation of significance of factors for random effects (E and GEI) by LRT (χ^2) and for fixed effect (G) by ANOVA and estimation of variance components by REML

SOV	Plant height, cm	Tiller number	Panicle length, cm	Number of filled grains per panicle	Number of unfilled grains per panicle	Fertility percentage	Hundred-grain weight, g	Grain yield, kg · ha ⁻¹
Mean Square								
Genotype	102.7**	9.39**	8.64**	59.31**	40.13**	19.76**	0.049**	224817**
LRT (χ^2)								
Environment	109.5**	39.63**	81.07**	102.1**	0.0000014 ^{ns}	2.11 ^{ns}	95.98**	71.75**
Gen:Env	384.3**	240.9**	184.9**	540.8**	931.3**	0.089**	379.2**	577.6**
Variance components								
REML Estimates (%)								
σ^2_E	592.1 (94.11)	5.19 (45.64)	9.96 (80.42)	421.9 (79.57)	0.00000016 (0.00000017)	3.62 (6.57)	0.047 (75.73)	306144 (52.25)
$\sigma^2_{E/R}$	0.40 (0.06)	0.19 (1.73)	0.012 (0.09)	0.14 (0.03)	0.00000021 (0.00000023)	0 (0)	0 (0)	1181 (0.20)
σ^2_i	30.14 (4.79)	4.23 (37.20)	1.52 (12.30)	96.48 (18.20)	87.66 (96.76)	49.66 (90.17)	0.012 (19.75)	252401 (43.08)
σ^2_e	6.49 (1.03)	1.75 (15.43)	0.88 (7.18)	11.68 (2.20)	2.93 (3.23)	1.79 (3.26)	0.0028 (4.51)	26182 (4.46)
σ^2_p	629.2	11.38	12.39	530.2	90.60	55.07	0.063	585907

Note. The numbers in parentheses indicate the percentage of variance accounted for by the phenotypic variance. LRT, likelihood ratio test. σ^2_E , environmental variance; $\sigma^2_{E/R}$, Block-within-Environment variance (or Environment/Replication variance); σ^2_i , Genotype × Environment interaction variance; σ^2_e , residual variance; σ^2_p , phenotypic variance.
ns, non-significant, *, **, significant at 5 % and 1 % probability levels, respectively.

component axis (IPCA) in AMMI, used to compute WAASB. The genotype with the lowest WAASB value is considered the most stable (Olivoto et al., 2019a).

The WAASBY index, which allows simultaneous selection based on grain yield and stability, was derived using the following formula (Olivoto et al., 2019a):

$$WAASBY_i = \frac{(rG_i \times \theta_Y) + (rW_i \times \theta_S)}{\theta_Y + \theta_S}.$$

(5)

Where $WAASBY_i$ is the superiority index for the i -th genotype, and θ_Y and θ_S are the weights for seed yield and stability (WAASB), respectively. The rescaled values (0–100) for the response variable (rG_i) and WAASB (rW_i) were calculated as follows:

$$rG_i = \frac{100 - 0}{G_{\max} - G_{\min}} \times (G_i - G_{\max}) + 100,$$

(6)

$$rW_i = \frac{0 - 100}{W_{\max} - W_{\min}} \times (W_i - W_{\max}) + 0.$$

(7)

The multi-trait stability index (MTSI) for the i -th genotype was calculated as follows (Olivoto et al., 2019b):

$$MTSI_i = \left[\sum_{j=1}^f (F_{ij} - F_j)^2 \right]^{0.5}.$$

(8)

Where, F_{ij} is the j -th score of the i -th genotype, and the ideotype (F_j) is an ideal reference genotype with the most desirable values for each trait, f represents the number of traits (variables) included in the analysis. The ideotype has the highest WAASBY (100) for all analyzed variables (Olivoto et al., 2019b). The genotype with the lowest MTSI is then closer to the ideotype and therefore presents a high mean performance and stability (MPE) for all analyzed variables.

Selection differential (SD), selection gain (SG), and percent- age of selection differential (% SD) represent the difference

between the mean of selected genotypes ($\bar{X}$) and the entire population, the expected genetic progress, and the relative selection intensity, respectively.

$$SD = \bar{X}_S - \bar{X}_0,$$

(9)

$$SG = h^2 \times SD,$$

(10)

$$\% SD = \frac{SD}{\bar{X}_0} \times 100.$$

(11)

$\bar{X}_S$: mean of selected genotypes; $\bar{X}_0$: mean of the entire popula- tion; h^2 : heritability of the trait.

The selection differential for mean performance and the WAASBY index was calculated for each trait considering a selection intensity of 15 %.

All statistical analyses were performed using the ‘metan’ (multi-environment trial analysis) R package (Olivoto, Lúcio, 2020). An example of statistical analysis using R can be seen in the Supplementary file¹.

Results

Analysis of variance, estimation of variance components and predicted grain yield

Single-environment ANOVA indicated significant genotypic effects for plant height, tiller number, panicle length, number of filled grains per panicle, number of unfilled grains per panicle, fertility percentage, hundred-grain weight, and grain yield. Likelihood ratio tests (LRTs) confirmed significant environ- ment and genotype-by-environment interaction (GEI) effects for all traits. Analysis of variance also indicated a significant effect of genotype on all traits (Table 2).

¹ Supplementary file is available at: <https://vavilovj-icg.ru/download/pict-2026-30/appx16.pdf>

Principal component analysis
and biplot interpretation

The first three principal components (PCs) captured 68.13, 14.46, and 9.76 % of the GEI variation, respectively (Fig. 1). A type III biplot (WAASB values vs grain yield, Fig. 2 and Fig. 3) was used to visualize genotypic stability.

Genotypes G14 and G15, along with environments E3 and E6, positioned in the first quadrant, exhibited below-average grain yield and high WAASB values, indicating instability and a substantial contribution to GEI. Conversely, genotypes G2, G7, and G8, and environments E1, E4, and E5, in the second quadrant, showed above-average grain yield but were also unstable with significant GEI contributions. Environments E3, E4, and E6 displayed the highest WAASB values, indicating greater discriminating ability. Genotypes G4, G6, G9–G11, G13, G16, and G17, located in the third quadrant, demonstrated low yield and high stability (low WAASB). The most desirable genotypes – G1, G3, G5, G12, and G18 – occupied the fourth quadrant, characterized by high yield and stability. Environment E2, also in this quadrant, showed high yield performance but low WAASB values, indicating poor genotypic discrimination ability (Fig. 2).

Genotype ranking based on stability
and performance weights

Figure 3 illustrates genotype rankings based on varying weights assigned to WAASB (stability) and grain yield. Ranking of 18 rice genotypes was based on the WAASBY index, illustrating variations in genotype performance under different weighting schemes for stability (WAASB) versus grain yield (BLUPs). The weights (θ) range from 0 (prioritizing yield alone) to 1 (prioritizing stability alone), with $\theta = 0.5$ representing equal emphasis on both traits. Ranking based solely on stability (WAASB weight = 1) identified genotypes G6, G9, G3, G10, G13, and G17 as most stable. Ranking based solely on yield (yield weight = 1) identified G1, G2, G5, G7, and G12 as the highest-yielding genotypes.

Hierarchical clustering revealed distinct genotype groupings based on stability and yield: (1) G4, G8, G14, and G15 were the least desirable; (2) G10, G11, G13, G16, and G17 were stable but low-yielding; (3) G1, G2, G7, and G18 were high-yielding but unstable; and (4) G3, G5, G6, G9, and G12 (particularly G5) were both high-yielding and stable.

Factor analysis, multi-trait stability index (MTSI)
and genotype selection

Factor analysis of the WAASBY index retained two principal components (eigenvalue > 1), explaining 71.5 % of the total variation (Table 3). Factorial loadings after varimax rotation and communalities obtained in the factor analysis are represented in Table 4. The eight traits were grouped into two factors: FA1 (agronomic and plant stature traits including plant height (PH), filled grains per panicle (FG), unfilled grains per panicle (UNFG), fertility percentage (FP), hundred-grain weight (HGW), grain yield (GY)), and FA2 (panicle-related traits contain tiller number (TN) and panicle length (PL)). While primarily associated with FA1, FG, UNFG, and GY also exhibited high scores on FA2 (Table 5).

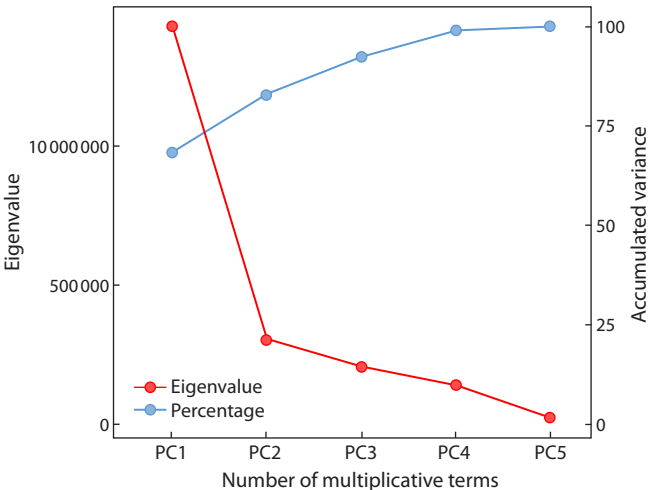


Fig. 1. Eigenvalues of the BLUP_GEI matrix for grain yield of 18 rice genotypes across six environments.

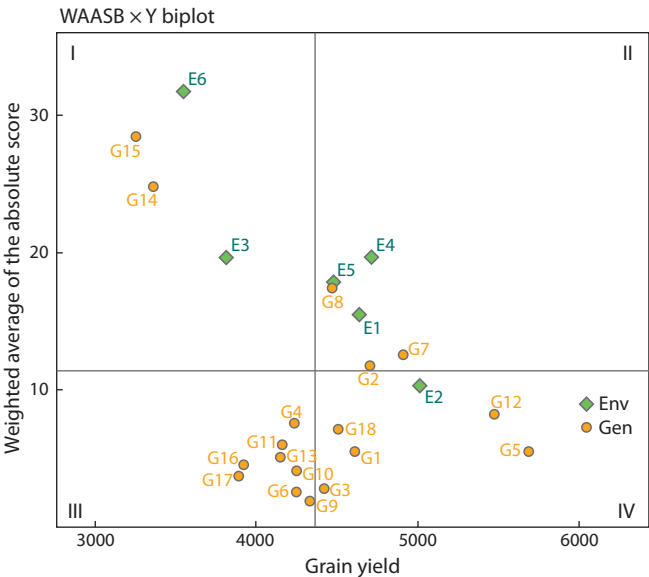


Fig. 2. Biplot of grain yield vs WAASB of 18 rice genotypes evaluated over six environments (combinations of two cultivation years in three locations).

Ranking based on MTSI identified G7 (MTSI = 1.97), G5 (MTSI = 2.07), and G1 (MTSI = 2.24) as superior genotypes (Fig. 4, cutoff at MTSI = 2.24). The cutoff value (MTSI = 2.24) was determined following the methodology of Olivoto et al. (2019b), where the Multi-Trait Stability Index (MTSI) ranks genotypes based on their stability and multi-trait performance. This threshold represents the upper limit of the top-performing genotypes (G1, G5, G7), beyond which stability and performance decline significantly. The selection of this cutoff aligns with standard MTSI applications, where natural breaks in the index distribution or percentile-based thresholds are used to distinguish superior genotypes. The genotype with the lowest MTSI is then closer to the ideotype and therefore presents a high MPE for all analyzed variables (Olivoto et al., 2019b).

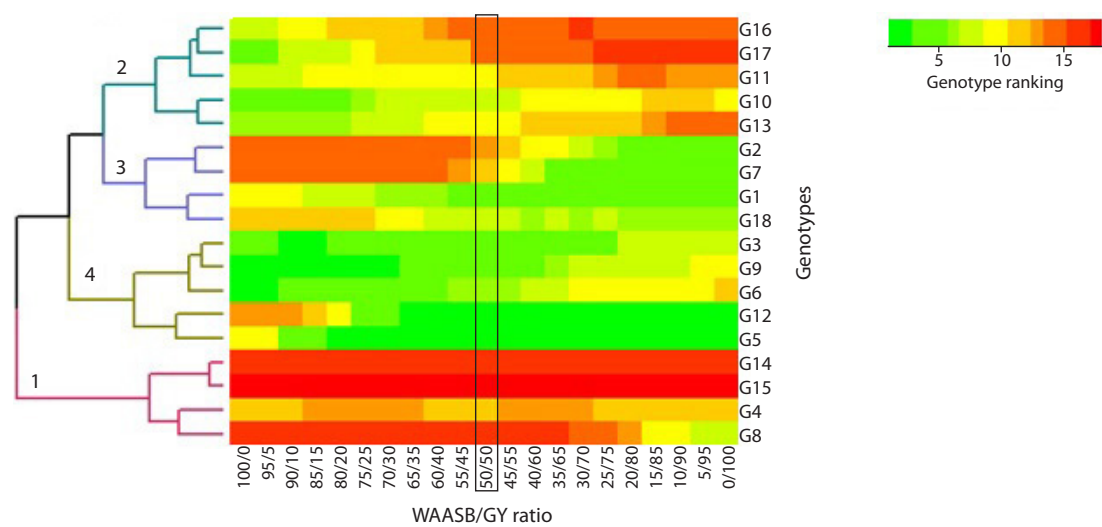


Fig. 3. Ranks of 18 rice genotypes with different weights for stability and grain yield.

Table 3. Eigenvalues and explained variance obtained in the factor analysis

PC	Eigenvalues	Variance, %	Cum. variance, %
PC1	4.61	57.6	57.6
PC2	1.11	13.9	71.5
PC3	0.997	12.5	84.0
PC4	0.636	7.95	91.9
PC5	0.343	4.29	96.2
PC6	0.22	2.75	99.0
PC7	0.073	0.913	99.9
PC8	0.00898	0.112	100.0

Table 4. Factorial loadings after varimax rotation, and communalities obtained in the factor analysis

Traits	FA1	FA2	Communality	Uniquenesses
Plant height (PH)	−0.908	0.133	0.841	0.159
Tiller number (TN)	−0.152	−0.872	0.784	0.216
Panicle length (PL)	−0.0767	−0.670	0.455	0.545
Filled grains per panicle (FG)	−0.570	−0.330	0.434	0.566
Unfilled grains per panicle (UNFG)	−0.717	−0.623	0.902	0.0981
Fertility percentage (FP)	−0.713	−0.639	0.917	0.0829
Hundred-grain weight (HGW)	−0.728	−0.229	0.582	0.418
Grain yield (GY)	−0.653	−0.617	0.807	0.193

Table 5. The related traits and genotypes for each factor

Factor	Related traits to FA	Related genotypes to FA
FA1	PH, FG, UNFG, FP, HGW, GY	G18, G2, G4, G5, G6, G1, G10, G14, G15
FA2	TN, PL	G7, G8, G9, G11, G12, G13, G17, G3

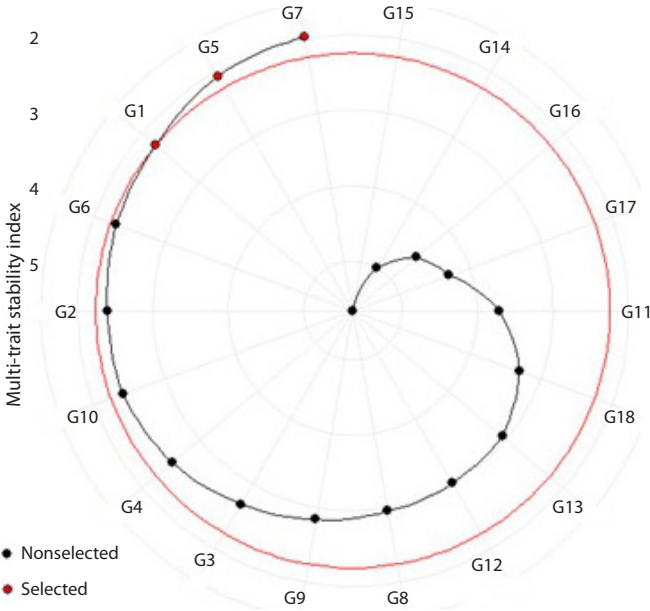


Fig. 4. Genotype ranking and genotypes selected for MTSI.

The MTSI resulted in positive selection differentials (SD) for WAASBY of all traits, ranging from 13 % (TN) to 49.5 % (HGW) (Table 6). Negative SDs for mean performance were observed for PH and UNFG, while positive SDs were noted for the remaining traits. This led to negative selection gains (SG) for PH (–0.35 %) and UNFG (–47.5 %), and positive SG for FG, FP, HGW, GY, TN, and PL (0.996 % ≤ SG ≤ 16.2 %). These results indicate that the selected genotypes exhibited improved genotypic stability compared to the original population, and indicate that MTSI is effective for genetic breeding.

Visualization of strengths and weaknesses

Figure 5 visualizes the strengths and weaknesses of each genotype. The contribution of each factor (FA1 and FA2) to the MTSI is ranked for each genotype, with the most contributing

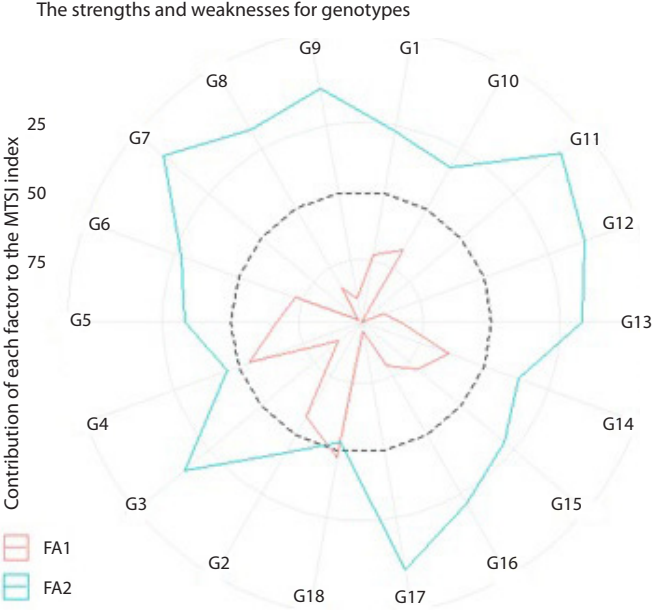


Fig. 5. The strengths and weaknesses of genotypes are shown as the proportion of each factor on the computed multi-trait stability index (MTSI) of all genotypes.

The smaller the proportion explained by a factor (closer to the external edge), the closer the traits within that factor are to the ideotype. FA1: PH, FG, UNFG, FP, HGW and GY; and FA2: TN and PL.

factor near the plot center and the least contributing factor near the plot edge (Olivoto et al., 2019b). Genotypes associated with a specific factor demonstrate high mean performance and stability (low WAASB) for the traits primarily influenced by that factor. Ten genotypes (G1, G2, G4, G5, G6, G10, G14, G15, G16 and G18) showed strengths related to FA1 (Fig. 5), indicating desirable combinations of high mean performance and stability for FA1-related traits: reduced PH and UNFG, and increased FG, FP, HGW, and GY. Eight genotypes (G3, G7, G8, G9, G11, G12, G13, and G17) exhibited strengths related to FA2 (Fig. 5). For G5, the second-ranked genotype by MTSI, significant FA1 contribution confirms its favorable

Table 6. Selection gains for mean performance of 18 rice genotypes obtained in a multi-trait selection with the multi-trait stability index (MTSI)

Traits	Factor	Goal	Mean performance (BLUP)							WAASBY index			
			Xo	Xs	SD	% SD	SG	SG %		Xo	Xs	SD	% SD
Plant height (PH)	FA1	Decrease	123	123.437	–0.437	–0.354	–0.409	–0.332		55.7	64.2	8.49	15.2
Filled grains per panicle (FG)		Increase	91.7	94.2	2.49	2.72	2	2.18		44.4	53.9	9.56	21.6
Unfilled grains per panicle (UNFG)		Decrease	13.5	7.1	–6.43	–47.5	–5.95	–43.9		83.3	95.8	12.5	14.9
Fertility percentage (FP)		Increase	88	92.7	4.63	5.26	4.21	4.78		78.1	93.7	15.6	19.9
Hundred-grain weight (HGW)			2.55	2.8	0.246	9.64	0.23	9.09		57.2	85.7	28.5	49.5
Grain yield (GY)	FA2		4360	5065	705	16.2	623	14.3		59.9	75.1	15.3	25.5
Tiller number (TN)		Increase	16.6	16.8	0.166	0.996	0.13	0.81		61.7	69.8	8.06	13
Panicle length (PL)			25.8	26.2	0.469	1.82	0.42	1.63		51.7	61.6	9.86	19.1

Note. Xo: mean performance (BLUP) and WAASBY index of the original population; Xs: mean performance (BLUP) and WAASBY index of the selected genotypes; SD: selection differential; % SD: percentage of selection differential; SG: selection gain.

Table 7. Ranking of the genotypes in all environments evaluated for adaptability parameters of genotypic values for the grain yield of rice genotypes evaluated in six environments

GEN	Grain yield, kg · ha ⁻¹	HMGV [†]	HMGV_order	RPGV	RPGV* $\bar{\mu}$	RPGV_order	HMRPGV	HMRPGV* $\bar{\mu}$	HMRPGV_order
G1	4607	4537	5	1.058	4612	5	1.054	4597	5
G2	4699	4640	4	1.084	4727	4	1.073	4680	4
G3	4418	4333	8	1.012	4414	8	1.009	4401	8
G4	4227	4096	12	0.965	4207	12	0.959	4183	12
G5	5684	5652	1	1.313	5725	1	1.306	5696	1
G6	4237	4181	10	0.974	4247	10	0.972	4236	10
G7	4904	4884	3	1.138	4962	3	1.124	4901	3
G8	4466	4430	7	1.039	4531	6	1.017	4436	7
G9	4323	4276	9	0.994	4333	9	0.993	4330	9
G10	4240	4164	11	0.973	4241	11	0.969	4225	11
G11	4157	4023	14	0.948	4134	14	0.942	4105	14
G12	5470	5448	2	1.266	5519	2	1.257	5481	2
G13	4148	4061	13	0.950	4144	13	0.947	4130	13
G14	3353	2365	17	0.742	3236	17	0.589	2568	17
G15	3252	2288	18	0.716	3121	18	0.568	2477	18
G16	3908	3845	15	0.898	3917	15	0.894	3897	15
G17	3883	3832	16	0.892	3890	16	0.891	3882	16
G18	4504	4433	6	1.037	4521	7	1.028	4481	6

Note. [†] Performance genetic value (RPGV); RPGV* $\bar{\mu}$: stability of genotypic values (HMGV), adaptability and stability of genotypic values (HMRPGV); HMRPGV* $\bar{\mu}$: general mean of all environments; the parameters are expressed in kg · ha⁻¹, except for RPGV and HMRPGV.

combination of reduced PH and UNFG, and increased FG, FP, HGW, and GY. This genotype exhibited lower plant height and unfilled grains per panicle, and higher tiller number, hundred-grain weight, and grain yield, compared to the overall average. Genotype G12 showed higher tiller number, fertility percentage, and grain yield, along with lower plant height and unfilled grains per panicle. While identified as superior by other indices, G12 did not rank highly using MTSI. This discrepancy is likely due to the low selection gain for traits in FA1, specifically hundred-grain weight.

Genotypic stability and adaptability assessment using HMRPGV

Based on the Harmonic Mean of Relative Performance of Genotypic Values (HMRPGV), the top five genotypes for stability and adaptability, relative to check varieties, were G1, G2, G7, G12, and G5. The HMRPGV multiplied by the overall mean (HMRPGV* $\bar{\mu}$) for these genotypes were 4597, 4680, 4901, 5481, and 5696 kg · ha⁻¹, respectively (Table 7). Selecting these genotypes would result in a 16.35 % increase in grain yield compared to the overall mean (4360 kg · ha⁻¹). This selection, based on stability, adaptability, and yield, highlights the positive response of these genotypes to improved environmental conditions and their consistent performance across diverse environments.

Discussion

A broad genetic base is crucial for successful rice breeding programs, and induced mutation represents a valuable technique for expanding genetic diversity (Cheema, 2006). The significant effect of genotype observed across all environments, as indicated by simple analysis of variance, underscores the genetic diversity present within the experimental material. This result suggests that mutation has effectively contributed to a reasonable level of diversity within the genetic material under investigation. This finding aligns with the work of other researchers who have assessed the stability of rice mutant lines and demonstrated a significant impact of mutation in generating diversity with respect to grain yield and other important traits, particularly yield components (Dushyanthakumar, Shadadashari, 2007; Donoso-Nanculao et al., 2015; Dewi, Dwimahyani, 2019; Rahayu, 2020).

Experimental trials designed to evaluate yield performance, stability, and adaptability are essential for characterizing genetic materials such as mutant lines before they are released as new breeding lines or varieties (Shu, 2012). Conversely, the observed variation in genotype grain yield across different environments highlights the need to carefully determine the yield performance and stability of genotypes in a range of environmental conditions (Ebadi et al., 2019). Therefore, it is crucial to investigate the yield stability of rice mutant lines and

to identify those lines that exhibit both high yield potential and stable performance across diverse environmental conditions. The likelihood ratio test (LRT) conducted in this study revealed significant effects of both the environment and genotype-by-environment interaction (GEI). Similar results have been reported by other researchers conducting multi-environment trials with rice (Bose et al., 2011; Akter et al., 2015; Sharifi et al., 2017; Rahayu, 2020). Variations in environmental conditions from year to year and from location to location contribute to a pronounced GEI effect, which underscores the role of genetic factors in influencing genotype performance across different environmental conditions (Dia et al., 2016).

The restricted maximum likelihood (REML) analysis conducted in this study indicated that environmental variance contributed more significantly to the phenotypic variance observed for traits such as plant height, number of tillers, panicle length, number of filled grains per panicle, hundred-grain weight, and grain yield. In contrast, GEI had a greater contribution to the observed variance in the number of unfilled grains per panicle and fertility percentage. The first two principal components derived from the GEI analysis accounted for 68.13 and 14.46 % of the total genotype-by-environment interaction variation, respectively. In line with this finding, other researchers have acknowledged the substantial contribution of the first two principal components in explaining genotype-by-environment interaction effects on rice grain yield (Nayak et al., 2008; Akter et al., 2015; Allahgholipour, 2017; Rahayu, 2020).

Genotype-by-environment interaction reduces the predictability of genotype performance in target environments based on observations made in test environments (Yan et al., 2011). Numerous procedures and statistical methods have been developed to quantify genotype by environment interaction through the evaluation of genotypes across multiple environments. In addition to conventional multivariate methods like AMMI and GGE biplot analysis, the best linear unbiased predictor (BLUP) method has also been recommended for evaluating genotypes in different environments and for quantifying genotypic stability. Several researchers have employed the BLUP method to assess the stability of rice genotypes (Balestre et al., 2010; Donoso-Nanculao et al., 2015). The use of restricted maximum likelihood (REML) for estimating variance components offers significant flexibility in analysis (Patterson, Thompson, 1971) and can effectively handle complex data structures (Searle et al., 1992).

The WAASBY index (Olivoto et al., 2019a) facilitates the weighting between mean performance and stability (MPE) in genotype evaluation. The primary advantage of this procedure is its integration of AMMI's graphical tools (Gauch, 2013) with the predictive accuracy of BLUP for stability analysis (Piepho, 1994; van Eeuwijk et al., 2016). Consequently, the WAASB approach, which proposes the SVD (singular value decomposition) of a two-way table with BLUPs for GEI interaction (instead of the residual of the additive model as in standard AMMI), allows for a graphical representation of a random GEI structure. The WAASB index accounts for more than one (stability based on IPCA1) or two (AMMI-stability value) interaction principal component axes. In the present research, the first three principal components accounted for

over 82 % of the GEI variation in grain yield, making the WAASB and WAASBY indices more effective in identifying superior genotypes. On the other hand, the variation of certain genotypes is explained by more IPCAs, making indices such as WAASB necessary, which is calculated using additional principal components. Consequently, we utilized this index and identified genotypes G1, G3, G5, G12, and G18 as superior (Fig. 2). The biplot based on this index was also employed to assess the discriminating ability of environments, revealing that the environments in the first two sections of biplot type III (Fig. 2), particularly E6, exhibited the highest WAASB value and strong genotype discrimination capability. However, the genotypes in these two quadrants with a significant contribution to the GEI were unstable. In the fourth quadrant of this biplot, genotypes G18, G1, G3, G5, and G12 emerged as the most desirable, demonstrating high yield stability and performance. The ranking of rice genotypes with varying weights for grain yield performance and stability in Figure 3 illustrated the diversity among rice mutant lines. In this heatmap, scenarios with different weights of yield stability and performance were simulated to display changes in genotype ranking. The grain yield of mutant lines G5, G12, G2, G7, and G1 surpassed that of parental varieties and Gilaneh, the control variety.

The HMRPGV method evaluates grain yield performance, stability, and adaptability simultaneously within a genotypic context (Resende, 2007). The HMRPGV of 1.306 for G5 indicated a 30.6 % increase in grain yield over the general mean. It appears that yield performance is prioritized over stability in the HMRPGV, and the selection of genotypes is based more on yield performance (as compared to Fig. 3) than on stability.

When data on several traits are available, the multi-trait stability index (MTSI) has many useful applications for simultaneous selection of mean performance and stability. This index is an MTSI-based on factor analysis, designed for simultaneous selection using both fixed and mixed effect models for multiple traits (Olivoto et al., 2019b). The factor-analytic variance-covariance structure accommodates different variances across environments and covariances between pairs of environments by approximating a completely unstructured covariance matrix through factor analysis. Therefore, it is reasonable to assume that the model will fit better than a standard mixed effect model. This factor analysis of WAASBY, which is the first step in computing the MTSI index, revealed that the eight traits were grouped into two factors. A Euclidean distance was then used to compute the distance between the genotypes' scores and the ideotype's score. This plot (Fig. 4) displays the genotype ranking in ascending order for the MTSI index. The selected genotypes using MTSI, including G7, G5, and G1, are shown in red, and the spiral depicts the cutpoint according to the selection pressure. A higher weight for mean performance was assigned to tiller number, panicle length, filled grains per panicle, fertility percentage, hundred grain weight, and grain yield. It is preferable to select highly productive genotypes that may not perform well in terms of broad stability and then consider the adaptability of these genotypes to specific environments to make better recommendations. The MTSI index enables simultaneous selection of genotypes based on performance and stability across all traits, making it a unique

index (Olivoto et al., 2019a). This index has been used for the selection of superior genotypes in *Solanum melongena* (Koundinya et al., 2019), *Brassica* spp. (Bocianowski et al., 2019), chickpea (Sellami et al., 2021), maize (Olivoto et al., 2021), and rice (Sharifi et al., 2021).

MTSI serves as an effective tool for identifying the strengths and weaknesses of genotypes and selecting those with desired mean performance and stability (Olivoto et al., 2021). This approach enabled the identification of genotypes that combine the desired mean performance and stability (MPE) for important traits such as grain yield (GY), hundred-grain weight (HGW), fertility percentage (FP), and tiller number (TN) (Fig. 5).

Conclusion

The significant effect of genotype indicated that the mutation process has successfully created a reasonable level of diversity in the genetic material. The first three principal components accounted for 92.35 % of the GEI variation with respect to grain yield, and the WAASB index was used to evaluate the stability of the genotypes. Biplot type III analysis revealed that genotypes G18, G1, G3, G5, and G12 were both high-yielding and stable across the tested environments. The heatmap visualization of yield performance and WAASB stability scores indicated that genotypes G3, G9, G6, G12, and G5 were highly productive and stable. The MTSI (multi-trait stability index) identified genotypes G7, G5, and G1 as the selected genotypes based on the overall stability and performance across all measured traits. According to the HMRPGV index, G5, G12, G7, G2, and G1 were identified as the top-performing genotypes. In conclusion, two genotypes, G12 and G5, exhibited a significant advantage over all other genotypes; however, only G5 was selected based on all traits in the MTSI index and could therefore undergo selection or cultivar introduction processes.

References

- Ahakpaz F., Abdi H., Neyestani E., Hesami A., Mohammadi B., Nader Mahmoudi K., Abedi-Asl G., Jazayeri Noshabadih M.R., Ahakpaz F. Genotype-by-environment interaction analysis for grain yield of barley genotypes under dryland conditions and the role of monthly rainfall. *Agric Water Manage.* 2021;245. doi 10.1016/j.agwat.2020.106665
- Akter A., Hasan M.J., Kulsum M.U., Rahman M.H., Paul A.K., Lipi L.F., Akter S. Genotype $\times$ environment interaction and yield stability analysis in hybrid rice (*Oryza sativa* L.) by AMMI biplot. *Bangladesh Rice J.* 2015;19(2):83-90. doi.org/10.3329/brj.v19i2.28168
- Allahgholipour M. Analysis of grain yield stability of new rice (*Oryza sativa* L.) genotypes originated from Iranian local cultivars. *Iranian J Crop Sci.* 2017;18(4):288-301. (in Persian with English abstract).
- Balestre M., dos Santos V.B., Soares A.A., Reis M.S. Stability and adaptability of upland rice genotypes. *Crop Breed Appl Biotech.* 2010;10:357-363. doi 10.1590/S1984-70332010000400011
- Bocianowski J., Niemann J., Nowosad K. Genotype-by-environment interaction for seed quality traits in interspecific cross-derived *Brassica* lines using additive main effects and multiplicative interaction model. *Euphytica.* 2019;215(1):7. doi 10.1007/s10681-018-2328-7
- Bose L.K., Jambhulkar N.N., Pande K., Singh O.N. Use of AMMI and other stability statistics in the simultaneous selection of rice genotypes for yield and stability under direct-seeded conditions. *Chil J Agric Res.* 2011;74(1):3-9. doi 10.4067/S0718-58392014000100001
- Cheema A.A. Mutation breeding for rice improvement in Pakistan: achievements and impact. *Plant Mut Rep.* 2006;1(1):36-39. https://www-pub.iaea.org/MTCD/publications/PDF/Newsletters/PMR-01-01.pdf
- Coan M.M.D., Marchioro V.S., Franco F.A., Pinto R.J.B., Scapim C.A., Baldissera J.N.C. Determination of genotypic stability and adaptability in wheat genotypes using mixed statistical models. *J Agric Sci Tech.* 2018;20:1525-1540. https://jast.modares.ac.ir/article-23-20190-en.pdf
- Colombari-Filho J.M., Resende M.D.V., de Moraes O.P., Castro A.P., Guimaraes E.L., Pereira J.M., Utumi M.M., Breseghello F. Upland rice breeding in Brazil: a simultaneous genotypic evaluation of stability, adaptability and grain yield. *Euphytica.* 2013;192:117-129. doi 10.1007/s10681-013-0922-2
- Dewi A.K., Dwimahyani I. Grain yield stability analysis of Jembar local rice mutant lines generated from mutation breeding. *IOP Conf Ser Earth Environ Sci.* 2019;230:012112. doi 10.1088/1755-1315/230/1/012112
- Dia M., Wehner T.C., Arellano C. Analysis of genotype $\times$ environment interaction (G $\times$ E) using SAS programming. *Agron J.* 2016;108:1838-1852. doi 10.2134/agronj2016.02.0085
- Donoso-Nanculao G., Paredes M., Becerra V., Arrepol C., Balzarini M. GGE biplot analysis of multi-environment yield trials of rice produced in a temperate climate. *Agric Res.* 2015;76(2):152-157. doi 10.4067/S0718-58392016000200003
- Dushyanthakumar B.M., Shadadashari Y.G. Stability analysis of P.U. Belliyappa local rice mutants. *Karnataka J Agric Sci.* 2007;20(4):724-726. http://14.139.155.167/test5/index.php/kjas/article/viewFile/1032/1024
- Ebadi A.A., Hallajian M.T., Kordrostami M. The genetic variation and stability analysis of rice mutant lines using AMMI model under normal and drought stress conditions. *Genetika.* 2019;51(2):687-699. doi 10.2298/GENSR1902687E
- Farshadfar E. Incorporation of AMMI stability value and grain yield in a single non-parametric index (GSI) in bread wheat. *Pakistan J Biol Sci.* 2008;11(14):1791-1796. doi 10.3923/pjbs.2008.1791.1796
- Gauch H.G. A simple protocol for AMMI analysis of yield trials. *Crop Sci.* 2013;53(5):1860-1869. doi 10.2135/cropsci2013.04.0241
- Kang M.S. A rank-sum method for selecting high-yielding, stable corn genotypes. *Cereal Res Comm.* 1988;16:113-115
- Karimizadeh R., Pezeshkpour P., Barzali M., Mehraban A., Sharifi P. Evaluation the mean performance and stability of lentil genotypes by combining features of AMMI and BLUP techniques. *J Crop Breed.* 2020;12(36):160-170. (In Persian with English abstract). doi 10.52547/jcb.12.36.160
- Khush G.S. What it will take to feed 5.0 billion rice consumers in 2030. *Plant Mol Biol.* 2005;59:1-6. doi 10.1007/s11103-005-2159-5
- Koundinya A.V.V., Pandit M.K., Ramesh D., Mishra P. Phenotypic stability of eggplant for yield and quality through AMMI, GGE and cluster analyses. *Sci Hort.* 2019;247:216-223. doi 10.1016/j.scienta.2018.12.019
- Nayak D., Bose L.K., Singh S., Nayak P. Additive main effects and multiplicative interaction analysis of host-pathogen relationship in rice-bacterial blight pathosystems. *Plant Path J.* 2008;24(3):337-351
- Olivoto T., Lúcio A.D. Metan: an R package for multi-environment trial analysis. *Meth Ecol Evol.* 2020;11:783-789. doi 10.1111/2041-210X.13384
- Olivoto T., Lúcio A.D.C., da Silva J.A.G., Marchioro V.S., de Souza V.Q., Jost E. Mean performance and stability in multi-environment trials I: combining features of AMMI and BLUP techniques. *Agron J.* 2019a;111(6):2949-2960. doi 10.2134/agronj2019.03.0220
- Olivoto T., Lúcio A.D.C., da Silva J.A.G., Sari B.G., Diel M.I. Mean performance and stability in multi-environment trials II: selection based on multiple traits. *Agron J.* 2019b;111(6):2961-2969. doi 10.2134/agronj2019.03.0221
- Olivoto T., Nardino M., Meira D., Meier C., Follmann D.N., de Souza V.Q., Konflanz V.A., Baretta D. Multi-trait selection for mean per-

- formance and stability in maize. *Agron J.* 2021;113:3968-3974. doi 10.1002/agj2.20741
- Patterson H.D., Thompson R. Recovery of interblock information when block sizes are unequal. *Biometrics.* 1971;58:545-554. doi 10.2307/2334389
- Piepho H.P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. *Theor Appl Genet.* 1994;89(5):647-654. doi 10.1007/BF00222462
- Rahayu S. Yield stability analysis of rice mutant lines using AMMI method. *J Phys Conf Ser.* 2020;1436:012019. doi 10.1088/1742-6596/1436/1/012019
- Resende M.D.V. Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético. Embrapa Florestas, Colombo, Brazil, 2007
- Rodvalho M.A., Coan M.M.D., Scapim C.A., Pinto R.J.B., Contreras-Soto R.I. Comparison of HMRPGV, Lin and Binns's and Annichiarico's methods for maize hybrid selection for high and stable yield. *Maydica.* 2015;60(1):1-7
- Santos F., Marza F. Selection of forage oat genotypes through GGE Biplot and BLUP. *bioRxiv.* 2020. doi 10.1101/2020.03.10.986422
- Searle S.R., Casella G., McCulloch C.E. Variance Components. John Wiley and Sons, New York, USA, 1992
- Sellami M.H., Lavini A., Pulvento C. Phenotypic and quality traits of chickpea genotypes under rainfed conditions in South Italy. *Agronomy.* 2021;11:962. doi 10.3390/agronomy11050962
- Sharifi P. Evolution, domestication, breeding methods and the latest breeding findings in rice. Agric Natural Res Engining Org of Iran. 2020. (in Persian)
- Sharifi P., Aminpanah H., Erfani R., Mohaddesi A., Abbasian A. Evaluation of genotype×environment interaction in rice based on AMMI model in Iran. *Rice Sci.* 2017;24(3):173-180. doi 10.1016/j.rsci.2017.02.001
- Sharifi P., Erfani A., Abbasian A., Mohaddesi A. Stability of some of rice genotypes based on WAASB and MTSI indices. *Iranian J Gen Plant Breed.* 2021;9(2):1-11. doi 10.30479/IJGPB.2021.14432.1283
- Shu Q.Y. Plant Mutation and Biotechnology. CABI, 2012
- van Eeuwijk F.A., Bustos-Korts D.V., Malosetti M. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Sci.* 2016;56(5):2119-2140. doi 10.2135/cropsci2015.06.0375
- Verma A., Singh G.P. Simultaneous application of AMMI measures and yield for stability analysis of wheat genotypes evaluated under irrigated late sown conditions of Central Zone of India. *J Appl Nat Sci.* 2020;12(4):541-549. doi 10.31018/jans.v12i4.2391
- Yan W. Analysis and handling of G × E in a practical breeding program. *Crop Sci.* 2016;56:2106-2118. doi 10.2135/cropsci2015.06.0336
- Yan W. A systematic narration of some key concepts and procedures in plant breeding. *Front Plant Sci.* 2021;12:724517. doi 10.3389/fpls.2021.724517
- Yan W., Hunt L.A., Shen Y., Szlavnick Z. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 2000;40:597-605 doi 10.2135/cropsci2000.403597x
- Yan W., Pageau D., Frégeau-Reid J., Durand J. Assessing the representativeness and repeatability of test locations for genotype evaluation. *Crop Sci.* 2011;51:1603-1610. doi 10.2135/cropsci2011.01.0016

Data availability statement. The data that support this study cannot be publicly shared due to ethical or privacy reasons and may be shared upon reasonable request to the corresponding author if appropriate.

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

Received March 11, 2025. Revised August 8, 2025. Accepted August 8, 2025.

doi 10.18699/vjgb-26-26

The role of explant type and selective agent application in the initial transformation rate of *Lens culinaris* Medik.

T.V. Dyubenko ¹, K.V. Smirnov ^{2, 3}, V.E. Tvorogova ^{2, 4} ¹ Federal Research Center the N.I. Vavilov All-Russian Institute of Plant Genetic Resources (VIR), St. Petersburg, Russia² Saint Petersburg State University, St. Petersburg, Russia³ All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia⁴ Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia krubaza@mail.ru

Abstract. *Lens culinaris* Medik. (lentil) is an agronomically important leguminous species, but its genome modification is rarely used for obtaining new varieties, probably due to a low efficiency of transformation protocols. Development of universal, genotype-independent protocols for obtaining transgenic plants usually relies, among other factors, on the possibility of obtaining a substantial number of transgenic cells *in vitro*. This study aimed to adapt a previously developed *Agrobacterium*-mediated transformation protocol, used for a related legume, for the production of transgenic callus tissue in *L. culinaris*. We used two different markers of transgenic tissue, beta-glucuronidase and green fluorescent protein, to find an optimal type of explant for obtaining transgenic tissue in lentil. We also evaluated the impact of hygromycin, a common selective agent, on the amount of transgenic tissue in developing transformed explants of *L. culinaris*. According to our results, the transformation protocol commonly used for *Medicago truncatula* Gaertn. leaf explants can also be applied for obtaining transgenic calli from *L. culinaris* shoot apices. Explants from shoot apices demonstrated higher initial transformation rate in comparison with explants from roots, stems and leaves. Moreover, explants of different types, which were cultivated on medium without hygromycin, developed significantly fewer calli expressing reporter genes than those grown on hygromycin-containing medium, confirming that hygromycin may be used as an effective selection agent for lentil. During our analysis, we noticed GUS-like staining in calli which didn't contain plasmids for *GUS* gene expression. This can be explained with so-called intrinsic GUS-like activity, which was described in previous research. These data can be used for further development of effective and universal *L. culinaris* transformation and genome editing protocols.

Key words: lentil; hygromycin; GUS; GFP; transformation; *Lens culinaris*

For citation: Dyubenko T.V., Smirnov K.V., Tvorogova V.E. The role of explant type and selective agent application in the initial transformation rate of *Lens culinaris* Medik. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed.* 2026;30(2):233-240. doi 10.18699/vjgb-26-26

Funding. The study is performed in the framework of the State Assignments for FRC VIR No. FGEM-2022-0002 "Identification of the possibilities of the gene pool of legumes to optimize their selection and diversification of use in various sectors of the national economy". The research was also funded by the Sirius University of Science and Technology project: PBB-RND-2243.

Acknowledgements. Authors are grateful to Veronika Y. Simonova for the help with literature analysis.

Роль типа экспланта и применения селективного агента в первичной эффективности трансформации *Lens culinaris* Medik.

Т.В. Дюбенко ¹, К.В. Смирнов ^{2, 3}, В.Е. Творогова ^{2, 4} ¹ Федеральный исследовательский центр Всероссийский институт генетических ресурсов растений им. Н.И. Вавилова (ВИР), Санкт-Петербург, Россия² Санкт-Петербургский государственный университет, Санкт-Петербург, Россия³ Всероссийский научно-исследовательский институт сельскохозяйственной микробиологии, Пушкин, Санкт-Петербург, Россия⁴ Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия krubaza@mail.ru

Аннотация. *Lens culinaris* Medik. (чечевица) – важная сельскохозяйственная культура из семейства бобовых, но модификация ее генома редко используется для создания новых сортов. Вероятно, это связано с низкой эффективностью существующих протоколов трансформации. Как правило, разработка универсальных, независимых от генотипа протоколов получения трансгенных растений зависит, среди прочего, от возможности образования существенного количества трансгенных клеток *in vitro*. В настоящем исследовании мы предприняли попытку адаптировать разработанный ранее для другого бобового протокол агробактериальной трансформации для получения трансгенной каллусной ткани у *L. culinaris*. Чтобы оценить эффективность трансформации, мы выбрали две независимые репортерные системы: бета-глюкуронидазу и зеленый флуоресцентный белок. С помощью этих систем мы нашли оптимальный тип экспланта для создания каллусной ткани чечевицы, экспрессирующей рекомбинантную ДНК. Мы также оценили влияние гигромицина, одного из распространенных селективных агентов, на количество

трансгенной ткани в развивающихся трансформированных эксплантах *L. culinaris*. Согласно нашим результатам, протокол трансформации, который обычно применяется для эксплантов листьев *Medicago truncatula* Gaertn., может быть использован и для получения трансгенных каллусов из верхушек побегов *L. culinaris*. Экспланты из апексов побегов продемонстрировали более высокую первичную эффективность трансформации по сравнению с эксплантами из корней, стеблей и листьев. Кроме того, экспланты разных типов, культивируемые на среде без гигромицина, образовывали значительно меньше каллусов, экспрессирующих репортерные гены, по сравнению с эксплантами, выращиваемыми на среде с гигромицином. Это подтверждает возможность использования гигромицина в качестве эффективного селективного агента для чечевицы. В ходе нашего анализа мы также обнаружили GUS-подобное окрашивание в каллусах, не содержащих плазмид для экспрессии гена *GUS*, что можно объяснить так называемой внутренней GUS-подобной активностью, которая была описана в предыдущих исследованиях. Эти данные могут быть полезны для дальнейшей разработки эффективных и универсальных протоколов трансформации и редактирования генома *L. culinaris*.

Ключевые слова: чечевица; гигромицин; GUS; GFP; трансформация; *Lens culinaris*

Introduction

Lentils (*Lens culinaris* Medik.) are one of the oldest domesticated crops (Erskine et al., 2009), the cultivation history of which dates back over 8,000 years (Liber et al., 2021). Lentil production has doubled since the beginning of the 21st century (Kaale et al., 2023), coinciding with a rising societal interest in plant-based meat alternatives and growing trends towards more sustainable food production (Runte et al., 2024). Being a leguminous plant, *L. culinaris* is used both as a nitrogen-fixing crop improving the soil fertility and as a source of protein for human nutrition. The level of carbohydrates, including starch and dietary fiber, is also quite high in lentils, making them a valuable crop among other legumes. Furthermore, biochemical properties, specific for different lentil cultivars, provide a wide range of lentil applications in food production (Kaale et al., 2023).

Despite many features, which make lentils a valuable crop, its production and usage are related with diverse obstacles, including undesirable traits common for most legumes and also drawbacks which are specific for this culture. The features which require alleviation in lentil, include high content of different anti-nutrients, such as phytate, as well as low levels of some amino acids (Joshi et al., 2017), pod shattering (Cao et al., 2024), etc.

Advancements in next-generation sequencing led to the discovery of many genetic markers and genes related to different desirable and undesirable traits in lentils (Kumar et al., 2023). These data can be widely used for selection and breeding of lentil cultivars with specific useful features. However, currently this information may rarely be used for genome modification of lentils, because, similarly to many legumes, the transformation and regeneration effectiveness of this species is quite low. There are only a few studies on *L. culinaris* transformation, and, to our knowledge, all of them report on low effectiveness of obtaining transgenic lentil plants (Gulati et al., 2002; Celikkol Akcay et al., 2009; Khatib et al., 2011; Chopra et al., 2012; Das et al., 2012, 2019; Polowick, Yan, 2023), which doesn't exceed 7 percents (i.e. 7 transgenic plants per 100 explants) for individual cultivars. Moreover, these protocols are developed for specific lentil varieties and, therefore, their effectiveness may depend greatly on genotype transformed.

The transformation effectiveness depends both on the success of foreign DNA introduction into plant cells and on further fate of the transformed cell, i.e. its ability to give rise

to a new plant. Consequently, these parameters are influenced by many factors accompanying transformation, including the method of transformation, explant type, cultivation medium and the method of transformant selection. A significant part of plant transformation protocols includes the stage of obtaining transgenic callus. At this stage, modified cells proliferate and some of them acquire the capacity for regeneration. Therefore, the efficiency of transgenic callus formation plays an important role in overall transformation efficiency. Moreover, maximization of transgenic tissue yield for a plant species facilitates the screening for morphogenic regulators, i.e. genes that are able to induce plant regeneration from transgenic tissue and can be used for effective cultivar-independent transformation (Bakulin et al., 2025).

In the current study, we made an attempt to find an optimal type of explant for obtaining transgenic callus tissue in lentil. We also evaluated the impact of hygromycin, a common selective agent, on the amount of transgenic tissue in developing transformed explants of *L. culinaris*. We built our transformation protocol on the highly effective transformation system used for *Medicago truncatula*, a well-studied model leguminous plant (Cosson et al., 2006). This system includes the formation of a significant amount of transgenic callus tissue. Moreover, as a model legume, *M. truncatula* provides an efficient platform for identifying key morphogenic regulators (Yakovleva et al., 2024), which can subsequently be validated in *L. culinaris*. The optimized protocol presented in this study provides a foundation for such validation experiments.

Materials and methods

Plant material and bacterial strains. *L. culinaris* Ekaterinskaya zelenosemyannaya cultivar (k-2942) seeds from the collection of the N.I. Vavilov All-Russian Institute of Plant Genetic Resources (VIR) were used in the study. For cloning, the *Escherichia coli* DH10B strain was used. For *Agrobacterium*-mediated transformation, the *Rhizobium radiobacter* (*Agrobacterium tumefaciens*) AGL1 strain was used. *E. coli* and *A. tumefaciens* cultivation and transformation were performed as described in (Potsenkovskaia et al., 2022).

Plasmids used in the study. For *L. culinaris* transformation, two plasmids were used: pMDC32_GUS (Tvorogova et al., 2019) with the beta-glucuronidase (GUS) gene and pMDC32_eGFPper obtained in this study. To construct the pMDC32_eGFPper plasmid, the *eGFPper* gene, encoding the enhanced green fluorescent protein (GFP) with endoplasmic re-

ticulum localization signal, was amplified from the pB7WG2D vector (Karimi et al., 2002) with primers containing attB1 and attB2 sites (forward primer: GGGGACAAGTTTGTACAAAAAAGCAGGCTTCATGGTGAAGACTAATCTTTTCTC; reverse primer: GGGGACCACTTTGTACAGAAAGCTGGGTCTTACAGCTCGTCCTTCTTG) and cloned into the pDONR207 entry vector (Invitrogen) and then to the pMDC32 vector (Curtis, Grossniklaus, 2003) using Gateway technology (Invitrogen).

Agrobacterium-mediated transformation. For explants preparation, seeds of *L. culinaris* were incubated in concentrated sulfuric acid for 7 minutes, rinsed with tap water 7 times, then incubated in 5–7 % sodium hypochlorite solution (the “Belizna” bleach) for 8 minutes and rinsed with sterile water 10 times in a laminar hood. After sterilization, seeds were put on modified Fahraeus medium (Fähræus, 1957; Tvorogova et al., 2019) and left to germinate for 12 days at 21–24 °C, 16 h (light)/8 h (dark) photoperiod. After that, germinated seedlings were used to prepare explants. Root explants (about 10–15 mm length) were obtained from root tips, shoot apex explants (about 10–15 mm length) were taken from the main and/or lateral shoots. Stem explants (about 3–8 mm length) were obtained from stem fragments between the first and the second leaves of the seedlings. For leaf explants, one or two first true leaves were taken. Seedling growth was asynchronous, and the explant size depended on the seedling development stage.

For *Agrobacterium*-mediated transformation itself, a modified protocol of Cosson et al. (2006) for *M. truncatula* Gaertn. was used. *Agrobacterium* strain carrying either pMDC32_GUS or pMDC32_eGFPper plasmids was recovered from glycerol stock (500 µl of 50 % glycerol mixed with 500 µl of overnight stationary-phase culture) stored at –80 °C and cultured on solid YEP medium (Table S1)¹ containing kanamycin (50 mg/l) and rifampicin (40 mg/l). Kanamycin resistance was provided by the pMDC32_GUS or pMDC32_eGFPper plasmid, whereas rifampicin resistance was characteristic of the AGL1 strain. On such medium, bacteria were cultivated at 30 °C for 1–2 days. Then, a small amount of bacteria from the Petri dish was sown to the liquid YEP with the same concentrations of antibiotics, in which they were grown overnight in a thermoshaker at 30 °C and 200 rpm. On the next day, 2 ml of bacterial culture were transferred to 50 ml of AB-MES medium (Wu et al., 2014, Table S1) with the addition of 40 mg/l (200 µM) acetosyringone, kanamycin (50 mg/l) and rifampicin (40 mg/l). Bacteria were cultivated in this medium at 30 °C and 200 rpm for about 3 hours. Then, they were centrifuged (4,000g, 15 minutes, room temperature) and resuspended in liquid ABM-SH medium (Table S1) with 4 mg/l (18 µM) 2,4-dichlorophenoxyacetic acid (2,4-D), 0.5 mg/l (2.22 µM) 6-benzylaminopurine (BAP), and 40 mg/l (200 µM) acetosyringone to the final OD600 0.1. ABM-SH represented a mixture of 0.5X AB-MES components and 0.5X modified liquid PCI-4 medium (Hoffmann et al., 1997) components, similarly to the infiltration medium from the AGROBEST protocol for *Arabidopsis thaliana* (L.) Heynh. (Wu et al., 2014).

Prepared explants were incubated in *Agrobacterium* suspension in ABM-SH medium for 15 minutes with slow gentle agitation. After the inoculation stage, explants were blotted on sterile filter paper, then put on solid ABM-SH medium (Table S1) and cultivated in the dark at 21–24 °C for about 36 hours. After cocultivation, explants were put on solid PCI-4 medium (Hoffmann et al., 1997, Table S1) for callus formation with 4 mg/l (18 µM) 2,4-D, 0.5 mg/l (2.22 µM) BAP, and with the addition of cefotaxime (250 mg/l), to eliminate agrobacteria. In some experiments, hygromycin (10 mg/l), necessary for the selection of transgenic plant cells, was added to the medium. On this medium, explants were cultivated for 44 days, with transfer on the fresh medium every 14–16 days. At the end of cultivation, explants were either stained with X-Gluc and analysed with stereomicroscope to detect GUS expression, or analysed with the fluorescent stereomicroscope to detect GFP fluorescence.

GUS staining. For GUS staining, X-Gluc substrate (Sisco Research Laboratories Pvt. Ltd.) was used. Plant tissues were collected in NT buffer (100 mM Tris-HCl, 50 mM NaCl) and then transferred to X-Gluc buffer (100 mM Tris-HCl, 50 mM NaCl, 0.2 mM K₃[Fe(CN)₆], 2.5 mM X-Gluc), where they were incubated at 37 °C overnight. After staining, tissues were transferred to 70 % ethanol. Photos of single explants were taken with a ZEISS SteREO Discovery.V12 stereomicroscope. Photos of Petri dishes with explants were taken with the Vilber Fusion-FX6.Edge imaging system. For microscopic images, pieces of callus tissue were squashed and examined under a Leica DM500 microscope.

Statistical analysis and software. For evaluation of explant type and hygromycin presence impact, 20–28 explants per group were used. Statistical significance of differences was evaluated with Fisher’s exact test with Bonferroni adjustment for multiple comparisons, if necessary. Statistical analysis of results and diagram drawing were performed in the R studio using the vcd (Meyer et al., 2006, 2024) and dplyr (Wickham et al., 2025) packages. Final figures were created using the Inkscape software.

Results

Evaluation of transgenic tissue development from different explant types using the GFP marker

We decided to build our transformation protocol on the transformation system used for *M. truncatula* (Cosson et al., 2006), in which transgenic plants are developed from transformed leaf explants through somatic embryogenesis. To find the optimal variant of explant type for transformation, we performed *Agrobacterium*-mediated transformation of root, stem, leaf and shoot apex explants with plasmid pMDC32_eGFPper for eGFPper overexpression. After cocultivation, explants were transferred to the callus-inducing medium with or without hygromycin. The concentrations of plant growth regulators and cefotaxime for agrobacteria elimination were identical to the ones used for *M. truncatula*, but the concentration of hygromycin was lower (10 mg/l instead of 25 mg/l used for *M. truncatula*), because 10 mg/l concentration is successfully used for other legumes to obtain transgenic tissue (Olhoft et

¹ Supplementary Table S1 and Figures S1, S2 are available at: <https://vavilovj-icg.ru/download/pict-2026-30/appx17.pdf>

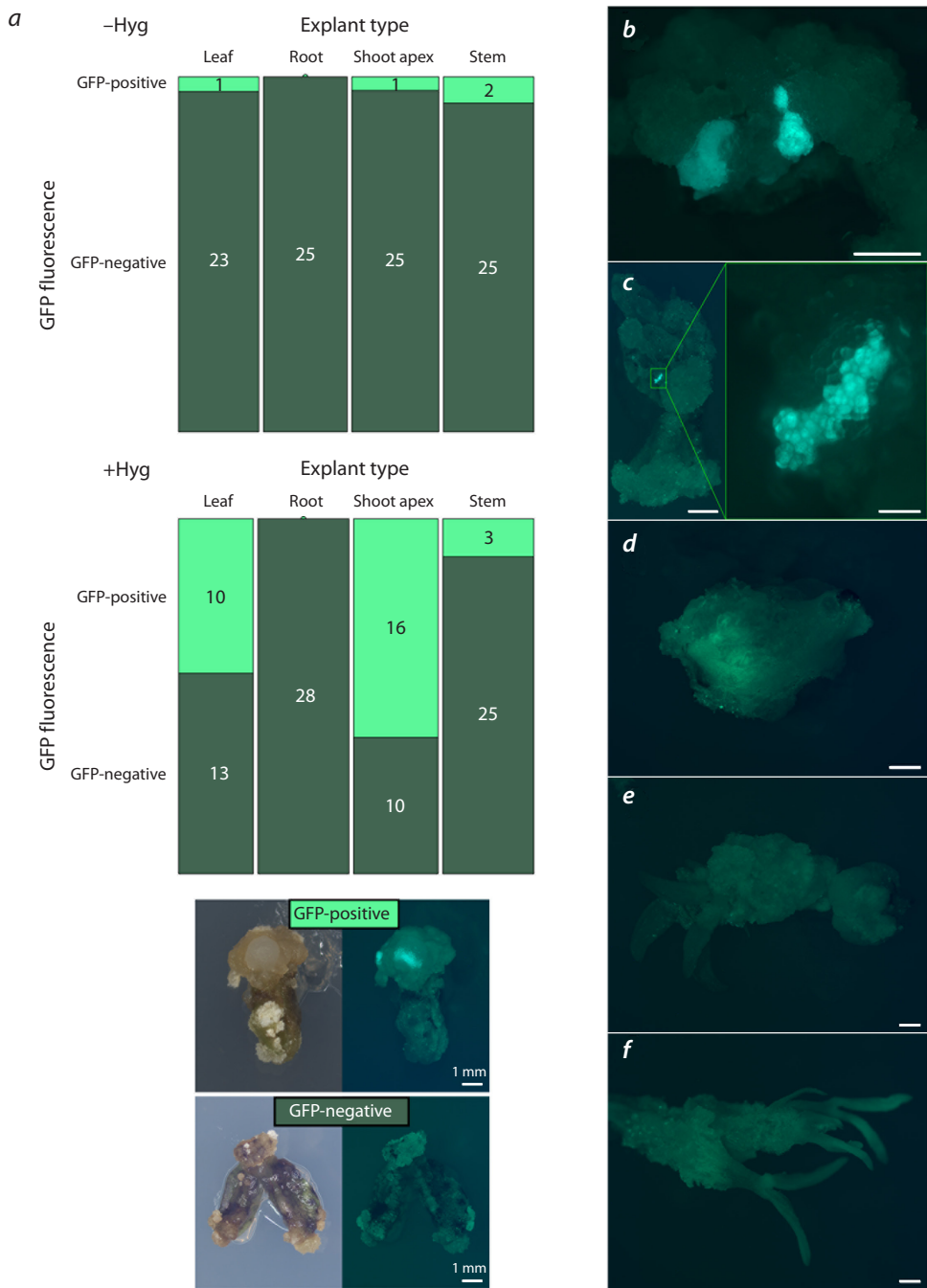


Fig. 1. *a*, Mosaic plots demonstrating the number of calli with GFP fluorescence (GFP-positive) and without GFP fluorescence (GFP-negative) developed from different types of explants, on the medium without hygromycin (above) or on the medium with addition of 10 mg/l hygromycin (below). In a legend below, the photos of explants were taken in white light (left) and fluorescent (right) light. *b–d*, Calli with GFP fluorescence. *e, f*, Calli without detectable GFP fluorescence, obtained after explant transformation with the pMDC32_GFP (*e*) or pMDC32_GUS (*f*) plasmids. Size bar is 1 mm for *b–f* and 100 µm for the inset in *c*.

al., 2003; Li et al., 2023; Tisseyre et al., 2024). After 46 days of cultivation on this callus-inducing medium, calli were developed from all types of explants both on the medium with hygromycin and hygromycin-free medium (Figs. S1, S2). We analysed GFP fluorescence to evaluate the effectiveness of transformation.

To avoid false positive results, we also analysed the green fluorescence in the explants transformed with pMDC32_GUS

construction, which didn't contain the *GFP* gene. Bright GFP fluorescence in separate tissue areas (Fig. 1*b, c*), as well as less bright diffuse fluorescence (Fig. 1*d*), were easily distinguishable from autofluorescence (Fig. 1*e, f*), which was observed in control pMDC32_GUS explants. It is worth noting that some fluorescent areas with weak signal, as well as autofluorescent areas, can only be seen with high magnification (60× and more). Therefore, for robust comparative analysis, it was

crucial to evaluate the presence of fluorescence for all explants without changing magnification and other visualization settings on the stereomicroscope.

According to our results, on the medium supplemented with 10 mg/l hygromycin, the type of explant had an impact on the frequency of fluorescent calli (Fisher’s exact test, p -value $4.359\text{e-}08$). The highest frequency of transformed calli with GFP fluorescence was detected for explants from shoot apices (Fig. 1a). Leaf explants also demonstrated a rather high frequency of transformation, whereas only 3 out of 28 stem explants developed fluorescent calli. Calli from the root explants didn’t demonstrate any transgenic fluorescent tissue at all (Fig. 1a). Overall, shoot apex explants formed significantly more fluorescent calli in comparison with explants from roots or stems (Fisher’s exact test with Bonferroni adjustment for multiple comparisons, p -values $1.51\text{e-}06$ and $8.46\text{e-}04$, respectively), whereas differences between explants from shoot apices and leaves were not statistically significant.

At the same time, explants of all types, cultivated on medium without hygromycin, developed significantly fewer fluorescent calli than those grown on hygromycin-containing medium (Fisher’s exact test, p -value $2.09\text{e-}06$), confirming

that hygromycin may be used as an effective selection agent for lentil (Fig. 1a).

Evaluation of transgenic tissue development from different explant types using GUS marker

We performed a similar experiment with transformation of different explant types using GUS, another transgenic tissue marker. Explants were transformed with the pMDC32_GUS plasmid (Tvorogova et al., 2019) using the protocol mentioned above. After 46 days of cultivation (starting from the day of transformation), GUS activity was evaluated in explants. Usage of GUS allowed us to detect single stained cells (at about 80× magnification on stereomicroscope), which we were not able to visualize with the GFP marker. The results of transgenic tissue detection were similar to those obtained using the GFP marker. On the media with hygromycin, the highest transformation efficiency was shown for shoot apex explants, followed by leaf explants. Both root and stem explants demonstrated a low transformation rate (Fig. 2). In total, shoot apex explants formed significantly more stained calli in comparison with explants from leaves, roots or stems (Fisher’s exact test with Bonferroni adjustment for multiple comparisons,

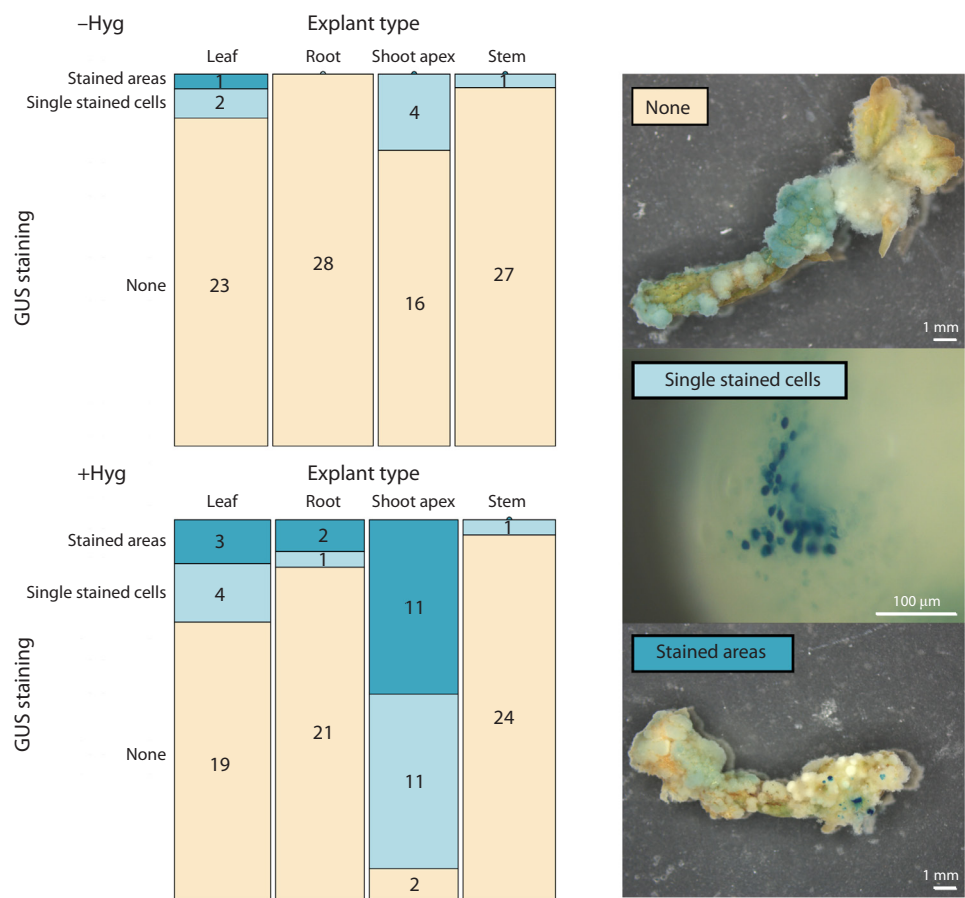


Fig. 2. Mosaic plots demonstrating the number of stained and non-stained calli developed from different types of explants after transformation with pMDC32_GUS plasmid, on the medium without hygromycin (above) or on the medium with addition of 10 mg/l hygromycin (below). The legend on the right demonstrates calli belonging to different categories relative to GUS staining: none – no detectable GUS staining apart from the light-blue color characteristic of endogenous GUS-like activity; single stained cells – small spots of GUS staining visible only with high (about 80×) magnification, stained areas – patches of GUS staining visible with naked eye or with low magnification.

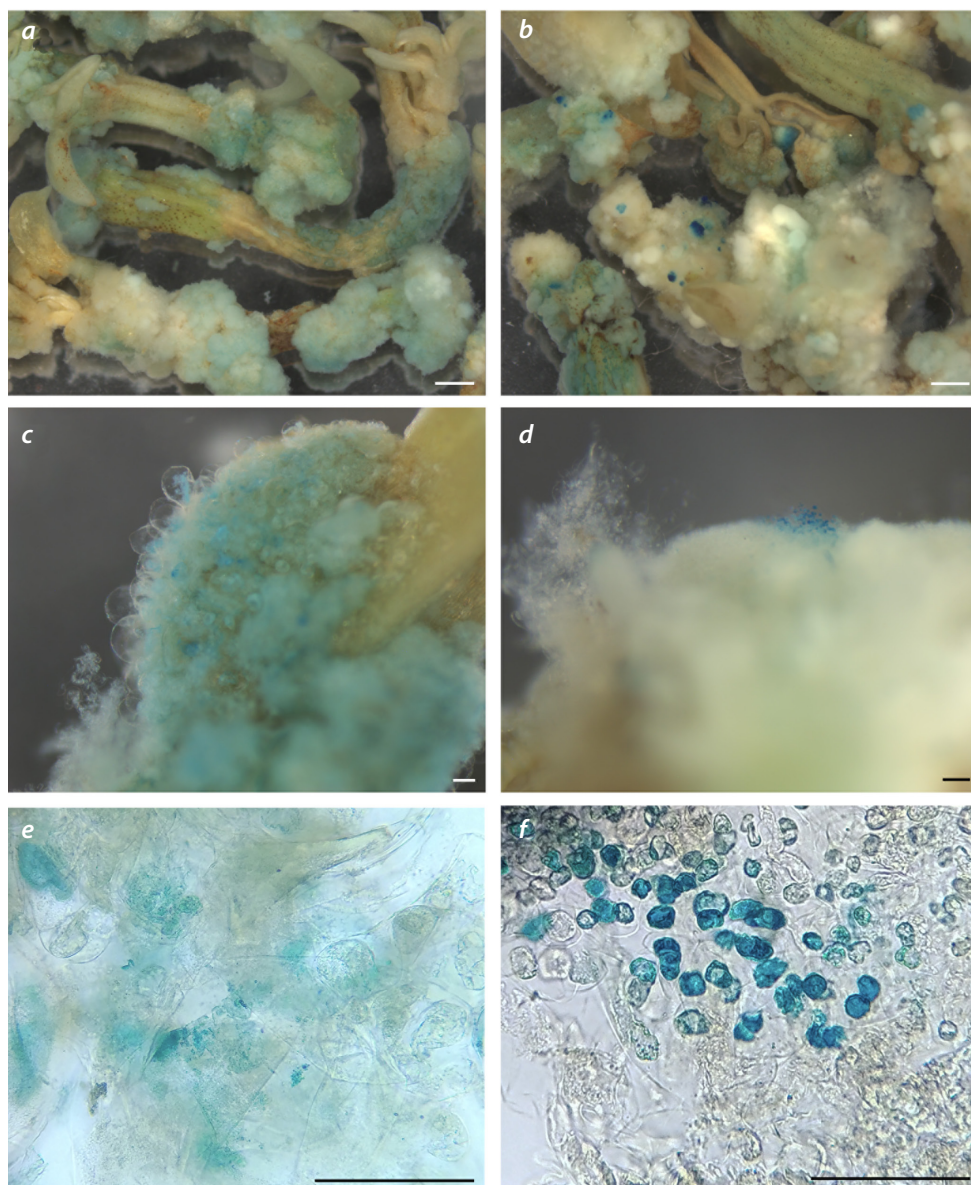


Fig. 3. GUS staining of control calli obtained from explants transformed with the pMDC32_eGFP_{er} plasmid (a, c, e) or calli obtained from explants transformed with the pMDC32_GUS plasmid (b, d, f).

Size bar is 1 mm for a, b and 100 μ m for c–f.

p -values $6.243734\text{e-}05$, $2.191140\text{e-}07$ and $7.429123\text{e-}10$, respectively). In contrast, all types of explants grown on the hygromycin-free media mostly didn't develop detectable transgenic tissues. The difference was statistically significant when compared to explants cultivated on hygromycin-containing medium (Fisher's exact test, p -value $5.842\text{e-}12$), confirming our results obtained for the GFP marker (Fig. 1).

Evaluation of GUS false positive staining

During the experiment with GUS activity analysis described above, we also performed GUS staining of explants transformed with the pMDC32_eGFP_{er} construction, which didn't carry the *GUS* gene. Unexpectedly, these control explants also demonstrated blue staining, although the staining pattern differed from the one provided by true *GUS* expression (Hu

et al., 1990). In false positives, the blue coloration was more diffuse and didn't have distinct borders compared with true GUS staining (Fig. 3).

Microscopic analysis also revealed differences between false positive and true GUS staining. *GUS*-expressing cells had deep blue coloring (Fig. 3f), which was not the case for false positive GUS-like staining (Fig. 3e). Therefore, the GUS marker can be used to detect transgenic tissue in lentil, but the staining of negative controls is necessary.

Discussion

Transformation is a powerful tool for obtaining new varieties of different plant species (Garcia Ruiz et al., 2018; Noack et al., 2024), which is why improving its effectiveness for lentil is important. Nevertheless, current *L. culinaris* transforma-

tion protocols have rather low efficiency, probably due to the low regeneration rate. One of the possible solutions for this problem is the usage of morphogenic regulators, i.e. genes that can be added to transformation vectors and that are able to induce proliferation of transgenic cells and plant regeneration from transgenic tissue. Still, the search for such morphogenic regulators, suitable for lentils, can be hampered with low initial transformation rate, i.e. low effectiveness of transgenic callus formation. Therefore, the development of effective protocols for maximizing the transgenic tissue amount can be very useful for further optimization of obtaining transformed and edited plants.

The protocols for obtaining transgenic lentil plants usually involve cotyledons and/or cotyledonary nodes, parts of embryos or embryo axes (Gulati et al., 2002; Celikkol Akcay et al., 2009; Khatib et al., 2011; Das et al., 2012, 2019; Polowick, Yan, 2023) or whole seeds (Chopra et al., 2012). In our study, we tried to perform *Agrobacterium*-mediated transformation using different types of explants obtained from 12-day old seedlings. According to our results, the transformation protocol, developed for *M. truncatula* leaf explants, can also be used for obtaining transgenic tissue from *L. culinaris* explants. Still, although the leaf explants are used in this protocol for *M. truncatula*, a different type of explant, shoot apex, gave better results for *L. culinaris*. It can possibly be related to a relatively high proportion of stem cells and, in general, proliferating cells in the shoot apex in comparison with other explant types. It would be interesting to evaluate the efficiency of usage of seedling shoot apices for *M. truncatula* transformation. The usage of hygromycin at the concentration 10 mg/l allowed us to obtain a significant amount of transgenic tissue from shoot apex explants. Nevertheless, the optimal hygromycin concentration has to be defined in the future.

There are also several studies aiming to obtain transgenic tissues in lentil from diverse explant types, including different parts of germinated seedlings (Warkentin, McHughen, 1992), cotyledonary nodes (Mahmoudian et al., 2002) and embryo axes (Lurquin et al., 1998), although in most cases, no selection agent and no 2,4-D-mediated callus induction was applied, in contrast to our research. As far as we know, the current study is also the first in which optimization of lentil transformation was performed with the usage of two different transgenic tissue markers.

In a study by Warkentin and McHughen (1992) *Agrobacterium*-mediated transformation of lentil shoot apex explants yielded a high percentage of explants, which contained transgenic tissue, without selection for transgenic cells at early stages of cultivation. In our study, cultivation on the medium without a selection agent was not effective, whereas hygromycin application made it possible to obtain a significant amount of transgenic tissue. The discrepancy between our results and results obtained previously may be due to differences in genotypes, *in vitro* cultivation and transformation conditions.

In experiments on evaluation of *in vitro* regeneration capacity in lentil, explants from shoot tips were shown to give callus on the medium with 2,4-D, as well as regenerating shoots on other types of media, with the addition of BAP, IAA or NAA (Polanco et al., 1988). Shoot regeneration from shoot apex

explants was also observed after *Agrobacterium*-mediated transformation, although shoots regenerated during kanamycin selection were non-transgenic (Warkentin, McHughen, 1992). Therefore, the development of a protocol for producing transgenic lentil plants from shoot tip explants may be feasible upon identification of suitable selection and regeneration conditions. Based on our data, hygromycin selection appears to be a promising approach.

During our analysis, we noticed the GUS-like staining in calli that didn't contain plasmids for *GUS* gene expression. This can be explained with so-called intrinsic GUS-like activity, which was described in previous research (Hu et al., 1990). Indeed, beta-glucuronidase (GUS) genes or proteins were described in several species, including *Scutellaria baicalensis* (Zuo et al., 2025) and *Zea mays* (Muhitch, 1998), whereas endogenous GUS activity was detected in *Brassica napus* (Abdollahi et al., 2011), *Oryza sativa*, *A. thaliana*, *Nicotiana tabacum* and other plants (Sudan et al., 2006). Still, in other studies involving GUS staining of lentil explants, no false positive GUS staining was detected on control explants, which may be due to different staining buffer composition or different tissue features (Warkentin, McHughen, 1992; Lurquin et al., 1998; Mahmoudian et al., 2002).

Conclusion

Our results show that *Agrobacterium*-mediated transformation of lentil shoot apices according to the protocol developed for *M. truncatula*, can be successfully used for obtaining transgenic callus tissues. These data can be used for further development of effective and universal lentil transformation and editing protocols.

References

- Abdollahi M.R., Memari H.R., van Wijnen A.J. Factor affecting the endogenous β -glucuronidase activity in rapeseed haploid cells: how to avoid interference with the GUS transgene in transformation studies. *Gene*. 2011;487(1):96-102. doi 10.1016/j.gene.2011.07.007
- Bakulin S.D., Monakhos S.G., Bruskin S.A. Morphogenetic factors as a tool for enhancing plant regeneration capacity during *in vitro* transformation. *Int J Mol Sci*. 2025;26(17):8583. doi 10.3390/ijms26178583
- Cao Z., Socquet-Juglard D., Daba K., Vandenberg A., Bett K.E. Understanding genome structure facilitates the use of wild lentil germplasm for breeding: a case study with shattering loci. *Plant Genome*. 2024;17(2):e20455. doi 10.1002/tpg2.20455
- Celikkol Akcay U., Mahmoudian M., Kamci H., Yucel M., Oktem H.A. *Agrobacterium tumefaciens*-mediated genetic transformation of a recalcitrant grain legume, lentil (*Lens culinaris* Medik). *Plant Cell Rep*. 2009;28(3):407-417. doi 10.1007/s00299-008-0652-4
- Chopra R., Aparna, Saini R. Use of sonication and vacuum infiltration for *Agrobacterium*-mediated transformation of an Indian lentil (*Lens culinaris* Medik.) cultivar. *Sci Hortic*. 2012;143:127-134. doi 10.1016/j.scienta.2012.06.019
- Cosson V., Durand P., d'Erfurth I., Kondorosi A., Ratet P. *Medicago truncatula* transformation using leaf explants. *Methods Mol Biol*. 2006;343:115-127. doi 10.1385/1-59745-130-4:115
- Curtis M.D., Grossniklaus U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol*. 2003;133(2):462-469. doi 10.1104/pp.103.027979
- Das S.K., Shethi K.J., Hoque M.I., Sarker R.H. *Agrobacterium*-mediated genetic transformation in lentil (*Lens culinaris* Medik.) followed by *in vitro* flowering and seed formation. *Plant Tissue Cult Biotechnol*. 2012;22(1):13-26. doi 10.3329/ptcb.v22i1.11243

- Das S.K., Shethi K.J., Hoque M.I., Sarker R.H. *Agrobacterium*-mediated genetic transformation of lentil (*Lens culinaris* Medik.) with chitinase gene followed by *in vitro* flower and pod formation. *Plant Tissue Cult Biotechnol.* 2019;29(1):99-109. doi 10.3329/ptcb.v29i1.41982
- Erskine W., Muehlbauer F., Sarker A., Sharma B. (Eds) The Lentil: Botany, Production and Uses. Wallingford: CABI, 2009
- Fähraeus G. The infection of clover root hairs by nodule bacteria studied by a simple glass slide technique. *J Gen Microbiol.* 1957; 16(2):374-381. doi 10.1099/00221287-16-2-374
- Garcia Ruiz M.T., Knapp A.N., Garcia-Ruiz H. Profile of genetically modified plants authorized in Mexico. *GM Crops Food.* 2018;9(3): 152-168. doi 10.1080/21645698.2018.1507601
- Gulati A., Schryer P., McHughen A. Production of fertile transgenic lentil (*Lens culinaris* Medik) plants using particle bombardment. *In Vitro Cell Dev Biol Plant.* 2002;38(4):316-324. doi 10.1079/IVP.2002303
- Hoffmann B., Trinh T.H., Leung J., Kondorosi A., Kondorosi E. A new *Medicago truncatula* line with superior *in vitro* regeneration, transformation, and symbiotic properties isolated through cell culture selection. *Mol Plant-Microbe Interact.* 1997;10(3):307-315. doi 10.1094/MPMI.1997.10.3.307
- Hu C.Y., Chee P.P., Chesney R.H., Zhou J.H., Miller P.D., O'Brien W.T. Intrinsic GUS-like activities in seed plants. *Plant Cell Rep.* 1990; 9(1):1-5. doi 10.1007/BF00232123
- Joshi M., Timilsena Y., Adhikari B. Global production, processing and utilization of lentil: a review. *J Integr Agric.* 2017;16(12):2898-2913. doi 10.1016/S2095-3119(17)61793-3
- Kaale L.D., Siddiq M., Hooper S. Lentil (*Lens culinaris* Medik) as nutrient-rich and versatile food legume: a review. *Legume Sci.* 2023; 5(2):e169. doi 10.1002/leg3.169
- Karimi M., Inzé D., Depicker A. GATEWAY™ vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci.* 2002;7(5): 193-195. doi 10.1016/S1360-1385(02)02251-3
- Khatib F., Makris A., Yamaguchi-Shinozaki K., Kumar S., Sarker A., Erskine W., Baum M. Expression of the *DREB1A* gene in lentil (*Lens culinaris* Medik. subsp. *culinaris*) transformed with the *Agrobacterium* system. *Crop Pasture Sci.* 2011;62(6):488-495. doi 10.1071/CP10351
- Kumar J., Gela T.S., Gupta D.S., Chandra A., Khazaei H. Recent advances in lentil genetics, genomics, and molecular breeding. In: Lentils: Production, Processing Technologies, Products, and Nutritional Profile. Hoboken: John Wiley & Sons Ltd., 2023;25-43. doi 10.1002/9781119866923.ch2
- Li G., Liu R., Xu R., Varshney R.K., Ding H., Li M., Yan X., ... Luo Y., Gao S., Wei P., Zong X., Yang T. Development of an *Agrobacterium*-mediated CRISPR/Cas9 system in pea (*Pisum sativum* L.). *Crop J.* 2023;11(1):132-139. doi 10.1016/j.cj.2022.04.011
- Liber M., Duarte I., Maia A.T., Oliveira H.R. The history of lentil (*Lens culinaris* subsp. *culinaris*) domestication and spread as revealed by genotyping-by-sequencing of wild and landrace accessions. *Front Plant Sci.* 2021;12:628439. doi 10.3389/fpls.2021.628439
- Lurquin P.F., Cai Z., Stiff C.M., Fuerst E.P. Half-embryo cocultivation technique for estimating the susceptibility of pea (*Pisum sativum* L.) and lentil (*Lens culinaris* Medik.) cultivars to *Agrobacterium tumefaciens*. *Mol Biotechnol.* 1998;9(2):175-179. doi 10.1007/BF02760819
- Mahmoudian M., Yücel M., Öktem H.A. Transformation of lentil (*Lens culinaris* M.) cotyledonary nodes by vacuum infiltration of *Agrobacterium tumefaciens*. *Plant Mol Biol Rep.* 2002;20(3):251-257. doi 10.1007/BF02782460
- Meyer D., Zeileis A., Hornik K. The struclplot framework: visualizing multi-way contingency tables with vcd. *J Statistical Software.* 2006; 17(3):1-48. doi 10.18637/jss.v017.i03
- Meyer D., Zeileis A., Hornik K., Friendly M. vcd: visualizing categorical data. R package version 1.4-12. 2024. URL: <https://CRAN.R-project.org/package=vcd>
- Muhitch M.J. Characterization of pedicel β -glucuronidase activity in developing maize (*Zea mays*) kernels. *Physiol Plant.* 1998;104(3): 423-430. doi 10.1034/j.1399-3054.1998.1040318.x
- Noack F., Engist D., Gantois J., Gaur V., Hyjazie B.F., Larsen A., M'Gonigle L.K., Missirian A., Qaim M., Sargent R.D., Souza-Rodrigues E., Kremen C. Environmental impacts of genetically modified crops. *Science.* 2024;385(6712):eado9340. doi 10.1126/science.ado9340
- Olhoft P.M., Fligel L.E., Donovan C.M., Somers D.A. Efficient soybean transformation using hygromycin B selection in the cotyledonary-node method. *Planta.* 2003;216(5):723-735. doi 10.1007/s00425-002-0922-2
- Polanco M.C., Peláez M.I., Ruiz M.L. Factors affecting callus and shoot formation from *in vitro* cultures of *Lens culinaris* Medik. *Plant Cell Tissue Organ Cult.* 1988;15(2):175-182. doi 10.1007/BF00035759
- Polowick P.L., Yan W. A protocol for *Agrobacterium*-mediated genetic transformation of *Lens culinaris* Medik (lentil). *Plant Cell Tissue Organ Cult.* 2023;152(3):605-618. doi 10.1007/s11240-022-02434-x
- Potsenkovskaia E., Tvorogova V., Yakovleva D., Zlydneva N., Lutova L. Novel *NF-Y* genes expressed during somatic embryogenesis in *Medicago truncatula*. *Plant Gene.* 2022;31:100364. doi 10.1016/j.plgene.2022.100364
- Runte M., Guth J.N., Ammann J. Consumers' perception of plant-based alternatives and changes over time. A linguistic analysis across three countries and ten years. *Food Qual Preference.* 2024;113:105057. doi 10.1016/j.foodqual.2023.105057
- Sudan C., Prakash S., Bhomkar P., Jain S., Bhalla-Sarin N. Ubiquitous presence of β -glucuronidase (GUS) in plants and its regulation in some model plants. *Planta.* 2006;224(4):853-864. doi 10.1007/s00425-006-0276-2
- Tisseyre P., Cartiaux F., Chabrilange N., Gully D., Hoche V., Svistoonoff S., Gherbi H. Setting up *Agrobacterium tumefaciens*-mediated transformation of the tropical legume *Aeschynomene evenia*, a powerful tool for studying gene function in nod factor-independent symbiosis. *PLoS One.* 2024;19(4):e0297547. doi 10.1371/journal.pone.0297547
- Tvorogova V.E., Fedorova Y.A., Potsenkovskaya E.A., Kudriashov A.A., Efremova E.P., Kvitkovskaya V.A., Wolabu T.W., Zhang F., Tadege M., Lutova L.A. The WUSCHEL-related homeobox transcription factor MtWOX9-1 stimulates somatic embryogenesis in *Medicago truncatula*. *Plant Cell Tissue Organ Cult.* 2019;138(3): 517-527. doi 10.1007/s11240-019-01648-w
- Warkentin T.D., McHughen A. *Agrobacterium tumefaciens*-mediated beta-glucuronidase (GUS) gene expression in lentil (*Lens culinaris* Medik.) tissues. *Plant Cell Rep.* 1992;11(5-6):274-278. doi 10.1007/BF00235081
- Wickham H., François R., Henry L., Müller K., Vaughan D. dplyr: a grammar of data manipulation. R package version 1.1.4. 2025. URL: <https://CRAN.R-project.org/package=dplyr>
- Wu H.-Y., Liu K.-H., Wang Y.-C., Wu J.-F., Chiu W.-L., Chen C.-Y., Wu S.-H., Sheen J., Lai E.-M. AGROBEST: an efficient *Agrobacterium*-mediated transient expression method for versatile gene function analyses in *Arabidopsis* seedlings. *Plant Methods.* 2014; 10(1):19. doi 10.1186/1746-4811-10-19
- Yakovleva D., Efremova E., Smirnov K., Simonova V., Konstantinov Z., Tvorogova V., Lutova L. The *WOX* genes from the intermediate clade: influence on the somatic embryogenesis in *Medicago truncatula*. *Plants.* 2024;13(2):223. doi 10.3390/plants13020223
- Zuo X., Li P., Ren G., Bai Z., Jiang D., Liu C. Functional characterization of β -glucuronidase genes involved in baicalein biosynthesis from *Scutellaria baicalensis* based on transcriptome analysis. *Int J Mol Sci.* 2025;26(5):4410. doi 10.3390/ijms26051793

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

Received November 4, 2025. Revised January 12, 2026. Accepted January 13, 2026.

doi 10.18699/vjgb-26-27

Generation and characterization of a double-knockout *Arabidopsis thaliana* line lacking expression of *AOX1a* and *VTC2*

E.V. Garmash , E.S. Belykh , K.V. Yadrinhinskiy , R.V. Malyshev , I.O. Velegzhaninov 

Institute of Biology of Komi Science Centre of the Ural Branch of the Russian Academy of Sciences, Syktyvkar, Russia

 garmash@ib.komisc.ru

Abstract. In higher plants, the L-galactose pathway is the main pathway for the biosynthesis of vitamin C (ascorbate, Asc), the final step of which is connected with the functioning of the mitochondrial electron transport chain (ETC). In addition to the main cytochrome pathway, plant ETC includes an alternative pathway (AP) via alternative terminal oxidase (AOX). The engagement of AOX promotes Asc synthesis, and it is hypothesized that AOX suppression under conditions of Asc deficiency may reduce plant viability. The aim of this work was to examine the consequences of simultaneously knocking out two genes in *Arabidopsis thaliana*: *AOX1a*, the most stress-inducible AOX gene, and *VTC2*, encoding a key enzyme of the L-galactose pathway of Asc synthesis. Two lines of *A. thaliana* with T-DNA insertions in the target genes were crossed to generate hybrid lines. Seed characteristics of the first (F₁) and second (F₂) generations were analyzed. F₁ seeds were larger than those of parent lines, possibly due to heterosis. In the F₂ generation, self-pollination of F₁ plants resulted in seeds with significant size variation, including a group of small seeds with degenerative morphological abnormalities. Most of small seeds failed to germinate or died at the seedling stage. PCR genotyping of these seeds revealed the absence of native *AOX1a* and *VTC2* indicating a lethal mutation caused by simultaneous knockout of both genes. One likely genetic cause is the interaction of mutations in non-allelic genes. At the physiological level, irreversible respiratory damage may occur, possibly including the impact of a cryptic mutation in the *vtc2* line. Further studies are necessary to confirm these hypotheses. In general, the results obtained indicate the vital co-functioning of the AP and the L-galactose pathway of Asc biosynthesis and may be useful for the development of genetically engineered techniques for the control of vitamin C synthesis in plants.

Key words: *Arabidopsis thaliana*; alternative oxidase; ascorbate; GDP-L-galactose phosphorylase; *AOX1a* and *VTC2* knockout lines; crossing; genotyping; double mutants; lethal mutation

For citation: Garmash E.V., Belykh E.S., Yadrinhinskiy K.V., Malyshev R.V., Velegzhaninov I.O. Generation and characterization of a double-knockout *Arabidopsis thaliana* line lacking expression of *AOX1a* and *VTC2*. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed.* 2026;30(2):241-249. doi 10.18699/vjgb-26-27

Funding. The work was supported by the Ministry of Science and Higher Education of the Russian Federation, State Assignment "Physiological and molecular mechanisms of integration of cellular processes and the integrity of the plant organism: photosynthesis and respiration" (No. 125020301262-2).

Acknowledgements. The study was performed using the equipment of the Collective Use Center (CUC) "Molecular Biology", Institute of Biology, Komi Science Centre, Ural Branch, RAS. Authors also thank PhD M.A. Shelyakin, scientific researcher of the Institute of Biology, Komi Science Centre, Ural Branch, RAS, for assistance in *A. thaliana* lines crossing.

Получение и характеристика двойной нокаутной линии *Arabidopsis thaliana*, лишенной экспрессии *AOX1a* и *VTC2*

Е.В. Гармаш , Е.С. Белых , К.В. Ядрихинский , Р.В. Малышев , И.О. Велегжанинов 

Институт биологии Коми научного центра Уральского отделения Российской академии наук, Сыктывкар, Россия

 garmash@ib.komisc.ru

Аннотация. У высших растений L-галактозный путь является основным путем биосинтеза витамина С (аскорбата, Аск), последний этап которого связан с функционированием электрон-транспортной цепи митохондрий. Помимо основного цитохромного пути, электрон-транспортная цепь растений содержит альтернативный путь через альтернативную терминальную оксидазу (АОХ). Вовлечение АОХ способствует синтезу Аск, поэтому предполагается, что подавление альтернативного пути в условиях дефицита Аск может привести к снижению жизнеспособности растений. В связи с этим была поставлена цель – рассмотреть последствия нокаутирования в растениях *Arabidopsis thaliana* одновременно двух генов: *AOX1a*, наиболее стресс-индуцибельного гена АОХ, и *VTC2*, гена ключевого фермента L-галактозного пути синтеза Аск. Для этого проведено скрещивание двух линий

A. thaliana с Т-ДНК инсерцией в целевых генах и получены гибридные линии. Изучены характеристики семян первого (F_1) и второго (F_2) поколений. Семена F_1 отличались более крупными размерами по сравнению с родительскими линиями, что могло быть следствием гетерозиса. В поколении F_2 в результате самоопыления растений F_1 сформировались семена, значительно варьирующие по размерам, в том числе группа мелких семян, имеющих дегенеративные морфологические отклонения. Большинство мелких семян не проросло или погибло на стадии появления проростка. Генотипирование этих семян с помощью ПЦР установило отсутствие нормальных копий *AOX1a* и *VTC2* в геноме, что указывает на появление летальной мутации при одновременном нокауте обоих генов. Обсуждаются причины летального исхода двойного нокаута. Одной из генетических причин, по-видимому, стало взаимодействие мутаций (неаллельных генов). На физиологическом уровне, возможно, возникли необратимые нарушения дыхания, в том числе из-за влияния в линии *vtc2* криптической мутации. Для подтверждения данных гипотез требуются дальнейшие исследования. В целом полученные результаты свидетельствуют о жизненно важном совместном функционировании альтернативного и L-галактозного путей биосинтеза Аск и могут быть полезны для развития генно-инженерных приемов управления синтезом витамина С в растениях.

Ключевые слова: *Arabidopsis thaliana*; альтернативная оксидаза; аскорбат; ГДФ-Л-галактозофосфорилаза; линии с нокаутом *AOX1a* и *VTC2*; скрещивание; генотипирование; двойная нокаутная линия; летальная мутация

Introduction

Ascorbate (vitamin C) is an important multifunctional antioxidant that participates in stress resistance and redox signaling (Smirnov, 2018; Foyer et al., 2020; Matos et al., 2022, and others). Plants replenish their ascorbate pool through several mechanisms, including biosynthetic processes and regeneration via the ascorbate-glutathione cycle. Among these pathways, the L-galactose pathway is the dominant one (Dowdle et al., 2007; Wheeler et al., 2015). In the L-galactose pathway, GDP-D-mannose is converted into GDP-L-galactose, L-galactose-1-phosphate, L-galactose, L-galactono-1,4-lactone, and finally into L-ascorbate (Asc) (Fig. 1). A key step in this pathway is the reaction catalyzed by GDP-L-galactose phosphorylase (GGP) (Yoshimura et al., 2014; Matos et al., 2022). The GGP enzyme is encoded by the paralogous genes *VTC2* and *VTC5* (VITAMIN C), which demonstrate light-dependent expression patterns (Yoshimura et al., 2014; Smirnov, 2018). Among these, *VTC2* plays the primary role in Asc biosynthesis. Knockout of *VTC2* (unlike its paralog *VTC5*) results in a significant decrease (up to 80 % or more) in Asc levels and plant growth (Dowdle et al., 2007; Lim et al., 2016).

The final stage of the L-galactose pathway is associated with the mitochondrial electron transport chain (ETC), where L-galactono-1,4-lactone dehydrogenase (GLDH) oxidizes L-galactono-1,4-lactone to Asc by transferring electrons between complexes III and IV via the labile electron carrier cytochrome *c* (Fig. 1). Unlike mammalian ETC, plants, in addition to the main energy-conserving cytochrome pathway (CP), possess an alternative pathway for electron transport (AP) through an alternative terminal cyanide-resistant oxidase (AOX) (Garmash, 2022) (Fig. 1). The AP bypasses two sites of proton translocation, complexes III and IV, and therefore represents a non-energy-conserving branch of electron transport (Vanlerberghe et al., 2020). The main function of the AP is to maintain redox balance during mitochondrial electron transport and to reduce reactive oxygen species (ROS) production. This impacts the metabolism of not only mitochondria but also cells and plants as a whole (Del-Saz et al., 2017; Vanlerberghe et al., 2020; Garmash, 2022). Moreover, the AP promotes Asc

synthesis by accepting electrons from ubiquinone and maintaining the cytochrome *c* pool in a more oxidized state (Bartoli et al., 2006; Matos et al., 2022).

Studies of mutant lines with *VTC2* suppression reveal a number of defects and metabolic disturbances, especially under stressful conditions (Lim et al., 2016; Matos et al., 2022). Ascorbate deficiency and the stunted phenotype observed in these mutant plants likely result from long-term impact of the mutation on respiratory activity, mitochondrial ETC function, and, in particular, on electron transfer via AOX (Matos et al., 2022).

Nuclear genome of *A. thaliana* includes five genes of AOX, with the highest expression in response to various types of stress attributed to the gene *AOX1a* (Del-Saz et al., 2017; Garmash, 2022). *A. thaliana* plants overexpressing *AOX1a* demonstrated increased Asc synthesis under stress conditions (high light) (Garmash et al., 2022; Sweetman et al., 2022), whereas mutant lines with *AOX1a* suppression showed decreased Asc levels (Vishwakarma et al., 2015; Garmash et al., 2022). Additionally, *A. thaliana* plants of the *vtc2* line exhibited AOX activation compared with wild-type plants, and treatment of mutant plants with a specific inhibitor of the AP resulted in a significant (up to 40 %) decrease in ATP concentration (Garmash et al., 2024). Other authors (Talla et al., 2011) studied responses of *A. thaliana* plants with knockout of the GDP-mannose phosphorylase gene catalyzing GDP-D-mannose synthesis (*vtc1* line) on mitochondrial and chloroplast ETC inhibition, demonstrating mutual complementation between Asc and AOX, preventing excess accumulation of ROS and protecting photosynthesis from photooxidation. These data confirm the participation of AOX in maintaining Asc synthesis and optimizing energy balance and indicate that suppression of the AP under conditions of Asc deficiency can result in decreased plant viability. Therefore, we hypothesized that *AOX1a* and *VTC2* play a key role in providing energy and vital functions of the plant organism.

The aim of this paper was to study the consequences of simultaneous knockout of two genes, *AOX1a* and *VTC2*, in *A. thaliana* plants. To this end, mutant lines were crossed, and the characteristics of seeds from the first (F_1) and second (F_2)

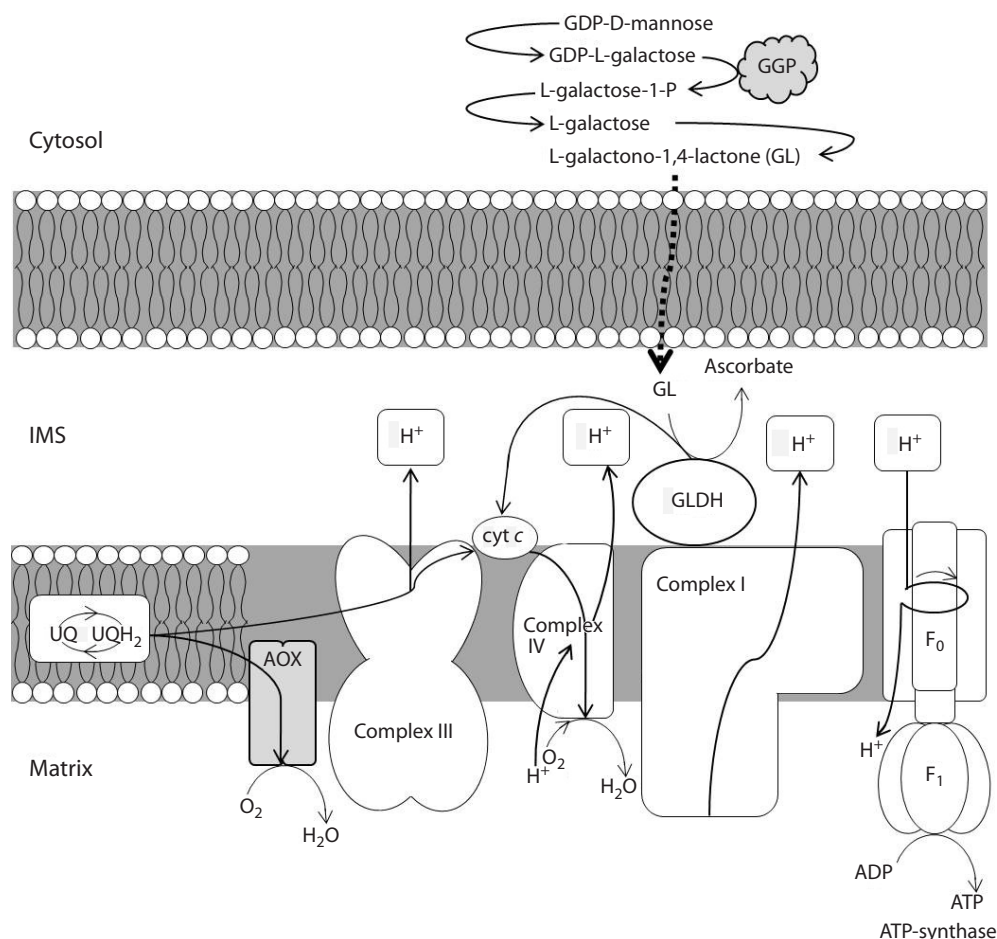


Fig. 1. Schematic representation of the relationship between ascorbate synthesis and the electron transport chain (ETC) of plant mitochondria.

The ETC components in the inner mitochondrial membrane are shown: from the ubiquinone pool (UQ/UQH₂) electrons are transferred via the alternative pathway (AP) catalyzed by alternative oxidase (AOX) and via the cytochrome pathway through complex III, cytochrome *c*, and complex IV. GGP denotes GDP-L-galactose phosphorylase. L-galactono-1,4-lactone (GL), synthesized through the L-galactose pathway and transported from the cytosol into the mitochondrial intermembrane space (IMS), is oxidized to ascorbate by L-galactono-1,4-lactone dehydrogenase (GLDH), which transfers electrons to cytochrome *c*. Electron transport along the ETC is coupled with proton pumping into the intermembrane space and the "return" of protons into the matrix via ATP synthase, driving ATP production.

generations were analyzed. Double knockouts were confirmed by T-DNA insertion localization using primers specific to the target genes.

Material and methods

Mutant lines with T-DNA insertion into *AOX1a* – SALK_084897 and into *VTC2* – SALK_146824C were used for hybridization (Fig. 2).

Seeds of mutant lines were obtained from the Arabidopsis Biological Resources Centre (ABRC, Ohio University, USA). The line with knockout of *AOX1a* (*aox1a*) contains a T-DNA insertion in one copy of the *AOX1a* gene, making the line *aox1a* heterozygous. This line also has an insertion in the *EFS* gene, which encodes the flowering inhibitor histone-lysine N-methyltransferase. The knockout line for *VTC2* (*vtc2*) has insertions in both copies of the gene and is homozygous. The wild-type line Columbia-0 (Col-0) was used to control

for relative transcript content. Seeds of Col-0 were obtained from the Nottingham Arabidopsis Stock Centre (NASC, UK) and kindly provided by O.I. Grabelnykh and V.I. Tarasenko (SIPB SB RAS).

Seeds were planted, stratified, and germinated in 200 cm³ pots containing a mixture of perlite, vermiculite, and soil in a 1 : 1 : 2 ratio. Seedlings were grown under controlled conditions at 22 °C with a light intensity of 90 μmol quanta/(m² · s) and a 16-hour photoperiod until the budding phase (stage 5.1; Boyes et al., 2000). Lighting was provided by luminescent lamps (TL-D 30W and TL-D 30WAquarelle, Philips, Netherlands).

Hybridization was performed by transferring pollen from the line SALK_146824C (*vtc2*) onto the line SALK_084897 (*aox1a*) under microscopic control. For this purpose, unblown buds (one day before blooming) on the maternal plant were carefully opened and dissected to remove all stamens with

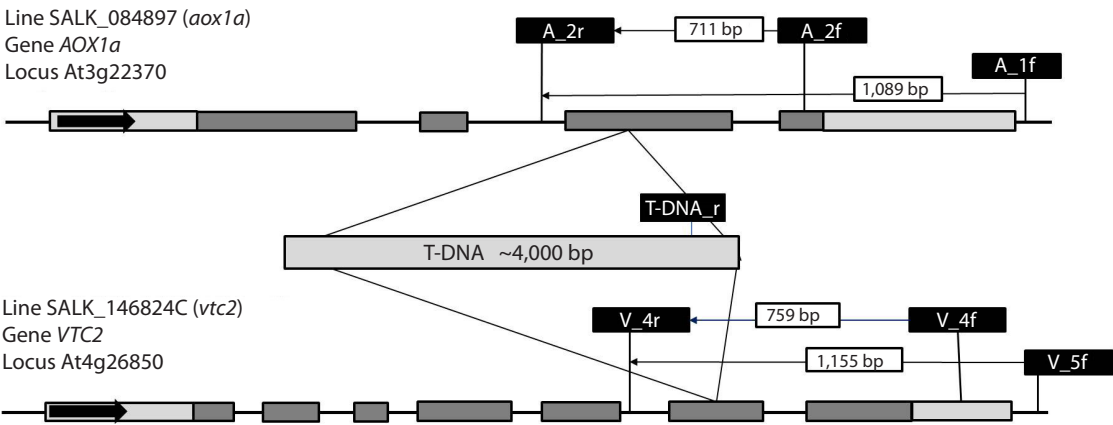


Fig. 2. Scheme of primer locations, PCR product lengths, and T-DNA insertion sites in the *AOX1a* gene from line SALK_084897 (*aox1a*) and *VTC2* from line SALK_146824C (*vtc2*).
Primer pairs V_4f-V_4r and V_5f-V_4r, as well as A_2f-A_2r and A_1f-A_2r, allow identification of the wild type *VTC2* or *AOX1a* genes, respectively. Primer pairs V_4f-T-DNA_r and V_5f-T-DNA_r, as well as A_2f-T-DNA_r and A_1f-T-DNA_r, allow identification of the corresponding mutant genes.

Table 1. Primers for genotyping and checking for T-DNA insertions in the *VTC2* gene (V) of line SALK_146824C (*vtc2*), and the *AOX1a* gene (A) of line SALK_084897 (*aox1a*)

Name	Sequence
T-DNA_r	CTTCCTTCTCGCCACGTTC
V_4f	ACTGAAGGACAAGGCACTCG
V_4r	ACCTTTTCACCCTTGCTCATCT
V_5f	TCATAACGAGAGCTTGTTGGTGG
A_1f	GATCCACATCCATTGGGTCCT
A_2r	CACTTCGTTAGTGCTTTGGATTCTTGG
A_2f	GCTTCCTTTAGTTCACGACCTTGG

pollen with tweezers. Pollen from the paternal plant was transferred onto the stigma of the maternal plant's pistil. The fertilized flower was marked, and fertilization was repeated for three consecutive days. All surrounding flowers were regularly removed during the budding phase. The plant was isolated with a transparent plastic cylinder until seed maturation.

Seed size from the F_1 and F_2 generations was measured ($n = 111$ and 496 for F_1 and F_2 , respectively). The length (major semi-axis of the ellipse, a , mm) and width (minor semi-axis of the ellipse, b , mm) of the seeds were measured under a microscope. The ellipse area (S , mm²) was calculated as $S = a \cdot b \cdot \pi$. F_2 seeds were divided according to size into three groups: large, medium, and small ($n = 49$, 365 , and 82 for each group, respectively).

To check for embryo presence, some seeds from each group, swollen after soaking in water, were dissected under a microscope.

To analyze the relative levels of gene transcripts, genotyping and qPCR were performed on plant samples at growth stage 1.14 (Boyes et al., 2000) and on seeds.

To genotype the parental lines (*aox1a*, *vtc2*) and crossbred lines, DNA from the corresponding samples was used along with primers (Fig. 2, Table 1). DNA was extracted from 50–100 mg of air-dried leaves or dry seeds using the Sorb-GMO-B kit (Sintol, Russia) according to the manufacturer's instructions. The PCR reaction mixture contained 4 μ L of Screen Mix (Evrogen, Russia), 4 μ L of each corresponding forward and reverse primer (0.3 μ M) (Sintol, Russia), 7 μ L of ddH₂O (PanEco, Russia), and 1 μ L of DNA. Amplification consisted of a preliminary denaturation for 5 minutes at 95 °C, followed by 35 cycles of denaturation (15 s) at 95 °C, primer annealing (30 s) at 55 °C, and elongation (60 s) at 72 °C, with a final elongation for 2 minutes at 72 °C. PCR products were analyzed by electrophoresis on a 1 % agarose gel in Tris-Acetate buffer. DNA fragments were stained with Safe Green Stain (APGENA, Russia) and visualized with blue light using the SuperRay Maxi system (APGENA, Russia). Fragment lengths were determined using a 100+ bp DNA ladder (100–1,500 bp) (Evrogen, Russia).

Relative levels of gene transcripts were analyzed by qPCR using the CFX96 system (Bio-Rad, USA). Rosette leaves were frozen in liquid nitrogen and stored at –80 °C prior to analysis. RNA was extracted using the HiPure Plant RNA Kit (Magen, China) according to the manufacturer's instructions. RNA concentration was measured using Qubit (Thermo Fisher Scientific, USA) with the QubitTM RNA HS Assay Kit (Thermo Fisher Scientific, USA). For cDNA synthesis, the MMLV RT kit with oligo(dT) primers (Evrogen, Russia) was used. qPCR reactions were performed with “qPCRMix-HS SYBR” (Evrogen, Russia) and the corresponding primers (Table 2). Target gene expression was normalized to the geometric mean of the *AT2G28390* and *AT4G34270* reference genes (Czechowski et al., 2005). Relative transcript levels were calculated using the $2^{-\Delta\Delta C_T}$ method (Livak, Schmittgen, 2001). Analyses were performed on three biological replicates, including leaves from one crossbred plant or 3–5 individual plants of the original lines, with two technical replicates for each sample.

Table 2. The list of primers for qPCR in *A. thaliana*

Gene	Locus	Name (by NCBI)	Forward primer (5'–3') Reverse primer (5'–3')	PCR product length, bp
<i>AOX1a</i>	At3g22370	Alternative oxidase 1A	ATGATAACTCGGGTGGAGC TCGTCGGAGCTCTAGTCCAT	166
<i>VTC2</i>	At4g26850	GDP-L-galactose phosphorylase 1	GGACCTCCATGGGTCTGT GAGGTTACTGCTCTCGCCTT	131
<i>AT2G28390*</i>	At2g28390	SAND family protein	AACTCTATGCAGCATTTGATCCACT TGATTGCATATCTTTATCGCCATC	61
<i>AT4G34270*</i>	At4g34270	TIP41-like family protein	GTGAAAACGTGTGGAGAGAAGCAA TCAACTGGATACCCCTTTCGCA	61

Note. Primers were designed using Primer-BLAST (Ye et al., 2012). * – reference genes (from Czechowski et al., 2005).

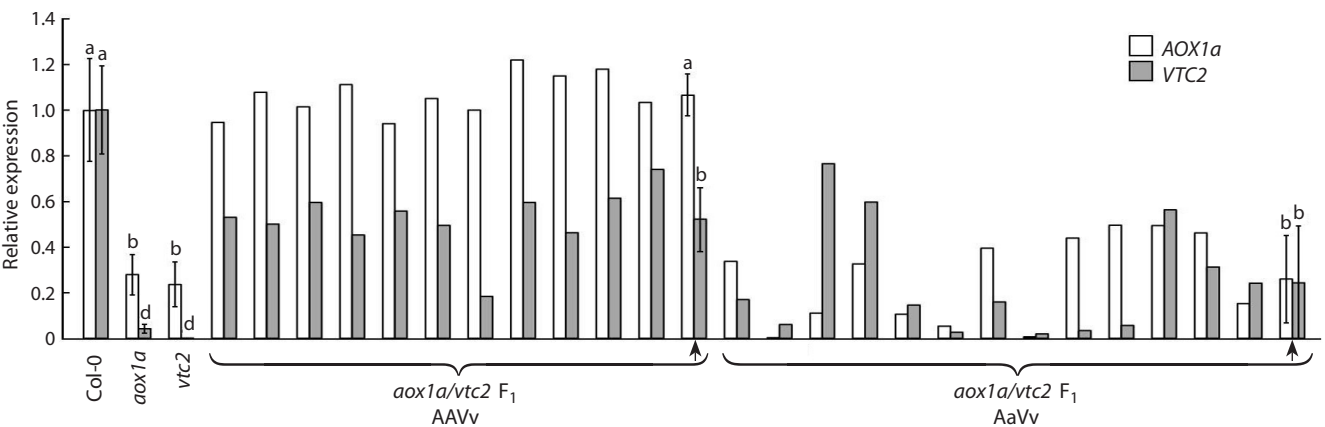


Fig. 3. Relative levels of *AOX1a* and *VTC2* transcripts in leaves of *A. thaliana* plants from the wild-type line (Col-0), parental lines (*aox1a* and *vtc2*), and 24F₁ hybrids (*aox1a/vtc2*). Arrows indicate mean values and their standard errors for the presumed genotypes (AAVv and AaVv). Statistical significance of differences in gene expression levels is indicated by different letters (ANOVA, Duncan's test, $p \leq 0.05$).

Results

Genes *AOX1a* (At3g22370), *VTC2* (At4g26850), and *EFS* (At1g77300) are located on chromosomes 3, 4, and 1 of *A. thaliana*, respectively. The offspring of heterozygotes obtained by crossing lines SALK_084897 and SALK_146824C, according to unlinked inheritance and Mendel's third law, allow obtaining all possible combinations of variants in the F₁ generation, and homozygous offspring for both target mutations in the F₂ generation.

Relative transcript levels of *AOX1a* obtained by qPCR in the *aox1a* line were significantly lower than in the Col-0 line, and the level of *VTC2* mRNA in the *vtc2* line was nearly undetectable (Fig. 3). Seed size of the *aox1a* line also varied notably compared to the *vtc2* line (Fig. 4). These data indirectly confirm the heterozygosity of *aox1a* and the homozygosity of *vtc2* mutant alleles, which was further verified by the presence of T-DNA insertions in the corresponding lines (Fig. 5). Bands corresponding to PCR products with primer pairs V_4f-V_4r and V_5f-V_4r, specific to wild-type genes, were absent, while bands with primer pairs V_4f-T-DNA_r and V_5f-T-DNA_r, confirming T-DNA insertion, were present in

the electrophoregram of the *vtc2* line. The T-DNA insertion was located precisely at position 13,499,579 in the sixth exon of the gene At4g26850 in the *vtc2* line (Belykh et al., 2024). The presence of a band with primer pair A_2f-A_2r and the absence of one with A_2f-T-DNA_r indicated the wild-type status of the *AOX1a* gene.

The *aox1a* line, in contrast, is characterized by a band corresponding to the PCR product of the V_4f-V_4r primer pair and the absence of a band corresponding to V_4f-T-DNA_r, confirming the presence of a wild-type copy of the *VTC2* gene (Fig. 5). Bands corresponding to both primer pairs for the *AOX1a* gene (A_2f-A_2r and A_2f-T-DNA_r; A_1f-A_2r and A_1f-T-DNA_r) indicate the coexistence of a wild-type copy together with a T-DNA insertion-disrupted copy, characteristic of a heterozygous line.

To predict segregation in the F₁ generation, a scheme based on a Punnett square was drawn up to determine possible genotype variants of the offspring based on parental genotypes and their gametes, calculated by dividing the number of cells of a given genotype by the total number of cells. According to the scheme, segregation by genotype was 1:1:1:1 (four

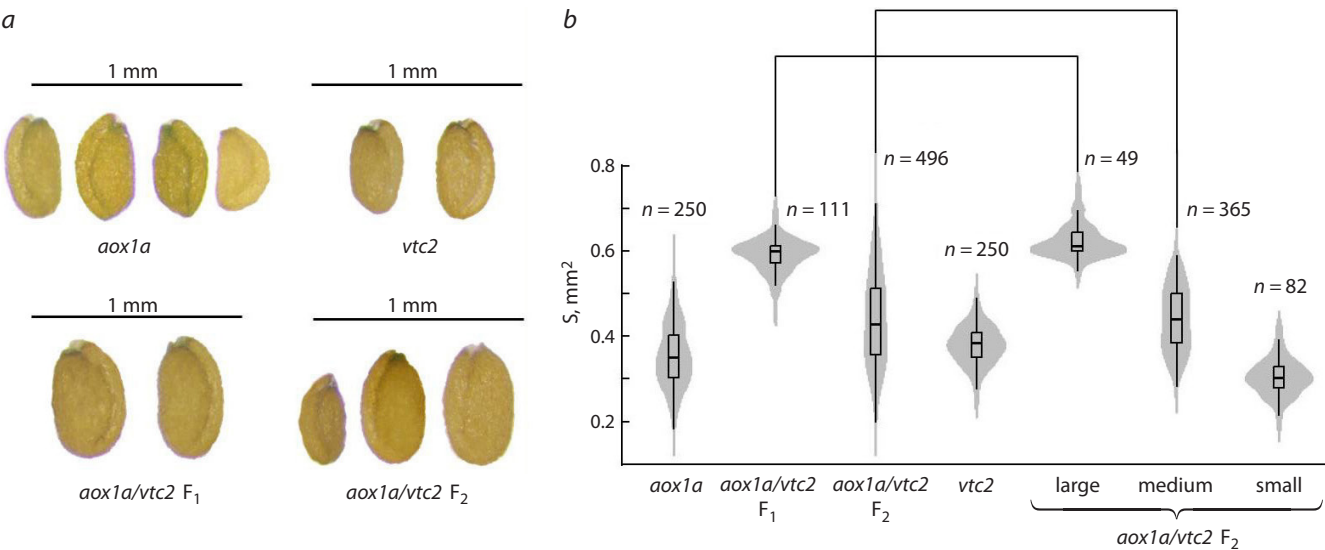


Fig. 4. Appearance (a) and seed size (S) (b) of parental and crossbred (*aox1a/vtc2*) *A. thaliana* lines from the F₁ and F₂ generations. Panel (b) shows seed size distribution presented as violin plots combined with boxplots, including arithmetic mean values and standard errors. *n* indicates sample size. Differences between paired parameters not marked with a square bracket are statistically significant at *p* < 0.01 (Dunn's test with the Bonferroni correction).

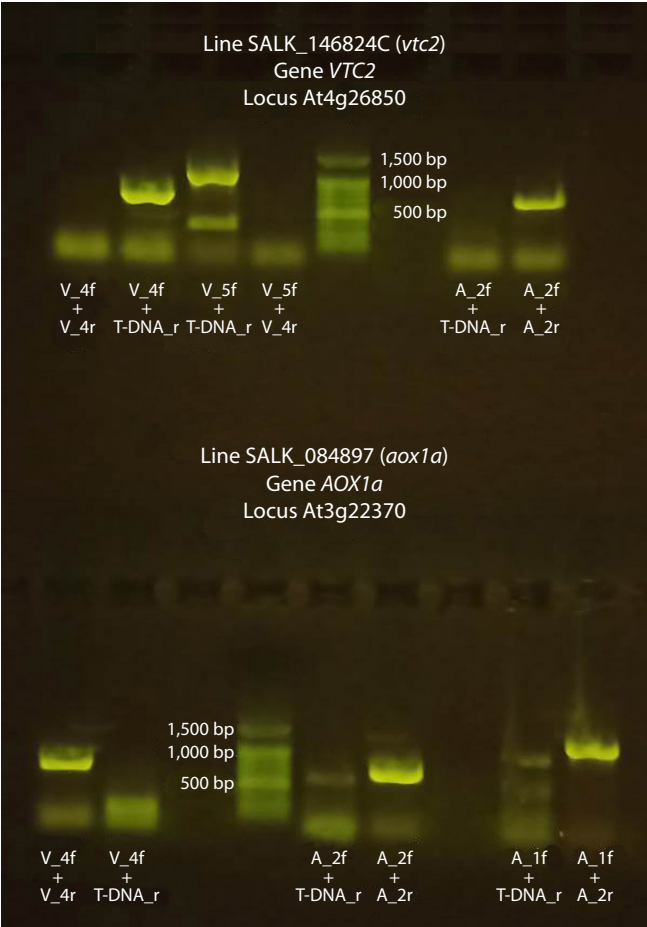


Fig. 5. Electrophoregram of DNA amplicons from *A. thaliana* plants of the *vtc2* and *aox1a* lines, demonstrating the presence of T-DNA insertions in the *VTC2* or *AOX1a* genes correspondently. Primer descriptions are provided in Table 1.

phenotypes in equal proportions) (Table 3). In other words, 25 % of F₁ seeds belonged to the mutant line AaVvEE, deficient in *VTC2* and *AOX1a*; 25 % had genotype AaVvEe, deficient in all genes (*AOX1a*, *VTC2*, *EFS*); the remaining seeds were represented by genotypes deficient in *VTC2* or in *VTC2* and *EFS*.

F₁ seeds obtained from the cross were significantly larger than those from the parental lines (Fig. 4). This likely results from heterosis, manifested as increased hybrid viability due to the inheritance of diverse alleles from genetically distinct parents.

Plants grown from 24 F₁ seeds exhibited decreased *VTC2* transcript levels (Fig. 3). More than half (54 %) also showed decreased *AOX1a* transcript levels. The genome of the remaining 46 % of plants presumably contained both copies of *AOX1a*, as in wild-type plants. Thus, the hybridization corresponded well with the predicted segregation.

Following self-pollination, the F₁ plants produced fruits containing F₂ seeds. These seeds varied considerably in size (Fig. 4) and were classified into three significantly different groups. Notably, the largest seed group did not differ in size from the F₁ seeds, suggesting genotypes of either AAVv or AaVv, as in the F₁ generation. The smallest seeds exhibited degenerative morphological abnormalities; dissection revealed undeveloped embryos (Fig. 6). Most small seeds failed to germinate or died at the seedling stage.

Genotyping of large seeds from the F₂ generation revealed the presence of wild-type copies of the *AOX1a* and *VTC2* genes in their genomes (Fig. 7). In contrast, bands corresponding to wild-type genes were absent in the electrophoregram of DNA extracted from small seeds. This finding confirms that the genotype of small seeds corresponds to the double knockout mutant *aox1a/vtc2*.

Table 3. Punnett square for the F₁ generation from crossing lines SALK_146824C (*vtc2*) and SALK_084897 (*aox1a*)

Line genotype	SALK_084897 (<i>aox1a</i>) AaVVEe			
SALK_146824C (<i>vtc2</i>) AAvvEE	AVE	aVE	AVe	aVe
	AvE	AAVvEE	AaVvEE	AAVvEe
	AvE	AAVvEE	AaVvEE	AAVvEe
	AvE	AAVvEE	AaVvEE	AAVvEe
	AvE	AAVvEE	AaVvEE	AAVvEe
		Generation deficient in <i>VTC2</i>	Generation deficient in <i>VTC2</i> and <i>AOX1a</i>	Generation deficient in <i>VTC2</i> and <i>EFS</i>
			Generation deficient in <i>VTC2</i> and <i>EFS</i>	Generation deficient in <i>VTC2</i> , <i>AOX1a</i> , <i>EFS</i>

Note. A – wild-type (as in Col-0) *AOX1a* gene; a – insertion in *AOX1a*; V – wild-type (as in Col-0) *VTC2* gene; v – insertion in *VTC2*; E – wild-type (as in Col-0) *EFS* gene; e – insertion in *EFS*.

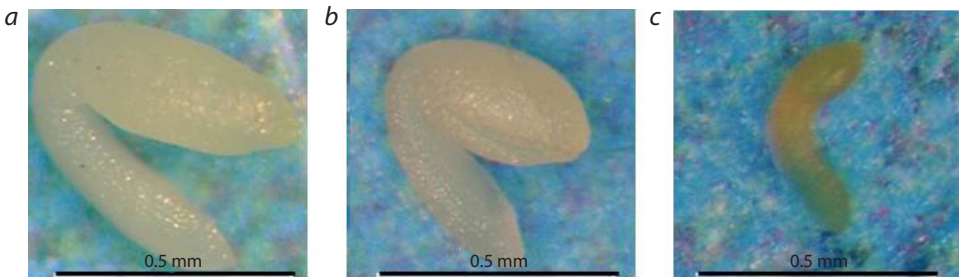


Fig. 6. Embryos dissected from large (a), medium (b) and small (c) seeds of *A. thaliana* crossbreed lines *aox1a/vtc2* of the F₂ generation.

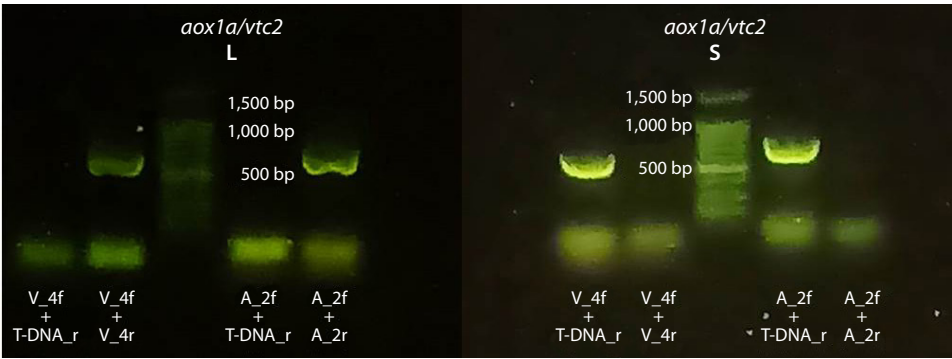


Fig. 7. Electrophoregram of DNA amplicons from large (L) and small (S) seeds of the F₂ generation from the *A. thaliana* crossbreed *aox1a/vtc2* lines.

The presence of bands corresponding to primer pairs V_4f-V_4r and A_2f-A_2r, characteristic of the wild-type variants of the *VTC2* and *AOX1a* genes, respectively, is evident in large seeds and absent in small seeds. Primer descriptions are provided in Table 1.

Discussion

Studies of vitamin C-deficient lines, primarily *vtc2*, have demonstrated the important role of the L-galactose pathway of vitamin C biosynthesis, and in particularly that of the key enzyme GGP, in replenishing the main pool of this metabolite in plants. However, it is believed that the reduced growth and viability of mutants are associated more with metabolic disturbances and an imbalance in the respiratory ETC than with decreased Asc levels alone (Lim et al., 2016; Matos et al., 2022).

It has been shown that these abnormalities may be linked to a cryptic mutation in the mutant line (Lim et al., 2016). However, the same authors demonstrated that the growth defects of *vtc2* mutants segregate independently of the mutation in the *VTC2* gene, indicating that the potential cryptic mutation is located on a different chromosome. This allows the analysis of the relationship between the phenotype of the offspring and the status of the *VTC2* gene.

The causes of metabolic and growth disturbances of the *vtc2* line have been identified in several studies using mutant

combinations with this line. The growth-reduced phenotype of *vtc2* has been described in combination with mutations causing chronic photooxidative stress (Müller-Moulé et al., 2004), knockout of chloroplast ascorbate peroxidase (Giacomelli et al., 2007), and suppression of the autoimmune response (Zhu et al., 2013). A double *vtc2-1/abi4* mutant, generated by crossing the *A. thaliana vtc2-1* line (a loss-of-function mutant with significantly decreased *VTC2* transcription) with an abscisic acid-insensitive mutant line (*abi4*), exhibited growth and morphology similar to those of wild-type plants (Kerchev et al., 2011). This suggests that abscisic acid contributes to the stunted phenotype of *vtc2-1*. However, all described double knockout mutants retained the ability to grow. The only lethal phenotype was observed in the double knockout of both GGP genes (*VTC2* and *VTC5*) (Dowdle et al., 2007).

Previously, vitamin C-deficient mutants, *vtc1* (Conklin et al., 1996) and *vtc2* (Jander et al., 2002), were initially selected for their sensitivity to ozone. Ozone sensitivity likely results from imbalances in respiratory pathways and increased ROS generation due to low Asc synthesis. Inhibition of the cytochrome pathway under elevated ozone levels occurred simultaneously with a sharp increase in AOX activity (Ederli et al., 2006; Pasqualini et al., 2007). These findings support the hypothesis that enhanced AOX activity in the *vtc2* line protects against oxidative stress and possible plant death.

This hypothesis was confirmed in our study, where the F₂ generation was nonviable when both *AOX1a* and *VTC2* were knocked out simultaneously. It is possible that the lethal mutation resulted from the interaction of mutations in non-allelic genes, including a cryptic mutation. However, the role of the interaction between the *AOX1a* and *VTC2* mutations – rather than between *AOX1a* and the cryptic mutation – is supported by the observation that all F₂ seeds that failed to germinate carried a mutation in *VTC2*, which appears to segregate independently from the cryptic mutation (Lim et al., 2016). To clarify this issue, more detailed molecular genetic studies are needed, including genomic DNA sequencing and RNA analysis to identify mutations affecting pre-mRNA splicing.

Conclusion

As a result of crossbreeding of the SALK_084897 and SALK_146824C lines, seeds of varying sizes were obtained in the F₂ generation. The smallest seed group was characterized as aberrant, displaying degenerative morphological deviations and an inability to germinate. Genotyping of these seeds revealed the absence of both *AOX1a* and *VTC2*, indicating a lethal mutation arising from the simultaneous double knockout of these genes. This finding highlights the essential joint functioning of two systems: the alternative pathway through AOX and the L-galactose pathway of Asc biosynthesis linked to the mitochondrial ETC. Further investigation into the interrelationship and coordination between the AP and the L-galactose Asc synthesis pathway could focus on producing double mutants with defects in genes involved in ascorbate synthesis, other respiratory electron transport components, and participants in hormonal and redox signaling. Such research may aid in the development of genetic engineering strategies to enhance ascorbate synthesis, thereby improving stress tolerance in significant food-value crops.

References

- Bartoli C.G., Yu J., Gómez F., Fernández L., McIntosh L., Foyer C.H. Inter-relationships between light and respiration in the control of ascorbic acid synthesis and accumulation in *Arabidopsis thaliana* leaves. *J Exp Bot.* 2006;57(8):1621-1631. doi 10.1093/jxb/erl005
- Belykh E., Yadrikhinskiy K., Golubev M., Belykh N., Velegzhaninov I., Garmash E. Localization data of the T-DNA insertion site in *Arabidopsis* line SALK_146824C. *Data Brief.* 2024;54:110309. doi 10.1016/j.dib.2024.110309
- Boyce D.C., Zayed A.M., Ascenzi R., McCaskill A.J., Hoffman N.E., Davis K.R., Görlach J. Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell.* 2001;13(7):1499-1510. doi 10.2307/3871382
- Conklin P.L., Williams E.H., Last R.L. Environmental stress sensitivity of an ascorbic acid-deficient *Arabidopsis* mutant. *Proc Natl Acad Sci USA.* 1996;93(18):9970-9974. doi 10.1073/pnas.93.18.9970
- Czechowski T., Stitt M., Altmann T., Udvardi M.K., Scheible W.-R. Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol.* 2005;139(1):5-17. doi 10.1104/pp.105.063743
- Del-Saz N.F., Ribas-Carbo M., McDonald A.E., Lambers H., Fernie A.R., Florez-Sarasa I. An *in vivo* perspective of the role(s) of the alternative oxidase pathway. *Trends Plant Sci.* 2018;23(3):206-219. doi 10.1016/j.tplants.2017.11.006
- Dowdle J., Ishikawa T., Gatzek S., Rolinski S., Smirnov N. Two genes in *Arabidopsis thaliana* encoding GDP-L-galactose phosphorylase are required for ascorbate biosynthesis and seedling viability. *Plant J.* 2007;52(4):673-689. doi 10.1111/j.1365-313X.2007.03266.x
- Ederli L., Morettini R., Borgogni A., Wasternack C., Miersch O., Reale L., Ferranti F., Tosti N., Pasqualini S. Interaction between nitric oxide and ethylene in the induction of alternative oxidase in ozone-treated tobacco plants. *Plant Physiol.* 2006;142(2):595-608. doi 10.1104/pp.106.085472
- Foyer C.H., Kyndt T., Hancock R.D. Vitamin C in plants: novel concepts, new perspectives, and outstanding issues. *Antioxid Redox Signal.* 2020;32(7):463-485. doi 10.1089/ars.2019.7819
- Garmash E.V. Signaling pathways of alternative oxidase genes expression in plants. *Russ J Plant Physiol.* 2022;69(1):1. doi 10.1134/S1021443722010058
- Garmash E.V. Suppression of mitochondrial alternative oxidase can result in upregulation of the ROS scavenging network: some possible mechanisms underlying the compensation effect. *Plant Biol.* 2023;25(1):43-53. doi 10.1111/plb.13477
- Garmash E.V., Dymova O.V., Silina E.V., Malyshev R.V., Belykh E.S., Shelyakin M.A., Velegzhaninov I.O. *AOX1a* expression in *Arabidopsis thaliana* affects the state of chloroplast photoprotective systems under moderately high light conditions. *Plants (Basel).* 2022;11(22):3030. doi 10.3390/plants11223030
- Garmash E.V., Silina E.V., Belykh E.S., Shelyakin M.A., Malyshev R.V. Alternative oxidase affects ascorbate metabolism in *Arabidopsis thaliana* plants under high-light conditions: possible links between respiratory electron transport pathways in mitochondria. *J Biosci.* 2024;49:94. doi 10.1007/s12038-024-00471-6
- Giacomelli L., Masi A., Ripoll D.R., Lee M.J., van Wijk K.J. *Arabidopsis thaliana* deficient in two chloroplast ascorbate peroxidases shows accelerated light-induced necrosis when levels of cellular ascorbate are low. *Plant Mol Biol.* 2007;65:627-644. doi 10.1007/s11103-007-9227-y
- Jander G., Norris S.R., Rounsley S.D., Bush D.F., Levin I.M., Last R.L. *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiol.* 2002;129(2):440-450. doi 10.1104/pp.003533
- Kerchev P.I., Pellny T.K., Vivancos P.D., Kiddle G., Hedden P., Driscoll S., Vanacker H., Verrier P., Hancock R.D., Foyer C.H. The transcription factor ABI4 is required for the ascorbic acid-dependent regulation of growth and regulation of jasmonate-dependent defense signaling pathways in *Arabidopsis*. *Plant Cell.* 2011;23(9):3319-3334. doi 10.1105/tpc.111.090100

- Lim B., Smirnoff N., Cobbett C.S., Golz J.F. Ascorbate-deficient *vtc2* mutants in *Arabidopsis* do not exhibit decreased growth. *Front Plant Sci.* 2016;7:1025. doi 10.3389/fpls.2016.01025
- Livak K.J., Schmittgen T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods.* 2001;25(4):402-408. doi 10.1006/meth.2001.1262
- Matos I.F., Morales L.M.M., Santana D.B., Silva G.M.C., Gomes M.M.A., Ayub R.A., Costa J.H., de Oliveira J.G. Ascorbate synthesis as an alternative electron source for mitochondrial respiration: possible implications for the plant performance. *Front Plant Sci.* 2022;13: 987077. doi 10.3389/fpls.2022.987077
- Müller-Moulé P., Golan T., Niyogi K.K. Ascorbate-deficient mutants of *Arabidopsis* grow in high light despite chronic photooxidative stress. *Plant Physiol.* 2004;134(3):1163-1172. doi 10.1104/pp.103.032375
- Pasqualini S., Paolucci F., Borgogni A., Morettini R., Ederli L. The overexpression of an alternative oxidase gene triggers ozone sensitivity in tobacco plants. *Plant Cell Environ.* 2007;30(12):1545-1556. doi 10.1111/j.1365-3040.2007.01730.x
- Smirnoff N. Ascorbic acid metabolism and functions: a comparison of plants and mammals. *Free Radic Biol Med.* 2018;122:116-129. doi 10.1016/j.freeradbiomed.2018.03.033
- Sweetman C., Waterman C.D., Wong D.C.J., Day D.A., Jenkins C.L.D., Soole K.L. Altering the balance between AOX1A and NDB2 expression affects a common set of transcripts in *Arabidopsis*. *Front Plant Sci.* 2022;13:876843. doi 10.3389/fpls.2022.876843
- Talla S., Riazunnisa K., Padmavathi L., Sunil B., Rajsheel P., Raghavendra A.S. Ascorbic acid is a key participant during the interactions between chloroplasts and mitochondria to optimize photosynthesis and protect against photoinhibition. *J Biosci.* 2011;36:163-173. doi 10.1007/s12038-011-9000-x
- Van Aken O. Mitochondrial redox systems as central hubs in plant metabolism and signaling. *Plant Physiol.* 2021;186(1):36-52. doi 10.1093/plphys/kiab101
- Vanlerberghe G.C., Dahal K., Alber N.A., Chadee A. Photosynthesis, respiration and growth: a carbon and energy balancing act for alternative oxidase. *Mitochondrion.* 2020;52:197-211. doi 10.1016/j.mito.2020.04.001
- Vishwakarma A., Tetali S.D., Selinski J., Scheibe R., Padmasree K. Importance of the alternative oxidase (AOX) pathway in regulating cellular redox and ROS homeostasis to optimize photosynthesis during restriction of the cytochrome oxidase pathway in *Arabidopsis thaliana*. *Ann Bot.* 2015;116(4):555-569. doi 10.1093/aob/mcv122
- Wheeler G., Ishikawa T., Pornsaksit V., Smirnoff N. Evolution of alternative biosynthetic pathways for vitamin C following plastid acquisition in photosynthetic eukaryotes. *eLife.* 2015;13(4):e06369. doi 10.7554/eLife.06369
- Ye J., Coulouris G., Zaretskaya I., Cutcutache I., Rozen S., Madden T.L. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics.* 2012;13:134. doi 10.1186/1471-2105-13-134
- Yoshimura K., Nakane T., Kume S., Shiomi Y., Maruta T., Ishikawa T., Shigeoka S. Transient expression analysis revealed the importance of *VTC2* expression level in light/dark regulation of ascorbate biosynthesis in *Arabidopsis*. *Biosci Biotechnol Biochem.* 2014;78(1): 60-66. doi 10.1080/09168451.2014.877831
- Zhu Y., Du B., Qian J., Zou B., Hua J. Disease resistance gene-induced growth inhibition is enhanced by *rcd1* independent of defence activation in *Arabidopsis*. *Plant Physiol.* 2013;161(4):2005-2013. doi 10.1104/pp.112.213363

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

Received July 18, 2025. Revised November 13, 2025. Accepted November 14, 2025.

doi 10.18699/vjgb-26-28

Pleiotropic genes underlying genetic correlations across human diseases

I.V. Zorkoltseva , N.M. Belonogova , A.V. Kirichenko , Y.A. Tsepilov , T.I. Axenovich 

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

 zor@bionet.nsc.ru

Abstract. Genetic correlation is a key characteristic of the global genetic similarity of human traits. Its primary underlying mechanism is pleiotropy, which operates at various biological levels. Gene-level pleiotropy is of particular interest, as genes are the fundamental functional units of the genome. Using publicly available results from genome-wide association studies for 324 diseases, we selected a set of 45 diseases in which every pair exhibited a significant genetic correlation. These diseases belonged to 10 nosological categories. The search for genes with pleiotropic effects was carried out using three approaches: (1) gene-based association analysis, (2) selection of single nucleotide polymorphisms (SNP) within gene coding regions significantly associated with at least two diseases, and (3) a cross-trait meta-analysis of SNP association signals followed by the identification of independent loci and gene prioritization within those loci. A comprehensive bioinformatic analysis was performed on all genes identified through these methods. We identified 167 pleiotropic genes implicated in 39 diseases. The most pleiotropic genes in our study were *LPA*, *TCF7L2*, *SLC22A3*, *FES*, *CDKN2B*, and *APOE*, which were associated with 7 to 9 diseases each. Bioinformatic analysis revealed that the pleiotropic genes identified for these 39 diseases are also involved in the genetic architecture of 501 other diseases and traits. This indicates a high degree of pleiotropy, facilitated by the involvement of these genes in diverse biological processes – including homeostasis, cell-cell signaling, regulation of cell proliferation, transport, and catalytic activity – and various molecular functions, such as signaling receptor binding. Thus, we demonstrated that 87% of diseases within a fully connected correlation network share associated genes with at least one other disease. This finding strongly suggests that genetic correlations between human diseases are largely driven by the pleiotropic effects of shared genes.

Key words: genetic correlation; common diseases; pleiotropic genes; gene-based association analysis; cross-trait meta-analysis; functional enrichment analysis

For citation: Zorkoltseva I.V., Belonogova N.M., Kirichenko A.V., Tsepilov Y.A., Axenovich T.I. Pleiotropic genes underlying genetic correlations across human diseases. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov J Genet Breed.* 2026;30(2): 250-258. doi 10.18699/vjgb-26-28

Funding. This work was supported by the budget project No. FWNR 2026-0023.

Генетические корреляции между болезнями человека и плейотропные гены

И.В. Зоркольева , Н.М. Белоногова , А.В. Кириченко , Я.А. Цепилов , Т.И. Аксенович 

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

 zor@bionet.nsc.ru

Аннотация. Характеристикой глобальной генетической общности признаков человека считается их генетическая корреляция. Ее основной механизм – плейотропия, проявляющаяся на различных уровнях. Наиболее интересна плейотропия на уровне генов, поскольку именно они являются фундаментальными функциональными единицами генома. Используя результаты полногеномного анализа ассоциаций 324 болезней, находящиеся в открытом доступе, мы отобрали группу из 45 болезней, в которой каждая пара показала значимую генетическую корреляцию. Эти болезни принадлежали 10 нозологическим категориям. Поиск генов с плейотропными эффектами осуществляли с помощью трех подходов: полногеномного анализа ассоциаций на уровне гена; выбора однонуклеотидных полиморфизмов (SNP) внутри кодирующей части гена, значимо ассоциированных хотя бы с двумя болезнями; и метаанализа сигналов ассоциации SNP со всеми болезнями с последующей идентификацией независимых локусов и приоритизации генов в них. Для всех отобранных таким образом генов мы провели биоинформатический анализ. Всего было идентифицировано 167 плейотропных генов, вовлеченных в контроль 39 болезней. Наиболее плейотропными в нашем исследовании были гены *LPA*, *TCF7L2*, *SLC22A3*, *FES*, *CDKN2B* и *APOE*, каждый из которых контролировал семь-девять болезней. Мы провели биоинформатический анализ всех генов и показали, что найденные нами для 39 болезней плейотропные гены участвуют в контроле еще 501 заболевания. Полученные данные указывают на высокую плейотропную способность этих генов,

которая обеспечивается их участием в различных биологических процессах, таких как поддержание гомеостаза, межклеточная сигнализация, регуляция клеточной пролиферации, транспорт веществ и каталитическая активность, а также выполнением ими различных молекулярных функций, в частности связывания с сигнальными рецепторами. Таким образом, мы показали, что 87 % болезней, представляющих полностью связанную группу, имеют общие гены с хотя бы еще одной болезнью. Это свидетельствует о том, что генетические корреляции между болезнями в значительной степени обусловлены плейотропными эффектами генов.

Ключевые слова: генетическая корреляция; распространенные болезни; плейотропные гены; анализ ассоциаций на уровне гена; метаанализ; биоинформатический анализ

Introduction

Traditionally, diseases have been studied in isolation, as independent entities. However, patients frequently present with multiple chronic conditions. This poses a significant challenge in medicine, as treatment strategies and prognosis are highly dependent on the presence of concomitant pathologies. Certain groups of diseases co-occur in patients more frequently than would be expected by chance, suggesting non-random associations. This phenomenon, known as comorbidity, is largely driven by the genetic similarity between diseases (Rzhetsky et al., 2007; Wang et al., 2017; Jia et al., 2023). Genetic sharing has been shown to account for 46 % of observed comorbidity (Dong et al., 2021).

The advancement of high-throughput genotyping techniques and the increasing availability of genome-wide association study (GWAS) results for a vast number of human diseases have enabled the investigation of their genetic sharing. The development of methods to estimate genetic correlations between traits has allowed for the assessment of global pairwise genetic similarities across human traits and diseases, leading to the creation of an atlas of genetic correlations (Bulik-Sullivan et al., 2015). It was discovered that almost every disease is genetically correlated with at least one other human disease or trait.

However, knowledge of genetic correlations alone is insufficient for understanding the mechanisms underlying genetic sharing, the primary mechanism of which is pleiotropy, manifesting at various biological levels. Widespread pleiotropy at the level of genetic variants has recently been demonstrated for human traits (Mackay, Anholt, 2024). An analysis of GWAS results for 116 complex traits identified 2,293 independent loci and revealed that in nearly all of these loci, the lead variants associated with one trait were also significant for at least one other trait (Qi et al., 2024).

Of particular interest is gene-level pleiotropy, as genes are the fundamental functional units of the genome. The vast majority of studies investigate the pleiotropic effects of genes on a limited set of pathologies belonging to one or two nosological categories. For instance, pleiotropy has been studied for cardiovascular (Song J. et al., 2024), psychiatric (Song Q. et al., 2025), and respiratory (Chen et al., 2024) diseases. A large group of studies comprises comparisons of a disease from one category with a set of diseases from another; for example, diabetes and cardiovascular (Adebekun et al., 2024) or gastrointestinal (Adewuyi et al., 2024) diseases, and post-traumatic stress disorder with a set of cardiovascular diseases (Shen et al., 2025). Some studies compare two sets of diseases from different categories, such as gastrointestinal and psychiatric disorders (Gong et al., 2023).

Although the findings of these studies are valuable for specialists working with specific pathologies, they do not allow for an assessment of the contribution of pleiotropy to the genetic sharing across a large number of diseases.

In this study, we estimate the contribution of pleiotropic genes to the genetic sharing among a large number of diseases from different nosological categories. To this end, we utilize publicly available GWAS summary statistics for 324 diseases from the UK Biobank cohort. From these, we select a subset of diseases in which every pair of diseases exhibits a significant genetic correlation. This subset will hereafter be referred to as the “fully connected” group.

Materials and methods

The study design is presented in Figure 1.

For 324 diseases, we estimated pairwise genetic correlations between them. A fully connected disease cluster was selected for further analysis. The analysis was performed using three approaches leveraging the GWAS results for these diseases. The first approach employs gene-level genome-wide association analysis to identify genes implicated in the etiology of at least two diseases. The second approach selects single-nucleotide polymorphisms (SNPs) significantly associated with at least two diseases and located within gene coding regions. The third approach is based on a meta-analysis of the association signals of each SNP across all diseases, followed by the identification of independent loci and prioritization of candidate genes within them. For all genes selected through these methods, we performed a bioinformatic analysis.

Materials. We utilized SAIGE summary statistics for UK Biobank phenotypes, derived from a sample of approximately 400,000 individuals of European ancestry (Zhou et al., 2018). These phenotypes corresponded to codes and nosological categories in accordance with the International Classification of Diseases PheCodes (Table S1 in Supplementary Materials)¹.

From 1,403 nosological disease codes, we selected those with more than 1,000 cases in the sample and excluded highly overlapping phenotypes. This resulted in a final set of 324 of the most common diseases. Summary statistics for each of these diseases were available for approximately 28 million single-nucleotide polymorphisms (SNPs).

The analysis included autosomal SNPs with a minor allele frequency (MAF) $> 1 \times 10^{-5}$. The major histocompatibility complex (MHC) region on chromosome 6 (positions 24 to 35 million base pairs) was excluded from the analysis.

¹ Supplementary Tables S1–S11 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Zork_Engl_30_2.xlsx

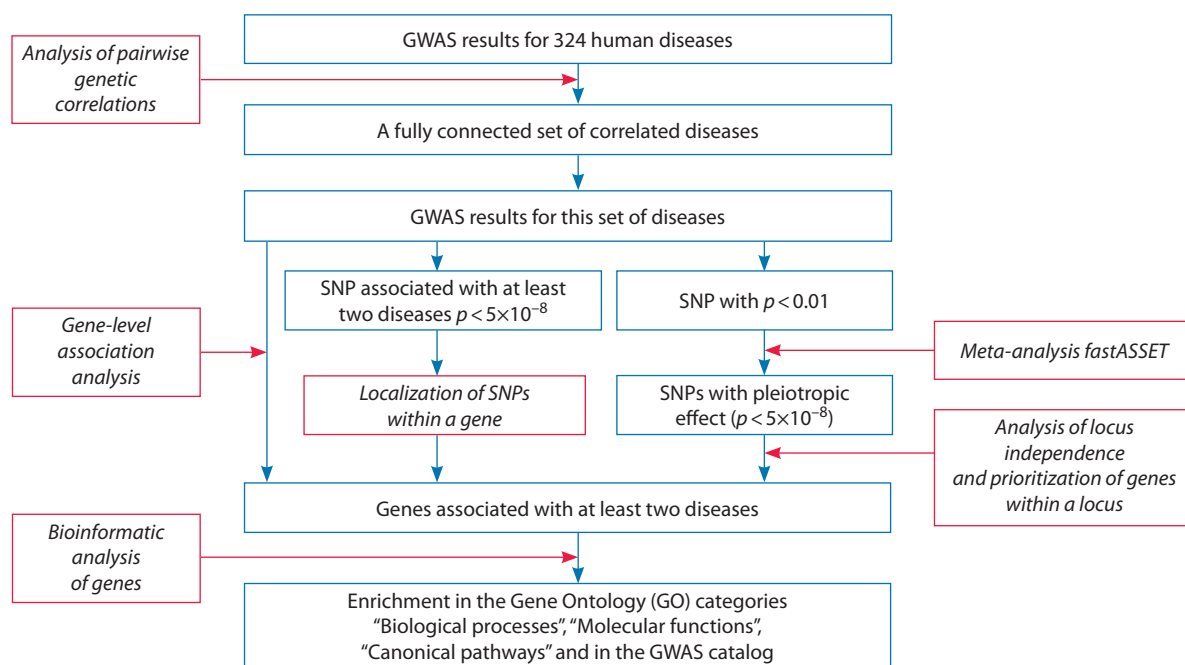


Fig. 1. Study design.

Data and results are indicated by blue boxes, while analysis methods are shown in red boxes and italicized type.

For gene-level association analysis, we used matrices of genetic correlations between SNP genotypes within genes, which we had generated previously (Belonogova et al., 2022).

Analysis of genetic correlations. For the 324 selected diseases, we estimated genome-wide genetic correlations using the LDSC software (Linkage Disequilibrium Score Regression; Bulik-Sullivan et al., 2015). We identified a fully connected cluster of diseases where all pairwise genetic correlations were nominally significant ($p \leq 0.05$).

Gene-level pleiotropy analysis. To identify genes with pleiotropic effects, we performed gene-based association analysis separately for each disease using summary statistics (z-scores and effect sizes) for each SNP and matrices of genetic correlations between all SNPs within a gene. The analysis included all SNPs located between the transcription start and end sites.

The analysis was conducted using the sumSTAAR platform (Belonogova et al., 2022). We employed two summary statistics-based methods: SKAT-O (Liu D.J. et al., 2014) and PCA (Principal Component Analysis) (Wang, Abbott, 2008). These methods were implemented in the sumFREGAT package (Svishcheva et al., 2019). We used a weight function based on minor allele frequency (MAF), defined by a beta distribution with parameters (1, 1). For the PCA-based approach, we used the first principal components that explained at least 85 % of the total trait variance.

The results from both analyses were combined using the ACAT-O method (Liu Y. et al., 2019). The analysis was restricted to protein-coding genes containing at least two SNPs with available summary statistics. The significance threshold was set at 2.5×10^{-6} (Bonferroni-corrected for 20,000 tests, $0.05/20,000$).

Some genes were located within 500 kilobases of each other. For genes with an identical pattern of association signals (i. e., associated with the same set of diseases), we were unable to prioritize a single causal gene. However, to exclude false positive association signals within these gene clusters arising from linkage disequilibrium (LD), we utilized the Open Targets Platform database (Buniello et al., 2025). We excluded genes that showed no association (association score < 0.3) with any of the diseases common to the entire gene cluster.

SNP-level pleiotropy analysis. In the first stage, we selected statistically significant associations from the GWAS data using a genome-wide significance threshold (p -value $< 5 \times 10^{-8}$). For further analysis, we considered only those SNPs that were associated with at least two different diseases. To determine the functional role of significant SNPs, we used annotations from the Variant Effect Predictor (VEP, Ensembl) (McLaren et al., 2016). If a variant was located in a gene coding region (classified as a stop-gained, missense, or synonymous variant), the corresponding gene was considered associated with the diseases for which that SNP showed a significant association. Such genes were included in the list of potentially pleiotropic genes.

Locus-level pleiotropy analysis consisted of two stages. In the first stage, we performed a meta-analysis of the association of each SNP with the full set of diseases using the fastASSET method (Fast Association Tests for Multiple Traits), implemented in the ASSET v.1.0.0 package (R/Bioconductor). In addition to assessing the significance of the pleiotropic effects for each SNP, this method identifies the specific subset of traits that contributes to the smallest meta-analysis p -value. To reduce computational time, the meta-analysis included only traits with a p -value below a specified threshold, and all traits were partitioned into relatively independent groups.

The methodological foundation of this approach is detailed in (Bhattacharjee et al., 2012; Qi et al., 2024).

For the meta-analysis of each SNP, we selected diseases with a p -value < 0.1 . A pleiotropic effect was considered significant at the genome-wide level if its meta-analysis p -value was $< 5 \times 10^{-8}$. We used the number of traits yielding significant p -values in the meta-analysis for a given SNP as a metric of its pleiotropy.

In the second stage, we defined independent loci. First, we performed clumping using PLINK v1.9 (www.cog-genomics.org/plink/1.9) with the parameters: ‘--clump-p1 5e-8 --clump-p2 1e-5 --clump-r2 0.1 --clump-kb 1000’. The analysis used linkage disequilibrium (LD) data from the 1000 Genomes Project (Phase 3, European population) and the coordinates of the reference genome GRCh37/hg19.

Subsequently, we identified independent genetic loci using the following algorithm to process the clumping results: if two index SNPs were located within 500 Kb of each other, their regions were merged, and the SNP with the smallest p -value was selected as the new index SNP. For SNPs with comparable significance (a difference in p -value of less than one order of magnitude), priority was given to the SNP located within a gene or the one associated with a larger number of traits. For SNPs within the same gene, the smallest p -value rule was applied.

For each index SNP in the independent loci, we assigned the gene in which it was located or the nearest gene within the locus boundaries. This gene was considered pleiotropic.

Bioinformatic analysis. Gene set enrichment analysis was performed using the GENE2FUNC module of the FUMA platform (Watanabe et al., 2017) (<http://fuma.ctglab.nl/>). All genes identified across the conducted analyses were used as the input gene list.

For the functional enrichment analysis of the identified genes, we utilized gene sets annotated in the Gene Ontology (GO) database (Biological Processes and Molecular Functions), canonical pathways, and the GWAS Catalog.

Default parameters were applied in all analyses. Multiple testing correction was performed using the Benjamini–Hochberg method. Results with an adjusted p -value < 0.05 were considered statistically significant.

Results

Analysis of genetic correlations

Among the 324 diseases, 220 were genetically correlated with at least one other disease. However, only 45 of these diseases formed a fully connected cluster (Fig. 2).

These diseases represented ten nosological categories based on the PheCodes classification (Fig. 3). The list of diseases and categories is provided in Table S1. The most represented categories were gastrointestinal and musculoskeletal diseases.

Gene-level pleiotropy analysis

Significant gene-based association signals were identified for 26 out of the 45 diseases (Table S2). The largest number of associated genes was found in the categories of cardiovascular pathologies (400) and metabolic disorders (165). In total,

680 genes were associated with at least one disease. Among them, 129 genes (19 %) demonstrated an association with at least two diseases.

We excluded eight genes that were in strong linkage disequilibrium with other genes and did not pass the filtering criteria based on the Open Targets Platform (see “Materials and methods” section). Consequently, 121 genes with pleiotropic effects were included in the subsequent analysis. These genes were implicated in the etiology of two to six diseases. The total number of diseases associated with at least one of the 121 genes was 19, belonging to six nosological categories (Table S3). The largest number of genes was shared between diseases of the cardiovascular category and metabolic disorders, as well as among diseases within these categories.

Analysis of SNP-level pleiotropic effects

From the GWAS results, we selected 1,389 SNPs significantly associated (p -value $< 5 \times 10^{-8}$) with at least two traits. For each of these SNPs, we determined their location relative to genes. Thirty-six SNPs were located within the protein-coding regions of 23 genes. These genes were subsequently considered as candidate genes with potential pleiotropic effects. Shared genes were identified for 11 diseases from three nosological categories (Table S4).

Analysis of locus-level pleiotropic effects

Meta-analysis identified 3,210 SNPs significantly associated ($p < 5 \times 10^{-8}$) with two or more traits. The complete list of associations is presented in Table S5. Following clumping and filtration, 71 independent loci were identified (Table S6). The largest number of independent loci was observed on chromosomes 2 and 10 (9 and 7 loci, respectively). For 11 loci, no gene could be assigned to the index SNP. That is, out of 71 index SNPs, only 60 tagged genes. Shared genes were identified for 36 diseases from nine nosological categories.

Comparison of results from different analyses

Figure 4a presents a diagram illustrating the number of pleiotropic genes detected by each method and their overlap. As can be seen, the gene-level analysis identified the maximum number of pleiotropic genes, while the SNP-level analysis identified the minimum.

Figure 4b shows the distribution of the number of diseases for which at least one gene shared with other diseases was identified, depending on the method used. The results demonstrate that the locus-level analysis revealed the largest number of such diseases, while the SNP-level analysis revealed the smallest.

Integrated analysis of results from all methods

We combined the results obtained by all methods. It turned out that six diseases – poisoning by psychotropic agents, peripheral vascular disease, rheumatoid arthritis, shortness of breath, constipation, and chemotherapy – did not share genes with any other disease. The first four diseases had the lowest prevalence among the 45 selected diseases (Table S1). It ranged from 0.005 to 0.015. Prevalence was measured as the proportion of patients in a large population sample from the UK Biobank.

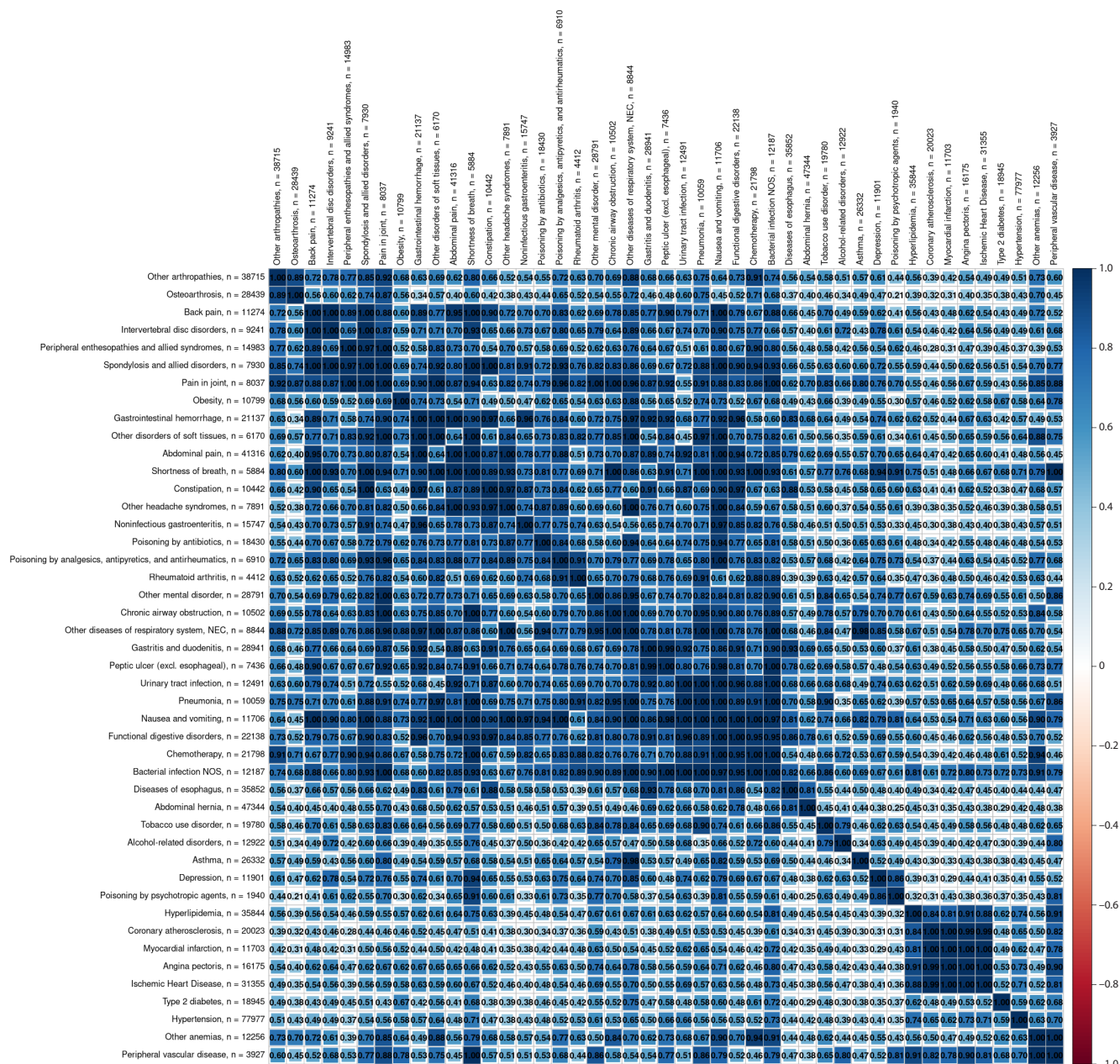


Fig. 2. Heatmap of genetic correlations among 45 traits forming a fully connected cluster.

n – the number of cases for each trait. The 'hclust' method was used for clustering.

The distribution of the remaining 39 diseases across nosological categories was similar to the distribution of the original 45 diseases presented in Figure 3.

The total number of pleiotropic genes identified by all three methods combined was 167 (Table S7). Figure 5 shows the genomic distribution of these genes. As can be seen, they are located on all autosomes but with varying density. The number of diseases associated with each gene ranges from 2 to 9.

We constructed a network of disease connections via pleiotropic genes (Fig. 6). As can be seen, diseases of the cardiovascular system and hyperlipidemia are most closely interconnected. All diseases except two (urinary tract infections and intervertebral disc disorders) are connected to at least two

other diseases. Diseases belonging to the same category are, in most cases, directly connected to each other or linked via a single disease from another category. The majority of diseases exhibit multiple connections to other diseases, both within the same category and across different categories.

Pleiotropic genes are often categorized into those shared by diseases within a single nosological category and those, the pleiotropic effects of which span diseases across different categories. Forty-four of the 167 pleiotropic genes were common only to diseases within a single category, while the remainder connected diseases from two to six categories. The most pleiotropic genes were *LPA*, *TCF7L2*, *SLC22A3*, *FES*, *CDKN2B*, and *APOE*, each associated with 7 to 9 diseases.

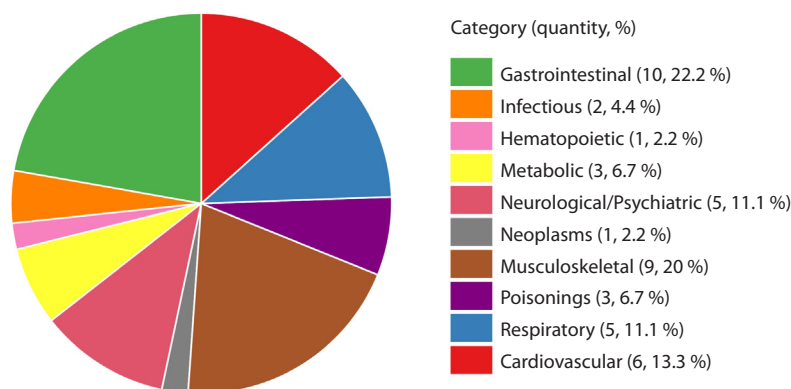


Fig. 3. Distribution of the analyzed diseases across nosological categories.

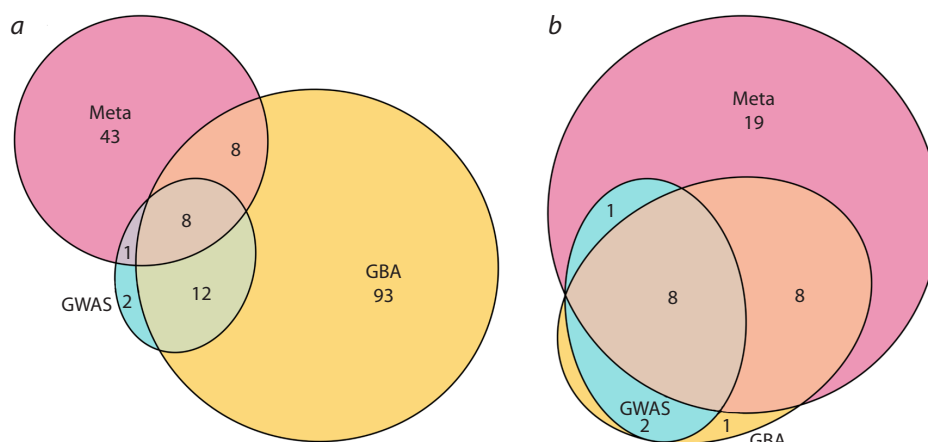


Fig. 4. Diagram of the results from the three analytical methods (Gene-Based Analysis (GBA), Locus-Level (Meta), and SNP-Level (GWAS)).

The circles represent: *a* – the number of genes identified by each method; *b* – the number of diseases sharing at least one gene with other diseases for each method.

Bioinformatic analysis

We identified 261 statistically significantly enriched Gene Ontology (GO) categories: 252 categories in the “Biological Process” (BP) ontology and nine categories in the “Molecular Function” (MF) ontology. Among the enriched biological processes, the largest number of genes were associated with the regulation of homeostasis, response to endogenous stimuli, intercellular signaling, small molecule metabolic processes, regulation of cell proliferation, transport, and catalytic activity. Among the enriched molecular functions, the largest number of genes were assigned to the category “signaling receptor binding”. The enriched canonical pathways with the largest number of genes from our set were involved in “signaling receptor binding” and “transport of small molecules”. However, the canonical pathways demonstrating the highest statistical significance of enrichment were those related to lipid metabolism. These included: statin-mediated inhibition of cholesterol synthesis, cholesterol metabolism, plasma lipoprotein assembly, remodeling, and clearance, LDL, HDL, and triglyceride metabolic pathways (including associated diseases), plasma lipoprotein remodeling, and a pathway associated with familial hyperlipidemia type 2. Complete gene lists for the enriched biological processes,

molecular functions, and canonical pathways are provided in Tables S8–S10.

Enrichment analysis using gene sets from the GWAS Catalog identified 501 significantly enriched traits. The most significant associations were observed with cardiovascular diseases (ischemic heart disease, myocardial infarction), blood pressure traits, lipid metabolism phenotypes, and body mass index (Table S11). This aligns well with the finding that the most significantly enriched canonical pathways are related to lipid metabolism.

Discussion

In this study, we investigated the contribution of pleiotropic genes to the shared genetic basis of common human diseases. We examined a fully connected cluster of 45 diseases. The analysis included not only classic multifactorial diseases but also conditions, the development of which is largely associated with external factors. However, the very fact that these diseases exhibit significant genetic correlations with other diseases indicates the existence of a common genetic component. We showed that 39 of them (87 %) share genes with at least one other disease and can be represented as a connected graph (Fig. 6). For the remaining six diseases, we did not



Fig. 5. Genomic map of gene-disease associations.
Each point corresponds to one disease, and its color represents the disease's nosological category.

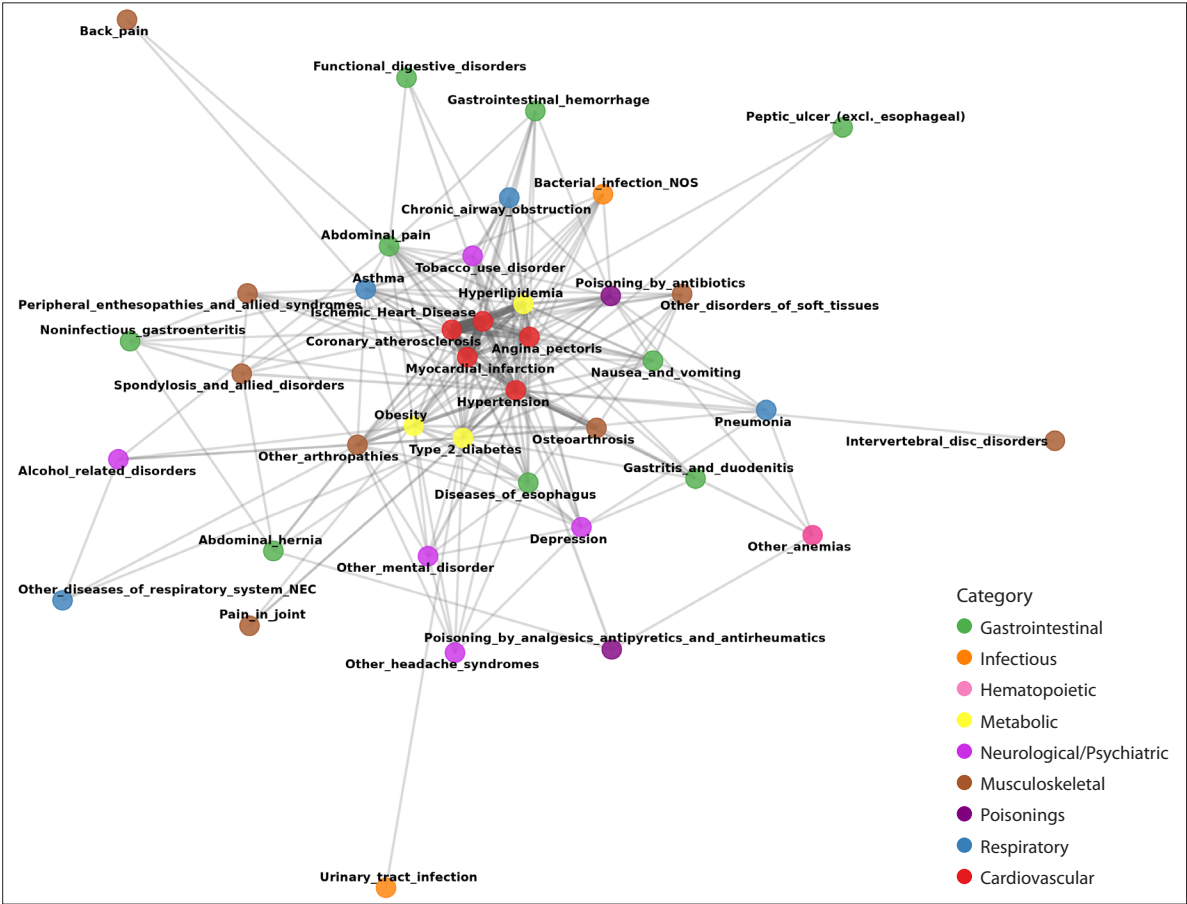


Fig. 6. Network of diseases connected by shared genes.
The color of the nodes corresponds to the disease's nosological category. The edges (lines) represent genes shared by a pair of diseases, and the edge thickness corresponds to the number of shared genes.

detect shared genes due to insufficient information about the diseases because of their low prevalence or the manifestation of pleiotropy at levels other than the gene level. This is supported by the fact that gene-level association analysis did not reveal a single significant gene for five of these diseases.

We employed three distinct approaches to identify pleiotropic genes, based on SNP-level, locus-level, and gene-level analyses. Collectively, these methods identified 167 genes with pleiotropic effects. A comparison of the approaches revealed that the largest number of pleiotropic genes was identified using gene-based association analysis. This method simultaneously leverages information from all SNPs within a gene during association testing. It is considered the most powerful approach for detecting associated genes, particularly when the analysis incorporates rare SNPs (Lee et al., 2012). The least powerful approach was the SNP-based analysis for identifying pleiotropic genes. However, this is likely less attributable to a fundamental limitation of the approach itself and more a consequence of the stringent gene annotation criteria we applied to SNPs to minimize false-positive results.

The genes we identified exhibited varying degrees of pleiotropy, with the number of associated diseases ranging from two to nine. The most pleiotropic genes were *LPA*, *TCF7L2*, *SLC22A3*, *FES*, *CDKN2B*, and *APOE*, each associated with 7–9 diseases. It is well-established that these genes play a key role in fundamental cellular and metabolic processes, which explains their broad influence on diverse phenotypes.

For instance, the *TCF7L2* gene encodes a transcription factor that is a key component of the Wnt signaling pathway. This pathway regulates cell growth, motility, survival, and differentiation during development (Jin, Liu, 2008; Facchinello et al., 2017). The *CDKN2B* gene encodes a protein that acts as a cell growth regulator and controls the progression through the G1 phase of the cell cycle (Xia et al., 2021). The *FES* gene encodes a tyrosine kinase involved in regulating the actin cytoskeleton, microtubule assembly, as well as processes of cell adhesion and migration (Laurent et al., 2004). Disruption of such fundamental biological processes logically leads to an increased risk of developing a wide spectrum of diseases.

We performed a bioinformatic analysis of all 167 pleiotropic genes by assessing their representation in various functional annotations. The analysis revealed significant enrichment across multiple categories, indicating the broad functional involvement of these genes. Bioinformatic analysis showed that the identified pleiotropic genes are significantly enriched in associations with hundreds (501) of traits and diseases from the GWAS catalog, confirming their high pleiotropic potential. This universality is explained by the involvement of these genes in key biological processes, including the maintenance of homeostasis, intercellular signaling, regulation of cell proliferation, transport, and catalytic activity, as well as by their diverse molecular functions, such as signaling receptor binding.

Pleiotropic gene effects largely explain the phenomenon of comorbidity – the simultaneous development of several diseases in a single patient, which significantly complicates diagnosis and treatment (Gratten, Visscher, 2016). Identifying pleiotropic genes allows for the identification of new therapeutic

targets and the development of drugs that simultaneously address multiple pathologies (Bao et al., 2024). Pleiotropy also opens the possibility of repurposing existing drugs for new diseases (Pushpakom et al., 2019). However, gene pleiotropy also complicates the development of targeted drugs, requiring the assessment of potential adverse effects on other diseases (Nguyen et al., 2019).

Conclusion

In conclusion, we have demonstrated that 87 % of the diseases forming a fully connected cluster share genes with at least one other disease. Furthermore, all these diseases were integrated into a single connected network. The genes we identified influence diseases belonging to both single and multiple nosological categories. Collectively, these findings indicate that the genetic correlations between diseases are largely driven by the pleiotropic effects of genes.

References

- Adebekun J., Nadig A., Saarah P., Asgari S., Kachuri L., Alagpulis D.A. Genetic relations between type 1 diabetes, coronary artery disease and leukocyte counts. *Diabetologia*. 2024;67(11):2518–2529. doi 10.1007/s00125-024-06247-9
- Adewuyi E.O., Porter T., O'Brien E.K., Olaniru O., Verdile G., Laws S.M. Genome-wide cross-disease analyses highlight causality and shared biological pathways of type 2 diabetes with gastrointestinal disorders. *Commun Biol*. 2024;7(1):643. doi 10.1038/s42003-024-06333-z
- Bao C., Tan T., Wang S., Gao C., Lu C., Yang S., Diao Y., Jiang L., Jing D., Chen L., Lv H., Fang H. A cross-disease, pleiotropy-driven approach for therapeutic target prioritization and evaluation. *Cell Rep Methods*. 2024;4(4):100757. doi 10.1016/j.crmeth.2024.100757
- Belonogova N.M., Svishcheva G.R., Kirichenko A.V., Zorkoltseva I.V., Tsepilov Y.A., Axenovich T.I. sumSTAAR: a flexible framework for gene-based association studies using GWAS summary statistics. *PLoS Comput Biol*. 2022;18(6):e1010172. doi 10.1371/journal.pcbi.1010172
- Bhattacharjee S., Rajaraman P., Jacobs K.B., Wheeler W.A., Melin B.S., Hartge P.; GliomaScan Consortium; Yeager M., Chung C.C., Channock S.J., Chatterjee N. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet*. 2012;90(5):821–835. doi 10.1016/j.ajhg.2012.03.015
- Bulik-Sullivan B., Finucane H.K., Anttila V., Gusev A., Day F.R., Loh P.R.; ReproGen Consortium; ... Patterson N., Robinson E.B., Daly M.J., Price A.L., Neale B.M. An atlas of genetic correlations across human diseases and traits. *Nat Genet*. 2015;47(11):1236–1241. doi 10.1038/ng.3406
- Buniello A., Suveges D., Cruz-Castillo C., Llinares M.B., Cornu H., Lopez I., Tsukanov K., ... Ghoussaini M., Dunham I., Hulcoop D.G., McDonagh E.M., Ochoa D. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res*. 2025;53(D1):D1467–D1475. doi 10.1093/nar/gkaf1128
- Chen J., Zhang X., Sun G. Causal relationship between type 2 diabetes and common respiratory system diseases: a two-sample Mendelian randomization analysis. *Front Med (Lausanne)*. 2024;11:1332664. doi 10.3389/fmed.2024.1332664
- Dong G., Feng J., Sun F., Chen J., Zhao X.M. A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank. *Genome Med*. 2021;13(1):110. doi 10.1186/s13073-021-00927-6
- Facchinello N., Tarifeno-Saldivia E., Grisan E., Schiavone M., Peron M., Mongera A., Ek O., Schmitner N., Meyer D., Peers B., Tiso N., Argenton F. Tcf7l2 plays pleiotropic roles in the control of glucose

- homeostasis, pancreas morphology, vascularization and regeneration. *Sci Rep.* 2017;7(1):9605. doi 10.1038/s41598-017-09867-x
- Gong W., Guo P., Li Y., Liu L., Yan R., Liu S., Wang S., Xue F., Zhou X., Yuan Z. Role of the gut-brain axis in the shared genetic etiology between gastrointestinal tract diseases and psychiatric disorders: a genome-wide pleiotropic analysis. *JAMA Psychiatry.* 2023;80(4):360-370. doi 10.1001/jamapsychiatry.2022.4974
- Gratten J., Visscher P.M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* 2016;8(1):78. doi 10.1186/s13073-016-0332-x
- Jia G., Li Y., Zhong X., Wang K., Pividori M., Alomairy R., Espósito A., ... Kubo M., Cox N.J., Evans J., Gao X., Rzhetsky A. The high-dimensional space of human diseases built from diagnosis records and mapped to genetic loci. *Nat Comput Sci.* 2023;3(5):403-417. doi 10.1038/s43588-023-00453-y
- Jin T., Liu L. The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. *Mol Endocrinol.* 2008;22(11):2383-2392. doi 10.1210/me.2008-0135
- Laurent C.E., Delfino F.J., Cheng H.Y., Smithgall T.E. The human c-Fes tyrosine kinase binds tubulin and microtubules through separate domains and promotes microtubule assembly. *Mol Cell Biol.* 2004;24(21):9351-9358. doi 10.1128/MCB.24.21.9351-9358.2004
- Lee S., Wu M.C., Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics.* 2012;13(4):762-775. doi 10.1093/biostatistics/kxs014
- Liu D.J., Peloso G.M., Zhan X., Holmen O.L., Zawistowski M., Feng S., Nikpay M., ... Willer C.J., Hveem K., Melander O., Kathiresan S., Abecasis G.R. Meta-analysis of gene-level tests for rare variant association. *Nat Genet.* 2014;46(2):200-204. doi 10.1038/ng.2852
- Liu Y., Chen S., Li Z., Morrison A.C., Boerwinkle E., Lin X. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet.* 2019;104(3):410-421. doi 10.1016/j.ajhg.2019.01.002
- Mackay T.F.C., Anholt R.R.H. Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nat Rev Genet.* 2024;25(9):639-657. doi 10.1038/s41576-024-00711-3
- McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R., Thormann A., Flicek P., Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122. doi 10.1186/s13059-016-0974-4
- Nguyen P.A., Born D.A., Deaton A.M., Nioi P., Ward L.D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat Commun.* 2019;10(1):1579. doi 10.1038/s41467-019-09407-3
- Pushpakom S., Iorio F., Eyers P.A., Escott K.J., Hopper S., Wells A., Doig A., Williams T., Latimer J., McNamee C., Norris A., Sanseau P., Cavalla D., Pirmohamed M. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov.* 2019;18(1):41-58. doi 10.1038/nrd.2018.168
- Qi G., Chhetri S.B., Ray D., Dutta D., Battle A., Bhattacharjee S., Chatterjee N. Genome-wide large-scale multi-trait analysis characterizes global patterns of pleiotropy and unique trait-specific variants. *Nat Commun.* 2024;15(1):6985. doi 10.1038/s41467-024-51075-5
- Rzhetsky A., Wajngurt D., Park N., Zheng T. Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci USA.* 2007;104(28):11694-11699. doi 10.1073/pnas.0704820104
- Shen J., Valentim W., Frilighou E., Overstreet C., Choi K.W., Koller D., O'Donnell C.J., ... Lv H., Sun L., Falcone G.J., Polimanti R., Pathak G.A. Shared genetic architecture of posttraumatic stress disorder with cardiovascular imaging, risk, and diagnoses. *Nat Commun.* 2025;16(1):5631. doi 10.1038/s41467-025-60487-w
- Song J., Gao N., Chen Z., Xu G., Kong M., Wei D., Sun Q., Dong A. Shared genetic etiology of vessel diseases: a genome-wide multi-traits association analysis. *Thromb Res.* 2024;241:109102. doi 10.1016/j.thromres.2024.109102
- Song Q., Zhang C., Wang W., Wang C., Yi C. Exploring the genetic landscape of the brain-heart axis: a comprehensive analysis of pleiotropic effects between heart disease and psychiatric disorders. *Prog Neuropsychopharmacol Biol Psychiatry.* 2025;136:111172. doi 10.1016/j.pnpbp.2024.111172
- Svishcheva G.R., Belonogova N.M., Zorkoltseva I.V., Kirichenko A.V., Axenovich T.I. Gene-based association tests using GWAS summary statistics. *Bioinformatics.* 2019;35(19):3701-3708. doi 10.1093/bioinformatics/btz172
- Wang K., Abbott D. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol.* 2008;32(2):108-118. doi 10.1002/gepi.20266
- Wang K., Gaitsch H., Poon H., Cox N.J., Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat Genet.* 2017;49(9):1319-1325. doi 10.1038/ng.3931
- Watanabe K., Taskesen E., van Bochoven A., Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826. doi 10.1038/s41467-017-01261-5
- Xia Q., Liu Y., Yang C., Simeone D.M., Sun T.T., DeGraff D.J., Tang M.S., Zhang Y., Wu X.R. Dominant role of *CDKN2B/p15INK4B* of 9p21.3 tumor suppressor hub in inhibition of cell-cycle and glycolysis. *Nat Commun.* 2021;12(1):2047. doi 10.1038/s41467-021-22327-5
- Zhou W., Nielsen J.B., Fritsche L.G., Dey R., Gabrielsen M.E., Wolford B.N., LeFaive J., ... Hveem K., Kang H.M., Abecasis G.R., Willer C.J., Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet.* 2018;50(9):1335-1341. doi 10.1038/s41588-018-0184-y

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

Received September 4, 2025. Revised November 1, 2025. Accepted December 1, 2025.

doi 10.18699/vjgb-26-29

Novel regulatory SNPs that can be activated due to metformin treatment may orchestrate liver gluconeogenesis and add to the variability in AMPK-dependent mechanisms of metformin response

E.E. Korbolina , I.S. Damarov, T.I. Merkulova 

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

 lungry@bionet.nsc.ru

Abstract. Metformin is a first-line therapy for type 2 diabetes, yet individual response varies significantly, with over 30 % of patients failing to achieve optimal glycemic control. The specific regulatory mechanisms of this phenomenon remain poorly understood and genetic variants involved are mainly undiscovered. There are multiple lines of evidence that the leading role in determining the variance in phenotypic outcome belongs to regulatory SNPs (rSNPs) as they directly modify gene expression. Therefore, the genome-wide search for such functional variants and deciphering associated phenotypes stands as a fundamental challenge. Previously, based on the results of bioinformatics analysis of allele-specific expression and binding landscape in human peripheral blood mononuclear cells, we have established an original panel of 14 796 rSNPs within promoters of 5132 genes. Aiming to pinpoint functional variants most likely linked to metformin hepatic response and impacts on liver gluconeogenesis, we analyzed the relevant open-access data as well as rSNPs from our panel and the corresponding genes. 1196 genes reported to be regulated by metformin in human hepatocytes and 115 genes involved in gluconeogenesis and/or its regulation via Gene Ontology annotations were intersected. Free R software and STRING v.11 tools were used for functional annotation. A number of genes harboring rSNPs within promoter regions were found to be particularly implicated in the mechanisms of metformin's action. Functional enrichment analyses revealed enrichment in critical pathways including FoxO, TNF- α and TGF- β signaling, also implicated in diabetes complications. Among these, six genes (*ARPP19*, *ATF4*, *NR3C1*, *PFKFB3*, *TCF7L2*, and *WDR5*) were strongly associated with regulation of gluconeogenesis, and may be modulated by metformin in the liver. We conclude that metformin therapy response may be influenced by the newly identified functional SNPs including rSNPs within the promoters of genes for gluconeogenic enzymes and transcription regulators.

Key words: metformin; individual drug response; TD2M; regulatory SNPs; gluconeogenesis; AMPK-dependent mechanisms; signal transduction pathways (FoxO, TNF- α , TGF- β)

For citation: Korbolina E.E., Damarov I.S., Merkulova T.I. Novel regulatory SNPs that can be activated due to metformin treatment may orchestrate liver gluconeogenesis and add to the variability in AMPK-dependent mechanisms of metformin response. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):259-266. doi 10.18699/vjgb-26-29

Funding. This study was supported by the RSF, Russian Science Foundation (No. 23-15-00113).

Новые регуляторные однонуклеотидные полиморфизмы (rSNPs) потенциально участвуют в регуляции глюконеогенеза в печени и вносят вклад в вариабельность АМФК-зависимых механизмов индивидуального ответа на метформин

E.E. Корболина , И.С. Дамаров, Т.И. Меркулова 

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

 lungry@bionet.nsc.ru

Аннотация. Метформин является препаратом первого выбора для лечения сахарного диабета второго типа, однако более чем у 30 % пациентов не достигается оптимальный гликемический контроль. Лежащие в основе этого феномена регуляторные механизмы и ассоциированные генетические варианты остаются неизученными. Накопленные данные свидетельствуют о том, что ключевую роль в определении фенотипической вариабельности играют функциональные генетические варианты, в особенности регуляторные однонуклеотидные полиморфизмы (rSNP), непосредственно влияющие на уровень экспрессии генов. Полногеномный поиск таких функциональных вариантов и выявление их фенотипических последствий – актуальная научная задача. Ранее на основе результатов биоинформатического анализа аллель-специфической экспрессии и профиля распределения активирующих гистоновых модификаций в мононуклеарных клетках периферической крови человека нами была сформирована

оригинальная панель из 14 796 rSNP, локализованных в промоторах 5132 генов. Целью настоящей работы был поиск в нашей панели rSNP вариантов, потенциально связанных с регуляцией процессов глюконеогенеза в печени, в том числе под действием монотерапии метформином, с помощью анализа независимых релевантных полногеномных данных. Анализировали 1196 генов, экспрессия которых в гепатоцитах человека изменялась под влиянием метформина АМФК-зависимым образом, а также 115 генов, вовлеченных в глюконеогенез и/или его регуляцию, согласно данным генных онтологий. Для функциональной аннотации использовали инструменты свободного программного обеспечения R и базы данных STRING (v.11). В результате анализа был выделен ряд генов, которые, вероятно, играют особую роль в механизмах действия метформина и индивидуальной чувствительности к препарату, в промоторных районах которых нами были картированы rSNP. Функциональный анализ обогащения выявил вовлеченность этих генов в ключевые сигнальные пути, ассоциированные также с осложнениями сахарного диабета второго типа. Для шести генов (*ARPP19*, *ATF4*, *NR3C1*, *PFKFB3*, *TCF7L2* и *WDR5*) экспрессия не только тесно связана с регуляцией глюконеогенеза, но также может модулироваться в гепатоцитах под действием метформина. Мы предполагаем, что rSNP в промоторах генов ферментов, участвующих в глюконеогенезе, и транскрипционных факторов могут играть существенную роль в механизмах формирования индивидуального ответа на терапию метформином.

Ключевые слова: метформин; индивидуальный ответ на препарат; сахарный диабет второго типа; регуляторные полиморфизмы (rSNPs); глюконеогенез; АМФК-зависимая киназа; FoxO, TNF- α , TGF- β -сигнальные пути

Introduction

Metformin, an oral biguanide agent, is the first-line medication to reduce blood glucose levels in patients with T2DM. Although its mechanism of action is still not fully understood, metformin has been shown to decrease liver gluconeogenesis and enhance insulin sensitivity in muscle and adipose tissue (Rena et al., 2017; LaMoia, Shulman, 2021). Moreover, its beneficial pleiotropic effects beyond glycemic control are well known, which include lowering the incidence of cardiovascular events (Bu et al., 2022), correcting energy metabolism in neurologic disorders (Loan et al., 2024), and regulating inflammatory markers (Karbalaei-Hasani et al., 2021).

At the molecular level, metformin can inhibit complex I of the mitochondrial electron transport chain in hepatocytes (El-Mir et al., 2000), leading to decreases in ATP and increases in AMP levels and in turn activating AMP-activated protein kinase (AMPK), a master controller of metabolic homeostasis (Smiles et al., 2024). The activation of AMPK results in suppression of gluconeogenic gene expression and reduction in glucose output (LaMoia, Shulman, 2021). Moreover, there is evidence that the metformin glucose-lowering effects in hepatocytes can occur via AMPK-independent mechanisms. In particular, increased AMP levels can directly suppress gluconeogenesis via allosteric inhibition of fructose-1,6-bisphosphatase 1, FBP1 (Hunter et al., 2018). There is also an increasing interest in examining the effects of AMPK activity modulation in chronic metabolic diseases and cancer (Strang et al., 2025), including using the activators and inhibitors with effects independent of AMPK, i. e. compound C, an ATP-competitive AMPK inhibitor (which was reported to disrupt various biological events).

However, metformin's response varies among individuals. As is shown, metformin is ineffective in over 30 % of cases because of various factors, including genetic ones (Cook et al., 2007; St-Amour et al., 2025). Currently, the search for genetic variants (mainly SNPs) associated with drug response efficacy is mainly focused on the genes coding for phase I and II drug metabolizing enzymes, drug transporters, and a number of upstream transcription regulators (Gaedigk et al., 2020; Rykova et al., 2022). Metformin is not metabolized in the human body (Gong et al., 2012); correspondingly, most

studies aimed to identify the SNPs associated with the response to metformin were focused on the genes coding for organic cation transporters (OCTs) and the multidrug and toxin extrusion (MATE) proteins. This allowed for detection of a number of SNPs in the genes of these groups (Zhou et al., 2015; Xiao et al., 2016). The use of unbiased genome-wide approach (GWAS) succeeded in finding several SNPs associated with the response to metformin in genes, the products of which are involved in other functions. These SNPs have been discovered in regulatory regions of the *ATM* gene, the product of which is involved in the maintenance of redox homeostasis in the cell (GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group et al., 2011); the *PRPF31* gene, coding for pre-mRNA processing factor 31; the *CPA6* (carboxypeptidase A6) gene, associated with the regulation of the Akt/mTOR signaling pathway; and the *STAT3* gene, coding for eponymous transcription factor, which is a well-known regulator of metabolic and immune processes (Rotroff et al., 2018).

Nevertheless, GWAS application in pharmacogenomics yet failed to provide a considerable progress, mainly because of the difficulties (and more often, impossibility) in forming large cohorts (tens and hundreds of thousands) of patients to determine the effects of each drug. Modern methods of functional genomics – eQTL analysis and detection of allele-specific events in omics data – do not suffer from this limitation (Degtareva et al., 2021). However, the studies utilizing these approaches are few and have even absent at all for antidiabetic drugs, including metformin. Earlier, we measured the allelic imbalance at the level of gene expression and the profiles of active chromatin marks in the paired RNA-seq and ChIP-seq data for peripheral blood mononuclear cells (PBMCs) of nine healthy donors and constructed a panel of 14 796 regulatory SNPs (rSNPs) potentially able to influence the expression levels of the genes harboring these SNPs in their promoters (Damarov et al., 2024). Then, using the RNA-seq data (GSE153315, GEO DataSets) on the blood cells of T2DM non-responders to metformin ($N = 10$) and T2DM responders ($N = 10$) (Vohra et al., 2022), we found 367 rSNPs from our panel in the promoters of 131 corresponding differentially expressed genes (DEGs). Primary analysis of this gene set revealed a number of transcriptional regulators both known

to be involved in T2DM pathogenesis and still poorly studied (Damarov et al., 2024). Here, we performed further analysis of the obtained rSNP list using independent transcriptomic data on metformin hepatic response (Luizon et al., 2016) and available data on metformin impacts on liver gluconeogenesis.

Materials and methods

Gene sets analyzed. Three original gene sets were compared, and first was the set of 5132 genes harboring the rSNPs in promoter regions based on our previous analysis of allele-asymmetric events in omics data (RNA-seq and ChIP-seq) for human PBMCs of nine healthy donors (Damarov et al., 2024). Each rSNP in the reported panel was located within ± 1000 bp from known TSS (transcription start site) and associated with allele-asymmetric binding of activating histone modifications (H3K4me3 and/or H3K27ac, ChIP-seq). Moreover, when analyzing RNA-seq data obtained for the same PBMCs samples, allele-specific expression events were observed within the transcribed sequences of the corresponding genes. Based on the association of the identified regulatory variants with two types of allele-specific events, we assumed that our rSNPs were most likely affecting the expression of certain human genes.

The second gene set comprised 1196 AMPK-dependent genes that were identified through clustering analysis on all DEGs ($p_{adj} < 0.05$) identified from RNA-seq data across the three conditions (treated with metformin, metformin + an AMPK inhibitor (compound C) or no treatment for 8 hours in triplicate) of cultured primary human hepatocytes from a sole male donor (Luizon et al., 2016). The DEGs were termed AMPK-dependent when found differently expressed between hepatocytes with metformin treatment and controls; however, their expression was restored to the normal level (or similarly altered) in hepatocytes with the combined treatment of compound C and metformin. And the third gene set consisted of 115 genes that were classified as involved only in gluconeogenesis, only in regulation of gluconeogenesis, or shared both Gene ontology terms (Ashburner et al., 2000). The analyzed human datasets were under ethical consent agreements as stated in authorized submissions.

Functional annotation. To predict the biological processes and metabolic pathways associated with the analyzed gene lists, a search for associated terms from GO and KEGG gene ontologies and metabolic pathways (Kanehisa, 2000) was performed using the clusterProfiler package (v4.15.1) (Yu et al., 2012) when taking into account the Benjamini–Hochberg correction ($p_{adj} < 0.1$).

A functional protein association network (PPI) was built and analyzed by STRING v.11 tools (Szklarczyk et al., 2023). The k-means clustering was applied in STRING on the basis of evidence score and connection cutoff at 0.40; the enrichments of the KEGG and Wiki pathways were analyzed with a p -value cutoff of < 0.05 and minimum count in the network setting of 2.

Results and discussion

Matching well the generally accepted model for metformin action, our panel of rSNPs located within the promoters of DEGs associated with metformin response (Damarov et al., 2024) included rSNPs within promoters of genes for key

enzymes of mitochondrial electron transfer chain (NDUFA11 and NDUFB1) and upstream/downstream AMPK regulators (Fig. 1). So, AK5 adenylate kinase increases the phosphorylation of AMPK (Dzeja, Terzic, 2009); and N-myristoyltransferase (NMT1) controls the N-myristoylation and lysosomal localization of the AMPK protein (Chen et al., 2020). Activated AMPK inhibits the mTOR (mammalian target of rapamycin) pathway, a pivotal regulator of cell metabolism and growth (Gwinn, Shaw, 2010; Smiles et al., 2024).

To gain a better understanding of the possible involvement of our rSNPs in the mechanisms underlying the antihyperglycemic effects of metformin treatment, we compared and analyzed three gene sets (Fig. 2). The ‘AMPKdep’ set of 1196 AMPK-dependent genes associated with metformin hepatic response was obtained from (Luizon et al., 2016), the ‘rSNPs’ set of 5132 genes with the regulatory variants originally identified within promoter regions was previously reported in (Damarov et al., 2024). Then, the ‘Glu’ set consisted of genes related to gluconeogenesis in GO terms (see the Materials and Methods section for details).

Combining the results of our previous analysis with the AMPKdep gene set allowed to identify 402 genes, the expression of which in hepatocytes may be influenced by our rSNPs. The results of functional enrichment analysis using R signified that Protein processing in endoplasmic reticulum, Apoptosis and FoxO signaling KEGG pathways were enriched in a given set (Fig. 3A). Using the enrichment analysis in STRING, we observed the enrichment of the same gene list belonging to several Wiki metabolic pathways: similarly, apoptosis and, in addition, adipogenesis and signaling pathways including TNF-alpha-, TGF-beta- and angiogenesis-related VEGFA\VEGFR2-signaling (Fig. 3B).

It is worth to mention that the signaling pathways above were widely reported to be affected by metformin (Yi et al., 2016; Kristófi, Eriksson, 2021; Pan et al., 2024). For example, some mice data indicate the Foxo1 protein to be one of the key targets of metformin in regulating blood glucose and hepatic glucose production under therapeutic concentrations (Guo et al., 2021). Moreover, the same pathways are commonly known as mediators that orchestrate and propagate inflammatory responses and thereby play a crucial role in pathological signaling mechanisms shared among common T2DM complications, including diabetic retinopathy (Behl et al., 2022) and diabetic kidney disease (Ansari et al., 2025; Hou et al., 2025).

When the genes annotated to the ‘Glu’ set were put together with our rSNP panel, we discovered 28 genes in the overlap. The encoded products include important enzymes, i. e. MDH2, malate dehydrogenase 2, which catalyzes a reversible NAD-dependent dehydrogenase reaction and glycolysis enzymes (i. e. ENO1, enolase 1, and PGP, phosphoglycolate phosphatase), as well as a number of regulatory proteins (i. e. FoxO1, USP7, SIRT6). Thus, FoxO1 has a critical role in the insulin-mediated regulation of de novo glucose synthesis affecting the transcription of rate-limiting enzymes, G6Pc and Pck1 (Hall et al., 2014). Notably, multiple studies support that sirtuin 6, SIRT6, an NAD-dependent deacetylase/deacetylase/mono-ADP ribosyltransferase, is involved in inhibition of FoxO1 leading

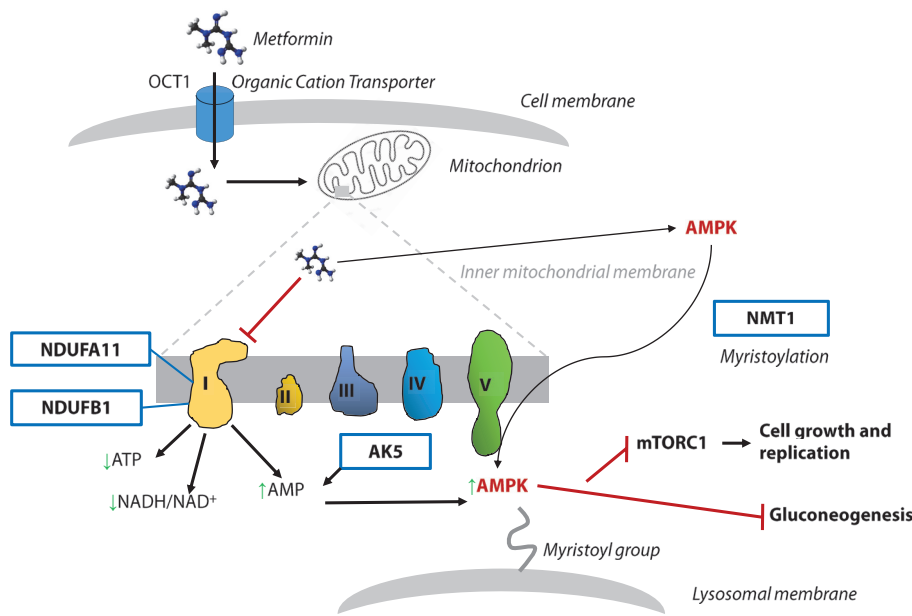


Fig. 1. Antihyperglycemic effects of metformin are implemented through inhibition of complex I of mitochondrial electron transport chain and activation of AMP kinase. The proteins associated with the AMPK pathway are denoted with blue frames. Green vertical arrows indicate a decrease or an increase in the amount of metabolite or enzyme activity; blunt red arrows indicate inhibition; black arrows indicate membrane transport or a functional connectivity; the complexes of the electron transport chain embedded in the inner mitochondrial membrane are labeled with roman numerals (I through V); NDUFA11, NADH:ubiquinone oxidoreductase subunit A11; NDUFB1, NADH:ubiquinone oxidoreductase subunit B1; AK5, adenylate kinase 5; AMPK, AMP-activated protein kinase; NMT1, N-myristoyltransferase 1; and OCT1, organic cation/carnitine transporter 1.

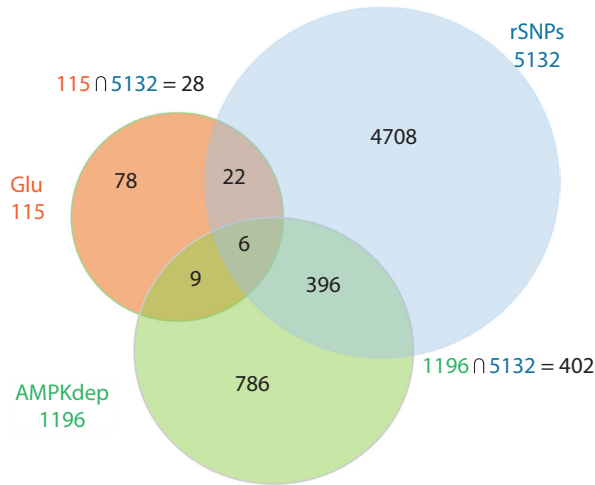


Fig. 2. Venn diagram representing the intersection between the 'Glu' (orange), 'rSNP' (blue) and 'AMPKdep' (green) gene sets. $\cap$ indicates overlapping; the numerical values in the intersections show the counts for corresponding gene groups.

to suppression of gluconeogenesis (Liu et al., 2021; Wang et al., 2023). Moreover, it was shown that the activation of deubiquitinating enzyme USP7 similarly results in a decrease in FoxO1 activity (Jiang et al., 2017).

Using STRING, we constructed and analyzed a PPI network for this subset of 28 genes (Fig. 4). Three clusters were identified by the KMeans algorithm: nine protein products of

cluster 1 were found to be involved in NADH metabolism and gluconeogenesis; seven from cluster 2, in regulation of gluconeogenesis, circadian rhythms and AMP-activated protein kinase activity (PRKAG2 and PRKACA); and cluster 3 was comprised of two subunits of protein phosphatase 4 complex: PPP4R3A and PPP4R3B (Fig. 4A). It is also to be noted that the protein products of four genes were functionally related to KEGG AMPK-signaling pathway (Fig. 4B). The results of functional enrichment analysis based on Gene Ontology are visualized in Figure 4C.

For a number of these genes, the functional links with regulation of gluconeogenesis have been well described as given on the schematic representation in Figure 5. From this background we hypothesized that our rSNPs located within the corresponding promoters may be involved in regulatory signatures.

Adding the data on AMPK-dependent metformin hepatic response allows us to highlight six genes (*ARPP19*, *ATF4*, *NR3C1*, *PFKFB3*, *TCF7L2*, and *WDR5*) from a list of 28. These genes belong to the intersection of all three compared gene sets, which means the corresponding promoters harbor rSNPs; the genes were shown to be expressed in AMPK-dependent manner in hepatocytes; and the resulting proteins were involved in gluconeogenesis.

Thus, NR3C1, better known as the glucocorticoid receptor (GR), has been functionally linked to stress response, immune regulation, inflammation and control of hepatic gluconeogenesis (Zhang et al., 2019). The encoded protein functions as a transcription factor enhancing the transcription of glu-

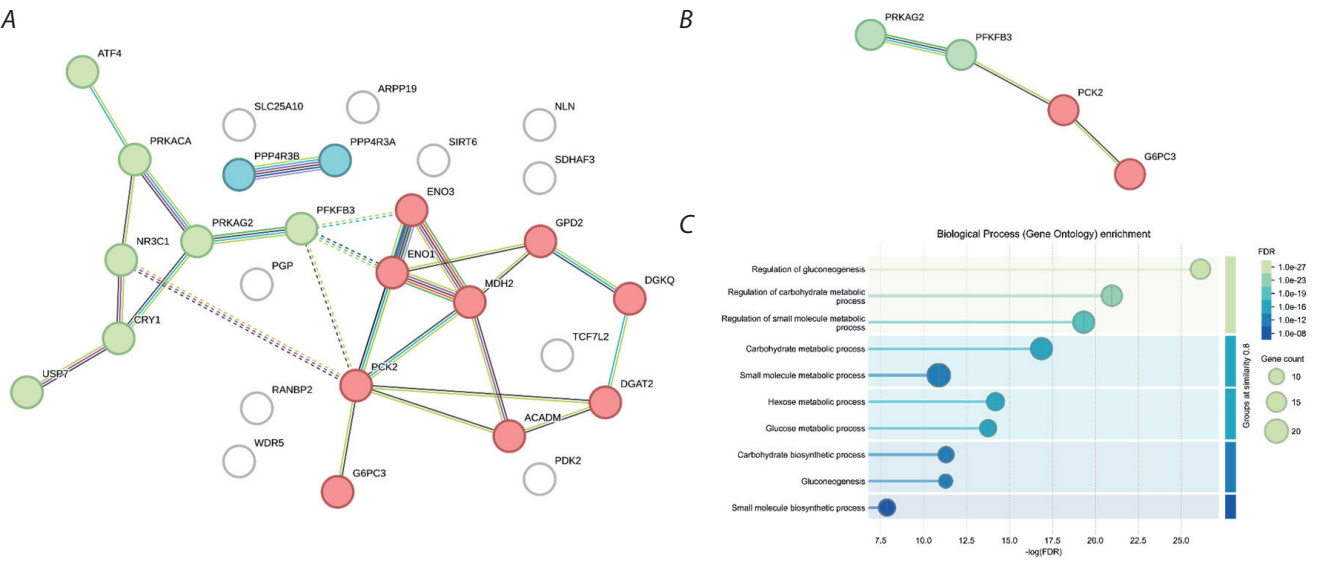


Fig. 3. Interpreting a list of 402 genes with pathway enrichment analysis. (A) The most relevant enriched KEGG terms obtained using R (Materials and Methods). (B) Wiki pathway enrichment determined and visualized by STRING. ID, gene or pathway identifier from the database; FE, Fold Enrichment; p.Adjust, a *p*-value as corrected with the Benjamini–Hochberg procedure; count, the number of genes presented in the list; and ER, endoplasmic reticulum (A). Different colors indicate top enriched terms with corresponding *p*-values (FDR) according to the color panel (B).

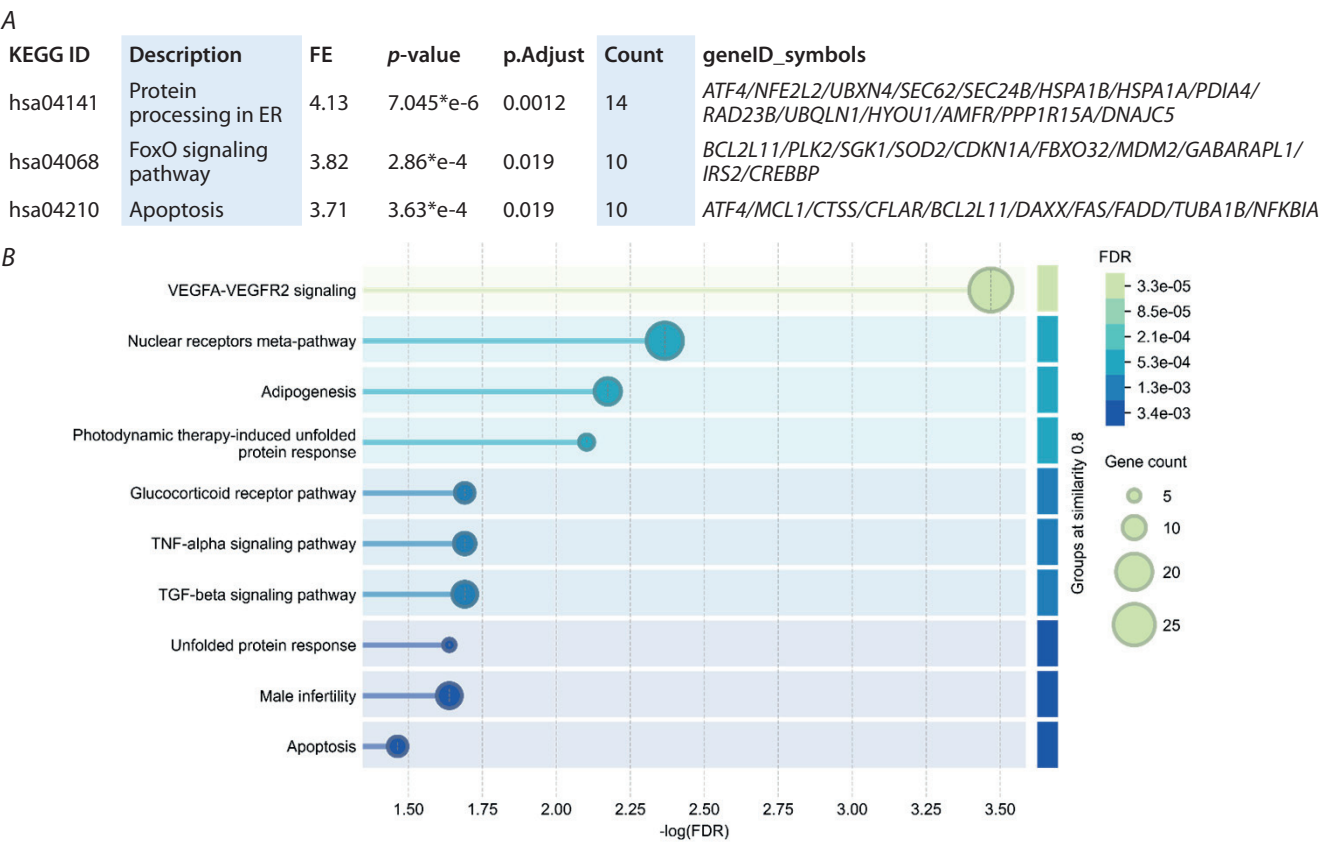


Fig. 4. PPI network for the intersection of 28 genes as by STRING. (A) The clusters identified with K-means cluster analysis. (B) Four closely connected PPI members belonging to KEGG AMPK-signaling pathway (hsa04152). (C) A visualization of the most relevant enriched Biological Process Gene Ontology terms. Nodes of different color indicate the members of different clusters (A and B): red – cluster 1, green – cluster 2, and blue – cluster 3. The solid and the dotted lines (A) indicate connections within the same and with a different cluster respectively. Edges of different color indicate both functional and physical protein interactions (A and B): cyan – from curated databases; pink – experimentally determined; blue – gene co-occurrence; khaki – from text mining; black – coexpression; light blue – protein homology. Different enriched terms with corresponding *p*-values (FDR) are highlighted in different colors according to the color panel (C).

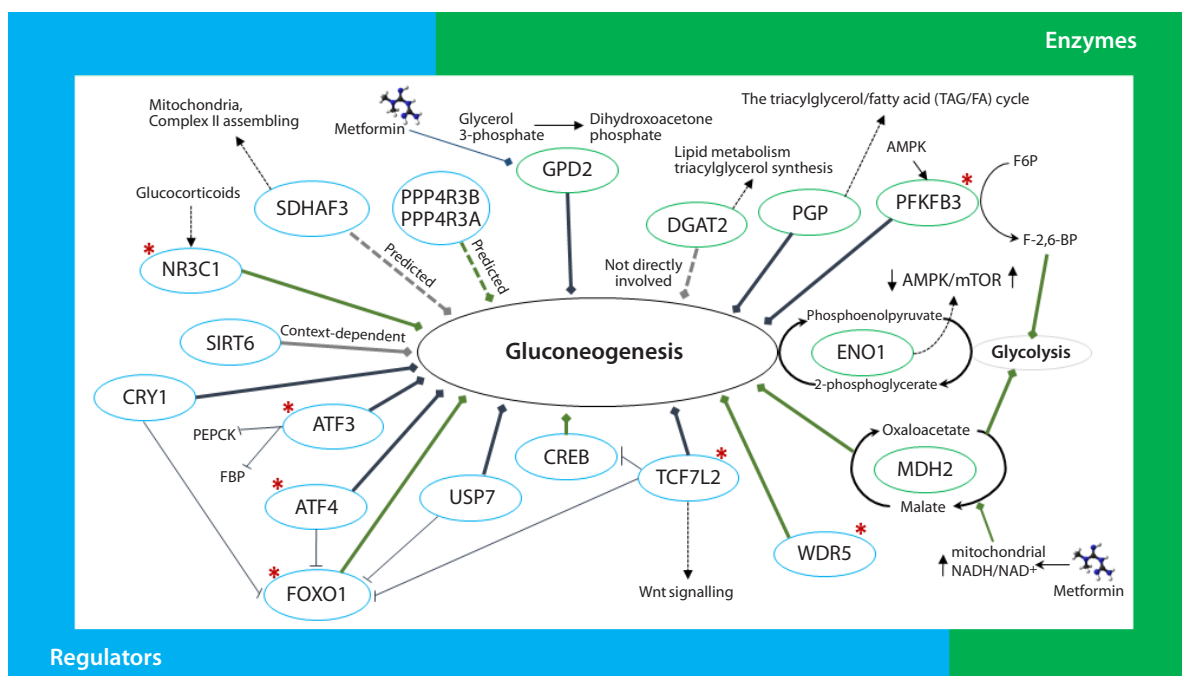


Fig. 5. Multiple potential rSNP targets, including enzymes and regulatory proteins, known to be involved in regulation of gluconeogenesis. Blunt orange arrows indicate the activation of gluconeogenesis and deep-blue ones stand for the suppression of the process; T-shaped blue arrows indicate the suppression of transcription/translation or protein activity; red stars indicate genes from the AMPKdep gene set; AMPK, 5' adenosine monophosphate-activated protein kinase; mTOR, mammalian target of rapamycin; Wnt, Wingless-related integration site; PEPCK, phosphoenolpyruvate carboxykinase; FBP, fructose-1,6-bisphosphatase; F6P, fructose 6-phosphate; F-2,6-BP, fructose-2,6-bisphosphate; NADH, nicotinamide adenine dinucleotide; and NAD⁺, the oxidized form of NAD.

coneogenic genes such as PEPCK and G6Pase under fasting conditions (Beaupere et al., 2021).

ATF4, activating transcription factor 4, is a key protein in metabolic regulation involved in insulin secretion and sensitivity control and regulation of gluconeogenesis by affecting the transcriptional activity of FoxO1 (Li et al., 2022). Interestingly, a number of findings demonstrate epigenetic regulatory mechanisms targeting ATF4. For example, it was reported that microRNA (miR)-214 upregulates ATF4 expression, leading to elevated expression of FoxO1 and thus to suppression of gluconeogenesis (Zhu et al., 2018; Wang et al., 2024).

WRD5 performs multiple scaffolding functions in the context of chromatin and was reported to modulate gluconeogenic gene expression by potentiating the activity of KAT2B histone lysine acetyltransferase. Although these two genes seem to regulate a relatively small subset of gene targets in hepatocytes, both KAT2B and WDR5 are required for the expression of a large fraction of glucagon-inducible genes, indicating that they represent the important cofactors for the CREB pathway (Ravnskjaer et al., 2013).

The TCF7L2 protein is a transcription factor that plays an important role in the Wnt signaling pathway and has been implicated in blood glucose homeostasis (Ip et al., 2012) through actions even beyond pancreatic beta cells (Bailey et al., 2015).

Finally, the protein encoded by *PFKFB3* is a bifunctional enzyme involved in both the synthesis and degradation of fructose-2,6-bisphosphate (F-2,6-BP), and is a potent activa-

tor of 6-phosphofructokinase-1 (PFK-1) – a trigger of aerobic oxidation for glucose metabolism. Recent studies have reported the pivotal role of PFKFB3 in the regulation of insulin resistance (Yang et al., 2023).

Conclusion

This work identifies several novel genes and gene regulatory variants that can be activated due to metformin treatment and thus provides candidates in the human genome where nucleotide variation can lead to differences in metformin response. We conclude that metformin therapy response may depend on the identified functional SNPs within the promoters of genes for gluconeogenic enzymes and regulatory proteins.

Список литературы / References

- Ansari Z., Chaurasia A., Neha, Sharma N., Bachheti R.K., Gupta P.C. Exploring inflammatory and fibrotic mechanisms driving diabetic nephropathy progression. *Cytokine Growth Factor Rev.* 2025;84: 120-134. doi 10.1016/j.cytogfr.2025.05.007
- Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., ... Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25-29. doi 10.1038/75556
- Bailey K.A., Savic D., Zielinski M., Park S.-Y., Wang L., Witkowski P., Brady M., Hara M., Bell G.I., Nobrega M.A. Evidence of non-pancreatic beta cell-dependent roles of Tcf7l2 in the regulation of glucose metabolism in mice. *Hum Mol Genet.* 2015;24(6):1646-1654. doi 10.1093/hmg/ddu577

- Beaupere C., Liboz A., Fève B., Blondeau B., Guillemain G. Molecular mechanisms of glucocorticoid-induced insulin resistance. *Int J Mol Sci.* 2021;22(2):623. doi 10.3390/ijms22020623
- Behl T., Wadhwa M., Sehgal A., Singh S., Sharma N., Bhatia S., Al-Harrasi A., Aleya L., Bungau S. Mechanistic insights into the role of FOXO in diabetic retinopathy. *Am J Transl Res.* 2022;14(6):3584-3602
- Bu Y., Peng M., Tang X., Xu X., Wu Y., Chen A.F., Yang X. Protective effects of metformin in various cardiovascular diseases: Clinical evidence and AMPK-dependent mechanisms. *J Cell Mol Med.* 2022;26(19):4886-4903. doi 10.1111/jcmm.17519
- Chen Y.-C., Navarrete M.S., Wang Y., McClintock N.C., Sakurai R., Wang F., Chen K.T., Chou T.-F., Rehan V.K., Lee D.J., Diaz B. N-myristoyltransferase-1 is necessary for lysosomal degradation and mTORC1 activation in cancer cells. *Sci Rep.* 2020;10(1):11952. doi 10.1038/s41598-020-68615-w
- Cook M.N., Girman C.J., Stein P.P., Alexander C.M. Initial monotherapy with either metformin or sulphonylureas often fails to achieve or maintain current glycaemic goals in patients with Type 2 diabetes in UK primary care. *Diabet Med.* 2007;24(4):350-358. doi 10.1111/j.1464-5491.2007.02078.x
- Damarov I.S., Korboлина E.E., Rykova E.Y., Merkulova T.I. Multi-omics analysis revealed the rSNPs potentially involved in T2DM pathogenic mechanism and metformin response. *Int J Mol Sci.* 2024;25(17):9297. doi 10.3390/ijms25179297
- Degtyareva A.O., Antontseva E.V., Merkulova T.I. Regulatory SNPs: altered transcription factor binding sites implicated in complex traits and diseases. *Int J Mol Sci.* 2021;22(12):6454. doi 10.3390/ijms22126454
- Dzeja P., Terzic A. Adenylate kinase and AMP signaling networks: metabolic monitoring, signal communication and body energy sensing. *Int J Mol Sci.* 2009;10(4):1729-1772. doi 10.3390/ijms10041729
- El-Mir M.-Y., Nogueira V., Fontaine E., Avéret N., Rigoulet M., Leverve X. Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. *J Biol Chem.* 2000;275(1):223-228. doi 10.1074/jbc.275.1.223
- Gaedigk A., Whirl-Carrillo M., Pratt V.M., Miller N.A., Klein T.E. PharmVar and the landscape of pharmacogenetic resources. *Clin Pharmacol Ther.* 2020;107(1):43-46. doi 10.1002/cpt.1654
- GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group; Wellcome Trust Case Control Consortium 2; Zhou K., Bellenguez C., Spencer C.C., Bennett A.J., Coleman R.L., ... McCarthy M.I., Holman R.R., Palmer C.N., Donnelly P., Pearson E.R. Common variants near *ATM* are associated with glycemic response to metformin in type 2 diabetes. *Nat Genet.* 2011;43(2):117-120. doi 10.1038/ng.735
- Gong L., Goswami S., Giacomini K.M., Altman R.B., Klein T.E. Metformin pathways. *Pharmacogenet Genomics.* 2012;22(11):820-827. doi 10.1097/FPC.0b013e3283559b22
- Guo X., Li X., Yang W., Liao W., Shen J.Z., Ai W., Pan Q., Sun Y., Zhang K., Zhang R., Qiu Y., Dai Q., Zheng H., Guo S. Metformin targets Foxo1 to control glucose homeostasis. *Biomolecules.* 2021;11(6):873. doi 10.3390/biom11060873
- Gwinn D.M., Shaw R.J. AMPK Control of mTOR signaling and growth. In: Tamanoi F., Hall M.N. (Eds). *The Enzymes*. Vol. 28. Academic Press, 2010;49-75. doi 10.1016/S1874-6047(10)28003-4
- Hall J.A., Tabata M., Rodgers J.T., Puigserver P. USP7 attenuates hepatic gluconeogenesis through modulation of FoxO1 gene promoter occupancy. *Mol Endocrinol.* 2014;28(6):912-924. doi 10.1210/me.2013-1420
- Hou G., Dong Y., Jiang Y., Zhao W., Zhou L., Cao S., Li W. Immune inflammation and metabolic interactions in the pathogenesis of diabetic nephropathy. *Front Endocrinol. (Lausanne).* 2025;16:1602594. doi 10.3389/fendo.2025.1602594
- Hunter R.W., Hughey C.C., Lantier L., Sundelin E.I., Pegg M., Zeqiraj E., Sicheri F., Jessen N., Wasserman D.H., Sakamoto K. Metformin reduces liver glucose production by inhibition of fructose-1-6-bisphosphatase. *Nat Med.* 2018;24(9):1395-1406. doi 10.1038/s41591-018-0159-7
- Ip W., Shao W., Chiang Y.A., Jin T. The Wnt signaling pathway effector TCF7L2 is upregulated by insulin and represses hepatic gluconeogenesis. *Am J Physiol Metab.* 2012;303(9):E1166-E1176. doi 10.1152/ajpendo.00249.2012
- Jiang L., Xiong J., Zhan J., Yuan F., Tang M., Zhang C., Cao Z., ... Wang Hui, Wang L., Wang J., Zhu W.-G., Wang Haiying. Ubiquitin-specific peptidase 7 (USP7)-mediated deubiquitination of the histone deacetylase SIRT7 regulates gluconeogenesis. *J Biol Chem.* 2017;292(32):13296-13311. doi 10.1074/jbc.M117.780130
- Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000;28(1):2730. doi 10.1093/nar/28.1.27
- Karbalaee-Hasani A., Khadive T., Eskandari M., Shahidi S., Mosavi M., Nejadbrahimi Z., Khalkhali L., Sangdari A., Mohammadi D., Soltani A., Khodabandehloo H., Hosseini H., Koushki M. Effect of metformin on circulating levels of inflammatory markers in patients with type 2 diabetes: A systematic review and meta-analysis of randomized controlled trials. *Ann Pharmacother.* 2021;55(9):1096-1109. doi 10.1177/1060028020985303
- Kristófi R., Eriksson J.W. Metformin as an anti-inflammatory agent: a short review. *J Endocrinol.* 2021;251(2):R11-R22. doi 10.1530/JOE-21-0194
- LaMoia T.E., Shulman G.I. Cellular and molecular mechanisms of metformin action. *Endocr Rev.* 2021;42(1):77-96. doi 10.1210/endo/rev/bnaa023
- Li J., Yan H., Xiang R., Yang W., Ye J., Yin R., Yang J., Chi Y. ATP secretion and metabolism in regulating pancreatic beta cell functions and hepatic glycolipid metabolism. *Front Physiol.* 2022;13:918042. doi 10.3389/fphys.2022.918042
- Liu G., Chen H., Liu H., Zhang W., Zhou J. Emerging roles of SIRT6 in human diseases and its modulators. *Med Res Rev.* 2021;41(2):1089-1137. doi 10.1002/med.21753
- Loan A., Syal C., Lui M., He L., Wang J. Promising use of metformin in treating neurological disorders: biomarker-guided therapies. *Neural Regen Res.* 2024;19(5):1045-1055. doi 10.4103/1673-5374.385286
- Luizon M.R., Eckalbar W.L., Wang Y., Jones S.L., Smith R.P., Laurance M., Lin L., ... Molony C., Innocenti F., Yee S.W., Giacomini K.M., Ahituv N. Genomic characterization of metformin hepatic response. *PLoS Genet.* 2016;12(11):e1006449. doi 10.1371/journal.pgen.1006449
- Pan Q., Ai W., Guo S. TGF- β 1 signaling impairs metformin action on glycemic control. *Int J Mol Sci.* 2024;25(4):2424. doi 10.3390/ijms25042424
- Ravnskjaer K., Hogan M.F., Lackey D., Tora L., Dent S.Y.R., Olefsky J., Montminy M. Glucagon regulates gluconeogenesis through KAT2B- and WDR5-mediated epigenetic effects. *J Clin Invest.* 2013;123(10):4318-4328. doi 10.1172/JCI69035
- Rena G., Hardie D.G., Pearson E.R. The mechanisms of action of metformin. *Diabetologia.* 2017;60(9):1577-1585. doi 10.1007/s00125-017-4342-z
- Rotroff D.M., Yee S.W., Zhou K., Marvel S.W., Shah H.S., Jack J.R., Havener T.M., ... Giacomini K.M., Pearson E.R., Wagner M.J., Buse J.B., Motsinger-Reif A.A. Genetic variants in *CPA6* and *PRPF31* are associated with variation in response to metformin in individuals with type 2 diabetes. *Diabetes.* 2018;67(7):1428-1440. doi 10.2337/db17-1164
- Rykova E., Ershov N., Damarov I., Merkulova T. SNPs in 3'UTR mRNA target sequences associated with individual drug susceptibility. *Int J Mol Sci.* 2022;23(22):13725. doi 10.3390/ijms232213725
- Smiles W.J., Ovens A.J., Kemp B.E., Galic S., Petersen J., Oakhill J.S. New developments in AMPK and mTORC1 cross-talk. *Essays Biochem.* 2024;68(3):321-336. doi 10.1042/EBC20240007
- St-Amour S., Tessier L., Harnois J., Allard C., Lavoie A., Caron P., Bouchard L., Perron P., Tremblay K. PCK1 and SLC22A2 gene variants associated with response to metformin treatment in type 2 diabetes. *PLoS One.* 2025;20(2):e0305511. doi 10.1371/journal.pone.0305511

- Strang J., Astridge D., Nguyen V., Reigan P. Small molecule modulators of AMP-Activated Protein Kinase (AMPK) activity and their potential in cancer therapy. *J Med Chem.* 2025;68(3):2238-2254. doi 10.1021/acs.jmedchem.4c02354
- Szklarczyk D., Kirsch R., Koutrouli M., Nastou K., Mehryary F., Hachilif R., Gable A.L., Fang T., Doncheva N.T., Pyysalo S., Bork P., Jensen L.J., von Mering C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51(D1): D638-D646. doi 10.1093/nar/gkac1000
- Vohra M., Sharma A.R., Mallya S., Prabhu N.B., Jayaram P., Nagri S.K., Umakanth S., Rai P.S. Implications of genetic variations, differential gene expression, and allele-specific expression on metformin response in drug-naïve type 2 diabetes. *J Endocrinol Invest.* 2022;46(6):1205-1218. doi 10.1007/s40618-022-01989-y
- Wang Y., Liu T., Cai Y., Liu W., Guo J. SIRT6's function in controlling the metabolism of lipids and glucose in diabetic nephropathy. *Front Endocrinol (Lausanne).* 2023;14:1244705. doi 10.3389/fendo.2023.1244705
- Wang Y., Wang X., Du C., Wang Z., Wang J., Zhou N., Wang B., Tan K., Fan Y., Cao P. Glycolysis and beyond in glucose metabolism: exploring pulmonary fibrosis at the metabolic crossroads. *Front Endocrinol (Lausanne).* 2024;15:1379521. doi 10.3389/fendo.2024.1379521
- Xiao D., Guo Y., Li X., Yin J.-Y., Zheng W., Qiu X.-W., Xiao L., Liu R.-R., Wang S.-Y., Gong W.-J., Zhou H.-H., Liu Z.-Q. The impacts of *SLC22A1* rs594709 and *SLC47A1* rs2289669 polymorphisms on metformin therapeutic efficacy in chinese type 2 diabetes patients. *Int J Endocrinol.* 2016;2016:4350712. doi 10.1155/2016/4350712
- Yang Q., Huo E., Cai Y., Zhang Z., Dong C., Asara J.M., Wei Q. PFKFB3-mediated glycolysis boosts fibroblast activation and subsequent kidney fibrosis. *Cells.* 2023;12(16):2081. doi 10.3390/cells12162081
- Yi Q.-Y., Deng G., Chen N., Bai Z.-S., Yuan J.-S., Wu G.-H., Wang Y.-W., Wu S.-J. Metformin inhibits development of diabetic retinopathy through inducing alternative splicing of VEGF-A. *Am J Transl Res.* 2016;8(9):3947-3954
- Yu G., Wang L.-G., Han Y., He Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omi A J Integr Biol.* 2012;16(5):284-287. doi 10.1089/omi.2011.0118
- Zhang X., Yang S., Chen J., Su Z. Unraveling the regulation of hepatic gluconeogenesis. *Front Endocrinol (Lausanne).* 2019;9:802. doi 10.3389/fendo.2018.00802
- Zhou Y., Ye W., Wang Y., Jiang Z., Meng X., Xiao Q., Zhao Q., Yan J. Genetic variants of OCT1 influence glycemic response to metformin in Han Chinese patients with type-2 diabetes mellitus in Shanghai. *Int J Clin Exp Pathol.* 2015;8(8):9533-9542
- Zhu X., Li H., Wu Y., Zhou J., Yang G., Wang W. lncRNA MEG3 promotes hepatic insulin resistance by serving as a competing endogenous RNA of miR-214 to regulate ATF4 expression. *Int J Mol Med.* 2018;43(1):345-357. doi 10.3892/ijmm.2018.3975

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

Received October 17, 2025. Revised November 21, 2025. Accepted December 1, 2025.

doi 10.18699/vjgb-26-30

Features of connectivity of default mode network depending on polymorphism of serotonin transporter gene (5-HTTLPR)

A.V. Bocharov , A.N. Savostyanov , S.S. Tamozhnikov , A.E. Saprygin , D.A. Lebedkin ,
E.A. Merkulova , G.G. Knyazev 

Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia

 bocharovav@neuronm.ru

Abstract. Serotonin transporter gene polymorphism is important in the regulation of the serotonergic system that affects mood and the regulation of emotions and behavior. In this study, 128 channel electroencephalogram recordings were performed, and buccal epithelium samples were obtained from 53 volunteers (32 females). *La*, *Lg*, and *S* alleles were identified by polymerase chain reaction. The aim of the study was to investigate the connectivity of the default mode network, measured using resting state electrophysiologic data, depending on the serotonin transporter gene polymorphism. Localization of the sources of bioelectrical activity of the cerebral cortex was performed by the beamformer method. Comparisons of *LaLa* genotype carriers and *S* or *Lg* allele carriers were performed using T-contrast of connectivity indices calculated between the nodes of the default mode network and the rest of the brain. It was found that carriers of the *S* allele were characterized by increased connectivity of the default mode network with the visual association cortex and with structures forming the posterior node of the default mode network, as well as increased connectivity of the posterior node of the default mode network with the right parahippocampal gyrus, and this pattern of connectivity may predispose to the onset and/or maintenance of intrusive thoughts. Whereas carriers of the *LaLa* genotype had higher connectivity of the anterior node of the default mode network with the right ventromedial prefrontal cortex, with the medial frontal gyrus, and with the posterior cingulate cortex, which is the structure of the posterior node of the default mode network, compared to carriers of the *S* or *Lg* allele. Also, carriers of the *LaLa* genotype had higher connectivity of the posterior node of the default mode network with the cluster involving the right dorsolateral prefrontal cortex compared to carriers of the *S* or *Lg* allele. It could be hypothesized that increased connectivity of the default mode network with brain structures (i.e., dorsolateral and ventromedial prefrontal cortex) involved in cognitive regulation processes may contribute to the regulation of the processes of the default mode network associated with autobiographical memory.

Keywords: EEG; 5-HTTLPR; serotonin transporter polymorphism; default mode network; resting-state networks; connectivity

For citation: Bocharov A.V., Savostyanov A.N., Tamozhnikov S.S., Saprygin A.E., Lebedkin D.A., Merkulova E.A., Knyazev G.G. Features of connectivity of default mode network depending on polymorphism of serotonin transporter gene (5-HTTLPR). *Vavilovskii Zhurnal Genetiki i Selektzii* = *Vavilov J Genet Breed*. 2026;30(2):267-273. doi 10.18699/vjgb-26-30

Funding. This work was supported by budgetary funding for basic scientific research (theme No. 126020316369-9, 2026–2028).

Особенности коннективности дефолтной сети мозга в зависимости от полиморфизма гена транспортера серотонина (5-HTTLPR)

А.В. Бочаров , А.Н. Савостьянов , С.С. Таможников , А.Е. Сапрыгин , Д.А. Лебедин ,
Е.А. Меркулова , Г.Г. Князев 

Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия

 bocharovav@neuronm.ru

Аннотация. Полиморфизм гена транспортера серотонина имеет большое значение в регуляции серотонинергической системы, влияющей на настроение и регуляцию эмоций и поведения. В исследовании была проведена запись 128-канальной электроэнцефалограммы и взяты образцы буккального эпителия у 53 добровольцев (32 женщины). Аллели *La*, *Lg* и *S* были определены методом полимеразной цепной реакции. Целью работы было изучение особенностей коннективности дефолтной сети мозга, измеренной с помощью электрофизиологических данных состояния покоя, в зависимости от полиморфизма гена транспортера серотонина. Локализация источников биоэлектрической активности коры мозга осуществлялась методом формирователя пучка. Сравнение групп носителей *LaLa* генотипа и носителей одного из аллелей *S* или *Lg* было выполнено с помощью Т-контраста показателей

коннективности, рассчитанных между узлами дефолтной сети мозга и остальным мозгом. Обнаружено, что для носителей *S* или *Lg* аллеля характерна повышенная коннективность дефолтной сети мозга со зрительной ассоциативной корой и со структурами, образующими задний узел дефолтной сети мозга, а также повышенная коннективность заднего узла дефолтной сети мозга с правой парагиппокампальной извилиной, что может predispose к возникновению и/или поддержанию навязчивых мыслей. У носителей *LaLa* генотипа по сравнению с носителями *S* или *Lg* аллеля была выше коннективность переднего узла дефолтной сети мозга с правой вентромедиальной лобной корой, с медиальной лобной извилиной и с задней поясной корой, являющейся структурой заднего узла дефолтной сети мозга. Также у носителей *LaLa* генотипа по сравнению с носителями *S* или *Lg* аллеля была выше коннективность заднего узла дефолтной сети мозга с кластером, охватывающим правую дорсолатеральную префронтальную кору. Можно предположить, что увеличение коннективности дефолтной сети мозга со структурами мозга (дорсолатеральная и вентромедиальная префронтальная кора), участвующими в процессах когнитивной регуляции, способствует регуляции процессов дефолтной сети мозга, связанных с автобиографическими воспоминаниями.

Ключевые слова: ЭЭГ; 5-HTTLPR; полиморфизм гена транспортера серотонина; дефолтная сеть мозга; дефолт-система мозга; коннективность; сети покоя

Introduction

A large number of studies using functional magnetic resonance imaging (fMRI) have found a characteristic pattern of activation of brain structures forming a network at rest. This network was named the default mode network (DMN), and among the structures that comprise it are the medial prefrontal cortex, posterior cingulate cortex, precuneus, medial, lateral, and inferior parietal lobes (Raichle et al., 2001). It has been repeatedly shown that the DMN exhibits a sustained decrease in activity/connectivity during most externally oriented tasks requiring concentration on external stimuli. However, in the resting state, when attention is not focused on the external world, an increase in activity/connectivity has been observed in DMN structures (Raichle, 2015). Further studies have greatly expanded the understanding of DMN functions. For example, patterns of increased DMN activity/connectivity have been identified during self-referential information processing and retrieval of autobiographical memories, when reflecting on self and relationships with loved ones, when planning for the future, and when experiencing internal emotional states (Raichle et al., 2001; Knyazev, 2013; Raichle, 2015; Buckner, DiNicola, 2019; Hutchinson-Wong et al., 2024; Si et al., 2025). In addition, DMN dysfunction has been found to be associated with a number of psychiatric disorders (Hutchinson-Wong et al., 2024). Thus, in people with depression symptoms and in patients with major depressive disorder (MDD), there is an increase in DMN activity/connectivity, which may be associated with disturbances in the processes of self-reflection and emotional regulation (Berman et al., 2011; Hamilton et al., 2011, 2013; Knyazev et al., 2016).

The attention of researchers is attracted to the study of serotonin transporter gene polymorphism, which is of great importance in the regulation of the serotonergic system that affects mood, emotional regulation and behavior (Cools et al., 2008). Serotonin transporter (5-HTT) performs an important function – serotonin reuptake, thereby controlling the level of serotonin in the synaptic cleft and determining the intensity and duration of its action. The 5-HTT gene polymorphism has been extensively studied and represents a variation in the promoter region of the gene. This polymorphism has two major alleles: long (*L*) and short (*S*). The difference in promoter region length affects the transcription efficiency of the gene, i. e. the amount of transporter protein synthesized. The

S allele is associated with reduced expression of the 5-HTT gene, resulting in less serotonin transporter in the synapse (Lesch et al., 1996). The long *L* allele contains *a/g* single nucleotide polymorphism. According to (Hu et al., 2006), the *Lg* allele is functionally similar to the *S* allele. The *La/La* genotype is classified as *LL* (high level of transcriptional efficiency), the *La/S* and *La/Lg* genotypes are classified as *LS* (intermediate level of transcriptional efficiency), and the *Lg/S*, *Lg/Lg* and *S/S* genotypes are classified as *SS* (low level of transcriptional efficiency) (Hu et al., 2006).

Studies have found an association between 5-HTT gene polymorphisms and susceptibility to various psychiatric disorders including MDD, post-traumatic stress disorder (PTSD), and suicidal behavior. It was found that individuals with at least one *S* allele showed increased sensitivity to stressful situations and negative emotions and were more likely to experience symptoms of MDD (Caspi et al., 2003) and PTSD (Koenen et al., 2009) in the presence of adverse factors. It should be noted that other studies have not confirmed associations between psychiatric disorders and the 5-HTT gene polymorphism (Risch et al., 2009). Also, in some studies, it was shown that carriers of the *S* allele are more prone to impulsive behavior (Walderhaug et al., 2010), have a higher expression of the personality trait neuroticism, the opposite pole of which is emotional stability (Greenberg et al., 2000), and also showed a reduced ability to cognitive control in the Stop Signal task (Landro et al., 2015).

fMRI study revealed that carriers of the *S* allele in response to negative stimuli showed increased activation of the amygdala, a brain structure responsible for detecting threats and experiencing emotions of fear and anxiety (Adolphs, 2008). At the same time, carriers of the *S* allele showed decreased activity in the ventromedial prefrontal cortex, a brain structure responsible for emotion regulation processes (Rao et al., 2007). In another study involving 23 women, the volume of the lateral prefrontal cortex, a brain structure involved in emotion regulation and cognitive control, was negatively associated with attention to emotional stimuli (both negative and positive) only in carriers of the *S* allele, but no polymorphism was associated with the volume of the amygdala and medial prefrontal cortex (Beevers et al., 2010). In a study of children and adolescents using the functional connectivity method, lower connectivity of the posterior node of the DMN with the superior medial

prefrontal cortex was found in *SS* genotype carriers compared to *L* allele carriers (Wiggins et al., 2012). In a study by A. Meyer-Lindenberg (2009), it was found that indicators of functional connectivity in the brain are more reliable predictors of the effect of 5-HTT gene polymorphism than indicators of activity of individual brain regions (Meyer-Lindenberg, 2009). An fMRI study by J. Wiggins and colleagues (2012) examined the severity of connectivity between the anterior and posterior DMN nodes in children and adolescents and its association with age in groups dependent on serotonin transporter gene polymorphism (Wiggins et al., 2012).

In this study, we will not limit ourselves to examining connectivity between DMN nodes, but will analyze DMN connections within the cortex in the resting state depending on the serotonin transporter gene polymorphism in adult study participants. The aim of the study was to examine the connectivity features of the DMN measured from EEG data while at rest in relation to the serotonin transporter gene polymorphism.

Materials and methods

Study participants. This study included 53 right-handed volunteers (32 females and 21 males) with normal or corrected-to-normal vision (mean age = 28 years, standard deviation = 10). Prior to the study, participants completed a questionnaire in which they answered questions about their mental and physical health, general well-being, and psychoactive substance use prior to the study. Women additionally answered questions about their current menstrual cycle phase and hormonal contraceptive use. Exclusion criteria were the presence of mental illness, head injuries, and use of narcotic, psychoactive, and hormonal substances. Participants were informed about the study methods and then signed an informed consent to participate in the experiment. The study was organized in accordance with the ethical standards established by the Declaration of Helsinki and was approved by the local bioethical committee of Scientific Research Institute of Neurosciences and Medicine (SRINM, Novosibirsk).

Genotyping. DNA was isolated from buccal epithelial cells using the Biosilica isolation kit (Russia). The *La*, *Lg*, and *S* alleles of the 5-HTT polymorphism were determined in DNA samples by polymerase chain reaction with primers: 5'-gagggactgagctggacaaccac-3' and 5'-gcgttgccgctctgaattgc-3' (Lesch et al., 1996). The resulting products were separated by agarose gel electrophoresis. The sizes of the *L* and *S* alleles for the 5-HTT gene were 529 and 489 bp. To detect the *La* and *Lg* alleles, hydrolysis of the amplification products was carried out for three hours by MspI endonuclease. Thereafter, the product sizes were 340, 127, and 62 bp for the *La* allele and 174, 166, 127, and 62 bp for the *Lg* allele. The sample included 17 *LL*, 28 *LS*, and 8 *SS* carriers.

EEG recording. For EEG recording, a Brain Products (Germany) multichannel biopotential amplifier was used, which was equipped with a cap containing 128 electrodes; one of the electrodes was used to record the oculogram. The electrodes were arranged according to the international 10–5 system. The bandwidth was between 0.1 and 100 Hz and the sampling rate was 1,000 Hz. Electrode Cz was used as a reference.

Data analysis. Artefacts were manually removed using independent component analysis (ICA) under visual control. Consistently with previous studies of oscillatory resting state networks (Knyazev et al., 2016; Bocharov et al., 2021, 2022), analysis was performed in the delta frequency range. After removing artefacts, EEG data were filtered in the delta frequency range (1–4 Hz) using the Butterworth filter and the filtfilt function. The boundary element model was used as the head model (Fuchs et al., 2001). The cortical grid used in the study contained 5,124 vertices and was created using a template based on the Montreal Neurological Institute (MNI) brain model. Localization of cortical sources of bioelectrical activity was performed using the beamformer method (Van Veen et al., 1997) with SPM-12 (<https://www.fil.ion.ucl.ac.uk/spm/>). Covariance matrices for the analysis were calculated on resting state EEG data recordings of 5 minutes. The orthogonalization method was used to correct for signal leakage that may arise due to insufficient spatial resolution of the source localization method (Brookes et al., 2011; Hipp et al., 2012).

After orthogonalization, a Hilbert transform was applied and the signal envelope was extracted, which is a curve of signal amplitudes over time. Connectivity maps were calculated between the time series of the selected regions of interest (ROIs) and the time series of the rest of the brain voxels. The selection of ROIs was based on data from previous studies. The medial prefrontal cortex (–1, 49, –2), posterior cingulate cortex (–5, –53, 41), and left (–45, –71, 35) and right (45, –71, 35) lateral parietal lobes were selected for the DMN (Gusnard, Raichle, 2001).

A Fisher transform was applied to the Pearson correlation coefficients between the temporal activity of ROIs and the rest of the brain. The connectivity maps were spatially smoothed (FWHM 8 mm) and converted to NIFTI format. Second-level statistical analysis was performed using T-contrast to detect differences between groups (carriers of the *S* allele vs carriers of the *LL* genotype). To assess the statistical significance of the detected effects, a double threshold was used: at the voxel level ($p < 0.005$) and at the cluster level (cluster size greater than 100 voxels).

Results

Statistical analysis (T-contrast) of DMN connectivity in two groups (*S* allele carriers vs *LL* genotype carriers) was performed in the SPM 12 program.

The Figure A shows the results comparing DMN connectivity in *S* allele carriers and *LL* genotype carriers. So, it was found that *S* allele carriers had greater DMN connectivity in the left hemisphere with precuneus (BA (Brodmann Area) 19, $x = -17, y = -82, z = 41$, cluster size = 166, $T = 3.10, p = 0.001$) and with a cluster in the left hemisphere encompassing the posterior cingulate cortex (BA 30, $x = -29, y = -70, z = 15$, cluster size = 943, $T = 3.08, p = 0.001$), cuneus (BA 30, $x = -25, y = -72, z = 7$, cluster size = 943, $T = 3.32, p < 0.001$), and middle occipital gyrus (BA 19, $x = -29, y = -70, z = 15$, cluster size = 943, $T = 3.05, p = 0.001$), and with a cluster in the right hemisphere encompassing the posterior cingulate cortex (BA 30, $x = 17, y = -54, z = 9$, cluster size = 341,

$T = 3.19, p = 0.001$) and the parahippocampal gyrus (BA 19, $x = 17, y = -46, z = -3$, cluster size = 341, $T = 3.03, p = 0.001$) (see the Figure A).

The Figure B shows the results of the T-contrast of DMN connectivity in carriers of the *LL* genotype more than in carriers of the *S* allele. *LL* genotype carriers had more DMN connectivity with the medial frontal gyrus (BA 10, $x = -11, y = 40, z = -7$, cluster size = 251, $T = 2.97, p = 0.002$), with the superior frontal gyrus (BA 11, $x = 9, y = 62, z = -21$, cluster size = 110, $T = 2.76, p = 0.003$) and with a cluster located in the right dorsolateral prefrontal cortex (BA 10, $x = 51, y = 54, z = 5$, cluster size = 473, $T = 3.06, p = 0.001$) (see the Figure B).

According to (Raichle, 2015), anterior and posterior DMN nodes may have different functional significance. To understand the contribution of the anterior and posterior DMN nodes to the identified effects, we performed comparisons for the anterior and posterior DMN nodes separately in *LL* genotype carriers and *S* allele carriers.

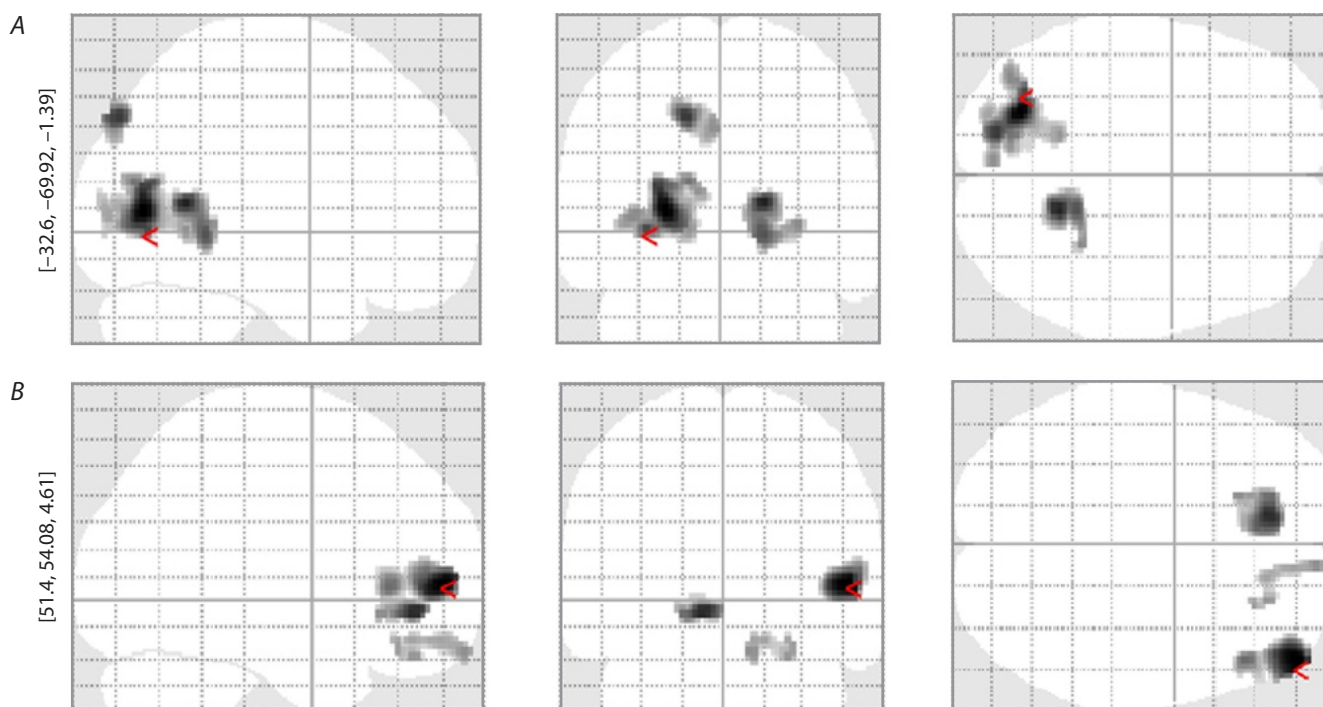
Comparison of anterior DMN connectivity in *S* allele carriers, which was greater than in *LL* genotype carriers, showed no significant effects. Whereas *LL* genotype carriers had higher anterior node DMN connectivity with posterior cingulate cortex compared to *S* allele carriers (BA 23, $x = 2, y = -44, z = 29$, cluster size = 1,800, $T = 3.77, p < 0.001$), with the right superior frontal gyrus (BA 10, $x = 23, y = 60, z = -3$, cluster size = 581, $T = 3.08, p = 0.001$) and the medial frontal gyrus (BA 6, $x = -11, y = -12, z = 57$, cluster size = 292, $T = 3.05, p = 0.001$).

S allele carriers compared to *LL* genotype carriers had greater connectivity of the posterior DMN node with a cluster in the right hemisphere encompassing the posterior cingulate cortex (BA 30, $x = 17, y = -56, z = 9$, cluster size = 249, $T = 2.95, p = 0.002$) and the parahippocampal gyrus (BA 19, $x = 19, y = -46, z = -5$, cluster size = 249, $T = 2.79, p = 0.002$), and with a cluster in the left hemisphere encompassing the posterior cingulate cortex (BA 30, $x = -29, y = -70, z = 15$, cluster size = 782, $T = 3.08, p = 0.001$) and cuneus (BA 17, $x = -11, y = -76, z = 13$, cluster size = 782, $T = 3.10, p = 0.002$).

LL genotype carriers compared to *S* allele carriers had greater connectivity of the posterior DMN node with a cluster of voxels spanning the right dorsolateral prefrontal cortex (BA 46, $x = 43, y = 30, z = 17$, cluster size = 234, $T = 2.77, p = 0.003$).

Discussion

In our study, the pattern of DMN connectivity with the precuneus, posterior cingulate cortex, parietal and occipital cortex was revealed in carriers of the *S* allele as compared to representatives of the *LL* genotype. The precuneus and posterior cingulate cortex are known to be part of the neural substrate of self-consciousness and form the posterior node of the DMN (Raichle et al., 2001; Utevsky et al., 2014). It has been repeatedly shown that the DMN is associated with internally directed thought processes and increases its activity during self-referential processing, autobiographical memories,



Results of statistical analysis (T-contrast) of the connectivity of the DMN between the group of carriers of the *S* allele and the group of carriers of the *LL* genotype.

A – T-contrast DMN connectivity in *S* allele carriers is greater than DMN connectivity in *LL* genotype carriers; B – T-contrast DMN connectivity in *LL* genotype carriers is greater than DMN connectivity in *S* allele carriers.

thinking about the future, and making plans (Knyazev, 2013; Menon, D'Esposito, 2022). According to (Raichle, 2015), there is a functional separation between the anterior and posterior DMN nodes and the posterior DMN node is predominantly associated with processes of remembering past events (Raichle, 2015). It has also been shown in studies that joint activation of posterior cingulate cortex and precuneus has been observed during memory retrieval and reliving of traumatic events in patients with PTSD (Ramage et al., 2013; Thome et al., 2020). And under conditions of emotion regulation compared to passive observation of emotional stimuli, reduced activation of posterior cingulate cortex and precuneus was observed in patients with PTSD and in a group of healthy subjects (Nicholson et al., 2022).

Interestingly, in addition to increased DMN connectivity in the precuneus and posterior cingulate cortex, *S* allele carriers were also characterized by increased connectivity to the parietal and occipital cortex, which are part of the visual association cortex. In studies, the thickness of the visual associative cortex has been found to correlate with the severity of depression and anxiety (Peterson et al., 2009). It has also been shown that the visual associative cortex is involved in rumination processes (Cooney et al., 2010).

In carriers of the *S* allele compared to carriers of the *LL* genotype, an increase in DMN connectivity with the right parahippocampal gyrus, which is part of the temporal cortex, was found. The DMN is known to be associated with introspective thinking activity, while the temporal cortex and parahippocampal gyrus are involved in memory retrieval processes (Buckner, DiNicola, 2019). In a previous study, we found a positive association of depression symptom severity with the pattern of DMN connectivity with the right temporal cortex (Bocharov et al., 2021). Also, in a subsequent study, we found that an increase in this connectivity pattern correlated with the severity of intrusive thoughts: rumination and preference for a non-adaptive emotion regulation strategy – thought suppression (Bocharov et al., 2022). It should be noted that in the study (Rassin, 2003), it was found that when using such a strategy, the desire to suppress intrusive thoughts paradoxically leads to their strengthening (Rassin, 2003). It is known that carriers of the *S* allele are more prone to develop MDD and have difficulties in regulating their emotional state (Caspi et al., 2003; Walderhaug et al., 2010). It can be assumed that the identified pattern of higher DMN connectivity with the posterior node of the DMN, posterior associative cortex and right temporal cortex (parahippocampal gyrus), which is characteristic of *S* allele carriers, may be associated with and/or predispose to the occurrence of intrusive thoughts and may underlie the predisposition to the emergence and maintenance of symptoms of depression.

In contrast to *S* allele carriers, who showed increased connectivity of the posterior node of the DMN, *LL* genotype carriers had higher DMN connectivity in the anterior node: with the medial frontal gyrus and with the superior frontal gyrus, which is part of the ventromedial frontal cortex. According to (Raichle, 2015), the structures of the anterior node of the DMN may have different functional importance, so the dorsomedial prefrontal cortex is involved in self-referential processes, while

the ventromedial part of the frontal cortex is involved in the processes of social behavior, motivation, mood and emotional processing. A study (Rao et al., 2007) found reduced activation of the ventromedial frontal cortex and increased activation of the amygdala in response to negative stimuli in carriers of the *S* allele compared to carriers of the *LL* genotype. Dysfunction of the ventromedial frontal cortex is thought to play a key role in the pathogenesis of affective and anxiety disorders. It is assumed that insufficient activity of the ventromedial frontal cortex may lead to impaired inhibition of the activity of the amygdala and, as a consequence, to pathologically increased levels of negative affect (Motzkin et al., 2015).

Notably, *LL* genotype carriers were characterized by increased connectivity of the anterior DMN node with the posterior cingulate cortex, which is part of the posterior DMN node, i. e. connectivity between the anterior and posterior DMN nodes was increased. To some extent, this result is consistent with the study (Cha et al., 2018). In it, they found that carriers of genotypes with low transcriptional efficiency had reduced connectivity between the superior frontal gyrus, which is part of the anterior DMN node, and the posterior DMN node. Reduced connectivity between these structures mediated increased impulsivity in *SS* genotype carriers (Cha et al., 2018).

In addition, our study revealed that carriers of the *LL* genotype were characterized by increased connectivity of the posterior node of the DMN, involved in autobiographical memory processes, with the right dorsolateral prefrontal cortex, responsible for cognitive control and regulatory functions (Sculth et al., 2017). Thus, it has been shown in studies that the right dorsolateral prefrontal cortex is often used as a target area for the application of transcranial magnetic stimulation (TMS) in the treatment of PTSD, obsessive-compulsive disorder, and addictions. In a study conducted by B.D. Greenberg and U. Ziemann (1998), it was shown that exposure with TMS to an area of the right dorsolateral prefrontal cortex in patients with obsessive-compulsive disorder promoted a reduction in compulsive symptoms and improved mood. H. Kober and colleagues (2010) found that in smokers, decreased cravings were associated with increased activity of the dorsolateral prefrontal cortex (Kober et al., 2010). A study conducted by H. Cohen and colleagues (2004) demonstrated the effectiveness of TMS in reducing PTSD symptoms and the importance of this brain region in the regulation of stress and emotional responses (Cohen et al., 2004). According to a study (Wu et al., 2020), stimulation of the right dorsolateral prefrontal cortex of the brain can improve a person's self-regulatory abilities, especially in the context of managing one's desires and urges, as well as in the regulation of negative emotions (Wu et al., 2020). It is conceivable that in carriers of the *LL* genotype, who are known to be more resistant to stress and the development of depression (Caspi et al., 2003; Koenen et al., 2009) and the manifestation of impulsive behavior (Walderhaug et al., 2010), increased DMN connectivity with brain structures involved in cognitive regulation processes (dorsolateral and ventromedial prefrontal cortex) may contribute to more effective regulation of DMN processes related to autobiographical memories.

Conclusion

The revealed features of DMN connectivity in the resting state in carriers of *LL* genotypes and carriers of the *S* allele contribute to the understanding of the links between serotonin transporter gene polymorphism and predisposition to affective disorders. Carriers of the *S* allele compared to carriers of the *LL* genotype showed a pattern of higher connectivity of the posterior DMN, involved in autobiographical memory processes, with the precuneus, posterior cingulate and visual association cortex, and the right parahippocampal gyrus, which may predispose to the emergence and maintenance of obsessive thoughts. Whereas *LL* genotype carriers had higher DMN connectivity in the anterior node of the DMN: with the medial frontal gyrus, ventromedial frontal cortex, and posterior cingulate cortex, which is part of the posterior node of the DMN. Also, carriers of the *LL* genotype had higher resting-state connectivity of the posterior DMN with the cluster encompassing the right dorsolateral cortex compared to carriers of the *S* allele. It can be assumed that increased DMN connectivity with brain structures involved in cognitive regulation processes (dorsolateral and ventromedial prefrontal cortex) may contribute to the regulation of DMN processes associated with autobiographical memories.

References

- Adolphs R. Fear, faces, and the human amygdala. *Curr Opin Neurobiol.* 2008;18:166-172. doi 10.1016/j.conb.2008.06.006
- Beevers C.G., Pacheco J., Clasen P., McGeary J.E., Schnyer D. Prefrontal morphology, 5-HTTLPR polymorphism and biased attention for emotional stimuli. *Genes Brain Behav.* 2010;9(2):224-233. doi 10.1111/j.1601-183X.2009.00550.x
- Berman M.G., Peltier S., Nee D.E., Kross E., Deldin P.J., Jonides J. Depression, rumination and the default network. *Soc Cogn Affect Neurosci.* 2011;6:548-555. doi 10.1093/scan/nsq080
- Bocharov A.V., Knyazev G.G., Savostyanov A.N., Saprygin A.E., Proshina E.A., Tamozhnikov S.S. Relationship of depression, anxiety and rumination scores with EEG connectivity of resting state networks. *Hum Physiol.* 2021;47(2):123-127. doi 10.1134/S0362119721010023
- Bocharov A.V., Savostyanov A.N., Tamozhnikov S.S., Proshina E.A., Knyazev G.G. The association between emotion regulation strategy and oscillation balance of resting state networks. *Hum Physiol.* 2022;48(1):30-36. doi 10.1134/S0362119722010030
- Brookes M.J., Woolric M., Luckhoo H., Price D., Hale J.R., Stephenson M.C., Barnes G.R., Smith S.M., Morris P.G. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proc Natl Acad Sci USA.* 2011;108(40):16783-16788. doi 10.1073/pnas.111268510
- Buckner R.L., DiNicola L.M. The brain's default network: updated anatomy, physiology and evolving insights. *Nat Rev Neurosci.* 2019;20(10):593-608. doi 10.1038/s41583-019-0212-7
- Caspi A., Sugden K., Moffitt T.E., Taylor A., Craig I.W., Harrington H., McClay J., Mill J., Martin J., Braithwaite A., Poulton R. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science.* 2003;301(5631):386-389. doi 10.1126/science.10839
- Cha J., Guffanti G., Gingrich J., Talat A., Wickramaratne P., Weissman M., Posner J. Effects of serotonin transporter gene variation on impulsivity mediated by default mode network: a family study of depression. *Cerebral Cortex.* 2018;28(6):1911-1921. doi 10.1093/cercor/bhx097
- Cohen H., Kaplan Z., Kotler M., Kouperman I., Moisa R., Grisaru N. Repetitive transcranial magnetic stimulation of the right dorsolateral prefrontal cortex in posttraumatic stress disorder: a double-blind, placebo-controlled study. *Am J Psychiatry.* 2004;161(3):515-524. doi 10.1176/appi.ajp.161.3.515
- Cools R., Roberts A.C., Robbins T.W. Serotonergic regulation of emotional and behavioural control processes. *Trends Cogn Sci.* 2008;12:31-40. doi 10.1016/j.tics.2007.10.011
- Cooney R.E., Joormann J., Eugène F., Dennis E.L., Gotlib I.H. Neural correlates of rumination in depression. *Cogn Affect Behav Neurosci.* 2010;10(4):470-478. doi 10.3758/CABN.10.4.470
- Fuchs M., Wagner M., Kastner J. Boundary element method volume conductor models for EEG source reconstruction. *Clin Neurophysiol.* 2001;112:1400-1407. doi 10.1016/S1388-2457(01)00589-2
- Greenberg B.D., Ziemann U. Decreased neuronal inhibition in cerebral cortex in obsessive-compulsive disorder on transcranial magnetic stimulation. *Lancet.* 1998;352(9131):881-882. doi 10.1016/S0140-6736(05)60009-8
- Greenberg B.D., Li Q., Lucas F.R., Hu S., Sirota L.A., Benjamin J., Murphy D.L. Association between the serotonin transporter promoter polymorphism and personality traits in a primarily female population sample. *Am J Med Genet.* 2000;96:202-216. doi 10.1002/(SICI)1096-8628(20000403)96:2<202::AID-AJMG16>3.0.CO;2-J
- Gusnard D.A., Raichle M.E. Searching for a baseline: functional imaging and the resting human brain. *Nat Rev Neurosci.* 2001;2(10):685-694. doi 10.1038/35094500
- Hamilton J.P., Furman D.J., Chang C., Thomason M.E., Dennis E., Gotlib I.H. Default-mode and task-positive network activity in major depressive disorder: implications for adaptive and maladaptive rumination. *Biol Psychiatry.* 2011;70(4):327-333. doi 10.1016/j.biopsych.2011.02.003
- Hamilton J.P., Chen M.C., Gotlib I.H. Neural systems approaches to understanding major depressive disorder: an intrinsic functional organization perspective. *Neurobiol Dis.* 2013;52:4-11. doi 10.1016/j.nbd.2012.01.015
- Hipp J.F., Hawellek D.J., Corbetta M., Siegel M., Engel A.K. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nat Neurosci.* 2012;15(6):884-890. doi 10.1038/nn.3101
- Hu X.Z., Lipsky R.H., Zhu G., Akhtar L.A., Taubman J., Greenberg B.D., Xu K., Arnold P.D., Richter M.A., Kennedy J.L., Murphy D.L., Goldman D. Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *Am J Hum Genet.* 2006;78:815-826. doi 10.1086/503850
- Hutchinson-Wong N., Glue P., Adhia D., de Ridder D. How does depressive cognition develop? A state-dependent network model of predictive processing. *Psychol Rev.* 2025;132(2):442-469. doi 10.1037/rev0000512. Epub 2024 Nov 14. PMID: 39541510.
- Knyazev G.G. EEG correlates of self-referential processing. *Front Hum Neurosci.* 2013;7:264. doi 10.3389/fnhum.2013.00264
- Knyazev G.G., Savostyanov A.N., Bocharov A.V., Tamozhnikov S.S., Saprygin A.E. Task-positive and task-negative networks and their relation to depression: EEG beamformer analysis. *Behav Brain Res.* 2016;306:160-169. doi 10.1016/j.bbr.2016.03.033
- Kober H., Kross E.F., Mischel W., Hart C.L., Ochsner K.N. Regulation of craving by cognitive strategies in cigarette smokers. *Drug Alcohol Depend.* 2010;106(1):52-55. doi 10.1016/j.drugalcdep.2009.07.017
- Koenen K.C., Aiello A.E., Bakshis E., Amstadter A.B., Ruggiero K.J., Acierno R., Galea S. Modification of the association between serotonin transporter genotype and risk of posttraumatic stress disorder in adults by county-level social environment. *Am J Epidemiol.* 2009;169(6):704-711. doi 10.1093/aje/kwn397
- Landro N.I., Jonassen R., Clark L., Haug K.F., Aker M., Bo R., Berg J.P., Neumeister A., Stiles T.C. Serotonin transporter polymorphisms predict response inhibition in healthy volunteers. *Neurosci Lett.* 2015;584:109-112. doi 10.1016/j.neulet.2014.10.006
- Lesch K.P., Bengel D., Heils A., Sabol S.Z., Greenberg B.D., Petri S., Murphy D.L. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science.* 1996;274(5292):1527-1531. doi 10.1126/science.274.5292.1527

- Menon V., D'Esposito M. The role of PFC networks in cognitive control and executive function. *Neuropsychopharmacol.* 2022;47(1):90-103. doi 10.1038/s41386-021-01152-w
- Meyer-Lindenberg A. Neural connectivity as an intermediate phenotype: brain networks under genetic control. *Hum Brain Mapp.* 2009;30(7):1938-1946. doi 10.1002/hbm.20639
- Motzkin J.C., Philippi C.L., Wolf R.C., Baskaya M.K., Koenigs M. Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. *Biol Psychiatry.* 2015;77(3):276-284. doi 10.1016/j.biopsych.2014.02.014
- Nicholson A.A., Rabellino D., Densmore M., Frewen P.A., Steyerl D., Schamowski F., Theberge J., Neufeld R.W.J., Schmahl C., Jetly R., Lanius R.A. Differential mechanisms of posterior cingulate cortex downregulation and symptom decreases in posttraumatic stress disorder and healthy individuals using real-time fMRI neurofeedback. *Brain Behav.* 2022;12(1):e2441. doi 10.1002/brb3.2441
- Peterson B.S., Warner V., Bansal R., Zhu H., Hao X., Liu J., Weissman M.M. Cortical thinning in persons at increased familial risk for major depression. *Proc Natl Acad Sci USA.* 2009;106(15):6273-6278. doi 10.1073/pnas.0805311106
- Raichle M.E. The brain's default mode network. *Annu Rev Neurosci.* 2015;38:433-447. doi 10.1146/annurev-neuro-071013-014030
- Raichle M.E., MacLeod A.M., Snyder A.Z., Powers W.J., Gusnard D.A., Shulman G.L. A default mode of brain function. *Proc Natl Acad Sci USA.* 2001;98(2):676-682. doi 10.1073/pnas.98.2.676
- Ramage A.E., Laird A.R., Eickhoff S.B., Acheson A., Peterson A.L., Williamson D.E., Telch M.J., Fox P.T. A coordinate-based meta-analytic model of trauma processing in posttraumatic stress disorder. *Hum Brain Mapp.* 2013;34(12):3392-3399. doi 10.1002/hbm.22155
- Rao H.Y., Gillihan S.J., Wang J.J., Korczykowski M., Sankoorikal G.M.V., Kaercher K.A., Brodtkin E.S., Detre J.A., Farah M.J. Genetic variation in serotonin transporter alters resting brain function in healthy individuals. *Biol Psychiatry.* 2007;62(6):600-606. doi 10.1016/j.biopsych.2006.11.028
- Rassin E. The white bear suppression inventory (WBSI) focuses on failing suppression attempts. *Eur J Pers.* 2003;17(4):285-298. doi 10.1002/per.478
- Risch N., Herrell R., Lehner T., Liang K.Y., Eaves L., Hoh J., Griem A., Kovacs M., Jurg J., Merikangas K.R. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *Jama.* 2009;301(23):2462-2471. doi 10.1001/jama.2009.878
- Scult M.A., Knodt A.R., Swartz J.R., Brigidi B.D., Hariri A.R. Thinking and feeling: individual differences in habitual emotion regulation and stress-related mood are associated with prefrontal executive control. *Clin Psychol Sci.* 2017;5(1):150-157. doi 10.1177/2167702616654688
- Si Q., Tian J., Savostyanov V.A., Lebedkin D.A., Bocharov A.V., Savostyanov A.N. Comparison of brain activity indexes in the Chinese and Russian students under recognition of self- and other-related information. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2025;28(8):982-992. doi 10.18699/vjgb-24-105
- Thome J., Terpou B.A., McKinnon M.C., Lanius R.A. The neural correlates of trauma-related autobiographical memory in posttraumatic stress disorder: a meta-analysis. *Depress Anxiety.* 2020;37(4):321-345. doi 10.1002/da.22977
- Utevsy A.V., Smith D.V., Huettel S.A. Precuneus is a functional core of the default-mode network. *J Neurosci.* 2014;34(3):932-940. doi 10.1523/JNEUROSCI.4227-13.2014
- Van Veen B.D., van Drongelen W., Yuchtman M., Suzuki A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans Biomed Eng.* 1997;44:867-880. doi 10.1109/10.623056
- Walderhaug E., Herman A.I., Magnusson A., Morgan M.J., Landro N.I. The short (S) allele of the serotonin transporter polymorphism and acute tryptophan depletion both increase impulsivity in men. *Neurosci Lett.* 2010;473:208-211. doi 10.1016/j.neulet.2010.02.048
- Wiggins J.L., Bedoyan J.K., Peltier S.J., Ashinoff S., Carrasco M., Weng S.J., Welsh R.C., Martin D.M., Monk C.S. The impact of serotonin transporter (5-HTTLPR) genotype on the development of resting-state functional connectivity in children and adolescents: a preliminary report. *Neuroimage.* 2012;59(3):2760-2770. doi 10.1016/j.neuroimage.2011.10.030
- Wu L.L., Potenza M.N., Zhou N., Kober H., Shi X.H., Yip S.W., Xu J., Zhu L., Wang R., Liu G., Zhang J.T. A role for the right dorsolateral prefrontal cortex in enhancing regulation of both craving and negative emotions in internet gaming disorder: a randomized trial. *Eur Neuropsychopharmacol.* 2020;36:29-37. doi 10.1016/j.euroneuro.2020.04.003

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

Received February 21, 2025. Revised April 17, 2025. Accepted April 21, 2025.

doi 10.18699/vjgb-26-31

Phylogeography and taxonomic status of the *Formica picea* complex (Hymenoptera: Formicidae)

Z.A. Zhigulskaia ¹, S.V. Shekhovtsov ^{1, 2} , S.V. Chesnokova³, A.P. Burnasheva ⁴, A.A. Gurina ³, R.Yu. Dudko ³, T.V. Poluboyarova ², S.V. Reshetnikov³, Yu.N. Sundukov⁵, D.I. Berman ¹

¹ Institute of Biological Problems of the North of the Far Eastern Branch of the Russian Academy of Sciences, Magadan, Russia

² Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Institute of Systematics and Ecology of Animals of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

⁴ Institute of Biological Problems of Cryolithozone of the Siberian Branch of the Russian Academy of Sciences, Yakutsk, Russia

⁵ Federal Scientific Center of the East Asia Terrestrial Biodiversity, Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia

 shekhovtsov@bionet.nsc.ru

Abstract. The black bog ant *Formica picea* complex is widespread from the Atlantic to the Pacific coasts of Eurasia. This complex was earlier believed to consist of one or two species (*F. picea* and *F. candida*). However, molecular analysis suggested that it includes three cryptic species. One is *F. picea* from Europe, another, *F. candida*, is currently known exclusively from Kyrgyzstan, while the third one, temporarily designated here as *Formica* sp., inhabits the easternmost part of Eurasia from China to Kamchatka. It is unknown how *F. picea* and *Formica* sp. are distributed in Siberia and whether their ranges intersect. Here we studied a sample of this complex from Siberia using mtDNA and found that their ranges overlap. The distribution of *Formica* sp. extends from the south of West Siberia, including Altai, to China, and the Russian Far East. No phylogeographic structure was detected, suggesting their recent dispersal from a single source. *F. picea* was found as far as East Siberia, but was relatively rare. While the European and West Siberian populations were genetically closely related, the specimens from Zabaykalsky Krai differed, suggesting a putative East Siberian refugium. We also determined that ecologically *F. picea* inhabits peat bogs in lowland areas and grassy communities above the tree line in the European mountains; in Altai, it is found in mountain steppes, while in Transbaikalia, in waterlogged areas along riverbanks. *Formica* sp. thrives in dry steppes and low riverbanks, but avoids bogs. Thus, *F. picea* and *Formica* sp. differ genetically, and have different distribution ranges, as well as habitat preferences. This supports the opinion that *Formica* sp. should be recognized as a distinct species.

Key words: *Formica picea*; *Formica candida*; cryptic species; Northern Eurasia; Siberia

For citation: Zhigulskaia Z.A., Shekhovtsov S.V., Chesnokova S.V., Burnasheva A.P., Gurina A.A., Dudko R.Yu., Poluboyarova T.V., Reshetnikov S.V., Sundukov Yu.N., Berman D.I. Phylogeography and taxonomic status of the *Formica picea* complex (Hymenoptera: Formicidae). *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):274-283. doi 10.18699/vjgb-26-31

Funding. The work was funded solely through the budgets of Federal State Budgetary Scientific Institutions: the Institute of Biological Problems of the North FEB RAS (No. 122041900011-9), the Institute of Cytology and Genetics SB RAS (FWNR-2026-0031), and the Institute of Systematics and Ecology of Animals SB RAS (FWGS-2026-0008). Sanger sequencing was performed in the SB RAS Genomics Core Facility (ICBFM SB RAS, Novosibirsk, Russia).

Acknowledgements. The authors are grateful to T.A. Novgorodova for critical comments on the manuscript, and to O.V. Vaulin, T.M. Krugova, and P.V. Golyakov for providing us with the ant samples. The photo of the Kansk Steppe (Central Altai) was kindly provided by R.M. Shershnev.

Филогеография и таксономический статус видового комплекса *Formica picea* (Hymenoptera: Formicidae)

З.А. Жигульская ¹, С.В. Шеховцов ^{1, 2} , С.В. Чеснокова³, А.П. Бурнашева ⁴, А.А. Гурина ³, Р.Ю. Дудко ³, Т.В. Полубоярова ², С.В. Решетников³, Ю.Н. Сундуков⁵, Д.И. Берман ¹

¹ Институт биологических проблем Севера Дальневосточного отделения Российской академии наук, Магадан, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, Россия

⁴ Институт биологических проблем криолитозоны Сибирского отделения Российской академии наук, Якутск, Россия

⁵ Федеральный научный центр биоразнообразия наземной биоты Восточной Азии Дальневосточного отделения Российской академии наук, Владивосток, Россия

 shekhovtsov@bionet.nsc.ru

Аннотация. Комплекс *Formica picea*, или черный болотный муравей, широко распространен по всему северу Евразии, от Испании и Ирландии до побережья Тихого океана и Тибета, примерно между 35 и 66.5° северной широты. Ранее считалось, что он состоит из одного или двух видов (*F. picea* и *F. candida*). Однако молекулярный

анализ показал, что в него входят три криптических вида. Один из них – *F. picea* из Европы, другой – *F. candida*, известный исключительно из Киргизии, а третий, временно обозначенный как *Formica* sp., обитает на востоке Евразии: от Китая до Камчатки. До сих пор было неизвестно, каково распределение *F. picea* и *Formica* sp. в Сибири, пересекаются ли их ареалы. Мы изучили образцы этого комплекса из Сибири с использованием митохондриальной ДНК и обнаружили, что их ареалы перекрываются. Ареал *Formica* sp. охватывает юг Западной Сибири, включая Алтай, на восток – до Тихоокеанского побережья. Филогеографической структуры не обнаружено, что указывает на недавнее расселение из одного источника. Мы нашли мало образцов *F. picea* в Сибири, это свидетельствует, что здесь этот вид встречается редко. Самая восточная находка вида была сделана в Забайкальском крае. *F. picea* обитает на торфяных болотах в пределах равнин и травяных сообществах выше линии леса в европейских горах; на Алтае – в горных степях, а в Забайкалье – на влажных участках у речных берегов. *Formica* sp. предпочитает сухие степи и низкие берега рек, но избегает болот. Таким образом, оказалось, что *F. picea* и *Formica* sp. отличаются генетически, имеют разные ареалы и биотопические предпочтения по местообитаниям. Это свидетельствует в пользу того, что *Formica* sp. можно считать отдельным видом.

Ключевые слова: *Formica picea*; *Formica candida*; виды-двойники; Северная Евразия; Сибирь

Introduction

The *Formica picea* Nylander, 1846 species complex, i. e. black bog ant, is widespread in the Palearctic, from Spain and Ireland to the Pacific coast and Tibet, and between approximately 35 and 66.5° N latitude (Ruzsky, 1905; Pleshanov, 1966; Dlussky, 1967; Dmitrienko, Petrenko, 1976; Kupyanskaya, 1990; Berman et al., 2007; Antonov, Pleshanov, 2008; Radchenko, 2016; Seifert, 2018) (Fig. 1). For a long time, *F. picea* was considered a single species with high geographic variation (Dlussky, 1967). Bolton (1995) proposed that the name *F. candida* Smith, 1878 takes precedence over *F. picea*. Not all specialists adopted this suggestion, leading to confusing parallel usage of both names in the literature.

Seifert (2004) suggested that *F. picea* and *F. candida* are two distinct species, based on their pubescence and certain morphometric parameters. Seifert (2004) also designated a neotype of *F. candida* collected from Kyrgyzstan due to the

loss of the original type specimen. However, Zakharov et al. (2019) disagreed with Seifert's conclusions, pointing out the smooth gradient of morphological variation from Europe to the Far East.

There were several molecular studies that included members of the *F. picea*/*F. candida* complex (Goropashnaya, 2003; Goropashnaya et al., 2012; Antonov, Bukin, 2016; Chen, Zhou, 2017; Schär et al., 2018). However, none of these aimed to explore the differences between the two taxa, instead simply using one or the other name. Moreover, some authors used the mitochondrial *cytb* gene, and others, *cox1*/*COI*, so their results could not be directly compared.

A recent study on the *F. picea* complex showed that it included one more species, in addition to *F. picea* and *F. candida* (Zhigul'skaya et al., 2022). While *F. picea* turned out to be restricted to Europe, as well as to one location from Tibet, ants from all east Eurasian locations (lowland China and the Far

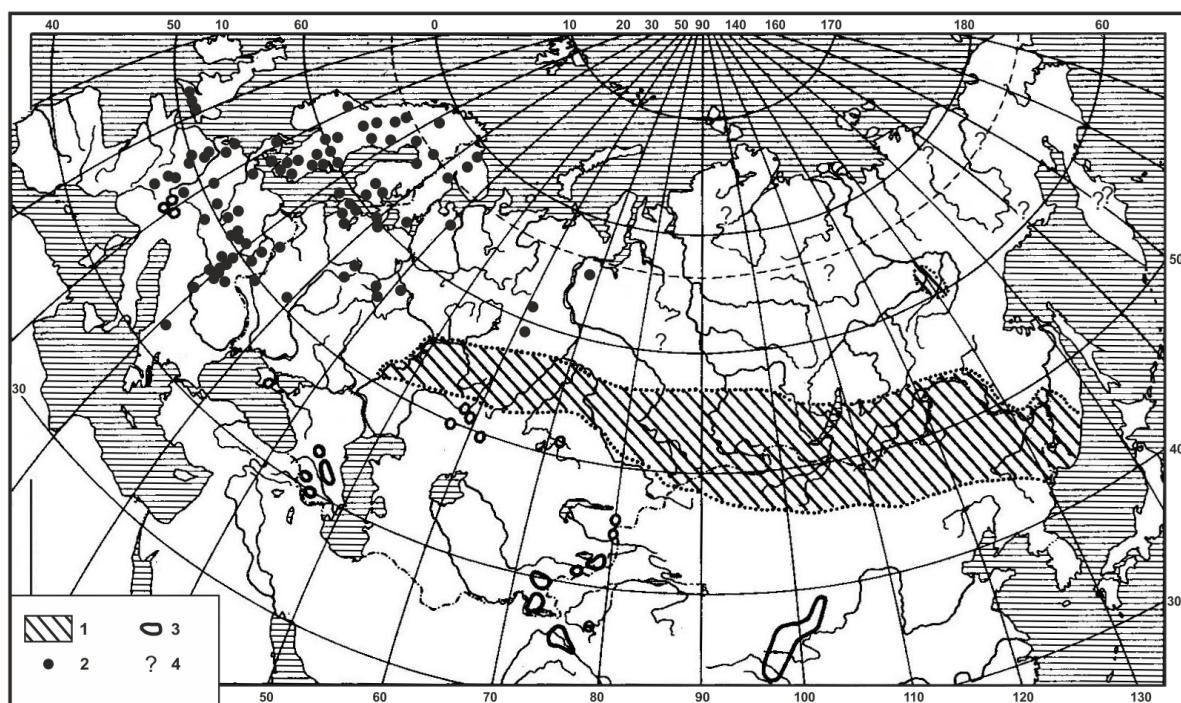


Fig. 1. Geographical distribution and ecology of the *F. picea* complex.

Characteristic habitats: 1 – diverse habitats; 2 – peat bogs; 3 – mountain meadows; 4 – no details provided (scheme from (Dlussky, 1967)).

East of Russia) were highly genetically divergent from other known *Formica* species. Geographic and genetic differences suggest that it should be regarded as a distinct species, although the formal description was not yet performed. Thus, this new taxon is hereafter referred to as *Formica* sp. There seem to be no obvious morphological distinctions for *Formica* sp. that could clearly differentiate it from *F. picea* and *F. candida*, supporting the observations of Zakharov and Dlussky (Zakharov et al., 2019) on high levels of intrapopulation and geographic variation for this complex.

Siberia currently remains a blank spot in this respect: it is unknown how far *F. picea* goes eastwards from the Urals, and how far west can *Formica* sp. be found. In this study we collected an extensive sample from different parts of Siberia in order to determine the ranges of both taxa, as well as to infer phylogeographic patterns. In order to search for putative ecological differences between *F. picea* and *Formica* sp., we also documented habitat preferences, which had been previously well described for European *F. picea* (Dlussky, 1967; Zakharov, 2015; Radchenko, 2016; Seifert, 2018; Zakharov et al., 2019).

Materials and methods

We surveyed a wide range of habitat types, from steppes to bogs, in both lowland and mountainous landscapes of Siberia and the Russian Far East (see the Table, Figs. 2, 3). Ants were collected into 96 % ethanol in a series of 6–10 specimens from nests, as well as from forage plants. The collection material is stored in the Institute of Systematics and Ecology of Animals, Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia).

Ants of the *F. picea* complex were taken for genetic analysis from 85 locations (Fig. 2): for 47, obtained by us in this study, for the rest, they were taken from our previous study (Zhigulskaya et al., 2022) and from GenBank.

DNA was extracted from individual worker ants preserved in ethanol using commercial silica columns (BioSilica, Novosibirsk, Russia), following the method outlined in (Shekhovtsov et al., 2020). We amplified a segment of the mitochondrial cytochrome c oxidase subunit 1 (*cox1*) gene with the universal primers LCO1490m (5'-TACTC-AACAA-ATCAC-AAAGA-TATTG-G-3'; modified from (Folmer et al., 1994)) and HCO2198 (5'-TAAAC-TTCAG-GGTGA-CCAAA-AAATC-A-3') (Folmer et al., 1994). Additionally, we amplified a portion of the cytochrome b (*cytb*) gene using primers Fcbl-F (5'-ACCCT-CACCT-GTAAA-TATTT-CTT-3') and Fcbl-R (5'-GGAAT-AGATC-GTAAA-ATTGC-AT-3') (Zhigulskaya et al., 2022). PCR was conducted with the Biomaster HS-Taq PCR Mix (Biolabmix, Novosibirsk, Russia).

The DNA fragments obtained were visualized through agarose gel electrophoresis. Unincorporated primers and nucleotide phosphates were eliminated using shrimp alkaline phosphatase and *Escherichia coli* exonuclease I mixture (New England Biolabs, Ipswich, MA, USA). Sanger sequencing was conducted on a 3130xl DNA Analyzer (Applied Biosystems, Framingham, MA, USA) at the SB RAS Genomics Core Facility (ICBFM SB RAS, Novosibirsk, Russia) using both forward and reverse primers. The resulting sequences were submitted to GenBank with accession numbers PQ807734–PQ807780 (*cox1*) and PQ810811–PQ810857 (*cytb*). Sequences from other researchers were also incorporated into this study (refer to the Table for GenBank accessions). This dataset includes the

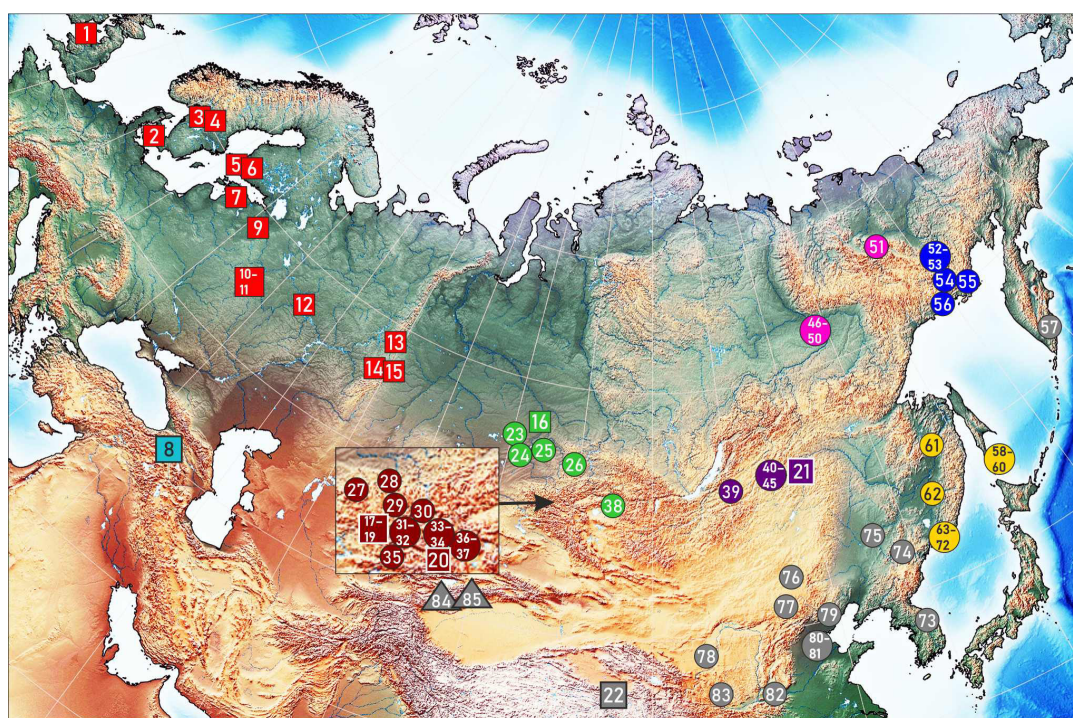


Fig. 2. *F. picea* complex specimens used in this study.

Location numbers refer to the Table. Rectangles: *F. picea*; circles: *Formica* sp.; triangles: *F. candida*.

Representatives of the *F. picea* complex with available *cox1*/*cytb* sequences

No.	Loc.	Coord.	Altitude	<i>cox1</i>	<i>cytb</i>
<i>Formica picea</i>					
1	United Kingdom: Cors Goch Nature Reserve	N53.308 E–4.253	78	–	AY786145 (as <i>F. candida</i>)
2	Denmark	N55.830 E12.560	30	LT977411	–
3	Sweden: Varmland	N59.426 E13.271	50	–	JX170886
4	Sweden: Trajmossen, Brattforsheden	N59.695 E13.889	170	–	AY786144 (as <i>F. candida</i>)
5	Finland: Nylandia, Raasepori, Harpar Stortraesket	N59.975 E23.436	4	MZ607401*	–
6	Finland: Helsinki, Vihti	N60.417 E24.317	39	–	AY786146* (as <i>F. candida</i>)
7	Estonia: Torma	N58.813 E26.738	86	–	AY786147 (as <i>F. candida</i>)
8	Armenia: Aragats Mountains	N40.432 E44.236	2728	ON220876	ON228281
9	Russia: Novgorod oblast	N58.517 E31.267	36	–	AY786148 (as <i>F. candida</i>)
10	Russia: Moscow oblast, Odintsovo district	N55.733 E36.850	184	ON220886-87	ON228278-79
11	Russia: Moscow oblast	N55.751 E37.617	148	–	AY786149 (as <i>F. candida</i>)
12	Russia: Novgorod oblast, Kerzhenskiy Natural Reserve, swamp	N56.467 E44.500	ca. 120	ON220875	ON228280
13	Russia: Perm Krai, Kachkanar	N58.700 E59.483	307	–	AY786152 (as <i>F. candida</i>)
14	Russia: Revda, Sverdlovsk oblast	N56.800 E59.917	260	–	AY786151 (as <i>F. candida</i>)
15	Russia: Yekaterinburg	N56.833 E60.583	369	–	AY786150 (as <i>F. candida</i>)
16	Russia: Tomsk oblast, Melnikovo village, swamp	N56.550 E84.067	90	PQ810826	PQ807749
17	Russia: Altai Republic, Abay village, swamp	N50.419 E85.066	1068	PQ810820	PQ807743
18	Russia: Altai Republic, Amur village, residual hills, stony steppe	N50.402 E85.107	1100	PQ810821	PQ807744
19	Russia: Altai Republic, Amur village, residual hills, stony steppe	N50.402 E85.107	1100	PQ810835	PQ807758
20	Russia: Altai Republic, Kosh-Agach village, steppe	N49.992 E88.664	2275	PQ810842	PQ807765
21	Russia: Zabaikalsky Krai, Gazimursky Zavod village, Kotikha River floodplain, swamp with rare trees	N51.545 E118.342	770	PQ810811	PQ807734
22	China: Quinghai, Tibet	N35.000 E96.000	ca. 4500	–	AY786157 (as <i>F. candida</i>)
<i>Formica</i> sp.					
23	Russia: Novosibirsk oblast, Kabinetnoye village, halophytic meadow	N55.117 E81.233	140	PQ810851	PQ807774
24	Russia: Novosibirsk oblast, Ordynsky district, steppe meadow	N54.320 E81.850	116	PQ810852	PQ807775
25	Russia: Novosibirsk oblast, Iskitim district, steppe	N54.450 E83.470	230	ON220885	–
26	Russia: Republic of Khakassia, Lake Shira, steppe	N54.534 E90.199	350	PQ810857	PQ807780
27	Russia: Altai Krai, Tigirek village, meadow with shrubs	N51.146 E83.010	500	PQ810831	PQ807754
28	Russia: Altai Republic, Ilyinka village, meadow by the river	N51.517 E85.033	690	PQ810843	PQ807766
29	Russia: Altai Republic, Yabogan village, steppe	N50.940 E85.021	1090	PQ810834	PQ807757
30	Russia: Altai Republic, Ongudai village, stony steppe	N50.733 E86.133	810	ON220880	ON228273
31	Russia: Altai Republic, Abay village, steppe near the swamp	N50.419 E85.066	1070	PQ810830	PQ807753
32	Russia: Altai Republic, Ust-Koksa village, stony steppe	N50.270 E85.615	1080	PQ810841	PQ807764
33	Russia: Altai Republic, Kurai village, steppe	N50.224 E87.932	1500	PQ810822	PQ807745
34	Russia: Altai Republic, Kurai village, steppe	N50.224 E87.932	1500	PQ810823	PQ807746
35	Russia: Altai Republic, Amur village, meadow near the lake	N50.223 E84.898	1980	PQ810840	PQ807763
36	Russia: Altai Republic, Kosh-Agach village, solonchak	N49.992 E88.664	1800	PQ810824	PQ807747
37	Russia: Altai Republic, Kosh-Agach village, solonchak	N49.992 E88.664	1800	PQ810825	PQ807748

Table (continued)

No.	Loc.	Coord.	Altitude	cox1	cytb
38	Russia: Republic of Tuva, Lake Dus-Khol, desertified steppe	N50.384 E94.870	990	PQ810836	PQ807759
39	Russia: Buryatia Republic, Kizhinga village, steppe	N51.846 E109.913	700	ON220883-84	ON228270, ON228274
40	Russia: Zabaikalsky Krai, Gazimur River, mixed-grass meadow	N51.764 E118.425	655	PQ810832	PQ807755
41	Russia: Zabaikalsky Krai, Gazimur River, mixed-grass meadow	N51.764 E118.425	655	PQ810833	PQ807756
42	Russia: Zabaikalsky Krai, Mount Kozlikha, meadow on a slope near a stream	N51.760 E118.713	730	PQ810837	PQ807760
43	Russia: Zabaikalsky Krai, Talovka River, swampy meadow	N51.450 E118.791	815	PQ810853	PQ807776
44	Russia: Zabaikalsky Krai, Uktycha River, steppe slope	N51.545 E118.342	750	PQ810854	PQ807777
45	Russia: Zabaikalsky Krai, Solonechnaya River, steppe meadow	N51.440 E118.907	750	PQ810855	PQ807778
46	Russia: Republic of Yakutia, Yakutsk, hygromesophytic meadow	N62.087 E129.289	260	PQ810845	PQ807768
47	Russia: Republic of Yakutia, Yakutsk, xerophytic meadow	N62.055 E129.644	100	PQ810846	PQ807769
48	Russia: Republic of Yakutia, Yakutsk, solonchak meadow	N62.021 E129.627	250	PQ810847	PQ807770
49	Russia: Republic of Yakutia, Yakutsk, lake Ytyk-Kyuel, meadow near the shore	N62.017 E129.618	100	PQ810848	PQ807771
50	Russia: Republic of Yakutia, Yakutsk, xerophytic meadow	N62.017 E129.618	100	PQ810849	PQ807772
51	Russia: Republic of Yakutia, Khayirdakh River, bank with <i>Chosenia</i>	N66.263 E144.268	606	PQ810850	PQ807773
52	Russia: Magadan oblast, Yagodnoye village, bank of Sokhatina River	N62.517 E149.617	465	ON220878	ON228277
53	Russia: Magadan oblast, Yagodnoye village, bank of Debin River	N62.750 E150.467	403	ON220879	ON228271
54	Russia: Magadan oblast, Madaun village, gravel high floodplain	N60.615 E150.696	520	PQ810839	PQ807762
55	Russia: Magadan oblast, Klepka village, gravel high floodplain	N59.754 E151.491	78	PQ810838	PQ807761
56	Russia: Magadan oblast, Arman village, gravel high floodplain	N59.667 E150.117	1	ON220872	ON228275
57	Russia: Kamchatka oblast	N53.190 E158.480	60	LT977377 (as <i>F. candida</i>)	–
58	Russia: Sakhalin, Aniva, dry meadow	N46.019 E143.413	< than 10	PQ810812	PQ807735
59	Russia: Sakhalin, Aniva, dry meadow	N46.019 E143.413	< than 10	PQ810829	PQ807752
60	Russia: Sakhalin, Aniva, dry meadow	N46.019 E143.413	< than 10	PQ810819	PQ807742
61	Russia: Komsomolsk-on-Amur, Silinka River, floodplain with poplars	N50.550 E137.000	30	PQ810856	PQ807779
62	Russia: Khabarovsk Krai, sandy area on forest road	N66.263 E134.330	74	PQ810844	PQ807767
63	Russia: Primorsky Krai, Lazo town, Lazovka river's bank	N43.380 E133.898	220	ON220881	ON228276
64	Russia: Primorsky Krai, Kievka River, sandy seaside meadow	N42.855 E133.675	2–3	PQ810817	PQ807740
65	Russia: Primorsky Krai, Kievka River, sandy seaside meadow	N42.855 E133.675	2–3	PQ810818	PQ807741
66	Russia: Primorsky Krai, Kievka River, sandy seaside meadow	N42.855 E133.675	2–3	PQ810828	PQ807751
67	Russia: Primorsky Krai, Gryznaya River, overgrown sand and gravel shoal	N43.361 E131.556	130–135	PQ810814	PQ807737

Table (end)

No.	Loc.	Coord.	Altitude	cox1	cytb
68	Russia: Primorsky Krai, Gryaznaya River, overgrown sand and gravel shoal	N43.361 E131.556	130–135	PQ810816	PQ807739
69	Russia: Primorsky Krai, Sinii Utes town, gravel area	N43.062 E131.356	40–45	ON220882	ON228272
70	Russia: Primorsky Krai, Andreyevka village, wasteland	N42.649 E131.140	7	PQ810813	PQ807736
71	Russia: Primorsky Krai, Andreyevka village, wasteland	N42.649 E131.140	7	PQ810815	PQ807738
72	Russia: Primorsky Krai, Andreyevka village, wasteland	N42.649 E131.140	7	PQ810827	PQ807750
73	South Korea	–	–	MT800217 (as <i>F. candida</i>)	–
74	China: Yanbian, Jilin	N43.133 E129.183	750	KX665026-27 (as <i>F. candida</i>)	–
75	China: Harbin	N45.750 E126.633	140	KX665031-32 (as <i>F. candida</i>)	–
76	China: Xilin Gol League, Inner Mongolia	N43.930 E116.090	1040	KX665022, KX665024 (as <i>F. candida</i>)	–
77	China: Ulanqab, Inner Mongolia	N40.994 E113.132	1370	KX665023, KX665025 (as <i>F. candida</i>)	–
78	China: Helan Mountains, Inner Mongolia	N38.900 E105.967	3260	KX665019-21 (as <i>F. candida</i>)	–
79	China: Hebei, Qinglong	N40.407 E118.950	245	KX665028 (as <i>F. candida</i>)	–
80	China: Hebei, Xiaowutai Mountain	N39.941 E115.043	2840	KX665029-30 (as <i>F. candida</i>)	–
81	China: Hebei	–	–	HQ619704, HQ619710 (as <i>F. candida</i>)	HQ651081, HQ651074 (as <i>F. candida</i>)
82	China: Zhong-Tiao Mountains	N35.116 E111.772	520	KX665033 (as <i>F. candida</i>)	–
83	China: Liupan Mountains, Ningxia	N35.657 E106.205	1300	KX665034 (as <i>F. candida</i>)	–
Formica candida					
84	Kyrgyzstan: Alai Valley	N39.703 E73.457	3200	–	AY786154*, AY78615, JX170887
85	Kyrgyzstan: Tian-Shan	–	–	–	AY786155–AY786156

Note. Loc. – location numbers referring to Fig. 2; altitude provided in m above sea level. Asterisks denote type of material/locations. Specimens sequenced in this study are indicated in boldface.

cytb sequence for the *F. candida* neotype (AY786154). The haplotype network was constructed using the median joining algorithm in Pop ART v. 1.7 (Leigh, Bryant, 2015).

Results

Out of the 47 ant specimens sequenced in this study for *cox1* and *cytb*, 6 belonged to *F. picea*, and 41, to *Formica* sp. We did not detect haplotypes of *F. candida* (Seifert, 2004) in any of the sampling points, despite the extensive survey area. *F. picea* and *Formica* sp. significantly differed at the DNA level (Fig. 4).

For *F. picea*, a notable geographic split was evident: samples from Europe and West Siberia formed a tight cluster, while samples from Armenia and Transbaikalia diverged from them

(Fig. 4). For *Formica* sp., the larger sample size resulted in a more informative network. However, no regions with a specific set of haplotypes were identified. Genetic diversity was substantial, especially for three samples (from different regions) that differed markedly from other haplotypes.

F. picea was found in five locations east of the Urals: in the Altai mountains, in the lowlands of West Siberia, as well as in East Siberia. The West Siberian location (Melnikovo, Tomsk oblast) was in a fen with a tall birch. The nest (with no external structures) was located in a sedge hummock with a fern; nest entry was in moss.

In the Altai mountains, we found nests of *F. picea*: three in Central Altai near the Abay (locations No. 17 in the Table and Fig. 2) and Amur (locations No. 18–19 in the Table and



Fig. 3. Habitats of *Formica* sp. and *F. picea*.

a – Central Altai, near Abay village, lowland swamp (*F. picea*); b – the same area; meadows on the site of drained swamps (co-occurrence of *Formica* sp. and *F. picea*); c – the same area; stony steppes on rocky outcrops (*F. picea*); d – Central Altai, near Ust-Kan village, steppe (*Formica* sp.); e – Magadan oblast, Nalednyy River (tributary of the Ola River) (*Formica* sp.); f – Zabaikalsky Krai, floodplain of the Talovka River (*Formica* sp.); g – Sakha Republic (Yakutia), Momsky District, Khairdakh River (*Formica* sp.); h – Magadan oblast, upper reaches of the Kolyma River, steppe (*Formica* sp.); i – Magadan oblast, an island in the upper reaches of the Kolyma River near the eastern end of the Big Ananchag Range (*Formica* sp.); j – same area; nest exits of *Formica* sp.; k – Zabaikalsky Krai, floodplain of the Kotikha River (*F. picea*); l – Novosibirsk oblast, Kabinetnoye village, halophytic meadow (*Formica* sp.); m – Southeast Altai, near Kosh-Agach village, Tabozhok town, stony steppe (*F. picea*). For more details, see the Table.

Fig. 2) villages (ca. 1100 m a.s.l.), and one in the Southeast Altai (locations No. 20 in the Table and Fig. 2), near the Kosh-Agach village (2275 m a.s.l.). Near the Abay village, a single nest was detected in a damp area on the edge of a bog, where it transitioned into a steppe meadow, 150–200 m away from a small lake. The nest had a ground part in the form of a small dome made of thin blades of grass, no more than 10 cm high and about 15 cm in diameter. Two other nests were found on southern steppe slopes, with no external structures and exits located under the rocks.

In East Siberia, *F. picea* was detected in Zabaykalsky Krai, east of the Gazimursky Zavod town (locations No. 21 in the Table and Fig. 2), at 770 m a.s.l., within the forest belt. The

nest was found on a swampy floodplain of the Kotikha River (a third-order tributary of the Gazimur River), with saplings of alder, birch, and larch. The nest, without visible external structures, was located under a layer of litter composed of fallen leaves and dry grass.

Specimens of *Formica* sp. were detected in 61 locations (Fig. 2). Their habitats included floodplain (including long-term and deeply flooded sandy-pebble spits) and non-floodplain terraces of different heights on rivers of various sizes, meadows of diverse composition and projective cover, sparse forests, steppes (from mixed grass to desert), foothills and highlands of Altai. This ant was found to avoid forests with closed canopies, as well as wetlands.

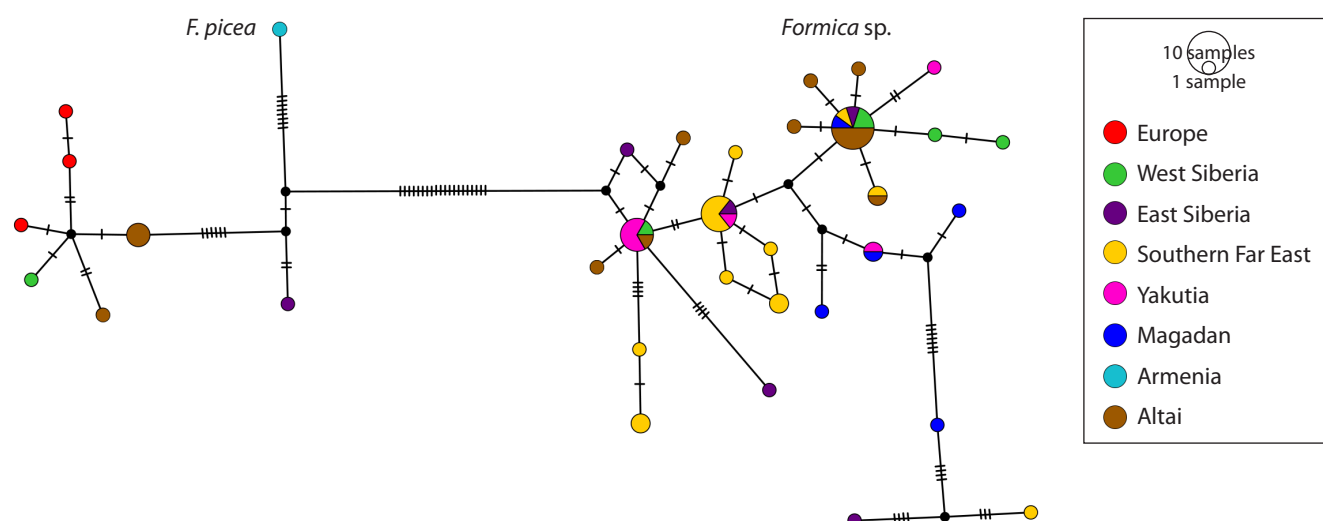


Fig. 4. Haplotype network for the combined *cox1*+*cytb* sequences. Dashes represent the number of nucleotide substitutions.

Discussion

Genetic patterns

As in the previous study (Zhigul'skaya et al., 2022), *F. picea* and *Formica* sp. were significantly different on mtDNA levels, with no intermediate haplotypes found (Fig. 4). Expectedly, we found no haplotypes of *F. candida* sensu Seifert (2004), which is probably limited to certain regions of Kyrgyzstan.

Throughout the distribution of *Formica* sp., from the Altai mountains to the Pacific coast and eastern China, there is a pretty high genetic variation, but with no obvious genetic structure (Fig. 4). This probably indicates recent colonization of this region without a bottleneck event and high dispersal abilities of the species. Unfortunately, there are no *cox1* data on *Formica* sp. from China, a region with a milder climate that could serve as a refugium/ancestral region of *Formica* sp.

Noteworthy, three haplotypes differed by 8–15 nucleotide substitutions from the main cluster (Fig. 4). However, all three were from different and geographically remote regions. This might imply that the history of the species is complex, with high dispersal rate and multiple colonization waves.

Although the data on the *cox1* gene in *F. picea* remain limited, several interesting insights can be gleaned from the haplotype network (Fig. 4). Most haplotypes form a tight cluster, indicating close genetic relationships. This cluster comprises specimens from the Altai and West Siberia, suggesting recent (likely Holocene) dispersal into these regions. The Armenian haplotype is distantly related to this group, as expected given the geographic distance and mountain barriers. Notably, the East Siberian haplotype is also distinct from the European cluster. Intriguingly, this pattern may indicate that East Siberian populations of *F. picea* belong to a different genetic lineage relative to the European/West Siberian group, and that an East Siberian refugium for the species may have existed. Of course, drawing firm conclusions based on a single specimen would be premature; thus, we intend to examine additional specimens from the region in future studies.

Geographic distribution

There are a lot of studies on the distribution and habitat preferences of the *F. picea*/*F. candida* complex. In various parts of its range, these ants are known to inhabit diverse habitats: fens, meadows, steppes, as well as sparse forests on plains, floodplains, and in highlands. In the south taiga zone of Siberia, black bog ants are found in swamps, in wet mixed forests, in pine woods (Omelchenko, Zhigul'skaya, 1998; Dlussky, 2001; Gridina, 2003). They are ubiquitous in Central Yakutia and the Baikal Region, from boggy larch forests, dry pine-larch forests, and steppe-like and steppe biotopes to bare rock formations, with nest densities reaching up to 30 nests per 100 square meters (Pleshanov, 1966; Dmitrienko, Petrenko, 1976).

However, they are particularly ecologically versatile and abundant in the mountain-steppe landscapes of Tuva, South-east and Central Altai, as well as in the steppes of Transbaikalia, where they dominate almost all habitats, from solonchaks and solonchaks to dry steppes on sandy and stony soils (Zhigul'skaya, 1968, 2009, 2011; Chesnokova, Omelchenko, 2011).

In Northeast Asia, the complex is associated primarily with river valleys; high abundance is observed in sparse *Populus* and *Chosenia* groves and on gravel-sandy spits in floodplains. Much less frequently, and only as isolated nests, *Formica* sp. can be found in mesoxeromorphic sparse forests (in Yakutia, specifically in pine forests), in steppe habitats, or southern slopes with sphagnum-moss bogs (Berman et al., 2007).

Since the black bog ant turned out to include several cryptic taxa, this raises a question if they have different habitat preferences. Here we obtained certain data on this point, which are, of course, far from exhaustive. For *F. picea*, we currently know that it is found from West Europe to West Siberia. We also have single findings from Transbaikalia (this study) and Tibet (AY786157), implying the presence of this taxon between Siberia and Tibet. *F. picea* appears to be rare in Asia, since we found only six locations of this species compared to 41 for *Formica* sp. Of these, only two were not from the mountains: a location in Tomsk oblast and one from mid-altitude highlands

of Transbaikalia. Based on our findings, we can suggest that *F. picea* appears to be associated with fens, but not peatlands. The latter is probably due to the fact that peatlands in Siberia are still insufficiently studied by myrmecologists.

Formica sp., on the other hand, appears to be found even more widely than previously thought (Zhigulskaya et al., 2022). In this study it was recorded across a vast area, from West Siberia to the Pacific Ocean, and from 66.5°N in Yakutia (Khayrydakh River) roughly to 35°N in the southern part of the Korean Peninsula (Fig. 2), as well as from sea level up to approximately 3500 m a. s. l., across a broad range of habitats (except for forests with closed canopy). Furthermore, unlike *F. picea*, *Formica* sp. has not been found in bogs in Europe, Siberia, or Altai. Conversely, only *Formica* sp. is known from river floodplains, and from these areas, it can spread to xeromorphic slopes and sections of low floodplain terraces.

The versatility of *Formica* sp. to moisture is striking: it can colonize both long-term flooded habitats on floodplains (that is, for about a month of the 4.5-month active period in the Subarctic) and dry zonal steppes. Notably, in the Altai Mountains, *Formica* sp. nests were found in both types of habitats.

In this study, we found that the distributions of *Formica* sp. and *F. picea* significantly overlap, probably from the Urals to Transbaikalia (Fig. 2), and we expect this overlap to expand when new data on bog ants are obtained. However, the ecological preferences of these two taxa seem to differ, which can be prospectively formulated as follows: *Formica* sp. avoids fens, while *F. picea* does not like low floodplains.

In this connection, the cases of syntopic occurrences of both taxa in the Central Altai are of particular note. Here, these two nests were found a few dozens of meters apart on the floodplain of the Abay River. The nest of *F. picea* was at the edge of a wet swampy depression about 1.5 m deep, while that of *Formica* sp. was on a dry steppe meadow that was formed in a dried fen. In the same location, about 3 km apart, we found a *F. picea* nest on a southern slope of the residual hills with rock outcrops, occupied by steppes.

We can suggest that in the Altai, *F. picea* inhabits meadows, and can move into both wet habitats and moderately dry steppes on southern slopes, as well as mountain steppes above the forest edge. However, it is unclear why this species can use both very wet and very dry habitats (Bönnner, 1915; Skwarra, 1929; Stitz et al., 1939; Seifert, 2004, 2018).

To this, we can put forward the hypothesis that the generalist *F. picea* has to constrict itself to the habitats where it can compete with other ant species (Dlussky, 1967). “Apparently, while spreading westward, *F. picea* (a eurytopic forest-steppe species of Mongolian origin) was displaced by species of the *F. rufibarbis* group into bogs. In these bogs, the *F. rufibarbis* species are virtually absent, whereas *F. picea* can survive, owing to its eurytopy and plasticity in nest construction. In the mountains, this species inhabits higher altitudes than other species of the *F. rufibarbis* group (except for *F. subpilosa pamirica* and *litoralis* in Central Asia, which exclusively occupy river thalwegs, rocky screes, and sands). *F. picea* cannot compete with *F. fusca* and its closely related species, as it is virtually absent in forests; instead, it is associated with more or less open habitats (meadows, steppified areas, sparse fo-

rests, river floodplains, and bogs)” (translated from: Dlussky, 1967, p. 63).

If this idea is correct, *Formica* sp. outcompetes *F. picea* on the vast floodplains and xeromorphic habitats of Asian Russia. So, *F. picea* is outcompeted both in Siberia and in Europe by different local ant species.

Conclusions

We can summarize the following important points inferred from our study: only *F. picea* is found in Europe, whereas both *F. picea* and *Formica* sp. are found in North Asia. We failed to detect *F. candida* on this vast territory, and we can surmise that in the current understanding proposed by Seifert (2004), it is a Central Asian taxon. In the south of West and East Siberia, the ranges of *F. picea* and *Formica* sp. overlap, but they have different ecological preferences. *F. picea* is probably constrained to fens on the plains, and to subalpine meadows in the mountains. *Formica* sp. is broadly tolerant to humidity but is not found in fens. The cases of syntopic distribution of these ants, which we described in Altai, are apparently associated with transient successional processes. The relative rarity of *F. picea* in our collections possibly reflects the insufficient study of raised bogs in Asia, the typical habitats of the species in Europe. Differences in the distribution and habitat preference of *F. picea* and *Formica* sp. are probably the result of competition, according to G.M. Dlussky (1967).

F. picea and *Formica* sp. differ genetically and have different distribution ranges, as well as habitat preferences. This supports the opinion that *Formica* sp. should be recognized as a distinct species.

References

- Antonov I.A., Pleshchikov A.S. Landscape ecological complexes of ants in Baikalian Siberia. *Sibirskij Ekologicheskij Zhurnal (Siberian Journal of Ecology)*. 2008;15(1):53-57 (in Russian)
- Antonov I.A., Bukin Y.S. Molecular phylogenetic analysis of the ant genus *Formica* L. (Hymenoptera: Formicidae) from Palearctic region. *Russ J Genet.* 2016;52(8):810-820. doi 10.1134/S1022795416080020
- Berman D.I., Alfimov A.V., Zhigulskaya Z.A., Leirikh A.N. Wintering and Cold Hardiness of Ants in the North-East of Asia. Moscow, 2007 (in Russian)
- Bolton B. A New General Catalogue of the Ants of the World. Cambridge, Mass.: Harvard University Press, 1995
- Bönnner W. Die Überwinterung von *Formica picea* und andere biologische Beobachtungen. *Biol Zent Bl.* 1915;35:65-77 (in German)
- Chen Y., Zhou S. Phylogenetic relationships based on DNA barcoding among 16 species of the ant genus *Formica* (Hymenoptera: Formicidae) from China. *J Insect Sci.* 2017;17(6):117. doi 10.1093/jisesa/ix092
- Chesnokova S.V., Omelchenko L.V. Ants of Central Altai: Spatial-topological structure and classification of communities. *Entomol Rev.* 2011;91(2):253-263. doi 10.1134/S0013873811020138
- Dlussky G.M. Ants of the Genus *Formica* (Hymenoptera: Formicidae g. *Formica*). Moscow, 1967 (in Russian)
- Dlussky G.M. Seasonal dynamics of *Formica candida* colonies in an isolated bog population. In: Ants and Protection of the Forest. Proceedings of the XI All-Russian Myrmecological Symposium. Perm, 2001;69-71 (in Russian)
- Dmitrienko V.K., Petrenko E.S. Ants of Taiga Biocenoses of Siberia. Novosibirsk, 1976 (in Russian)
- Folmer O., Black M., Hoeh W., Lutz R., Vrijenhoek R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from

- diverse metazoan invertebrates. *Mol Mar Biol Biotechnol.* 1994; 3(5):294-299
- Goropashnaya A. Phylogeographic Structure and Genetic Variation in *Formica* Ants. Upsala, 2003
- Goropashnaya A.V., Fedorov V.B., Seifert B., Pamilo P. Phylogenetic relationships of palaeartic *Formica* species (Hymenoptera, Formicidae) based on mitochondrial cytochrome *b* sequences. *PLoS One.* 2012;7(7):e41697. doi 10.1371/journal.pone.0041697
- Gridina T.I. Ants of the Urals and their geographical distribution. *Uspekhi Sovremennoi Biologii = Advances in Current Biology.* 2003; 123(3):289-298 (in Russian)
- Kupyanskaya A.N. Ants (Hymenoptera, Formicidae) of the Far East. Vladivostok, 1990 (in Russian)
- Leigh J.W., Bryant D. POPART: full-feature software for haplotype network construction. *Methods Ecol Evol.* 2015;6(9):1110-1116. doi 10.1111/2041-210X.12410
- Omelchenko L.V., Zhigul'skaya Z.A. Ecological-faunistic analysis of ants in the southern taiga of western and central Siberia. *Uspekhi Sovremennoi Biologii = Advances in Current Biology.* 1998;118(3): 299-305 (in Russian)
- Pleshanov A.S. Family Formicidae (Ants). In: Pests of Siberian Larch. Moscow, 1966;194-302 (in Russian)
- Radchenko A.G. Ants (Hymenoptera, Formicidae) of Ukraine. Kiev, 2016 (in Russian)
- Ruzsky M.D. Ants of Russia. Part 1. *Trudy Obshchestva Yestestvoispytatelej pri Imperatorskom Kazanskom Universitete = Proceedings of the Naturalist Society through the Kazan Imperial University.* 1905; 38(5-7):3-798 (in Russian)
- Schär S., Talavera G., Espadaler X., Rana J.D., Andersen A.A., Cover S.P., Vila R. Do Holarctic ant species exist? Trans-Beringian dispersal and homoplasy in the Formicidae. *J Biogeogr.* 2018;45(8): 1917-1928. doi 10.1111/jbi.13380
- Shekhovtsov S.V., Derzhinsky Y.A., Poluboyarova T.V., Golovanova E.V., Peltek S.E. Phylogeography and genetic lineages of *Aporrectodea rosea* (Lumbricidae, Annelida). *Eur J Soil Biol.* 2020;99: 103191. doi 10.1016/j.ejsobi.2020.103191
- Seifert B. The "Black Bog Ant" *Formica picea* Nylander, 1846 – a species different from *Formica candida* Smith, 1878 (Hymenoptera: Formicidae). *Myrmecologische Nachrichten.* 2004;6:29-38
- Seifert B. The Ants of Central and North Europe. Lutra Verlags – und Vertriebsgesellschaft. Tauer, Germany, 2018
- Skwarra E. *Formica fusca-picea* Nyl. als Moorageise Biologische. *Anzeiger.* 1929;82:46-55 (in German)
- Stitz H. Hautflügler oder Hymenoptera: Ameisen oder Formicidae. In: Die Tierwelt Deutschlands. Jena, 1939 (in German)
- Zakharov A.A. Ants of Forest Communities, their Life and Role in the Forest. Moscow, 2015 (in Russian)
- Zakharov A.A., Dlussky G.M., Goryunov D.N., Gilyov A.V., Zryanin V.A., Fedoseeva E.B., Gorokhovskaya E.A., Radchenko A.G. Monitoring of the *Formica* Ants. Moscow, 2019 (in Russian)
- Zhigul'skaya Z.A. Ant (Formicidae) population of steppe Tuva landscapes. In: Animal Population of Soils in Treeless Biogeocenoses of the Altai-Sayan Mountain System. Novosibirsk, 1968;115-139 (in Russian)
- Zhigul'skaya Z.A. The ants of the Chuya depression and the Yustyd River basin in Southeastern Altai. *Contemp Probl Ecol.* 2009;2(3): 210-215. doi 10.1134/S1995425509030084
- Zhigul'skaya Z.A. Ants (Hymenoptera, Formicidae) of the Kuraiskaya depression of the Southeastern Altai. *Tomsk State University Journal.* 2011;350:189-195 (in Russian)
- Zhigul'skaya Z.A., Shekhovtsov S.V., Poluboyarova T.V., Berman D.I. *Formica picea* and *F. candida* (Hymenoptera: Formicidae): synonyms or two species? *Diversity.* 2022;14(8):613. doi 10.3390/d14080613

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

Received August 29, 2025. Revised September 23, 2025. Accepted October 6, 2025.

doi 10.18699/vjgb-26-32

Functional symmetry and reproducibility of the evolutionary process

S.I. Bartsev ^{1, 2}¹ Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Krasnoyarsk, Russia² Siberian Federal University, Krasnoyarsk, Russia bartsev@yandex.ru

Abstract. The question on the reproducibility of evolutionary processes is primarily of fundamental importance; however, with the development of methods for modeling evolutionary processes on computer multilevel models, an answer to this question is necessary to clarify the status of the predictions obtained. Experimental obtaining of ensembles of evolutionary outcomes for subsequent statistical processing on real biological systems seems to be impracticable. At the same time, the results obtained on multilevel computer models are difficult to interpret due to their complexity and the dependence of modeling results on a variety of parameters. This work is aimed at identifying common properties of evolving systems using a simple heuristic model based on transparent general principles and ideas about the key properties of biological systems that are important for the evolutionary process. Agents undergoing evolutionary changes are recurrent neural networks with a well-defined structure, a given function, and a specific rule for modifying the structure in the direction of maximum fitness. A separate instance of a neural network formed during the evolutionary process is called neural network model object (NNMO). Computational experiments have been carried out to generate ensembles of NNMO structures performing a given function, and the patterns of NNMO distribution in the structural space have been analyzed. This analysis confirms the presence of functional symmetry in the structure of NNMOs performing the same function. An assessment of the stability and reproducibility of individual evolutionary trajectories has been carried out. It is shown that under certain constraints leading to a reduction of the complexity of the NNMO structure (analogous to a narrow environmental specialization), the final NNMO structures may be close, but not identical. This suggests an inaccurate reproduction of the evolution of the structure with functional equivalence. Nevertheless, it can be argued that in the general case, the very ability for evolutionary change is possible with the redundancy of the potential complexity of the structure over the functional complexity and automatically entails a multiplicity of evolutionary outcomes based on the fact that the same function can be implemented by different, but functionally invariant structures.

Key words: reproducibility of the evolutionary process; equifinality of evolutionary outcomes; functional symmetry; heuristic neural network model; functional complexity

For citation: Bartsev S.I. Functional symmetry and reproducibility of the evolutionary process. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):284-292. doi 10.18699/vjgb-26-32

Funding. The study was funded by State Assignment of the Ministry of Science and Higher Education of the Russian Federation (project No. 0287-2021-0018).

Функциональная симметрия и воспроизводимость эволюционного процесса

С.И. Барцев ^{1, 2}¹ Институт биофизики Сибирского отделения Российской академии наук – обособленное подразделение Федерального исследовательского центра «Красноярский научный центр СО РАН», Красноярск, Россия² Сибирский федеральный университет, Красноярск, Россия bartsev@yandex.ru

Аннотация. Вопрос о воспроизводимости эволюционных процессов имеет в первую очередь фундаментальное значение, однако с развитием методов моделирования эволюционных процессов на компьютерных многоуровневых моделях ответ на этот вопрос необходим для прояснения статуса получаемых прогнозов. Экспериментальное получение ансамблей эволюционных исходов для последующей статистической обработки на реальных биологических системах представляется неосуществимым. В то же время прогнозы, сгенерированные многоуровневыми компьютерными моделями, вследствие их сложности и зависимости результатов моделирования от множества параметров с трудом поддаются интерпретации. Данная работа направлена на выявление общих свойств эволюционирующих систем с помощью простой эвристической модели, построенной на прозрачных общих принципах и представлениях о ключевых свойствах биологических систем, значимых для эволюционного процесса.

Агенты, претерпевающие эволюционные изменения, являются рекуррентными нейронными сетями с четко определенной структурой, заданной функцией и определенным правилом модификации структуры в направлении максимальной приспособленности. Отдельный экземпляр нейронной сети, формируемой в ходе эволюционного процесса, назван нейросетевым модельным объектом (НМО). В работе проведены вычислительные эксперименты по генерации ансамблей структур НМО, выполняющих заданную функцию, и проанализированы закономерности распределения НМО в структурном пространстве. Этот анализ подтверждает наличие функциональной симметрии структуры НМО, выполняющих одну и ту же функцию. Оценены устойчивость и воспроизводимость индивидуальных эволюционных траекторий. Показано, что при определенных ограничениях, приводящих к редукции сложности структуры НМО (аналог – узкая экологическая специализация), финальные структуры НМО могут быть близки, но не идентичны. Это позволяет говорить о неточном воспроизведении эволюции структуры на фоне функциональной эквивалентности. Тем не менее можно утверждать, что в общем случае сама способность к эволюционным изменениям реализуется при избыточности потенциальной сложности структуры над функциональной сложностью и автоматически влечет за собой множественность эволюционных исходов, основанную на том, что одна и та же функция может реализовываться различными, но функционально инвариантными структурами.

Ключевые слова: воспроизводимость эволюционного процесса; эквивалентность эволюционных исходов; функциональная симметрия; эвристическая нейросетевая модель; функциональная сложность

A true seeker does not know what he is looking for,
a true wanderer does not know where he is going.

Attributed to Lao-tzu

Introduction

The extent of our misunderstanding of life can be seen from the abundance of problems associated with it. One of the key scientific problems of biology is the problem of predictability and/or possible equifinality of biological evolution.

According to M. Eigen, “every single system that has emerged as a result of mutation and selection is unpredictable in terms of its structure; nevertheless, the inevitable result is always the process of evolution – this is the law. ...The optimizing process of evolution is in principle inevitable, although the choice of a particular path is not deterministic” (Eigen, 1971).

At the same time, there is a certain parallelism in the evolution of various species and genera, which found its expression in the law of homological series in hereditary variability by N.I. Vavilov (Meyen et al., 1977). If there is some structural similarity in systems with the same function, but formed along different evolutionary trajectories, then we can talk about the equifinality (Meyen, 1974; Meyen et al., 1977) of evolution in a certain sense.

The question of whether evolution follows a deterministic path or can follow alternative trajectories attracts many researchers (Povolotskaya, Kondrashov, 2010; Lobkovsky et al., 2011; Orgogozo, 2015; Xue et al., 2017). Currently, there are contradictions between some theoretical concepts and experimental data on the reproducibility of evolutionary trajectories.

For example, the paper (Lobkovsky, Koonin, 2012) notes that with a high intensity of mutations, evolution can follow different trajectories on the fitness landscape, but, ultimately, they all converge at a single peak corresponding to the most adapted structure of the evolving biological system. And the paper (Orgogozo, 2015) considers the

possibility of predicting genotype by phenotype, i. e. the author admits the existence of a one-to-one mapping of the genotype into the phenotype, which means the uniqueness of the evolutionary outcome, that is, the similarity of the genotype with the same phenotype.

There are experiments in favor of the uniqueness of evolutionary outcomes (Weinreich et al., 2006; Dickins, Nekrutenko, 2009; Meyer et al., 2012), but there is evidence (Poelwijk et al., 2007; Dunham et al., 2009; Kvitek, Sherlock, 2011; Podgornaia, Laub, 2015; Starr et al., 2017) demonstrating the existence of several endpoints of evolution.

The complexity of research in this area is due to the fact that direct research of the evolutionary processes of biological systems in nature encounters three main obstacles: (A) the uniqueness of evolutionary outcomes, which does not allow using comparative analysis of biological structures; (B) the characteristic time of evolutionary changes, which usually exceeds the life span of the researcher many times; (C) the extreme complexity of real biological objects.

A possible approach to overcoming the stated obstacles immediately suggests itself: this is mathematical modeling and its special kind – computer modeling.

Computer modeling makes it possible to overcome the first and second of these obstacles: the researcher can obtain ensembles of model evolutionary trajectories to which statistical data processing can be applied to identify common patterns of evolutionary processes. But the status of the results of computer modeling depends on how well it is possible to cope with the complexity of the biological system and provide the adequacy of its description. There are two main obstacles.

First, when modeling real events, models are needed to be as close in complexity as possible to real systems, but

here we face the curse of dimensionality – the number of fitting parameters in conditions of uncertainty about the exact type of functional dependencies makes it impossible to determine them due to a lack of evolutionary data.

Secondly, there are doubts (Beckage et al., 2011; Garte et al., 2025) about the possibility of constructing adequate mathematical models of biological systems in general and evolutionary models in particular. The authors (Garte et al., 2025) note that “biological concepts resist clear definitions amenable to mathematical treatment. These include being alive and being an individual, as well as agency, inheritance, intelligence, sentience, and cognition” and on this basis they argue that the next “paradigm shift will require not just new mathematical tools but a new scientific EPISTEMOLOGY”, that is, ways of representing knowledge. At the same time, in their opinion, “choosing which data to pay attention to is far more valuable than collecting large amounts of data”.

One of the ways to solve the equifinality problem is to carefully analyze the theoretical arguments and experimental data available in the press in favor of a particular position.

Another way is to get expected and unexpected answers based on the development of initial principles and general ideas about the evolution of life. The study presented here is based on previously obtained results (Bartsev, Bartseva, 2002, 2006, 2010; Bartsev, Baturina, 2019).

Methods and materials

Since the approach to studying the properties of biological evolution used in this paper is quite unusual from the point of view of traditional molecular biology, it seems necessary to focus on the methodological foundations of the proposed approach.

Models and model objects are tools of scientific research, starting with Galileo, who, technically unable to directly study the free fall of bodies, derived formulas for the kinematics of equidistant motion using a physical model – balls rolling down an inclined chute. In the future, the following definition will be useful to us: “If similarity can be established between two objects in at least one specific sense, then between these objects there is original-model relationship” (Lerner, 1972).

The paper will use heuristic models (Von Neumann, 1966), which, unlike traditional models, are not aimed at modeling (describing) specific real systems, but are abstract models designed to identify convenient concepts, widely applicable principles, and build a general theory.

Von Neumann justified the usefulness of moving towards heuristic model objects as follows: “Appealing to the organic, living world does not help us greatly, because we do not understand enough about how natural organisms function. We will stick to automata which we know completely because we made them...” (Von Neumann, 1966).

Some ways of setting the problem with this approach can be distinguished from Von Neumann’s reasoning: “Automa-

ta theory seeks general principles of organization, structure, language, information, and control. Many of these principles are applicable to both natural and artificial systems, and so a comparative study of these two types of automata is a good starting point. Their similarities and differences should be described and explained. Mathematical principles applicable to both types of automata should be developed. ...The question that one can then hope to answer, or at least investigate, is: What principles are involved in organizing these elementary parts into functioning organisms, and what are the essential quantitative characteristics of such organisms?” (Von Neumann, 1966).

The requirements for a heuristic model of an evolving biological system have long been formulated by J. Bernal: “Biology differs methodologically from other natural sciences in that the focus is primarily on the *functioning* and evolution of systems. Structure is important here only in relation to function and origin...” (Bernal, 1968, p. 112).

Bernal’s identification of functioning as a special characteristic of living beings is consistent with the approach of N. Rashevsky and R. Rosen, who considered the organism as a “set or system of functional mappings” (Rosen, 1958; Rashevsky, 1968, p. 63).

Practically any study of living objects is first the identification of functional patterns such as “stimulus–response” or “impact–response”, and then the definition of “mechanisms” that provide the implementation of this function (catalysis, inheritance, locomotion, recognition, etc.). The mechanism of the system is considered disclosed if it is described in terms of interactions of the parts that make up this system.

Formally, the mechanism is revealed by decomposing the original function $F(x)$ into simpler, basic functions $\{g_i\}$ implemented by the corresponding subsystems $\{S_i\}$ – the elements of the structure. Coupling coefficients $\{\alpha_{ij}\}$ describe the interaction (relationship) between the elements of the structure and actually define the structure of the system.

It is important to emphasize that the decomposition procedure corresponds to the network description, where $\{S_i\}$ are the nodes of the network, and $\{\alpha_{ij}\}$ are the coupling coefficients between the nodes. The network description has long been widely used in various fields of biology: from the description of the structure-property correlation of a chemical compound (Golovanov et al., 1998) to metabolic pathways, protein-protein interactions, genetic networks, the nervous system, food chains of ecosystems, etc. (Albert et al., 2000; Amaral et al., 2000; Edelman, Gally, 2001; Strogatz, 2001; Dunne et al., 2002; Sole, 2002; Stumpf et al., 2008; De Las Rivas, Fontanillo, 2010).

To describe a biological system, a network model must be able to evolve. An evolutionary process will be understood as a change in the structure of a biological system in accordance with some functional criterion of optimality (Eigen, 1971; Forst et al., 1995; Schuster, 1996).

An analysis of existing formal models shows that networks of formal neurons, the so-called neural networks

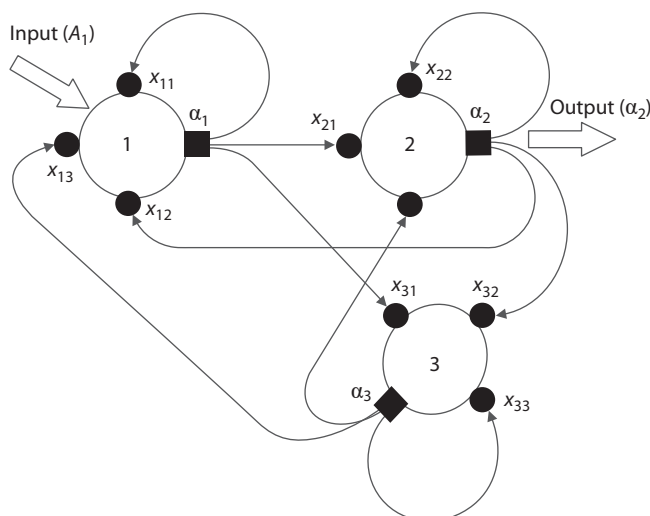


Fig. 1. A fully connected neural network of three neurons: 1 – input neuron; 2 – output neuron; 3 – associative (hidden) neuron, which is not directly connected to the input and output.

(NNs), are the most adequate objects for the purposes of this study. The advantage of NNs as a heuristic model of evolution is that it is easy for them to determine the FUNCTION, highlight the STRUCTURE and start the learning process, the formal description of which does not differ in key features from the process of biological EVOLUTION.

In this study, fully connected recurrent NNs were used, which can function in a stream of discrete events. Figure 1 illustrates a digraph of the simplest NN of 3 neurons, describing the ways in which signals are received and exchanged.

Formulas describing the functioning of the NN have the form:

$$\alpha_i^{k+1} = \frac{\rho_i^k}{a + |\rho_i^k|}, \quad \rho_i^k = \sum_j x_{ij} \alpha_j^k + A_i^k,$$

where α_i^k is the output signal of the i -th neuron at the k -th moment of time; ρ_i^k is the weighted sum of the output signals of the neurons received by the i -th neuron at the k -th moment of time, plus the external input signal A_i^k , received by the i -th input neuron at the k -th moment of time; x_{ij} is the matrix of weighting coefficients (synapses); a is the coefficient specifying the steepness of the activation function.

From the formula, it can be seen that the output signal of each neuron (its level of excitation) at each time depends on the state of the NN (and the external signal for the input neurons) at the previous time.

The modification of the NN structure was carried out by a random search algorithm, which was performed as follows: (1) a random increment was added to each synapse of the current NN, and two modified NNs were immediately created, differing in that the same modulo random increments were added to the corresponding synapses

with different signs; (2) all three NNs – the initial one and its two modified versions – were started to function in a stream of events with a duration of 100 symbols, and for each NN, the total goal function (loss function) for this fragment was estimated – the sum of the squares of the differences between the outputs of the NN and the required response; (3) based on the results of the comparison, the NN with the minimum goal function was chosen, then it was compared with the required level of accuracy, and if it exceeded it, the algorithm was restarted from step 1. Formally, this procedure corresponds to random mutational substitutions.

In order for the NN to act as a model of the evolutionary process, it is necessary to clearly state what the similarities are between it and a biological species in the process of evolution. The key components of the evolutionary process are heredity, variability, and natural selection, i. e. survival of the fittest. The measure of fitness for biological systems is difficult to deduce (Garte et al., 2025), but it is obvious that this measure should take into account the ability to reproduce and successful competition for resources.

In the course of natural selection, the fittest survive, therefore, to model natural selection, a measure of the fitness of one or another variant of the structure of NN is necessary. For neural networks, the goal function or loss function acts as a measure of fitness, which determines which variant of the NN structure will continue to evolve (learn). In other words, the goal function integrally includes an assessment of success in reproduction and competition for resources. In our case, the carrier of heredity is the structure of interneuronal connections, which is inherited by the next generation of the NN. The variability of the structure is provided by random increments of synapses, which leads to a variation in the quality of functioning of the neural network, i. e. a change in its adaptability to the environment in which it learns. We consider the evolutionary process in principle; the implementation features in the material are nuances.

During learning (evolution) and during functioning, a continuous, quasi-random sequence consisting of signals was applied to the input of the NN: “pause” – (00), A – (10), B – (01) and C – (11), where the numbers in parentheses indicate the presence or absence of a signal at the corresponding input neurons. The random parameters of the sequence were: the type of signal that will be received at the input and the duration of the pause between the signals. According to the training pattern, the neural network should output (11) if the input receives the “correct” signal, and (00) otherwise. The sequences may differ in the number of time cycles provided to the neural network for processing the input signal (3 or 4), which is indicated in the function designation (Fig. 2).

Despite their simplicity, such sequences are suitable for simulating a wide class of processes in biological systems.

Any biological system exists in a continuous flow of time and discrete flow of events; therefore, discrete states in which the system remains during specific time inter-

Input	+				+		+		+		+			
Output	AAA	BBB	BBB	CCC	AAA	AAA	BBB	AAA	CCC	BBB				
Input	+				+				+		+			
Output	CCCC	BBBB	AAAA	CCCC	AAAA	BBBB	CCCC	CCCC						

Fig. 2. Fragments of the input and output sequences for functions A3 and C4.

The "+" sign in the output sequence indicates the "correct" operation of the NN.

vals can be distinguished: the enzyme can be in a free or substrate-bound state; the glycolytic pathway can be in the mode of either glycolysis or gluconeogenesis, which, in particular, is determined by the state of the key enzyme phosphofructokinase/fructose-biphosphatase; the animal may be in a state of sleep, food intake, prey pursuit, etc. W. Ashby (1956) drew attention to the fact that a change in the types (patterns) of animal behavior can be described as a change in the states of a finite automation. In this case, the algebra of regular events (for example, Hopcroft et al., 2007, chapter 3) can formally describe the environment (and its complexity) in which the automation "lives" and the neural network "evolves".

Since computational experiments were aimed at identifying common patterns of the evolutionary process, ensembles of neural networks performing the same function were created during the experiments. For brevity, a separate instance of a trained neural network will be called a neural network model object (NNMO).

The initial structure of the NNMO (a matrix of weighting coefficients) was set by a random number generator modulo close to 0. Training ended when a given level of the goal function was reached, corresponding to the complete absence of errors in recognizing stimuli in the input stream. That is, all the trained instances implemented the required function equally well. The generation and reduction of the complexity of NNMOs was carried out using the original Lazarus (Object Pascal) program.

For visualization and statistical analysis, the structure of the NNMO was represented as a point in the space of weighting coefficients; the Euclidean distance between points corresponding to different NNMOs can be taken as a measure of the proximity of their structures. In addition to the distribution of NNMO structures in the structural space, the paper evaluated the divergence of trajectories when starting from the same initial state. Visualization was performed using the original program in the Scilab 6.1.1 environment.

Results

It should be immediately noted that the network description assumes the presence of addition and multiplication operations to describe the interaction of elements and their mutual influence on each other. But the specific values of a sum or product can be obtained in an infinite number of ways if we use non-integers. Hence, it's possible to assume, or rather say with certainty, that in the general case, the

maximum values of the fitness function can be achieved in different ways, i. e. by different functionally invariant structures. The results of computational experiments have confirmed this assumption.

Note that the minimum number of neurons in the NNMOs of the considered configuration, which provides the required quality of functioning, is 6, with two input neurons, two output neurons, and two associative or hidden neurons. Attempts to use NNs with fewer neurons made in this and earlier works (Bartsev, Bartseva, 2010) were unsuccessful. Thus, the minimum space of NNMO structures performing the specified functions is 36-dimensional. The localization of NNMO structures was visualized by projecting a 36-dimensional space into a 3-dimensional one, with various combinations of weight coefficient numbers being selected as the coordinates of the projection space.

It was previously shown that the structures of NNMOs in space are not distributed randomly (diffusely), but form clusters (Fig. 3). Since the points correspond to NNMO structures that perform the same function, we can talk about the presence of functional symmetry, that is, the preservation of an invariant (function) under certain transformations (displacements in the space of NNMO structures).

Note that displacements within clusters correspond to local symmetry, and jumps from cluster to cluster correspond to discrete transformations of permutation and sign change (Bartsev, Bartseva, 2006, 2010). The number of clusters N for NNMOs with an odd activation function is determined by the formula: $N = 2^K K!$, where K is the number of associative neurons. The power of 2 describes the number of sign-changing combinations, and the factorial describes the number of permutations of associative neurons in the weighting matrix. In the case of the C4 function, there are twice as many clusters, since the signals of the input neurons are identical and it is possible to rearrange these neurons in the synapse matrix.

Each cluster in Figure 3 corresponds to a peak with a cut-off top in the fitness landscape. We emphasize that the quality of functioning (the level of fitness) of the structures shown in the figure is identical, which follows from the nature of functionally invariant transformations and is confirmed in a computational experiment.

From the comparison of the left and right 3D images, a useful methodological conclusion can be drawn that representations of the properties of the fitness landscape critically depend on which structural elements are taken into account. In this case, 8 and 16 structural clusters appear only

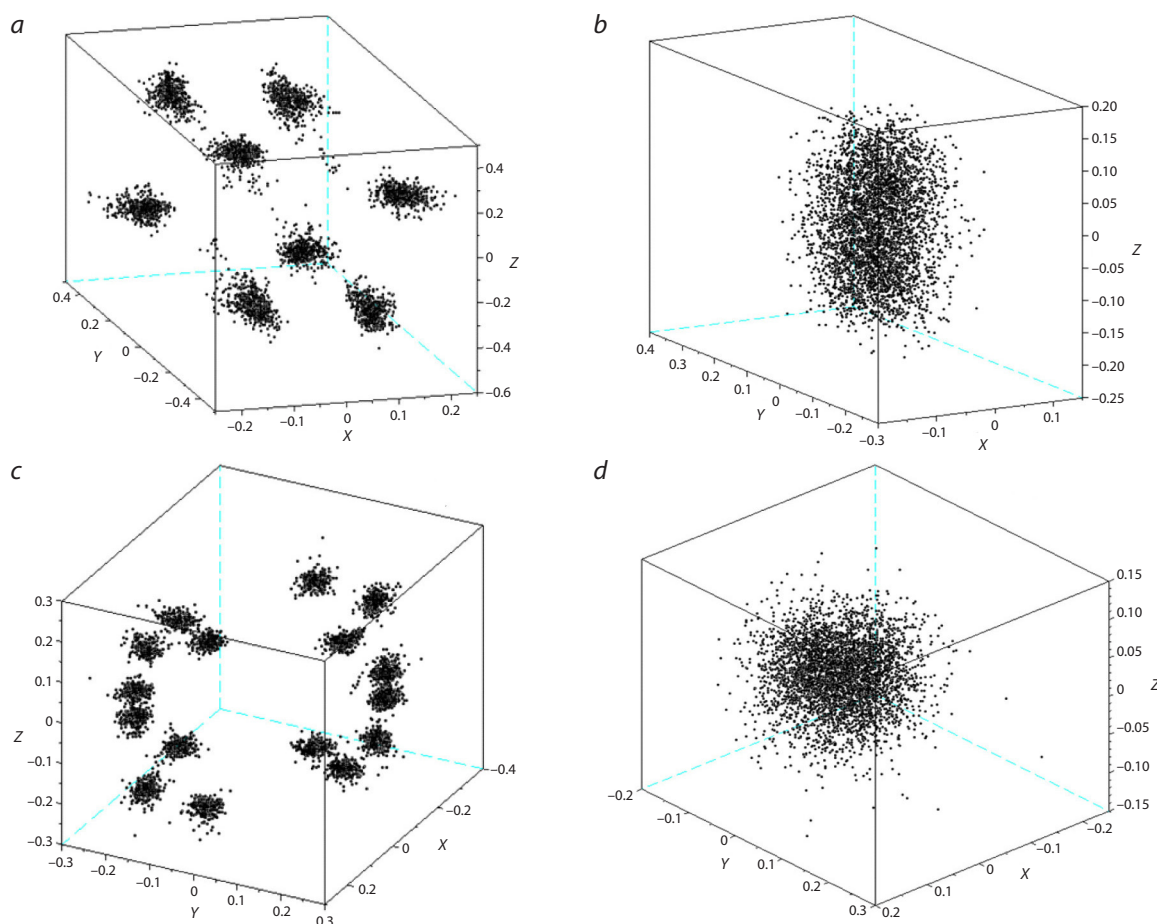


Fig. 3. The location of NNMOs structures with the function A3 (a, b) and C4 (c, d).

The XYZ coordinates for (a) and (c) correspond to the numbers of synapses 25/26/31. For (b) and (d) the coordinates are 9/12/22.

with a certain selection of parameters (weight coefficients) of the NNMO. Additionally, the bi- and polymodality of frequency distributions of pairwise Cartesian distances between NNMO structures is one of the indicators of cluster structure existence (Bartsev, Bartseva, 2010).

As already noted, the initial weight coefficients of the NNMOs were set by a random number generator, and all NNMOs started from different starting points (near 0) in the structure space. Then the presented result suggests the possibility of different outcomes of evolution, but it does not say anything about the stability and reproducibility of the evolutionary trajectory.

The reproducibility of the trajectory of simpler NNs for other tasks has been studied previously (Bartsev, Baturina, 2019). It has been shown that throughout the learning process, each point of the trajectory can generate a bundle of alternative trajectories, nevertheless directed towards the nearest cluster corresponding to maximum fitness. In this paper, to demonstrate the generally expected results, several implementations of training trajectories from the same initial state were obtained (Fig. 4). It is easy to see the rapid divergence of evolutionary trajectories, which is due:

firstly, to the randomness of perturbations (mutations) of the weight coefficients; and secondly, to the non-reproducibility of the flow of events (Fig. 2) (the surrounding world) in which NNMOs were trained.

As noted above, only NNMOs consisting of 6 or more neurons can adapt (learn this task) to a given environment. To estimate the minimum number of synapses capable of performing a given function (in given case, A3 and C4), the NNMO complexity reduction procedure was used. This procedure consisted in the sequential removal of minimal (in modulus) synapses and subsequent retraining of the NN to a given quality of functioning.

It was shown that the maximum number of weight coefficients for which the NNMO retains the same quality of functioning is 11 for sequence A3 (Fig. 5), and 13 for sequence C4. These differences in the minimum number of synapses are expected, since the C4 function is more complex than A3 due to the fact that the neural network needs to store the memory of the stimulus for longer before giving a response.

One more important point should be noted: if we take the structure of the reduced NNMO and replace the remaining

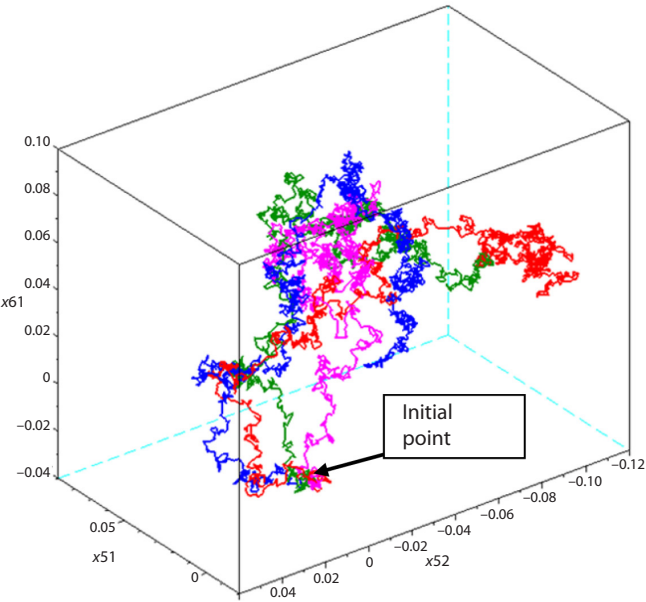


Fig. 4. An example of the evolutionary trajectories of NNMOs for the A3 function in coordinates 25/26/31.

non-zero synapses with random numbers (even conserving the sign), then NNMO learning does not occur. It follows from this that for learning (the evolutionary process), it is necessary to have degrees of freedom, i. e. weighting factors that are not fully involved in the implementation of the function and allow avoiding obstacles on the landscape of the fitness function via other dimensions.

It should be noted that the minimum number of synapses does not depend on the size of the initial neural network, but depends only on the function performed (Fig. 6). This fact allows us to assert that function complexity (for a

given transient characteristic of a neuron) is an invariant quantifiable property of the function itself.

The NNMO reduction procedure can be interpreted as the price paid for a non-zero weighting factor. In this case, the formation of a reduced structure corresponds to the rigid specialization of the organism, which subsequently prevents it from breaking out of the evolutionary impasse.

The minimum number of synapses required for the functioning of NNMOs can be taken as an estimate of functional complexity. Then the total number of NNMO synapses can be considered as a potential complexity, which is similar to the information capacity of the class (Eigen, 1971).

It follows from computational experiments that in order for the evolutionary process to be carried out, a redundancy of potential complexity over functional complexity is necessary, and this redundancy leads, in turn, to a multiplicity of evolutionary outcomes, which is possible only when the same function can be implemented by different structures.

Discussion

The results of the study of the heuristic evolutionary model have shown that, on the one hand, with a significant redundancy of potential complexity over functional complexity, it is impossible to predict the final structure of the NNMO. But, on the other hand, the introduction of restrictions, for example, the requirement of a minimum number of synapses, leads to a decrease in the number of possible final variants of the structure – their almost complete coincidence is possible, even with a difference in the initial size of the NNMOs. For example, comparing the right-hand weighting matrices in Figures 5 and 6, it is easy to see that the matrices of reduced 6- and 10-neuron NNMOs are very similar – it is enough to swap the bottom two rows and change the signs

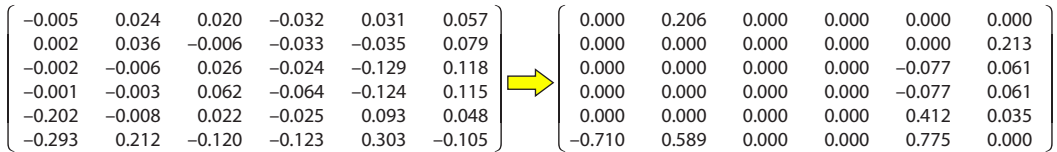


Fig. 5. An example of reducing the complexity of a 6-neural NNMO with the A3 function.

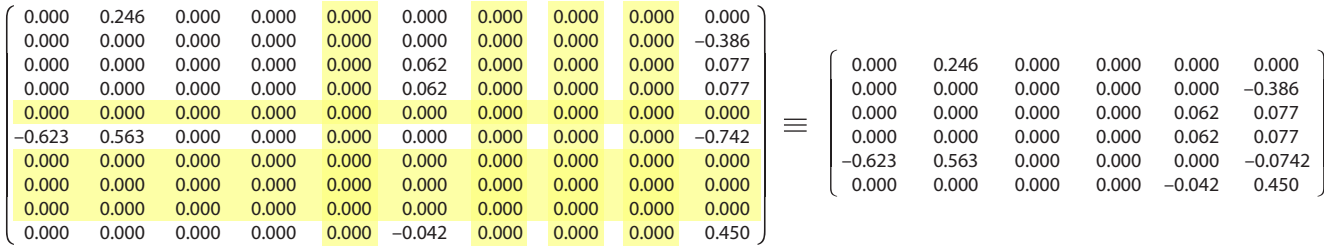


Fig. 6. Transfer of the already reduced structure of 10-neuron NNMO (A3 function) to the 6-neuron configuration by deleting the zero rows (and the corresponding columns) of the synapse matrix.

of synapses in the rows, that is, to make discrete functionally invariant transformations of the structure.

The conclusions by the heuristic model (from 2005) about the potential multiplicity of evolutionary outcomes were illustrated in the works (Podgornaia, Laub, 2015; Starr et al., 2017), where it was shown that in addition to the natural set of four amino acids, about 1 % of the combinations from the total set of 160 thousand can perform their function (of the same quality) in both proteins considered. At the same time, in the sequence space, the studied region splits into subdomains (clusters), within which evolutionary movements can occur relatively freely, while transitions between clusters are difficult, which corresponds well to the heuristic neural network model.

Herewith, if there are limitations and if the maximum of the fitness function is found in the vicinity, evolutionary trajectories can be reproduced (Weinreich et al., 2006; Meyer et al., 2012; Orgogozo, 2015), which also agrees with the heuristic model using the example of reduced NNMOs.

Note that the results obtained are not purely heuristic and computationally experimental. The possibility of functionally invariant transformations leads to the fact that “mutations” of the structure with identical or very similar quality of functioning (equal or close fitness) are possible, which corresponds to neutral mutations of Kimura (1979). The consequence of this is that at the next step, the evolutionary trajectory may take several paths equivalent in terms of fitness function.

In summary, we can agree with the claim that mathematical models (Garte et al., 2025) are generally unable to predict a specific evolutionary trajectory. Apparently, the requirement to predict a specific trajectory is in most cases an incorrect request.

Instead, the model, working with ensembles of trajectories, can calculate the probabilities and conditions of an evolutionary outcome, which will provide an understanding of the general properties of evolving systems. For example, the computational construction of the Agekyan–Anosova “homological map” (Lento et al., 2008) showed that the dependence of the lifetime of a 3-body system on the initial state has a fractal structure – and it becomes clear what can be expected from this system in principle.

Conclusion

An abstract heuristic model of evolution based on general assumptions about the nature of living things allows us to formulate assumptions (hypotheses) about the general properties of evolving systems. In particular, the very possibility of *evolutionary change* depends on the *redundancy of the structure* over the complexity of the function under selection, which generally leads to a *potential multiplicity of evolutionary outcomes* due to the *ambiguity of structural-functional relationships*. That is, in the general case, the specific path and the final structure (and its very achievement) of an evolving system are not deterministic given the functional equivalence of the evolutionary outcomes.

References

- Albert R., Jeong H., Barabasi A. Error and attack tolerance of complex networks. *Nature*. 2000;406:378–382. doi 10.1038/35019019
- Amaral L.A., Scala A., Barthelemy M., Stanley H.E. Classes of small-world networks. *Proc Natl Acad Sci USA*. 2000;97(21):11149–11152. doi 10.1073/pnas.200327197
- Ashby W.R. An Introduction to Cybernetics. New York, J. Wiley & Sons Inc, 1956
- Bartsev S.I., Bartseva O.D. Symmetric structures and equifinality of evolution outcomes in simple neural network models. *Dokl Biochem Biophys*. 2002;386(1-6):235–238. doi 10.1023/A:1020747225202
- Bartsev S.I., Bartseva O.D. Functional invariant approach to the biological system uniqueness: a simple neuronet model. *Doklady Biochemistry and Biophysics*. 2006;406(1):15–18. doi 10.1134/S1607672906010054
- Bartsev S.I., Bartseva O.D. Heuristic Neural Network Models in Biophysics: an application to the problem of structural and functional relations. Krasnoyarsk: Siberian Federal University, 2010 (in Russian)
- Bartsev S., Baturina P. A heuristic neural network model in the research of properties of evolutionary trajectories. *IOP Conf Ser Mater Sci Eng*. 2019;537(4):042001. doi 10.1088/1757-899X/537/4/042001
- Beckage B., Gross L.J., Kauffman S. The limits to prediction in ecological systems. *Ecosphere*. 2011;2(11):1–12. doi 10.1890/ES11-00211.1
- Bernal J.D. Molecular structure, biochemical physics and evolution. In: Theoretical and Mathematical Biology. Moscow: Mir Publ., 1968;110–151 (in Russian)
- De Las Rivas J., Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Computat Biol*. 2010;6(6):e1000807. doi 10.1371/journal.pcbi.1000807
- Dickins B., Nekrutenko A. High-resolution mapping of evolutionary trajectories in a phage. *Genome Biol Evol*. 2009;1:294–307. doi 10.1093/gbe/evp029
- Dunham E.J., Dugan V.G., Kaser E.K., Perkins S.E., Brown I.H., Holmes E.C., Taubenberger J.K. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J Virol*. 2009;83(11):5485–5494. doi 10.1128/JVI.02565-08
- Dunne J.A., Williams R.J., Martinez N.D. Food-web structure and network theory: the role of competence and size. *Proc Natl Acad Sci USA*. 2002;99(20):12917–12922. doi 10.1073/pnas.192407699
- Edelman G.M., Gally J.A. Degeneracy and complexity in biological systems. *Proc Natl Acad Sci USA*. 2001;98(24):13763–13768. doi 10.1073/pnas.231499798
- Eigen M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*. 1971;58(10):465–523. doi 10.1007/BF00623322
- Forst C.V., Reidys C., Weber J. Evolutionary dynamics and optimization. In: Morán F., Moreno A., Merelo J.J., Chacón P. (Eds) Advances in Artificial Life. ECAL 1995. Lecture Notes in Computer Science. Vol. 929. Springer, 1995;128–147. doi 10.1007/3-540-59496-5_294
- Garte S., Marshall P., Kauffman S. The reasonable ineffectiveness of mathematics in the biological sciences. *Entropy*. 2025;27(3):280. doi 10.3390/e27030280
- Golovanov I.B., Ivanitsky G.R., Tsygankova I.B. The simple form of the correlation relation and the physico-chemical meaning of its parameters. *Doklady Akademii Nauk*. 1998;359(2):258–262 (in Russian)
- Hopcroft J.E., Motwani R., Ullman J.D. Introduction to Automata Theory, Languages, and Computations. 3rd ed. Pearson, Addison-Wesley, USA, 2007
- Kimura M. The Neutral Theory of Molecular Evolution. *Sci American*. 1979;241(5):98–129. http://www.jstor.org/stable/24965339
- Kvitek D.J., Sherlock G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet*. 2011;7(4):e1002056. doi 10.1371/journal.pgen.1002056

- Lehto H.J., Kotiranta S., Valtonen M.J., Heinämäki P., Mikkola S., Chernin A.D. Mapping the three-body system – decay time and reversibility. *Mon Not R Astron Soc.* 2008;388(3):965-970. doi [10.1111/j.1365-2966.2008.13450.x](https://doi.org/10.1111/j.1365-2966.2008.13450.x)
- Lerner A.Y. Fundamentals of Cybernetics. Springer, 1972
- Lobkovsky A.E., Koonin E.V. Replaying the tape of life: quantification of the predictability of evolution. *Front Genet.* 2012;3:246. doi [10.3389/fgene.2012.00246](https://doi.org/10.3389/fgene.2012.00246)
- Lobkovsky A.E., Wolf Y.I., Koonin E.V. Predictability of evolutionary trajectories in fitness landscapes. *PLoS Computat Biol.* 2011;7(12):e1002302. doi [10.1371/journal.pcbi.1002302](https://doi.org/10.1371/journal.pcbi.1002302)
- Meyen S.V. The relationship between the nomogenetic and tychogenetic aspects of evolution. *J Gener Biol.* 1974;35(3):353-364 (in Russian)
- Meyen S.V., Sokolov B.S., Schrader Yu.A. Classical and nonclassical biology: the Lyubishchev phenomenon. *Vestnik Akademii Nauk USSR = Herald of the Academy of Sciences (USSR)*. 1977;10:112-124 (in Russian)
- Meyer J.R., Dobias D.T., Weitz J.S., Barrick J.E., Quick R.T., Lenski R.E. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science.* 2012;335(6067):428-432. doi [10.1126/science.1214449](https://doi.org/10.1126/science.1214449)
- Orgogozo V. Replaying the tape of life in the twenty-first century. *Interface Focus.* 2015;5(6):20150057. doi [10.1098/rsfs.2015.0057](https://doi.org/10.1098/rsfs.2015.0057)
- Podgornaia A.I., Laub M.T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science.* 2015;347(6222):673-677. doi [10.1126/science.1257360](https://doi.org/10.1126/science.1257360)
- Poelwijk F.J., Kiviet D.J., Weinreich D.M., Tans S.J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature.* 2007;445(7126):383-386. doi [10.1038/nature05451](https://doi.org/10.1038/nature05451)
- Povolotskaya I.S., Kondrashov F.A. Sequence space and the ongoing expansion of the protein universe. *Nature.* 2010;465(7300):922-926. doi [10.1038/nature09105](https://doi.org/10.1038/nature09105)
- Rashevsky N. Models and mathematical principles in biology. In: Theoretical and Mathematical Biology. Moscow: Mir Publ., 1968;48-66 (in Russian)
- Rosen R. A relational theory of biological systems. *Bull Math Biophys.* 1958;20:245-260. doi [10.1007/BF02478302](https://doi.org/10.1007/BF02478302)
- Schuster P. How does complexity arise in evolution? In: Götschl J. (Ed.) Evolution and Progress in Democracies. Theory and Decision Library. Vol. 31. Springer, 1996;147-159. doi [10.1007/978-94-017-1504-1_8](https://doi.org/10.1007/978-94-017-1504-1_8)
- Sole R.V. Modelling macroevolutionary patterns: an ecological perspective. In: Lässig M., Valleriani A. (Eds) Biological Evolution and Statistical Physics. Lecture Notes in Physics. Vol. 585. Springer, 2002;312-337. doi [10.1007/3-540-45692-9_18](https://doi.org/10.1007/3-540-45692-9_18)
- Starr T.N., Picton L.K., Thornton J.W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature.* 2017;549(7672):409-413. doi [10.1038/nature23902](https://doi.org/10.1038/nature23902)
- Strogatz S.H. Exploring complex networks. *Nature.* 2001;410(6825):268-276. doi [10.1038/35065725](https://doi.org/10.1038/35065725)
- Stumpf M.P.H., Thorne T., de Silva E., Stewart R., An H.J., Lappe M., Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci USA.* 2008;105(19):6959-6964. doi [10.1073/pnas.0708078105](https://doi.org/10.1073/pnas.0708078105)
- Von Neumann J. Theory of Self-Reproducing Automata. University of Illinois Press, USA, 1966
- Weinreich D.M., Delaney N.F., DePristo M.A., Hart D.L. Darwinian evolution can follow only few mutational paths to fitter proteins. *Science.* 2006;312(5770):111-114. doi [10.1126/science.1123539](https://doi.org/10.1126/science.1123539)
- Xue K.S., Stevens-Ayers T., Campbell A.P., Englund J.A., Pergam S.A., Boeckh M., Bloom J.D. Parallel evolution of influenza across multiple spatiotemporal scales. *eLife.* 2017;6:e26875. doi [10.7554/eLife.26875](https://doi.org/10.7554/eLife.26875)

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

Received August 1, 2025. Revised October 16, 2025. Accepted October 27, 2025.

doi 10.18699/vjgb-26-33

Alembic: a framework for converting disparate biological data into structured resources

I.V. Bezdvornyykh[#], K.I. Yuditskiy[#], N.A. Cherkasov, A.A. Samsonova , A.A. Kanapin  

Institute for Translational Biomedicine, Saint Petersburg State University, St. Petersburg, Russia

 a.kanapin@gmail.com

Abstract. The imperative to re-analyze existing public sequencing data is central to modern biology, driven by new hypotheses and advanced analytical methods. However, this effort is critically hampered by the profound heterogeneity of repository data, particularly the non-standardized, free-text descriptions of biological experiments. This lack of structural and semantic homogeneity prevents systematic search, integration, and comparative analysis, effectively locking away the full potential of accumulated datasets. Advances in Natural Language Processing (NLP) offer a pivotal pathway to overcome this bottleneck by transforming unstructured text into computable, homogeneous information. The integrated Entrez database system, maintained by the National Center for Biotechnology Information (NCBI), provides sophisticated programmatic access via an API to primary sequencing data and its associated metadata, including detailed experimental descriptions. This interface enables researchers to identify and retrieve relevant data through keyword searches, including those based on gene names, and to apply modern NLP techniques to transform textual metadata into structured information. The output is formatted data ready for integration into local databases, accompanied by a systematic list of links for downloading primary files. The Alembic software package offers a comprehensive and automated solution for the entire workflow. Designed as a locally deployable client-server system, Alembic incorporates state-of-the-art transformer-based AI algorithms for analyzing the biomedical text that accompanies sequencing data. Its core utilizes the openly available AIONER platform, which is built upon the PubMedBERT model trained on the PubMed repository, to ensure efficient and accurate recognition of biomedical named entities (e.g., genes, diseases). This provides users with structured and meaningful keyword search results. By delivering a curated list of datasets, Alembic streamlines the path from search to analysis. Researchers can efficiently identify high-value targets and obtain a complete package of metadata and primary data to construct a tailored local repository. This positions Alembic as a universal solution that overcomes the fragmented approach of existing tools, offering an integrated workflow for diverse public sequencing data.

Key words: natural language processing; biomedical text mining; semantic annotation; data harmonization; omics data integration

For citation: Bezdvornyykh I.V., Yuditskiy K.I., Cherkasov N.A., Samsonova A.A., Kanapin A.A. Alembic: a framework for converting disparate biological data into structured resources. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):293-298. doi 10.18699/vjgb-26-33

Funding. This work was supported by Russian Science Foundation grant No. 23-14-00134.

Alembic: от разрозненных биологических данных к структурированным ресурсам

И.В. Бездворных[#], К.И. Юдыцкий[#], Н.А. Черкасов, А.А. Самсонова , А.А. Канапин  

Институт трансляционной биомедицины, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

 a.kanapin@gmail.com

Аннотация. Развитие технологий высокопроизводительного секвенирования и методов анализа больших данных создает устойчивую потребность в повторном анализе накопленной в открытых репозиториях гетерогенной информации. Серьезной проблемой при этом остается преобладание свободного текстового описания биологических экспериментов, что затрудняет продуктивный поиск, систематизацию и дальнейшее использование соответствующих наборов данных. Прогресс в области искусственного интеллекта, особенно в развитии методов обработки естественного языка (natural language processing, NLP), обуславливает новые методологические возможности для эффективного решения этой задачи. Интегрированная система баз данных Entrez, поддерживаемая Национальным центром биотехнологической информации США (NCBI), предоставляет развитый и надежный доступ как к исходным данным секвенирования, так и к сопутствующей метаинформации, включающей детальное описание параметров экспериментов, через программный интерфейс (application programming interface, API). Это позволяет идентифицировать и загружать данные секвенирования и соответствующие им метаданные с описаниями экспериментов, используя поиск по ключевым словам и

различным терминам, таким, например, как имена генов, в репозиториях; преобразовывать и систематизировать текстовые описания с применением современных NLP-методов и обеспечивать исследователям структурированную информацию для интеграции в локальные базы данных и форматированный перечень ссылок для загрузки исходных данных. Программный пакет Alembic предлагает комплексное решение для поиска и загрузки данных, автоматизируя все указанные этапы. Платформа использует клиент-серверную архитектуру и предназначена для локальной установки. Для анализа биомедицинских текстов, сопровождающих данные секвенирования, в Alembic интегрированы современные алгоритмы искусственного интеллекта на основе архитектуры трансформеров. В частности, используется имеющаяся в открытом доступе платформа AIONER, обученная на данных репозитория PubMed с помощью модели PubMedBERT. Такой подход обеспечивает эффективное распознавание именованных сущностей (named entity recognition, NER) биомедицинского характера (гены, заболевания и др.), предоставляя пользователю структурированные результаты поиска по ключевым словам. Формируемый пакетом список дает возможность исследователю анализировать результаты, отбирать наиболее релевантные наборы данных и получать всю необходимую информацию (включая исходные данные) для создания локального репозитория, ориентированного на конкретную исследовательскую задачу. В отличие от имеющихся аналогов, Alembic является универсальным решением для интеграции данных из репозиториях открытого доступа и работы с разнородными типами данных секвенирования.

Ключевые слова: обработка естественных языков; анализ биомедицинских текстов; семантическая аннотация; гармонизация данных; интеграция омиксных данных

Introduction

The rapid accumulation of sequencing data from multiple modalities, including WGS, RNASeq, and BSSeq, as well as related publications, is driven by advances in high-throughput sequencing technologies. The analysis of these extensive datasets calls for specialized software systems that can automatically and efficiently extract biologically relevant information. One promising approach involves the application of Natural Language Processing (NLP) algorithms. Modern NLP techniques automate the analysis of scientific texts, significantly enhancing the efficiency of large-scale data processing. A notable example is the MetaMap system (Aronson, Lang, 2010), which utilizes Named Entity Recognition (NER) to identify concepts from the Unified Medical Language System (UMLS) metathesaurus within textual sources. Building upon this foundation, specialized tools such as the BIONER package have further extended NER capabilities for biomedical entity recognition (Wang et al., 2019).

Building on the foundational role of NER in biomedical text extraction, advancements in neural network architectures and word embedding techniques have markedly improved the accuracy of biomedical entity recognition. The introduction of context-sensitive models, such as ELMo (Embeddings from Language Models), enabled the generation of word representations that incorporate contextual information. Recent progress in the field is characterized by the dominance of transformer architectures, particularly BERT (Devlin et al., 2019), which have transformed approaches to NLP tasks. Adapting these architectures for biomedical applications has spurred the development of domain-specific models, including BioBERT and PubMedBERT (Lee et al., 2020), as well as specialized processing frameworks such as scispaCy (Neumann et al., 2019).

The most extensive integrated system of biomedical databases remains Entrez, developed by the US National Center for Biotechnology Information (NCBI). This platform consolidates 38 distinct repositories, including PubMed, PMC, and databases of nucleotide and protein sequences. Data are added to the system through both automated processes, such as the analysis of scientific publications, and manual

submissions by researchers. Depositing sequencing data into public archives (e. g., SRA, ENA) requires ones to provide essential metadata describing their experiments and biological samples. A principal advantage of Entrez is its programmatic interface, the Entrez Programming Utilities (E-utilities) (Sayers, 2022), which supports advanced querying and automated data extraction.

A common challenge in genomics involves identifying relevant raw datasets from public repositories. While programmatic access to resources like Entrez enables data retrieval, the integration of contemporary NLP for analyzing the associated biomedical text is key to building a unified system for this task. Existing tools, such as iSeq (Chao et al., 2024) and SampleExplorer (Chin, Lassman, 2024), have a narrow focus and offer limited functionality in this regard. To address this gap, we present Alembic, a system designed to analyze data within NCBI's public repositories and perform the structured extraction of targeted information. Its name reflects its core function: much like the distillation apparatus used to extract essential substances from raw materials, Alembic refines meaningful insights from large volumes of primary data.

Materials and methods

Alembic employs a client-server architecture, where a client module handles query construction and result visualization, and a server module is responsible for processing the request and extracting the named entity. Figure 1 illustrates the overall system module design and data processing workflow.

The system operates on data from the NCBI Short Read Archive (SRA), accessed through the public Entrez API, allowing users to search using a broad spectrum of biomedical terms. Upon query execution, the AlembicDump module processes the results through a sequential pipeline: it first extracts NCBI Universal Identifiers (UIDs), uses them to fetch the full corresponding metadata in XML format, and then standardizes and restructures this information into a tabular format with predefined columns. This final table adheres to the Entrez SRA metadata schema, which comprises 21 fields containing

either free-text descriptions (e. g., Study Abstract) or formally structured identifiers (e. g., Sample Accession).

The structured data from AlembicDump are then analyzed by the AlembicNLP algorithm to perform named entity recognition. First, the input text undergoes preprocessing with sciSpaCy (v 2.0.18), involving conversion to lowercase, removal of special characters and punctuation, filtering of isolated digits and stop-words (using the NLTK dictionary, www.nltk.org, v 3.8.1), lemmatization, and tokenization. The processed text is then analyzed by the core AIONER model (Luo et al., 2023), a Bioformer-based architecture.

Alembic implements this using the embedded, open-source AIONER code with the bioformer-cased-v1.0 and BiomedNLP-PubMedBERT-base-uncased-abstract model variants, which were pre-trained on PubMed abstracts (<https://huggingface.co/lingbionlp/AIONER-0415/tree/main>). These models were trained on information available in the article abstracts deposited in the PubMed catalog (Luo et al., 2023). The recognition stage outputs entities that are automatically categorized into predefined classes: Gene, Disease, Species, Cell Line, Variant, and Chemical. Consequently, AlembicNLP produces its final output: an annotated list that classifies each entity and records its textual coordinates.

The front-end application is implemented on the Vite JS framework (v 7.0) to optimize the build process. Its component-based architecture is developed in React (v 19), while the UI layer employs Material UI (v 7) components complemented by the MUI X DataGrid for advanced data table features. Asynchronous API communication is facilitated by the Axios library (v 1.6.7).

The application implements a clear workflow through two dedicated modules. The workflow begins with the Search Interface Module, where users query biomedical terms. This module returns results in a consolidated, paginated table optimized for large datasets.

The user then selects relevant fields, triggering the Data Preparation Module. This module automatically extracts and categorizes key terms for AI-driven analysis. This entire process is enabled by a client-server architecture: all complex data processing occurs on the server, while the client provides a clean, structured interface for interaction and presentation. This separation streamlines the analysis of complex descriptions and significantly reduces the user's cognitive load.

The system presents an interactive dashboard for grouped results, equipped with hierarchical term filters, a faceted quick-search field, and dynamic highlight-on-hover functionality. These tools enable drill-down analysis and rapid identification of patterns within large datasets, eliminating the need to manually parse complete experimental descriptions. The interface implements a fully responsive design, ensuring optimal usability and consistent functionality across desktop, tablet, and mobile platforms.

A researcher using Alembic typically follows a three-stage workflow: beginning with query formulation, followed by metadata structuring, and concluding with entity extraction. For instance, a search for Alzheimer's disease transcriptomic data starts by entering keywords such as "Alzheimer", "disease", and "RNASeq" into the dedicated search interface

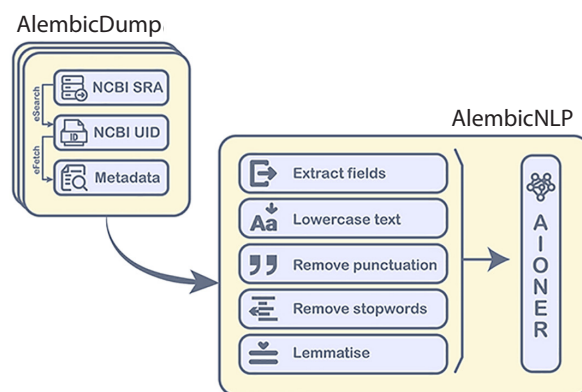


Fig. 1. Schematic diagram of the Alembic package, comprising the AlembicDump (search result preprocessing) and AlembicNLP (named entity extraction) modules.

(Fig. 2). The system returns the number of matching SRA records and, if results exceed 1,000 entries, prompts the user to refine terms or accept the default limit.

In the second stage, Alembic structures the retrieved metadata into a table for review (Fig. 3). The table presents standard Entrez metadata field names, the percentage of experiments containing each field, and sample data entries. Using checkboxes in the first column, users can select multiple metadata types for subsequent AI-powered analysis.

Upon clicking the CONTINUE button, the system executes the NLP pipeline and displays the final output (Fig. 4). The results in this example are presented as a word cloud, visualizing the frequency of recognized entities. The interface includes download links in the upper-right corner for two tab-separated value (TSV) files: RESULTS.TSV (containing data prior to AI processing) and AI_PROCESSED.TSV (containing the post-processed results). These structured files are suitable for direct import into standard relational databases, including SQLite, MySQL, or PostgreSQL. A separate file, DOWNLOADS.TXT, supplies direct URLs for retrieving the corresponding raw sequencing data from the SRA archive. The page concludes with a detailed listing of each experiment and its associated downloadable sequencing samples.

The project's front end is built using modular JavaScript. Its architecture prioritizes scalability, allowing new functional modules to be added without modification of the core system. All source code and comprehensive documentation are publicly accessible under the MIT license via the project's GitHub repository: <https://github.com/shitohana/Alembic>.

A local installation of Alembic requires Python v 3.12 or later and Docker v 24.0.2 or later. The setup procedure automatically installs all necessary dependencies, such as the TensorFlow (v 2.3.0), Transformers (v 4.18.0), and stanza (v 1.4.0) packages required for the integrated AIONER model. Once installation is complete, the system becomes accessible through a local web interface.

Results and discussion

The increasing scale of biomedical text processing has necessitated the development of numerous named entity recogni-

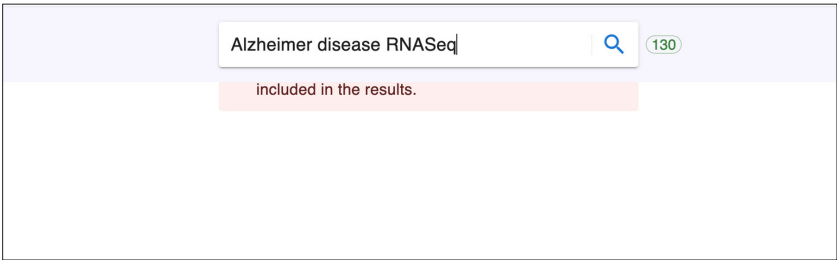


Fig. 2. The Alembic search interface.

Alzheimer disease RNASeq			
130			
Results found: 130.			
For further processing using AI, select columns for analysis or use default settings.			
CONTINUE			
	Column Title	Filled %	Rows Data sample
☐	Experiment Accession	100%	SRX22148999(1) SRX22149000(1) SRX22149001(1)
☐	Experiment Alias	100%	GSM7849611_r1(1) GSM7849612_r1(1) GSM7849613_r1(1)
☐	Experiment Title	100%	GSM7849611: 205_Neurons_VEH_A; Homo sapiens
☐	Experiment ID	100%	SRX22148999(1) SRX22149000(1) SRX22149001(1)
☐	Study Accession	100%	SRP46740(99+) SRP042143(10)
☒	Study Title	100%	Restoring hippocampal glucose metabolism.(99+) RN
☒	Study Abstract	100%	Impaired cerebral glucose metabolism is a.(99+) In th
☐	Sample Accession	100%	SRS19208683(1) SRS19208684(1) SRS19208685(1)
☐	Sample Alias	100%	GSM7849611(1) GSM7849612(1) GSM7849613(1)
☐	Scientific Name	100%	Homo sapien.(99+)
☐	Taxon ID	100%	960(99+)
☐	Library Name	100%	GSM7849611(1) GSM7849612(1) GSM7849613(1)
☐	Library Strategy	100%	RNA-Se.(99+)
☐	Library Source	100%	TRANSCRIPTOMIC(99+)
☐	Library Selection	100%	cDN/(99+) RANDOM(10)
☐	Run Accession	100%	SRR26444415(1) SRR26444414(1) SRR26444413(1)
☐	Run Alias	100%	GSM7849611_r1(1) GSM7849612_r1(1) GSM7849613_r1(1)

Fig. 3. Aggregated, tabular view of the results returned from a keyword search.

tion (NER) systems. Leading solutions in this area include scispaCy and AIONER. scispaCy provides models pretrained on biomedical corpora such as GENIA and MedMentions, enabling efficient handling of essential NLP tasks, including tokenisation, part-of-speech (POS) tagging, syntactic parsing, and NER, while accounting for domain-specific linguistic features. POS tagging, a foundational NLP technique, assigns grammatical categories (e. g., noun, verb, adjective) to each word in a text.

In contrast, AIONER specialises exclusively in biomedical NER (BioNER) through an innovative All-in-One (AIO) ap-

proach, which simultaneously recognises terms drawn from multiple annotated datasets. Unlike conventional methods that label tokens separately per category (e. g., Gene, Disease), the AIO architecture processes all categories concurrently using specially designed boundary tags. This reduces attribution errors for individual words or compound expressions, improves tokenisation accuracy, and addresses common challenges such as overfitting and poor generalisation, often stemming from sparsely annotated biomedical data. Built on modern pretrained language models like PubMedBERT and enhanced with a conditional random fields (CRF) layer for precise entity boundary

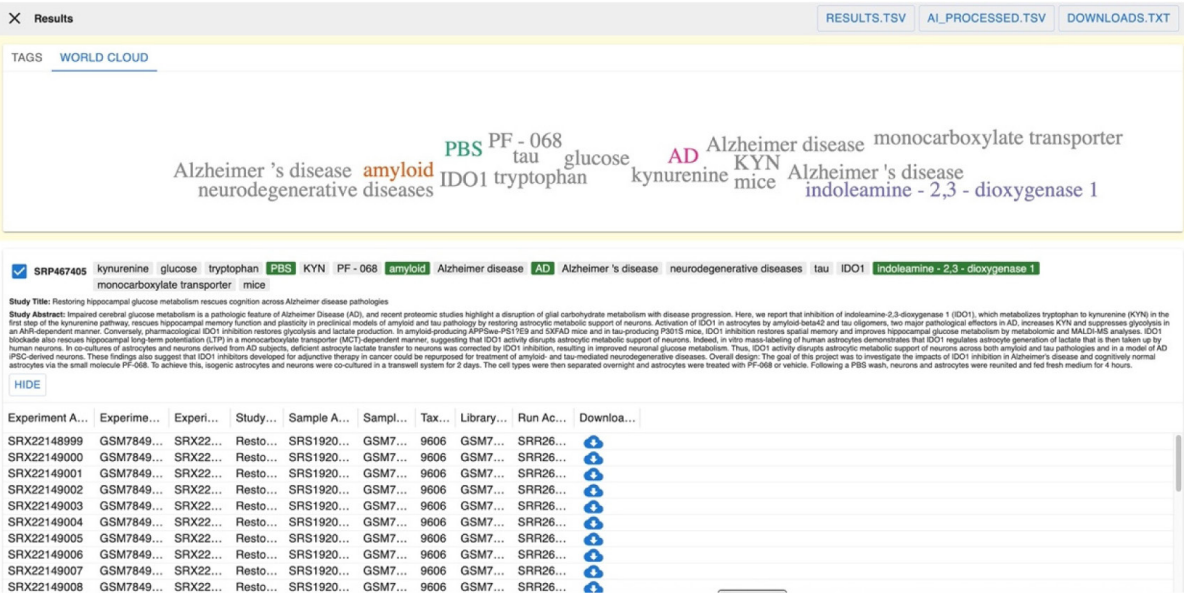


Fig. 4. Visualization of named entities identified by the PubMedBERT model, displayed as a word cloud where term size corresponds to frequency.

detection, AIONER was selected as the core NER engine for the Alembic package due to these combined strengths.

The growing volume of biomedical text has driven the development of numerous named entity recognition (NER) systems, with scispaCy and AIONER being leading examples. scispaCy offers models pretrained on biomedical corpora, such as GENIA and MedMentions. These models efficiently perform essential natural language processing tasks, namely tokenisation, part-of-speech (POS) tagging, syntactic parsing, and NER, while incorporating domain-specific linguistic features. Here, POS tagging refers to the foundational NLP technique of assigning a grammatical category (e.g., noun, verb, adjective) to each word in a text.

While tools like iSeq and SampleExplorer retrieve data from public repositories, Alembic introduces a fundamentally different approach grounded in natural language processing. The distinction reflects their core operational philosophies: iSeq functions as an automated data retrieval utility, whereas Alembic is designed for semantic analysis and entity extraction. iSeq operates via the command line to download batches of NGS data from major repositories, such as GSA, SRA, ENA, and DDBJ, using known accession numbers. Although it automates the download process and verifies data integrity, it performs no semantic analysis of metadata, relying entirely on pre-defined lists of identifiers. This approach treats data fundamentally as files to be transferred rather than information to be understood. SampleExplorer adopts a different strategy, employing language models to query repositories such as ARCHS4 for transcriptomic data. Its algorithm identifies semantically similar experiments by combining vectorised text metadata with transcriptomic similarity scores, using genes or textual descriptions as search queries. This design optimises discovery at the experiment level, enabling researchers to find analogous samples and related studies.

However, it does not parse individual descriptions to extract and classify specific biomedical entities – such as genes, variants, and chemicals – which form the core function of Alembic’s NLP pipeline.

Unlike iSeq and SampleExplorer, Alembic provides a fundamentally distinct solution through an integrated NLP stack purpose-built for biomedicine. It utilises established preprocessing frameworks, notably scispaCy, and incorporates transformer architectures, including BERT and BioBERT. Its primary advantage, however, derives from implementing PubMedBERT (a model further trained on the entire PubMed/PMC corpus) within the AIONER architecture (Luo et al., 2023). This specialised training enables Alembic to substantially outperform generic models, achieving superior accuracy in recognising complex biomedical terminology and interpreting its contextual meaning, thereby transforming unstructured text into structured, actionable knowledge.

Conclusion

Alembic enables systematic extraction and structuring of biomedical metadata from open repositories through deep NLP-driven analysis. The system generates import-ready metadata packages compatible with SQLite and PostgreSQL, automates retrieval of linked raw sequencing files, and supports on-demand generation of specialized, project-specific local databases. The integrated AIONER platform leverages state-of-the-art language models, particularly PubMedBERT pretrained on the PubMed/PMC corpus, to accurately identify and classify biomedical entities with higher precision than alternative tools. Unlike command-line utilities that require manual query formulation, Alembic’s graphical interface and interactive visualizations streamline data discovery, reducing operational friction and enabling researchers to focus on analysis rather than data wrangling.

By converting unstructured metadata into structured, machine-actionable knowledge, Alembic substantially reduces the time and expertise required to prepare public repository data for downstream analysis.

References

- Aronson A.R., Lang F.M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17(3): 229-236. doi 10.1136/jamia.2009.002733
- Chao H., Li Z., Chen D., Chen M. iSeq: an integrated tool to fetch public sequencing data. *Bioinformatics.* 2024;40(11):btac641. doi 10.1093/bioinformatics/btac641
- Chin W.L., Lassmann T. SampleExplorer: using language models to discover relevant transcriptome data. *Bioinformatics.* 2024;41(1): btac759. doi 10.1093/bioinformatics/btac759
- Devlin J., Chang M.W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv.* 2019. doi 10.48550/arXiv.1810.04805
- Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234-1240. doi 10.1093/bioinformatics/btz682
- Luo L., Wei C.-H., Lai P.-T., Leaman R., Chen Q., Lu Z. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics.* 2023;39(5):btad310. doi 10.1093/bioinformatics/btad310
- Neumann M., King D., Beltagy I., Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, 2019;319-327. doi 10.18653/v1/W19-5034
- Sayers E. The E-utilities in-depth: parameters, syntax and more. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US), 2022. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK25499/>. Accessed: Jul. 30, 2025
- Wang X., Zhang Y., Ren X., Zhang Y., Zitnik M., Shang J., Langlotz C., Han J. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics.* 2019;35(10):1745-1752. doi 10.1093/bioinformatics/bty869

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

Received August 6, 2025. Revised October 21, 2025. Accepted November 6, 2025.

doi 10.18699/vjgb-26-34

Development and validation of the PipeSeq program for RNA-seq data analysis in the *Chlamydomonas reinhardtii* as a model

A.M. Nerezenko , P.A. Virolainen , S.A. Tupitsyna , E.M. Chekunova 

Saint-Petersburg University, St. Petersburg, Russia

 alexnerezenko@gmail.com

Abstract. RNA sequencing (RNA-seq) is a highly sensitive method for transcriptome analysis that allows simultaneous assessment of expression of thousands of genes and identification of expression patterns under various conditions. The existing variety of RNA-seq data formats, normalization methods, and approaches to statistical processing of results complicates comparison of data from different studies and reduces reproducibility of the analysis. This study presents an automated pipeline PipeSeq that combines standard steps of RNA-seq data processing: loading (SRA Toolkit), read alignment to the reference genome (HISAT2), transcript assembly (StringTie), transcript counting (FeatureCounts) and statistical analysis of differential gene expression under various experimental conditions (DESeq2). PipeSeq has a simple visual interface, supports multithreading, and generates ready-to-analyze gene expression heat maps, tables and graphs. The functionality of the pipeline is demonstrated on three sets of raw RNA-seq data from the green alga *Chlamydomonas reinhardtii* cells available in the NCBI SRA database. The data from these experiments were used to analyze the differential expression of *C. reinhardtii* genes encoding the GATA family transcription factors under different light cultivation conditions. The data obtained by *in silico* methods were verified by real-time reverse transcription polymerase chain reaction (RT-qPCR) for 12 GATA genes, which allowed us to hypothesize their functions and evaluate the correlation between the bulk (RNA-seq) and targeted (RT-qPCR) approaches. Our results showed that RNA-seq and RT-qPCR methods reveal similar directions of gene expression changes, but demonstrate differences in the effect size and sensitivity, which emphasizes the need for a combined use of the two approaches. Thus, the PipeSeq program is a tool for conducting a full cycle of bioinformatic analysis of RNA-seq data, additionally providing the opportunity to process RT-qPCR data and perform a comparative statistical analysis of the results obtained.

Key words: RNA sequencing; RNA-seq; RT-qPCR; pipeline; transcriptome; gene expression; GATA family transcription factors; GATA TFs; *Chlamydomonas reinhardtii*

For citation: Nerezenko A.M., Virolainen P.A., Tupitsyna S.A., Chekunova E.M. Development and validation of the PipeSeq program for RNA-seq data analysis in the *Chlamydomonas reinhardtii* as a model. *Vavilovskii Zhurnal Genetiki i Selektcii* = *Vavilov J Genet Breed*. 2026;30(2):299-310. doi 10.18699/vjgb-26-34

Funding. The authors acknowledge Saint-Petersburg State University for a research project 124032000041-1.

Разработка и валидация программы PipeSeq для анализа данных секвенирования РНК на модели *Chlamydomonas reinhardtii*

А.М. Нерезенко , П.А. Виролайнен , С.А. Тупицына , Е.М. Чекунова 

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

 alexnerezenko@gmail.com

Аннотация. Секвенирование РНК (РНК-сек) – высокочувствительный метод анализа транскриптома, позволяющий одновременно оценивать экспрессию тысяч генов и выявлять паттерны экспрессии в различных условиях. Существующее разнообразие форматов данных РНК-сек, методов нормализации и подходов к статистической обработке результатов затрудняет сопоставление данных разных исследований и снижает воспроизводимость анализа. В настоящей работе представлен автоматизированный пайплайн PipeSeq, объединяющий стандартные этапы обработки данных РНК-сек – от загрузки (SRA Toolkit), выравнивания прочтений на референсный геном (HISAT2) и сборки транскриптов (StringTie) до подсчета транскриптов (FeatureCounts) и статистического анализа дифференциальной экспрессии генов в различных экспериментальных условиях (DESeq2). Программа PipeSeq имеет простой визуальный интерфейс, поддерживает многопоточность и формирует готовые для анализа тепловые карты экспрессии генов и отчеты в форме таблиц и графиков. Функциональность пайплайна продемонстрирована на трех наборах пакетов сырых данных секвенирования РНК клеток зеленой водоросли

Chlamydomonas reinhardtii, доступных в открытой базе данных NCBI SRA. Результаты этих экспериментов были использованы для анализа дифференциальной экспрессии генов *C. reinhardtii*, кодирующих факторы транскрипции семейства GATA, в различных световых условиях культивирования. Полученные методами *in silico* данные верифицированы методом полимеразной цепной реакции в реальном времени с обратной транскрипцией (ОТ-ПЦР-РВ) по 12 генам GATA, что позволило выдвинуть предположения об их функциях, а также оценить степень согласованности между массовым (РНК-сек) и таргетным (ОТ-ПЦР-РВ) подходами. Результаты нашего исследования показали, что методы секвенирования РНК и ОТ-ПЦР-РВ выявляют схожие направления изменения экспрессии генов, но демонстрируют различия по оценке степени размера эффекта и чувствительности, что подчеркивает необходимость совместного применения двух подходов. Таким образом, программа PipeSeq представляет собой инструмент для проведения полного цикла биоинформатического анализа данных РНК-сек, дает возможность обрабатывать данные ОТ-ПЦР-РВ и выполнять сравнительный статистический анализ полученных результатов.

Ключевые слова: секвенирование РНК; РНК-сек; ОТ-ПЦР-РВ; пайплайн; транскриптом; экспрессия генов; факторы транскрипции семейства GATA; ОТ GATA; *Chlamydomonas reinhardtii*

Introduction

In recent years, RNA sequencing (RNA-seq) has become widely adopted as a reliable approach for large-scale quantitative analysis of gene expression across diverse biological systems (Marioni et al., 2008; Mortazavi et al., 2008; Conesa et al., 2016; Li X., Wang, 2021). Due to the “digital” nature of the data, RNA-seq enables direct calculation of the number of reads for each transcript, providing a wide dynamic measurement range and high reproducibility of experimental results. In addition, this method makes it possible to detect previously unannotated transcripts and alternative splicing variants (Wang et al., 2009; Li X., Wang, 2021). These features establish RNA sequencing as a powerful tool for systems-level studies of the transcriptome, enabling reliable quantification of transcript levels under various experimental conditions and identification of differential gene expression in the study object (Wang et al., 2009).

The processing, accumulation and consolidation of these data can contribute to obtaining novel insights into the functioning of living systems. To implement this approach, it is essential to address the challenge of standardizing the processing of raw RNA-seq data (Conesa et al., 2016; Li X., Wang, 2021). Currently, outdated normalization methods are still employed for gene expression assessment. These methods are characterized by low reproducibility, fail to adequately correct for compositional biases, and hinder direct comparison of results across independent studies. Such methods include: RPKM (Reads Per Kilobase per Million mapped reads), FPKM (Fragments Per Kilobase of exon model per Million mapped fragments), and TPM (Transcripts Per Million). At present, methods based on the negative binomial distribution – DESeq2 and edgeR – are considered the standard for differential gene expression analysis. Normalized counts generated by DESeq2 have demonstrated the lowest coefficient of variation and highest reproducibility (Zhao S. et al., 2020; Zhao Y. et al., 2021; Elahimanesh, Najafi, 2024).

The existing diversity of RNA-seq data processing methods, formats, and representation approaches across various databases complicates the consolidation of information for comprehensive analysis. To address this issue, there is a need to develop integrated solutions capable of minimizing manual

labor, ensuring a high degree of reproducibility of results and ease of use for biologists without special training in information technology (Conesa et al., 2016).

Comparative analysis of RNA-seq data processing results and molecular biology data is complicated by several methodological challenges. Although RNA sequencing is recognized as a reliable method for global expression profiling, the results obtained through this technique typically represent relative changes in transcript levels across the entire genome. Historically, reverse transcription quantitative polymerase chain reaction (RT-qPCR) has been established as the “gold standard” for validating data derived from transcriptomic studies (Derveaux et al., 2010; Coenye, 2021). However, RT-qPCR-based methods and RNA sequencing technology rely on different protocols, which complicates their direct quantitative comparison. Several factors can influence the obtained results, including: efficiency of reverse transcription and amplification processes, data normalization methods, sensitivity of the techniques employed.

Currently, RNA-seq is utilized to cover the entire transcriptome of the study object and identify a set of differentially expressed candidate genes in response to a specific stimulus. In contrast, RT-qPCR is applied for precise quantitative assessment of these changes, specifically targeting a limited pool of genes of interest (Shi, He, 2014; He et al., 2015; Coenye, 2021).

To understand the mechanisms of metabolic regulation, special attention is devoted to the study of transcription factors (TFs) as key regulators of gene activity. TFs are proteins capable of binding to specific DNA sequences in promoter regions, thereby enhancing or repressing the transcription of target genes. These regulators coordinate gene activity in response to various environmental changes and signaling influences, participating in global processes such as growth, development, and adaptation to stress factors (Riechmann et al., 2000).

In photosynthetic organisms, genes encoding GATA family transcription factors are of particular interest. GATA TFs are proteins that carry a conserved zinc finger domain of type IV (general formula: CX₂CX_{18–20}CX₂C). This domain mediates binding to the consensus sequence (A/T)GATA(A/G) in the

promoters of target genes (Reyes et al., 2004). Plant GATA factors are involved in the regulation of photomorphogenesis, nitrogen and carbon metabolism, and hormonal control (Manfield et al., 2007; Naito et al., 2007; Luo et al., 2010; Schwechheimer et al., 2022; Schröder et al., 2023; Ren et al., 2025). In recent years, there has been a significant increase in scientific interest in studying GATA TFs in bryophytes and algae, as their biological functions and evolutionary roles remain insufficiently characterized (Schwechheimer et al., 2022; Virolainen, Chekunova, 2024).

The aim of the present study is to develop an integrated solution for standardizing the processing of raw RNA sequencing data and RT-qPCR results.

To achieve this goal, the following objectives must be addressed.

1. Developing an automated pipeline for conducting a full cycle of RNA-seq data analysis.
2. Testing the pipeline on raw RNA-seq data publicly available in the NCBI SRA database, using as a model the genes encoding GATA family TFs in a model object of photosynthesis genetics, the green alga *Chlamydomonas reinhardtii*, in response to changes in light growing conditions.
3. Conducting an analysis of GATA gene expression in *C. reinhardtii* using the RT-qPCR method in response to changes in light conditions during growth.
4. Performing a comparative statistical analysis of the obtained RNA-seq and RT-qPCR data.

Materials and methods

Quantitative analysis of gene expression using the PipeSeq program. To analyze the expression profile of 12 genes encoding GATA TFs in *C. reinhardtii* under various cultivation conditions by *in silico* methods, the following publicly available raw RNA-seq data packages were selected from the NCBI SRA database (Wheeler et al., 2005): SRX8380269, SRX8380270, SRX8380271 (acclimation to high light, 600 $\mu\text{mol}/\text{m}^2/\text{s}$, 1 h), SRX7413406, SRX7413407, SRX7413412, SRX7413413, SRX7413414, SRX7413415 (acclimation to light, 30 min), SRX5120530, SRX5120531, SRX5120532, SRX5120533, SRX5120534, SRX5120535 (acclimation to darkness).

The three selected datasets (PRJNA634446, PRJNA596622, PRJNA509798) include both experimental (altered growth conditions) and control (standard growth conditions) RNA-seq data for the wild-type strain CC-124 (*wt*, *mt*-) of *C. reinhardtii* performed in triplicate biological and technical replicates. For read alignment, the *C. reinhardtii* genome assembly (v5.6) (Merchant et al., 2007) and its annotation in GTF format were downloaded from the Phytozome portal (Goodstein et al., 2012).

Data upload and processing were carried out using PipeSeq, an automated software package consisting of 17 scripts and more than 2,000 lines of code. The tool is designed for system analysis of transcriptomic data obtained via RNA-seq, processing of RT-qPCR data using the $\Delta\Delta\text{Ct}$ method (Livak, Schmittgen, 2001; Schmittgen, Livak, 2008), and comparative analysis of results (including calculation of Pearson, Spearman, and Kendall correlation coefficients). The program was

developed in Python (version 3.9) using Pandas (McKinney, 2011), Matplotlib (Hunter, 2007), PyQt6 (<https://www.riverbankcomputing.com/software/pyqt/>), PyDESeq2 (Muzellec et al., 2023) libraries, and bioinformatics tools SRA Toolkit, FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), MultiQC (Ewels et al., 2016), Cutadapt (Martin, 2011), HISAT2 (Kim et al., 2019), SAMtools (Li H. et al., 2009), FeatureCounts (Liao et al., 2014), StringTie (Pertea et al., 2015), DESeq2 (Love et al., 2014).

Strains and cultivation conditions. The wild-type strain CC-124 (*wt*, *mt*-) of *C. reinhardtii* from the Peterhof Genetic Collection at St. Petersburg State University (Kvitko et al., 1983) was grown in Petri dishes on agarized TAP medium (1.5 %) (Harris, 1989), supplemented with arginine (50 mg per liter of medium) and yeast autolysate (4 g per liter of medium) at a temperature of 20–25 °C with a day (14 h, illumination 90 $\mu\text{mol}/\text{m}^2/\text{s}$) / night (10 h) cycle and subculturing every three days. Culture samples were grown and collected under standard illumination conditions (90 $\mu\text{mol}/\text{m}^2/\text{s}$) and darkness (control conditions), under conditions of increased illumination (high light, 215 $\mu\text{mol}/\text{m}^2/\text{s}$) for 30 min and 2 h (transfer of cultures from standard conditions), under conditions of standard illumination for 30 min and 2 h (transfer of cultures from darkness), under darkness for 30 min and 2 h (transfer of cultures from standard conditions) (experimental conditions).

Culture fixation and RNA isolation. Fixation of *C. reinhardtii* cell cultures grown under control and experimental conditions and total RNA isolation was performed using the “ExtractRNA” reagent (Evrogen, Russia) in strict accordance with the manufacturer’s protocol. RNA preparations were treated with DNase I (Thermo Fisher Scientific, USA) to remove genomic DNA contamination, and subsequently ethanol-precipitated. The concentration of total RNA was measured using an Eppendorf BioPhotometer plus spectrophotometer (Eppendorf, Germany).

Primer design. Gene-specific primers were designed using the IDT PrimerQuest Tool (<https://www.idtdna.com/pages/tools/primerquest>) and NCBI Primer-BLAST (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>). Primer design criteria included positioning of at least one primer at an exon–exon junction, or ensuring that an intron sequence was located between primer binding sites. The constitutively expressed genes *RPL19* (ribosomal protein L19) and *RPL32* (ribosomal protein L32) of *C. reinhardtii* were used as reference genes for data normalization (Liu et al., 2012). The *RBCS* (Ribulose biphosphate carboxylase small subunit) gene of *C. reinhardtii* was used as experimental condition change control (Sanchez-Tarre, Kiparissides, 2021). Primer sequences are listed in the Table.

Primer specificity assessment included melting curve analysis across a series of controls (no-template control, no-reverse-transcription control, positive control) and visualization of PCR products via gel electrophoresis (Derveaux et al., 2010). PCR efficiency was evaluated using the software of the QuantStudio 5 real-time PCR system (Thermo Fisher Scientific, USA).

Primer sequences for quantitative analysis of GATA gene expression

Gene	Identifier	Primer	Sequence (5'–3')	Reference
GATA-1	Cre01.g025050	F	GTTGTTTGGCGACCTCTTTGTG	The present study
		R	GATCAGCGGCGGCTATGTC	
GATA-2	Cre10.g435450	F	ACTACGACGAGCGGGAAGA	
		R	GCCTTCTCGCCATGTACTCC	
GATA-3	Cre08.g378800	F	GAGCTGGACGGGAACGAAAC	
		R	GTGCGGTGCCGAGTAGTTT	
GATA-4	Cre11.g467581	F	GATCCTATCACCACCAAGGTTGC	
		R	CCATGCCGCCCATGTTCA	
GATA-5	Cre03.g160600	F	TCACGGGACGACGACATCA	
		R	CGGGTGAAGAATATGCCACAGG	
GATA-6	Cre03.g160700	F	GAAAAGGCAGGACAAGTCCAAG	
		R	TGTGAGGCGGGATGAAGAT	
GATA-7	Cre05.g242600	F	AGGAGCAGCAGCAGCAATC	
		R	CTGGTTAGTGCGGCGGTATC	
GATA-8	Cre06.g266850	F	TGTGCAACGCATGTGGGATA	
		R	CGGTCTTGGCTGACACATAGTT	
GATA-9	Cre06.g266950	F	ACATCAGCGGCTGCGATAAT	
		R	CGCCTGAGCCACTTTCGG	
GATA-10	Cre07.g319701	F	TCCGCTGCTGCGTAGAGT	
		R	GCAAAGACATCCTCGTCGGC	
GATA-11	Cre08.g358532	F	TCAGCAACAGCCCTCACTTC	
		R	CGCTCAAACCACTTGACCTCTAT	
GATA-12	Cre08.g358534	F	TGTCAAGTGTTCCACGACAAGA	
		R	GCACCAGAACCACTCGCA	
RPL32	Cre06.g289550	F	CCCAACGGCTTCTGAAGTA	
		R	AAGCGACGGTTGTGCATCAT	
RPL19	Cre02.g075700	F	CCTGAAGAAGTACCGGACTC	Liu et al., 2012
		R	AACACGTTACCCTTGACCTTCA	
RBCS	Cre02.g120150	F	ACCCCGGTCAACAACAAGATG	Sanchez-Tarre, Kiparissides, 2021
		R	GTCGTAGTACAGGCAAGACACG	

Quantitative analysis of gene expression by RT-qPCR. RT-qPCR reactions were performed in a one-step format using the “OneTube RT-PCR SYBR” reagent kit (Evrogen, Russia), strictly following the manufacturer’s instructions. Amplification was carried out on a QuantStudio 5 real-time PCR system (Thermo Fisher Scientific, USA), with fluorescence reading during the elongation and melting steps. The thermal cycling protocol was as follows: 55 °C – 15 min, 95 °C – 1 min, then 50 cycles: 95 °C – 15 s, 62 °C – 20 s, 72 °C – 20 s,

melting: 55–95 °C with 0.5 °C increments. Each sample was analyzed in triplicate biological replicates (averaged) (Der-veaux et al., 2010). Data processing and visualization were performed using the $\Delta\Delta C_t$ method (Livak, Schmittgen, 2001; Schmittgen, Livak, 2008) in the developed PipeSeq program.

Comparative analysis of RNA-seq and RT-qPCR results. A comparative statistical analysis of the data obtained by different methods was carried out in the PipeSeq program.

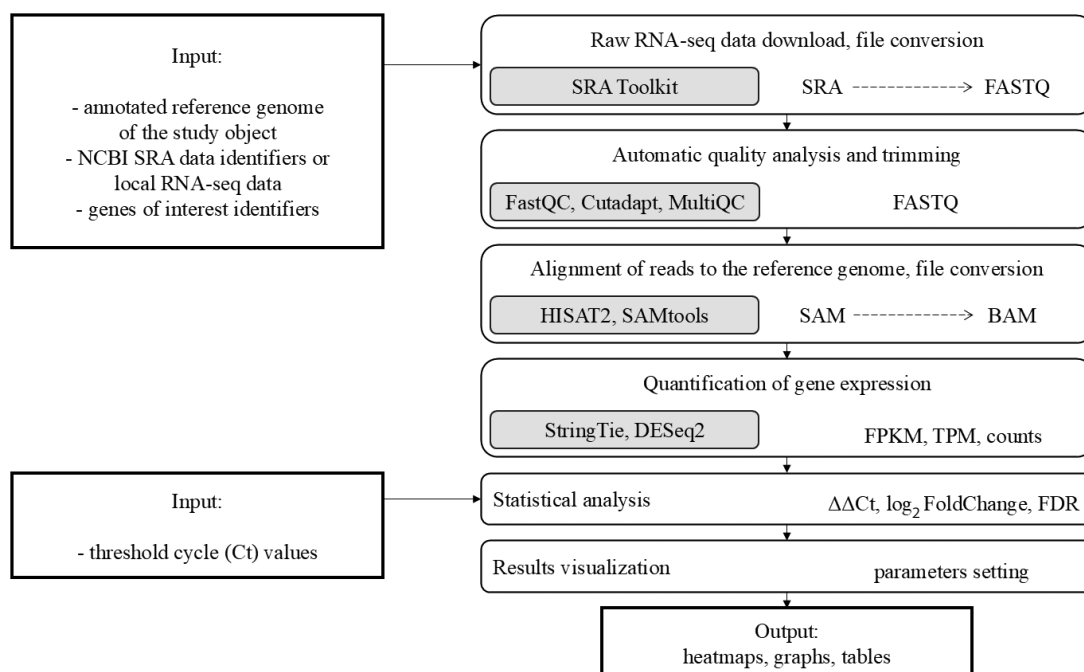


Fig. 1. Workflow diagram of the PipeSeq program.

Results

PipeSeq pipeline development

The aim of this work was to develop such a tool (pipeline) that would enable a full cycle of RNA-seq data analysis via a simple visual interface, requiring minimal user involvement in system administration. The pipeline is designed to run in a local Windows environment using the Windows Subsystem for Linux (WSL) to execute Linux commands. The workflow of the PipeSeq program is presented in Figure 1.

All steps of the algorithm are fully automated and allow processing large amounts of data through optimized commands and parallel computing. The input data are FASTQ files containing raw short sequencing reads obtained after conversion from the SRA format using the SRA Toolkit included in the software package. At the initial stage, automated preprocessing of reads is performed, which includes quality control (FastQC), removal of adapter sequences, low-quality nucleotides, and short reads (Cutadapt), aggregation of summary reports (MultiQC). Next, the reads are aligned to the reference genome using the HISAT2 tool (Kim et al., 2019), with the reference genome index being automatically created if it is absent. The advantages of HISAT2 include splicing consideration, accurate determination of transcript structure, and rapid analysis of alignment quality. After alignment, files are converted from the SAM format to the BAM format, sorted, and prepared for further analysis using the SAMtools (Li H. et al., 2009).

The developed PipeSeq program integrates three normalization approaches: DESeq2 (for differential expression analysis), TPM (absolute quantification), and FPKM (an outdated metric

included to ensure backward compatibility). At this stage, quantitative assessment of gene expression is performed by counting the number of reads mapping to genes and transcripts using the FeatureCounts package (Liao et al., 2014), as well as transcript assembly and calculation of FPKM and TPM values using the StringTie tool (Pertea et al., 2015). This tool analyzes read alignments, builds transcripts, and identifies exons, introns, and splice sites. The PipeSeq program provides the ability to disable the function of assembling new transcripts to minimize the likelihood of false results. These steps are performed considering various modes, such as strict annotation (the “-e” option in StringTie) and adjustable sensitivity (the “-c” parameter).

A package based on the negative binomial distribution, DESeq2 (Love et al., 2014), implemented in the PyDESeq2 library (Muzellec et al., 2023), is used for statistical analysis of differential gene expression. The PipeSeq pipeline automatically prepares input data for DESeq2, performs statistical calculations, and generates final tables with logarithmic changes in expression levels (log₂FoldChange) and corresponding values of the adjusted p-value – FDR (False Discovery Rate) using the Benjamini–Hochberg procedure. Statistical hypotheses about differential gene expression are tested based on a generalized linear model of the negative binomial distribution. After normalization for data counting and variance estimation (followed by averaging variance estimates across all genes using Bayesian shrinkage), the Wald test is applied.

The program also implements the possibility of processing RT-qPCR data using the ΔΔCt method (Livak, Schmittgen, 2001; Schmittgen, Livak, 2008) and conducting comparative statistical analysis with the three data normalization methods

used in RNA-seq (Pearson's linear correlation coefficient, Spearman's rank correlation coefficient, and Kendall's rank correlation coefficient).

As output, the PipeSeq program generates heatmaps of expression reflecting changes in the expression levels of genes of interest under different experimental conditions, summary tables with $\log_2$ FoldChange values, and graphs. A wide range of data display settings is available to the user. The program is available for download at the following link: <https://github.com/MarvinMarss/PipeSeq>.

Processing of available RNA sequencing data of *C. reinhardtii* cells using the PipeSeq pipeline

Using the developed program, we analyzed three available RNA-seq datasets of *C. reinhardtii* wild-type strain CC-124 (*wt*, *mt*-) cells under various light conditions: acclimation to high light (600 $\mu\text{mol}/\text{m}^2/\text{s}$) for 1 h, acclimation to light for 30 min, and acclimation to darkness. The generated heatmap demonstrates the complex dynamics of gene expression for GATA TFs in response to changes in light conditions (Fig. 2). Most of the obtained values of logarithmic change in gene expression levels were statistically insignificant.

Under conditions of increased illumination, a significant upregulation of the expression levels of the *GATA-7*, *GATA-9*, *GATA-10*, *GATA-11* genes is observed, whereas transcription of the *GATA-3*, *GATA-5*, *GATA-8* genes is repressed (statistically significantly). It is likely that excessively high light intensity has a negative impact on the viability of *C. reinhardtii* cells.

Under 30-minute light acclimation, a metabolic rearrangement of the cells occurs, which is marked by a significant increase in transcript levels of the *GATA-3*, *GATA-5*, *GATA-6*, *GATA-7*, *GATA-8* genes. In these conditions, the expression of the *GATA-2*, *GATA-4*, *GATA-10* genes is repressed (statistically significantly).

In the dark, there is a statistically significant suppression of the *GATA-2*, *GATA-3*, *GATA-7*, *GATA-8*, *GATA-11*, *GATA-12* genes expression, along with active expression of the *GATA-9* gene.

The publicly available datasets are characterized by limited experimental conditions and poor characterization, as well as a low level of reliability of changes. We decided to use the RT-qPCR method to obtain comprehensive data on GATA gene expression. Analysis of the RNA-seq data allowed us to select reference genes with stable expression under varying light conditions for our own experiments. The *RPL19* and *RPL32* genes encoding ribosomal proteins were discovered and then confirmed by literature data (Liu et al., 2012). According to published data (Sanchez-Tarre, Kiparissides, 2021), the expression of the *RBCS* gene varies depending on the light spectrum and intensity. Therefore, we selected this gene as a control of changes in experimental conditions.

Analysis of GATA gene expression in *C. reinhardtii* using RT-qPCR

The results of GATA gene expression analysis in *C. reinhardtii* obtained by the RT-qPCR method are presented in Figure 3.

The entire group of GATA genes responds to changes in light conditions rapidly and in a coordinated manner. The strongest

response is observed during the first 30 min of exposure to high light intensity (215 $\mu\text{mol}/\text{m}^2/\text{s}$) after transferring the culture from standard conditions (90 $\mu\text{mol}/\text{m}^2/\text{s}$): nearly all genes show a multi-fold increase in expression levels in response to the stress stimulus. Notably, significant activation of *GATA-6* gene expression occurs only under these conditions. High light intensifies the metabolism of *C. reinhardtii* cells, driving a unified shift in the expression profile of the studied genes. After 2 h of acclimation, stabilization is observed: the expression of several genes (*GATA-1*, *GATA-6*, *GATA-8*, *GATA-9*, *GATA-10*, *GATA-11*, *GATA-12*) decreases significantly, while the expression level of *GATA-5* and *GATA-7* increases.

During the first 30 min after transferring the culture from darkness to light (90 $\mu\text{mol}/\text{m}^2/\text{s}$), cell metabolism undergoes reprogramming. The transcription of the *GATA-1* and *GATA-3* genes is suppressed simultaneously with the activation of the *GATA-2*, *GATA-4*, *GATA-5*, *GATA-7*, *GATA-9*, and *GATA-10* genes. By 2 h of exposure, two profiles are formed that ensure the maintenance of growth and development processes in optimal light conditions: (1) actively transcribed genes (*GATA-1*, *GATA-4*, *GATA-5*, *GATA-8*, *GATA-9*, *GATA-11*, *GATA-12*); (2) repressed genes (*GATA-2*, *GATA-3*, *GATA-6*, *GATA-7*, *GATA-10*).

In the first 30 min after culture transfer from light (90 $\mu\text{mol}/\text{m}^2/\text{s}$) to darkness, the expression of all GATA genes, with the exception of the *GATA-7* gene, is suppressed. Apparently, its product is involved in the metabolic switching processes of *C. reinhardtii* cells during light/dark and dark/light transitions. After 2 h, cell metabolism stabilizes with the identification of three actively transcribed genes, apparently involved in ensuring adaptation to darkness – *GATA-1*, *GATA-3*, *GATA-9*.

The identified dynamics of GATA gene expression align with the standard model of stress response: a change in light conditions triggers a broad emergency cascade (a rapid response to the stimulus), while prolonged exposure narrows the response down to specific regulatory modules. A unique expression profile is observed under each cultivation condition, with some of the GATA genes presumably acting as “switches” between the light and dark metabolic programs.

Comparative analysis of the obtained results

A comparative analysis based on the results of RNA-seq data processed with three normalization methods – FPKM, TPM, DESeq2 – and RT-qPCR data processed using the $\Delta\Delta\text{Ct}$ method was conducted in the PipeSeq program. The constructed correlation matrix allows us to evaluate the concordance between data obtained by different methods and approaches (Fig. 4).

Cross-platform consistency of the results is limited and context-dependent. Despite methodological differences, reproducible agreements include induction of the *GATA-7*, *GATA-9*, *GATA-10* and *GATA-11* genes expression under high light conditions, repression of *GATA-2*, *GATA-11*, and *GATA-12* in the dark and upregulation of *GATA-9* expression in darkness. Early repression of *GATA-3*, *GATA-5*, and *GATA-8* under high light observed at 1 h in RNA-seq data appeared to be inverted in RT-qPCR results at the 2 h time point. This pattern is consistent with a potential phase shift in the cellular response;

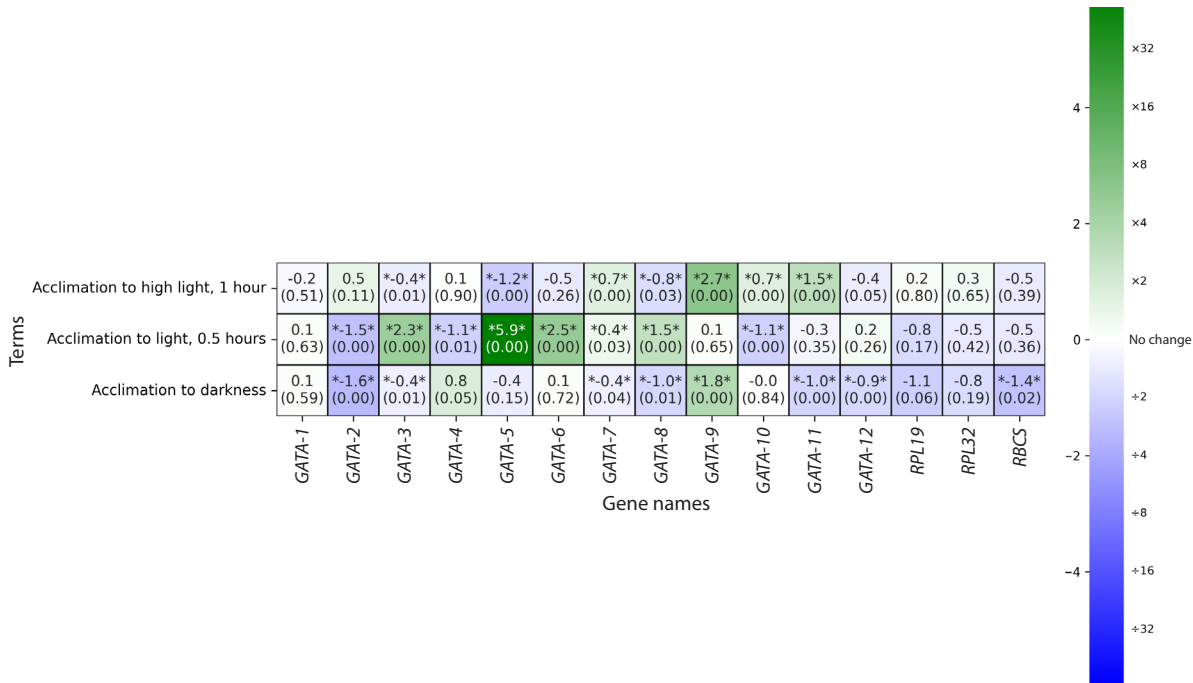


Fig. 2. Heatmap of transcript-level expression for 12 genes encoding GATA family transcription factors and three control genes (*RPL19*, *RPL32*, *RBCS*) in *C. reinhardtii* under various acclimation conditions according to the RNA-seq data. Genes are reflected from left to right, experimental conditions are shown from top to bottom: acclimation to high light for 1 h (row 1), acclimation to light for 30 min (row 2), acclimation to darkness (row 3). Here and in Fig. 3: the cells are colored according to log₂FoldChange values, where positive values (green scale) indicate the induction of expression, negative values (blue scale) indicate repression, and zero values (white color) signify no change in the expression level. The statistical significance of the changes was assessed by adjusted p-values with a threshold of 0.05. The control transcripts were included to verify the quality of data normalization. The images were created in the PipeSeq program.

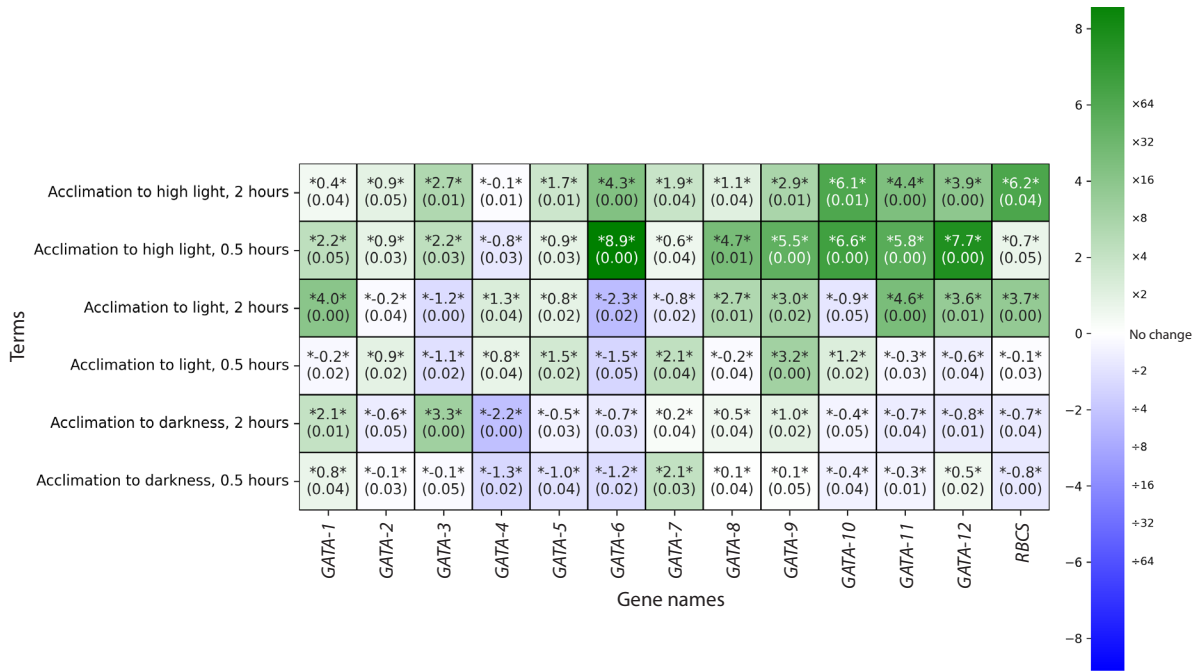


Fig. 3. Heatmap of gene expression (at the transcriptional level) for 12 genes encoding GATA family transcription factors and the experimental condition change control gene (*RBCS*) in *C. reinhardtii* under various acclimation conditions, based on RT-qPCR data. The data are normalized to the reference gene pair *RPL19–RPL32*. Genes are reflected from left to right, experimental conditions are shown from top to bottom: acclimation to high light for 2 h (row 1), for 30 min (row 2), acclimation to light for 2 h (row 3), for 30 min (row 4), acclimation to darkness for 2 h (row 5), for 30 min (row 6).

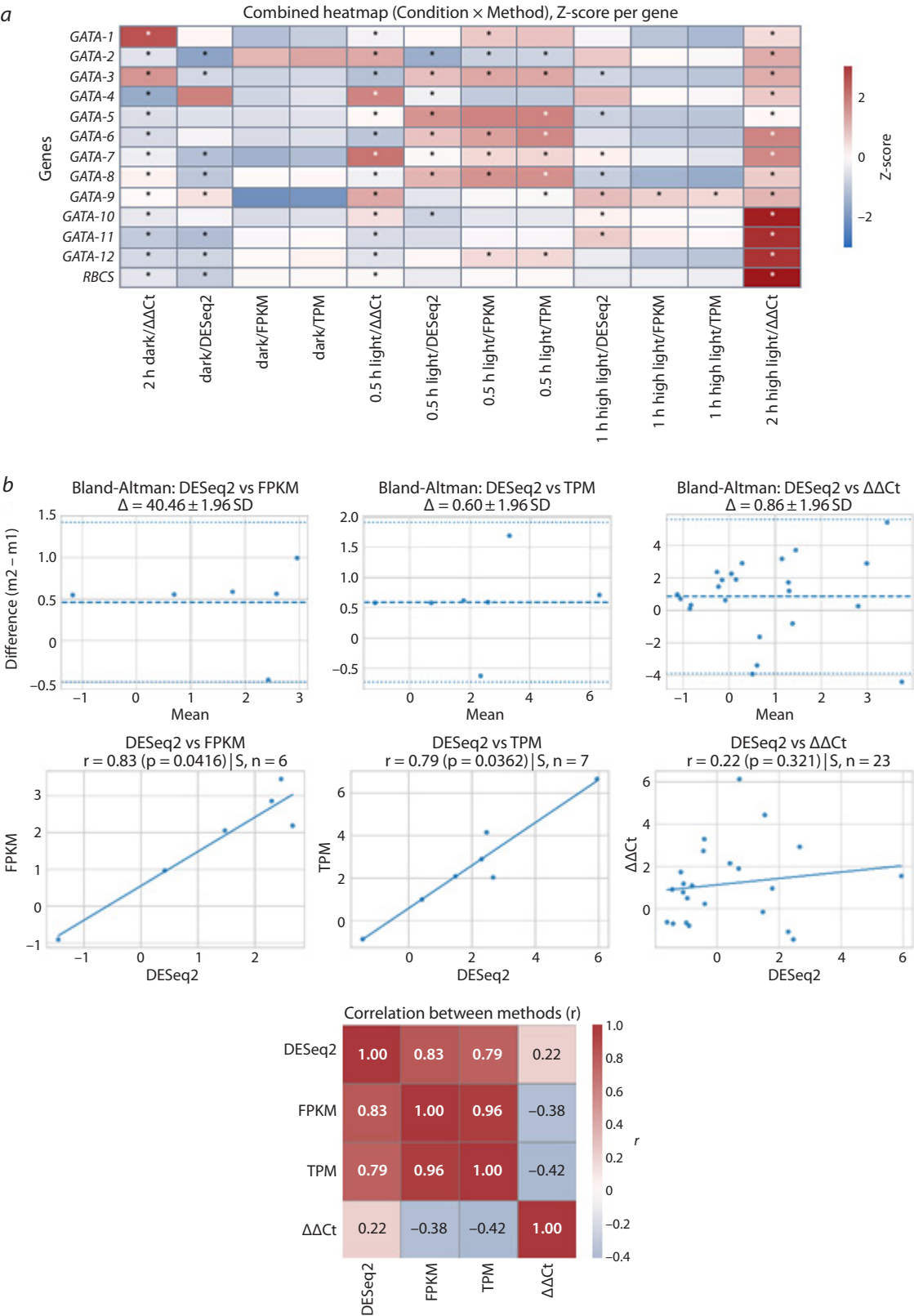


Fig. 4. Comparative statistical analysis of three RNA-seq normalization methods – FPKM, TPM, DESeq2 – and the $\Delta\Delta\text{Ct}$ method for RT-qPCR data.

a, A combined heatmap of Z-scores for log fold change in the expression ($\log_2\text{FoldChange}$) values of 12 genes encoding GATA family transcription factors and the *RBCS* gene in *C. reinhardtii* presented for each combination of experimental conditions and normalization method. Statistically significant changes ($\text{FDR} < 0.05$) are marked with an asterisk. *b*, Bland–Altman plots, correlation scatterplots, and a matrix of statistically significant ($\text{FDR} < 0.05$) $\log_2\text{FoldChange}$ estimates across the compared normalization methods. Color reflects the Pearson correlation coefficient (r), where the red scale indicates positive correlation, and the blue scale indicates negative correlation. The images were created in the PipeSeq program.

however, confirmation requires a time-matched dataset. In dark conditions, both datasets corroborated a common repression of photosynthesis-related transcripts of *RBCS*, *GATA-2*, *GATA-11*, and *GATA-12* alongside increased *GATA-9* expression. This convergence indicates a coordinated dark-adaptation program, as captured by both methods.

The correlation analysis revealed strong internal consistency among the RNA-seq-derived metrics (DESeq2, FPKM, TPM), while agreement with the $\Delta\Delta C_t$ method (RT-qPCR) was weaker. This discrepancy likely reflects differences in sampling time points, reference gene stability, and methodological resolution. The highest correlation was observed between DESeq2 and $\Delta\Delta C_t$ (Fig. 4). These findings demonstrate that although both platforms capture overlapping regulatory trends for certain GATA genes, each also reveals condition- and time-specific expression changes. This underscores the value of integrating both approaches.

Discussion

Advantages of the PipeSeq program

PipeSeq was developed with a focus on minimal system administration requirements and optimized for local execution on Windows-based systems. The program offers a user-friendly graphical interface that lowers the entry barrier, integrates state-of-the-art data analysis methods, and is suitable for small to medium-sized datasets on standard personal computers. This makes PipeSeq a practical solution for individual researchers and small laboratories that do not have significant computing resources or system administration experience.

Currently, there are several tools and platforms for transcriptomic analysis, such as Galaxy (Afgan et al., 2018), Nextflow (Di Tommaso et al., 2017), Snakemake (Mölder et al., 2021), as well as specialized pipelines like HISAT2-StringTie-Ballgown (Pertea et al., 2016) and Kallisto-Sleuth (Bray et al., 2016). Galaxy is a convenient web-based platform with a graphical interface, but it requires separate server administration and is not always suitable for individual researchers or small laboratories without IT support (Afgan et al., 2018).

The Nextflow and Snakemake software platforms offer high flexibility and scalability through parallel task execution and support for containerization, which ensures complete analysis reproducibility. However, using these systems requires programming and Linux environment administration skills, limiting their use by biologists without console and scripting experience (Di Tommaso et al., 2017; Mölder et al., 2021).

An important aspect is support for differential gene expression analysis. Modern approaches recommend using packages based on the negative binomial distribution (e. g., DESeq2), which provide high accuracy and false discovery rate control (Zhao S. et al., 2020; Zhao Y. et al., 2021; Elahimanesh, Najafi, 2024).

Our developed PipeSeq software pipeline (Fig. 1) uses one of the fastest tools for read alignment (HISAT2) and transcript assembly (StringTie), which significantly outperform previous methods (TopHat and Cufflinks) in terms of speed and computational resource requirements (Kim et al., 2015,

2019). The integration of DESeq2 allows to automatically obtain statistically valid results of differential expression with FDR control (Love et al., 2014). This places it on par with recognized solutions in RNA-seq data analysis, such as the HISAT2-StringTie-Ballgown pipeline, which uses a similar statistical approach, albeit less integrated (Pertea et al., 2016). Another advantage of the developed program is the ability to process and visualize qPCR data using the $\Delta\Delta C_t$ method and to conduct comparative statistical analysis (calculating the Pearson, Spearman, and Kendall correlation coefficients) with three RNA-seq normalization methods – FPKM, TPM, and DESeq2.

The developed PipeSeq program has the following features.

1. User interaction with the program occurs exclusively through a graphical interface (five windows).
2. All data processing stages are integrated within a single program (no need to connect additional packages) and do not require an internet connection.
3. The program includes automated selection of trimming parameters (adapter removal, quality threshold, minimum read length) based on quality control at the input and output.
4. A built-in module for processing qPCR data using the $\Delta\Delta C_t$ method (import of threshold cycle (Ct) values, calculation of normalized expression, comparative statistical analysis based on RNA-seq data processing results and qPCR results).

Expression profiles

of GATA genes in *C. reinhardtii*

Light is one of the key regulators of gene expression in photosynthetic organisms. Through a comprehensive study, we have obtained the most complete expression profiles of GATA genes in *C. reinhardtii*, supplementing publicly available RNA-seq data with our own experiments using the RT-qPCR method. Despite the fact that currently the results of transcriptomic studies in most cases do not require additional verification (Coenye, 2021), we encountered a limited amount of relevant data (three datasets) in open databases, accompanied by brief descriptions of the growth conditions for the cultures selected for analysis.

The resulting RNA-seq data (Fig. 2) for some GATA genes showed low statistical significance of changes ($FDR > 0.05$). Processing of the selected RNA-seq datasets using the PipeSeq program allowed us: (1) to identify and confirm a pair of constitutively expressed reference genes *RPL19–RPL32* (Liu et al., 2012); (2) to determine expression profiles of the target genes in response to specific stimuli, with the aim of subsequently applying the RT-qPCR method for precise quantitative assessment of these changes across an expanded range of experimental conditions.

Our findings (Fig. 3) confirm that changes in light conditions are a significant factor modulating GATA gene expression in photosynthetic organisms (Manfield et al., 2007; Naito et al., 2007; Luo et al., 2010; Schröder et al., 2023). A previous analysis of the protein interaction network (Virolainen, Chekunova, 2024) identified three functional clusters in which the 12 GATA factors of *C. reinhardtii* are likely involved.

The first functional cluster, consisting of the *GATA-1*, *GATA-2*, and *GATA-10* genes, links photoreception (genes *CHLAMYDOMONAS PHOTOLYASE HOMOLOG 1 (CPH1)*, *SHOC2/SUR8-like LRR (CSL)*), circadian regulation (genes of the *RHYTHM OF CHLOROPLAST (ROC)* family) and phosphorus metabolism (the *PHOSPHATE STARVATION RESPONSE 1 (PSR1)* gene), ensuring adaptation of the cell to light regimes.

The second cluster (*GATA-9*) is associated with the functioning of the aryl hydrocarbon receptor complex, which is activated under high light conditions, supporting detoxification processes in the cell.

The third cluster (*GATA-3*, *GATA-4*, *GATA-5*, *GATA-6*, *GATA-7*, *GATA-8*, *GATA-10*, *GATA-11*, *GATA-12*) coordinates nitrogen assimilation (via the *NITRATE REDUCTASE (NIT1)* gene), chromatin remodeling (via genes encoding histone deacetylases), DNA replication (via the gene encoding helicase (*CrRuvBL1*)), membrane transport and cell division (via genes encoding the Rab family protein (*RABF1*), kinesin-like protein, and a subunit 4 of the cyclosome), showing pronounced light dependence.

The observed dynamics of GATA gene expression represent a coordinated response that integrates key metabolic processes during the cell cycle (Voigt, Münzner, 1987; Müller et al., 2017; Salomé, Merchant, 2019). The response of all studied GATA genes to light stimuli suggests the presence of a cross-regulation mechanism among the three functional networks through by yet unidentified or uncharacterized genes and proteins.

The results of our studies allow us to make the first significant assumptions regarding the functions of GATA TFs in *C. reinhardtii* and green algae in general (Schwechheimer et al., 2022), laying the necessary foundation for future research.

A comparative analysis of the use of three RNA-seq normalization methods (FPKM, TPM, DESeq2), and the $\Delta\Delta C_t$ method based on RT-qPCR data demonstrated the highest correlation in the DESeq2– $\Delta\Delta C_t$ pair (Fig. 4). This finding confirms the literature reports on the high accuracy of tools based on the negative binomial distribution (Zhao S. et al., 2020; Zhao Y. et al., 2021; Elahimanesh, Najafi, 2024) and demonstrates the reliability of the $\Delta\Delta C_t$ normalization method for RT-qPCR data for accurate quantification of the expression levels of target genes (Livak, Schmittgen, 2001; Schmittgen, Livak, 2008; Shi, He, 2014; He et al., 2015; Coenye, 2021; Schröder et al., 2023). Collectively, both approaches reveal overlapping trends for a small number of GATA genes and condition-, method-, and time-dependent differences. This underscores the value of integrating massive (RNA-seq) and targeted (RT-qPCR) approaches to obtain a more comprehensive understanding of gene expression dynamics.

Conclusion

During the study, we developed and applied the PipeSeq pipeline. It includes automated steps for data loading, read alignment, and statistical processing of RNA-seq data. Additionally, the program enables the analysis and visualization of data generated by both RNA-seq and RT-qPCR methods.

The results of our study showed that RNA sequencing and RT-qPCR methods can reveal similar patterns of gene expression changes, but show differences in the effect size estimation and sensitivity in detecting expression changes.

The data obtained allow us to conclude that the GATA TFs in *C. reinhardtii* form three functionally specialized groups (clusters), the coordinated regulation of which constitutes a key mechanism ensuring proper progression of the cell cycle under changing environmental conditions. The expression profiles of the *GATA-2*, *GATA-4*, *GATA-5*, *GATA-6*, *GATA-8*, *GATA-10*, *GATA-11*, and *GATA-12* genes suggest their involvement in the regulation of light-dependent metabolic processes. The *GATA-1*, *GATA-3*, *GATA-7*, and *GATA-9* genes are involved in switching metabolism during the light/dark and dark/light transitions. Future research on GATA TFs in *C. reinhardtii* should be aimed at search and verification of the target genes and interactions in regulatory networks, as well as confirmation of the predicted functions in response to changes in other cultivation conditions.

The PipeSeq program has demonstrated its effectiveness in a comprehensive study of differential gene expression as a tool for conducting a full cycle of bioinformatic analysis of RNA-seq data with the ability to process RT-qPCR data and perform comparative statistical analysis of the results obtained using different methods. The developed pipeline can be used to study the gene expression profiles of any research object.

References

- Afgan E., Baker D., Batut B., van den Beek M., Bouvier D., Čech M., Chilton J., ... Soranzo N., Goecks J., Taylor J., Nekrutenko A., Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 2018;46(W1):W537–W544. doi 10.1093/nar/gky379
- Bray N.L., Pimentel H., Melsted P., Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016;4(5):525–527. doi 10.1038/nbt.3519
- Coenye T. Do results obtained with RNA-sequencing require independent verification? *Biofilm.* 2021;3:100043. doi 10.1016/j.biofilm.2021.100043
- Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szczesniak M.W., Gaffney D.J., Elo L.L., Zhang X., Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:13. doi 10.1186/s13059-016-0881-8
- Deraveaux S., Vandesompele J., Hellemans J. How to do successful gene expression analysis using real-time PCR. *Methods.* 2010;50(4):227–230. doi 10.1016/j.ymeth.2009.11.001
- Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E., Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35(4):316–319. doi 10.1038/nbt.3820
- Elahimanesh M., Najafi M. Differentially expressed genes of RNA-seq data are suggested on the intersections of normalization techniques. *Biochem Biophys Rep.* 2024;37:101618. doi 10.1016/j.bbrep.2023.101618
- Ewels P., Magnusson M., Lundin S., Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016;32(19):3047–3048. doi 10.1093/bioinformatics/btw354
- Goodstein D.M., Shu S., Howson R., Neupane R., Hayes R.D., Fazo J., Mitros T., Dirks W., Hellsten U., Putnam N., Rokhsar D.S. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 2012;40(D1):D1178–D1186. doi 10.1093/nar/gkr944

- Harris E.H. The *Chlamydomonas* Sourcebook. A Comprehensive Guide to Biology and Laboratory Use. San Diego: Academic Press, 1989
- He F., Liu Q., Zheng L., Cui Y., Shen Z., Zheng L. RNA-Seq analysis of rice roots reveals the involvement of post-transcriptional regulation in response to cadmium stress. *Front Plant Sci.* 2015;6:1136. doi 10.3389/fpls.2015.01136
- Hunter J.D. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9(3):90-95. doi 10.1109/MCSE.2007.55
- Kim D., Langmead B., Salzberg S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12(4):357-360. doi 10.1038/nmeth.3317
- Kim D., Paggi J.M., Park C., Bennett C., Salzberg S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37(8):907-915. doi 10.1038/s41587-019-0201-4
- Kvitko K.V., Borschevskaya T.N., Chunaev A.S., Tugarinov V.V. Peterhof genetic collection of green algae strains (*Chlorella*, *Scenedesmus*, *Chlamydomonas*). In: Cultivation of Collection Strains of Algae. Leningrad, 1983 (in Russian)
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079. doi 10.1093/bioinformatics/btp352
- Li X., Wang C.Y. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci.* 2021;13:36. doi 10.1038/s41368-021-00146-0
- Liao Y., Smyth G.K., Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923-930. doi 10.1093/bioinformatics/btt656
- Liu C., Wu G., Huang X., Liu S., Cong B. Validation of housekeeping genes for gene expression studies in an ice alga *Chlamydomonas* during freezing acclimation. *Extremophiles.* 2012;16:419-425. doi 10.1007/s00792-012-0441-4
- Livak K.J., Schmittgen T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods.* 2001;25(4):402-408. doi 10.1006/meth.2001.1262
- 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
- Luo X.M., Lin W.H., Zhu S., Zhu J.Y., Sun Y., Fan X.Y., Cheng M., ... Liu L., Zhang M., Xie Q., Chong K., Wang Z.Y. Integration of light- and brassinosteroid-signaling pathways by a GATA transcription factor in *Arabidopsis*. *Dev Cell.* 2010;19(6):872-883. doi 10.1016/j.devcel.2010.10.023
- Manfield I.W., Devlin P.F., Jen C.H., Westhead D.R., Gilmartin P.M. Conservation, convergence, and divergence of light-responsive, circadian-regulated, and tissue-specific expression patterns during evolution of the *Arabidopsis* GATA gene family. *Plant Physiol.* 2007;143(2):941-958. doi 10.1104/pp.106.090761
- Marioni J.C., Mason C.E., Mane S.M., Stephens M., Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008;18(9):1509-1517. doi 10.1101/gr.079558.108
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal.* 2011;17(1):10-12. doi 10.14806/ej.17.1.200
- McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing.* 2011;14(9):1-9.
- Merchant S.S., Prochnik S.E., Vallon O., Harris E.H., Karpowicz S.J., Witman G.B., Terry A., ... Werner G., Zhou K., Grigoriev I.V., Rokhsar D.S., Grossman A.R. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science.* 2007;318(5848):245-250. doi 10.1126/science.1143609
- Mölder F., Jablonski K.P., Letcher B., Hall M.B., Tomkins-Tinch C.H., Sochat V., Forster J., ... Wilm A., Holtgrewe M., Rahmann S., Nahnsen S., Köster J. Sustainable data analysis with Snakemake. *F1000Res.* 2021;10:33. doi 10.12688/f1000research.29032.2
- Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B.J. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008;5(7):621-628. doi 10.1038/nmeth.1226
- Müller N., Wenzel S., Zou Y., Künzel S., Sasso S., Weiß D., Prager K., Grossman A., Kottke T., Mittag M. A plant cryptochrome controls key features of the *Chlamydomonas* circadian clock and its life cycle. *Plant Physiol.* 2017;174(1):185-201. doi 10.1104/pp.17.00349
- Muzellec B., Teleńczuk M., Cabeli V., Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics.* 2023;39(9):btad547. doi 10.1093/bioinformatics/btad547
- Naito T., Kiba T., Koizumi N., Yamashino T., Mizuno T. Characterization of a unique GATA family gene that responds to both light and cytokinin in *Arabidopsis thaliana*. *Biosci Biotechnol Biochem.* 2007;71(6):1557-1560. doi 10.1271/bbb.60692
- Pertea M., Pertea G.M., Antonescu C.M., Chang T.C., Mendell J.T., Salzberg S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015;33(3):290-295. doi 10.1038/nbt.3122
- Pertea M., Kim D., Pertea G.M., Leek J.T., Salzberg S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;11(9):1650-1667. doi 10.1038/nprot.2016.095
- Ren W., Kong L., Jiang S., Ma L., Wang H., Li X., Liu Y., Ma W., Yan X. Genome-wide identification, evolution, and characterization of GATA gene family and GATA gene expression analysis post-MeJA treatment in *Platycodon grandiflorum*. *J Plant Growth Regul.* 2025;44:155-167. doi 10.1007/s00344-024-11468-8
- Reyes J.C., Muro-Pastor M.I., Florencio F.J. The GATA family of transcription factors in *Arabidopsis* and rice. *Plant Physiol.* 2004;134(4):1718-1732. doi 10.1104/pp.103.037788
- Riechmann J.L., Heard J., Martin G., Reuber L., Jiang C.Z., Keddie J., Adam L., ... Broun P., Zhang J.Z., Ghandehari D., Sherman B.K., Yu G.L. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science.* 2000;290(5499):2105-2110. doi 10.1126/science.290.5499.2105
- Salomé P.A., Merchant S.S. A series of fortunate events: introducing *Chlamydomonas* as a reference organism. *Plant Cell.* 2019;31(8):1682-1707. doi 10.1105/tpc.18.00952
- Sanchez-Tarre V., Kiparissides A. The effects of illumination and trophic strategy on gene expression in *Chlamydomonas reinhardtii*. *Algal Res.* 2021;54:102186. doi 10.1016/j.algal.2021.102186
- Schmittgen T., Livak K. Analyzing real-time PCR data by the comparative CT method. *Nat Protoc.* 2008;3:1101-1108. doi 10.1038/nprot.2008.73
- Schröder P., Hsu B.Y., Gutsche N., Winkler J.B., Hedtke B., Grimm B., Schwechheimer C. B-GATA factors are required to repress high-light stress responses in *Marchantia polymorpha* and *Arabidopsis thaliana*. *Plant Cell Environ.* 2023;46(8):2376-2390. doi 10.1111/pce.14629
- Schwechheimer C., Schröder P.M., Blaby-Haas C.E. Plant GATA factors: their biology, phylogeny, and phylogenomics. *Annu Rev Plant Biol.* 2022;73(1):123-148. doi 10.1146/annurev-arplant-072221-092913
- Shi Y., He M. Differential gene expression identified by RNA-Seq and qPCR in two sizes of pearl oyster (*Pinctada fucata*). *Gene.* 2014;538(2):313-322. doi 10.1016/j.gene.2014.01.031
- Virolainen P.A., Chekunova E.M. GATA family transcription factors in alga *Chlamydomonas reinhardtii*. *Curr Genet.* 2024;70(1):1. doi 10.1007/s00294-024-01280-y
- Voigt J., Münzner P. The *Chlamydomonas* cell cycle is regulated by a light/dark-responsive cell-cycle switch. *Planta.* 1987;172:463-472. doi 10.1007/BF00393861

- Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10:57-63. doi [10.1038/nrg2484](https://doi.org/10.1038/nrg2484)
- Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Church D.M., DiCuccio M., ... Suzek T.O., Tatusov R., Tatusova T.A., Wagner L., Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2005;33:D39-D45. doi [10.1093/nar/gki062](https://doi.org/10.1093/nar/gki062)
- Zhao S., Ye Z., Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA.* 2020;26(8):903-909. doi [10.1261/rna.074922.120](https://doi.org/10.1261/rna.074922.120)
- Zhao Y., Li M.C., Konaté M.M., Chen L., Das B., Karlovich C., Williams P.M., Evrard Y.A., Doroshow J.H., McShane L.M. TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *J Transl Med.* 2021;19(1):269. doi [10.1186/s12967-021-02936-w](https://doi.org/10.1186/s12967-021-02936-w)

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

Received August 12, 2025. Revised December 21, 2025. Accepted December 22, 2025.

doi 10.18699/vjgb-26-35

Plant virome analysis by high-throughput sequencing: concepts and approaches

D.D. Belkina , S.V. Vinogradova  

Federal Research Centre "Fundamentals of Biotechnology" of the Russian Academy of Sciences, Moscow, Russia

 coatprotein@bk.ru

Abstract. The metagenomic approach based on high-throughput sequencing is becoming increasingly prevalent for the detection of viral infections in plants. This method allows us to study the species composition of viruses associated with the plant, including novel species, describe their population genetic structure, and develop genetic test systems for routine diagnostics. A metagenomic approach to phytosanitary monitoring can help to determine the cause of unknown plant diseases, which is particularly important for preventing the spread of pathogens, such as viruses. Furthermore, as it is impossible to eliminate plant viruses in field conditions, comprehensive diagnostics using high-throughput sequencing is becoming an effective tool for complying with quarantine regulations on the import of foreign material, as well as for producing high-quality local planting material. High-throughput sequencing is becoming more affordable every year, with both the instrumentation and analytical capacity improving. This review summarizes key approaches to analyzing plant virome using high-throughput sequencing. The analysis process, from sample collection to bioinformatic data processing, validation and interpretation, is described in detail. The features of sequencing platforms and the factors affecting sequencing quality, including contamination, are discussed. Three complementary approaches to processing bioinformatic data are described: mapping reads to reference viral sequences; assembling and annotating contigs; taxonomic classification of reads without assembly. The importance of carefully interpreting the results is emphasized, considering the bioinformatic analysis and the validation by molecular genetic methods. This review will be useful for both researchers and specialists who have no experience with high-throughput sequencing, and those who have used this method for other applications.

Key words: metagenomics; high-throughput sequencing; viruses; plant virome; bioinformatics

For citation: Belkina D.D., Vinogradova S.V. Plant virome analysis by high-throughput sequencing: concepts and approaches. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov J Genet Breed.* 2026;30(2):311-320. doi 10.18699/vjgb-26-35

Funding. This work was supported by Russian Science Foundation grant No. 23-16-00232.

Анализ вирома растений с помощью высокопроизводительного секвенирования: принципы и подходы

Д.Д. Белкина , С.В. Виноградова  

Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Россия

 coatprotein@bk.ru

Аннотация. В последние годы метагеномный подход, основанный на высокопроизводительном секвенировании, находит все большее применение в диагностике вирусных инфекций растений. Этот метод позволяет изучить видовой состав вирусов, ассоциированных с исследуемым растением, и в том числе обнаружить ранее не описанные виды, охарактеризовать их популяционно-генетическую структуру и разработать генетические тест-системы для рутинной диагностики. Проведение фитосанитарного мониторинга с использованием метагеномного подхода может способствовать определению этиологии неизвестных заболеваний растений, что особенно важно для предотвращения распространения таких патогенов, как вирусы. Кроме того, с учетом невозможности элиминации вирусов растений в полевых условиях комплексная диагностика с помощью высокопроизводительного секвенирования становится эффективным инструментом как в целях соблюдения карантинного законодательства при ввозе импортного материала, так и для получения отечественного посадочного материала высоких категорий. При этом с каждым годом высокопроизводительное секвенирование становится более доступным: расширяется как приборно-техническая, так и аналитическая база. В настоящем обзоре систематизированы ключевые подходы к анализу вирома растений с помощью высокопроизводительного секвенирования. В основной части статьи

описаны этапы проведения анализа: от сбора образцов до биоинформатической обработки данных, ее валидации и интерпретации. Подробно рассмотрены особенности современных платформ секвенирования и факторы, влияющие на качество чтения, в том числе контаминация. Охарактеризованы три взаимодополняющих подхода для обработки биоинформатических данных: картирование чтений на референсные последовательности вирусов; сборка и аннотация контигов; таксономическая классификация чтений без дополнительной сборки. Особое внимание уделено необходимости тщательной интерпретации результатов с учетом как биоинформатического анализа, так и валидации идентификации молекулярно-генетическими методами. Обзор будет полезен как для исследователей и специалистов, не имеющих опыта работы с высокопроизводительным секвенированием, так и для тех, кто использовал этот инструмент для выполнения других задач.

Ключевые слова: метагеномика; высокопроизводительное секвенирование; вирусы; виром растений; биоинформатика

Introduction

Plant viral diseases cause substantial losses in crop production worldwide. In the context of globalization and climate change, the spread of plant viruses can lead to devastating epidemics, posing a serious threat to global food security (Jones, 2021). For instance, yield losses in tomato and pepper crops due to infection with the tomato mosaic virus (ToMV) are estimated at 25–70 % (Panno et al., 2021), whereas a reduction in grape yield caused by grapevine leafroll disease associated with viruses of the family *Closteroviridae* can reach 50 % (Atallah et al., 2012). In this regard, monitoring, early diagnostics and adherence to quarantine measures play a crucial role in limiting the spread of viral infections.

Currently, the identification of plant viruses is carried out mainly by serological and molecular methods, which have become widely used due to their assay speed, high sensitivity and specificity. However, these approaches enable the detection of only those virus species for which corresponding test systems have been developed (Maina et al., 2024).

Prior to the advent of metagenomics, knowledge of plant viruses was restricted to a few hundred species (Pappas et al., 2021), and consequently, the role of viruses in the microbiome of agricultural plants remained underestimated. The metagenomic approach allows us to obtain an in-depth understanding of all viruses of a plant, or its virome. As of 2025, the International Committee on Taxonomy of Viruses (ICTV) has registered 2,598 plant viruses and viroids, including 206 added in the past year (Rubino et al., 2025). Virome sequencing provides unique information on viral genome sequences, which can be used to conduct population studies and trace pathways of virus spreading, associate specific strains with symptom expression, and develop specific test systems for their rapid detection.

Plant virome analysis using high-throughput sequencing is a rapidly evolving method. Each year, new recommendations, protocol optimizations, and specialized data-processing tools are published. Although this analysis has not yet become a routine diagnostic method for plant viral infections, its fundamental principles have already been established.

This review systematizes key approaches to plant virome analysis using high-throughput sequencing. It will be useful for both researchers who are new to high-throughput sequencing and those who have used this technology for other applications.

General scheme of plant virome analysis

Structurally, virome analysis can be divided into two main stages. The first stage includes sample collection, material preparation, and sequencing (Fig. 1). This part of the analysis is crucial. If the material for sequencing is of poor quality, obtaining reliable results at subsequent steps becomes impossible. When collecting samples, it is important to consider that viruses are unevenly distributed in plant tissues, therefore, multiple fragments should be collected from each individual plant. This is particularly important for perennial plants. After collection, the samples should be transported to the laboratory as quickly as possible to minimize the risk of nucleic acid degradation, for instance, due to exposure to direct sunlight.

The second stage of plant virome analysis involves computer processing of sequencing data using bioinformatics tools, followed by validation and interpretation of the results. In contrast to the steps of the first stage, bioinformatics data processing can be repeated at any time to correct potential errors, apply additional analytical methods, or double-check the results of other studies. Therefore, it is extremely important to avoid errors during sampling, sample preparation, and library sequencing. For reliable virus identification, the results of bioinformatics analysis must be validated using molecular genetic methods.

Initial material for analysis

Plant viruses differ from other pathogens in the diversity of their genome types. They can use either DNA or RNA molecules as carriers of genetic information, and these molecules may be single-stranded or double-stranded, linear or circular (Lefeuvre et al., 2019). Single-stranded RNA viruses are subdivided into those with positive-sense genomes, in which the genomic RNA corresponds to mRNA, and those with negative-sense genomes, in which the genomic RNA is complementary to mRNA. Given this

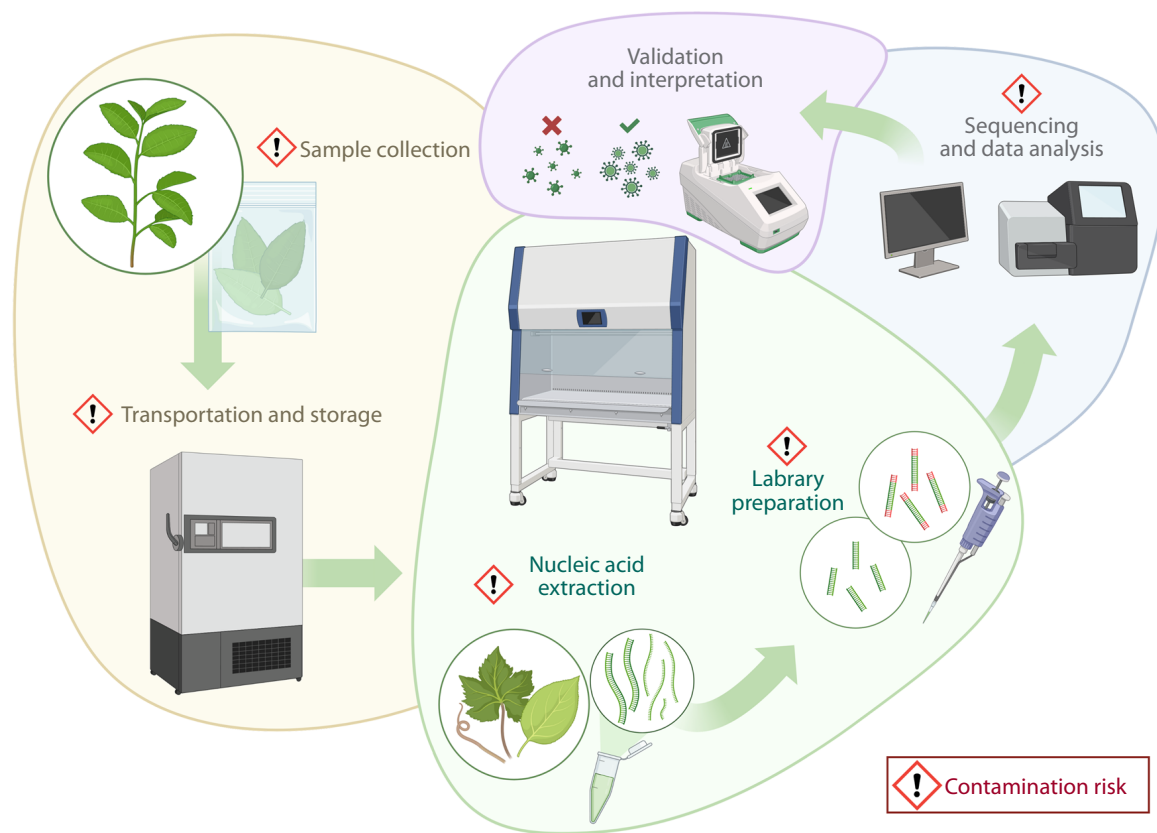


Fig. 1. The main stages of plant virus analysis.

genomic diversity, identifying all plant-infecting viruses using a single approach is challenging (Fitzpatrick et al., 2021). Below, we review various approaches to plant virome analysis depending on the type of template used.

Sequencing of total RNA extracted from plant tissues is a relatively simple, reliable, and widely used method for virome analysis (Lee et al., 2020; Nabeshima, Abe, 2021; Vinogradova et al., 2023). This approach enables the detection of not only RNA viruses but also transcripts of DNA viruses (Pecman et al., 2017; Cobbin et al., 2021), and viroids, making it optimal for diagnostics of viruses with different types of genomic nucleic acids.

However, the sensitivity of this method can be significantly reduced by the co-extraction of abundant host plant RNA during sample preparation (Maliogka et al., 2018). As a result, the detection of low-titer viruses becomes difficult (Roossinck et al., 2015). This problem can be mitigated by enriching the sample for virus-associated nucleic acids (Gaafar, Ziebell, 2020). Commonly employed strategies include the depletion of ribosomal and transfer RNAs or the enrichment of polyadenylated transcripts. The former method is generally preferable because not all groups of viruses have a polyadenylated 3' end.

Another approach for plant virome analysis is sequencing virion-associated nucleic acids (Filloux et al., 2015).

This approach has the advantage of targeting exclusively viral sequences, which makes it possible to analyze both RNA and DNA viruses, including those present at low titers (Moubset et al., 2022). However, it is limited to the detection of viruses with a capsid, which is absent in satellite viruses and viroids, and it requires more effort than the extraction of total RNA (Maliogka et al., 2018).

The next approach involves the analysis of short interfering RNAs (siRNAs), which are produced in plant cells as a result of the immune response to viral infection. These molecules represent fragments of viral genomes cleaved by plant enzymes, 21–24 nucleotides in length (Vivek et al., 2020; Zhuravlyov et al., 2022).

While siRNA sequencing can detect both RNA and DNA viruses, the short length of these molecules complicates the *de novo* assembly of complete viral genomes (Pecman et al., 2017; Maliogka et al., 2018; Turco et al., 2018).

Many viruses produce double-stranded RNA (dsRNA) as a replication intermediate, which can also be used in plant virome studies. Sequencing of dsRNA enables the identification of not only dsRNA viruses but also most single-stranded RNA viruses, viroids, and some DNA viruses (Gallo et al., 2021; Fall et al., 2025). Nevertheless, this method is less efficient for detecting DNA viruses and

typically yields positive results only at high levels of viral transcript expression (Gaafar, Ziebell, 2020).

Sequencing platforms

All metagenomic studies are based on high-throughput sequencing (HTS) of nucleic acids. This group of methods includes second- and third-generation sequencing that replaced Sanger sequencing. While Sanger sequencing remains the gold standard for reading sequences up to 1,000 nucleotides in length, its throughput is insufficient for metagenomic applications (Crossley et al., 2020).

Second-generation sequencing, also referred to as next-generation sequencing (NGS) or massively parallel sequencing, generates large volumes of short reads, typically several hundred nucleotides in length. This capability enables the rapid and comprehensive sequencing of all sequences in a sample. However, the short read length complicates the assembly of the complete viral genome during bioinformatics analysis (Maina et al., 2024).

The leader in the NGS market is Illumina (USA), whose platforms are characterized by high throughput and low error rates (Maina et al., 2024). In recent years, Illumina has faced growing competition from Chinese manufacturers such as MGI, GeneMind, and Cygnus Biosciences. Moreover, the Russian company Syntol has recently developed and commercialized the Nanophor SPS sequencer (Kurochkin et al., 2021). Most contemporary NGS platforms utilize fluorescence-based sequencing-by-synthesis technology within flow cells (Zubov et al., 2021) and share similar library preparation protocols, often with compatible reagents. The library preparation process typically involves several key steps: fragmentation of nucleic acids template (e. g., via ultrasonication or restriction enzyme digestion), ligation of platform-specific short oligonucleotide adapters to one or both ends of the fragments, and amplification of these fragments (Kutnjak et al., 2021). Adapters vary depending on the NGS platform and are required to initiate the sequencing process (Lebas et al., 2022). It is worth noting that second-generation sequencers can only read DNA molecules, therefore, if the template is RNA, fragmentation must be followed by reverse transcription to synthesize cDNA.

Third-generation sequencing (TSG), also known as single-molecule sequencing, is based on reading the sequences of individual molecules and does not require their fragmentation and amplification (Villamor et al., 2019). The maximum read length in this case is several hundred thousand nucleotides, which allows for the complete sequencing of small genomes, such as those of viruses, in a single read (Liu et al., 2025). Important advantages of TGS include the direct sequencing of RNA without reverse transcription to cDNA and the capacity for real-time sequencing (Sun et al., 2022).

Leading manufacturers of third-generation platforms are Pacific Biosciences (USA) and Oxford Nanopore Technologies (UK). Although both technologies allow to eliminate amplification-induced polymerase errors, they have relatively low read quality compared to NGS (Rose et al., 2016; Rang et al., 2018). However, the recently introduced PacBio HiFi technology provides read accuracy exceeding 99.5 % (Han et al., 2024). Nanopore sequencing, originally developed by Oxford Nanopore Technologies, has seen continuous improvements in accuracy and a significant reduction in cost, enhancing its accessibility (Javaran et al., 2021). The Chinese company Qitan Tech has also mastered this technology (Wang et al., 2022). It can be expected that in the future single-molecule sequencing will replace NGS.

Other factors affecting the quality of plant virome sequencing

As discussed above, the selection of both nucleic acid type (e. g., total RNA, siRNA, dsRNA) and sequencing platform plays a crucial role in plant virome analysis. Sequencing quality and read amount are also affected by a number of other factors.

Contamination is one a major negative factor in any metagenomic analysis (Piombo et al., 2021; Lebas et al., 2022). Sample contamination can occur at all stages, from plant tissue collection to the sequencing process itself (Fig. 1). To monitor contamination during nucleic acid extraction and library preparation, negative controls without template addition are recommended (Fitzpatrick et al., 2021). It is also essential to maintain physical separation between laboratory workflows, use sterile reagents, and ensure the cleanliness of glassware, equipment, and work surfaces (Kutnjak et al., 2021; Maina et al., 2024). Comprehensive guidance for preventing contamination using HTS in plant pathogen diagnostics is provided by S. Massart et al. (2022).

In addition to contamination, virome sequencing quality is strongly influenced by the choice of methods, reagents (particularly enzymes), and equipment. For example, library preparation for second-generation sequencing requires the use of reverse transcriptase and DNA polymerase, both of which can introduce errors during amplification (Cholet et al., 2020). Therefore, high-precision enzymes should be used to minimize artifacts.

The accuracy of virus identification is affected by the number of reads in the dataset. If the read amount is insufficient, viruses present at low titers may remain undetected. Conversely, excessive sequencing increases the risk of false positive identifications (Massart et al., 2014). To determine an optimal sequencing depth, it may be useful to review published virome analyses for the plant species of interest.

Bioinformatics analysis

Bioinformatics analysis of plant virome sequencing data is a multi-step process that can be performed on both personal computers and high-performance clusters. Difference in computing power affects the speed of analysis. Sequencing reads are processed by specialized tools in an order optimized for the characteristics of a given dataset.

Most bioinformatics tools are developed for Linux operating systems and lack a graphical user interface, requiring command-line use. For other operating systems, cross-platform applications, web-based tools, and commercial software packages are available. The latter are particularly convenient for novice users, as they allow nearly all steps of virome analysis to be performed in a single environment with an intuitive interface. Popular examples include Geneious Prime (New Zealand) and CLC Genomics Workbench (USA). Despite their convenience, the functionality of commercial solutions is limited by the set of built-in tools. While this is generally sufficient for routine virome analysis, the Linux command line offers broader capabilities.

Programs designed for Linux systems are typically published as open-source software and are available free of charge. Windows users can rely on the Windows Subsystem for Linux (WSL) or a virtual machine such as VirtualBox. Automated software suites for virus and phytopathogen identification include VirusDetect (Zheng et al., 2017), Virtool (Rott et al., 2017), Kodoja (Baizan-Edge et al., 2019), PhytoPipe (Hu et al., 2023), etc. Such solutions provide rapid results but may not account for dataset-specific characteristics. Better results can be achieved by using programming skills to write your own scripts. In biological data analysis, the most commonly used programming languages are Python and R.

Sequencing results are most often provided as FastQ files containing reads with associated quality scores. For visualizing this information, there is a free program, FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>), that generates detailed reports.

Regardless of the nucleic acid type or sequencing platform, an important step in virome bioinformatics analysis is preprocessing – a set of procedures designed to improve data quality. Inadequate preprocessing can lead to errors during viral genome assembly and hinder read mapping, ultimately distorting analysis results.

Preprocessing typically includes the following steps:

- Read filtering. This step involves trimming and removing low-quality and excessively short reads. The user specifies the required quality threshold and minimum read length. For Illumina data, commonly used thresholds include Q20 and Q30, corresponding to read accuracies of 99 and 99.9 %, respectively (Kutnjak et al., 2021). Filtering settings that are too strict can lead to partial

loss of informative data, while those that are too lenient can lead to errors during genome mapping and assembly. Filtering also includes the removal of adapter sequences added during library preparation. These operations can be performed using tools such as BBDuk (<https://sourceforge.net/projects/bbmap>), Trimmomatic (Bolger et al., 2014), etc.

- Merging paired reads based on the overlapping region. This step is only required for paired-end sequencing.
- Removing duplicate reads to reduce data volume and increase analysis speed.
- Removing reads mapping to the host genome. This step can significantly reduce data volume and improve data quality. However, it also removes viral sequences integrated into the host genome (Pappas et al., 2021). Moreover, the host plant genome assembly may contain errors, including misannotated viral fragments. Thus, host read removal should be applied with caution.

Virus identification in sequencing data relies on comparison of sequences with available databases (Lebas et al., 2022). Identification quality depends directly on database completeness and annotation accuracy. As the database is updated with new sequences, any dataset can be reassessed.

There are three main approaches that can complement each other (Fig. 2): read mapping to viral reference sequences; assembly and annotation of contigs; and taxonomic classification of reads without prior assembly.

The fastest and simplest approach to virus identification in HTS data is read mapping to viral reference sequences. Mapping is one of the basic operations in the analysis of second-generation sequencing results. The algorithm aligns reads to a reference sequence according to a specified identity threshold. Reads not meeting the threshold are discarded, while the remaining reads align to positions that match the reference sequence.

To detect plant viruses using this approach, it is necessary to first prepare a database of reference sequences, which should contain at least one complete genome of each virus capable of infecting the target plant.

Preprocessed reads are mapped to these reference sequences, producing information about coverage, identity levels, and other metrics (Lebas et al., 2022). With sufficient coverage, the mapped reads form a continuous sequence along the entire length of the reference genome.

Mapping is usually performed using tools such as BWA (Li, Durbin, 2009), Bowtie2 (Langmead, Salzberg, 2012), Minimap2 (Li, 2018), BBmap (<https://sourceforge.net/projects/bbmap>), STAR (Dobin, Gingeras, 2015), GraphMap (Sović et al., 2016), and Hisat2 (Kim et al., 2019). The settings and thresholds are very important. Viral genomes can differ markedly from available references due to their rapid mutation rates. If mapping settings are too

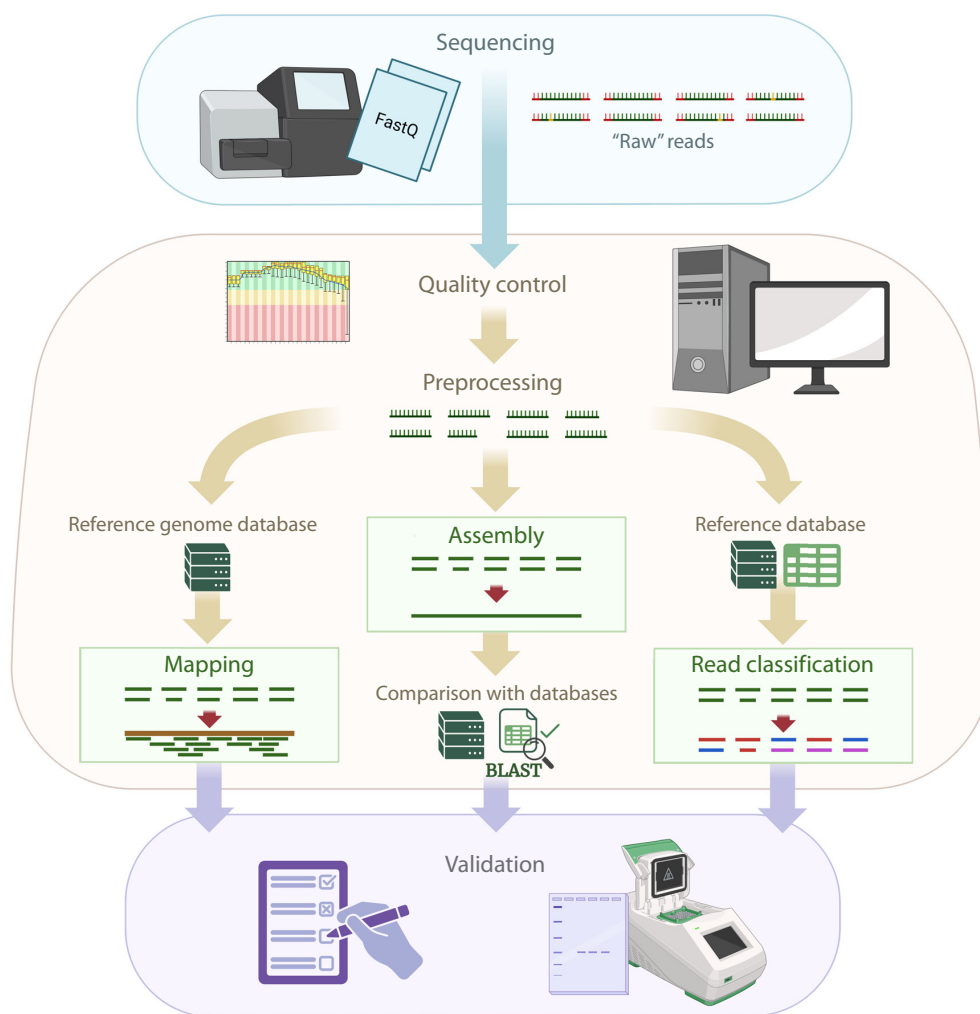


Fig. 2. A workflow for plant virome bioinformatics analysis.

strict, only isolates closely related to the reference sequence will be detected (Kutnjak et al., 2021). On the other hand, overly permissive settings may allow reads from one virus species to map to another, leading to false positives (Roy et al., 2018).

Mapping enables efficient detection of viruses even when viral read abundance is low. However, results should be interpreted cautiously, considering possible contamination during sample preparation and sequencing. For this reason, threshold values for read count or reference coverage should be established to determine when a sample is considered virus-negative (Belkina et al., 2023).

A second approach to virus identification using HTS is based on *de novo* assembly of sequences from overlapping reads – contigs – and their comparison with annotated databases. Assembly enables the discovery of novel virus species and new hosts for known viruses, as well as the characterization of genomic variants. Furthermore, this approach eliminates the risk of false negatives associated with a possible incompleteness of the database of reference genomes for viruses of the target plant.

Contig assembly can be performed using various algorithms, each with distinct advantages and limitations (White et al., 2017). With sufficient sequencing depth, contigs may correspond to nearly complete viral genomes, excluding short 5' and 3' end regions. The efficiency of assembly by the same algorithm varies across different datasets, therefore, it is impossible to choose a single universal one (Sutton et al., 2019; Shvets et al., 2022). Assembly quality can be positively affected by pre-removal of host plant reads and optimization of tool settings (Kutnjak et al., 2021).

Common assemblers for short reads include SPAdes (Bankevich et al., 2012), Velvet (Zerbino, Birney, 2008), Geneious (<https://www.geneious.com>), CLC (<https://digitalinsights.qiagen.com>), MIRA (Chevreux et al., 1999), ABySS (Simpson et al., 2009), IDBA-UI (Peng et al., 2012), SOAPdenovo2 (Luo et al., 2012); for long reads, Canu (Koren et al., 2017), Falcon (Chin et al., 2016), and Pomoxis (<https://github.com/nanoporetech/pomoxis>). Some algorithms, such as SPAdes and Unicycler (Wick et al., 2017), perform well with hybrid datasets combining short and long reads (Pappas et al., 2021).

To increase reliability, it is recommended to use several assembly algorithms. Note that depending on the library size and the chosen algorithm, the assembly may take a long time and require a lot of RAM.

After assembly, contigs are compared against annotated sequence databases. Result reliability depends largely on database completeness and quality. Many of them are available online, but it is often more convenient to deploy the database on a cloud server or locally when storage capacity allows.

The most convenient, up-to-date, and frequently used databases are maintained by NCBI; by early 2025, they contained approximately six million nucleotide sequences and one million amino acid sequences (Sayers et al., 2025). NCBI provides both the curated RefSeq database, which includes one genome per organism (<https://www.ncbi.nlm.nih.gov/refseq>), and the community-updated GenBank database, which contains the majority of publicly available sequences (<https://www.ncbi.nlm.nih.gov/genbank>). However, GenBank entries often lack thorough quality control or annotation verification. Thus, comparisons of contigs against the GenBank database can lead to misidentifications. Using the database of reference virus genomes, Viral RefSeq, will potentially yield the most reliable results, but its updates occur slowly and may lack a significant number of recently discovered viruses.

The most widely used tool for comparing contigs against databases is BLAST, which integrates several algorithms for analyzing nucleotide and amino acid sequences. The fastest algorithm that can be used to determine which virus in the database corresponds to a particular contig is megablast. For discovering novel viruses, blastx or tblastx are preferred due to their ability to detect distant sequence similarities. The main drawback of blastx and especially tblastx is the low speed of analysis; therefore, when time is limited, the faster DIAMOND algorithm can be used (Buchfink et al., 2015). However, some researchers believe that it is less effective for virus identification (Kutnjak et al., 2021).

BLAST can be run via the NCBI website (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>), which is convenient for a small number of sequences and requires no local computing resources. BLAST is also integrated into graphical environments such as Geneious Prime and CLC Genomics Workbench. For command-line analysis in Linux, macOS, and Windows, the BLAST+ suite is used (Camacho et al., 2009).

As a result of comparison with the virus sequence database, contigs are assigned to specific species with varying reliability. The level of reliability is typically indicated using the E-value, which reflects the probability of a random match between a contig and a database entry (VanderWeele, Ding, 2017). The closer the E-value to 0, the more reliable the identification.

Approaches based on taxonomic classification of reads without prior assembly or mapping are less common in

plant virome studies. However, if the necessary computing resources are available, they can produce results much faster than the approaches discussed above. Popular taxonomic read classifiers include Kraken2 (Wood et al., 2019), Kaiju (Menzel et al., 2016), CLARK-S (Ounit, Lonardi, 2016), and Centrifuge (Kim et al., 2016). These tools typically require large databases, substantial computing power, and proficiency with the Linux command line.

Validation and interpretation of virome analysis

After detecting reads from a particular virus in a library using any bioinformatics method, the results must be interpreted correctly. Because sequencing quality depends on many factors, no universal criteria exist for determining how many reads or contigs reliably indicate infection (Villamor et al., 2019). Moreover, analyzing the same dataset using different software or different settings may lead to different results.

Therefore, each case requires a comprehensive and critical evaluation of the data. During the final steps of virome analysis, the researcher's expertise and biological understanding become crucial. In particular, the researcher must understand that not all viruses detected in the samples are pathogens of the sampled plant. Virome data may include viruses of insects, fungi, and bacteria associated with the plant, as well as human or animal viruses (Cobbin et al., 2021). This is particularly important to remember when results appear to indicate the discovery of previously undescribed viruses.

To reliably identify plant viruses in metagenomic data, bioinformatics analysis results should be confirmed by molecular genetic methods such as PCR or reverse transcription-PCR (RT-PCR), depending on the viral genome type. When a virus is detected in a sample both by virome sequencing and by molecular diagnostics, identification can be considered highly reliable.

Conclusion

Virome analysis using high-throughput sequencing is a powerful method that supports both fundamental and applied research in agriculture. Each year, high-throughput sequencing becomes more accessible as analytical and instrumental capabilities expand. In the future, once an appropriate regulatory framework is established, this method may form the basis for decision-making in national food security strategies.

References

- Atallah S.S., Gómez M.I., Fuchs M.F., Martinson T.E. Economic impact of grapevine leafroll disease on *Vitis vinifera* cv. Cabernet Franc in Finger Lakes vineyards of New York. *Am J Enol Vitic.* 2012;63(1): 73-79. doi 10.5344/ajev.2011.11055
- Baizan-Edge A., Cock P., MacFarlane S., McGavin W., Torrance L., Jones S. *Kodoja*: a workflow for virus detection in plants using *k-mer* analysis of RNA-sequencing data. *J Gen Virol.* 2019;100(3):533-542. doi 10.1099/jgv.0.001210

- Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., ... Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-477. doi 10.1089/cmb.2012.0021
- Belkina D., Karpova D., Porotikova E., Lifanov I., Vinogradova S. Grapevine virome of the Don ampelographic collection in Russia has concealed five novel viruses. *Viruses.* 2023;15(12):2429. doi 10.3390/v15122429
- Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120. doi 10.1093/bioinformatics/btu170
- Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59-60. doi 10.1038/nmeth.3176
- Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10(1):421. doi 10.1186/1471-2105-10-421
- Chevreaux B., Wetter T., Suhai S. Genome sequence assembly using trace signals and additional sequence information. In: Proceedings of the German Conference on Bioinformatics (GCB) 99. Hannover, 1999;45-56
- Chin C.-S., Peluso P., Sedlazeck F.J., Nattestad M., Concepcion G.T., Clum A., Dunn C., ... Luo C., Ecker J.R., Cantu D., Rank D.R., Schatz M.C. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* 2016;13(12):1050-1054. doi 10.1038/nmeth.4035
- Cholet F., Ijaz U.Z., Smith C.J. Reverse transcriptase enzyme and priming strategy affect quantification and diversity of environmental transcripts. *Environ Microbiol.* 2020;22(6):2383-2402. doi 10.1111/1462-2920.15017
- Cobbin J.C., Charon J., Harvey E., Holmes E.C., Mahar J.E. Current challenges to virus discovery by meta-transcriptomics. *Curr Opin Virol.* 2021;51:48-55. doi 10.1016/j.coviro.2021.09.007
- Crossley B.M., Bai J., Glaser A., Maes R., Porter E., Killian M.L., Clement T., Toohey-Kurth K. Guidelines for Sanger sequencing and molecular assay monitoring. *J Vet Diagn Invest.* 2020;32(6):767-775. doi 10.1177/1040638720905833
- Dobin A., Gingeras T.R. Mapping RNA-seq reads with STAR. *Curr Protoc Bioinformatics.* 2015;51(1):11.14.1-11.14.19. doi 10.1002/0471250953.bi1114s1
- Fall M.L., Xu D., Lemoyne P., Clément G., Moffett P., Ritzenthaler C. An innovative binding-protein-based dsRNA extraction method: comparison of cost-effectiveness of virus detection methods using high-throughput sequencing. *Mol Ecol Resour.* 2025;25(7):e14111. doi 10.1111/1755-0998.14111
- Filloux D., Dallot S., Delaunay A., Galzi S., Jacquot E., Roumagnac P. Metagenomics approaches based on virion-associated nucleic acids (VANA): an innovative tool for assessing without a priori viral diversity of plants. In: Lacomme C. (Ed.) Plant Pathology. Methods in Molecular Biology. Vol. 1302. New York: Humana Press, 2015; 249-257. doi 10.1007/978-1-4939-2620-6_18
- Fitzpatrick A.H., Rupnik A., O'Shea H., Crispie F., Keaveney S., Cotter P. High throughput sequencing for the detection and characterization of RNA viruses. *Front Microbiol.* 2021;12:621719. doi 10.3389/fmicb.2021.621719
- Gaafar Y.Z.A., Ziebell H. Comparative study on three viral enrichment approaches based on RNA extraction for plant virus/viroid detection using high-throughput sequencing. *PLoS One.* 2020;15(8):e0237951. doi 10.1371/journal.pone.0237951
- Gallo Y., Marin M., Gutiérrez P. Detection of RNA viruses in *Solanum quitoense* by high-throughput sequencing (HTS) using total and double stranded RNA inputs. *Physiol Mol Plant Pathol.* 2021;113:101570. doi 10.1016/j.pmp.2020.101570
- Han Y., He J., Li M., Peng Y., Jiang H., Zhao J., Li Y., Deng F. Unlocking the potential of metagenomics with the PacBio high-fidelity sequencing technology. *Microorganisms.* 2024;12(12):2482. doi 10.3390/microorganisms12122482
- Hu X., Hurtado-Gonzales O.P., Adhikari B.N., French-Monar R.D., Malapi M., Foster J.A., McFarland C.D. PhytoPipe: a phytosanitary pipeline for plant pathogen detection and diagnosis using RNA-seq data. *BMC Bioinformatics.* 2023;24(1):470. doi 10.1186/s12859-023-05589-2
- Javaran V.J., Moffett P., Lemoyne P., Xu D., Adkar-Purushothama C.R., Fall M.L. Grapevine virology in the third-generation sequencing era: from virus detection to viral epitranscriptomics. *Plants.* 2021;10(11):2355. doi 10.3390/plants10112355
- Jones R.A.C. Global plant virus disease pandemics and epidemics. *Plants.* 2021;10(2):233. doi 10.3390/plants10020233
- Kim D., Song L., Breitwieser F.P., Salzberg S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 2016;26(12):1721-1729. doi 10.1101/gr.210641.116
- Kim D., Paggi J.M., Park C., Bennett C., Salzberg S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37(8):907-915. doi 10.1038/s41587-019-0201-4
- Koren S., Walenz B.P., Berlin K., Miller J.R., Bergman N.H., Phillippy A.M. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722-736. doi 10.1101/gr.215087.116
- Kurochkin V.E., Alekseev Y.I., Petrov D.G., Evstrapov A.A. Domestic devices for molecular genetic analysis: developments of the IAP RAS and SINTOL LLC. *Russ Mil Med Acad Rep.* 2021;40(3):69-74. doi 10.17816/rmmar76918 (in Russian)
- Kutnjak D., Tamisier L., Adams I., Boonham N., Candresse T., Chiumenti M., De Jonghe K., ... Rollin J., Rott M., Schumpp O., Massart S., Haegeman A. A primer on the analysis of high-throughput sequencing data for detection of plant viruses. *Microorganisms.* 2021;9(4):841. doi 10.3390/microorganisms9040841
- Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359. doi 10.1038/nmeth.1923
- Lebas B., Adams I., Al Rwahnih M., Baeyen S., Bilodeau G.J., Blouin A.G., Boonham N., ... Vicente C.S.L., Vossenbergh B.T.L.H., Wetzel T., Ziebell H., Massart S. Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: a step-by-step description. *EPPO Bull.* 2022;52(2):394-418. doi 10.1111/epp.12863
- Lee H.-K., Kim S.-Y., Yang H.-J., Lee D.-S., Kwon B., Lee D.-Y., Oh J., Lee S.-H. The detection of plant viruses in Korean ginseng (*Panax ginseng*) through RNA sequencing. *Plant Pathol J.* 2020;36(6):643-650. doi 10.5423/PPJ.NT.07.2020.0137
- Lefeuve P., Martin D.P., Elena S.F., Shepherd D.N., Roumagnac P., Varsani A. Evolution and ecology of plant viruses. *Nat Rev Microbiol.* 2019;17(10):632-644. doi 10.1038/s41579-019-0232-3
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094-3100. doi 10.1093/bioinformatics/bty191
- Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760. doi 10.1093/bioinformatics/btp324
- Liu S., Rodriguez J.S., Munteanu V., Ronkowski C., Sharma N.K., Alser M., Andreade F., ... Ganda E., Davenport E.R., Pop M., Koslicki D., Mangul S. Analysis of metagenomic data. *Nat Rev Methods Primers.* 2025;5(1):5. doi 10.1038/s43586-024-00376-6
- Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., ... Li Y., Yang H., Wang J., Lam T.-W., Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience.* 2012;1(1):18. doi 10.1186/2047-217X-1-18

- Maina S., Donovan N.J., Plett K., Bogema D., Rodoni B.C. High-throughput sequencing for plant virology diagnostics and its potential in plant health certification. *Front Hortic.* 2024;3:1388028. doi 10.3389/fhort.2024.1388028
- Maliogka V.I., Minafra A., Saldarelli P., Ruiz-García A.B., Glasa M., Katis N., Olmos A. Recent advances on detection and characterization of fruit tree viruses using high-throughput sequencing technologies. *Viruses.* 2018;10(8):436. doi 10.3390/v10080436
- Massart S., Olmos A., Jijakli H., Candresse T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* 2014;188:90-96. doi 10.1016/j.virusres.2014.03.029
- Massart S., Adams I., Al Rwahnih M., Baeyen S., Bilodeau G.J., Blouin A.G., Boonham N., ... van de Vossen B.T.L.H., Westenberg M., Wetzel T., Ziebell H., Lebas B.S.M. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. *Peer Commun J.* 2022;2:e62. doi 10.24072/pcjournal.181
- Menzel P., Ng K.L., Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun.* 2016;7(1):11257. doi 10.1038/ncomms11257
- Moubset O., François S., Maclot F., Palanga E., Julian C., Claude L., Fernandez E., ... Massart S., Ogliastro M., Martin D.P., Filloux D., Roumagnac P. Virion-associated nucleic acid-based metagenomics: a decade of advances in molecular characterization of plant viruses. *Phytopathology.* 2022;112(11):2253-2272. doi 10.1094/PHYTO-03-22-0096-RVW
- Nabeshima T., Abe J. High-throughput sequencing indicates novel *Varicosavirus*, *Emaravirus*, and *Deltapartitivirus* infections in *Vitis coignetiae*. *Viruses.* 2021;13(5):827. doi 10.3390/v13050827
- Ounit R., Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics.* 2016;32(24):3823-3825. doi 10.1093/bioinformatics/btw542
- Panno S., Davino S., Caruso A.G., Bertacca S., Crnogorac A., Mandić A., Noris E., Matić S. A review of the most common and economically important diseases that undermine the cultivation of tomato crop in the mediterranean basin. *Agronomy.* 2021;11(11):2188. doi 10.3390/agronomy11112188
- Pappas N., Roux S., Hölzer M., Lamkiewicz K., Mock F., Marz M., Dutilh B.E. Virus bioinformatics. In: Encyclopedia of Virology. Elsevier, 2021;124-132. doi 10.1016/B978-0-12-814515-9.00034-5
- Pecman A., Kutnjak D., Gutiérrez-Aguirre I., Adams I., Fox A., Boonham N., Ravnikar M. Next generation sequencing for detection and discovery of plant viruses and viroids: comparison of two approaches. *Front Microbiol.* 2017;8:1998. doi 10.3389/fmicb.2017.01998
- Peng Y., Leung H.C.M., Yiu S.M., Chin F.Y.L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420-1428. doi 10.1093/bioinformatics/bts174
- Piombo E., Abdelfattah A., Droby S., Wisniewski M., Spadaro D., Schena L. Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. *Microorganisms.* 2021;9(1):188. doi 10.3390/microorganisms9010188
- Rang F.J., Kloosterman W.P., de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 2018;19(1):90. doi 10.1186/s13059-018-1462-9
- Roossinck M.J., Martin D.P., Roumagnac P. Plant virus metagenomics: advances in virus discovery. *Phytopathology.* 2015;105(6):716-727. doi 10.1094/PHYTO-12-14-0356-RVW
- Rose R., Constantinides B., Tapinos A., Robertson D.L., Prosperi M. Challenges in the analysis of viral metagenomes. *Virus Evol.* 2016; 2(2):vew022. doi 10.1093/ve/vew022
- Rott M., Xiang Y., Boyes I., Belton M., Saeed H., Kesanakurti P., Hayes S., Lawrence T., Birch C., Bhagwat B., Rast H. Application of next generation sequencing for diagnostic testing of tree fruit viruses and viroids. *Plant Dis.* 2017;101(8):1489-1499. doi 10.1094/PDIS-03-17-0306-RE
- Roy S., Coldren C., Karunamurthy A., Kip N.S., Klee E.W., Lincoln S.E., Leon A., Pullambhatla M., Temple-Smolkin R.L., Voelkerding K.V., Wang C., Carter A.B. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. *J Mol Diagn.* 2018;20(1):4-27. doi 10.1016/j.jmoldx.2017.11.003
- Rubino L., Abrahamian P., An W., Aranda M.A., Ascencio-Ibañez J.T., Bejerman N., Blouin A.G., ... Whitfield A.E., Wylie S.J., Yang C., Zerbini F.M., Zhang S. Summary of taxonomy changes ratified by the International Committee on Taxonomy of Viruses from the Plant Viruses Subcommittee, 2025. *J Gen Virol.* 2025;106(7):002114. doi 10.1099/jgv.0.002114
- Sayers E.W., Beck J., Bolton E.E., Brister J.R., Chan J., Connor R., Feldgarden M., ... Wang J., Ye J., Zellers E., Schneider V.A., Pruitt K.D. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* 2025;53(D1):D20-D29. doi 10.1093/nar/gkae979
- Shvets D., Sandomirsky K., Porotikova E., Vinogradova S. Metagenomic analysis of ampelographic collections of Dagestan revealed the presence of two novel grapevine viruses. *Viruses.* 2022;14(12):2623. doi 10.3390/v14122623
- Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birrol I. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 2009;19(6):1117-1123. doi 10.1101/gr.089532.108
- Sović I., Šikić M., Wilm A., Fenlon S.N., Chen S., Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun.* 2016;7(1):11307. doi 10.1038/ncomms11307
- Sun K., Liu Y., Zhou X., Yin C., Zhang P., Yang Q., Mao L., Shentu X., Yu X. Nanopore sequencing technology and its application in plant virus diagnostics. *Front Microbiol.* 2022;13:939666. doi 10.3389/fmicb.2022.939666
- Sutton T.D.S., Clooney A.G., Ryan F.J., Ross R.P., Hill C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome.* 2019;7(1):12. doi 10.1186/s40168-019-0626-5
- Turco S., Golyaev V., Seguin J., Gilli C., Farinelli L., Bollner T., Schumpp O., Pooggin M.M. Small RNA-omics for virome reconstruction and antiviral defense characterization in mixed infections of cultivated *Solanum* plants. *Mol Plant Microbe Interact.* 2018; 31(7):707-723. doi 10.1094/MPMI-12-17-0301-R
- VanderWeele T.J., Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med.* 2017;167(4):268-274. doi 10.7326/M16-2607
- Villamor D.E.V., Ho T., Al Rwahnih M., Martin R.R., Tzanetakis I.E. High throughput sequencing for plant virus detection and discovery. *Phytopathology.* 2019;109(5):716-725. doi 10.1094/PHYTO-07-18-0257-RVW
- Vinogradova S., Porotikova E., Navrotskaya E., Galbacs Z.N., Massart S., Varallyay E. The first virome of a Russian vineyard. *Plants.* 2023;12(18):3292. doi 10.3390/plants12183292
- Vivek A.T., Zahra S., Kumar S. From current knowledge to best practice: a primer on viral diagnostics using deep sequencing of virus-derived small interfering RNAs (vsiRNAs) in infected plants. *Methods.* 2020;183:30-37. doi 10.1016/j.ymeth.2019.10.009
- Wang Z., Qin L., Liu J., Jiang L., Zou X., Chen X., Song F., Dai H., Hou Y. Forensic nanopore sequencing of microhaplotype markers using QitanTech's QNome. *Forensic Sci Int Genet.* 2022;57:102657. doi 10.1016/j.fsigen.2021.102657
- White D.J., Wang J., Hall R.J. Assessing the impact of assemblers on virus detection in a *de novo* metagenomic analysis pipeline. *J Comput Biol.* 2017;24(9):874-881. doi 10.1089/cmb.2017.0008

- Wick R.R., Judd L.M., Gorrie C.L., Holt K.E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 2017;13(6):e1005595. doi 10.1371/journal.pcbi.1005595
- Wood D.E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019;20(1):257. doi 10.1186/s13059-019-1891-0
- Zerbino D.R., Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821-829. doi 10.1101/gr.074492.107
- Zheng Y., Gao S., Padmanabhan C., Li R., Galvez M., Gutierrez D., Fuentes S., Ling K.-S., Kreuze J., Fei Z. VirusDetect: an automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology.* 2017;500:130-138. doi 10.1016/j.virol.2016.10.017
- Zhuravlyov V.S., Dolgikh V.V., Timofeev S.A., Gannibal F.B. RNA interference method in plant protection against insect pests. *Plant Prot News.* 2022;105(1):28-39. doi 10.31993/2308-6459-2022-105-1-15219
- Zubov V.V., Chemeris D.A., Vasilov R.G., Kurochkin V.E., Alekseev Ya.I. Brief history of high-throughput nucleic acid sequencing methods. *Biomics.* 2021;13(1):27-46. doi 10.31301/2221-6197.bmcs.2021-4

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

Received September 3, 2025. Revised October 31, 2025. Accepted October 31, 2025.

doi 10.18699/vjgb-26-36

Applicability of the StatFaRmer time series analysis tool in soybean (*Glycine max*) digital phenotyping

D.S. Ulyanov , A.A. Ulyanova , A.A. Kocheshkova , A.O. Blinkov , A.V. Arkhipov , Ya.S. Meglitskaya ,
N.Yu. Svistunova , G.I. Karlov , M.G. Divashuk 

All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia

 uldas1508@gmail.com

Abstract. Contemporary agrobiotechnology research increasingly relies on automated methods for capturing and interpreting morphophysiological and spectral plant characteristics – a field known as digital phenotyping. This approach aims to identify stable differences between genotypes cultivated under non-identical environmental conditions. We previously introduced StatFaRmer, an open-source tool that we further develop here for comprehensive analysis of temporal phenotypic datasets, with a primary focus on crops such as soybean (*Glycine max*). The tool implements automated data preprocessing procedures, including synchronization of timestamps across samples and removal of noise artifacts and outliers. These features are particularly relevant for multi-month experiments involving assessments of growth parameters, fluctuations in photosynthetic apparatus area, or other biometric indicators. Support for standardized data formats (XLSX, CSV) ensures compatibility with common phenotyping systems, simplifying cross-platform integration. Thus, the tool can integrate with widely used HTTP platforms (e.g., Traitmill, HyperAlxpert, Plant Accelerator), enabling data from diverse sources to be analyzed within a single pipeline. For soybean experiments, StatFaRmer provides customizable analysis of variance (ANOVA) with visualization of diagnostic parameters (normality of distribution, homogeneity of variances) and evaluation of effect significance between user-defined groups. An example application compares growth parameters across 20 soybean cultivars under controlled stress: the tool automatically aggregated data with uneven measurement frequencies (from 1 hour to 3 days), identified anomalies in hypocotyl elongation dynamics, and computed statistical significance between groups ($p < 0.01$). The tool has been tested on large-scale datasets (over 2,000 measurements per experiment). StatFaRmer is implemented as a Shiny-based web application, with step-by-step deployment guides for Windows and Linux. All processing stages – from raw data to final plots – are documented to ensure transparency and compliance with research reproducibility standards. Thus, StatFaRmer offers a specialized solution for statistical hypothesis testing in soybean digital phenotyping, reducing data preparation time and minimizing risks of error when handling non-stationary time series.

Key words: high-throughput plant phenotyping; phenotypic data visualization; time series analysis; digital phenotyping platforms; genotype-phenotype analysis; statistical analysis of phenotypic data; open-source software; automated data analysis

For citation: Ulyanov D.S., Ulyanova A.A., Kocheshkova A.A., Blinkov A.O., Arkhipov A.V., Meglitskaya Ya.S., Svistunova N.Yu., Karlov G.I., Divashuk M.G. Applicability of the StatFaRmer time series analysis tool in soybean (*Glycine max*) digital phenotyping. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):321-329. doi 10.18699/vjgb-26-36

Funding. This study was supported by the Ministry of Science and Higher Education of the Russian Federation (State Assignment FGUM-2025-0010).

Применимость инструмента для анализа временных рядов StatFaRmer при цифровом фенотипировании сои (*Glycine max*)

Д.С. Ульянов , А.А. Ульянова , А.А. Кочешкова , А.О. Блинков , А.В. Архипов , Я.С. Меглицкая ,
Н.Ю. Свистунова , Г.И. Карлов , М.Г. Дивашук 

Всероссийский научно-исследовательский институт сельскохозяйственной биотехнологии, Москва, Россия

 uldas1508@gmail.com

Аннотация. Современные исследования в области агrobiотехнологий все чаще опираются на методы автоматизированной фиксации и интерпретации морфофизиологических и спектральных характеристик растений – направление, известное как цифровое фенотипирование. Этот подход направлен на обнаружение устойчивых различий между генотипами, культивируемыми в неидентичных условиях среды. Ранее был предложен StatFaRmer – инструмент с открытым кодом, совершенствующийся в рамках настоящей работы для

комплексного анализа временных фенотипических наборов данных, преимущественно ориентированный на изучение сельскохозяйственных культур, таких как соя (*Glycine max*). Разработанный инструмент реализует автоматизированные процедуры предобработки данных, включая синхронизацию временных меток между образцами и устранение шумовых артефактов и выбросов. Эти функции особенно актуальны для долгосрочных экспериментов, связанных с оценкой параметров роста, колебаний площади фотосинтетического аппарата или других биометрических показателей. Поддержка стандартизированных форматов данных (XLSX, CSV) обеспечивает совместимость с распространенными системами фенотипирования, упрощая кроссплатформенную интеграцию. Таким образом, инструмент может поддерживать интеграцию с популярными HTPP-платформами (например, Traitmill, HyperAlxpert, Plant Accelerator), что позволяет использовать данные, полученные из различных источников, в едином аналитическом конвейере. Для экспериментов с соей StatFaRmer предоставляет настраиваемый дисперсионный анализ (ANOVA) с визуализацией диагностических параметров (нормальность распределения, гомогенность дисперсий) и оценкой значимости эффектов между пользовательскими группами. Пример применения включает сравнение параметров роста 20 сортов сои в условиях контролируемого стресса: инструмент автоматически агрегировал данные с неравномерной частотой измерений (от 1 часа до 3 суток), идентифицировал аномалии в динамике удлинения гипокотилия и рассчитал статистическую значимость различий между группами ($p < 0.01$). Инструмент протестирован на масштабных наборах данных (более 2000 измерений на эксперимент). StatFaRmer реализован в виде веб-приложения на платформе Shiny с пошаговыми инструкциями для установки и запуска в ОС Windows и Linux. Все этапы обработки – от первичных данных до итоговых графиков – документируются, что обеспечивает прозрачность анализа и соответствие требованиям воспроизводимости исследований. Таким образом, StatFaRmer предлагает специализированное решение для статистической верификации гипотез в цифровом фенотипировании сои, сокращая время на подготовку данных и минимизируя риски ошибок при работе с нестационарными временными рядами.

Ключевые слова: высокопроизводительное фенотипирование растений; визуализация фенотипических данных; анализ временных рядов; цифровые платформы фенотипирования; генотип-фенотипический анализ; статистический анализ фенотипических данных; программное обеспечение с открытым исходным кодом; автоматизированный анализ данных

Introduction

Despite the rapid progress of genomics, phenotyping – the assessment of external plant traits and characteristics – remains a key bottleneck in accelerating breeding (Fasoula et al., 2020; Patel et al., 2023). Traditional methods based on manual measurements are labor-intensive, subjective, low-throughput, and often require destructive sampling, making them unsuitable for large-scale programs aimed at developing more productive and resilient cultivars (Atefi et al., 2021; Li L. et al., 2023).

The solution is high-throughput plant phenotyping (HTPP), a breakthrough technology in phenotypic research. This approach integrates modern sensor systems (hyperspectral, thermal, or fluorescence imaging, 3D reconstruction) and machine learning methods for automated, non-destructive, and precise plant analysis.

Traditional manual phenotyping suffers from major time and cost demands, limited scalability due to throughput constraints, subjectivity and variability, and a narrow set of measurable traits (Thrash et al., 2022; Abebe et al., 2023; Anand et al., 2023; Yuan et al., 2023). By contrast, digital phenotyping offers substantial efficiency gains, continuous and non-destructive measurements, improved accuracy and objectivity, and the ability to quantify dozens of traits simultaneously, greatly increasing data volumes (Li D. et al., 2021; Anand et al., 2023; Buelvas et al., 2023; Gyan et al., 2023; Li L. et al., 2023; Lu et al., 2023; Wang et al., 2024).

Modern HTP platforms generate massive volumes of heterogeneous data, including images (RGB, hyperspectral, thermal), 3D models, and sensor readings, reaching terabytes per day. The key challenges are not only volume but

also multimodality, the spatiotemporal nature of the data, and the need for storage and data management. The lack of standardized formats and the challenges of ensuring data quality further complicate analysis (Coppens et al., 2017; Morota et al., 2019; Danilevicz et al., 2021; Gill et al., 2022; Ninomiya, 2022).

Statistical methods commonly used for phenotyping data include the following model families:

- univariate and multivariate methods for hypothesis testing and investigating relationships among plant traits. These include correlation analysis, factor analysis (FA), and principal component analysis (PCA) for dimensionality reduction and identification of key traits (Rahaman et al., 2015);
- random-effects regression models and models of multi-factor interactions (e.g., genotype–environment interaction) used to model dynamic phenotypic data (Morota et al., 2019);
- cluster analysis (hierarchical, k-means, etc.) to assess similarities and differences among plants and stress responses (Rahaman et al., 2015);
- machine learning and deep learning for processing large and complex data, including classification, segmentation, feature localization in images, and prediction of phenotypic responses. These approaches have advantages such as revealing non-obvious correlations, but they often require large training datasets and offer limited interpretability (Coppens et al., 2017; Ubbens et al., 2025).

A variety of tools exist for comprehensive application of the above methods. Among commercial solutions, one option is Genstat, which provides comprehensive statistical

analysis capabilities tailored for plant breeding. Its features include spatial analysis, multi-treatment experimental designs, genomic selection, and stability analysis tools adapted to agricultural data. HortControl is the default tool supplied with the TraitFinder platform (HortControl – Plant Data Management Software, 2025) used by the authors. It offers integrated data management for digital phenotyping with BrAPI compatibility and includes some analysis tools, but advanced analyses may require exporting .csv files and using third-party tools.

There are also open-source tools. For example, R packages such as ASReml-R and pcvr offer specialized functions for longitudinal growth modeling and spatial analysis of phenotyping data (Sumner, 2025). These tools implement advanced mixed-effects models and functional data analysis methods but require programming expertise. Another example is PIPPA (plant image phenotyping and analysis), which provides web access to high-performance computing resources for image analysis pipelines (Coppens et al., 2017).

To address similar tasks in longitudinal studies (a research approach in which the same group of objects is observed repeatedly over an extended period), we developed StatFaRmer – an open R-based web tool with a two-stage architecture: a Master Wizard for data processing and a Main Application for statistical analysis that does not require programming skills. At present, StatFaRmer includes:

- Master Wizard: project validation, DBSCAN clustering, outlier filtering (Z-criterion/IQR), and technical data aggregation;
- Main Application: automatic model selection (ANOVA/mixed models/splines), multifactor analysis with post hoc tests, diagnostics of normality and homogeneity of variances, and analysis of effect sizes and growth curves;
- interactive visualization with Plotly and high-resolution export (SVG, PNG, PDF);
- specialized modules: effect size analysis and growth summaries;
- export of results in tabular formats.

The tool is optimized for longitudinal studies under controlled conditions, supports two-factor interactions in ANOVA models, and ensures transparency of statistical conclusions. StatFaRmer fills the gap between specialized software and researchers' needs by offering flexible analysis of complex phenotypic data.

Compared with standard HortControl, our R/Shiny-based system offers greater flexibility in tailoring analyses to specific experiments and provides a deeper and more automated statistical pipeline, whereas HortControl focuses more on sensor data and basic calculations.

By comparison with existing statistical packages and platforms (Genstat, ASReml-R, pcvr), these are powerful but general-purpose statistical environments. They can do everything required but demand substantial statistical knowledge and/or programming skills to specify correct models, especially for complex spatial and temporal data. In contrast,

StatFaRmer automatically generates complex ANOVA models, including all interactions, freeing researchers from routine and error-prone coding. Also, unlike generic packages, StatFaRmer is designed out of the box for the specifics of phenotypic data (repeated measures, outliers). Another advantage is interactivity and visualization: the ability to subset data interactively (by genotypes, treatments, temporal clusters) and obtain instantaneous results for hypothesis testing, which is a key benefit over traditional packages, where such analyses require constantly rewriting code.

In this study, an updated version of StatFaRmer (Ulyanov et al., 2025) is applied to data obtained from digital phenotyping of a key agricultural crop – soybean (*Glycine max*). The tool enables integration with platforms employing diverse sensor systems (multispectral scanners, 3D cameras) thanks to support for standardized data formats (XLSX, CSV) and flexible parameter settings. This allows it to be adapted to data collected both under controlled conditions (e.g., climate chambers) and in field experiments, where irregular measurement schedules and heterogeneous sources require automated preprocessing. Standardization of input data and StatFaRmer's modular architecture ensure compatibility with systems such as Traitmill and PhytoOracle, opening avenues for cross-platform studies.

The aim of this work was to present the capabilities of StatFaRmer, to consistently demonstrate its functional potential for readers, and to describe enhancements implemented since the previous publication.

Materials and methods

Experimental design and plant material. The study employed 50 soybean (*G. max*) cultivars of diverse origin (Russian and international breeding), differing in maturity time, growth type, and recommended cultivation regions. Seeds were treated with a fungicide before sowing, then sown to a depth of 2 cm in moist peat in 500 ml pots (four seeds per pot); after emergence, three plants were retained per pot.

Plants were cultivated under controlled climate chamber conditions throughout the entire growing cycle. The experiment included two treatments differing in lighting regimes while keeping other microclimate parameters identical (temperature, humidity, nutrient levels). Each treatment comprised two biological replicates, yielding a total sample of 200 plants (50 cultivars × 2 treatments × 2 replicates). Routine care included watering with room-temperature water during the first 7–10 days, then as needed, and daily application of mineral fertilizer after the appearance of true leaves. For phenotyping, plants were randomized into groups of 12 individuals; the final number of groups was 9 per treatment. A detailed description of experimental procedures and cultivation parameters is provided in (Ulyanov et al., 2025).

Digital phenotyping. Scanning was performed on a TraitFinder phenotyping system (Phenospex, Netherlands) with two multispectral scanners operating in the following wavelength ranges: red (R) 624–634 nm, green (G)

530–540 nm, blue (B) 465–485 nm, and near-infrared (NIR) 720–750 nm.

Plant phenotyping followed a scheme in which the allocation of cultivars and treatments to blocks was recorded in the embedded HortControl software and duplicated in StatFaRmer tables for subsequent analysis. Plants were distributed across the experimental area represented as a layout – a virtual scheme reflecting the physical arrangement of blocks and plants. All plants were divided into blocks, each corresponding to one table with a unique identifier – 9 blocks in total. Blocks were further divided into units corresponding to individual plants for data collection. In HortControl, blocks were reserved for the experiment and linked to biological information (genotype, treatment) through metadata assignment.

Prior to scanning, plants were transported from climate chambers to a dedicated phenotyping table, maintaining identical positioning within predefined digital coordinates. The procedure was performed in a darkened room with the air conditioning system turned off to avoid artifacts caused by airflow or changes in illumination. Each plant was positioned so that its center of mass coincided with the geometric center of the scanning area, minimizing overlap of leaves and shoots with neighboring samples. To stabilize plants in pots, support structures made of green plastic rods and black wire (spectrally neutral in the PlantEye sensor range) were used.

Five consecutive scans were performed during the experiment, with a total observation period of 20 days. We recorded overall plant height and spectral indices (NDVI, PSRI), computed from the multispectral sensor as follows:

Plant height average (mm): the average height of the top 10 % of plant points, minimizing the influence of external factors (e. g., wind).

Spectral indices:

NDVI (normalized difference vegetation index):

$$NDVI = \frac{NIR - RED}{NIR + RED}.$$

Range: –1 to 1. Values > 0.66 indicate healthy photosynthetic tissues.

PSRI (plant senescence reflectance index):

$$PSRI = \frac{RED - BLUE}{NIR}.$$

Used to evaluate leaf senescence.

Data preparation. Users can upload their own experiments by placing projects in StatFaRmer/data/project_NAME/ after preparing the data as follows.

1. **Base experiment archive.** The project must contain a ZIP archive (*_data.zip) including a CSV table exported from HortControl with the raw data.
2. **Manual adjustment table.** A *_handmade.csv file is required with the following columns:
 - V.T.R – unique identifier of the biological sample (cultivar/treatment/replicate);

- Treatment – treatment parameters (overrides the corresponding column in the source table);
- Cultivar.

These CSV files are prepared manually to enter plant metadata. For each plant's unique identifier, specify its cultivar and treatment; other per-plant data may also be provided.

3. **Spatial mapping table.** The *_translation.csv file must include:

- V.T.R – sample identifier;
- T.X:Y – sample coordinates corresponding to the unit column in the source data.

These CSV files are prepared manually so the program can link the imported data in item 1 with the experimental sample descriptions in item 2. To generate this table, match each plant's V.T.R with the corresponding coordinates in the HortControl database. Unit coordinates are encoded as follows:

- (block number of the plant group):
- (row number starting from the barcode):
- (plant index in the row, left to right).

4. **Additional grouping (optional).** A groups.xlsx file with a mandatory Cultivar column.

Other columns can be used as ANOVA factors and must contain only Latin letters, digits, and underscores.

All additional columns in the specified tables are preserved and available for analysis as independent factors.

Analysis configuration is carried out via the Master Wizard GUI, including:

- selecting a project from the available list with validation;
- tuning DBSCAN eps (time interval between measurements);
- choosing the method for filtering/winsorizing/annotating outliers (IQR or Z-score) based on the specified experimental cell structure;
- configuring technical aggregation (median or mean);
- exporting and importing experiment configurations to ensure processing transparency.

StatFaRmer effectively processes high-frequency time series with non-uniform measurement intervals, automatically correcting artifacts and aggregating data to reduce errors in longitudinal studies. The tool includes a multifactor analysis of variance (ANOVA) module with automatic checking of statistical assumptions. The generated interactive visualizations allow researchers to focus on interpretation while minimizing technical analysis hurdles.

Results

Clustering, outlier filtering, and data grouping

Clustering and outlier filtering are critical steps in the initial processing of phenotypic data. Outliers (anomalies) are data points that deviate substantially from the overall pattern, potentially due to technical artifacts (e. g., transient sensor shading) or biological factors (local tissue damage). This step removes noise arising from different phenotyping start

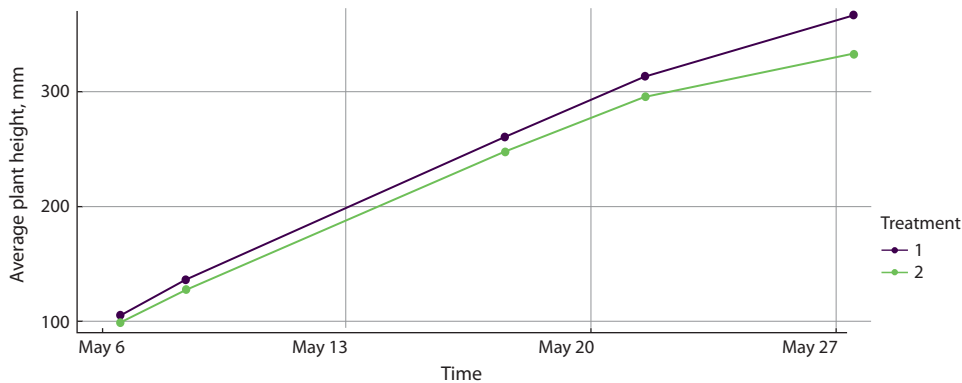


Fig. 1. Dynamics of the “average plant height” parameter between different experimental treatments. Figure 1 shows two plots of average plant height over time for treatments 1 and 2. All 50 cultivars in two replicates are presented for rapid preliminary analysis of height distribution patterns under treatments 1 and 2.

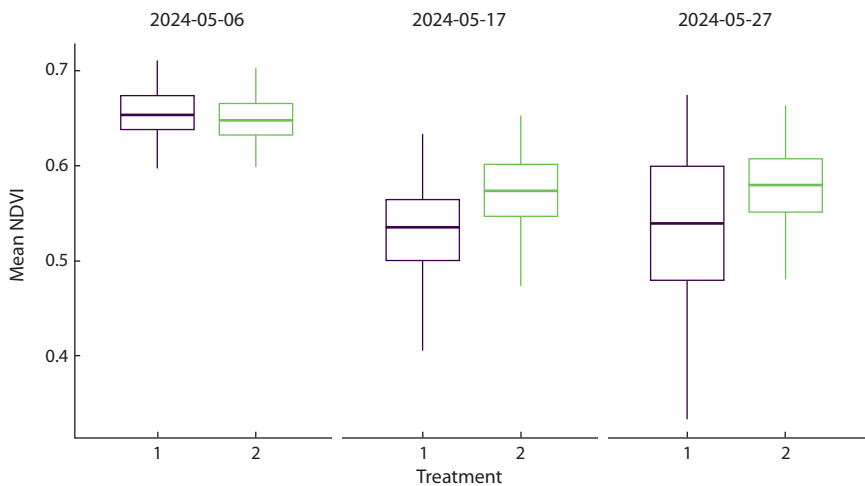


Fig. 2. Comparison of mean NDVI across three time points – at the beginning, middle, and end of the experiment – for the two treatments.

times across experimental blocks, sensor errors, and external influences (e.g., wind), as well as large sporadic outliers, such as foreign objects in the camera’s field of view. At this stage the program also performs data clustering; grouping and comparisons can be conducted under different parameters – conditions, treatments, and others. Outlier filtering in StatFaRmer can be toggled in the graphical interface by checking or unchecking “plot with outliers”, allowing users to control this aspect of the analysis. This processing step is illustrated in Figure 1.

StatFaRmer enables comparative analysis of plant sets by aggregating samples by treatment, cultivar, or any additional user-specified parameters in the accompanying data tables. The plots demonstrate that the mean height of plants grown under treatment 1 generally exceeds that under treatment 2. A substantial difference is apparent from the outset; given the large number of samples (100 per treatment) and phenotyping not starting on day one, this indicates that differences between treatments emerge rather quickly.

ANOVA

StatFaRmer provides automatic selection of statistical models: classical ANOVA for balanced designs (≤ 10 factor levels), mixed models for unbalanced designs (11–20 factor levels), and spline models for time series analysis. The program attempts to select an optimal analysis scheme given all specified factors and their pairwise interactions, while advanced settings allow the user to fix the model type.

When analyzing Figure 1, an important question arises: is the accelerated growth under treatment 1 a positive sign or an indicator of plant stress. The NDVI indicator can be used to assess plant condition, providing an approximate measure of photosynthetic activity and overall plant health. Higher NDVI values indicate better plant condition and more intense photosynthesis, whereas lower values may indicate stress. Figure 2 shows the change in mean NDVI over time, averaged across all cultivars.

In Figure 2, NDVI does not show significant differences between treatments at the first time point; subsequently,

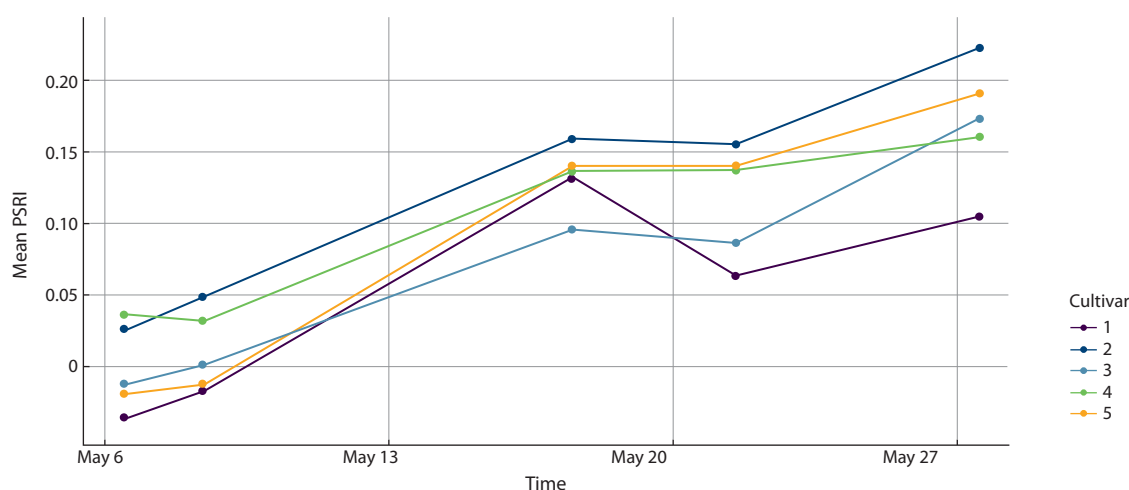


Fig. 3. Changes in general mean PSRI under treatment 1 for five cultivars.

NDVI for plants grown under treatment 2 is significantly higher, as indicated by the higher median. This can also be seen in the comparative group table in the “Export of results in tabular formats” section. This indicates more favorable growth conditions under treatment 2.

Subsampling for temporal cluster analysis

When working with large datasets, users need the ability to perform targeted analyses to uncover patterns independently. In the previous example, we compared NDVI values across treatments for all cultivars. Now suppose we need to select cultivars from this set that are more or less sensitive to the treatment. One could select field values to compare one cultivar with another, but checking each of the 50 cultivars in this manner would be laborious.

Instead, an alternative approach can be used. StatFaRmer’s data subsampling function allows, for example, comparing different cultivars under identical conditions rather than one cultivar across treatments. As an example, consider five cultivars and evaluate the PSRI parameter, which reflects plant senescence. The higher this parameter, the greater the stress experienced by the plant. Results of this comparison for treatment 1 are shown in Fig. 3.

Interactive visualization with grouping by features

Interactive visualization is key for users who need an adaptive tool for experiment analysis. For the selected cultivars, one can plot PSRI versus time with treatment taken into account. To study the dynamics of multiple groups, it is practical to use faceting – splitting plots into subplots generated by a consistent principle for subgroups of the same variable(s). Using the “grouping factor”, “selected cultivars”, and “faceting formula” fields (the latter being a syntactic template used in data visualization to describe how a dataset is split into subgroups (facets) when building multiple linked plots; a brief description can be found on the project’s GitHub page), we obtain a separate plot for each cultivar (Fig. 4).

This approach enables a detailed examination of how the senescence indicator changes over time across cultivars. For example, some cultivars exhibit a marked shift in dynamics – an increase in senescence until mid-experiment followed by stabilization – whereas in others, this inflection is weak or appears only at the final stages.

Export of results in tabular formats

StatFaRmer can present analysis results in tabular form and export plots at high resolution (300 DPI) in SVG, PNG, and PDF formats for publication. The Table below provides descriptive statistics for Figure 2, listing parameters characterizing each observed group at each time point.

Performance

StatFaRmer performance was tested on a laptop with the following configuration: AMD Ryzen 5 5500U with Radeon Graphics (2.10 GHz) and 8 GB RAM.

- Master Wizard: processing project project_NO3 (58,380 rows, 49 columns) takes 8–16 seconds with median aggregation, depending on the configuration file;
- Main Application: launch and analysis occur almost instantly due to caching of processed data;
- visualization is optimized for large datasets using density maps for > 2,000 measurements.

Logging (logger.R) and benchmarking (benchmark.R) of memory and time usage are implemented for module operation.

Thus, employing automated time series analysis methods reveals stable dynamic patterns in soybean data. Unlike tools that do not support time series analysis, StatFaRmer successfully handles this task while remaining a flexible tool that does not require programming skills. In long-term experiments, StatFaRmer demonstrates efficiency in processing high-frequency time series with non-uniform measurement intervals. In addition, the tool automatically performs artifact correction and data aggregation. These procedures are critical

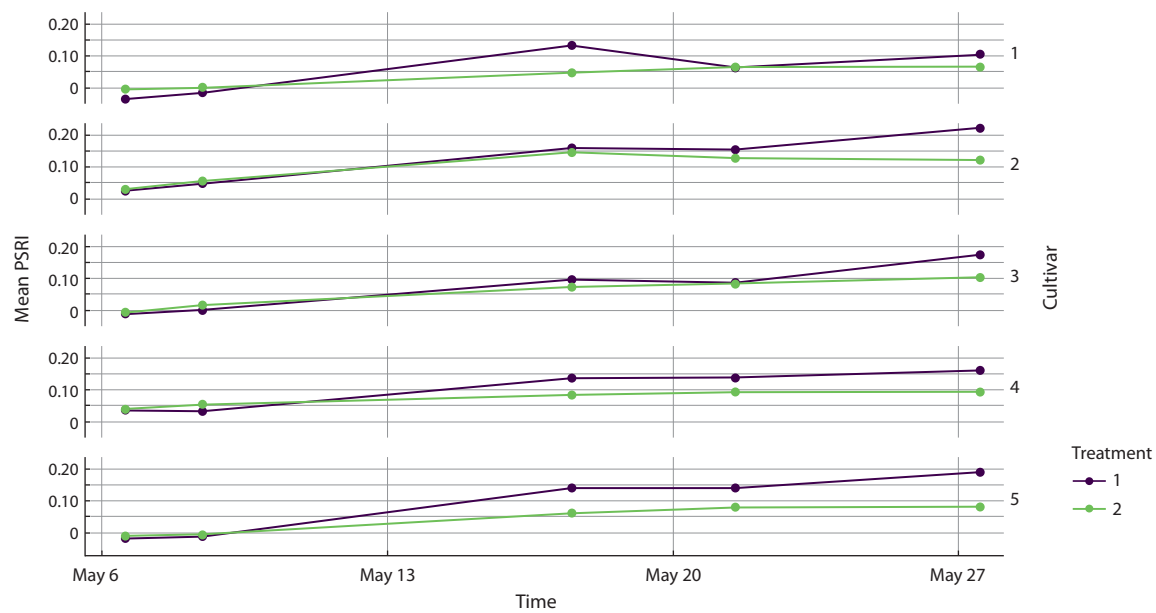


Fig. 4. Change in mean PSRI over time for cultivars 1–5.

Descriptive statistics for Figure 2

No.	Time point	Treatment	n	Avg.	Median	SD	SE	min	max	Q4	Q1	IQR	CV	Skewness	Kurtosis
1	1	1	100	0.655	0.653	0.0259	0.00259	0.596	0.731	0.637	0.674	0.0369	3.96	−0.00113	0.0668
2	1	2	100	0.648	0.647	0.0268	0.00268	0.56	0.724	0.633	0.665	0.0315	4.14	−0.0214	1.3
3	4	1	100	0.532	0.535	0.0467	0.00467	0.395	0.633	0.5	0.564	0.0636	8.77	−0.411	0.263
4	4	2	100	0.57	0.573	0.0457	0.00457	0.421	0.653	0.547	0.601	0.0542	8.02	−0.823	1.16
5	6	1	100	0.528	0.539	0.0838	0.00838	0.333	0.674	0.479	0.6	0.12	15.9	−0.477	−0.537
6	6	2	100	0.579	0.58	0.0396	0.00396	0.48	0.664	0.551	0.607	0.0567	6.84	−0.0886	−0.274

Note. The “Time point” column shows the analyzed groups, where 1, 4, and 6 denote the ordinal time points corresponding to the start, middle, and end of the experiment.

for minimizing errors in longitudinal studies. StatFaRmer also includes a multifactor ANOVA module with automatic diagnostics of statistical assumptions (e. g., tests of normality and homogeneity of variances). The tool’s high-resolution interactive visualizations compare favorably with most other platforms, where plots are typically static. As a result, researchers can focus on interpreting results rather than the technical details of analysis.

Discussion

Unlike several existing solutions such as HTPPheno (Hartmann et al., 2011) and IAP (Yang et al., 2020), which focus primarily on image processing, StatFaRmer provides extended integration with external phenotyping systems. This enables researchers to combine data from heterogeneous sources while minimizing manual effort. HTPPheno and IAP offer limited support for automated analysis of

large non-stationary time series; therefore, additional tools are often required for full statistical processing of such data. By contrast, StatFaRmer successfully automates the analysis of complex phenotypic time series generated by high-throughput phenotyping platforms.

In parallel with the development of StatFaRmer, other specialized tools have emerged to address adjacent digital phenotyping tasks. For example, the AllInOne Pre-processing package (Najafabadi et al., 2023) focuses on preprocessing field phenotyping data and provides image normalization, spatial analysis, and basic visualization functions. Although AllInOne effectively improves raw data quality and can handle large sets of images, it does not provide full time series analysis or statistical hypothesis testing. StatFaRmer complements such solutions by offering in-depth statistical analysis of cleaned data, including multifactor group comparisons and identification of temporal dynamics.

Another scalability-oriented approach is the PhytoOracle platform (Gonzalez et al., 2023), a modular processing pipeline for phenotyping data optimized for high-performance computing clusters. PhytoOracle can process multimodal data in parallel (e. g., RGB images, thermal maps, 3D point clouds) and scales efficiently to large data volumes. However, its complexity and requirement for specialized infrastructure make PhytoOracle less accessible to a broad range of biologists. StatFaRmer, by contrast, emphasizes accessibility: it is implemented as a web application with an intuitive interface, simplifying its use on standard desktop systems. At the same time, StatFaRmer supports processing of large numerical datasets and integration of heterogeneous information sources, striking a balance between analytical power and usability.

To quickly assess the program's capabilities, a demonstration dataset based on an experiment on nitrogen stress in cereal crops was created. It is publicly available on GitHub (<https://github.com/Stathmin/StatFaRmer>) and deployed for interactive testing on shinyapps.io (<https://stathmin.shinyapps.io/StatFaRmer>), where one can explore the data and test a working copy of StatFaRmer. In addition, we conducted extensive testing on diverse datasets, including wheat, triticale, sugar beet, maize, sunflower, and soybean. The program has proven to be a reliable and effective tool for analyzing large volumes of phenotypic data across various experimental conditions.

Conclusion

Modern methods of digital plant phenotyping generate colossal amounts of data, including high-frequency time series, structured metadata (cultivar, treatments, replicates, treatment variant), and external parameters integrated from third-party sources. These factors, together with the variety of traits measured by the platform, require specialized solutions for interpretation. Our StatFaRmer tool addresses this challenge through:

- interactive data visualization with support for temporal slices and artifact filtering;
- multifactor group comparisons (by cultivar, treatment, user tags) using various statistical analyses;
- flexible configuration of data slices for independent pattern discovery, accessible to typical users.

Thus, StatFaRmer streamlines the processing of complex phenotypic data, reducing the time needed to uncover patterns of adaptation in soybean and other crops under different treatments, stresses, and other influences.

References

- Abebe A.M., Kim Y., Kim J., Kim S.L., Baek J. Image-based high-throughput phenotyping in horticultural crops. *Plants*. 2023;12(10):2061. doi 10.3390/plants12102061
- Anand K.J., Nagre S.P., Shrivastava M.K., Amrate P.K., Patel T., Katar V.K. Enhancing crop improvement through synergistic integration of advanced plant breeding and proximal remote sensing techniques: a review. *Int J Plant Soil Sci*. 2023;35(19):121-138. doi 10.9734/ijps/2023/v35i193533
- Atefi A., Ge Y., Pitla S., Schnable J. Robotic technologies for high-throughput plant phenotyping: contemporary reviews and future perspectives. *Front Plant Sci*. 2021;12:611940. doi 10.3389/fpls.2021.611940
- Buelvas R.M., Adamchuk V.I., Lan J., Hoyos-Villegas V., Whitmore A., Stromvik M.V. Development of a quick-install rapid phenotyping system. *Sensors (Basel)*. 2023;23(9):4253. doi 10.3390/s23094253
- Coppens F., Wuyts N., Inzé D., Dhondt S. Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr Opin Syst Biol*. 2017;4:58-63. doi 10.1016/j.coisb.2017.07.002
- Danilevich M.F., Bayer P.E., Nestor B.J., Bennamoun M., Edwards D. Resources for image-based high-throughput phenotyping in crops and data sharing challenges. *Plant Physiol*. 2021;187(2):699-715. doi 10.1093/plphys/kiab301
- Fasoula D., Ioannides I., Omirou M. Phenotyping and plant breeding: overcoming the barriers. *Front Plant Sci*. 2020;10:1713. doi 10.3389/fpls.2019.01713
- Gill T., Gill S.K., Saini D., Chopra Y., de Koff J.P., Sandhu K. A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics*. 2022;2(3):156-183. doi 10.1007/s43657-022-00048-z
- Gonzalez E.M., Zarei A., Hendler N., Simmons T., Zarei A., Demieville J., Strand R., ... Swetnam T.L., Merchant N., Michelmore R.W., Lyons E., Pauli D. PhytoOracle: scalable, modular phenomics data processing pipelines. *Front Plant Sci*. 2023;14:1112973. doi 10.3389/fpls.2023.1112973
- Gyan F., Cudjoe D., Sadeghi-Tehran P., Virlet N., Riche A., Castle M., Greche L., Mohareb F., Simms D., Mhada M., Hawkesford M. Machine learning methods for automatic segmentation of images of field- and glasshouse-based plants for high-throughput phenotyping. *Plants*. 2023;12(10):2035. doi 10.3390/plants12102035
- Hartmann A., Czauderna T., Hoffmann R., Stein N., Schreiber F. HTPheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC Bioinformatics*. 2011;12:148. doi 10.1186/1471-2105-12-148
- HortControl – Plant Data Management Software. *PHENOSPEX*. URL: <https://www.phenospec.com/products/plant-phenotyping/science-hortcontrol-data-management-software/> (accessed 8.28.25)
- Li D., Quan C., Song Z., Li X., Yu G., Li C., Muhammad A. High-throughput plant phenotyping platform (HT3P) as a novel tool for estimating agronomic traits from the lab to the field. *Front Bioeng Biotechnol*. 2021;8:623705. doi 10.3389/fbioe.2020.623705
- Li L., Hassan M.A., Song J., Xie Y., Rasheed A., Yang S., Li H., Liu P., Xia X., He Z., Xiao Y. UAV-based RGB imagery and ground measurements for high-throughput phenotyping of senescence and QTL mapping in bread wheat. *Crop Sci*. 2023;63(6):3292-3309. doi 10.1002/csc2.21086
- Lu Y., Wang J., Fu L., Yu L., Liu Q. High-throughput and separating-free phenotyping method for on-panicle rice grains based on deep learning. *Front Plant Sci*. 2023;14:1219584. doi 10.3389/fpls.2023.1219584
- Morota G., Jarquín D., Campbell M.T., Iwata H. Statistical methods for the quantitative genetic analysis of high-throughput phenotyping data. In: Lorence A., Medina Jimenez K. (Eds) High-Throughput Plant Phenotyping. Methods in Molecular Biology. Vol. 2539. Humana, New York, 2019;269-296. doi 10.1007/978-1-0716-2537-8_21
- Najafabadi M.Y., Heidari A., Rajcan I. AllInOne pre-processing: a comprehensive preprocessing framework in plant field phenotyping. *SoftwareX*. 2023;23:101464. doi 10.1016/j.softx.2023.101464
- Ninomiya S. High-throughput field crop phenotyping: current status and challenges. *Breed Sci*. 2022;72(1):3-18. doi 10.1270/jbsbs.21069
- Patel T., Babbar A., Behera K., Katar V.K., Anand K.J., Vyshnavi R.G., Pachori S., Bichewar N. Exploring the potential of proximal remote sensing in plant stress phenotyping: a comprehensive review. *Int J Environ Clim Change*. 2023;13(9):2602-2621. doi 10.9734/ijec/2023/v13i92511

- Rahaman M., Chen D., Gillani Z., Klukas C., Chen M. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Front Plant Sci.* 2015;6:619. doi 10.3389/fpls.2015.00619
- Sumner J. Longitudinal Growth Modeling Options [WWW Document]. 2025. URL: <https://cran.r-project.org/web/packages/pcvr/vignettes/longitudinal.html> (accessed 28.08.25)
- Thrash T., Lee H., Baker R.L. A low-cost high-throughput phenotyping system for automatically quantifying foliar area and greenness. *Appl Plant Sci.* 2022;10(6):e11502. doi 10.1002/aps3.11502
- Ubbens J., Stavness I., Pound M.P., Guo W. Deep learning in plant phenotyping: the first ten years. *Plant Phenomics.* 2025;7(4):100062. doi 10.1016/j.plaphe.2025.100062
- Ulyanov D.S., Ulyanova A.A., Litvinov D., Kocheshkova A., Kroupina A.Yu., Syedina N.M., Voronezhskaya V.S., Vasilyev A.V., Karlov G.I., Divashuk M. StatFaRmer: cultivating insights with an advanced R shiny dashboard for digital phenotyping data analysis. *Front Plant Sci.* 2025;16:1475057. doi 10.3389/fpls.2025.1475057
- Wang Z., Hao J., Shi X., Wang Q., Zhang W., Li F., Mur L.A.J., Han Y., Hou S., Han J., Sun Z. Integrating dynamic high-throughput phenotyping and genetic analysis to monitor growth variation in foxtail millet. *Plant Methods.* 2024;20(1):168. doi 10.1186/s13007-024-01295-z
- Yang W., Feng H., Zhang X., Zhang J., Doonan J., Batchelor W., Xiong L., Yan J. Crop phenomics and high-throughput phenotyping: past decades, current challenges and future perspectives. *Mol Plant.* 2020;13(2):187-214. doi 10.1016/j.molp.2020.01.008
- Yuan H., Song M., Liu Y., Xie Q., Cao W., Zhu Y., Ni J. Field phenotyping monitoring systems for high-throughput: a survey of enabling technologies, equipment, and research challenges. *Agronomy.* 2023;13(11):2832. doi 10.3390/agronomy13112832

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

Received April 22, 2025. Revised October 20, 2025. Accepted November 1, 2025.

doi 10.18699/vjgb-26-37

Description of morphological characteristics of wheat spike as a digital certificate in the SpikeDroidDB database

E.G. Komyshev^{1, 2}, Yu.V. Kruchinina ^{1, 2}, V.S. Koval^{1, 2}, A.A. Poteshkina^{1, 3}, N.V. Petrash ^{1, 3},
V.V. Piskarev ^{1, 3}, N.P. Goncharov¹, D.A. Afonnikov ^{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

³ Siberian Research Institute of Plant Production and Breeding – Branch of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, Russia

 ada@bionet.nsc.ru

Abstract. It has been repeatedly shown that spike productivity is the main component of wheat yield. The main spike parameters related to productivity are size, the number of grains and spikelets per spike, and the presence or absence of awns. In modern genetic research, morphometric analysis of hundreds and thousands of spikes is required to determine the loci that control spike productivity traits. On the other hand, thousands of accessions in modern collections of wheat genetic resources need detailed description. These considerations motivate the development of digital technologies for describing spike traits in wheat, which can be achieved through image analysis methods. These methods allow for automated acquisition of trait values that can serve as the basis for digital plant collections. Here we propose an extended set of spike characteristics obtained both manually and through digital image analysis and present plant characterization. These data form the basis of the updated version of the SpikeDroidDB database (<http://spikedroid.biores.cytogen.ru/>). The digital description of the spike consists of two blocks. The block of uploaded data includes a description of the plant and contains five tables: collection; variety sample (year of cultivation (vegetation), sowing identifier, taxonomic information, etc.), planting site, and characteristics of the spike determined manually (length, width of frontal and lateral views, type and color of the spike, etc.) The block of extracted features includes spike characteristics obtained by digital phenotyping and contains six tables: characteristics of the spike outline in the image; characteristics of the quadrangle model, values of the color components of the spike, dominant colors of the spike, and texture characteristics of the spike in the image. The most illustrative and significant features of the spike have been identified, allowing for the formation of the spike digital certificate, which includes size, shape, and color features derived from the digital images. The features forming the digital certificate have been compared between two wheat species, *T. aethiopicum* and *T. carthlicum*. It is shown that the features of the digital certificate allow for a clear representation of the spike model and the identification of distinct parameters: colors of the spike and awns and roundness of the frontal view of the spike. The database interface has been supplemented with the ability to upload data on plant and spike characteristics, as well as their images, in the batch mode.

Key words: wheat; spike; morphometry; digital phenotyping; database; collection

For citation: Komyshev E.G., Kruchinina Yu.V., Koval V.S., Poteshkina A.A., Petrash N.V., Piskarev V.V., Goncharov N.P., Afonnikov D.A. Description of morphological characteristics of wheat spike as a digital certificate in the SpikeDroidDB database. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov J Genet Breed*. 2026;30(2):330-338. doi 10.18699/vjgb-26-37

Funding. The work was supported by the Russian Science Foundation, project 23-14-00150.

Acknowledgements. The hosting of the database website and image processing were carried out using the resources of the Bioinformatics Shared Access Center with support from State Budgeted Project FWNR-2026-0023.

Описание морфологических характеристик колосьев пшеницы в базе данных SpikeDroidDB в виде цифрового паспорта

Е.Г. Комышев^{1, 2}, Ю.В. Кручинина ^{1, 2}, В.С. Коваль^{1, 2}, А.А. Потешкина^{1, 3}, Н.В. Петраш ^{1, 3},
В.В. Пискарев ^{1, 3}, Н.П. Гончаров¹, Д.А. Афонников ^{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

³ Сибирский научно-исследовательский институт растениеводства и селекции – филиал Федерального исследовательского центра Институт цитологии и генетики Сибирского отделения Российской академии наук, р. п. Краснообск, Новосибирская область, Россия

 ada@bionet.nsc.ru

Аннотация. При анализе структуры урожая пшеницы неоднократно показано, что его выраженность зависит от продуктивности колоса. Основные характеристики колоса, которые связаны с продуктивностью, это прежде всего масса 1000 зерен, число зерен и колосков в колосе, его размеры, наличие/отсутствие остей и другие. В современных генетических исследованиях для идентификации локусов, контролирующих признаки продуктивности колоса, требуется морфометрия сотен и тысяч колосьев. С другой стороны, современные коллекции генетических ресурсов содержат тысячи образцов, которые также требуют своего детального описания. Все это обуславливает необходимость развития цифровых технологий описания признака колоса пшеницы, которые могут быть достигнуты на основе методов анализа изображений. Эти методы позволяют автоматически получать значения набора признаков, которые могут служить основой формирования цифровых коллекций растений. В настоящей работе предложено цифровое описание колоса пшеницы на основе характеристик, полученных как вручную, так и на основе анализа цифровых изображений, а также характеристик растений, у которых был взят колос. Эти данные положены в основу обновленной версии базы данных SpikeDroidDB (<http://spikedroid.biores.cytogen.ru/>). Цифровое описание колоса состоит из двух блоков. Блок загружаемых данных содержит описание растения и включает 5 таблиц: коллекция, сортообразец (год выращивания (вегетация), посевной номер, таксономическую информацию и др.), место выращивания, характеристики колоса растений, оцененные экспертом вручную (длину, ширину фронтальной и боковой проекций, тип и цвет колоса и др.). Блок извлекаемых характеристик включает признаки колоса, полученные в результате цифрового фенотипирования, и состоит из шести таблиц: характеристики контура колоса на изображении; характеристики модели четырехугольников, компоненты цвета колоса, доминантные цвета колоса, текстурные характеристики колоса на изображении. Было проведено выделение наиболее наглядных и информативных характеристик колоса, которые позволили сформировать цифровой паспорт колоса, включающий признаки размера, формы и цвета, определенные на основе анализа цифровых изображений. Проведено сравнение признаков, формирующих цифровой паспорт, у двух видов пшеницы, *T. aethiopicum* и *T. carthlicum*. Показано, что признаки цифрового паспорта позволяют наглядно представить модель колоса, а также выявить достоверно различающиеся параметры (цвет колоса и остей, округлость фронтальной проекции колоса). В интерфейс базы данных также добавлена возможность пакетной загрузки данных о характеристиках растения и колоса, а также их изображений.

Ключевые слова: пшеница; колос; морфометрия; цифровое фенотипирование; база данных; коллекция

Introduction

It has been repeatedly shown that spike productivity is the main component of wheat yield (Moiseeva, 2017; Romanov, Pimonov, 2018; Demina, 2022; Shuklina et al., 2022). The most significant characteristics for yield are spike size, the number of spikelets and grains, and degree of compactness. For wheat processing, it is also important to consider the presence or absence of awns, spike color, and so on. The investigation of genes that control these traits is of particular interest to geneticists and breeders (Guo et al., 2018; Garland-Campbell, 2022). In such studies, it is crucial to evaluate the diversity of spike traits across a large number of accessions, which allows for the identification of associations between varieties in spike characteristics and nucleotide substitutions in the wheat genome (Wu et al., 2012; Xu et al., 2024).

The study of wheat and its relatives also relies on the identification of various spike traits, particularly those that allow for plant taxonomy, biological material classification (Spagnoletti Zeuli, Qualset, 1987; Khanjari et al., 2008; Goncharov, 2009), and the evaluation of breeding trials (Methodology..., 2019). Such traits include spike shape (Konopatskaia et al., 2016), pubescence (Goncharov et al., 2007), the presence or absence of awns (Smolenskaya et al., 2022), spike and awn color (Lya-punova, 2017; Smolenskaya et al., 2022), and compactness (Vavilova et al., 2017). These morphological traits are used by geneticists and breeders in analyzing wheat collections and breeding programs, but their assessment is labor-consuming due to the large volume of material analyzed, often involving hundreds or thousands of specimens (Pakul, Sherina, 2009; Piskarev et al., 2018; Zuev et al., 2019).

To enhance the efficiency of trait determination in modern genetics and breeding, phenotyping technologies based on digital image analysis are employed. These technologies automate the process of phenotypic characteristic assessment, thereby reducing labor costs, enabling evaluations of thousands of samples (Afonnikov et al., 2016; Meraj et al., 2024), and permitting the application of machine learning methods (Murphy et al., 2024). For wheat spikes, image-based phenotyping methods have also been intensely developed recently. They primarily allow for the determination of morphological characteristics of the spike, such as length, width, and density (Genaev et al., 2019). X-ray imaging enables detailed three-dimensional descriptions of spike structures down to individual grains (Ling et al., 2023). Machine learning methods can automatically determine the number (Genaev et al., 2026) and shape (Niu et al., 2024) of spikelets, as well as assess the presence or absence of pubescence (glume hairs) (Artemenko et al., 2024).

Digital phenotyping allows for the determination of a significantly larger number of plant characteristics, much more than a breeder could previously examine (Afonnikov et al., 2016). Some of these characteristics, such as spike length or the number of grains per spike, are directly related to yield (Moiseeva, 2017; Romanov and Pimonov, 2018; Demina, 2022; Shuklina et al., 2022). Others are indirectly related to productivity, e. g., grain weight, which can be estimated from the area of its view on an image. Some digital characteristics are difficult to interpret agrobiologically, but they can be effectively used for spike classification employing machine learning methods (Bi et al., 2010). The choice of digital

traits is crucial for creating digital collections—databases that describe plant phenotypes (Conejo-Rodríguez et al., 2024). On the one hand, these traits should be familiar to breeders working with the databases (Methodology..., 2019), whereas on the other, they should be sufficiently representative to solve classification tasks.

Previously, we developed the SpikeDroidDB database, which stores information about wheat spike samples, including expert-assessed trait descriptions (Genaev et al., 2018). Here we provide an expansion and systematization of the set of digital features for describing spikes in this database, such as quadrangle model parameters (Genaev et al., 2019) and spike and awn color characteristics. The database interface has been supplemented with the ability to upload information, including expert-assessed sample characteristics and spike images, in the batch mode. As a result, the number of spike sample descriptions in the database exceeded 1700.

Materials and methods

Plant material. The database was supplemented with descriptions of spike samples from the Siberian Gene Pool bread wheat variety collection of the Siberian Research Institute of Plant Production and Breeding and the Institute of Cytology and Genetics, SB RAS (Piskarev et al., 2018). Plants were grown in 2019–2020 in experimental fields of the Siberian Research Institute of Plant Production and Breeding – Branch of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (Krasnoobsk, Novosibirsk region).

Digital phenotyping. To obtain spike images, we used the ‘on a clip’ protocol described previously (Genaev et al., 2018, 2019). The spike was placed vertically on a clip against a back-drop of blue paper. Images were taken using a Canon 350D digital camera with an EF-S 18-55mm f/3.5-5.6 lens. X-Rite Mini ColorChecker Classic color chart was positioned within the frame for scale and color calibration (Fig. 1a). The spikes were photographed in four views: frontal (the widest view), back, and two side views. To obtain digital characteristics of the spike, the images were segmented into background, color chart, spike body, and awns using deep machine learning methods (Artemenko et al., 2024). The segmented image was then processed with the WERecognizer program (Genaev et al., 2019) to estimate morphometric and color characteristics of the spike. The morphological characteristics included a set of features for a quadrangle model symmetrized relative to the spike main axis (Komyshev et al., 2024), Figure 1b, c; a set of outline features of the spike body in the image (length, perimeter, area, circularity, roundness, solidity, and rugosity); and the area of the awns. In total, 19 morphological features of the spike and awns were analyzed.

For the spike and its awns, digital color characteristics were determined. The method is based on a previously applied approach for analyzing wheat grains (Afonnikov et al., 2022). For all pixels of the spike body, the average values of color components were calculated in four color spaces: RGB, Lab, HSV, and YCrCb (Komyshev et al., 2020), totaling 12 characteristics. Additionally, three dominant colors



Fig. 1. Morphometric characteristics of the spike. (a) Image of the spike placed vertically on a clip, frontal view. (b) Results of image segmentation showing the spike body (light green), awns (yellow), and background (blue). The spike axis is indicated by a pink line. (c) Quadrangle model for the spike. Feature designations in (b) and (c) follow Table S1.

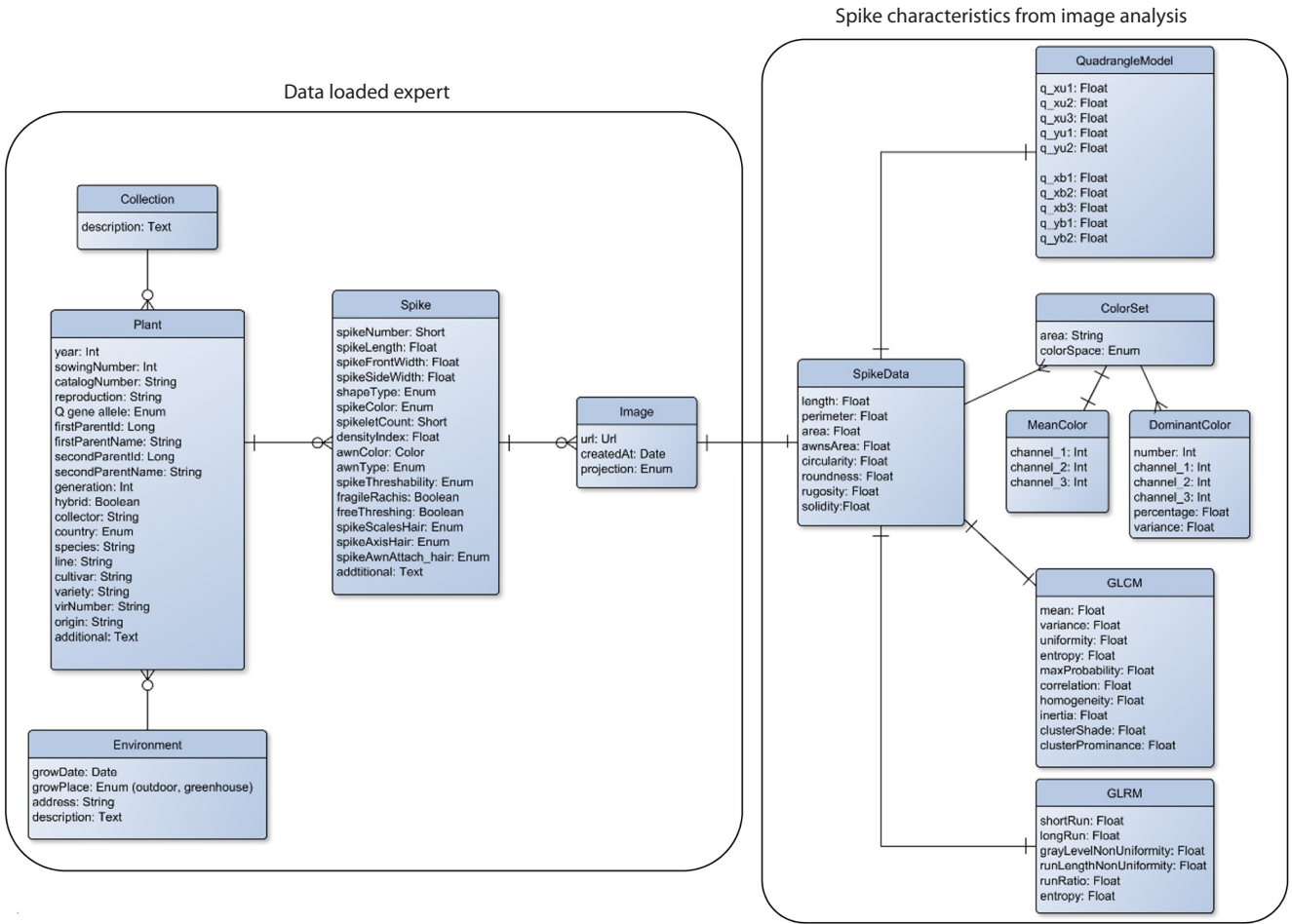


Fig. 2. Logical structure of the updated version of the SpikeDroidDB database. On the right is the user-input data block. On the left is the data block obtained through digital phenotyping of images.

and their corresponding component values were identified. Dominant colors are defined as centroids of clusters in the RGB component space obtained by the *k*-means method, assuming *k* = 3 (Afonnikov et al., 2022). These colors characterize the three largest groups of pixels similar in color in the spike image and satisfactorily describe its color heterogeneity. Typically, the first cluster (dominant color) has component values close to the average values across the entire spike image. Color characteristics for awn pixels were determined in a similar way.

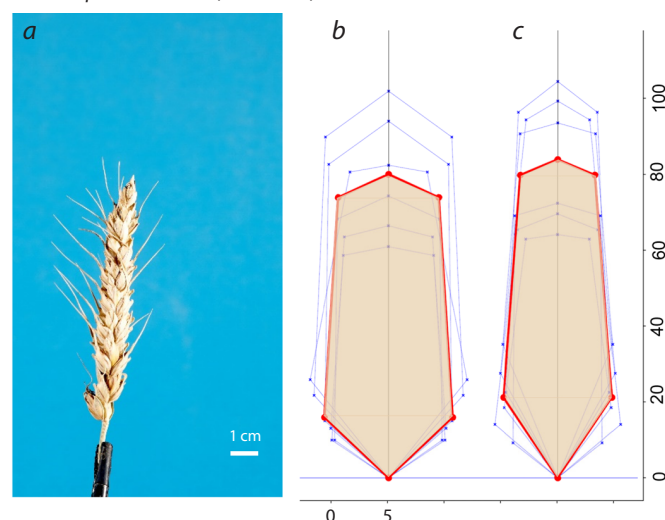
Additionally, texture features of the spike area in the image were calculated. These features characterize the structural heterogeneity of the object in the image, which may arise due to uneven coloring or the surface structure of the glumes (e.g., pubescence). A distinguishing feature of simple texture is regularity, with repetitive or partially reproducible elements on a certain surface or object. Two types of structural features were calculated: those based on the Grey Level Co-occurrence Matrix (GLCM, 10 characteristics) and the Gray Level Run-length Matrix (GLRM, 6 characteristics) (Komyshev et al., 2020).

SpikeDroid database structure. The aforementioned spike characteristics obtained through the analysis of digital images,

together with the characteristics determined by expert evidence and plant parameters, are stored in the updated version of the SpikeDroidDB database, whose logical structure is presented in Figure 2.

The structure of the updated version of the SpikeDroidDB database includes two data blocks. The block of data uploaded by experts contains five tables and describes the collection and the plant whose spike is analyzed. The Collection table contains a text field for the name of the collection and a description field. The Plant table includes 20 fields: year of cultivation (vegetation), accession number in the collection, sowing identifier, reproduction number, *Q*-gene allele (Vavilova et al., 2020), information on parent plants, taxonomic information, and origin of the variety/line. The Environment table contains information about the time and location of cultivation, including the address of the originating institution. The Spike table contains characteristics determined manually: spike number on the plant, spike length, frontal and lateral view widths, spike type, color, number of grains, awn color and type, pubescence characteristics of the glume, spike threshability, rachis fragility, and awn attachment type. Additionally, this block contains information about the spike image (Image table), in-

T. aethiopicum Jakubcz. (k-19301/2)



T. carthlicum Nevski (k-32496)

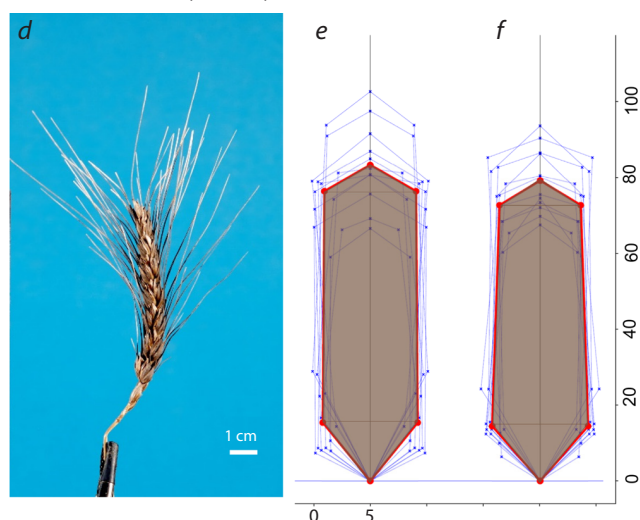


Fig. 3. Graphical representation of the spike digital certificate parameters for two wheat species: *T. aethiopicum* Jakubcz. (a–c) and *T. carthlicum* Nevski (d–f). Digital images of the spike are shown in panels a and d; models of the spike for the frontal view are shown in panels b and e; models of the spike for side views are shown in panels c and f. Thin blue lines represent the geometric models of the spike for individual plants, while thick red lines show polygons representing models for the mean values of spike parameters. The fill color of the red polygons corresponds to the averaged color of the spike outline pixels. The length scales in mm along the X- and Y-axes for the spike models differ by a factor of 5 for better visualization of the spike shape characteristics.

cluding the view number, date of acquisition, and a link to the image file.

The block of extracted features includes spike characteristics obtained by digital phenotyping (see above). It includes seven tables. The SpikeData table includes characteristics of the spike outline in the image. The QuadrangleModel table presents characteristics of the quadrangle model. The Mean-Color table stores mean color characteristics of the spike. The DominantColor table contains dominant color characteristics of the spike. The GLCMTexture table contains texture parameters for GLCM. The GLRMTexture table contains texture parameters for GLRM.

Digital spike certificate. Among the numerous shape and color characteristics of a spike presented in the SpikeDroidDB database (see above), several key features should be highlighted that can be used for a reliable assessment of similarities/differences in spike shape, size, and color. These characteristics can also be used for visually demonstrating the similarities or dissimilarities between spikes of different accessions from the plant collection. Previously, we demonstrated that simplified geometric model features of the spike based on quadrangles and outline characteristics of the spike from images can be used to identify both interspecies and intraspecies differences in spikes (Komyshev et al., 2024). These features allow for high-precision species identification of plants. We obtained grounds for forming the digital spike certificate by choosing those features of shape and size obtained from the analysis of digital images that were most visually informative and sufficient for classification tasks. These features were supplemented with color characteristics of the spike and awns. The list of

features forming the digital certificate is provided in Table S1 of the Supplementary File¹. For each spike image, it includes: 11 parameters of the quadrangle model (Komyshev et al., 2024), 7 outline features of the spike (Genaev et al., 2019), 3 color parameters of the spike (average R, G, B components of the spike body pixels), and 3 color parameters of the awns (average R, G, B components of the awn pixels).

For most spikes in the database, at least four views are stored. Therefore, the digital certificate of each spike includes averaged characteristics for the frontal/back and two side views. In cases where only one view is available for a spike image, usually the frontal view following the ‘on the table’ protocol (Genaev et al., 2018), the certificate contains only the features derived from that single image.

Database implementation. The database is implemented using the Drupal content management system (<https://www.drupal.org>). Data storage is provided by a relational database managed by the MySQL database system, which is deployed on the server of the “Bioinformatics” shared access computational center running CentOS Linux.

Results

Two examples of spike visualization based on the characteristics presented in their digital certificates are shown in Figure 3. They represent two wheat species from N.P. Goncharov’s collection: *Triticum aethiopicum* Jakubcz. (VIR accession number k-19301/2, represented by 6 plants) and *T. carthlicum* Nevski (VIR accession number k-32496, 12 plants).

¹ Tables S1 and S2 are available at:

<https://vavilovj-icg.ru/download/pict-2026-30/appx19.pdf>

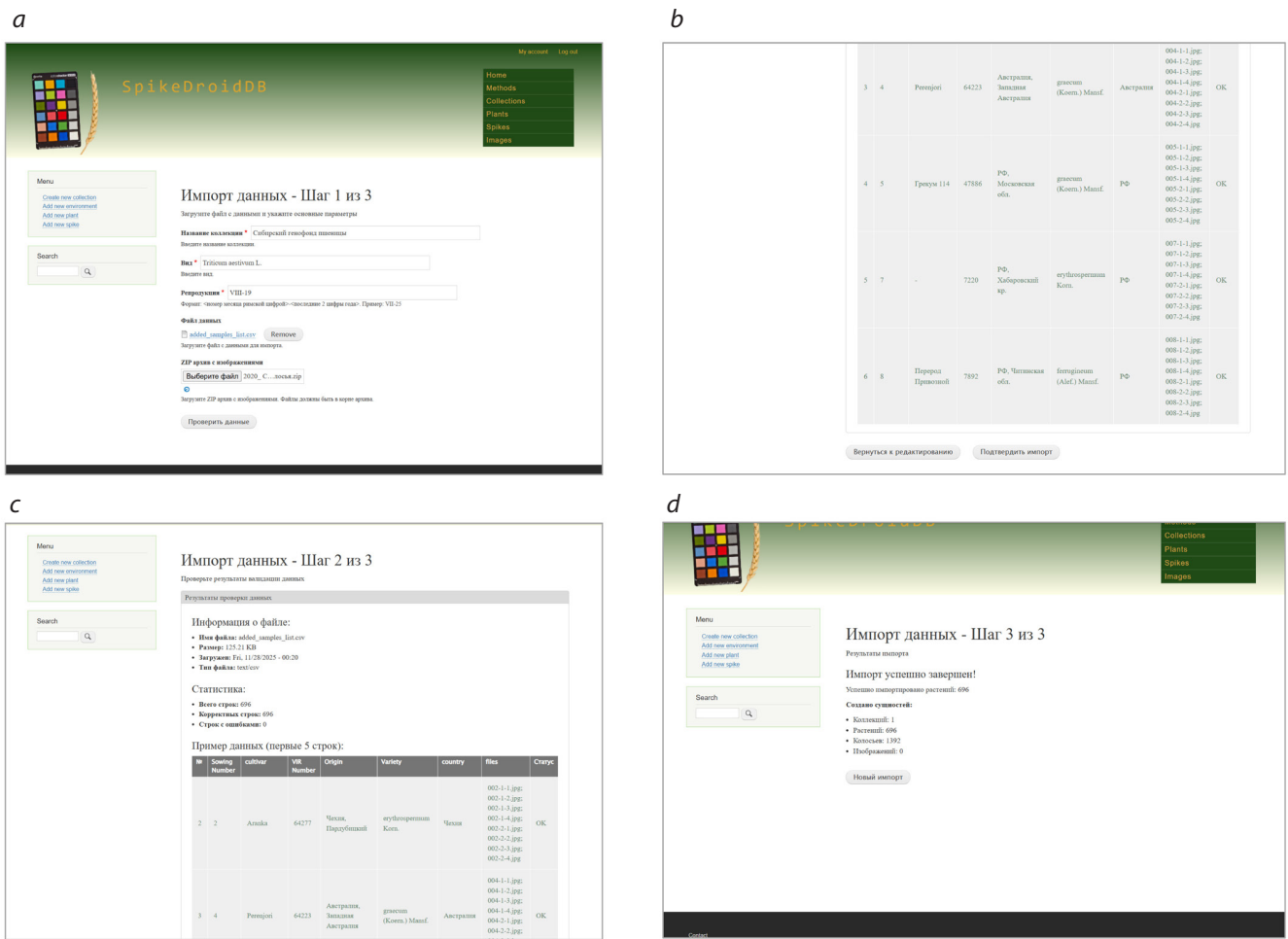


Fig. 4. Visualization of the user interface during the main steps of batch uploading data about spikes and their images into the SpikeDroidDB database. (a) Input of primary dataset characteristics (wheat species, reproduction number, field for uploading the collection description file). (b) Visualization of the list of uploaded data. (c) Visualization of information about the uploaded data. (d) Completion of data import.

The figure demonstrates the similarity in spike length between the two wheat species. However, their shapes differ slightly. The frontal view of *T. aethiopicum* appears more rounded, whereas that of *T. carthlicum* is more elongated. Differences in the size and number of awns are clearly visible in the images: *T. aethiopicum* has fewer and shorter awns compared to *T. carthlicum*. Additionally, the colors of the spikes and awns differ: They are lighter in *T. aethiopicum* and darker in *T. carthlicum*.

We conducted a statistical test to confirm the differences between the mean values of the digital certificate parameters for the two species. Significant differences were found in five shape features (all related to outline characteristics but not the model parameters) and all color characteristics (both for the spike and awns) in both the frontal and side views. The results are provided in the supplementary material (Table S2). They are in good agreement with the visual assessment (Figure 3). For the frontal view, significant differences were observed in the following indices: area of awns (c_{Sa}), circularity (c_{Ci}),

roundness (c_{Ro}), and integrity (c_{So}). For the side view, significant differences were observed only in the area of awns. Significant differences were noted in all color components of the spike and awns in both the frontal and side views.

Thus, the features forming the digital certificate of the spike allow for a clear representation of similarities and differences in spike characteristics between the two wheat species, as well as the recognition of statistically significant characteristics.

The possibility to upload data for multiple spikes at once simplifies the user's interaction with the database. The user interface for importing data in this case is shown in Figure 4. It demonstrates the step-by-step implementation of uploading multiple images and descriptions of spikes.

As an example, Figure 5 provides the statistics of the uploaded data for samples from the 'Siberian Wheat Collection' grown in 2019. A total of 696 samples were grown and analyzed. Their distributions by countries of origin, i.e., the country from which each specific variety sample originated

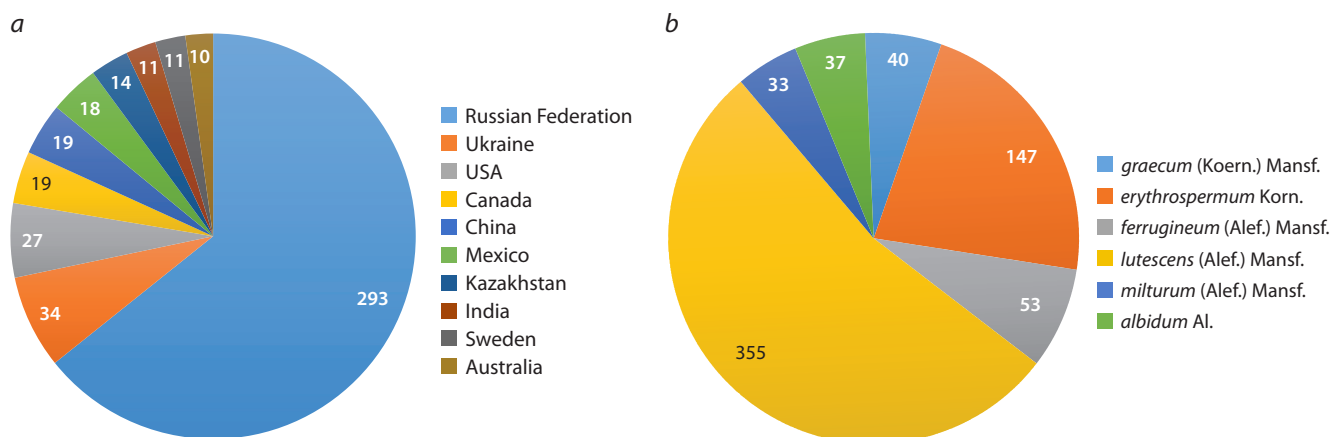


Fig. 5. Statistics of bread wheat samples (reproduction 2019) in the SpikeDroidDB database. Distribution of samples: (a) by countries of origin, (b) by varieties.

(the top 10 most represented in the analyzed collection), and by varieties (most represented in the analyzed material) are shown.

Discussion

The development of modern phenotyping methods allows for the digitization of spike characteristics for subsequent computer analysis. Previously we demonstrated that digital characteristics of spike shape and size correlate with equivalent biological characteristics measured manually (Komyshev et al., 2024). The same work showed that the combined use of digital and classical characteristics allows for high-precision classification of plants into species. Furthermore, such representation enables the identification of traits that significantly vary among specimens of the same species (Komyshev et al., 2024). All of this underscores the usefulness of both classical and digital spike characteristics for geneticists and breeders in describing spike traits. We applied these considerations to the development of the digital spike certificate, which includes descriptions of plant characteristics, their growing site locations, spike traits assessed by experts, and digital characteristics derived from image analysis. This comprehensive representation allows for a closer description of spike properties. Digital characteristics of the spike include quadrangle model parameters, outline properties of the spike in images, and a set of color and texture parameters. Thus, digital phenotype parameters, in conjunction with classical trait sets, make plant descriptions more detailed and help in classification (Conejo-Rodríguez et al., 2024).

The digital representation of phenotypic characteristics of plants is a crucial step in the development of modern plant genetic resource collections (Thormann et al., 2012; Heberling, 2022). On the one hand, it enables rapid assessment of the vast diversity of phenotypic traits in collections. On the other hand, digital data from collections can be more effectively utilized in breeding programs (Dipta et al., 2023). This highlights the advantage of describing the properties of germplasm banks and breeding collections in a digital form for the purpose of genetic research and breeding. Further development of

phenotyping methods and the selection of the most important digital characteristics will enable a deeper understanding of the manifestation of phenotypic traits, thereby potentially enhancing the efficiency of identifying genes that control them.

Conclusions

A comprehensive system for the digital description of wheat spike was developed. It integrated both traditional manual evaluation methods and modern digital phenotyping technologies. As a result, the structure of the SpikeDroidDB database was updated for the comprehensive representation of spike properties. The updated structure includes two main data blocks. The block of uploaded data contains fundamental information about the collection and specific variety sample characteristics, presented in five interrelated tables. Emphasis was placed on agronomic and taxonomic information, growing conditions, and basic spike characteristics determined manually by experts. The block of extracted features comprises an extended set of parameters obtained through digital image analysis. It includes six categories of data: outline characteristics of the spike, quadrangle model parameters, color characteristics (including dominant colors), and texture features. An important functional enhancement was the implementation of batch data upload capability, significantly simplifying the process of entering information about plants and uploading their images, as well as their subsequent statistical analysis.

The most visually informative and relevant characteristics of a spike were identified, allowing the formation of a digital certificate of the spike, which includes size, shape, and color features determined based on digital image analysis. We compared the features forming the digital certificate between two wheat species, *T. aestivum* and *T. carthlicum*, and demonstrated that the digital certificate features allow for a clear representation of the spike model and the identification of significantly differing parameters (color of the spike and awns and roundness of the frontal view of the spike). Thus, the efficiency of describing spike characteristics for research in breeding and genetics is enhanced, enabling a comprehensive analysis of plant morphological features.

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
- Afonnikov D.A., Komyshev E.G., Efimov V.M., Genaev M.A., Koval V.S., Gierke P.U., Börner A. Relationship between the characteristics of bread wheat grains, storage time and germination. *Plants.* 2022;11(1):35. doi 10.3390/plants11010035
- Al Khanjari S., Filatenko A.A., Hammer K., Buerkert A. Morphological spike diversity of Omani wheat. *Genet Resour Crop Evol.* 2008; 55(8):1185-1195. doi 10.1007/s10722-008-9319-9
- Artemenko N.V., Genaev M.A., Epifanov R.U., Komyshev E.G., Kruchinina Y.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
- Bi K., Jiang P., Lei L., Shi B., Wang C. Non-destructive measurement of wheat spike characteristics based on morphological image processing. *Trans Chin Soc Agric Eng.* 2010;26(12):212-216. doi 10.3969/j.issn.1002-6819.2010.12.036
- Conejo-Rodriguez D.F., Gonzalez-Guzman J.J., Ramirez-Gil J.G., Wenzl P., Urban M.O. Digital descriptors sharpen classical descriptors, for improving genebank accession management: A case study on *Arachis* spp. and *Phaseolus* spp. *PLoS One.* 2024;19(5):e0302158. doi 10.1371/journal.pone.0302158
- Demina I.F. Adaptability of spring wheat varieties and lines in terms of productivity and elements of its structure under forest-steppe conditions of the Middle Volga region. *Mezhdunarodnyy Sel'skhozaystvennyy Zhurnal = International Agricultural Journal.* 2022;65(5):530-534. doi 10.55186/25876740_2022_65_5_530 (in Russian)
- Dipta B., Sood S., Devi R., Bhardwaj V., Mangal V., Thakur A.K., Kumar V., Pandey N.K., Rathore A., Singh A.K. Digitalization of potato breeding program: Improving data collection and management. *Heliyon.* 2023;20(9):e12974. doi 10.1016/j.heliyon.2023.e12974
- Garland-Campbell K.A. Club Wheat – A review of club wheat history, improvement, and spike characteristics in wheat. *Plant Breed Rev.* 2022;46:421-465. doi 10.1002/9781119874157.ch7
- Genaev M.A., Komyshev E.G., Fu Hao, Koval V.S., Goncharov N.P., Afonnikov D.A. SpikeDroidDB: an information system for annotation of morphometric characteristics of wheat spike. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2018;22(1): 132-140. doi 10.18699/VJ18.340 (in Russian)
- Genaev M.A., Komyshev E.G., Kruchinina Y.V., Goncharov N.P., Afonnikov D.A., Smirnov N.V. Morphometry of the wheat spike by analyzing 2D images. *Agronomy.* 2019;9(7):390. doi 10.3390/agronomy9070390
- 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 Selekcii = Vavilov J Genet Breed.* 2026;30(1):27-35. doi 10.18699/vjgb-26-09
- Goncharov N.P. Manual book of common and durum wheat varieties. Novosibirsk: SB RAS Publ. House, 2009 (in Russian)
- Goncharov N.P., Bannikova S.V., Kawahara T. Wheat artificial amphiploids involving *Triticum timopheevii* genome: their studies, preservation and reproduction. *Genet Resour Crop Evol.* 2007;54(7): 1507-1516. doi 10.1007/s10722-006-9141-1
- Guo Z., Zhao Y., Röder M.S., Reif J.C., Ganai M.W., Chen D., Schnurbusch T. Manipulation and prediction of spike morphology traits for the improvement of grain yield in wheat. *Sci Rep.* 2018;8(1):14435. doi 10.1038/s41598-018-31977-3
- Heberling J.M. Herbaria as big data sources of plant traits. *Int J Plant Sci.* 2022;183(2):87-118. doi 10.1086/717623
- Komyshev E.G., Genaev M.A., Afonnikov D.A. Analysis of color and texture characteristics of cereals on digital images. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2020;24(4):340-347. doi 10.18699/VJ20.626
- Komyshev E.G., Genaev M.A., Kruchinina Yu.V., Koval V.S., Goncharov N.P., Afonnikov D.A. Evaluation of the spike diversity of seven hexaploid wheat species and an artificial amphidiploid using a quadrangle model obtained from 2D images. *Plants.* 2024;13(19):2736. doi 10.3390/plants13192736
- 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
- Ling Y., Zhao Q., Liu W., Wei K., Bao R., Song W., Nie X. Detection and characterization of spike architecture based on deep learning and X-ray computed tomography in barley. *Plant Methods.* 2023;19(1): 115. doi 10.1186/s13007-023-01096-w
- Lyapunova O.A. Intraspecific classification of durum wheat: new botanical varieties and forms. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2017;21(2):152-157. doi 10.18699/VJ17.233 (in Russian)
- Meraj T., Sharif M.I., Raza M., Alabrah A., Kadry S., Gandomi A.H. Computer vision-based plants phenotyping: A comprehensive survey. *iScience.* 2024;27(1):108709. doi 10.1016/j.isci.2023.108709
- Methodology of State Variety Testing of Agricultural Crops. Issue 1. General Part. Moscow: 2019. (Ministry of Agriculture of the Russian Federation. Federal State Budgetary Institution «Gossortkomissiya»). URL https://gossortrf.ru/upload/2019/08/metodica_1.pdf (Accepted 04.12.2025) (in Russian)
- Moiseeva K.V. The productivity of winter wheat varieties. *Agrar Bull Urals.* 2017;(9):30-34 (in Russian)
- Murphy K.M., Ludwig E., Gutierrez J., Gehan M.A. Deep learning in image-based plant phenotyping. *Annu Rev Plant Biol.* 2024;75(1): 771-795. doi 10.1146/annurev-arplant-070523-042828
- Niu Z., Liang N., He Y., Xu C., Sun S., Zhou Z., Qiu Z. A novel method for wheat spike phenotyping based on instance segmentation and classification. *Appl Sci.* 2024;14:6031. doi 10.3390/app14146031
- Pakul' V.P., Sherina A.N. Comparative study of spring wheat samples from the VIR collection according to the complex of agrobiological properties and characteristics. *Dostizheniya Nauki i Tekhniki APK = Achievements of Science and Technology of AIC.* 2009;9:7-8 (in Russian)
- Piskarev V.V., Zuev E.V., Brykova A.N. Sources for the breeding of soft spring wheat in the conditions of Novosibirsk region. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2018; 22(7):784-794. doi 10.18699/VJ18.422 (in Russian)
- Romanov B.V., Pimanov K.I. Phenogenomics of production traits of wheat varieties. Persianovsky, 2018 (in Russian)
- Shchuklina O.A., Zavgorodniy S.V., Alenicheva A.D., Ivanova L.P., Kvitko V.E., Pylnev V.V., Upelnik V.P. Relationship between spike structure elements and plant productivity of $\times$ *Triticotrigia cziczinii* Tzvel samples. *Izvestiya of Timiryazev Agricultural Academy.* 2022; (5):57-69. doi 10.26897/0021-342X-2022-5-57-69 (in Russian)
- Smolenskaya S.E., Efimov V.M., Kruchinina Yu.V., Nemtsev B.F., Chepurinov G.Yu., Ovchinnikova E.S., Belan I.A., Zuev E.V., Zhou Chenxi, Piskarev V.V., Goncharov N.P. Earliness and morphotypes of common wheat cultivars of Western and Eastern Siberia. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov J Genet Breed.* 2022;26(7):662-674. doi 10.18699/VJGB-22-81
- Spagnoletti Zeuli P.S., Qualset C.O. Geographical diversity for quantitative spike characters in a world collection of durum wheat. *Crop Sci.* 1987;27(2):235-241. doi 10.2135/cropsci1987.0011183X002700020022x

- Thormann I., Gaisberger H., Mattei F., Snook L., Arnaud E. Digitization and online availability of original collecting mission data to improve data quality and enhance the conservation and use of plant genetic resources. *Genet Resour Crop Evol.* 2012;59(5):635-644. doi 10.1007/s10722-012-9804-z
- 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., Konopatskaia I., Blinov A., Kondratenko E.Ya., Kruchinina Y.V., Goncharov N.P. Genetic variability of spelt factor gene in *Triticum* and *Aegilops* species. *BMC Plant Biol.* 2020;20(Suppl. 1): 310. doi 10.1186/s12870-020-02536-8
- Wu X., Chang X., Jing R. Genetic insight into yield-associated traits of wheat grown in multiple rain-fed environments. *PLoS One.* 2012; 7(2):e31249. doi 10.1371/journal.pone.0031249
- Xu H., Wang Z., Wang F., Hu X., Ma C., Jiang H., Xie C., Gao Y., Ding G., Zhao C., Qin R., Cui D., Sun H., Cui F., Wu Y. Genome-wide association study and genomic selection of spike-related traits in bread wheat. *Theor Appl Genet.* 2024;137(6):131. doi 10.1007/s00122-024-04640-x
- Zuev E.V., Brykova A.N., Kudryavtseva E.Yu. Results of analyzing the passport database 'Spring bread wheat landraces in the VIR collection'. *Proc Appl Bot Genet Breed.* 2019;180(1):7-11. doi 10.30901/2227-8834-2019-1-7-11 (in Russian)

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

Received November 30, 2025. Revised January 15, 2026. Accepted January 15, 2026.

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 СМЭ 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, Белый список (уровень 1), Russian Science Citation Index, Российский индекс научного цитирования, ВИНТИ, 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 30.03.2026. Формат 60 × 84 1/8. Уч.-изд. л. 24.2.
.....