

Научный рецензируемый журнал

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.

Периодичность 8 выпусков в год

DOI 10.18699/VJ21.001

Учредители

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

Межрегиональная общественная организация Вавиловское общество генетиков и селекционеров

Сибирское отделение Российской академии наук

Главный редактор

В.К. Шумный – академик РАН, д-р биол. наук, профессор (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)

Ответственный секретарь

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

Редакционный совет

*Л.И. Афтана*с – академик РАН, д-р мед. наук (Россия)
В.С. Баранов – чл.-кор. РАН, д-р мед. наук (Россия)
Л.А. Беспалова – академик РАН, д-р с.-х. наук (Россия)
А. Бёрнер – д-р наук (Германия)
М.И. Воевода – академик РАН, д-р мед. наук (Россия)
И. Гроссе – д-р наук, проф. (Германия)
Г.Л. Дианов – д-р биол. наук, проф. (Великобритания)
Ю.Е. Дуброва – д-р биол. наук, проф. (Великобритания)
Н.Н. Дыгало – чл.-кор. РАН, 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)
И.Е. Керкис – д-р наук (Бразилия)
А.В. Кильчевский – чл.-кор. НАНБ, д-р биол. наук (Беларусь)
С.В. Костров – чл.-кор. РАН, д-р хим. наук (Россия)
А.В. Кочетов – чл.-кор. РАН, д-р биол. наук (Россия)
Ж. Ле Гуи – д-р наук (Франция)
Б. Люгтенберг – д-р наук, проф. (Нидерланды)
В.И. Молодин – академик РАН, д-р ист. наук (Россия)
В.П. Пузырев – академик РАН, д-р мед. наук (Россия)
А.Ю. Ржецкий – канд. биол. наук, проф. (США)
И.Б. Rogozin – канд. биол. наук (США)
А.О. Рувинский – д-р биол. наук, проф. (Австралия)
Е.А. Салина – д-р биол. наук, проф. (Россия)
К.В. Славин – д-р наук, проф. (США)
В.А. Степанов – чл.-кор. РАН, д-р биол. наук (Россия)
И.А. Тихонович – академик РАН, д-р биол. наук (Россия)
Е.К. Хлесткина – д-р биол. наук, профессор (Россия)
Л.В. Хотылева – академик НАНБ, д-р биол. наук (Беларусь)
Э.К. Хуснутдинова – д-р биол. наук, проф. (Россия)
М.Ф. Чернов – д-р мед. наук (Япония)
С.В. Шестаков – академик РАН, д-р биол. наук (Россия)
Н.К. Янковский – академик РАН, д-р биол. наук (Россия)

Редакционная коллегия

Т.Г. Амстиславская – д-р биол. наук (Россия)
Е.Е. Андронов – канд. биол. наук (Россия)
Ю.С. Аульченко – д-р биол. наук (Россия)
Д.А. Афонников – канд. биол. наук, доцент (Россия)
Е.В. Березиков – канд. биол. наук, проф. (Нидерланды)
Н.П. Бондарь – канд. биол. наук (Россия)
С.А. Боринская – д-р биол. наук (Россия)
П.М. Бородин – д-р биол. наук, проф. (Россия)
Т.А. Гавриленко – д-р биол. наук (Россия)
В.Н. Даниленко – д-р биол. наук, проф. (Россия)
С.А. Демаков – д-р биол. наук (Россия)
Е.А. Долгих – д-р биол. наук (Россия)
Ю.М. Константинов – д-р биол. наук, проф. (Россия)
О. Кребс – д-р биол. наук, проф. (Германия)
И.Н. Лаврик – канд. хим. наук (Германия)
Д. Ларкин – д-р биол. наук (Великобритания)
И.Н. Лебедев – д-р биол. наук, проф. (Россия)
Л.А. Лутова – д-р биол. наук, проф. (Россия)
В.Ю. Макеев – чл.-кор. РАН, д-р физ.-мат. наук (Россия)
М.П. Мошкин – д-р биол. наук, проф. (Россия)
Л.Ю. Новикова – канд. техн. наук (Россия)
Е. Песцова – д-р биол. наук (Германия)
Н.А. Проворов – д-р биол. наук, проф. (Россия)
Д.В. Пышный – чл.-кор. РАН, д-р хим. наук (Россия)
А.В. Ратушный – канд. биол. наук (США)
М.Г. Самсонова – д-р биол. наук (Россия)
Е. Турусбеков – канд. биол. наук (Казахстан)
М. Чен – д-р биол. наук (Китайская Народная Республика)
Ю. Шавруков – д-р биол. наук (Австралия)

Scientific Peer Reviewed Journal

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

*Founded in 1997**Published 8 times annually*

DOI 10.18699/VJ21.001

Founders

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

Editor-in-Chief

V.K. Shumny, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), 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

Executive Secretary

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

Editorial council

L.I. Aftanas, Full Member of the RAS, Dr. Sci. (Medicine), Russia
V.S. Baranov, Corr. Member of the RAS, Dr. Sci. (Medicine), Russia
L.A. Beshpalova, Full Member of the RAS, Dr. Sci. (Agric.), Russia
A. Börner, Dr. Sci., Germany
M.F. Chernov, Dr. Sci. (Medicine), Japan
G.L. Dianov, Professor, Dr. Sci. (Biology), Great Britain
Yu.E. Dubrova, Professor, Dr. Sci. (Biology), Great Britain
N.N. Dygalo, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
J. Le Gouis, Dr. Sci., France
I. Grosse, Professor, Dr. Sci., Germany
S.G. Inge-Vechtomov, Full Member of the RAS, Dr. Sci. (Biology), Russia
I.E. Kerkis, Dr. Sci., Brazil
E.K. Khlestkina, Professor, Dr. Sci. (Biology), Russia
L.V. Khotyleva, Full Member of the NAS of Belarus, Dr. Sci. (Biology), Belarus
E.K. Khusnutdinova, Professor, Dr. Sci. (Biology), Russia
A.V. Kilchevsky, Corr. Member of the NAS of Belarus, Dr. Sci. (Biology), Belarus
A.V. Kochetov, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
S.V. Kostrov, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia
B. Lugtenberg, Professor, Dr. Sci., Netherlands
V.I. Molodin, Full Member of the RAS, Dr. Sci. (History), Russia
V.P. Puzyrev, Full Member of the RAS, Dr. Sci. (Medicine), Russia
I.B. Rogozin, Cand. Sci. (Biology), United States
A.O. Ruvinsky, Professor, Dr. Sci. (Biology), Australia
A.Yu. Rzhetsky, Professor, Cand. Sci. (Biology), United States
E.A. Salina, Professor, Dr. Sci. (Biology), Russia
S.V. Shestakov, Full Member of the RAS, Dr. Sci. (Biology), Russia
K.V. Slavin, Professor, Dr. Sci., United States
V.A. Stepanov, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
I.A. Tikhonovich, Full 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
I.K. Zakharov, Professor, Dr. Sci. (Biology), Russia
I.A. Zakharov-Gezekhus, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

Editorial board

D.A. Afonnikov, Associate Professor, Cand. Sci. (Biology), Russia
T.G. Amstislavskaya, Dr. Sci. (Biology), Russia
E.E. Andronov, Cand. Sci. (Biology), Russia
Yu.S. Aulchenko, Dr. Sci. (Biology), Russia
E.V. Berezikov, Professor, Cand. Sci. (Biology), Netherlands
N.P. Bondar, Cand. Sci. (Biology), Russia
S.A. Borinskaya, Dr. Sci. (Biology), Russia
P.M. Borodin, Professor, Dr. Sci. (Biology), Russia
M. Chen, Dr. Sci. (Biology), People's Republic of China
V.N. Danilenko, Professor, Dr. Sci. (Biology), Russia
S.A. Demakov, Dr. Sci. (Biology), Russia
E.A. Dolgikh, Dr. Sci. (Biology), Russia
T.A. Gavrilenko, Dr. Sci. (Biology), Russia
Yu.M. Konstantinov, Professor, Dr. Sci. (Biology), Russia
O. Krebs, Professor, Dr. Sci. (Biology), Germany
D. Larkin, Dr. Sci. (Biology), Great Britain
I.N. Lavrik, Cand. Sci. (Chemistry), Germany
I.N. Lebedev, Professor, Dr. Sci. (Biology), Russia
L.A. Lutova, Professor, Dr. Sci. (Biology), Russia
V.Yu. Makeev, Corr. Member of the RAS, Dr. Sci. (Physics and Mathem.), Russia
M.P. Moshkin, Professor, Dr. Sci. (Biology), Russia
E. Pestsova, Dr. Sci. (Biology), Germany
L.Yu. Novikova, Cand. Sci. (Engineering), Russia
N.A. Provorov, Professor, Dr. Sci. (Biology), Russia
D.V. Pyshnyi, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia
A.V. Ratushny, Cand. Sci. (Biology), United States
M.G. Samsonova, Dr. Sci. (Biology), Russia
Y. Shavrukov, Dr. Sci. (Biology), Australia
E. Turuspekov, Cand. Sci. (Biology), Kazakhstan

- 5 **ОТ РЕДАКТОРА**
Биоинформатика и системная компьютерная биология
- 7 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2. *А.В. Цуканов, В.Г. Левицкий, Т.И. Меркулова*
- 18 **ОБЗОР**
Геномная изменчивость в регуляторных районах генов, ассоциированная с заболеваниями человека: механизмы влияния на транскрипцию генов и полногеномные информационные ресурсы, обеспечивающие исследование этих механизмов. *Е.В. Игнатьева, Е.А. Матросова*
- 30 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Улучшение качества сборки *de novo* транскриптомов ячменя на основе гибридного подхода для линий с изменениями окраски колоса и стебля. *Н.А. Шмаков*
- 39 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Поиск участников сигнального пути ауксина к его транспортерам PIN на основе метаанализа транскриптомов, индуцированных ауксином. *В.В. Коврижных, З.С. Мустафин, З.З. Багаутдинова*
- 46 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Филостратиграфический анализ генных сетей заболеваний человека. *З.С. Мустафин, С.А. Лашин, Ю.Г. Матушкин*
- 57 **ОБЗОР**
Пангеномы сельскохозяйственных растений. *А.Ю. Пронозин, М.К. Брагина, Е.А. Салина*
- 64 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса. *Е.А. Урбанович, Д.А. Афонников, С.В. Николаев*
- 71 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Автоматическое фенотипирование морфологии колоса тетра- и гексаплоидных видов пшеницы методами компьютерного зрения. *А.Ю. Пронозин, А.А. Паулиш, Е.А. Заварзин, А.Ю. Приходько, Н.М. Прохошин, Ю.В. Кручинина, Н.П. Гончаров, Е.Г. Комышев, М.А. Генаев*
- 82 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19. *О.И. Криворотько, С.И. Кабанихин, М.И. Сосновская, Д.В. Андорная*
- 92 **ОБЗОР**
Механический стресс клеток мозга, локальная трансляция и нейродегенеративные заболевания: молекулярно-генетические аспекты. *Т.М. Хлебодарова*
- Биотехнология**
- 101 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Биоинформационный анализ сплайс-лидерного транс-сплайсинга у регенерирующего плоского червя *Macrostomum lignano* показал его преобладание среди консервативных генов и генов стволовых клеток. *К.В. Устьянцев, Е.В. Березиков (на англ. языке)*
- 108 **ОБЗОР**
Macrostomum lignano как модельный объект для исследования генетики и геномики паразитических плоских червей. *К.В. Устьянцев, В.Ю. Вавилова, А.Г. Блинов, Е.В. Березиков*
- 117 **ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ**
Трансгенная клеточная линия с индуцируемой транскрипцией для исследования механизмов экспансии (CGG)_n повторов. *И.В. Грищенко, А.А. Тулупов, Ю.М. Рымарева, Е.Д. Петровский, А.А. Савелов, А.М. Коростышевская, Ю.В. Максимова, А.Р. Шорина, Е.М. Шитик, Д.В. Юдкин*
- 125 **ОБЗОР**
Продукция субтилизиновых протеаз в бактериях и дрожжах. *А.С. Розанов, С.В. Шеховцов, Н.В. Богачева, Е.Г. Першина, А.В. Ряполова, Д.С. Бытяк, С.Е. Пельтек*

- 5 FROM THE EDITOR
Bioinformatics and computational systems biology
- 7 ORIGINAL ARTICLE
Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. A.V. Tsukanov, V.G. Levitsky, T.I. Merkulova
- 18 REVIEW
Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms. E.V. Ignatieva, E.A. Matrosova
- 30 ORIGINAL ARTICLE
Improving the quality of barley transcriptome *de novo* assembling by using a hybrid approach for lines with varying spike and stem coloration. N.A. Shmakov
- 39 ORIGINAL ARTICLE
The auxin signaling pathway to its PIN transporters: insights based on a meta-analysis of auxin-induced transcriptomes. V.V. Kovrizhnykh, Z.S. Mustafin, Z.Z. Bagautdinova
- 46 ORIGINAL ARTICLE
Phylostratigraphic analysis of gene networks of human diseases. Z.S. Mustafin, S.A. Lashin, Yu.G. Matushkin
- 57 REVIEW
Crop pangenomes. A.Yu. Pronozin, M.K. Bragina, E.A. Salina
- 64 ORIGINAL ARTICLE
Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm. E.A. Urbanovich, D.A. Afonnikov, S.V. Nikolaev
- 71 ORIGINAL ARTICLE
Automatic morphology phenotyping of tetra- and hexaploid wheat spike using computer vision methods. A.Yu. Pronozin, A.A. Paulish, E.A. Zavarzin, A.Yu. Prikhodko, N.M. Prokhoshin, Yu.V. Kruchinina, N.P. Goncharov, E.G. Komyshev, M.A. Genaev
- 82 ORIGINAL ARTICLE
Sensitivity and identifiability analysis of COVID-19 pandemic models. O.I. Krivorotko, S.I. Kabanikhin, M.I. Sosnovskaya, D.V. Andornaya
- 92 REVIEW
The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases. T.M. Khlebodarova
- Biotechnology**
- 101 ORIGINAL ARTICLE
Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes. K.V. Ustyantsev, E.V. Berezikov
- 108 REVIEW
Macrostomum lignano as a model to study the genetics and genomics of parasitic flatworms. K.V. Ustyantsev, V.Yu. Vavilova, A.G. Blinov, E.V. Berezikov
- 117 ORIGINAL ARTICLE
A transgenic cell line with inducible transcription for studying (CGG)_n repeat expansion mechanisms. I.V. Grishchenko, A.A. Tulupov, Y.M. Rymareva, E.D. Petrovskiy, A.A. Savelov, A.M. Korostyshevskaya, Y.V. Maksimova, A.R. Shorina, E.M. Shitik, D.V. Yudkin
- 125 REVIEW
Production of subtilisin proteases in bacteria and yeast. A.S. Rozanov, S.V. Shekhovtsov, N.V. Bogacheva, E.G. Pershina, A.V. Ryapolova, D.S. Bytyak, S.E. Peltek

Уважаемые коллеги, дорогие читатели! Настоящий выпуск журнала имеет биоинформатическую направленность. В последнее десятилетие в результате стремительного совершенствования методов расшифровки геномов произошёл информационный взрыв такой силы, что генетика стала самым большим источником данных не только в мировой науке, но и во всех других сферах человеческой деятельности, включая социальные сети. Всё шире разворачиваются исследования генома человека в рамках крупных международных проектов. В проекте «1000 геномов» (<https://www.internationalgenome.org/>) на 14.08.2020 просеквенировано 3202 генома. В проекте «100 000 геномов» (<https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/>) проанализированы геномы 85 000 пациентов с редкими заболеваниями/раком.

В рамках проекта «1000 геномов быков» (<http://www.1000bullgenomes.com>) на 31.07.2020 было отсеквенировано свыше 5000 животных, относящихся более чем к 200 породам и видам крупного рогатого скота. В результате в геномах этих животных идентифицировано более 155 млн генетических вариантов (ОНП и небольшие инсерции/делетии). На 01.08.2019 в проекте по секвенированию геномов овец SheepGenomesDB (<https://sheepgenomesdb.org>) было отсеквенировано 935 животных, относящихся к 69 породам, у которых идентифицировано более 50 млн генетических вариантов. Проект «1000 геномов коз» (http://www.goatgenome.org/vargoats_data_access.html) на 09.11.2020 содержит данные о 127852473 генетических вариантах, выявленных у 1159 животных, относящихся к 101 породе коз.

Маркёр-ориентированная и геномная селекция, а также геномное редактирование потребовали расшифровки геномов основных сельскохозяйственных растений: пшеницы, кукурузы, ячменя, риса, сои, фасоли, картофеля, широкого спектра овощных и фруктовых культур и других (<http://plants.ensembl.org/species.html>; <http://www.plantgdb.org/prj/GenomeBrowser/>). Выполнен крупный проект по исследованию генетической изменчивости генома риса на основе секвенирования коллекции из 3000 образцов этого сельскохозяйственного растения из 89 стран (The 3,000 rice

genomes project. *Gigascience*. 2014;3:7. DOI 10.1186/2047-217X-3-7). С 2018 г. осуществляется проект по полному секвенированию геномов 10000 растений, относящихся к основнымкладам эмбриофитов, зелёных водорослей и протистов (за исключением грибов) (Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.M., Li F.W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K. 10KP: A phylodiverse genome sequencing plan. *Gigascience*. 2018;7(3):1-9. DOI 10.1093/gigascience/giy013). Расшифровано более 36000 полных геномов вирусов (<https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/>), секвенировано 163 645 полных геномов бактерий и 1886 полных геномов архей (<https://gold.jgi.doe.gov/distribution>). Прочитано более 2590 геномов грибов (проект «1000 геномов грибов», <https://mycoscosm.jgi.doe.gov/pages/fungi-1000-projects.jsf>).

Получены огромные объёмы информации по структуре белков. В базе данных UniProt <https://www.uniprot.org/> (хранилище данных аминокислотных последовательностей) содержится описание 563 082 экспериментально подтверждённых первичных структур белков, а в базе TrEMBL (<https://www.uniprot.org/statistics/TrEMBL>) хранится более 190 млн аминокислотных последовательностей, полученных на основе автоматической компьютерной аннотации геномов. Совершенствование методов физико-химического изучения белков обеспечило стремительное накопление сведений по их пространственным структурам (174 507 записей в базе данных PDB <https://www.rcsb.org/>). Ценнейшая информация о структуре белков представлена в базе масс-спектрометрических данных Chemdata.nist.gov (<https://chemdata.nist.gov/>), включающей описание свыше 100 млн масс-спектров химических пептидов и метаболитов из различных тканей, биологических жидкостей и клеток.

Следует отметить, что к настоящему времени реконструировано более 70 000 геномных сетей, путей передачи сигналов и метаболических путей, представленных в базе KEGG Pathway (ручная аннотация), базах данных систем STRING (<https://string-db.org/>), GeneMANIA (<https://genemania.org/>), Pathway Commons (<https://www.pathwaycommons.org/>) и других.

Особенно значимы для медицины гигантские объёмы информации по генетической изменчивости человека: в базе dbSNP (<https://www.ncbi.nlm.nih.gov/>) на текущий момент содержится более 72 млн записей об SNP в геномах человека (из них ~24 тыс. связаны с развитием заболеваний), а в базе Ensembl (http://www.ensembl.org/Homo_sapiens) приведено более 667 млн записей об SNP в геномах человека.

Информационный взрыв в генетике стал грандиозным вызовом, поскольку темпы накопления геномных данных на порядок опережают возможности их компьютерного анализа, из-за чего большая часть геномных проектов заканчивается их формальной сборкой с очень поверхностной аннотацией (или даже без неё) (<https://gold.jgi.doe.gov/>).

Всё это свидетельствует о фундаментальной значимости информационных технологий и биоинформатики для хранения, обработки и анализа геномных данных в интересах решения фундаментальных и прикладных задач генетики, медицины, фармакологии, сельского хозяйства, биотехнологии и биобезопасности.

Понимание и практическое применение огромных объёмов генетических экспериментальных данных исключительно высокой сложности потребовало разработки современных информационных технологий, эффективных методов компьютерного анализа больших данных

и математического моделирования биологических систем и процессов на различных иерархических уровнях организации живых систем, начиная с геномов, генов, белков, метаболических путей и генных сетей, включая клетки и ткани и заканчивая целостными организмами, популяциями и экосистемами. В этом номере вниманию читателей предложены статьи, посвящённые конкретным исследованиям по таким направлениям биоинформатики, как компьютерная геномика и транскриптомика, системная компьютерная биология, эволюционная компьютерная биология и автоматический анализ фенотипов растений.

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

Английский текст <https://vavilov.elpub.ru/jour>

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

А.В. Цуканов¹✉, В.Г. Левицкий^{1, 2}, Т.И. Меркулова^{1, 2}

¹ Федеральное исследовательское учреждение Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ tsukanov@bionet.nsc.ru

Аннотация. В настоящее время самой распространенной моделью поиска сайтов связывания транскрипционных факторов (ССТФ) в пиках ChIP-seq является позиционная весовая матрица (position weight matrix, PWM). Но эта модель не учитывает взаимосвязи между частотами встреч нуклеотидов в разных позициях ССТФ, поэтому не способна гарантировать определение всех возможных структурных вариантов ССТФ. На сегодняшний день уже предложены альтернативные модели, например BaMM и InMoDe, которые учитывают такие взаимосвязи. Однако применение этих моделей обычно сводилось к сравнению их точности с точностью традиционной модели PWM, тогда как анализ совместной встречаемости и относительного расположения ССТФ разных моделей в пиках не производился. В нашей работе мы предлагаем конвейер программ MultiDeNA, позволяющий сочетать разные модели *de novo* поиска ССТФ для выявления структурной гетерогенности ССТФ в данных ChIP-seq. Разработанный конвейер включает этапы построения моделей на основе заданного набора пиков, оценки точности распознавания моделей с помощью перекрестных тестов, выбора порогов, сканирования пиков ChIP-seq и классификацию пиков по результатам сканирования. С применением конвейера нами проведен анализ 22 экспериментов ChIP-seq для ТФ FOXA2 с помощью четырех моделей: PWM, diPWM, BaMM и InMoDe. Показано, что сочетание моделей позволяет существенно увеличить общее количество распознанных пиков (на 26.3 %) по сравнению с применением только PWM; при этом основной вклад в распознавание внесла модель BaMM. В значительной доле пиков разные модели распознают совпадающие ССТФ; однако для моделей PWM, diPWM, BaMM и InMoDe медианы доли пиков, которые содержали ССТФ только одной модели, составили 1.08, 0.49, 4.15 и 1.73 % соответственно. Таким образом, совокупность ССТФ FOXA2 не описывается полностью только одной моделью, что свидетельствует о наличии структурной гетерогенности в ССТФ у FOXA2. Ключевые слова: сайты связывания транскрипционных факторов (ССТФ); *de novo* поиск ССТФ; ChIP-seq; гетерогенность ССТФ.

Для цитирования: Цуканов А.В., Левицкий В.Г., Меркулова Т.И. Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2. Вавиловский журнал генетики и селекции. 2021;25(1):7-17. DOI 10.18699/VJ21.002

Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites

A.V. Tsukanov¹✉, V.G. Levitsky^{1, 2}, T.I. Merkulova^{1, 2}

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

² Novosibirsk State University, Novosibirsk, Russia

✉ tsukanov@bionet.nsc.ru

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

were 1.08, 0.49, 4.15 and 1.73 % for PWM, diPWM, BaMM and InMoDe, respectively. Thus, FOXA2 BSs were not fully described by only a sole model, which indicates their heterogeneity. We assume that the BaMM model is the most successful in describing the structure of the FOXA2 BS in ChIP-seq datasets under study.

Key words: transcription factor binding sites (TFBS); TFBS *de novo* searching; ChIP-seq; heterogeneity of TFBS.

For citation: Tsukanov A.V., Levitsky V.G., Merkulova T.I. Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):7-17. DOI 10.18699/VJ21.002

Введение

Транскрипционные факторы (ТФ) – белки, способные распознавать определенные участки ДНК (сайты связывания ТФ, ССТФ) (Lambert et al., 2018) и как повышать, так и снижать уровень транскрипции генов (Latchman, 2001). Этап связывания ТФ с ДНК является ключевым для регуляции экспрессии генов, поскольку инициирует цепь молекулярных событий, обеспечивающих сборку/регуляцию активности преинициаторного комплекса РНК-полимеразы II за счет непосредственных или опосредованных контактов с компонентами этого комплекса, а также благодаря привлечению различных модифицирующих хроматин и ремоделирующих белков и, как следствие, локальных изменений структуры хроматина (Iwafuchi-Doi, 2019; Srivastava, Mahony, 2020). Поэтому одна из важнейших задач современной молекулярной биологии – это идентификация всего массива ССТФ в геноме.

В настоящее время для решения этой задачи широко применяется метод, основанный на иммунопреципитации хроматина с использованием антител к исследуемому ТФ с последующим высокопроизводительным секвенированием преципитированной ДНК – ChIP-seq (Farnham, 2009; Park, 2009). Первичная обработка данных экспериментальных методов позволяет выявлять участки ДНК, или пики, для которых ТФ напрямую или через некоторого посредника был связан с ДНК (Furey, 2012). Поскольку длина пиков обычно исчисляется в сотнях пар оснований (п. о.), а протяженность ССТФ не превышает 20–25 п. о. (Levitsky et al., 2007; Kulakovskiy et al., 2018), следующим этапом биоинформатической обработки данных ChIP-seq является поиск ССТФ в полученных пиках. Для этого разработано множество инструментов, большинство из которых основано на использовании позиционных весовых матриц (position weight matrix, PWM) (Stormo, 2000), включая такие популярные, как ChIPMunk (Kulakovskiy, Makeev, 2009) и Homer (Heinz et al., 2010). Без преувеличения можно сказать, что применение разных реализаций модели PWM входит практически в каждый конвейер обработки полногеномных данных (Lloyd, Bao, 2019).

Применение стандартного подхода, основанного на использовании PWM, к обработке данных ChIP-seq показывает, что примерно в половине пиков для большинства ТФ не обнаруживаются соответствующих мотивов (Worsley Hunt, Wasserman, 2014; Gheorghe et al., 2019). Традиционно это связывают с главным недостатком PWM – гипотезой независимости частот встреч нуклеотидов в разных позициях ССТФ, которая не всегда подтверждается, что негативно сказывается на точности распознавания (Venos et al., 2002; Keilwagen, Grau, 2015). Поэтому разрабатываются альтернативные методы распознавания ССТФ, где

тем или иным способом учитываются зависимости между нуклеотидами в модели сайта (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). С одной стороны, самой простой моделью, которая старается учитывать зависимости между соседними нуклеотидами, является динуклеотидная PWM (dinucleotide position weight matrix, diPWM) (Zhang M., Marr, 1993; Kulakovskiy et al., 2013). С другой стороны, предложены такие модели, как BaMM (Siebert, Söding, 2016) и InMoDe (Eggeling et al., 2017). Они построены с использованием марковских цепей, которые учитывают зависимости позиций с помощью концепции порядка марковской цепи, т. е. участка, длина которого обычно не превышает 5 п. о. и в пределах которого частоты нуклеотидов могут быть зависимыми.

Авторы альтернативных моделей часто доказывают, что их модели могут иметь более высокую точность, чем PWM, однако ни одна из этих моделей сама по себе не решает проблему неполного распознавания ССТФ в пиках ChIP-seq. Мы предполагаем, что частично проблема обусловлена структурной гетерогенностью сайтов связывания для одного и того же ТФ и число распознанных пиков может быть значительно увеличено при одновременном использовании разных моделей. При этом данные ChIP-seq будут содержать как ССТФ, предсказываемые одновременно двумя и более моделями, так и ССТФ, предсказываемые только одной из моделей (Ignatieva et al., 2004; Levitsky et al., 2014, 2016). Ранее при анализе двух независимых экспериментов ChIP-seq для ТФ FOXA2 (Wederell et al., 2008; Wallerman et al., 2009) с помощью альтернативных моделей ChIPMunk (*de novo* PWM (Kulakovskiy, Makeev, 2009)) и SiteGA (по выборке обучения из 53 известных сайтов ТФ подсемейства FOXA (Levitsky et al., 2007)) и экспериментально подобранных порогов моделей (эксперимент EMSA, electrophoretic mobility shift assay – сдвиг в анализе электрофоретической подвижности) удалось обнаружить FOXA2 сайты более чем в 95 % пиков (Levitsky et al., 2014), что согласуется с отсутствием в литературе каких-либо данных о непрямом взаимодействии этого хорошо изученного ТФ с ДНК.

Приведенный пример указывает на перспективность сочетания альтернативных методов поиска ССТФ с матричной моделью для анализа ChIP-seq данных. Однако до сих пор не было систематических исследований на эту тему. Альтернативные модели для поиска ССТФ не получили широкого применения, несмотря на то что уже около 20 лет известно о наличии зависимости частот встреч нуклеотидов в разных позициях ССТФ (Bulyk et al., 2002). В качестве косвенного показателя популярности разных моделей можно привести количество цитирований статей, в которых обсуждаются конкретные программы *de novo*

поиска ССТФ. Так, на конец 2020 г. статьи, посвященные реализации матричной модели в виде программ MEME (Bailey, Elkan, 1994; Machanick, Bailey, 2011), HOMER (Heinz et al., 2010) и ChIPMunk (Kulakovskiy et al., 2010), имеют суммарное количество цитирований более 6000, а статьи, посвященные альтернативным моделям BaMM (Siebert, Söding, 2016; Kiesel et al., 2018), InMoDe (Eggeling et al., 2017) и diChIPMunk (Kulakovskiy et al., 2013), – чуть более 50. При этом конкретные исследования (отдельные эксперименты ChIP-seq) почти всегда анализируются только с использованием стандартной модели PWM. Такое положение можно объяснить следующими причинами: 1) простота применения PWM и доступность в понимании результатов этой модели; 2) недостаточное понимание преимуществ альтернативных моделей, которые, помимо лучшей точности в сравнении с PWM, способны находить ССТФ иной структуры.

В данной работе мы предлагаем конвейер программ, который сочетает четыре модели *de novo* поиска ССТФ, а именно: PWM, реализованную в программе ChIPMunk (Kulakovskiy et al., 2010); diPWM, реализованную в программе diChIPMunk (Kulakovskiy et al., 2013); и две марковские модели – InMoDe (Eggeling et al., 2017) и BaMM (Siebert, Söding, 2016). Конвейер оценивает точность распознавания моделей, выбирает их пороги и проводит классификацию ChIP-seq пиков, сравнивая результаты сканирования всех моделей. Такой подход позволит расширить наши представления о структурном разнообразии ССТФ при прямом связывании ТФ с ДНК, особенно для случаев, когда модель PWM не способна найти ССТФ. Работа конвейера апробирована в ходе анализа данных 22 экспериментов ChIP-seq для ТФ FOXA2.

Материал и методы

Исходные данные. Для анализа использовали набор предобработанных ChIP-seq данных в виде разметки пиков в формате bed из базы данных ReMap <http://remap.univ-amu.fr/> (Chèneby et al., 2020). Набор данных включал 22 ChIP-seq эксперимента для ТФ FOXA2 (см. таблицу). Из каждого эксперимента для анализа брали только лучшие 4000 пиков (см. далее раздел «Подготовка первичных данных»).

Помимо ChIP-seq пиков, на вход в конвейер программ указывали список доступных программ (PWM, diPWM, BaMM, InMoDe) *de novo* поиска ССТФ, включая путь к программам. Также устанавливали версию генома – mm10 или hg38; этот параметр позволяет выбрать список промоторов в формате fasta – 5'-участки кодирующих белок генов (2000 п. о. от сайта старта транскрипции). Общий объем выборки составил 19795 генов для версии генома человека GRCh38.p13 и 19991 ген для версии генома мыши GRCm38.p6. Для извлечения последовательностей нуклеотидов по координатам пиков использовали референсный геном в формате fasta.

Конвейер программ для выявления структурной гетерогенности ССТФ. Нами был разработан конвейер программ MultiDeNA (multiple *de novo* analysis, <https://github.com/ubercomrade/MultiDeNA>) для поиска ССТФ с помощью нескольких *de novo* моделей в данных ChIP-seq. Данный конвейер программ позволяет получить класси-

Список ChIP-seq экспериментов, используемых в работе

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

Примечание. GEO/ENCODE – уникальный идентификатор баз данных (GSE*/ENC*); TomTom – результат фильтрации данных с помощью программы TomTom; «+»/«–» – частотная матрица, построенная на основе ССТФ, найденных ChIPMunk (PWM), значимо похожа (p -value < 0.001)/не похожа (p -value > 0.001) на частотную матрицу ССТФ FOXA2 из HOCOMO FOXA2_HUMAN.H11MO.0.A.

фикацию пиков ChIP-seq, по результатам которой можно оценить структурное разнообразие ССТФ. В настоящее время конвейер использует модели ChIPMunk (PWM), diChIPMunk (diPWM), BaMM и InMoDe, а также вспомогательные программы bedtools (Quinlan, Hall, 2010) и TomTom (Gupta et al., 2007). Принципиальная схема конвейера программ представлена на рис. 1. Конвейер включает в себя следующие этапы: подготовка данных; построение моделей; оценка точности моделей; выбор порогов, поиск ССТФ в пиках ChIP-seq с фиксированным порогом; классификация ChIP-seq пиков по результатам распознавания ССТФ *de novo* моделями. Каждый этап конвейера программ детально описан ниже.

Подготовка первичных данных включала сортировку пиков по округленному значению $-10 \cdot \log_{10}(p\text{-value})$, которое было ранее вычислено для каждого пика программой MACS (Zhang Y. et al., 2008) и характеризовало

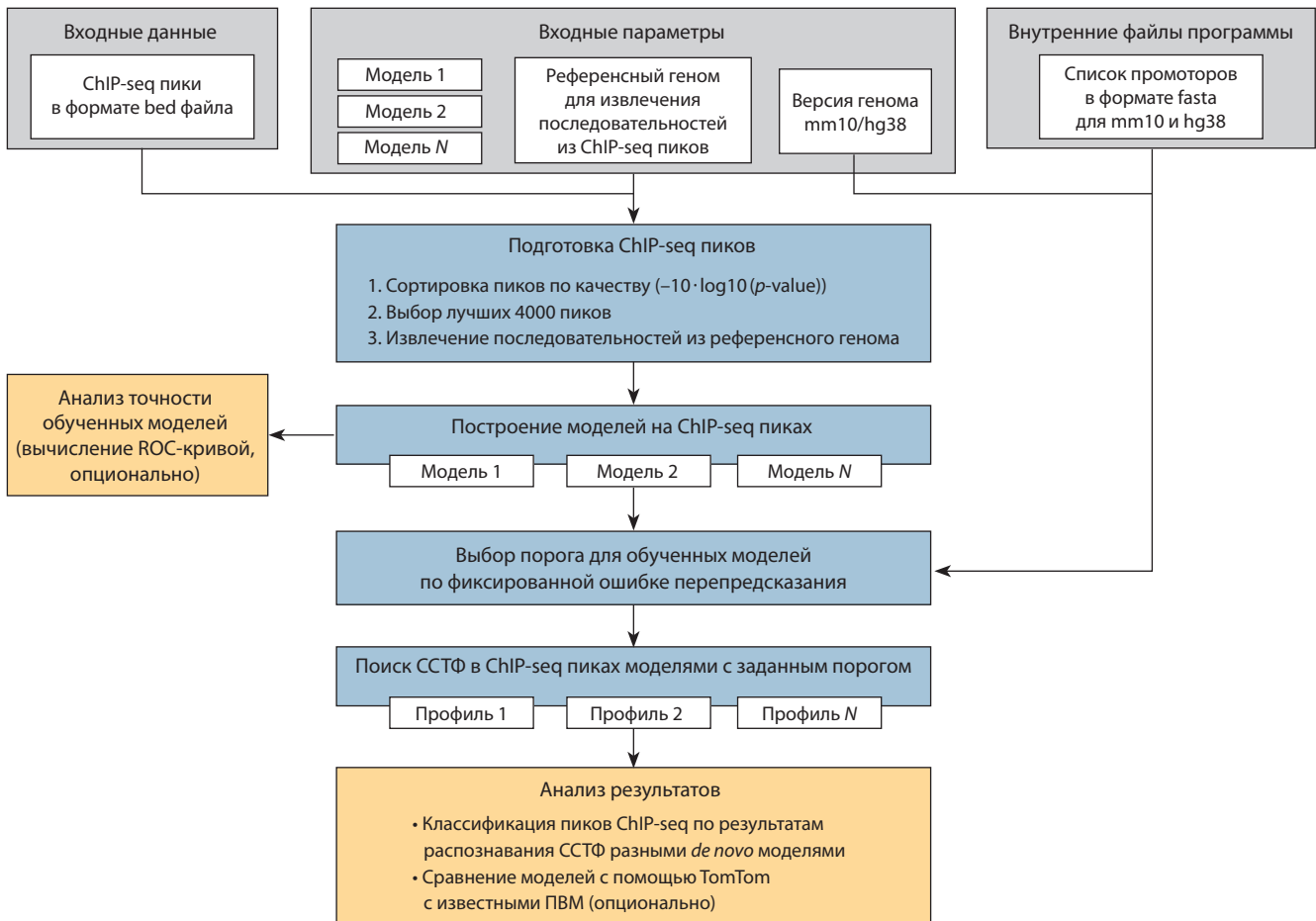


Рис. 1. Принципиальная схема работы конвейера программ.

качество пика. Эту программу конвейер базы ReMap (Chèpeby et al., 2020) использовал для обработки сырых данных ChIP-seq. Из каждого набора данных ChIP-seq для анализа мы взяли 4000 лучших по качеству пиков. Далее извлекали нуклеотидные последовательности пиков из генома с помощью bedtools (Quinlan, Hall, 2010).

Построение *de novo* моделей и оценка их точности распознавания ССТФ. Для того чтобы распознавать ССТФ в пиках, необходимо построить *de novo* модели. Построение нетрадиционных моделей ССТФ осуществлялось программами BaMM (Siebert, Söding, 2016) и InMoDe (Eggeling et al., 2017), а модели PWM и diPWM строили соответственно с помощью ChIPMunk и diChIPMunk (Kulakovskiy et al., 2010, 2013).

Чтобы улучшить точность распознавания ССТФ для PWM, подбирали ее оптимальную длину методом перекрестных тестов; эту же длину использовали и при построении других моделей. Метод оценки точности включал следующие этапы: 1) разделение данных на выборку обучения – случайно отобранные 90 % пиков от исходных данных, и контрольную выборку, включавшую оставшиеся 10 % пиков; 2) построение модели на выборке обучения; 3) проверка модели на контрольной выборке для оценки доли верноположительных результатов (ДВР); 4) генерация выборки случайных последовательностей путем случайной перестановки нуклеотидов в последо-

вательностях контрольной выборки; 5) проверка модели на выборке случайных последовательностей для оценки доли ложноположительных результатов (ДЛР); 6) повторение этапов 1–5 несколько раз; 7) вычисление ROC-кривой (receiver operating characteristic) на основе полученных данных. Разные длины модели сравнивали по показателю pAUC (partial area under curve), вычисленному как часть площади под кривой ROC для всех значений ДЛР, меньших 0.001 (McClish, 1989; Siebert, Söding, 2016). Описанный выше способ выбора оптимальной длины PWM на основе наилучшей точности распознавания ССТФ был разработан ранее (Levitsky et al., 2007; Kulakovskiy et al., 2013). Аналогичным методом оценивали точность всех моделей.

После того как модель построена, ее можно применять к последовательности нуклеотидов, равной длине модели. Результатом применения модели является значение функции распознавания. Чем больше это значение, тем выше вероятность того, что оцениваемая последовательность нуклеотидов является функциональным ССТФ.

Выбор порога для моделей на основе фиксированной ошибки перепредсказания. Чтобы корректно сравнивать результаты поиска ССТФ разных моделей, необходимо единообразно установить для всех моделей пороговые значения их функций распознавания. Эти пороги определяли по фиксированной ошибке перепредсказания. Для ее

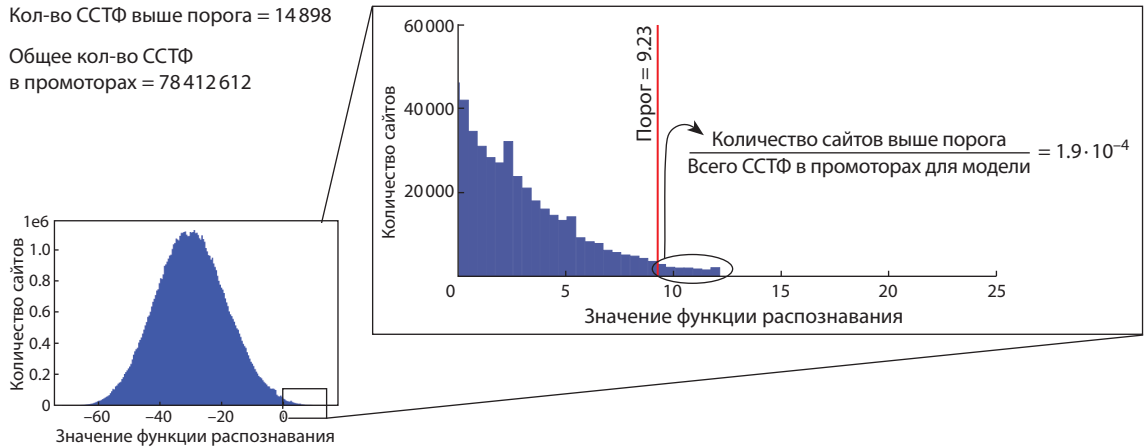


Рис. 2. Выбор порога для модели по фиксированной ошибке перепредсказания с использованием в качестве негативной выборки последовательности промоторов.

вычисления использовали негативную выборку, в которую входили 5'-участки кодирующих белок генов (2000 п. о. от сайта старта транскрипции).

Величину ошибки перепредсказания вычисляли следующим образом. Определяли значение функции распознавания модели для каждого сайта в негативной выборке в каждой позиции и цепи ДНК. Затем оценивали величину ошибки перепредсказания для каждого уникального значения функции распознавания как отношение количества предсказанных ССТФ, для которых значение функции выше этого порога, к общему числу позиций в выборке, доступных для таких ССТФ. При распознавании ССТФ для всех моделей в качестве порога использовали такое значение функции распознавания, при котором ошибка перепредсказания составляла $1.9 \cdot 10^{-4}$. После того как порог выбран для каждой модели, сканировали пики ChIP-seq. Пример выбора порога для PWM длиной 20 п. о. на данных GSE92491 приведен на рис. 2.

Классификация пиков ChIP-seq по результатам распознавания ССТФ разными моделями. После того как для каждой модели был выбран порог, мы искали ССТФ в пиках ChIP-seq. Результаты сканирования записывали в файл bed формата. Далее пики классифицировали на фракции в зависимости от присутствия/отсутствия сайтов, найденных разными моделями (PWM, diPWM, BaMM, InMoDe), как с учетом расположения ССТФ разных моделей в позициях пиков, так и без такого учета (на основе присутствия или отсутствия сайтов в пиках), согласно ранее разработанной методике (Levitsky et al., 2014, 2016). В частности, классификацию пиков с учетом позиций ССТФ разных моделей проводили для каждой пары моделей. Всего было шесть пар моделей: PWM и diPWM, PWM и BaMM, PWM и InMoDe, BaMM и diPWM, BaMM и InMoDe, InMoDe и diPWM. Если в пике присутствовали ССТФ, предсказанные только одной моделью, то данный пик классифицировался как пик соответствующей модели. Если в пике найдены ССТФ, предсказанные двумя разными моделями, то возможны два исхода (рис. 3).

В первом случае, если существует хотя бы одна пара сайтов от разных моделей, которые имеют как минимум одну общую позицию, такой пик классифицируется как

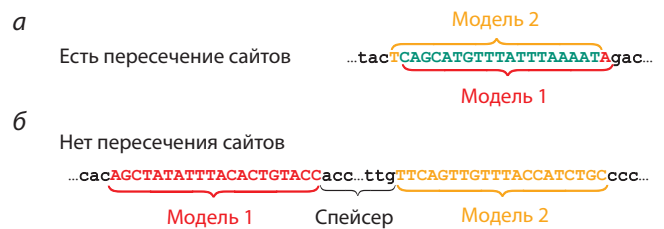


Рис. 3. Пример классификации ChIP-seq двух пиков, в которых обнаружены сайты двух разных моделей: *а* – в пике сайты пересекаются; *б* – пересечения нет.

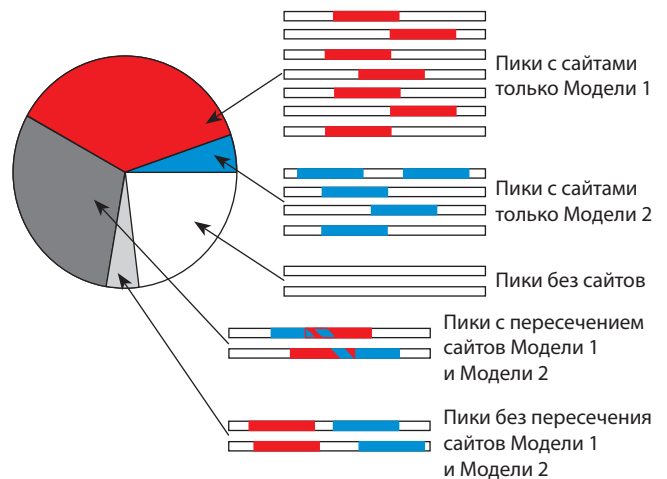


Рис. 4. Классификация ChIP-seq пиков для двух моделей с учетом пересечения ССТФ.

«пересечение сайтов». В другом случае, когда в пике присутствуют ССТФ, найденными разными методами, но их последовательности не пересекаются, пик классифицируется как «нет пересечения». Если в пике нет сайтов, то он классифицируется как «нет сайтов». Представить такую классификацию ChIP-seq пиков для двух моделей можно в виде круговой диаграммы (рис. 4).

Классификацию пиков без учета позиций ССТФ от разных моделей проводили следующим образом. Выделяли группы пиков, где присутствуют только сайты одной из моделей, пики, содержащие сайты всех моделей, а также пики, содержащие сайты комбинации моделей.

Сравнение найденных ССТФ с известными с помощью программы TomTom. Чтобы оценить, соответствуют ли ССТФ, которые находят модели, известным сайтам FOXA2, мы применили программу сравнения мотивов TomTom (Gupta et al., 2007). Эта программа предназначена для оценки значимости схожести частотных матриц. Для каждой PWM модели на основе найденных с ее помощью сайтов строили матрицу частот нуклеотидов. Далее с помощью TomTom оценивали схожесть этой матрицы с частотной матрицей ССТФ FOXA2 из базы данных HOCOMOFO FOXA2_HUMAN.H1MO.0.A (Kulakovskiy et al., 2018). Если при сравнении матриц значение *p*-value было меньше 0.001, то считали, что ChIP-seq обогащен ССТФ FOXA2 (см. таблицу).

Статистический анализ и визуализацию данных выполняли на языке программирования Python 3.8 в среде Jupyter с использованием пакетов numpy, matplotlib, seaborn и statannot. Распределения сравнили с помощью U-критерия Манна–Уитни с поправкой на множественные сравнения Бонферрони.

Результаты и обсуждение

Фильтрация данных на основе сравнения мотивов программой TomTom

Чтобы убедиться, что построенные модели сайтов соответствуют известным сайтам FOXA2 и последующий анализ является корректным, применили фильтр на основе программы оценки сходства мотивов TomTom. Для этого частотные матрицы ССТФ для модели PWM сравнивали с соответствующими матрицами известных ССТФ из базы данных HOCOMOFO. Только в шести из 22 ChIP-seq наборов, согласно TomTom, построенная матричная модель не обладала сходством с известными сайтами FOXA2 (см. таблицу), поэтому в дальнейшем анализе использовали оставшиеся 16 наборов.

Классификация пиков ChIP-seq без учета пересечения ССТФ, найденных разными *de novo* моделями

Основным результатом работы MultiDeNA является классификация пиков, которая позволяет установить, как соотносятся модели между собой, по способности выявлять пики с ССТФ. Всего используются два типа классификации пиков: с учетом пересечения позиций ССТФ разных моделей и без него. Результаты классификации приведем на примере данных GSE90454.FOXA2.KerCT (рис. 5).

Рассмотрим более детально классификацию ChIP-seq пиков по результатам поиска ССТФ четырьмя моделями без учета позиций сайтов. Можно видеть, что все четыре модели совместно распознали 88.35 % пиков (3534 из 4000, сумма всех областей на диаграмме Венна, см. рис. 5, а, б). Общая для всех методов группа пиков, в которых ССТФ были найдены четырьмя моделями одновременно, составила 34.25 % (1370 из 4000 пиков). Зна-

чительный вклад в распознавание пиков (34.55 %) вносят нематричные методы (BaMM и InMoDe): $696 + 647 + 39 = 1382$ из 4000, что сопоставимо с фракцией перекрытия всех моделей (1370). При этом самый крупный независимый вклад в распознавание вносит модель BaMM, которая добавляет 17.4 % пиков (696), в отличие от моделей PWM, InMoDe и diPWM, которые добавляют 0.525 % (21), 0.975 % (39) и 0.2 % (8) соответственно.

Чтобы оценить структурное разнообразие ССТФ, мы построили лого для фракций пиков «только PWM», «только diPWM», «только BaMM», «только InMoDe» и «все модели» (см. рис. 5, в). Во всех полученных лого можно выделить стандартный консенсус GTAAACA, однако для первых двух нуклеотидов консенсуса у фракций «только PWM», «только diPWM» и «только InMoDe» частота встречаемости GT меньше, чем AT. Можно также отметить, что 5'-концы всех лого разнообразны по информационному и нуклеотидному содержанию.

Классификация пиков ChIP-seq с учетом пересечения ССТФ, найденных разными моделями

Описанная выше классификация пиков без учета позиций ССТФ не учитывает тот факт, что используемые нами модели могут находить сайты в разных позициях одного и того же пика. Чтобы принять во внимание данное обстоятельство, была проведена классификация пиков с учетом позиций ССТФ для каждой пары моделей (PWM–diPWM, PWM–BaMM, PWM–InMoDe, diPWM–BaMM, diPWM–InMoDe, InMoDe–BaMM). Результаты классификации пиков на примере данных GSE90454.FOXA2.KerCT показаны в виде круговых диаграмм (рис. 6).

Все пары сочетаний моделей имеют незначительный класс пиков «нет пересечения», который варьирует от 0.3 до 6.9 %. С другой стороны, для всех случаев характерна большая фракция пиков «только пересечение»: BaMM–InMoDe – 53.6 %, PWM–diPWM – 44.4 %, diPWM–BaMM – 41.0 %, PWM–BaMM – 37.3 %, diPWM–InMoDe – 35.4 %, PWM–InMoDe – 31.6 %; при этом данная фракция больше для методологически близких пар моделей BaMM–InMoDe и PWM–diPWM (см. рис. 6). Класс пиков, где ССТФ находятся только одной из моделей, наиболее выражен для BaMM. В парах PWM–BaMM, diPWM–BaMM и InMoDe–BaMM он преобладает относительно второй модели пары (39.2, 36.4 и 26.8 % соответственно).

Оценка точности распознавания ССТФ для FOXA2 разными моделями

Чтобы сравнить, насколько точно разные модели способны распознавать ССТФ для FOXA2, по каждому эксперименту для всех четырех моделей рассчитали меру точности распознавания *rAUC* по кривой ROC, полученной с помощью перекрестного теста (см. выше раздел «Построение *de novo* моделей и оценка их точности распознавания ССТФ») (рис. 7, а). Согласно полученным данным, значения медиан *rAUC* для моделей PWM, diPWM, BaMM и InMoDe равны $8.0E-4$, $8.1E-4$, $7.3E-4$ и $5.6E-4$ соответственно. Полученные значения *rAUC* в парных сравнениях для PWM, diPWM и BaMM значимо не отличаются ($p > 0.05$), однако для InMoDe оно достоверно меньше, чем у остальных моделей ($p < 0.05$).

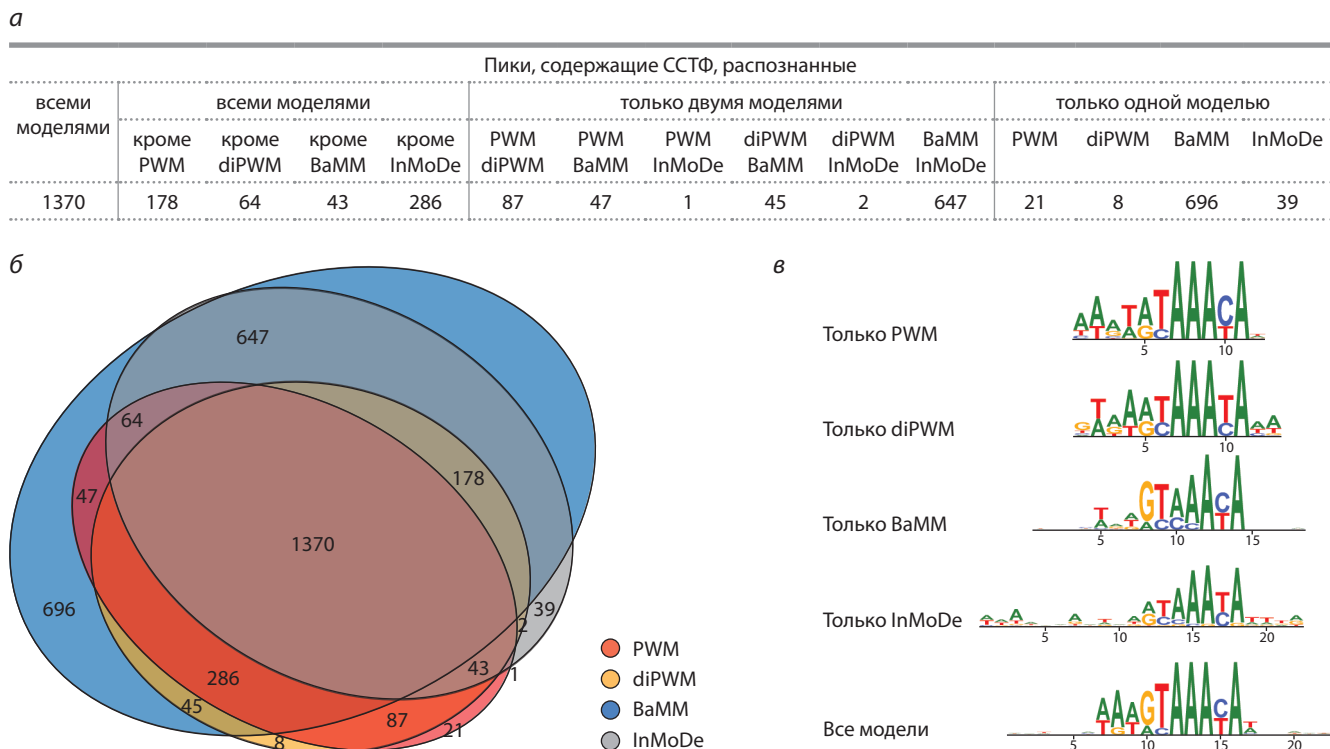


Рис. 5. Классификация пиков по результатам сканирования всеми четырьмя моделями.

а – таблица; *б* – диаграмма Венна; *в* – лого для фракций пиков, содержащих сайты только одной из моделей, и для фракции, где сайты всех моделей пересечены. Проанализирован набор данных GSE90454.FOXA2.KerCT.

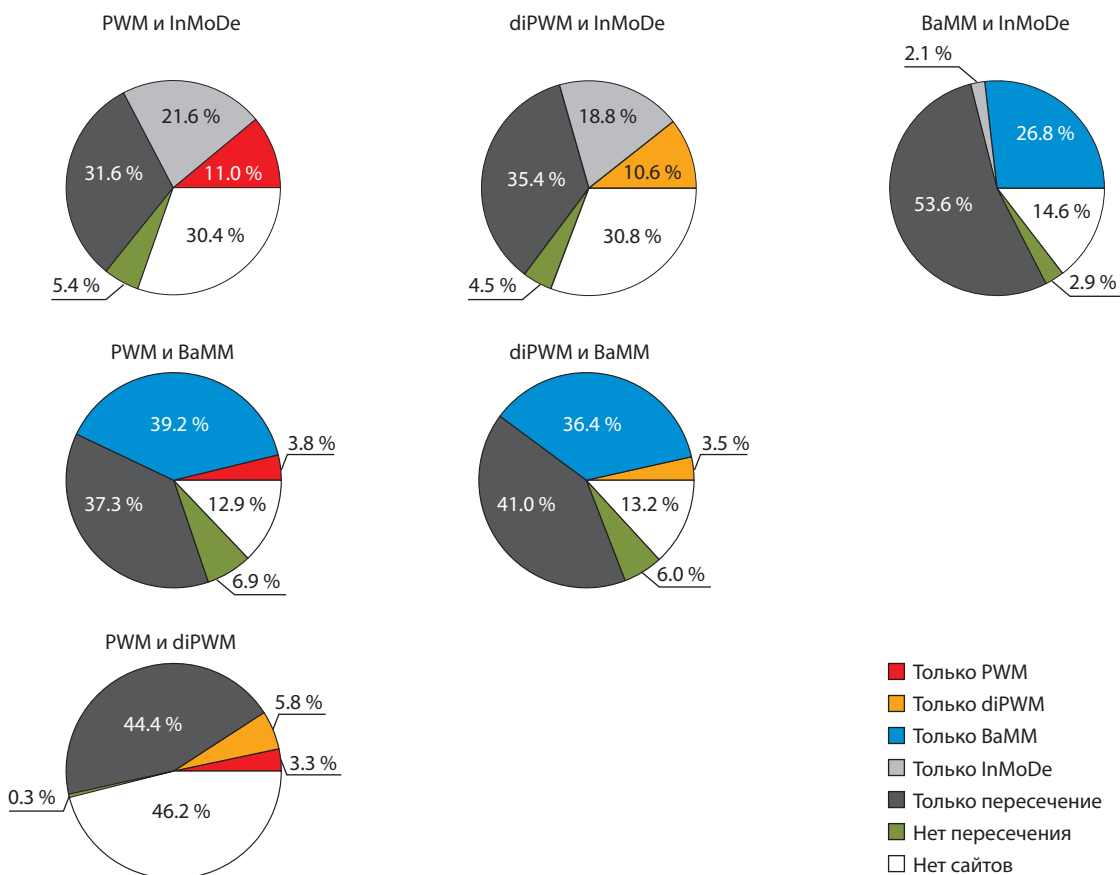


Рис. 6. Классификация ChIP-seq пиков с учетом пересечения ССТФ, распознанных разными моделями на примере данных GSE90454.FOXA2.KerCT.

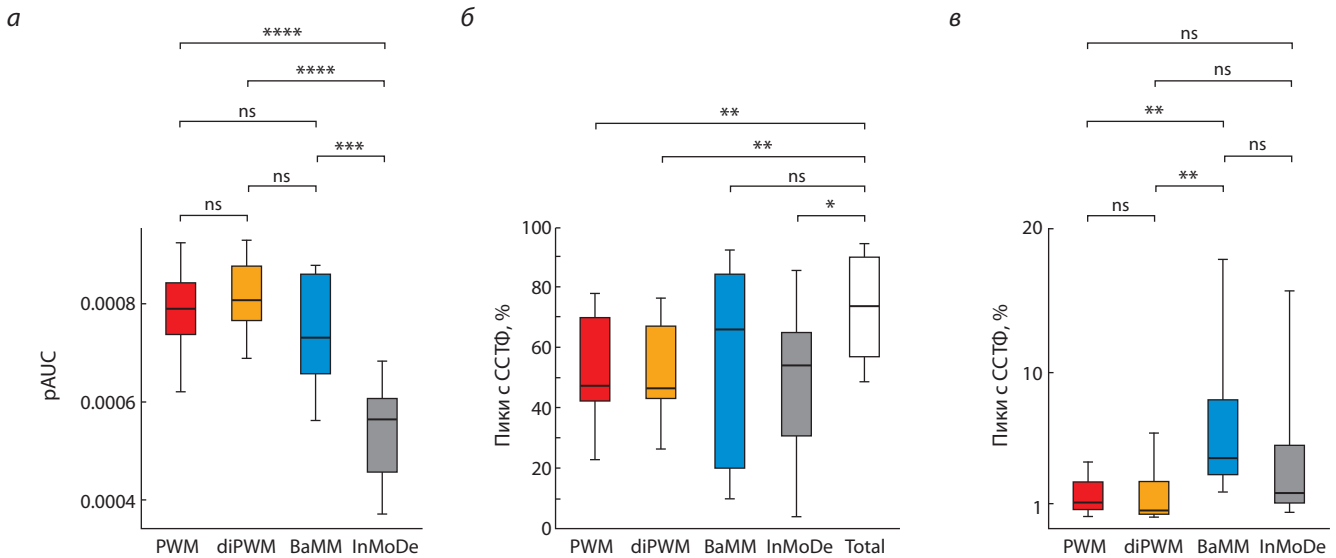


Рис. 7. Диаграммы распределения квартилей для данных: а – значения pAUC для всех моделей по всем ChIP-seq экспериментам; б – значения доли пиков с ССТФ, распознанных каждой моделью в отдельности (PWM, diPWM, BaMM, InMoDe) и всеми моделями (Total); в – значения доли пиков, в которых ССТФ находится только одной из моделей.

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

Сравнение долей пиков с ССТФ, найденных каждой моделью и всеми моделями. Чтобы исследовать вклады разных моделей в эффективность поиска ССТФ FOXA2 и оценить общий результат использования нескольких моделей для поиска ССТФ, мы определили, в какой доле пиков каждая модель и все модели вместе распознают хотя бы один ССТФ для FOXA2 (см. рис. 7, б). Значения медиан доли распознанных пиков составили 47.3, 46.4, 65.8 и 54 % для PWM, diPWM, BaMM и InMoDe соответственно, а медиана доли распознанных пиков при сочетании результатов всех четырех моделей равна 73.6 %. Следовательно, совместно все модели находят на 26.3 % больше пиков, содержащих ССТФ, чем модель PWM, что согласуется с ранее полученным результатом применения двух принципиально разных моделей PWM и SiteGA (Levitsky et al., 2014). При этом доли распознанных пиков для моделей PWM, diPWM и InMoDe значимо отличаются ($p < 0.05$) от результата, полученного сочетанием четырех моделей. Таким образом, подход с сочетанием разных моделей позволяет лучше выявлять пики с ССТФ для FOXA2, чем использование только одной модели. Однако для BaMM доля распознанных пиков статистически не отличается ($p > 0.05$) от результата, полученного сочетанием четырех моделей. Можно предположить, что модель BaMM вносит основной вклад в распознавание пиков FOXA2 и, возможно, лучше описывает структуру сайтов FOXA2. Тем не менее остальные модели добавляют еще 7.8 % пиков к результатам BaMM, что доказывает эффективность совместного использования разных моделей.

Сравнение долей пиков, содержащих ССТФ, распознанные только одной из моделей. Как показано выше, сочетание разных моделей увеличивает количество пиков с ССТФ, соответственно каждая модель должна распознавать ССТФ, которые не распознаются остальными. Чтобы оценить вклады в поиск ССТФ, специфичных

только для конкретной модели, были определены доли пиков, содержащих ССТФ только одной из моделей (см. рис. 7, в). Как видно из представленных данных, каждая модель (PWM, diPWM, BaMM, InMoDe) способна находить ССТФ, которые не обнаруживаются остальными моделями. Значения медиан по доле пиков, содержащих ССТФ только одной из моделей, для PWM, diPWM, BaMM и InMoDe составили 1.08, 0.49, 4.15 и 1.73 % соответственно. При этом данные по BaMM значимо отличаются ($p < 0.05$) как от PWM, так и от diPWM. Полученный результат подтверждает предположение, что модель BaMM может лучше описывать ССТФ FOXA2. Тем не менее каждая модель вносит вклад в распознавание сайтов. Следовательно, каждая из моделей может выявлять один из структурных вариантов ССТФ, который другие модели не находят.

Перекрестная проверка моделей PWM на данных ChIP-seq, на которых модели не обучались

Чтобы понять, насколько специфика одного ChIP-seq набора, в котором обучалась модель, может повлиять на точность распознавания ССТФ этой же моделью в других ChIP-seq данных, мы провели перекрестную проверку. Оценили точность каждой модели PWM не только внутри того же набора данных, где обучалась модель (для этого случая проводили несколько итераций разделения всей выборки обучения, так что модель обучалась на 90 % пиков, а тестировалась на оставшихся 10 % пиков), но и на остальных 15 наборах данных (контрольных). Для каждого случая рассчитали оценку точности pAUC (см. выше раздел «Построение *de novo* моделей и оценка их точности распознавания ССТФ»), результаты представили в виде тепловой карты (рис. 8). Из тепловой карты видно, что только в трех случаях – ENCSR000BRE.A-549, ENCSR000BNI.Hep-G2 и ERP008682.pancreas – другие

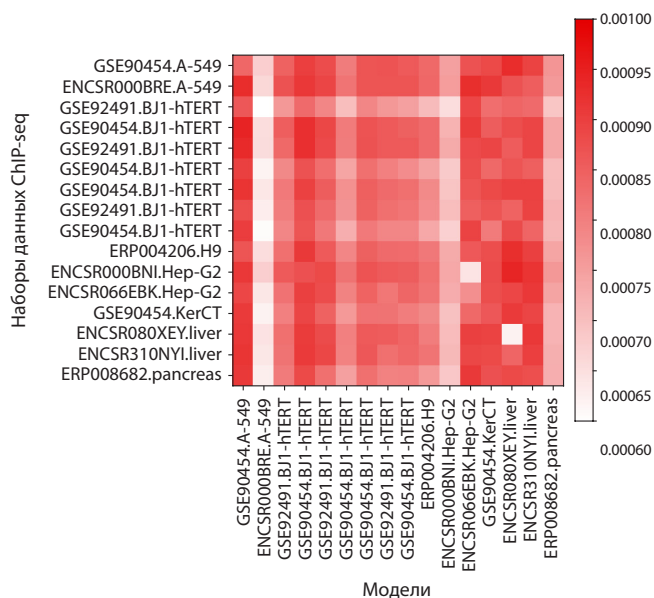


Рис. 8. Тепловая карта сравнения pAUC.

Цвета соответствуют значениям pAUC. Для ячеек, расположенных по диагонали, контрольные и обучающие наборы данных совпадают. В остальных ячейках они различаются. Строки означают модели, столбцы – наборы данных ChIP-seq.

модели имеют очень низкую оценку pAUC, а для случаев GSE90454.A-549, ENCSR066EBK.Hep-G2, GSE90454.KerCT, ENCSR080XEY.liver и ENCSR310NYL.liver все модели имеют высокое значение pAUC.

Обсуждение

На основе полученных данных можно заключить, что совместное использование альтернативных моделей с PWM позволяет расширить количество выявляемых пиков, содержащих ССТФ, относительно PWM.

Такой результат можно объяснить наличием разных структурных типов ССТФ для FOXA2, т.е. их гетерогенностью. Это хорошо согласуется с экспериментальными данными, полученными для ряда других TF, включая представителей семейства FOX. Так, было показано, что TF NOXB13 и FOXC2 способны связываться с одинаковой аффинностью с совершенно отличными последовательностями CAATAAA/TCGTAAA (Morgunova et al., 2018) и GTAAACA/ACAAATA (Chen et al., 2019) соответственно. Недавно обнаружено, что TF FOXN3 может связываться с двумя принципиально различными типами ССТФ, которые имеют разную длину (Rogers et al., 2019). Помимо этого, небольшие изменения в структуре ССТФ зависят от кооперативного взаимодействия между TF (Morgunova, Taipale, 2017). Очевидно, что FOXA2 также связывается с разными структурными типами СС.

Чтобы учесть все варианты ССТФ, одной модели PWM для распознавания сайта может быть уже недостаточно. Эта проблема частично решается использованием нескольких PWM (Bi et al., 2011; Mitra et al., 2018) или альтернативных моделей (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). Однако ранее альтернативные

модели обычно сравнивали с PWM только по точности поиска ССТФ (Siebert, Söding, 2016) либо по количеству распознанных сайтов (Samee et al., 2019). В настоящей работе мы не только сравнили точность и количество распознаваемых пиков, но и оценили, сколько каждая модель привносит своих пиков с ССТФ, совместный вклад моделей в поиск ССТФ, а также как соотносятся между собой пики с ССТФ от разных моделей. Результаты по оценке точности (см. рис. 7, а) показали, что на данных по FOXA2 модель InMoDe имеет самую низкую точность относительно других моделей, а модели VaMM, diPWM и PWM сопоставимы между собой. С точки зрения расширения общей доли пиков с ССТФ в рассмотренном наборе данных лучше всего себя показала модель VaMM, поскольку она находит самую крупную долю пиков с ССТФ, которые не выявляются другими моделями. Тем не менее все альтернативные модели (diPWM, VaMM и InMoDe) позволяют расширить набор распознанных ССТФ относительно PWM, а PWM вносит свой независимый вклад в общее количество пиков с распознанными ССТФ.

Заключение

Нами разработан конвейер программ MultiDeNA, который позволяет единообразно обрабатывать данные ChIP-seq с использованием разных моделей поиска ССТФ. В настоящее время с его помощью можно строить модели PWM, diPWM, InMoDe, VaMM. MultiDeNA включает в себя этапы подготовки данных, построения моделей, оценки точности моделей, сканирования пиков, сочетания результатов и их анализа. Разработанным конвейером программ был обработан набор данных из базы ReMap, включающий 22 ChIP-seq эксперимента для TF FOXA2. Мы показали, что совместное применение разных моделей позволяет увеличить общее количество распознанных пиков до 73.6 %, относительно модели PWM количество распознанных пиков увеличилось на 26.3 %. Разные модели распознают совпадающие ССТФ в значительной доле пиков, тем самым выявляя наиболее общий структурный тип ССТФ в этих пиках. Также каждая модель находила ССТФ, которые не выявлялись другими моделями. Лучше всего себя показала модель VaMM с 4.15 % пиков, содержащих только ее сайты, против 1.08, 0.49, 1.73 % для PWM, diPWM и InMoDe соответственно. Исходя из результатов можно предположить, что гетерогенность сайтов для FOXA2 не учитывается полностью только одной из моделей. Хуже всего себя в этом плане проявила модель diPWM, которая распознает ССТФ только в 46.4 % пиков. Оптимальной моделью для сайтов FOXA2 оказалась модель VaMM, которая нашла ССТФ в 65.8 % пиков. На основании полученных данных мы предположили, что модель VaMM может лучше описывать ССТФ для FOXA2.

Список литературы / References

- Bailey T.L., Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proc. Int. Conf. Intell. Syst. Mol. Biol. 1994;2:28-36. DOI citeulike-article-id:878292. PMID 7584402.
- Benos P.V., Bulyk M.L., Stormo G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* 2002;30(20):4442-4451. DOI 10.1093/nar/gkf578.

- Bi Y., Kim H., Gupta R., Davuluri R.V. Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One*. 2011;6(9):e24210. DOI 10.1371/journal.pone.0024210.
- Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002;30(5):1255-1261. DOI 10.1093/nar/30.5.1255.
- Chen X., Wei H., Li J., Liang X., Dai S., Jiang L., Guo M., Qu L., Chen Z., Chen L., Chen Y. Structural basis for DNA recognition by FOXC2. *Nucleic Acids Res.* 2019;47(7):3752-3764. DOI 10.1093/nar/gkz077.
- Chèneby J., Ménétrier Z., Mestdagh M., Rosnet T., Douida A., Rhaloussi W., Bergon A., Lopez F., Ballester B. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 2020;48(D1):D180-D188. DOI 10.1093/nar/gkz945.
- Eggeling R., Grosse I., Grau J. InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics.* 2017;33(4):580-582. DOI 10.1093/bioinformatics/btw689.
- Farnham P.J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 2009;10(9):605-616. DOI 10.1038/nrg2636.
- Furey T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* 2012;13(12):840-852. DOI 10.1038/nrg3306.
- Gheorghe M., Sandve G.K., Khan A., Chèneby J., Ballester B., Mathelier A. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* 2019;47(4):e21. DOI 10.1093/nar/gky1210.
- Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol.* 2007;8(2):R24. DOI 10.1186/gb-2007-8-2-r24.
- Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 2010;38(4):576-589. DOI 10.1016/j.molcel.2010.05.004.
- Ignatieva E.V., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina T.V., Merkulova T.I. Comparison of the results of search for the SF-1 binding sites in the promoter regions of the steroidogenic genes, using the SiteGA and SITECON methods. In: Proc. Fourth Int. Conf. Bioinform. Genome Regul. Struct. (BGRS). 2004;1:69-72.
- Iwafuchi-Doi M. The mechanistic basis for chromatin regulation by pioneer transcription factors. *WIREs Syst. Biol. Med.* 2019;11(1):e1427. DOI 10.1002/wsbm.1427.
- Keilwagen J., Grau J. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* 2015;43(18):e119. DOI 10.1093/nar/gkv577.
- Kiesel A., Roth C., Ge W., Wess M., Meier M., Söding J. The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.* 2018;46(W1):W215-W220. DOI 10.1093/nar/gky431.
- Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics.* 2010;26(20):2622-2623. DOI 10.1093/bioinformatics/btq488.
- Kulakovskiy I., Levitsky V., Oshchepkov D., Bryzgalov L., Vorontsov I., Makeev V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.* 2013;11(01):1340004. DOI 10.1142/S0219720013400040.
- Kulakovskiy I.V., Makeev V.J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics (Oxf.)*. 2009;54(6):667-674. DOI 10.1134/S0006350909060013.
- Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.
- Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell.* 2018;172(4):650-665. DOI 10.1016/j.cell.2018.01.029.
- Latchman D.S. Transcription factors: bound to activate or repress. *Trends Biochem. Sci.* 2001;26(4):211-213. DOI 10.1016/S0968-0004(01)01812-6.
- Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C.T. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinform.* 2007;8(1):1-20. DOI 10.1186/1471-2105-8-481.
- Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.
- Levitsky V.G., Oshchepkov D.Y., Klimova N.V., Ignatieva E.V., Vasiliev G.V., Merkulov V.M., Merkulova T.I. Hidden heterogeneity of transcription factor binding sites: a case study of SF-1. *Comput. Biol. Chem.* 2016;64:19-32. DOI 10.1016/j.compbiolchem.2016.04.008.
- Lloyd S.M., Bao X. Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr. Protoc. Cell Biol.* 2019;84(1):e89. DOI 10.1002/cpcb.89.
- Machanick P., Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics.* 2011;27(12):1696-1697. DOI 10.1093/bioinformatics/btr189.
- Mathelier A., Wasserman W.W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 2013;9(9):e1003214. DOI 10.1371/journal.pcbi.1003214.
- McClish D.K. Analyzing a portion of the ROC curve. *Med. Decis. Mak.* 1989;9(3):190-195. DOI 10.1177/0272989X8900900307.
- Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol.* 2018;14(4):1-20. DOI 10.1371/journal.pcbi.1006090.
- Morgunova E., Taipale J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* 2017;47:1-8. DOI 10.1016/j.sbi.2017.03.006.
- Morgunova E., Yin Y., Das P.K., Jolma A., Zhu F., Popov A., Xu Y., Nilsson L., Taipale J. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife.* 2018;7:1-21. DOI 10.7554/eLife.32963.
- Park P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 2009;10(10):669-680. DOI 10.1038/nrg2641.
- Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842. DOI 10.1093/bioinformatics/btq033.
- Rogers J.M., Waters C.T., Seegar T.C.M., Jarrett S.M., Hallworth A.N., Blacklow S.C., Bulyk M.L. Bispecific forkhead transcription factor FoxN3 recognizes two distinct motifs with different DNA shapes. *Mol. Cell.* 2019;74(2):245-253. DOI 10.1016/j.molcel.2019.01.019.
- Samee M.A.H., Bruneau B.G., Pollard K.S. A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.* 2019;8(1):27-42. DOI 10.1016/j.cels.2018.12.001.
- Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.* 2016;44(13):6055-6069. DOI 10.1093/nar/gkw521.
- Srivastava D., Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta – Gene Regul. Mech.* 2020;1863(6):e194443. DOI 10.1016/j.bbagr.2019.194443.

- Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics*. 2000;16(1):16-23. DOI 10.1093/bioinformatics/16.1.16.
- Wallerman O., Motallebipour M., Enroth S., Patra K., Bysani M.S.R., Komorowski J., Wadelius C. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res.* 2009;37(22):7498-7508. DOI 10.1093/nar/gkp823.
- Wederell E.D., Bilenky M., Cullum R., Thiessen N., Daggpinar M., Delaney A., Varhol R., Zhao Y., Zeng T., Bernier B., Ingham M., Hirst M., Robertson G., Marra M.A., Jones S., Hoodless P.A. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* 2008;36(14):4549-4564. DOI 10.1093/nar/gkn382.
- Worsley Hunt R., Wasserman W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* 2014;15(7):412. DOI 10.1186/s13059-014-0412-4.
- Yang L., Zhou T., Dror I., Mathelier A., Wasserman W.W., Gordân R., Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* 2014;42(D1):D148-D155. DOI 10.1093/nar/gkt1087.
- Zhang M.O., Marr T.G. A weight array method for splicing signal analysis. *Bioinformatics*. 1993;9(5):499-509. DOI 10.1093/bioinformatics/9.5.499.
- Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137. DOI 10.1186/gb-2008-9-9-r137.

ORCID ID

A.V. Tsukanov orcid.org/0000-0002-5174-6609

V.G. Levitsky orcid.org/0000-0002-4905-3088

Благодарности. Работа поддержана Российским фондом фундаментальных исследований (№ 18-29-13040) и бюджетным проектом № 0259-2019-0008.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 10.10.2020. После доработки 10.01.2021. Принята к публикации 12.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

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

Е.В. Игнатьева^{1,2}✉, Е.А. Матросова^{1,2}

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

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ eignat@bionet.nsc.ru

Аннотация. Полногеномные и полноэкзомные технологии секвенирования играют важную роль в исследовании генетических аспектов патогенеза различных заболеваний. Широкое применение методов полногеномного и полноэкзомного анализа ассоциаций позволяет идентифицировать множество вариантов геномной изменчивости (ГИ), ассоциированных с заболеваниями. Эта информация накапливается в базах данных GWAS central, GWAS catalog, OMIM, ClinVar и др. Большинство вариантов, идентифицированных методикой полногеномного анализа ассоциаций, располагается в некодирующих областях генома человека. По данным проекта ENCODE, доля участков в геноме человека, потенциально задействованных в регуляции транскрипции, во много раз превышает долю кодирующих областей. Таким образом, геномная изменчивость в некодирующих областях генома может повышать предрасположенность к заболеваниям, нарушая функционирование различных регуляторных элементов (промоторов, энхансеров, участков, определяющих 3D структуру хроматина и т. д.). Однако идентификация механизмов влияния патогенных вариантов ГИ на риск развития заболеваний затруднена ввиду большого разнообразия регуляторных элементов. В обзоре рассмотрены молекулярно-генетические механизмы влияния патогенных вариантов ГИ на экспрессию генов. При этом внимание сосредоточено на транскрипционном уровне регуляции как ключевой стадии, запускающей последовательность этапов экспрессии любого гена. Пусковым событием, опосредующим влияние патогенного варианта ГИ на уровень экспрессии гена, может быть, например, изменение функциональной активности сайтов связывания транскрипционных факторов или уровня метилирования ДНК, что, в свою очередь, отражается на функциональной активности промоторов или энхансеров. Выявление регуляторных эффектов полиморфных локусов невозможно без тесной интеграции современных экспериментальных подходов с компьютерным анализом больших массивов генетических данных, получаемых на основе омиксных технологий. В обзоре кратко описаны наиболее известные открытые полногеномные информационные ресурсы, содержащие данные, полученные на основе омиксных технологий, в том числе: ресурсы, накапливающие сведения о состоянии хроматина и участках его связывания с транскрипционными факторами, выявленными с помощью технологии ChIP-seq; ресурсы по геномным локусам, для которых на основе данных ChIP-seq выявлено аллель-специфичное связывание с транскрипционными факторами; а также ресурсы, содержащие предсказанные *in silico* данные о потенциальном влиянии геномной изменчивости на сайты связывания транскрипционных факторов.

Ключевые слова: регуляция транскрипции; геномная изменчивость; патогенные геномные варианты; районы, регулирующие транскрипцию; сайты связывания транскрипционных факторов; геномные базы данных.

Для цитирования: Игнатьева Е.В., Матросова Е.А. Геномная изменчивость в регуляторных районах генов, ассоциированная с заболеваниями человека: механизмы влияния на транскрипцию генов и полногеномные информационные ресурсы, обеспечивающие исследование этих механизмов. *Вавиловский журнал генетики и селекции*. 2021;25(1): 18-29. DOI 10.18699/VJ21.003

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

E.V. Ignatieva^{1,2}✉, E.A. Matrosova^{1,2}

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

² Novosibirsk State University, Novosibirsk, Russia

✉ eignat@bionet.nsc.ru

Abstract. Whole genome and whole exome sequencing technologies play a very important role in the studies of the genetic aspects of the pathogenesis of various diseases. The ample use of genome-wide and exome-wide association study methodology (GWAS and EWAS) made it possible to identify a large number of genetic variants associated with diseases. This information is accumulated in the databases like GWAS central, GWAS catalog, OMIM, ClinVar, etc. Most of the vari-

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

Key words: transcription regulation; genetic variability; pathogenic genetic variants; transcription regulatory regions; transcription factor binding sites (TFBSs); genomic databases.

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

Введение

В настоящее время, во многом благодаря развитию технологии полногеномного и полноэкзомного анализа ассоциаций (ПГАА и ПЭАА), идентифицировано большое количество полиморфизмов, ассоциированных с заболеваниями. Так, ресурс GWAS central (<https://www.gwascentral.org/>) содержит информацию о 3.2 млн вариантов генетической изменчивости, ассоциированных с заболеваниями либо фенотипическими характеристиками (Beck et al., 2020). Сравнимые по объему массивы экспериментальных данных накоплены в ряде других баз по ассоциациям генотип-фенотип (GWAS catalog, OMIM, ClinVar, HGMD, PheGenI, EGA, GAD, dbGaP).

Несмотря на наличие значительных объемов экспериментальной информации о вариантах геномной изменчивости (ГИ), ассоциированных с заболеваниями, молекулярные механизмы, лежащие в основе этих ассоциаций, изучены крайне недостаточно. Это обусловлено тем, что лишь малая доля патогенных вариантов ГИ находится в кодирующих областях генома человека, изменение нуклеотидной последовательности которых нарушает строение и функцию белков. Огромная масса полиморфных локусов, связанных с заболеваниями, располагается в некодирующих участках генома (интронах, 5'- и 3'-фланкирующих районах генов, межгенных областях). Например, из общего числа вариантов, ассоциированных с заболеваниями по данным ПГАА, ~90 % локализованы вне кодирующих районов генов (Maurano et al., 2012; Farh et al., 2015).

Известно, что некодирующие области генома содержат участки, выполняющие широкий спектр регуляторных функций. Это промоторные районы, энхансеры, негативные регуляторные элементы, районы прикрепления к ядерному матриксу, районы, определяющие структуру топологически ассоциированных доменов хроматина (topologically associating domain, TAD) и другие особенности 3D укладки генома (Mathelier et al., 2015; Meddens et al., 2019; Ibrahim, Mundlos, 2020). В геноме человека

доля участков, потенциально вовлеченных в регуляцию транскрипции, чрезвычайно высока. По данным проекта ENCODE, участки хроматина, соответствующие пикам связывания с транскрипционными факторами (ТФ), выявленным методикой ChIP-seq, занимают ~8.1 % всей геномной ДНК (ENCODE Project Consortium, 2012), что заметно больше, чем доля кодирующих областей генома человека (~1.2 %). С учетом того, что в проекте ENCODE изучались не все известные ТФ и не все линии клеток, во взаимодействии с ТФ вовлечена заведомо большая доля геномной ДНК. Суммарная протяженность участков генома человека, которые имеют характеристики хроматина, свойственные энхансерам, также существенно превышает общий размер кодирующих областей: например, только в одном исследованном типе клеток (H1-ES) энхансерные участки занимают ~3.2 % (Roadmap Epigenomics Consortium et al., 2015).

Исследования, направленные на выявление механизмов влияния патогенных вариантов ГИ на предрасположенность к заболеваниям, ведутся очень активно, что нашло отражение в целом ряде публикаций обзорного характера (Mathelier et al., 2015; Merkulov et al., 2018; Smith et al., 2018; Wang et al., 2019; Vohra et al., 2020). Наиболее обсуждаемое проявление эффекта патогенных вариантов ГИ – изменение функциональных характеристик сайтов связывания транскрипционных факторов (ССТФ) (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Mathelier et al., 2015; Gorbacheva et al., 2018). Показано также, что полиморфные локусы могут быть ассоциированы с особенностями паттернов метилирования ДНК (Howard et al., 2014; Kumar D. et al., 2017; Rahbar et al., 2018; Schmitz et al., 2019) и модификаций гистоновых белков (Kilpinen et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019), формированием петель хроматина (Visser et al., 2015; Zhang et al., 2018) и, как одним из проявлений этого процесса, – с изменениями структуры TAD (Cong et al., 2019; Mei et al., 2019). Примеры подобных эффектов будут рассмотрены ниже (табл. 1).

Таблица 1. Примеры ассоциаций полиморфных локусов с патологиями и механизмы их влияния на уровень экспрессии генов

Заболевание/ патология	Полиморфный локус	Локализация	Механизм	Литературный источник
Атопическая астма	rs928413 A→G	Промоторный район гена <i>IL33</i>	Появление сайта связывания CREB1 приводит к повышению уровня экспрессии гена <i>IL33</i>	Gorbacheva et al., 2018
Ожирение	rs1421085 T→C	Интрон гена <i>FTO</i> содержит регуляторный район генов <i>IRX3</i> и <i>IRX5</i> (удаленных от него на 517 и 1164 тыс. нуклеотидов)	Нарушение сайта связывания фактора-репрессора ARID5B приводит к повышению уровня экспрессии генов <i>IRX3</i> и <i>IRX5</i>	Claussnitzer et al., 2015
Рак поджелудочной железы	rs2001389 A→G	Участок, ограничивающий TAD, на 10-й хромосоме человека	Нарушение сайта связывания CTCF приводит к изменению 3D структуры хроматина, что вызывает снижение уровня экспрессии гена-супрессора опухолевого роста <i>MFSD13A</i>	Mei et al., 2019
Нарушения липидного метаболизма	rs174537 G→T	Энхансер генов <i>FADS1</i> и <i>FADS2</i>	Повышение уровня метилирования регуляторных районов генов <i>FADS1</i> и <i>FADS2</i> приводит к снижению уровня экспрессии генов <i>FADS1</i> и <i>FADS2</i>	Howard et al., 2014
Атопический дерматит	rs612529 T→C	Промоторный район гена <i>VSTM1</i>	Нарушение сайта связывания фактора PU.1, иницирующего деметилирование ДНК (посредством привлечения деметилаз), приводит к повышению уровня метилирования промоторного участка гена <i>VSTM1</i> , что вызывает снижение уровня экспрессии гена <i>VSTM1</i>	Kumar D. et al., 2017
Синдром ломкой X-хромосомы	Увеличение числа тринуклеотидных повторов CGG с 5–55 (норма) до 100 и более	5'-НТФ гена <i>FMR1</i>	Нарушение структуры TAD, включающего ген <i>FMR1</i> . Граница между двумя TAD смещается в 3'-направлении относительно <i>FMR1</i> , ввиду чего ген <i>FMR1</i> оказался включенным в «чужой» TAD. В этом случае наблюдается гиперметилирование CpG островков в районе промотора, что приводит к снижению уровня экспрессии гена <i>FMR1</i>	Sun et al., 2018
Ревматоидный артрит, диабет 2-го типа	rs7873784 G→C	3'-НТФ гена <i>TLR4</i>	Появление сайта связывания фактора PU.1 повышает активность энхансера, расположенного в 3'-НТФ гена <i>TLR4</i> . Повышается активность промотора гена <i>TLR4</i> , что приводит к активации его экспрессии	Korneev et al., 2020
Рак молочной железы	rs4321755 C→T	Энхансер генов <i>MRPS30</i> и <i>RP11-53019.1</i>	Появление сайта связывания фактора GATA3 повышает активность энхансера. Усиливаются контакты энхансера с двунаправленным промотором генов <i>MRPS30</i> и <i>RP11-53019.1</i> , что приводит к активации их экспрессии	Zhang et al., 2018

Влияние геномной изменчивости на сайты связывания транскрипционных факторов

Ключевую роль в процессе регуляции транскрипции играют транскрипционные факторы – белки, способные специфически взаимодействовать с ДНК регуляторных районов генов и иницирующие формирование транскрипционных комплексов. В геноме человека содержится более 1500 генов, кодирующих ТФ (Wingender et al., 2013). Сайты связывания ТФ, как правило, имеют протяженность 10–25 нуклеотидов (Levitsky et al., 2014; Kulakovskiy et al., 2018).

Замены нуклеотидов, а также короткие инсерции/делеции в полиморфных локусах могут приводить к повреждению (разрушению) ССТФ либо их возникновению *de novo* (см. табл. 1), и это, в свою очередь, может оказывать как негативный, так и позитивный эффект на уровень транс-

крипции генов (Chen L. et al., 2013; Gorbacheva et al., 2018). Такие варианты ГИ (и соответствующие полиморфные локусы), влияющие на транскрипционную активность генов, принято называть регуляторными (Kumar S. et al., 2017; Guo, Wang, 2018; Merkulov et al., 2018).

Патологическим (т. е. ассоциированным с заболеванием) может оказаться как аллельный вариант последовательности ДНК, содержащий нарушенный ССТФ (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Kumar D. et al., 2017; Mei et al., 2019), так и аллельный вариант, в котором идентифицируется возникновение сайта *de novo* (Gorbacheva et al., 2018; Zhang et al., 2018; Korneev et al., 2020) (см. табл. 1).

Патологические варианты ГИ, влияющие на функциональность ССТФ, могут располагаться не только в промоторных участках, но также в удаленных регуляторных

районах: энхансерах (Lewinsky et al., 2005; Zhang et al., 2018; Meddens et al., 2019), регуляторных районах с репрессорной функцией (Claussnitzer et al., 2015) и участках, ограничивающих TAD (Mei et al., 2019) (см. табл. 1). Так, например, ассоциированная с ожирением замена Т→С в полиморфном локусе rs1421085 нарушает функционирование негативного регуляторного района генов *IRX3* и *IRX5* (Claussnitzer et al., 2015). Локус rs1421085 расположен в интроне гена *FTO* (рис. 1) и находится на значительном расстоянии от стартов транскрипции генов *IRX3* и *IRX5* (520 и 1164 тыс. нуклеотидов). В норме участок ДНК, содержащий rs1421085 (аллель Т), связывается с транскрипционным фактором-репрессором ARID5B, что способствует снижению транскрипционной активности генов *IRX3* и *IRX5*. У носителей мутантного варианта последовательности ДНК (аллель С) сайт связывания фактора-репрессора ARID5B оказывается нарушенным, что приводит к чрезмерно высокой экспрессии генов *IRX3* и *IRX5* и ускоренному росту клеток жировой ткани (Claussnitzer et al., 2015).

Возможны ситуации, когда замена нуклеотида в полиморфном локусе нарушает ССТФ, и это, в свою очередь, отражается на функционировании TAD (см. табл. 1). Такой эффект выявлен в случае локуса rs2001389 (А→G), ассоциированного с риском рака поджелудочной железы (рис. 2). Локус rs2001389 расположен в участке, определяющем структуру петель хроматина в пределах TAD, содержащего 91 ген и образованного пространственно сближенными участками хроматина (Mei et al., 2019). Участок ДНК, содержащий патологический аллель G, характеризуется пониженной способностью к связыванию с транскрипционным фактором CTCF, который в данном случае выполняет функцию структурного белка хроматина. В норме связывание с фактором CTCF обеспечивает функционирование одного из участков, определяющих структуру петель хроматина в пределах рассматриваемого TAD. Нарушение связывания с фактором CTCF влечет за собой изменение 3D структуры хроматина, что нарушает экспрессию генов, содержащихся в данном TAD. При этом в наибольшей степени оказывается подавленной экспрессия гена-супрессора опухолевого роста *MFSD13A*.

Влияние геномной изменчивости на метилирование ДНК и транскрипционная активность генов

Метилирование ДНК – это модификация, которая не изменяет нуклеотидной последовательности и заключается в присоединении метильной группы к пятому атому углерода цитозина (Angeloni, Bogdanovic, 2019). Повышение уровня метилирования ДНК, как правило, приводит к долгосрочной инактивации экспрессии генов, лежащих в метилированном участке, так как, согласно общепризнанной концепции, метилирование участка ДНК способствует посадке на этот участок белковых комплексов, включающих гистон деацетилазу (HDAC) (Jones et al., 1998; Nan et al., 1998). Метилирование может также препятствовать взаимодействию транскрипционных факторов с ДНК: известно, что такой чувствительностью к метилированию обладают факторы CTCF и факторы из семейства ETS (Wang et al., 2019). Другой транскрипци-

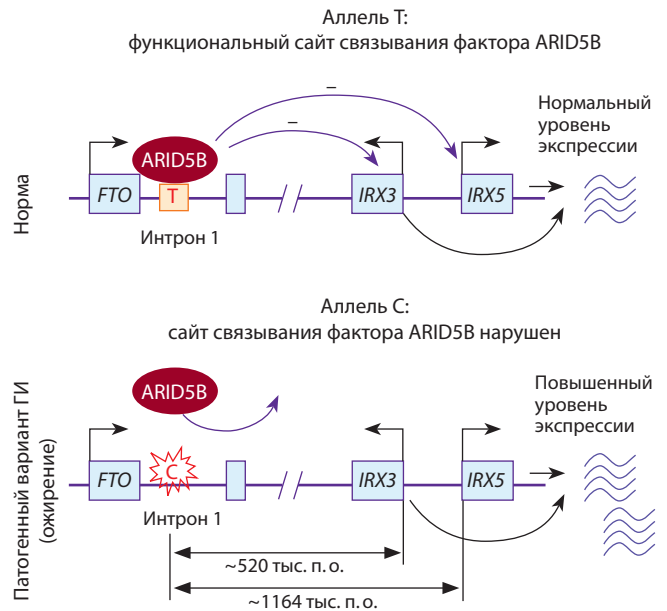


Рис. 1. Нарушение сайта связывания, вызванное заменой нуклеотида Т→С в локусе rs1421085, препятствует взаимодействию фактора-репрессора ARID5B с регуляторным районом генов *IRX3* и *IRX5*, в результате чего уровень экспрессии *IRX3* и *IRX5* повышается.

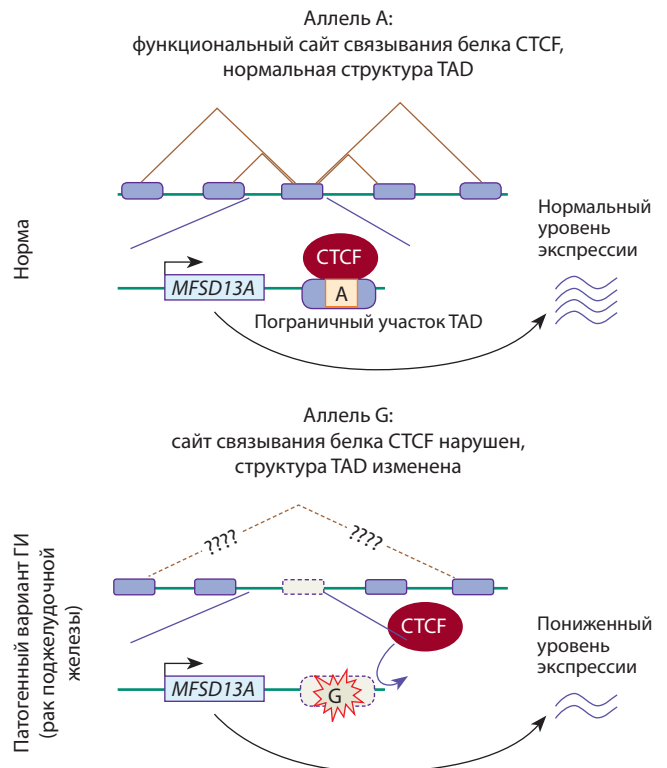


Рис. 2. Нарушение сайта связывания белка CTCF, вызванное заменой нуклеотида rs2001389, приводит к исчезновению одного из пограничных участков, определяющих структуру TAD, вследствие чего происходит снижение уровня экспрессии гена-супрессора опухолевого роста *MFSD13A*.

Контакты между участками хроматина в пределах TAD показаны коричневыми линиями. Знаки вопроса на нижнем рисунке обозначают отсутствие точных сведений о структуре TAD.

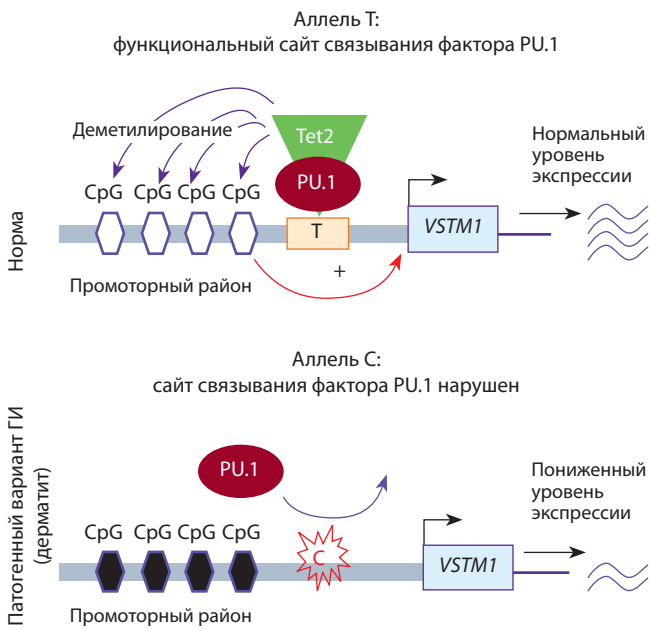


Рис. 3. Нарушение сайта связывания PU.1, вызванное заменой нуклеотида Т→С (rs612529), снижает активность деметилаз (например, Tet2), поддерживающих промоторный участок гена *VSTM1* в активном состоянии, в связи с чем экспрессия *VSTM1* снижается.

онный фактор, ZFP57, напротив, взаимодействует только с метилированной ДНК (Quenneville et al., 2011). Таким образом, метилирование цитозина может активировать различные механизмы регуляции транскрипции генов, и не всегда повышение уровня метилирования регуляторного участка ДНК сопряжено со снижением экспрессии соответствующего гена (Izzi et al., 2016; Wang et al., 2019).

Геномная изменчивость оказывает существенное влияние на метилирование участков ДНК, обладающих регуляторным потенциалом. Так, в ходе полногеномного анализа паттернов метилирования геномов 24 жителей острова Норфолк (Benton et al., 2019) был выявлен 12761 район, содержащий не менее двух CpG-динуклеотидов и имеющий аллель-специфичный уровень метилирования. В большинстве случаев (98 %) расположение районов совпадало с позициями однонуклеотидных замен, представленных в dbSNP (Benton et al., 2019).

В этой же работе (Benton et al., 2019) был проведен анализ совместного расположения районов аллель-специфичного метилирования и набора полиморфных локусов из базы данных GWAS catalogue, ассоциированных с заболеваниями человека. Оказалось, что полиморфные локусы, ассоциированные с заболеваниями, в два раза чаще перекрываются с районами аллель-специфичного метилирования, чем можно было ожидать по случайным причинам. Это означает, что изменение уровня метилирования, обусловленное ГИ, является одним из факторов, повышающих риск развития заболеваний.

В качестве примера рассмотрим полиморфный локус rs174537 (G→T), расположенный в энхансере генов *FADS1* и *FADS2*, кодирующих десатуразы жирных кислот 1 и 2. Вариант Т локуса rs174537 ассоциирован с повышенным риском патологических нарушений липидного метаболиз-

ма (см. табл. 1). Было показано, что носители аллеля Т в локусе rs174537 имели более высокий уровень метилирования регуляторных районов генов *FADS1* и *FADS2* в печени (Howard et al., 2014), что приводило к подавлению транскрипционной активности генов *FADS1* и *FADS2*.

Возможны ситуации, когда в одном из аллельных вариантов происходит деметилирование ДНК, инициированное связыванием с транскрипционным фактором (см. табл. 1). Такой механизм выявлен, например, при исследовании полиморфного локуса rs612529 Т→С. Локус расположен в промоторном районе гена *VSTM1* (рис. 3), низкая экспрессия которого в моноцитах провоцирует развитие дерматита. В этом типе клеток промоторный участок, содержащий протективный вариант Т, более активно взаимодействует с транскрипционным фактором PU.1, инициирующим деметилирование ДНК посредством привлечения деметилаз (например, Tet2). Как следствие, у носителей аллеля Т промоторный участок гена *VSTM1* оказывается полностью деметилированным, и наблюдается активная экспрессия данного гена. У носителей патогенного варианта С взаимодействие фактора PU.1 с ДНК нарушается, в результате чего промоторный участок более сильно метилирован, что сопровождается снижением экспрессии гена *VSTM1* (Kumar D. et al., 2017).

Влияние геномной изменчивости на состояние хроматина и его пространственную структуру

Присутствие патогенных вариантов ГИ отражается и на состоянии хроматина (Kilpinen et al., 2013). Известны случаи, когда наличие патогенного варианта ГИ сопровождалось изменением паттернов модификации гистонов и появлением (либо исчезновением) участков гиперчувствительности к ДНКазе I типа (McVicker et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019). В этих случаях были выявлены аллель-специфичные контакты между промоторами и энхансерами, количество которых коррелировало с активностью энхансерных районов.

Также известны ситуации, когда структурные варианты генома (инсерции, делеции, дупликации, инверсии, транслокации длиной более 50 нуклеотидов) приводят к изменению пространственной организации хроматина, нарушая тем самым экспрессию генов, связанных с патологическими процессами (Sun et al., 2018; Ibrahim, Mundlos, 2020). Например, ассоциированная с синдромом ломкой Х-хромосомы экспансия тринуклеотидных повторов CGG в 5'-нетранслируемом районе (НТР) гена *FMRI* нарушает структуру TAD, включающего *FMRI* (рис. 4, см. табл. 1). В норме *FMRI* находится очень близко к 5'-пограничной области TAD (на рис. 4 этот домен обозначен как TAD1), а участок ДНК, соответствующий 5'-пограничной области TAD1, гипометилирован и взаимодействует с фактором CTCF. У индивидов, имеющих повышенное количество триплетных повторов CGG (более 100), данный участок ДНК перестает выполнять барьерную функцию (он гиперметилирован и не связывается с фактором CTCF). Нарушение барьерной функции 5'-пограничной области TAD1 приводит к исчезновению TAD1 и расширению границ другого TAD (см. TAD2 на рис. 4). Ген *FMRI* оказывается включенным в чужеродный ему TAD2. В этом случае

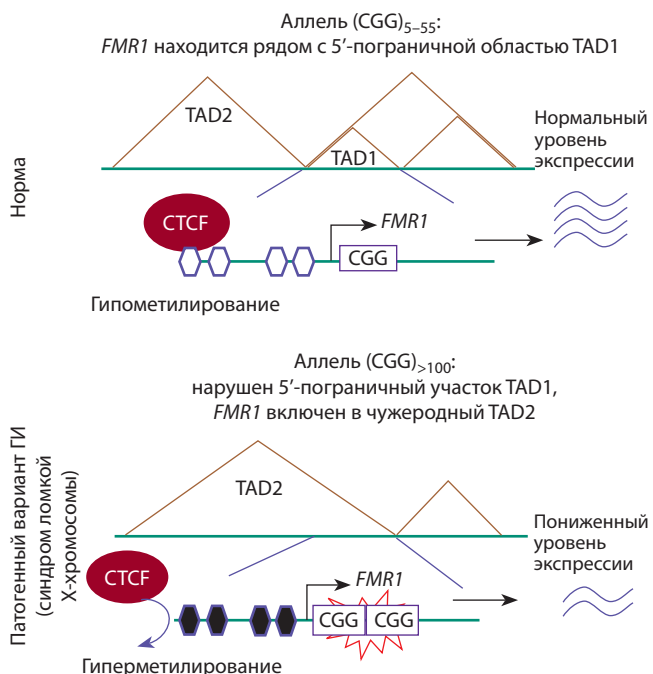


Рис. 4. При увеличении количества триплетных повторов CGG в 5'-нетранслируемом районе гена *FMR1* происходит гиперметилирование участка ДНК, соответствующего пограничной области TAD1. Это приводит к нарушению связывания факторов CTCF и нарушению барьерной функции пограничного участка.

Коричневыми линиями показаны контакты между участками хроматина в пределах TAD.

промоторный район гена *FMR1* также оказывается гиперметилированным, а экспрессия гена – резко сниженной (Park et al., 2015; Sun et al., 2018).

Для исследования молекулярно-генетических механизмов влияния геномной изменчивости на 3D структуру хроматина необходимо реконструировать пространственную укладку генома. В 3D структуре геномов выявлены такие базовые уровни организации, как: регуляторные петли ДНК со сближенными промоторами и энхансерами; топологически ассоциированные домены, внутри которых участки ДНК контактируют друг с другом чаще, чем с соседними доменами; А и В компартменты, соответствующие транскрипционно-активному и конденсированному хроматину; и наконец, хромосомные территории (Fishman et al., 2018; Hansen et al., 2018). Нарушение 3D контактов между промоторами и энхансерами в пределах TAD, вызванное, например, хромосомными перестройками, может существенно повлиять на транскрипционную активность гена, приводя к развитию патологий (Lupiáñez et al., 2015).

В Институте цитологии и генетики СО РАН разработан экспериментально-компьютерный подход к предсказанию физических контактов между промоторами и энхансерами в 3D структуре хроматина (Fishman et al., 2018; Belokopytova et al., 2020; Belokopytova, Fishman, 2021). Подход основан на использовании следующей информации: 1) тип клеток; 2) клеточно-специфическая локализация энхансеров в линейном геноме (из базы данных ENCODE); 3) транскрипционная активность промоторов (из экспериментов RNA-seq); 4) границы экструзии пе-

тель хроматина (из экспериментов ChIP-seq с фактором CTCF); 5) ориентация мотивов связывания CTCF-белков (из анализа геномов); 6) А либо В компартмент хроматина (по данным экспериментов Hi-C). Анализ этих данных с помощью оригинальной программы 3DPredictor (Belokopytova et al., 2020), разработанной на основе алгоритмов машинного обучения, позволяет предсказывать частоты физических контактов между промоторами и энхансерами в 3D структуре генома с точностью, превышающей точность других известных методов предсказания.

С помощью программы 3DPredictor была проанализирована 3D структура генома у мышей *DelB/DelB*, гомозиготных по делеции геномного участка длиной 1.5 Мб, содержащего ген *Epha4*. Такая делеция сопровождается появлением добавочных контактов между геном *Pax3* и энхансерным районом гена *Epha4*, что нарушает экспрессию гена *Pax3* и приводит к брахидактилии. Мыши с генотипом *DelB/DelB* являются генетической моделью патологии человека, выражающейся в нарушениях формирования конечностей (Lupiáñez et al., 2015). Тестирование 3DPredictor на этом модельном объекте продемонстрировало высокую эффективность программы: у мышей с нарушенным генотипом были предсказаны новые добавочные контакты между генами и удаленными регуляторными элементами (Belokopytova et al., 2020), и предсказания хорошо соответствовали экспериментальным данным.

Геномная изменчивость: комплексный анализ больших гетерогенных генетических данных

Как отмечено выше, многие варианты ГИ, ассоциированные с заболеваниями, находятся на значительном расстоянии от кодирующих областей генов (ENCODE Project Consortium, 2012; Maurano et al., 2012). Для идентификации молекулярно-генетических механизмов влияния таких вариантов ГИ на предрасположенность к заболеваниям необходимы дополнительные исследования. Их целью является выяснение регуляторной роли вариантов ГИ. Типичный пример – работа (Zhang et al., 2018), позволившая найти функционально значимый регуляторный вариант rs4321755, ассоциированный с риском рака молочной железы. Лocus rs4321755 располагается в удаленном энхансере, регулирующем экспрессию генов *MRPS30* и *RP11-53019.1* (см. табл. 1). Оказалось, что при наличии патогенного варианта Т в локусе rs4321755 формируется новый сайт связывания фактора GATA3. Транскрипционный фактор GATA3 повышает функциональную активность энхансера, что проявляется в более интенсивных контактах энхансера с двунаправленным промотором генов *MRPS30* и *RP11-53019.1* и активации их экспрессии. Чтобы выявить этот функционально значимый регуляторный вариант ГИ, авторы разработали интегрированный экспериментально-компьютерный метод, основанный на комплексном анализе больших гетерогенных генетических данных, включая данные об аллель-специфичной экспрессии генов, полученные на основе технологии RNA-seq в сочетании с данными о гаплотипах; о локусах количественных характеристик экспрессии (eQTL); об участках генома, чувствительных к ДНКазе I; о локализации ChIP-seq пиков из баз ENCODE и GEO; о локализации

Таблица 2. Информационные ресурсы по геномным данным, полученным на основе современных высокопроизводительных экспериментальных методов

Ресурс/база данных	URL	Характеристика ресурса
Аннотация генома человека		
GENCODE*	https://www.encodegenes.org/	Аннотация генома человека, составленная на основе слияния результатов, полученных с помощью ручной аннотации, с результатами компьютерной аннотации генов
Геномная изменчивость в популяциях человека		
НарМaп (Haplotype Map)	https://www.genome.gov/10001688/international-hapmap-project ftp://ftp.ncbi.nlm.nih.gov/hapmap/	Гаплотипы и гапоблоки генома человека, а также маркирующие их репрезентативные полиморфные локусы
1000 Genomes Project (1KGP)	https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/	Генетические варианты (однонуклеотидные полиморфизмы, инсерции/делеции, структурные варианты) и генотипы, выявленные у индивидов из 26 популяций
International Genome Sample Resource (IGSR)	https://www.internationalgenome.org	Объединение данных проекта «1000 геномов» с данными, полученными методикой RNA-Seq (проект GEUVADIS), и данными проекта ENCODE, полученными на линии клеток NA12878
dbSNP	https://www.ncbi.nlm.nih.gov/snp/	Однонуклеотидные генетические варианты, микросателлиты, инсерции и делеции в геномах различных видов организмов (включая человека). Локализация полиморфных локусов на хромосомах, популяционные частоты. dbSNP накапливает как данные массового анализа, полученные в геномных проектах, так и результаты исследования отдельных локусов, представленные в публикациях
Генетические варианты, ассоциированные с заболеваниями		
GWAS central (Genome-wide association studies central)	https://www.gwascentral.org/	Частоты аллелей и генотипов человека, а также их ассоциации с фенотипами (либо патологиями) из различных источников (публикаций и исследовательских проектов)
GWAS catalog (Genome-wide association studies catalog)	https://www.ebi.ac.uk/gwas/home	Ассоциации между полиморфными локусами и фенотипическими признаками, полученные с помощью методики ПГАА
OMIM (Online Mendelian Inheritance in Man)	https://www.ncbi.nlm.nih.gov/omim	Каталог, описывающий гены человека, их генетические варианты и генетически обусловленные заболевания и синдромы человека. Данные внесены в каталог командой экспертов на основе анализа научных публикаций
ClinVar (Clinical Variations)	https://www.ncbi.nlm.nih.gov/clinvar/	Ассоциации между генами и генетическими вариантами генома человека и фенотипическими признаками
HGMD (The Human Gene Mutation Database)	http://www.hgmd.cf.ac.uk/ac/index.php	Генетические нарушения, связанные с наследственными заболеваниями человека
PheGenI (The Phenotype-Genotype Integrator)	https://www.ncbi.nlm.nih.gov/gap/phegeni	Ресурс, интегрирующий данные из GWAS catalog с данными из нескольких баз, размещенных в NCBI, включая Gene, dbGaP, OMIM, eQTL и dbSNP
EGA (The European Genome-phenome Archive)	https://ega-archive.org/	Данные о связи генотипов и фенотипов, полученные различными экспериментальными методами (ПГАА, экзомное и полногеномное секвенирование, генотипирование, секвенирование геномов отдельных клеток)
dbGaP (The database of Genotypes and Phenotypes)	https://www.ncbi.nlm.nih.gov/gap/	Данные по ассоциациям между генотипами и фенотипами человека, полученные различными методами (ПГАА, экзомное секвенирование, генотипирование когорт, исследования на близнецах и т.д.)
Данные о состоянии хроматина		
ENCODE (The Encyclopedia of DNA Elements)	http://genome.ucsc.edu/ENCODE/ https://www.encodeproject.org/	Полногеномные профили модификации гистонов, метилирования ДНК, районы связывания с ТФ (по данным ChIP-seq), области контактов между удаленными участками ДНК, участки открытого хроматина, экспрессионные данные для более 300 типов клеток
NIH Roadmap Epigenomics Mapping Consortium	http://www.roadmapepigenomics.org/	Данные, полученные с помощью методик ChIP-seq, RNA-seq, бисульфитного секвенирования. Аннотация генома человека в соответствии с классификациями состояний хроматина (15, 18 и 25 типов хроматина)

Окончание табл. 2

Ресурс/база данных	URL	Характеристика ресурса
Локусы количественных характеристик экспрессии eQTL		
Genotype-Tissue Expression (GTEx) project	https://www.gtexportal.org/home/	Данные по экспрессии и eQTL в 54 типах клеток человека, имеющих здоровый фенотип
eQTL databases	https://www.hsph.harvard.edu/liming-liang/software/eqtl/	eQTL в лимфобластоидных линиях клеток
exSNP	http://www.exsnp.org/	eQTL и их связи с заболеваниями человека
eQTL Catalogue	https://www.ebi.ac.uk/eqtl/	Cis-eQTL и QTL, выявленные на основе анализа данных публикаций, а также из проекта GTEx
eQTL Browser	http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/	eQTL, выявленные на основе анализа данных из научных публикаций
Коллекции экспериментов ChIP-seq, направленных на идентификацию ССТФ		
Cistrome Data Browser	http://cistrome.org/db/#/	Данные экспериментов ChIP-seq, DNase-seq и ATAC-seq, выявляющих в геномах человека и мыши (1) участки хроматина, взаимодействующие с ТФ; (2) участки хроматина, доступные для действия эндонуклеазы; (3) участки, содержащие посттрансляционные модификации гистонов. Данным присвоены статусы согласно шести критериям качества
Gene Transcription Regulation Database (GTRD)	https://gtrd.biouml.org/#!	Коллекция экспериментов ChIP-seq, направленных на поиск сайтов связывания ТФ в геноме человека и мыши
ReMap (Global map of regulatory elements)	http://remap.univ-amu.fr/	Коллекция экспериментов ChIP-seq, ChIP-exo, DAP-seq из публичных ресурсов (GEO, ENCODE, ENA). Участки хроматина, контактирующие с ТФ, транскрипционными коактиваторами, хроматин-ремоделлирующими факторами
Аллель-специфичное связывание ТФ, выявленное на основе анализа данных экспериментов ChIP-seq в комбинации с данными о генотипах исследованных клеток		
AlleleDB	http://alleledb.gersteinlab.org/	Данные по аллель-специфичному связыванию ТФ, полученные на основе анализа экспериментов ChIP-seq для образцов от 383 индивидов, чьи геномы были секвенированы в ходе выполнения проекта «1000 геномов»
AlleleSeq	http://alleleseq.gersteinlab.org/	Данные по аллель-специфичному связыванию для шести ТФ (cFos, cMyc, JunD, Max, NfkB, CTCF), полученные при анализе данных ChIP-seq в лимфобластоидной клеточной линии GM12878
Потенциальные эффекты вариантов геномной изменчивости на ССТФ		
HaploReg	https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php	Аннотация вариантов ГИ. Приведены данные о состоянии хроматина, сцеплении, консервативности, перекрытии с регуляторными мотивами, eQTL
SNP2TFBS	http://ccg.vital-it.ch/snp2tfbs/	Варианты ГИ, идентифицированные проектом «1000 геномов», изменяющие сходство ССТФ с весовыми матрицами
rSNPBase	http://rsnp3.psych.ac.cn/index.do	Эффекты ОНП на ССТФ, регулируемые гены, регуляторные сети
rVarBase	http://rv.psych.ac.cn/	Эффекты вариантов ГИ (включая вариации числа копий) на потенциальные ССТФ, данные о состоянии хроматина и регулируемых генах
Информационные ресурсы широкого профиля		
UCSC Genome Browser	https://genome.ucsc.edu/	Интеграция на основе графического интерфейса данных о первичных последовательностях и аннотации геномов, а также характеристиках геномных районов (нуклеотидном составе, геномной изменчивости, состояниях хроматина, экспрессии, контактах между участками хроматина и т.д.). Программное средство UCSC table browser позволяет извлекать данные в текстовом виде
Ensembl Genome Browser	https://www.ensembl.org/index.html	Интеграция на основе графического интерфейса данных о первичных последовательностях и аннотации геномов, а также характеристиках геномных районов (нуклеотидном составе, геномной изменчивости, состояниях хроматина и т.д.). Программное средство BioMart data mining tool позволяет извлекать данные в текстовом виде
GEO (Gene Expression Omnibus)	https://www.ncbi.nlm.nih.gov/gds	Крупнейший репозиторий данных по функциональной геномике человека и других видов организмов, полученных на основе омиксных технологий (экспрессия, профили состояния хроматина, генотипирование и др.)

* Аннотация генома человека из базы GENCODE доступна также через UCSC Genome Browser (<https://genome.ucsc.edu/>) и Ensembl genome browser (<https://www.ensembl.org/index.html>).

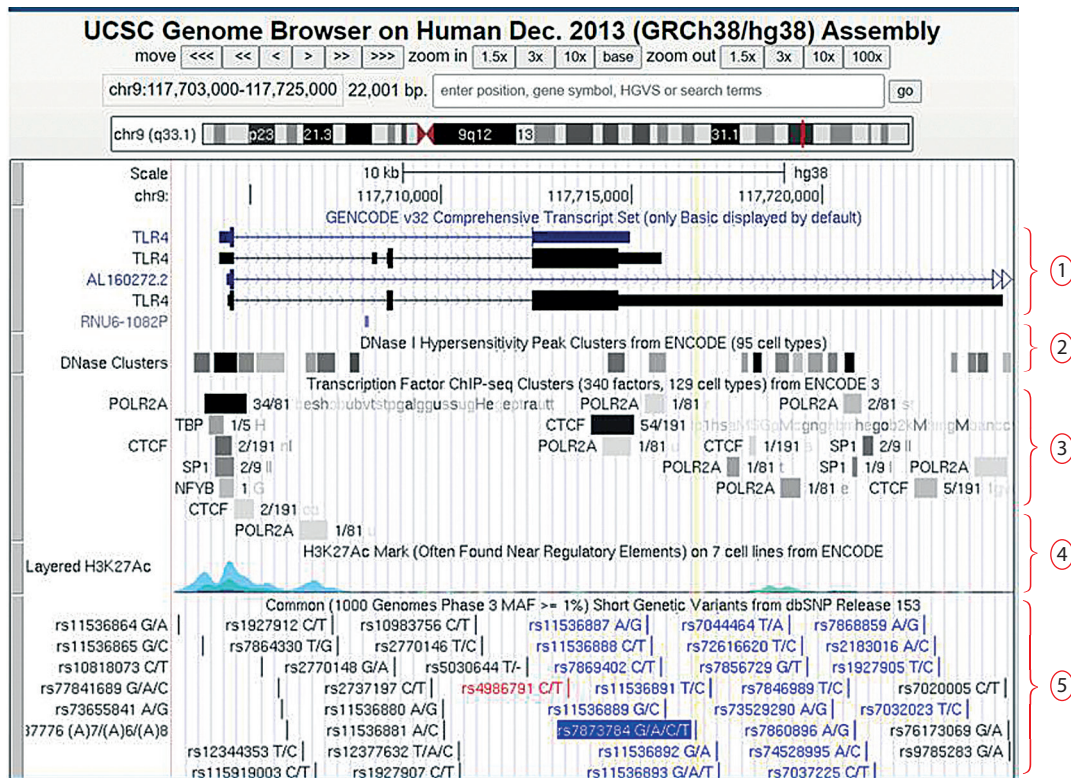


Рис. 5. Пример графического представления информации об участке 9-й хромосомы человека (хромосомные координаты chr9:117,703,000–117,725,000) в геномном браузере Университета г. Санта-Круз, США (UCSC Genome Browser, <https://genome.ucsc.edu/>).

1 – транскрипты гена *TLR4*, позиции которых приведены согласно данным базы GENCODE; 2 – участки гиперчувствительности к ДНКазе I типа; 3 – участки хроматина, для которых с помощью методики ChIP-seq (данные проекта ENCODE) показано взаимодействие с ТФ; 4 – участки хроматина, содержащие модифицированный гистоновый белок H3 (модификация H3K27Ac часто присутствует в участках, имеющих регуляторные функции); 5 – позиции вариантов геномной изменчивости. Желтой вертикальной линией отмечено расположение варианта ГИ rs7873784, локализованного в 3'-HTP гена *TLR4* и ассоциированного с ревматоидным артритом и диабетом 2-го типа (см. табл. 1). По данным (Korneev et al., 2020), замена G→C в локусе rs7873784 приводит к возникновению сайта связывания транскрипционного фактора PU.1, что повышает активность энхансера, расположенного в 3'-HTP гена *TLR4*.

регуляторных мотивов, предсказанных компьютерными программами. Сходные сценарии интегрированных экспериментально-компьютерных исследований были реализованы и в других работах (Chen C.-Y. et al., 2014; Claussnitzer et al., 2015; Zhao et al., 2019; Li et al., 2020).

Проведение исследований подобного рода стало возможным благодаря развитию современных высокопроизводительных экспериментальных подходов, позволяющих получать различные типы данных в масштабе всего генома (параллельное высокопроизводительное секвенирование, методики ChIP-seq, 3C, Hi-C, ChIA-PET, футпринтинг ДНК с использованием ДНКазы I типа, бисульфитное секвенирование и т. д.), и наличию открытых информационных ресурсов, накапливающих подобные экспериментальные данные. В табл. 2 приведена краткая характеристика информационных ресурсов, содержащих геномные данные, полученные на основе омиксных технологий, и используемых для изучения механизмов повреждающего влияния ГИ на транскрипцию генов. Это данные по аннотации генома человека (GENCODE); о геномной изменчивости в популяциях человека (MapMap, 1000 Genomes Project, IGSR, dbSNP); данные о вариантах ГИ, ассоциированных с заболеваниями (GWAS central, GWAS catalog, ClinVar,

HGMD, OMIM и др.); о состоянии хроматина (ENCODE, NIH Roadmap Epigenomics Mapping Consortium); о локусах количественных характеристик экспрессии (GTEx project, eQTL databases, exSNP и др.); об экспериментах ChIP-seq, направленных на идентификацию ССТФ (Cistrome Data Browser, GTRD, ReMap); об аллель-специфичном связывании ТФ, выявленном на основе анализа данных экспериментов ChIP-seq в комбинации с данными о генотипах исследованных клеток (AlleleDB, AlleleSeq); об эффектах вариантов ГИ на ССТФ, предсказанных на основе компьютерного анализа (HaploReg, SNP2TFBS, rSNPBase, rVarBase).

К отдельной категории информационных ресурсов относятся геномный браузер Университета г. Санта-Круз, США (UCSC Genome Browser, <https://genome.ucsc.edu/>) и геномный браузер базы Ensembl совместного научного проекта Европейского института биоинформатики и Института Сенгера (Ensembl Genome Browser, <https://www.ensembl.org/index.html>). Они используются как средства интеграции данных о характеристиках геномных районов, полученных разными экспериментальными методами и из разных информационных источников (Lee et al., 2020; Yates et al., 2020). Веб-сайты этих браузеров обеспечивают

доступ к первичным последовательностям и аннотациям геномов многих видов организмов, включая позвоночных животных и ряд других модельных видов. Графические интерфейсы браузеров позволяют в интерактивном режиме получать масштабируемые карты геномных районов и отображать на картах различные характеристики (например, локализацию транскриптов, вариантов геномной изменчивости, участков хроматина, взаимодействующих с ТФ по данным методики ChIP-seq, участков гиперчувствительности к ДНКазе I типа и т. д.) (рис. 5).

Веб-сайты геномных браузеров UCSC Genome Browser и Ensembl Genome Browser оснащены программными средствами доступа к данным в текстовом виде: UCSC table browser (<https://genome.ucsc.edu/cgi-bin/hgTables>) и BioMart data mining tool (<https://www.ensembl.org/info/data/biomart/index.html>).

Информационные ресурсы по аллель-специфичному связыванию транскрипционных факторов и предсказанным *in silico* эффектам вариантов геномной изменчивости на ССТФ

Как отмечалось выше, достаточно часто влияние патогенных вариантов ГИ на экспрессию генов реализуется через изменение функциональной активности сайтов связывания транскрипционных факторов. В связи с этим чрезвычайно полезными могут оказаться информационные ресурсы, включающие полногеномные данные об аллель-специфичном связывании ТФ, идентифицированном с помощью методики ChIP-seq. Разработано несколько подходов к выявлению аллель-специфичного связывания ТФ (Rozowsky et al., 2011; Reddy et al., 2012; Waszak et al., 2014; Younesy et al., 2014). Эти подходы основаны на анализе данных экспериментов ChIP-seq в комбинации с данными секвенирования, позволяющими выявлять гетерозиготные позиции в геноме и генотип исследуемых клеток. Таким образом, для каждого обследованного типа клеток может быть выявлен свой набор геномных локусов, взаимодействующих с конкретным транскрипционным фактором аллель-специфичным образом. Например, в работе (Cavalli et al., 2016a) были проанализированы данные экспериментов ChIP-seq для 55 ТФ в клеточной линии HepG2 и 57 ТФ в линии HeLa-S3. В клетках HepG2 был найден 3001 локус ГИ, имеющий аллель-специфичные сигналы, а в клетках HeLa-S3 обнаружено 712 таких локусов. Авторы отмечают выраженный тканеспецифичный характер аллель-специфичного связывания ТФ: из всего набора выявленных локусов только 34 были обнаружены в обеих клеточных линиях (Cavalli et al., 2016a).

Данные об аллель-специфичном связывании ТФ представлены в информационных ресурсах: AlleleDB (<http://alleledb.gersteinlab.org/>) (Chen J. et al., 2016), AlleleSeq (<http://alleleseq.gersteinlab.org/>) (Rozowsky et al., 2011) (см. табл. 2), а также в виде дополнительных материалов к публикациям (Cavalli et al., 2016a, b, 2019; Shi et al., 2016).

Исследования, направленные на идентификацию аллель-специфичного связывания ТФ, позволили оценить количество генетических вариантов, влияющих на связывание конкретного транскрипционного фактора на ДНК в конкретном типе клеток. Среднее количество таких событий, зарегистрированных для отдельного транскрип-

ционного фактора, может составлять от 19 до 37 для клеток с нормальным кариотипом (GM12878, H1-hESC) и от 12 до 55 для клеток с раковым кариотипом (SK-N-SH, K562) (Cavalli et al., 2016a, b).

При построении гипотез о механизмах влияния генетических вариантов на риск развития патологий также могут быть использованы данные об эффектах вариантов ГИ на ССТФ, предсказанных *in silico* на основе компьютерных программ распознавания ССТФ. Подобные сведения содержатся в специализированных базах данных: HaploReg (<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>) (Ward, Kellis, 2012), SNP2TFBS (<http://ccg.vital-it.ch/snp2tfbs/>) (Kumar S. et al., 2017), rSNPBase (<http://rsnp3.psych.ac.cn/index.do>) (Guo, Wang, 2018), rVarBase (<http://rv.psych.ac.cn>) (см. табл. 2).

Заключение

Существенная доля патогенных генетических вариантов, ассоциированных с заболеваниями, локализована в некодирующих областях генома. Такие генетические варианты могут с большой долей вероятности нарушать функционирование регуляторных районов, контролирующих транскрипционную активность генов. Наглядным подтверждением этой возможности являются рассмотренные в нашем обзоре примеры механизмов влияния патогенных генетических вариантов на экспрессию генов. Исследования, позволившие идентифицировать такие механизмы, носят комплексный характер и основаны на анализе больших гетерогенных генетических данных. Имеющийся в настоящее время арсенал информационных ресурсов, содержащих омиксные данные, обеспечивает широкие возможности для подобных исследований. В будущем, с развитием экспериментальных технологий и биоинформатических методов анализа полученных с их помощью данных, а также с расширением спектра исследуемых типов клеток, эти возможности еще более возрастут.

Список литературы / References

- Angeloni A., Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem.* 2019;63(6):707-715. DOI 10.1042/EBC20190030.
- Beck T., Shorter T., Brookes A.J. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res.* 2020;48(D1):D933-D940. DOI 10.1093/nar/gkz895.
- Belokopytova P., Fishman V. Predicting genome architecture: challenges and solutions. *Front. Genet.* 2021. DOI 10.3389/fgene.2020.617202.
- Belokopytova P.S., Nuriddinov M.A., Mozheiko E.A., Fishman D., Fishman V. Quantitative prediction of enhancer-promoter interactions. *Genome Res.* 2020;30(1):72-84. DOI 10.1101/gr.249367.119.
- Benton M.C., Lea R.A., Macartney-Coxson D., Sutherland H.G., White N., Kennedy D., Mengersen K., Haupt L.M., Griffiths L.R. Genome-wide allele-specific methylation is enriched at gene regulatory regions in a multi-generation pedigree from the Norfolk Island isolate. *Epigenetics Chromatin.* 2019;12(1):60. DOI 10.1186/s13072-019-0304-7.
- Cavalli M., Baltzer N., Umer H.M., Grau J., Lemnian I., Pan G., Wallerman O., Spalinskas R., Sahlén P., Grosse I., Komorowski J., Wadelius C. Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci. Rep.* 2019;9(1):2695. DOI 10.1038/s41598-019-39633-0.

- Cavalli M., Pan G., Nord H., Wallén Artzt E., Wallerman O., Wadelius C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics*. 2016a;107(6):248-254. DOI 10.1016/j.ygeno.2016.04.006.
- Cavalli M., Pan G., Nord H., Wallerman O., Wallén Artzt E., Berggren O., Elvers I., Eloranta M.L., Rönnblom L., Lindblad Toh K., Wadelius C. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* 2016b;135(5):485-497. DOI 10.1007/s00439-016-1654-x.
- Chen C.-Y., Chang I.-S., Hsiung C.A., Wasserman W.W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics*. 2014;7:34. DOI 10.1186/1755-8794-7-34.
- Chen J., Rozowsky J., Galeev T.R., Harmanci A., Kitchen R., Bedford J., Abyzov A., Kong Y., Regan L., Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.* 2016;18(7):11101. DOI 10.1038/ncomms11101.
- Chen L., Liang Y., Qiu J., Zhang L., Chen X., Luo X., Jiang J. Significance of rs1271572 in the estrogen receptor beta gene promoter and its correlation with breast cancer in a southwestern Chinese population. *J. Biomed. Sci.* 2013;20:32. DOI 10.1186/1423-0127-20-32.
- Claussnitzer M., Dankel S.N., Kim K.-H., Quon G., Meuleman W., Haugen C., Glunk V., Sousa I.S., Beaudry J.L., Puviondrand V., Abdennur N.A., Liu J., Svensson P.-A., Hsu Y.-H., Drucker D.J., Mellgren G., Hui C.-Ch., Hauner H., Kellis M. *FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 2015; 373:895-907. DOI 10.1056/NEJMoa1502214.
- Cong Z., Li Q., Yang Y., Guo X., Cui L., You T. The SNP of rs6854845 suppresses transcription via the DNA looping structure alteration of super-enhancer in colon cells. *Biochem. Biophys. Res.* 2019;514: 734-741. DOI 10.1016/j.bbrc.2019.04.190.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. DOI 10.1038/nature11247.
- Farh K.K.-H., Marson A., Zhu J., Kleinewietfeld M., Housley W.J., Beik S., Shores N., Whitton H., Ryan R.J.H., Shishkin A.A., Hatan M., Carrasco-Alfonso M.J., Mayer D., Luckey C.J., Patshopoulos N.A., De Jager P.L., Kuchroo V.K., Epstein C.B., Daly M.J., Hafler D.A., Bernstein B.E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343. DOI 10.1038/nature13835.
- Fishman V.S., Salnikov P.A., Battulin N.R. Interpreting chromosomal rearrangements in the context of 3-dimensional genome organization: a practical guide for medical genetics. *Biochemistry*. 2018; 83(4):393-401. DOI 10.1134/S0006297918040107.
- Gorbacheva A.M., Korneev K.V., Kuprash D.V., Mitkin N.A. The risk G allele of the single-nucleotide polymorphism rs928413 creates a CREB1-binding site that activates *IL33* promoter in lung epithelial cells. *Int. J. Mol. Sci.* 2018;19(10):2911. DOI 10.3390/ijms19102911.
- Guo L., Wang J. rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res.* 2018;46(D1):D1111-D1116. DOI 10.1093/nar/gkx1101.
- Hansen A.S., Cattoglio C., Darzacq X., Tjian R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*. 2018; 9(1):20-32. DOI 10.1080/19491034.2017.1389365.
- Howard T.D., Mathias R.A., Seeds M.C., Herrington D.M., Hixson J.E., Shimmin L.C., Hawkins G.A., Sellers M., Ainsworth H.C., Sergeant S., Miller L.R., Chilton F.H. DNA methylation in an enhancer region of the *FADS* cluster is associated with *FADS* activity in human liver. *PLoS One*. 2014;9(5):e97510. DOI 10.1371/journal.pone.0097510.
- Ibrahim D.M., Mundlos S. Three-dimensional chromatin in disease: what holds us together and what drives us apart? *Curr. Opin. Cell Biol.* 2020;64:1-9. DOI 10.1016/j.ceb.2020.01.003.
- Izzi B., Pistoni M., Cludts K., Akkor P., Lambrechts D., Verfaillie C., Verhamme P., Freson K., Hoylaerts M.F. Allele-specific DNA methylation reinforces *PEAR1* enhancer activity. *Blood*. 2016;128: 1003-1012. DOI 10.1182/blood-2015-11-682153.
- Jones P.L., Veenstra G.J., Wade P.A., Vermaak D., Kass S.U., Landsberger N., Strouboulis J., Wolffe A.P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.* 1998;19:187-191. DOI 10.1038/561.
- Kilpinen H., Waszak S.M., Gschwind A.R., Raghav S.K., Witwicki R.M., Orioli A., Migliaiavacca E., Wiederkehr M., Gutierrez-Arcelus M., Panousis N., Yurovsky A., Lappalainen T., Romano-Palumbo L., Planchon A., Bielser D., Bryois J., Padiouleau I., Udin G., Thurnheer S., Hacker D., Core L.J., Lis J.T., Hernandez N., Raymond A., Deplancke B., Dermitzakis E.T. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342:744-747. DOI 10.1126/science.1242463.
- Korneev K.V., Sviriaeva E.N., Mitkin N.A., Gorbacheva A.M., Uvarova A.N., Ustiugova A.S., Polanovsky O.L., Kulakovskiy I.V., Afanasyeva M.A., Schwartz A.M., Kuprash D.V. Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression. *Biochim. Biophys. Acta Mol. Basis Dis.* 2020;1866(3):165626. DOI 10.1016/j.bbdis.2019.165626.
- Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.
- Kumar D., Puan K.J., Andiappan A.K., Lee B., Westerlaken G.H., Haase D., Melchioni R., Li Z., Yusof N., Lum J., Koh G., Foo S., Yeong J., Alves A.C., Pekkanen J., Sun L.D., Irwanto A., Fairfax B.P., Naranbhai V., Common J.E., Tang M., Chuang C.K., Jarvelin M.R., Knight J.C., Zhang X., Chew F.T., Prabhakar S., Jianjun L., Wang Y., Zolezzi F., Poidinger M., Lane E.B., Meyaard L., Röttschke O. A functional SNP associated with atopic dermatitis controls cell type-specific methylation of the *VSTM1* gene locus. *Genome Med.* 2017;9(1):18. DOI 10.1186/s13073-017-0404-6.
- Kumar S., Ambrosini G., Bucher P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* 2017;45(D1):D139-D144. DOI 10.1093/nar/gkw1064.
- Lee C.M., Barber G.P., Casper J., Clawson H., Diekhans M., Gonzalez J.N., Hinrichs A.S., Lee B.T., Nassar L.R., Powell C.C., Ranev B.J., Rosenbloom K.R., Schmelter D., Speir M.L., Zweig A.S., Haussler D., Haeussler M., Kuhn R.M., Kent W.J. UCSC Genome Browser enters 20th year. *Nucleic Acids Res.* 2020;48(D1):D756-D761. DOI 10.1093/nar/gkz1012.
- Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.
- Lewinsky R.H., Jensen T.G.K., Møller J., Stensballe A., Olsen J., Troelsen J.T. T₋₁₃₉₁₀ DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity *in vitro*. *Hum. Mol. Genet.* 2005;14(24):3945-3953. DOI 10.1093/hmg/ddi418.
- Li S., Li Y., Li X., Liu J., Huo Y., Wang J., Liu Z., Li M., Luo X.-J. Regulatory mechanisms of major depressive disorder risk variants. *Mol. Psychiatry*. 2020;25(9):1926-1945. DOI 10.1038/s41380-020-0715-7.
- Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Kloppock E., Horn D., Kayserli H., Opitz J.M., Laxova R., Santos-Simarro F., Gilbert-Dussardier B., Wittler L., Borschiwer M., Haas S.A., Osterwalder M., Franke M., Timmermann B., Hecht J., Spielmann M., Visel A., Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161(5):1012-1025. DOI 10.1016/j.cell.2015.04.004.

- Mathelier A., Shi W., Wasserman W.W. Identification of altered cis-regulatory elements in human disease. *Trends Genet.* 2015;31(2): 67-76. DOI 10.1016/j.tig.2014.12.003.
- Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., Sandstrom R., Qu H., Brody J., Shafer A., Neri F., Lee K., Kutayavin T., Stehling-Sun S., Johnson A.K., Canfield T.K., Giste E., Diegel M., Bates D., Hansen R.S., Neph S., Sabo P.J., Heimfeld S., Raubitschek A., Ziegler S., Cotsapas C., Sotoodehnia N., Glass I., Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012;337(6099):1190-1195. DOI 10.1126/science.1222794.
- McVicker G., van de Geijn B., Degner J.F., Cain C.E., Banovich N.E., Raj A., Lewellen N., Myrthil M., Gilad Y., Pritchard J.K. Identification of genetic variants that affect histone modifications in human cells. *Science.* 2013;342:747-749. DOI 10.1126/science.1242429.
- Meddens C., van der List A.C.J., Nieuwenhuis E.E.S., Mokry M. Non-coding DNA in IBD: from sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut.* 2019;68(5):928-941. DOI 10.1136/gutjnl-2018-317516.
- Mei S., Ke J., Tian J., Ying P., Yang N., Wang X., Zou D., Peng X., Yang Y., Zhu Y., Gong Y., Zhong R., Chang J., Miao X. A functional variant in the boundary of a topological association domain is associated with pancreatic cancer risk. *Mol. Carcinog.* 2019;58(10): 1855-1862. DOI 10.1002/mc.23077.
- Merkulov V.M., Leberfarb E.Y., Merkulova T.I. Regulatory SNPs and their widespread effects on the transcriptome. *J. Biosci.* 2018;43(5): 1069-1075. DOI 10.1007/s12038-018-9817-7.
- Nan X., Ng H.H., Johnson C.A., Laherty C.D., Turner B.M., Eisenman R.N., Bird A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature.* 1998;393:386-389. DOI 10.1038/30764.
- Park C.-Y., Halevy T., Lee D.R., Sung J.J., Lee J.S., Yanuka O., Benvenisty N., Kim D.-W. Reversion of *FMR1* methylation and silencing by editing the triplet repeats in fragile X iPSC-derived neurons. *Cell. Rep.* 2015;13(2):234-241. DOI 10.1016/j.celrep.2015.08.084.
- Qunneville S., Verde G., Corsinotti A., Kapopoulou A., Jakobsson J., Offner S., Baglivo I., Pedone P.V., Grimaldi G., Riccio A., Trono D. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell.* 2011;44(3):361-372. DOI 10.1016/j.molcel.2011.08.032.
- Rahbar E., Waits C.M.K., Kirby E.H., Jr., Miller L.R., Ainsworth H.C., Cui T., Sergeant S., Howard T.D., Langefeld C.D., Chilton F.H. Allele-specific methylation in the *FADS* genomic region in DNA from human saliva, CD4+ cells, and total leukocytes. *Clin. Epigenetics.* 2018;10:46. DOI 10.1186/s13148-018-0480-5.
- Reddy T.E., Gertz J., Pauli F., Kucera K.S., Varley K.E., Newberry K.M., Marinov G.K., Mortazavi A., Williams B.A., Song L., Crawford G.E., Wold B., Willard H.F., Myers R.M. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* 2012;22(5):860-869. DOI 10.1101/gr.131201.111.
- Roadmap Epigenomics Consortium, Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M.J., ... Hirst M., Meissner A., Milosavljevic A., Ren B., Stamatoyannopoulos J.A., Wang T., Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015; 518(7539):317-330. DOI 10.1038/nature14248.
- Rozowsky J., Abyzov A., Wang J., Alves P., Raha D., Harmanci A., Leng J., Bjornson R., Kong Y., Kitabayashi N., Bhardwaj N., Rubin M., Snyder M., Gerstein M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* 2011;7:522. DOI 10.1038/msb.2011.54.
- Schmitz R.J., Lewis Z.A., Goll M.G. DNA methylation: shared and divergent features across eukaryotes. *Trends Genet.* 2019;35(11): 818-827. DOI 10.1016/j.tig.2019.07.007.
- Shi W., Fornes O., Mathelier A., Wasserman W.W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* 2016;44(21):10106-10116. DOI 10.1093/nar/gkw691.
- Smith A.J.P., Deloukas P., Munroe P.B. Emerging applications of genome-editing technology to examine functionality of GWAS-associated variants for complex traits. *Physiol. Genomics.* 2018;50(7): 510-522. DOI 10.1152/physiolgenomics.00028.2018.
- Sun J.H., Zhou L., Emerson D.J., Phyto S.A., Titus K.R., Gong W., Gilgenast T.G., Beagan J.A., Davidson B.L., Tassone F., Phillips-Cremins J.E. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell.* 2018;175(1):224-238. DOI 10.1016/j.cell.2018.08.005.
- Visser M., Palstra R.J., Kayser M. Allele-specific transcriptional regulation of *IRF4* in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the *IRF4* promoter. *Hum. Mol. Genet.* 2015;24(9):2649-2661. DOI 10.1093/hmg/ddv029.
- Vohra M., Sharma A.R., Prabhu B.N., Rai P.S. SNPs in sites for DNA methylation, transcription factor binding, and miRNA targets leading to allele-specific gene expression and contributing to complex disease risk: a systematic review. *Public Health Genomics.* 2020;23: 1-16. DOI 10.1159/000510253.
- Wang H., Lou D., Wang Z. Crosstalk of genetic variants, allele-specific DNA methylation, and environmental factors for complex disease risk. *Front. Genet.* 2019;9:695. DOI 10.3389/fgene.2018.00695.
- Ward L.D., Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 2012;40(Database issue):D930-D934. DOI 10.1093/nar/gkr917.
- Waszak S.M., Kilpinen H., Gschwind A.R., Orioli A., Raghav S.K., Witwicki R.M., Migliavacca E., Yurovsky A., Lappalainen T., Hernandez N., Reymond A., Dermitzakis E.T., Deplancke B. Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics.* 2014;30(2):165-171. DOI 10.1093/bioinformatics/btt667.
- Wingender E., Schoeps T., Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 2013;41(D1):D165-D170. DOI 10.1093/nar/gks1123.
- Yates A.D., Achuthan P., Akanni W., Allen J., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Azov A.G., Bennett R., Bhai J., ... Perry E., Ruffier M., Trevanion S.J., Cunningham F., Howe K.L., Zerbino D.R., Flicek P. Ensembl 2020. *Nucleic Acids Res.* 2020; 48(D1):D682-D688. DOI 10.1093/nar/gkz966.
- Younesy H., Möller T., Heravi-Moussavi A., Cheng J.B., Costello J.F., Lorincz M.C., Karimi M.M., Jones S.J.M. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics.* 2014;30(8): 1172-1174. DOI 10.1093/bioinformatics/btt744.
- Zhang Y., Manjunath M., Zhang S., Chasman D., Roy S., Song J.S. Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. *Cancer Res.* 2018;78(7):1579-1591. DOI 10.1158/0008-5472.CAN-17-3486.
- Zhao T., Hu Y., Zang T., Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet.* 2019;10:1021. DOI 10.3389/fgene.2019.01021.

ORCID ID

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

Благодарности. Исследование поддержано из средств бюджетного проекта 0259-2021-0009.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 28.12.2020. После доработки 18.01.2021. Принята к публикации 18.01.2021.


Английский текст <https://vavilov.elpub.ru/jour>

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

Н.А. Шмаков^{1, 2} 

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

² Курчатовский геномный центр, Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

 shmakov@bionet.nsc.ru

Аннотация. Реконструкция транскриптома *de novo* – важная стадия биоинформатического анализа данных RNA-seq, которая позволяет получить последовательности транскриптов, присутствующих в изучаемом биологическом образце. Наличие точной и полной последовательности транскриптома организма, в свою очередь, является необходимым условием для дальнейшей работы с данными RNA-seq. Биоинформатическим сообществом было создано множество программ-сборщиков для реконструкции транскриптома из коротких прочтений RNA-seq. Сборщики позволяют проводить как *de novo* реконструкцию транскриптома, так и реконструкцию, основанную на картировании коротких прочтений RNA-seq на последовательность референсного генома организма. Большинство *de novo* сборщиков, работающих с данными RNA-seq, применяют технологию реконструкции последовательностей методом графов де Брейна. Однако детали их работы могут существенно различаться, поэтому различия могут встречаться и в результатах. Некоторые авторы рекомендуют для получения более полной и качественной сборки использовать гибридную сборку транскриптома – подход, основанный на комбинации результатов работы нескольких сборщиков. Преимущество такого подхода было продемонстрировано в ряде исследований по анализу транскриптомов на платформе Illumina. Нами предложен гибридный подход по созданию сборок транскриптома ячменя *Hordeum vulgare* изогенной линии Vowman и двух почти изогенных линий, полученных на основе Vowman и контрастных по окраске колоса, используя данные, полученные при секвенировании матричной РНК на платформе IonTorrent. В данном подходе применяются несколько индивидуальных сборщиков: Trans-ABYSS, maSPAdes и Trinity. Были оценены некоторые показатели, характеризующие полноту и точность сборки: доля обнаруженных в сборке известных транскриптов ячменя, доля задействованных в сборке прочтений из библиотек RNA-seq, значение критерия BUSCO. По совокупности этих показателей метасборки демонстрируют более высокое качество полученного транскриптома по сравнению с индивидуальными сборщиками.

Ключевые слова: RNA-seq; транскриптомика; *de novo* реконструкция транскриптома; IonTorrent.

Для цитирования: Шмаков Н.А. Улучшение качества сборки *de novo* транскриптомов ячменя на основе гибридного подхода для линий с изменениями окраски колоса и стебля. *Вавиловский журнал генетики и селекции*. 2021;25(1):30-38. DOI 10.18699/VJ21.004

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

N.A. Shmakov^{1, 2} 

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

² Kurchatov Genomics Center, Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

 shmakov@bionet.nsc.ru

Abstract. *De novo* transcriptome assembly is an important stage of RNA-seq data computational analysis. It allows the researchers to obtain the sequences of transcripts presented in the biological sample of interest. The availability of accurate and complete transcriptome sequence of the organism of interest is, in turn, an indispensable condition for further analysis of RNA-seq data. Through years of transcriptomic research, the bioinformatics community has developed a number of assembler programs for transcriptome reconstruction from short reads of RNA-seq libraries. Different assemblers makes it possible to conduct a *de novo* transcriptome reconstruction and a genome-guided reconstruction. The majority of the assemblers working with RNA-seq data are based on the De Bruijn graph method of sequence reconstruction. However, specifics of their procedures can vary drastically, as do their results. A number of authors recommend a hybrid approach to transcriptome reconstruction based on combining the results of several assemblers in order to achieve a better transcriptome assembly. The advantage of this approach has been demonstrated in a number

of studies, with RNA-seq experiments conducted on the Illumina platform. In this paper, we propose a hybrid approach for creating a transcriptome assembly of the barley *Hordeum vulgare* isogenic line Bowman and two nearly isogenic lines contrasting in spike pigmentation, based on the results of sequencing on the IonTorrent platform. This approach implements several *de novo* assemblers: Trinity, Trans-ABYSS and rnaSPAdes. Several assembly metrics were examined: the percentage of reference transcripts observed in the assemblies, the percentage of RNA-seq reads involved, and BUSCO scores. It was shown that, based on the summation of these metrics, transcriptome meta-assembly surpasses individual transcriptome assemblies it consists of.

Key words: RNA-seq; transcriptomics; *de novo* transcriptome reconstruction; IonTorrent.

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

Введение

В настоящее время лидирующую позицию в транскриптомных исследованиях занимает технология массового высокопроизводительного секвенирования второго поколения, применяемая к РНК (RNA-seq). Она заключается в выделении тотальной матричной РНК биологического образца, ее фрагментировании и дальнейшем секвенировании одновременно большого числа полученных коротких фрагментов (Engström et al., 2013; Hrdlickova et al., 2017).

Сборка *de novo* последовательностей транскриптов из секвенированных фрагментов является одной из важнейших стадий анализа эксперимента по профилированию транскриптома (Chang et al., 2014). Она позволяет получить последовательности, соответствующие мРНК, представленным в изучаемом образце. Существуют два основных подхода к реконструкции последовательностей транскриптома из библиотек коротких прочтений – так называемый метод OLC (overlap–layout–consensus) и метод графов де Брёйна (Li et al., 2012; Schliesky et al., 2012). Метод OLC заключается в попарном выравнивании прочтений и создании ориентированных графов, где каждый узел – это одно прочтение. В качестве ребер выступают перекрытия между прочтениями. Таким образом, путь по графу позволяет реконструировать контиг, который можно собрать из перекрывающихся прочтений. Использование метода OLC предпочтительнее для сборки контигов из сравнительно малого количества прочтений большой длины с большими участками перекрытия и поэтому используется чаще для сбора последовательностей, полученных методом Сэнгера, или методами секвенирования третьего поколения (Cui et al., 2020).

Второй метод заключается в построении графа де Брёйна, в котором вершинами выступают k -меры, т. е. последовательности нуклеотидов заданной длины k . Затем на графе отмечают все пути, составляющие последовательности коротких прочтений, полученных в результате секвенирования. После чего отмечают все пути, содержащие непрерывные последовательности перекрывающихся прочтений. Таким образом, находят последовательности контигов, которые можно собрать из прочтений библиотеки. Этот метод используется в таких программах-сборщиках транскриптома, как Trinity (Grabherr et al., 2013), Trans-ABYSS (Robertson et al., 2010), SOAPdenovo-Trans (Xie et al., 2014), Oases (Schulz et al., 2012).

Для сборщиков, основанных на методе графов де Брёйна, существует важный параметр k – длина k -меров, использованных при создании данного графа. Под k -мером

понимается длина слов, являющихся вершинами графа де Брёйна. Этот параметр может устанавливаться пользователем при запуске программ-сборщиков. Увеличение k повышает точность сборки, но одновременно увеличивает сложность вычисления (Fu et al., 2018). При более высоких значениях k сборщик может не обнаружить ограниченное пересечение между прочтениями, размер которого меньше k . Нередко применяется следующая стратегия – проведение предварительных сборок при разных значениях k , после чего из них путем объединения отдельных сборок и последующего удаления избыточности (см. ниже) составляется финальная *de novo* сборка транскриптома (Wang, Gribskov, 2017).

Поскольку на сегодняшний день разработано множество программ, осуществляющих сборку транскриптома *de novo*, отдельные исследования были посвящены вопросу о производительности и точности этих сборщиков. Обзоры, в которых сравниваются несколько программ для сборки транскриптома *de novo*, как правило, выделяют в качестве лучших и наиболее популярных программы Trinity, SOAPdenovo-Trans, Velvet-Oases (Jain et al., 2013; Honaas et al., 2016; Wang, Gribskov, 2017). Trinity, помимо непосредственно сборщика, включает в себя широкий набор утилит для оценки качества сборки, удаления слабо представленных контигов и других манипуляций с *de novo* сборкой. SOAPdenovo-Trans отмечают как программу, подходящую для сборки растительных транскриптомов (Payá-Milans et al., 2018).

При всем разнообразии современных сборщиков транскриптомов *de novo* ни один из них не идеален настолько, чтобы полностью удовлетворить требованиям качества и полноты сборки. Поэтому было высказано предположение, что применение нескольких сборщиков и дальнейшее создание одной «метасборки» дополнительно могут улучшить чувствительность и точность получения последовательностей транскриптома (Cerveau, Jackson, 2016). Под метасборкой при этом понимается совокупность всех *de novo* собранных разными программами контигов после удаления избыточности. Удаление избыточности состоит в удалении каждого контига, который является подсловом хотя бы одного другого контига в данном множестве контигов. Такой подход был опробован для реконструкции транскриптома немодельных растений с использованием трех сборщиков – Trinity, Trans-ABYSS, rnaSPAdes (Evangelistella et al., 2017). Были также предприняты попытки создания метасборок транскриптома, отталкиваясь от геном-ориентированныхборок (Venturini et al., 2018).

Однако, насколько нам известно, попыток оценить производительность такого подхода, как формирование метасборок транскриптома из индивидуальных *de novo* сборок, на данных, полученных на платформе секвенирования IonTorrent, до сих пор не было предпринято. При этом платформа IonTorrent, хотя и уступает в популярности платформам Illumina, остается востребованной в биологических исследованиях, в том числе в изучении микробных метагеномов (Lee et al., 2019), внутривидового генетического разнообразия дождевых червей (Shekhovtsov et al., 2019), трансгенных линий крыс (Bürckert et al., 2017), секвенировании геномов растений (Salina et al., 2018). Ряд авторов сравнивают платформы Illumina и IonTorrent, указывая, что прочтения IonTorrent, в отличие от прочтений Illumina, в среднем имеют несколько более низкую точность и некоторый разброс по длинам прочтений (Lahens et al., 2017).

Целью нашей работы является создание вычислительного конвейера, основанного на построении метасборки транскриптома с помощью программ сборки *de novo* rnaSPAdes, Trans-ABYSS, Trinity, а также версии сборки Trinity с использованием референсного генома. Вычислительный конвейер был апробирован на задаче сборки транскриптомов ячменя *Hordeum vulgare* L. изогенной линии Bowman и почти изогенных линий i:BwAlm с частичным альбинизмом колоса и стебля и BLP с частичным меланизмом колоса. Установлено, что качество сборки транскриптомов у разных сборщиков различается, однако в целом их результаты дополняют друг друга. Наилучшее качество сборки обеспечивает метасборка транскриптома, которая превосходит индивидуальные сборки по ряду параметров, характеризующих качество сборок транскриптома.

Материалы и методы

Библиотеки коротких прочтений. Использовались библиотеки транскриптомов ячменя *H. vulgare* изогенной линии Bowman и двух почти изогенных линий: i:BwAlm (характеризуется частичным альбинизмом колоса и стеб-

ля) и BLP (характеризуется частичным меланизмом колоса). Данные были загружены из базы данных SRA NCBI BioProject PRJNA342150 (библиотеки почти изогенной линии i:BwAlm и изогенной линии Bowman) и PRJNA399215 (библиотеки почти изогенной линии BLP и изогенной линии Bowman).

Эксперимент PRJNA342150 состоит в сравнении транскриптомов леммы почти изогенной линии i:BwAlm, полученной на основе изогенной линии Bowman, и самой линии Bowman, взятой в качестве контроля (Shmakov et al., 2016). Для каждой из линий было взято по три биологических повторности. Таким образом, в эксперименте задействовано шесть библиотек коротких прочтений RNA-seq. Этот эксперимент для краткости и удобства далее будем называть «эксперимент alm».

В эксперименте PRJNA399215 сравнивались транскриптом почти изогенной линии ячменя BLP, полученной на основе изогенной линии Bowman, и сама линия Bowman, взятая в качестве контроля (Glagoleva et al., 2017). Для каждой линии ячменя было взято по три биологических повторности. Таким образом, в эксперименте были использованы шесть библиотек RNA-seq. Для краткости будем называть его «эксперимент blp».

Все библиотеки были получены с помощью секвенирования на платформе IonTorrent. Далее библиотеки прошли процедуру фильтрации, которая состояла в удалении адаптерных последовательностей с помощью программы CutAdapt версии 1.9.1 (Martin, 2011) и удалении прочтений со средним значением качества ниже 20, длинами ниже 50 или больше 270 с помощью программы PRINSEQ-lite версии 0.20.4 (Schmieder, Edwards, 2011). Характеристики использованных в исследовании библиотек приведены в табл. 1.

Получение сборки транскриптомов. Использовались три сборщика транскриптома: Trinity (Grabherr et al., 2013) версии 2.2.0, Trans-ABYSS (Robertson et al., 2010) версии 2.0.1 и rnaSPAdes (Bushmanova et al., 2018) версии 3.12.0. Все указанные программы в исследованиях по сравнению производительности и качеству сборщиков

Таблица 1. Характеристики использованных библиотек коротких прочтений

Эксперимент	Линия	Библиотека	Сырой размер	Очищенный размер	Средняя длина прочтения
PRJNA342150	i:BwAlm	Alm_1	4596395	3874912	166.94
		Alm_2	3056413	2372255	199.52
		Alm_3	5794644	5332600	181.47
	Bowman	A_bow_1	4122599	2450068	175.49
		A_bow_2	4023501	2356572	126.56
		A_bow_3	6887599	6523266	201.68
PRJNA399215	BLP	Blp_1	3583148	1311442	185.39
		Blp_2	4710862	1687289	156.96
		Blp_3	4070591	1864073	146.02
	Bowman	B_bow_1	1769261	438702	164.66
		B_bow_2	3740926	1092191	199.48
		B_bow_3	5253524	2364034	209.00

транскриптома *de novo* были отмечены в числе лучших (Honaas et al., 2016; Lafond-Lapalme et al., 2017; Fu et al., 2018; Hölzer, Marz, 2019).

Работу с библиотеками из двух экспериментов проводили по отдельности. Индивидуальные сборки транскриптомов для каждого эксперимента были получены следующим образом.

Запуск сборщика Trinity проходил с параметрами «по умолчанию», на ввод программы были поданы шесть библиотек, относящихся к данному эксперименту. При запуске программы SPAdes на ввод тоже были поданы все шесть библиотек коротких прочтений, относящихся к этому эксперименту, и указаны опции ‘-iontorrent’ и ‘-only-assembler’.

Сборка программой Trans-ABySS была проведена по отдельности для каждой из библиотек, относящихся к данному эксперименту, после чего программой transabyss-merge, входящей в пакет Trans-ABySS, полученные сборки были объединены. Эта сборка проходила с параметрами «по умолчанию», при которых длина k -мера равна 32. Аналогичным образом проведены сборки со значениями параметра k 48 и 64. Таким образом, с помощью Trans-ABySS были созданы три сборки *de novo*, различающиеся длинами k -меров. Затем эти три сборки были объединены программой transabyss-merge. Результирующую сборку далее использовали как индивидуальную сборку транскриптома *de novo*, полученную с помощью программы trans-ABySS.

Дополнительно была проведена геном-ориентированная сборка программой Trinity. Для этого сначала библиотеки коротких прочтений были картированы на геном ячменя. Затем из файлов картирования библиотек в формате sam (sequence alignment/mapping) был скомпонован общий файл, объединяющий все шесть картирований, при помощи команды merge программы samtools версии 1.6. Этот файл, вместе с шестью библиотеками, относящимися к данному эксперименту, был использован для сборки программой Trinity в режиме геном-ориентированной сборки транскриптома, с указанием при этом максимальной длины интрона в 500 000 нуклеотидов.

Для удаления избыточности сборок была задействована программа tr2aacds.pl из линейки программ Evidential Gene версии 20.05.2020 (Gilbert, 2019). Каждую из сборок обрабатывали этой программой по отдельности. Таким образом, получили три избыточные сборки транскриптома *de novo* и одну избыточную геном-ориентированную сборку. В дальнейшем для простоты будем называть *de novo* сборки сокращенными названиями соответствующих программ: abyss, spades и trinity – для сборок, созданных с помощью Trans-ABySS, maSPAdes и Trinity. Геном-ориентированную сборку будем называть сокращенно GG (от англ. genome-guided – геном-ориентированная).

Для получения оптимального метатранскриптома сборки были конкатенированы в один файл, после чего этот файл для удаления избыточности также был обработан программой tr2aacds.pl. Следует отметить, что здесь и далее рассматриваются контиги, имеющие открытые рамки считывания, так как tr2aacds.pl использует для дальнейшего анализа только те контиги, в которых были

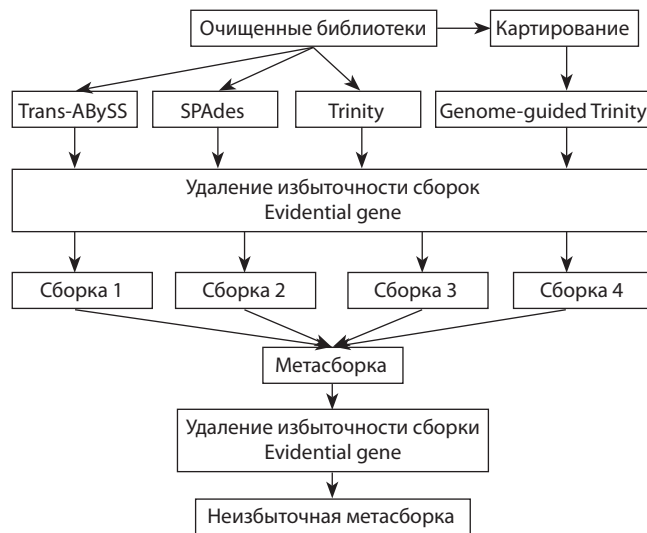


Рис. 1. Схема получения индивидуальных сборок *de novo* и метасборки транскриптома ячменя.

предсказаны открытые рамки считывания, имеющие длину не меньше пороговой. Основные этапы получения избыточной метасборки показаны на рис. 1.

Таким образом, для каждого из двух экспериментов было создано по четыре индивидуальные сборки транскриптома: spades и trinity, составленные каждая из всех шести библиотек коротких прочтений, входящих в этот эксперимент; abyss, проведенная для каждой из библиотек по отдельности с разными значениями k -меров, после чего сборки для разных библиотек были объединены в одну сборку abyss с помощью программы abyss-merge; геном-ориентированная сборка GG, составленная из всех шести библиотек, входящих в этот эксперимент, и файла картирования, объединенного из файлов картирования всех шести библиотек, входящих в эксперимент, на геном ячменя. Далее из четырех индивидуальных сборок для каждого из экспериментов была получена одна метасборка транскриптома ячменя.

Оценка качества сборок транскриптомов. Все индивидуальные и метасборки прошли обработку программой BUSCO версии 3.0.2 (Simão et al., 2015) для оценки полноты сборок исходя из представленности характерных для растений последовательностей и TransRate версии 1.0.3 (Smith-Unna et al., 2016) для аннотации контигов и оценки полноты наличия генов ячменя в сборке. После этого проведено сравнение наборов CDS ячменя, обнаруженных программой TransRate в каждой из индивидуальных сборок. На основании перекрытия множеств CDS, выявленных в каждой из индивидуальных сборок, были построены диаграммы Венна, иллюстрирующие вклад каждого из сборщиков транскриптома *de novo* в структуру метасборки.

Далее контиги двух метасборок транскриптома ячменя, относящиеся к двум экспериментам, были выровнены на последовательность генома ячменя *H. vulgare* с помощью программы rnaQUAST (Bushmanova et al., 2016). rnaQUAST подсчитывает и предоставляет для оценки пользователя различные параметры, основываясь на вы-

равнинности контигов и референса, благодаря чему можно оценить качество сборки. В частности, эта программа разделяет контиги на три категории: контиги, выровненные на референс и совпадающие с аннотированными генами; контиги, выровненные на референс, но не совпадающие с известными аннотированными генами; и контиги, не имеющие существенной гомологии к референсному генному. Эту последнюю группу будем называть «новыми контигами».

Сравнение качества сборок транскриптома. С целью количественного сравнения качества сборок использовали подход, предложенный в (Holzner, Marz, 2019). Он состоит в том, чтобы для ряда выбранных параметров, отражающих качество сборки транскриптома *de novo*, провести процедуру нормализации по формуле

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

Здесь R_j^i – значение параметра i для сборки транскриптома j до нормализации; N_j^i – значение этого параметра после нормализации; V^i – вектор, составленный из всех значений параметра i для всех k сборок транскриптома *de novo* до нормализации: $V^i = (V_1^i, \dots, V_k^i)$. Таким образом, после нормализации каждый из параметров принимает значение от 0 до 1 для каждой сборки *de novo*. После этого для каждой из сборок все значения нормализованных параметров суммируются и проводится градация сборок по значению суммы всех нормализованных параметров. Сборка, имеющая наибольшую сумму нормализованных параметров, считается наиболее качественной.

Для сравнения качества индивидуальных сборок и метасборок транскриптома ячменя, полученных при работе с библиотеками коротких прочтений, относящихся к двум экспериментам, были использованы семь параметров, характеризующих разные аспекты качества сборки транскриптома: 1) N50; 2) медиана распределения длин

контигов; 3) количество обнаруженных (как целиком, так и фрагментарно) генов из списка BUSCO; 4) доля контигов, для которых с помощью TransRate была выявлена гомология с известными CDS ячменя; 5) количество CDS ячменя, с которыми контиги из сборки *de novo* имеют гомологию; 6) количество CDS ячменя, не менее 95 % длины которых покрыто выравниванием с контигами из сборки *de novo*; 7) доля прочтений из библиотек, использованных для создания сборки *de novo*, псевдокартированных на эту сборку с помощью программы kallisto. Параметры 1 и 2 отражают распределение длин контигов, 3–6 – полноту сборки транскриптома, а параметр 7 – полноту использования библиотек коротких прочтений при составлении этой сборки.

Результаты

Эксперимент alm

Для линии ячменя i:BwAlm и использованной в качестве контроля изогенной линии Bowman были получены четыре индивидуальные сборки *de novo* транскриптома леммы и перикарпа и одна метасборка, составленная из четырех индивидуальных сборок. В табл. 2 приведены результаты сборки *de novo* транскриптома ячменя линий i:BwAlm и Bowman, включая метасборки, а также общей для двух линий генеральной сборки.

Метасборка транскриптома ячменя линий i:BwAlm и Bowman, полученная из сборок *de novo*, созданных с помощью rnaSPAdes, Trans-ABYSS и Trinity, и геном-ориентированной сборки trinity, до удаления избыточности состоит из 169232 контигов. Избыточность метасборки включает 68414 контигов суммарной длиной 46440750 оснований. Максимальная длина контига в сборке – 9920 нуклеотидов, средняя длина – 678.8 нуклеотида, N50 – 936 нуклеотидов. Удаление избыточности уменьшило размер метасборки до 40.4 % от исходного.

Таблица 2. Характеристики *de novo* сборок транскриптома ячменя в эксперименте alm

Сборка	Размер сборки, контигов		N50	Средняя длина	Прочтений картировано, %
	Избыточная	Неизбыточная			
abyss	705 015	40 806	1076	723.60	67.08
spades	22 649	19 181	1130	1072.65	39.13
trinity	267 201	52 005	976	741.19	64.97
GG	451 309	57 240	766	594.82	61.37
Метасборка	169 232	68 414	936	678.82	61.47

Таблица 3. Количество известных CDS ячменя, обнаруженных в *de novo* сборках транскриптома в эксперименте alm

Сборка	Контиги		CDS найдено	p_95
	кол-во	%		
abyss	30 530	0.748	22 420	2542
spades	17 323	0.903	14 989	644
trinity	35 547	0.684	27 173	1779
GG	38 686	0.676	26 978	2240
Метасборка	42 887	0.627	29 790	3073

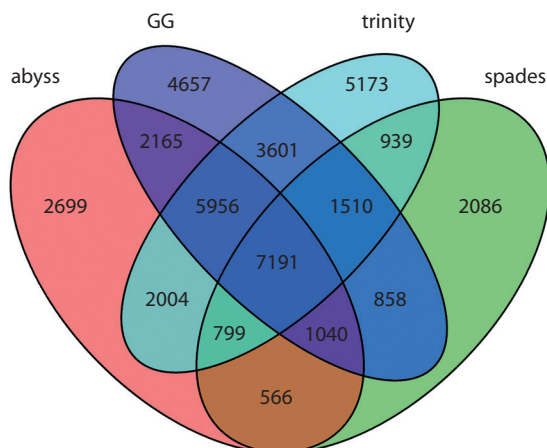


Рис. 2. Диаграмма Венна, показывающая перекрывание множеств CDS, обнаруженных в индивидуальных сборках транскриптома *de novo* в эксперименте alm.

Проведена оценка покрытия контигов прочтениями библиотек в индивидуальных сборках и метасборке транскриптома с помощью технологии псевдокартирования. Установлено, что наибольшая доля прочтений была выровнена на сборку транскриптома abyss, тогда как наименьшая – на сборку spades. На метасборку транскриптома было выровнено 61.47 % всех коротких прочтений (см. табл. 2).

Был проведен поиск известных CDS ячменя в сборках транскриптома *de novo* с помощью программы TransRate. Результаты идентификации CDS для разных сборок представлены в табл. 3.

Наибольшее количество известных CDS (29 790) обнаружено в метасборке транскриптома. Также здесь выявлено самое большое количество CDS, покрытых контигами сборки не менее чем на 95 %. Однако при этом максимальная доля контигов, для которых выявлена значимая гомология с CDS ячменя, представлена в сборке spades – 90.3 %. В метасборке этот показатель составил всего 62.7 % – меньше, чем во всех индивидуальных сборках.

Далее для оценки вклада каждого из сборщиков в структуру метасборки транскриптома была проведена оценка перекрывания множеств CDS ячменя, встреченных в каждой из индивидуальных сборок (рис. 2). Как можно видеть, 7191 CDS ячменя был обнаружен во всех четырех индивидуальных сборках транскриптома, еще 9305 CDS найдены в трех сборках из четырех. 14 615 CDS были обнаружены только в одной из четырех сборок, из которых наибольшее количество (5173) выявлено только в сборке trinity, наименьшее (2086) – только в сборке spades. Максимальное перекрывание множеств, обнаруженных CDS, наблюдалось между индивидуальными сборками trinity и GG – 18 258 CDS.

В контигах каждой из сборок были предсказаны открытые рамки считывания (ОРС). Найденные в контигах общей сборки ОРС кодируют 58 636 белковых продуктов длинами не менее 30 аминокислотных остатков. Эти белковые продукты были использованы для того, чтобы оценить полноту сборок при помощи программы BUSCO (рис. 3). В метасборке транскриптома количество выяв-

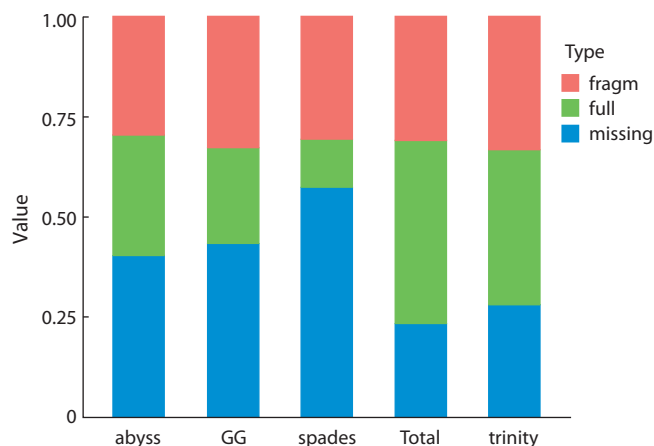


Рис. 3. Полнота сборок транскриптома по критерию BUSCO в эксперименте alm.

ленных полных последовательностей BUSCO оказалось больше, чем в индивидуальных сборках, а количество фрагментированных – меньше, как и количество отсутствующих. Это говорит о преимуществе метасборки транскриптома по полноте и качеству.

Эксперимент blp

Для библиотек RNA-seq из эксперимента blp были построены индивидуальные сборки транскриптома *de novo* и метасборка транскриптома, после чего проведено сравнение их качества (табл. 4).

Исходная избыточная метасборка транскриптома ячменя линий Bowman и BLP состоит из 133 070 контигов. После удаления избыточности в метасборке осталось 32 466 контигов суммарной длиной 25 184 753 основания. Таким образом, в ходе удаления избыточности количество контигов было уменьшено до 24.4 % от исходного. Отметим также, что метасборка транскриптома в эксперименте blp имеет более высокое значение длин контигов N50, чем индивидуальные сборки, из которых она составлена. 72.1 % всех прочтений из библиотек эксперимента blp было картировано на метасборку транскриптома. По этому показателю метасборка уступает сборке GG (77.6 %), но опережает три другие индивидуальные сборки.

В сборке транскриптома *de novo* исследуемых линий был проведен поиск известных CDS с помощью программы TransRate (табл. 5). Гомологию к известным CDS ячменя показывают от 19 848 контигов в сборке spades до 29 412 контигов в сборке GG. При этом наибольшее количество CDS ячменя обнаружено в сборке trinity, а максимальное количество CDS ячменя, покрытых контигами сборки не менее чем на 95 % своей длины, – в метасборке транскриптома (1825). Доля контигов из сборки, для которых была установлена гомология к известным CDS ячменя, в метасборке составляет 74.5 %, что ниже, чем у всех индивидуальных сборок, кроме trinity.

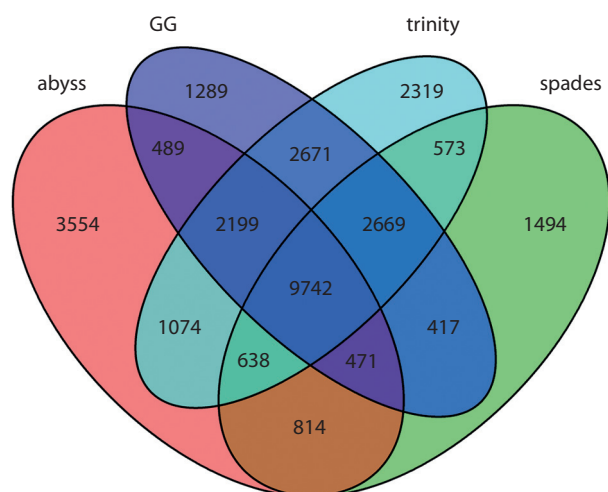
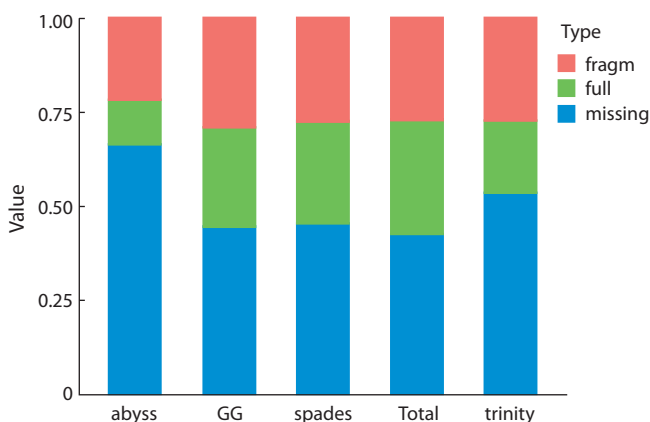
Далее был проведен поиск перекрывания полученных для индивидуальных сборок транскриптома списков CDS и оценен вклад каждой индивидуальной сборки в общую структуру (рис. 4). Во всех четырех индивидуальных сборках транскриптома *de novo* были обнаружены

Таблица 4. Характеристики *de novo* сборок транскриптома ячменя в эксперименте blp

Сборка	Размер сборки, контигов		N50	Средняя длина	Прочтений картировано, %
	Избыточная	Неизбыточная			
abyss	214 465	34 987	606	490.32	68.75
spades	31 453	24 401	1046	824.60	58.25
trinity	116 897	34 363	891	661.59	66.55
GG	122 304	39 319	976	707.83	77.55
Метасборка	133 070	32 466	1056	775.73	72.07

Таблица 5. Количество известных CDS ячменя, обнаруженных в *de novo* сборках транскриптома в эксперименте blp

Сборка	Контиги		CDS найдено	p_95
	количество	%		
abyss	25 804	0.738	18 981	1224
spades	19 848	0.813	16 818	1017
trinity	22 793	0.663	21 885	1478
GG	29 412	0.748	19 947	1597
Метасборка	24 194	0.745	19 665	1825

**Рис. 4.** Пересечение множеств CDS, обнаруженных в индивидуальных сборках транскриптома *de novo* эксперимента blp.**Рис. 5.** Полнота сборок транскриптома в эксперименте blp по BUSCO.

9742 CDS. 8656 CDS были обнаружены только в одной из индивидуальных сборок, из которых максимальное количество (3554 CDS) было уникальным для сборки abyss, а наименьшее (1289 CDS) – для сборки GG. Наибольшее количество общих CDS (17281) имеют сборки GG и trinity.

При оценке полноты сборок с помощью программы BUSCO установлено, что полнота метасборки транскриптома превышает полноту индивидуальных сборок (рис. 5). В ней обнаружено наибольшее количество полных последовательностей BUSCO, а количество невыявленных последовательностей BUSCO меньше, чем в индивидуальных сборках. Суммарно в неизбыточной метасборке транскриптома встречаются в полном или частичном виде 57.6 % всех последовательностей BUSCO из набора для покрытосеменных организмов.

Сравнение качества сборок *de novo*

С целью определения качества сборок были оценены семь параметров индивидуальных сборок *de novo* и метасборок транскриптома. Это длины контигов в полученных сборках *de novo* (N50 и медиана распределения длин контигов); наличие в сборке *de novo* известных CDS ячменя (доля контигов, имеющих сходство с CDS ячменя, количество обнаруженных CDS и количество CDS, покрытых не менее чем на 95 % от их длины) и генов, характерных для сосудистых растений (BUSCO-значения); полнота использования библиотек коротких прочтений при создании сборки *de novo* (доля псевдокартированных прочтений). Значения этих параметров были нормализованы и приведены в диапазон от 0 до 1 (Hölzer, Mars, 2019), после чего просуммированы для каждой индивидуальной сборки транскриптома *de novo* и для метасборки. Наибольшие значения суммы нормализованных параметров будут указывать на самую оптимальную сборку транскриптома (табл. 6).

Таблица 6. Суммарные нормализованные значения качества индивидуальных сборок транскриптома и метасборок

Сборка	Эксперимент (линии iBwAlm и Bowman)	Эксперимент (линии BLP и Bowman)
abyss	4.16	1.72
spades	3.00	3.86
trinity	4.07	3.61
GG	2.85	5.22
Метасборка	4.32	5.56

Наибольшие значения суммы нормализованных параметров в обоих экспериментах принадлежат метасборке транскриптома (см. табл. 6). Это, вкпе с максимальной среди всех имеющихсяборок полнотой представленности генов, характерных для сосудистых растений, обнаруженных с помощью программы BUSCO, и наибольшим количеством полно реконструированных CDS ячменя, указывает на то, что метасборки транскриптома, полученные путем объединения индивидуальныхборок *de novo* и удаления избыточности, опережают по своему качеству все индивидуальные сборки транскриптома.

Обсуждение

В нашей работе был протестирован подход к реконструкции транскриптома *de novo*, состоящий в создании метасборки из нескольких индивидуальныхборок транскриптома. Установлено, что метасборки транскриптома имеют большую полноту, исходя из таких критериев, как количество обнаруженных фрагментов BUSCO, количество CDS ячменя, гомологичные которым последовательности были обнаружены в сборке транскриптома, и доля псевдокартированных на сборки прочтений из библиотек RNA-seq. Таким образом, можно заключить, что описанный выше подход к *de novo* реконструкции транскриптома, состоящий в создании нескольких индивидуальныхборок транскриптома *de novo* и последующем объединении их в метасборку, повышает качество реконструированного транскриптома.

Сравнение нескольких программ для реконструкции транскриптома показало, что программа rnaSPAdes реконструирует наименьшее количество контигов, в то время как Trans-ABuSS – самое большое количество контигов. Сборщик Trinity реконструирует сравнимые количества контигов при запуске в двух режимах – *de novo* и genome-guided. При этом удаление избыточности уменьшает размерборок Trans-ABuSS сильнее всего: в эксперименте alm было удалено 94.3 % всех контигов, реконструированных Trans-ABuSS, в эксперименте blp – 83.7 %. В случае со сборками spades было удалено 15.3 и 22.4 % всех контигов соответственно. В сборках trinity удаляется в среднем 80.5 и 70.6 % всех контигов, в геном-ориентированных сборках – 87.3 и 67.8 % контигов соответственно. Геном-ориентированные сборки в обоих экспериментах имеют наибольший размер после удаления избыточности, сборки spades – наименьший.

Spades реконструирует самые длинные контиги из всех индивидуальныхборок, что характеризуется самыми

большими значениями N50 и медианы распределения длин контигов. Наименьшее значение N50 в эксперименте alm наблюдается у сборки GG, тогда как в эксперименте blp – у сборки abyss.

Наибольшей полнотой, согласно параметру BUSCO, в эксперименте alm из всех индивидуальныхборок обладает сборка trinity. В эксперименте blp это сборка GG. Наименьшей полнотой по BUSCO обладают сборки spades в эксперименте alm и сборка abyss в эксперименте blp.

Заключение

Таким образом, в двух экспериментах наблюдается разная производительность сборщиков транскриптома *de novo*, несмотря на то что в обоих случаях используются библиотеки коротких прочтений, полученные на платформе IonTorrent, и реконструируемый транскриптом принадлежит одному организму – ячменю *H. vulgare*. Это указывает на чувствительность задействованных сборщиков к входным данным, т.е. их производительность может сильно различаться в зависимости от данных.

Однако в обоих случаях метасборки транскриптома, составленные из индивидуальныхборок, имеют более высокое качество, чем любая из индивидуальныхборок транскриптома. Это говорит об эффективности такого подхода реконструкции транскриптомов, как создание метасборок, объединяющих в себе результаты работы нескольких сборщиков транскриптома *de novo*.

Список литературы / References

- Bürckert J.P., Dubois A.R.S.X., Faison W.J., Farinelle S., Charpentier E., Sinner R., Wienecke-Baldacchino A., Muller C.P. Functionally convergent B cell receptor sequences in transgenic rats expressing a human B cell repertoire in response to tetanus toxoid and measles antigens. *Front. Immunol.* 2017. DOI 10.3389/fimmu.2017.01834.
- Bushmanova E., Antipov D., Lapidus A., Przhibelskiy A.D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *BioRxiv.* 2018. DOI 10.1101/420208.
- Bushmanova E., Antipov D., Lapidus A., Suvorov V., Przhibelski A.D. rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies. *Bioinformatics.* 2016;32(14):2210-2212. DOI 10.1093/bioinformatics/btw218.
- Cerveau N., Jackson D.J. Combining independent *de novo* assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC Bioinform.* 2016;17:525. PMID: 27938328. DOI 10.1186/s12859-016-1406-x.
- Chang Z., Wang Z., Li G. The impacts of read length and transcriptome complexity for *de novo* assembly: a simulation study. *PLoS One.* 2014;9(4):e94825. PMID: 24736633. DOI 10.1371/journal.pone.0094825.
- Cui J., Shen N., Lu Z., Xu G., Wang Y., Jin B. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the *Arabidopsis* transcriptome. *Plant Methods.* 2020;16:85. DOI 10.1186/s13007-020-00629-x.
- Engström P.G., Steijger T., Sipos B., Grant G.R., Kahles A., Rättsch G., Goldman N., Hubbard T.J., Harrow J., Guigó R., Bertone P., Alioto T., Behr J., Bohnert R., Campagna D., Davis C.A., Dobin A., Gingeras T.R., Jean G., Kosarev P., Li S., Liu J., Mason C.E., Molodtsov V., Ning Z., Ponstingl H., Prins J.F., Ribeca P., Seledtsov I., Solovyev V., Valle G., Vitulo N., Wang K., Wu T.D., Zeller G. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods.* 2013;10:1185-1191. PMID: 24185836. DOI 10.1038/nmeth.2722.
- Evangelistella C., Valentini A., Ludovisi R., Firrincieli A., Fabbrini F., Scalabrini S., Cattonaro F., Morgante M., Mugnozza G.S., Keuren-

- tjes J.J.B., Harfouche A. De novo assembly, functional annotation, and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock. *Biotechnol. Biofuels*. 2017;10:138. DOI 10.1186/s13068-017-0828-7.
- Fu S., Ma Y., Yao H., Xu Z., Chen S., Song J., Au K.F. IDP-denovo: *de novo* transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*. 2018;34(13):2168-2176. PMID: 28407034. DOI 10.1093/bioinformatics/bty098.
- Gilbert D.G. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. *PeerJ*. 2019;7:e6374. DOI 10.7717/peerj.6374.
- Glagoleva A.Y., Shmakov N.A., Shoeva O.Y., Vasiliev G.V., Shatskaya N.V., Börner A., Afonnikov D.A., Khlestkina E.K. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the *Black lemma and pericarp (Blp)* gene. *BMC Plant Biol*. 2017;17:182. DOI 10.1186/s12870-017-1124-1.
- Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol*. 2013;29:644-652. PMID: 21572440. DOI 10.1038/nbt.1883.
- Hölzer M., Marz M. *De novo* transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*. 2019;8(5):giz039. PMID: 31077315. DOI 10.1093/gigascience/giz039.
- Honaas L.A., Wafula E.K., Wickett N.J., Der J.P., Zhang Y., Edger P.P., Altman N.S., Chris Pires J., Leebens-Mack J.H., DePamphilis C.W. Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS One*. 2016;11(1):e0146062. PMID: 26731733. DOI 10.1371/journal.pone.0146062.
- Hrdlickova R., Toloue M., Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*. 2017;8:e1364. PMID: 27198714. DOI 10.1002/wrna.1364.
- Jain P., Krishnan N.M., Panda B. Augmenting transcriptome assembly by combining *de novo* and genome-guided tools. *PeerJ*. 2013;1:e133. PMID: 24024083. DOI 10.7717/peerj.133.
- Lafond-Lapalme J., Duceppe M.O., Wang S., Moffett P., Mimee B. A new method for decontamination of *de novo* transcriptomes using a hierarchical clustering algorithm. *Bioinformatics*. 2017;33(9):1293-1300. PMID: 28011783. DOI 10.1093/bioinformatics/btw793.
- Lahens N.F., Ricciotti E., Smirnova O., Toorens E., Kim E.J., Baruzzo G., Hayer K.E., Ganguly T., Schug J., Grant G.R. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genom*. 2017;18:602. PMID: 28797240. DOI 10.1186/s12864-017-4011-0.
- Lee S., La T.M., Lee H.J., Choi I.S., Song C.S., Park S.Y., Lee J.B., Lee S.W. Characterization of microbial communities in the chicken oviduct and the origin of chicken embryo gut microbiota. *Sci. Rep*. 2019;9:6838. PMID: 31048728. DOI 10.1038/s41598-019-43280-w.
- Li Z., Chen Y., Mu D., Yuan J., Shi Y., Zhang H., Gan J., Li N., Hu X., Liu B., Yang B., Fan W. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief Funct. Genomics*. 2012;11(1):25-37. PMID: 22184334. DOI 10.1093/bfpg/blr035.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBNet.Journal*. 2011;17(1):10-12. PMID: 100006697. DOI 10.14806/ej.17.1.200.
- Payá-Milans M., Olmstead J.W., Nunez G., Rinehart T.A., Staton M. Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species. *GigaScience*. 2018;7(12):giy132. PMID: 30418578. DOI 10.1093/gigascience/giy132.
- Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S.D., Mungall K., Lee S., Okada H.M., Qian J.Q., Griffith M., Raymond A., Thiessen N., Cezard T., Butterfield Y.S., Newsome R., Chan S.K., She R., Varhol R., Kamoh B., Prabhu A.L., Tam A., Zhao Y., Moore R.A., Hirst M., Marra M.A., Jones S.J.M., Hoodless P.A., Birol I. *De novo* assembly and analysis of RNA-seq data. *Nat. Methods*. 2010;7(11):909-912. DOI 10.1038/nmeth.1517.
- Salina E.A., Nesterov M.A., Frenkel Z., Kiseleva A.A., Timonova E.M., Magni F., Vrána J., Šafář J., Šimková H., Doležel J., Korol A., Sergeeva E.M. Features of the organization of bread wheat chromosome 5BS based on physical mapping. *BMC Genom*. 2018;19:80. PMID: 29504906. DOI 10.1186/s12864-018-4470-y.
- Schliesky S., Gowik U., Weber A.P.M., Bräutigam A. RNA-seq assembly – are we there yet? *Front. Plant Sci*. 2012;3:220. DOI 10.3389/fpls.2012.00220.
- Schmieder R., Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27:863-864. PMID: 21278185. DOI 10.1093/bioinformatics/btr026.
- Schulz M.H., Zerbino D.R., Vingron M., Birney E. *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012;28(8):1086-1092. PMID: 22368243. DOI 10.1093/bioinformatics/bts094.
- Shekhovtsov S.V., Ershov N.I., Vasiliev G.V., Peltek S.E. Transcriptomic analysis confirms differences among nuclear genomes of cryptic earthworm lineages living in sympatry. *BMC Evol. Biol*. 2019;19:50. PMID: 30813890. DOI 10.1186/s12862-019-1370-y.
- Shmakov N.A., Vasiliev G.V., Shatskaya N.V., Doroshkov A.V., Gordeeva E.I., Afonnikov D.A., Khlestkina E.K. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq. *BMC Plant Biol*. 2016;16. DOI 10.1186/s12870-016-0926-x.
- Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210-3212. PMID: 26059717. DOI 10.1093/bioinformatics/btv351.
- Smith-Unna R., Boursnell C., Patro R., Hibberd J.M., Kelly S. TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res*. 2016;26:1134-1144. PMID: 27252236. DOI 10.1101/gr.196469.115.
- Venturini L., Caim S., Kaithakottil G.G., Mapleson D.L., Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*. 2018;7(8):giy093. PMID: 30052957. DOI 10.1093/gigascience/giy093.
- Wang S., Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics*. 2017;33(3):327-333. PMID: 27694201. DOI 10.1093/bioinformatics/btw625.
- Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Huang W., He G., Gu S., Li S., Zhou X., Lam T.W., Li Y., Xu X., Wong G.K.S., Wang J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2014;30(12):1660-1666. DOI 10.1093/bioinformatics/btu077.

Благодарности. Работа поддержана грантом РНФ № 18-14-00293 (формулировка задачи, создание алгоритмов, анализ данных). Выполнена с использованием вычислительных ресурсов ЦКП «Биоинформатика» при поддержке бюджетного проекта № 0259-2021-0009.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию 24.11.2020. После доработки 15.01.2021. Принята к публикации 15.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

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

В.В. Коврижных^{1, 2}✉, З.С. Мустафин¹, З.З. Багаутдинова¹

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

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ vasilinakovr@gmail.com

Аннотация. Активный полярный транспорт гормона растений ауксина, осуществляемый его транспортерами, – ключевое звено в формировании и поддержании распределения ауксина, которое, в свою очередь, определяет морфогенез растения. Пластичность распределения ауксина в большой степени реализуется через молекулярно-генетическую регуляцию им экспрессии транспортеров семейства PIN-FORMED (PIN) белков. Регуляция ауксином экспрессии чувствительных к нему генов происходит через ARF-Aux/IAA-зависимый сигнальный путь. Однако неизвестно, какие ARF-Aux/IAA белки участвуют в регуляции ауксином экспрессии генов PIN. У *Arabidopsis thaliana* семейства белков PIN, ARF и Aux/IAA многочисленны, возможны различные комбинации представителей этих семейств в реализации сигнального пути, что создает сложность для понимания механизмов этого процесса. Использование данных высокопроизводительного секвенирования транскриптомов, индуцированных ауксином (RNA-Seq), делает возможным обнаружение генов-кандидатов, участвующих в регуляции экспрессии белков PIN. Мы разработали алгоритм метаанализа ауксин-индуцированных транскриптомов, с помощью которого отобрали гены, изменяющие свою экспрессию в ответе на ауксин вместе с PIN1, PIN3, PIN4, PIN7, и предсказали возможные регуляторы ARF-Aux/IAA сигнального пути для каждого из дифференциально экспрессирующихся PIN. Применяя сравнительный анализ, мы определили общие и специфичные аспекты в регуляторных контурах, исследуемых PIN. Реконструкция генных сетей и их оценка показали возможные взаимодействия между генами и послужили дополнительным подтверждением большинства сигнальных путей, полученных в метаанализе. С помощью комплексного подхода мы предсказали, что регуляция ауксином экспрессии PIN происходит через несколько ARF-Aux/IAA регуляторных контуров, опосредованных комбинацией ARF4, ARF10 и IAA4, IAA12, IAA17, IAA18 и IAA32. Часть из них являются специфичными при формировании ауксинового ответа с участием отдельных белков PIN, тогда как другие – общими для нескольких белков PIN. Разработанный алгоритм метаанализа можно применять для решения других задач поиска регуляторов экспрессии генов с привлечением полногеномных данных.

Ключевые слова: *Arabidopsis thaliana*; ауксин; PIN-FORMED; ауксин-регулируемые гены; метаанализ полногеномных данных; генные сети.

Для цитирования: Коврижных В.В., Мустафин З.С., Багаутдинова З.З. Поиск участников сигнального пути ауксина к его транспортерам PIN на основе метаанализа транскриптомов, индуцированных ауксином. *Вавиловский журнал генетики и селекции*. 2021;25(1):39-45. DOI 10.18699/VJ21.005

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

V.V. Kovrizhnykh^{1, 2}✉, Z.S. Mustafin¹, Z.Z. Bagautdinova¹

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

² Novosibirsk State University, Novosibirsk, Russia

✉ vasilinakovr@gmail.com

Abstract. Active polar transport of the plant hormone auxin carried out by its PIN transporters is a key link in the formation and maintenance of auxin distribution, which, in turn, determines plant morphogenesis. The plasticity of auxin distribution is largely realized through the molecular genetic regulation of the expression of its transporters belonging to the PIN-FORMED (PIN) protein family. Regulation of auxin-response genes occurs through the ARF-Aux/IAA signaling pathway. However, it is not known which ARF-Aux/IAA proteins are involved in the regulation of PIN gene expression by auxin. In *Arabidopsis thaliana*, the PIN, ARF, and Aux/IAA families contain a larger number of members; their various combinations are possible in realization of the signaling pathway, and this is a challenge for understanding the mechanisms of this process. The use of high-throughput sequencing data on auxin-induced transcriptomes makes it possible to identify candidate genes involved in the regulation of PIN expression. To address this problem, we created an approach for the meta-analysis of auxin-induced transcriptomes, which helped us select genes that change their expression during the auxin response together with PIN1, PIN3, PIN4 and PIN7. Possible regulators of ARF-Aux/IAA signal-

ing pathway for each of the PINs under study were identified, and so were the aspects of their regulatory circuits both common for groups of *PIN* genes and specific for each *PIN* gene. Reconstruction of gene networks and their analysis predicted possible interactions between genes and served as an additional confirmation of the pathways obtained in the meta-analysis. The approach developed can be used in the search for gene expression regulators in other genome-wide data.

Key words: *Arabidopsis thaliana*; auxin; PIN-FORMED; auxin-response genes; meta-analysis; gene network.

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

Введение

Ключевая роль ауксина в регуляции роста и развития растений – давно признанный факт (Mroue et al., 2018). Значительная часть ауксина синтезируется в апикальных меристемах надземной части растения и затем переносится к корню, обеспечивая там развитие боковых и придаточных корней, а также поддержание ниши ствольных клеток в меристеме главного корня. На клеточном уровне роль ауксина в физиологических процессах осуществляется за счет концентрационно-зависимого действия на скорость деления и удлинения клеток (Campanoni, Nick, 2005). Поэтому формирование и поддержание градиентов концентрации ауксина в тканях имеют решающую роль в морфогенезе. Например, в экспериментах по отсечению кончика корня было показано, что через несколько часов может вновь сформироваться распределение ауксина с максимумом, отстоящим от нового кончика корня на определенном расстоянии (Grieneisen et al., 2007; Mironova et al., 2010). При этом регенерация меристемы и нормальное функционирование корня происходят только после восстановления паттерна распределения ауксина (Xu et al., 2006).

Гены семейства *PIN-FORMED* (*PIN*), которые у *Arabidopsis thaliana* кодируют восемь трансмембранных белков-транспортёров, осуществляют отток ауксина из клетки (Weijers et al., 2001; Petrasek, 2006). Транспортёры *PIN1-4*, *PIN7* локализируются неравномерно (полярно) на плазматической мембране клетки, за счет чего в ткани формируются направленные потоки ауксина. Так, в результате объединения потоков ауксина на уровне отдельных клеток в кончике корня *A. thaliana* формируется распределение этого гормона с максимумом в покоящемся центре, определяющем поддержание ниши ствольных клеток в корне (Ferguson, Friml, 2008). В большинстве случаев функция белков *PIN* является основополагающей в формировании и поддержании распределения ауксина. Известно, что существует сложная сеть регуляции ауксином экспрессии генов *PIN*, которая включает положительные и отрицательные обратные связи (Geldner et al., 2001; Friml, 2004; Sauer et al., 2006; Vieten et al., 2007). В работе A. Vieten с коллегами (2005) экспериментально показано, что обработка экзогенным ауксином приводит к увеличению транскрипции всех генов *PIN* (*PIN1-PIN4*, *PIN6*, *PIN7*) в корне, и оптимальная концентрация ауксина для максимального повышения уровня различается у каждого из этих генов (Vieten et al., 2005). Позднее мы сообщили, что транскрипционная и посттранскрипционная регуляция экспрессии *PIN1* ауксином имеют свои особенности (Omelyanchuk et al., 2016). На транскрипционном уровне повышение экспрессии *PIN1* происходит в широком диа-

пазоне концентраций экзогенно применяемого ауксина, в то время как уровень белка *PIN1* изменяется нелинейно, повышаясь с ростом концентрации ауксина при низких дозах гормона и понижаясь при увеличении концентрации ауксина в районе его высоких доз.

Основной механизм регуляции ауксином экспрессии чувствительных к нему генов происходит через ARF-Aux/IAA-зависимый сигнальный путь (Ulmasov et al., 1997). В отсутствие ауксина транскрипционные факторы ARF связаны корепрессорами Aux/IAA. При поступлении в клетку ауксин взаимодействует с рецептором TIR1, который формирует убиквитин-лигазный комплекс SCF^{TIR1} вместе с другими белками (Dharmasiri et al., 2005; Kepinski, Leyser, 2005). Далее этот комплекс связывается с белками Aux/IAA, регулируя их деградацию в протеасоме 26S (Calderon-Villalobos et al., 2010; Hayashi, 2012). Таким образом, транскрипционные факторы ARF начинают функционировать и активировать или подавлять транскрипцию генов ответа на ауксин. В геноме *A. thaliana* найдено 29 генов *Aux/IAA* и 23 *ARF*, их экспрессия в разных типах клеток отличается, создавая достаточную молекулярную сложность для обеспечения множества различных ответов на ауксин (Remington et al., 2004; Teale et al., 2006). При этом не установлено, какие именно ARF-Aux/IAA белки участвуют в регуляции экспрессии генов *PIN* ауксином. Известно лишь, что в промоторах всех генов *PIN* биоинформатическими методами были обнаружены сайты связывания транскрипционных факторов ARF (Habets, Offringa, 2014).

Реконструкция сигнального пути ауксина к его *PIN* транспортёрам имеет сложности для прямого решения экспериментальными методами. В связи с этим мы провели метаанализ полногеномных данных транскриптомов, индуцированных ауксином, с целью получить список наиболее вероятных регуляторов экспрессии генов *PIN* ауксином. Мы разработали алгоритм метаанализа комбинированных полногеномных данных и определили списки генов, значимо меняющих экспрессию совместно с генами *PIN* в ответ на ауксин. Комплексный подход, включающий сравнительный анализ этих списков и на их основе реконструированных генных сетей, предсказал участников ARF-Aux/IAA сигнального пути, вовлеченных в регуляцию экспрессии *PIN* ауксином. Так, общие сигнальные пути для *PIN1*, *PIN3*, *PIN7* опосредованы комбинацией *ARF4* с *IAA12* и *IAA18*. В то время как специфичная регуляция ауксином экспрессии отдельных генов *PIN*, вероятно, осуществляется другими белками ARF-Aux/IAA сигнального пути. Например, *ARF10* и *IAA32*, по результатам нашего анализа, присутствовали только в списках генов, значимо меняющих экспрессию

вместе с *PIN4*. Кроме того, в списках генов-кандидатов мы отметили дифференциальную экспрессию генов, которые вовлечены в посттранскрипционную регуляцию активности PIN.

Материалы и методы

Информация, используемая в метаанализе. Для исследования были взяты общедоступные данные по транскриптомам (микрочипам и секвенированию РНК) *A. thaliana*, индуцированным ауксином, большая часть которых использована ранее в работе (Cherenkov et al., 2018). Сводная таблица была дополнена данными из статьи (Omelyanchuk et al., 2017). В итоге для метаанализа мы взяли результаты 22 экспериментов. Гены определяли как дифференциально экспрессирующиеся гены (ДЭГ), если критерий *p* (по Бенджамини–Хохбергу) был меньше 0.05. На основе этих данных в соответствии с разработанным нами алгоритмом (см. раздел «Результаты. Алгоритм метаанализа») для каждого *PIN* были выделены свои наборы экспериментов (Приложение 1)¹.

Работа со сводной таблицей и списками проведена в программе Excel с помощью стандартных инструментов (фильтры, условное форматирование).

Реконструкция генных сетей. Генные сети на основе списков дифференциально экспрессирующихся генов были реконструированы с помощью ресурса String (<https://string-db.org/>) (Szklarczyk et al., 2019). Этот ресурс строит генные сети, используя заданные пользователем критерии, объединяя заданные гены по следующим типам связей: *experimentally determined* (экспериментальные данные, такие как, например, аффинная хроматография), *databases* (связь, полученная из записей из различных баз данных), *textmining* (упоминание генов/белков вместе в публикациях), *co-expression* (связь, полученная из сходства паттернов экспрессии мРНК), *neighborhood* (вычисляется на основании близости расстояния между генами в различных геномах), *gene fusion* (выявление при сравнении геномов гибридных генов, образованных в ходе эволюции из ранее независимых генов в результате хромосомных перестроек), *co-occurrence* (определение при сравнении геномов совместного переноса, потери или дубликации генов в эволюции, что может указывать на участие в общей функции), *protein homology* (гомология белков). Каждая связь обладает собственным значением достоверности, вычисленной посредством алгоритмов String.

Результаты

Алгоритм метаанализа

Этап 1: сбор данных. Формируем сводную таблицу из всех общедоступных микрочип-экспериментов и данных по РНК секвенированию на интересующую тему. В нашем случае это информация о дифференциальной экспрессии генов *A. thaliana* в ответ на обработку экзогенным ауксином. Собранные данные могут быть разнородными, например в 22 экспериментах, используемых нами в метаанализе, содержится два типа образцов (корень, весь проросток), три стадии развития (3-, 5–7-, 10–12-дневные

проростки), пять временных отрезков обработки (0.5–1 ч, 2–4, 6–8, 12, 24 ч), шесть типов ауксина и концентраций (0.1; 1; 5; 10 мкМ ИУК; 10 мкМ НУК; 10 мкМ ИБК).

Этап 2: отбор экспериментов, соответствующих задаче. Находим в сводной таблице, полученной на этапе 1, в каких экспериментах происходило изменение экспрессии генов, для которых мы ищем регуляторы. Согласно нашей задаче, известно, что у *A. thaliana* есть восемь транспортеров PIN. Дифференциально экспрессирующимися в данных общедоступных транскриптомах, индуцированных ауксином, мы обнаружили *PIN1* (в пяти экспериментах), *PIN3* (в восьми экспериментах), *PIN4* (в одном эксперименте) и *PIN7* (в шести экспериментах).

Этап 3: определение генов, меняющих свою экспрессию под действием ауксина вместе с генами PIN. Отдельно для каждого *PIN* мы отбирали только те ДЭГ, которые изменялись исключительно в экспериментах, где этот *PIN* меняет экспрессию, а в других экспериментах экспрессия этих генов была прежней. Таким образом, мы выявляем гены, потенциально участвующие в регуляции ауксином экспрессии генов *PIN*. Здесь также могут оказаться гены – непосредственные мишени изменения градиента ауксина под действием белков PIN. Для каждого исследуемого гена *PIN* формируется таблица, в которой отмечены гены, изменяющие свою экспрессию совместно с ним хотя бы в одном эксперименте, а также содержится информация о направлении изменения экспрессии каждого гена (активация или подавление) под воздействием исследуемого фактора.

Этап 4: формирование списков ДЭГ, значимо изменяющих экспрессию совместно с PIN. Мы применили биномиальный тест для определения, в скольких экспериментах ДЭГ должен присутствовать, чтобы считать совместное с геном *PIN* изменение экспрессии значимым с вероятностью более 95 %. Для разных списков генов пороговые величины значимости отличаются в соответствии с разным числом экспериментов, в которых определенный *PIN* дифференциально экспрессируется (см. этап 2). В нашем случае для *PIN3* ДЭГ считается значимым, если меняет свою экспрессию совместно в трех и более экспериментах, для *PIN1* и *PIN7* – в двух и более. Поскольку изменение экспрессии *PIN4* происходит в одном эксперименте, то список из этапа 3 в его случае не изменится.

Этап 5: определение общих и специфичных групп генов. Сравнивая между собой списки ДЭГ из предыдущего этапа, выделяем гены, встречающиеся в нескольких списках, т.е. общие для нескольких *PIN*, а также отмечаем гены, обнаруженные только в одном списке, тем самым определяя гены, специфично изменяющие экспрессию совместно с определенным *PIN*.

Этап 6: построение генных сетей. Используя подготовленные списки дифференциально-экспрессирующихся генов из этапа 4, реконструируем сеть взаимодействий между всеми генами. Наличие связности в этой сети отображает набор генов, для которых уже было найдено одно из доступных в базе String взаимодействий (*textmining*, *co-expression*, *co-occurrence* и т.д.).

Этап 7: анализ состава генных сетей. В первую очередь выделяли гены, для которых в String найдены связи к исследуемым генам, обращая внимание, на основе ка-

¹ Приложения 1–3 см. по адресу:

<http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx1.pdf>

ких данных определено взаимодействие. Затем в списке генов онтологий выбираем биологические процессы, имеющие отношение к исследуемому вопросу. В нашем исследовании мы выбирали активируемый ауксином сигнальный путь.

Применив описанный выше алгоритм метаанализа данных, мы получили несколько генов-кандидатов, с высокой долей вероятности осуществляющих регуляцию экспрессии исследуемых генов *PIN*. Далее описываем результаты для задачи реконструкции сигнального пути ауксина к его транспортерам *PIN*.

Метаанализ транскриптомов, индуцированных ауксином

Изначально в собранных транскриптомах, индуцированных ауксином, было более 20 тыс. ДЭГ, изменяющих экспрессию в ответ на обработку ауксином, среди этих генов было четыре представителя из семейства *PIN*: *PIN1*, *PIN3*, *PIN4*, *PIN7*. Выполнив описанный выше алгоритм метаанализа, мы получили четыре списка ДЭГ, совместно изменяющих экспрессию с *PIN1*, *PIN3*, *PIN4*, *PIN7* соответственно (Приложение 2). Суммарно значимо повышал экспрессию 531 ген, 236 генов снижали экспрессию (рис. 1). Совместно с *PIN1* значимо изменяли экспрессию 378 генов, 375 из которых так же, как *PIN1*, повысили уровень экспрессии в ответ на обработку ауксином. Для остальных генов *PIN* разница в числе подавляемых и активируемых ауксином потенциальных регуляторов была не такой резкой.

Затем, сравнивая списки между собой, мы определили ДЭГ, общие для нескольких *PIN* и специфичные для каждого гена *PIN*. Получено 12 групп генов: для каждого *PIN* были найдены специфичные активируемые ауксином гены и специфичные подавляемые, а также две группы генов, активируемых ауксином, общие для групп (*PIN1*, *PIN3*, *PIN7*) и (*PIN1*, *PIN7*), две группы генов, подавляемых ауксином, общие для (*PIN3*, *PIN7*) и (*PIN1*, *PIN3*). Активируемые и подавляемые потенциальные регуляторы *PIN4* не пересекаются с таковыми для других *PIN*. Поскольку среди потенциальных генов-регуляторов активности *PIN* присутствовали участники сигнального пути ответа на ауксин, то мы провели их поиск в списках (см. Приложение 2) и оценили, к каким из 12 групп ДЭГ, описанных выше, они принадлежат.

Предсказание ауксин-зависимых регуляторов экспрессии генов *PIN*

Поскольку метаанализ направлен на предсказание ауксин-зависимых регуляторов экспрессии генов *PIN*, в списках ДЭГ мы выделили гены транскрипционной и посттранскрипционной регуляции. Поиск возможных транскрипционных регуляторов мы ограничили транскрипционными факторами ARF и белками IAA. Возможные посттранскрипционные регуляторы определяли среди членов известных семейств белков, влияющих на локализацию белков *PIN* на мембране клетки.

Возможные регуляторы экспрессии *PIN* на транскрипционном уровне. В результате метаанализа мы нашли, что общими потенциальными регуляторами для (*PIN1*, *PIN3*, *PIN7*) являются *ARF4* и *IAA12*, *IAA18*. Спе-

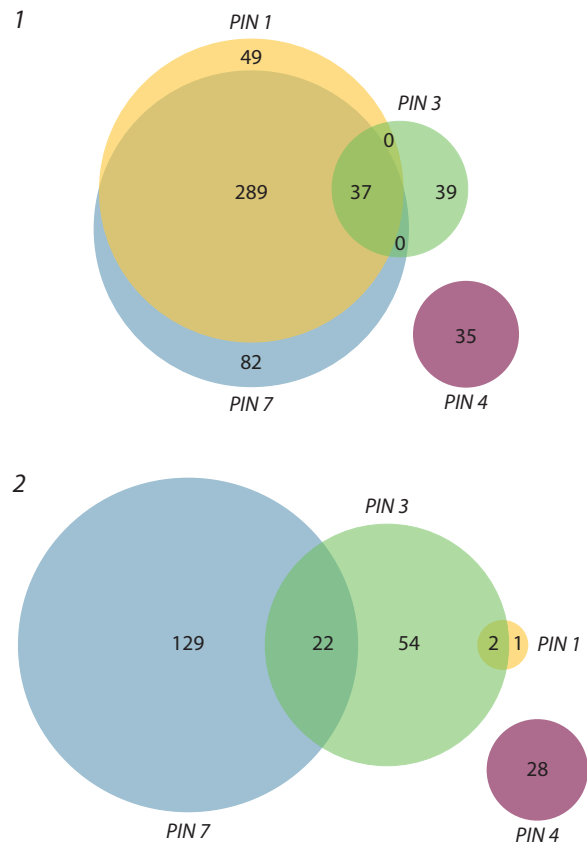


Рис. 1. Выявленные в метаанализе 12 групп генов, значимо меняющих свою экспрессию совместно с *PIN1*, *PIN3*, *PIN4*, *PIN7*, экспрессия которых активируется (1) и подавляется (2) в ответ на ауксин.

цифичными были определены *IAA4* для *PIN1*, *ARF10* и *IAA32* для *PIN4*. Кроме того, *IAA17* был найден в группе генов, изменяющих свою экспрессию с *PIN1* и *PIN7*. Интересно, что среди специфичных регуляторов *PIN3* и *PIN7* мы не обнаружили транскрипционные факторы, относящиеся к семейству Aux/IAA, но нашли регуляторы, принадлежащие к другим семействам транскрипционных факторов. Следовательно, очевидны различия в наборах ARF-Aux/IAA для исследуемых генов *PIN*, что может также обуславливать различия в дозозависимой регуляции ауксином.

Возможные регуляторы полярной локализации белков *PIN*. Согласно опубликованным данным, белки *PIN* циркулируют между мембраной и цитоплазмой в везикулах. Этот процесс регулируется белками BIG, GN, ARF1, семейством киназ AGC, PID, функционирование которых находится под контролем ауксина (Dhonukshe, 2011). На полярную локализацию белков *PIN* влияют также такие ключевые регуляторы, как ABCB1 и ABCB19, семейства ROPGEF (Pan et al., 2015). В ходе метаанализа данных среди дифференциально экспрессирующихся генов в ответ на обработку ауксином мы обнаружили снижение экспрессии *BIG4* и *ROPGEF11* в списках генов, изменяющих свою экспрессию вместе с *PIN7* и *PIN4* соответственно. Повышение экспрессии было отмечено для *WAG2* (семейство киназ AGC) в группе генов, изменяющих свою экспрессию вместе с *PIN1* и *PIN7*.

Кроме того, на наш взгляд, интересно, что сигнальный пептид из RGF/GLV/CLE семейства *RGF6/GLV1/CLEL6* повышает свою активность в ответе на ауксин в экспериментах, где увеличивается активность *PIN1* и *PIN7*, а другой сигнальный пептид из этого семейства, *RGF8/GLV6/CLEL2*, – в экспериментах, где повышалась экспрессия только *PIN7*.

Таким образом, формирование реакций в ответ на ауксин для (*PIN1*, *PIN3*, *PIN7*) группы обусловлено двумя общими сигнальными путями – *ARF4* с *IAA12* или *IAA18*. Дополнительно существуют специфичные ARF-Аух/IAA пути для *PIN1* и *PIN4*. Также среди известных ауксинчувствительных генов, влияющих на экспрессию *PIN*, мы обнаружили подавление экспрессии *BIG4* и *ROPGEF11*, что, вероятно, вносит вклад в специфичные ответы *PIN7* и *PIN4* соответственно.

Реконструкция генных сетей

Мы использовали списки ДЭГ для каждого гена *PIN* и реконструировали генные сети, что позволило оценить, для каких ДЭГ описано взаимодействие и, главное, как все эти ДЭГ могут влиять на активность генов *PIN*. В результате связанные сети, в которых были найдены взаимодействия с генами *PIN*, получились только для *PIN1*, *PIN3* и *PIN7*. Метаанализ, на основе которого были сделаны списки генов для генных сетей, сам по себе обеспечивает значимость, поэтому мы использовали порог для построения связей 0.4. Поскольку нас интересует реконструкция сигнального пути ауксина, в String мы отметили только этот биологический процесс. Примечательно, что большинство связей образовано на основе автоматического анализа текстов статей. В генной сети, реконструированной на основе ДЭГ, меняющих свою экспрессию вместе с *PIN1*, было найдено 12 генов, относящихся к активации сигнального пути ауксина (Приложение 3). При этом напрямую с *PIN1* были связаны *IAA12*, *IAA17* (*AXR3*), *WAG2*, *AUX1*, остальные гены ответа на ауксин были связаны с *PIN1* опосредованно (рис. 2). Можно отметить также, что напрямую с *PIN1* были связаны *AIL6/PLT3* и *AVP1*, относящиеся к процессам развития органов арабидопсиса, регулируемых ауксином (Krizek, 2011). Эти гены можно отнести к генам, являющимся непосредственными мишенями изменения градиента ауксина под действием белков PIN. Среди этих генов связи *PIN1* с *AIL6* и *WAG2* построены на основе данных о коэкспрессии в полногеномных экспериментах РНК секвенирования.

В генной сети, реконструированной для ДЭГ, изменяющих экспрессию вместе с *PIN3*, было восемь генов, традиционно относящихся к сигнальному пути ауксина (см. Приложение 3). Прямые связи с *PIN3* были найдены для *AUX1* и *IAA12*, *SAUR9*. В генной сети для *PIN7* 14 генов относились к традиционному сигнальному пути ауксина. При этом *PIN7* напрямую взаимодействует с *IAA12*, *IAA17* (*AXR3*), *AUX1*, *LRP1*, *WAG2* (см. Приложение 3). Кроме того, у *PIN7* были прямые связи с *ABCG33*, *NFA6*, *PHOT1*, *YUC2*, *YUC6*, относящимися к другим биологическим процессам, контролируемым ауксином. Реконструкция генных сетей – дополнительное подтверждение того, что регуляция ауксином экспрессии генов *PIN*, вероятно, происходит с участием *IAA12*,

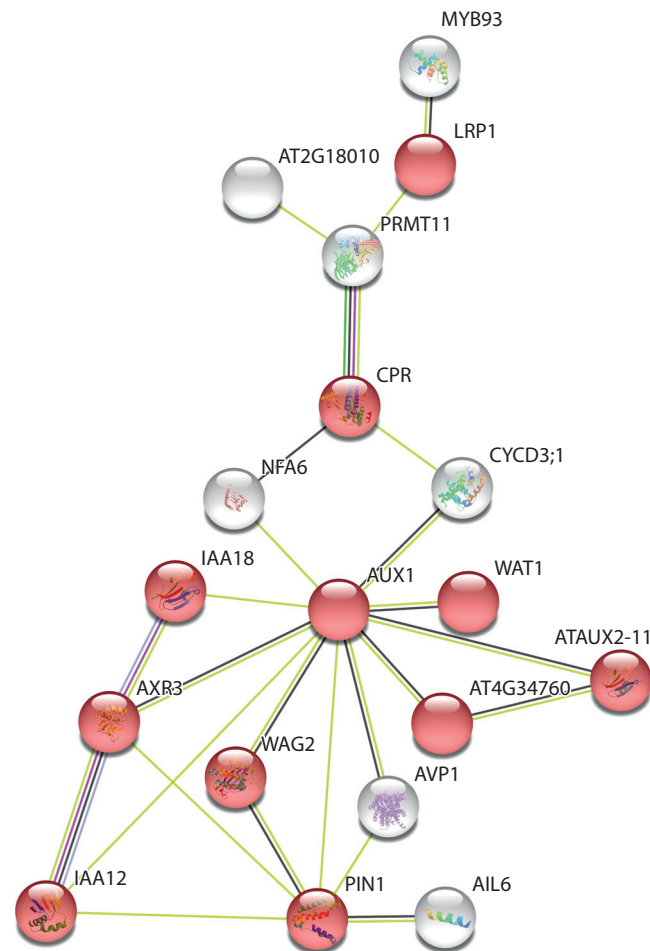


Рис. 2. Фрагмент генной сети, содержащий гены, связанные с *PIN1*, и гены сигнального пути ауксина.

Красными кругами обозначены гены, традиционно относящиеся к сигнальному пути ауксина, серыми – гены, выявленные в метаанализе, для которых в String были найдены прямые или опосредованные связи к *PIN1*. Цвет связи отражает, на основе каких данных из String построено взаимодействие. Желтым цветом показаны связи, построенные исходя из анализа текстов статей, черным – по данным о коэкспрессии, голубым – по принципу гомологии белков, розовым – на основе данных о белок-белковом взаимодействии.

IAA17. При этом следует отметить, что отсутствие прямых связей с *PIN* для остальных предсказанных метаанализом регуляторов ARF-Аух/IAA сигнального пути не исключает их из списка кандидатов для экспериментальной проверки в будущем.

Обсуждение

Фитогормоны активно участвуют в процессах роста и морфогенеза растений. Действие ауксина в этих процессах хорошо изучено и основано на изменении распределения ауксина в тканях (Mroque et al., 2018). Следовательно, концентрация ауксина в клетке – лимитирующий фактор в определении ее судьбы. Белки-транспортеры семейства PIN играют важную роль в реализации морфогенетического действия ауксина, так как создают направленные потоки этого гормона в тканях и, таким образом, опосредуют формирование градиентов концентрации ауксина (Vanneste, Friml, 2009).

Важным аспектом в описанном выше процессе является наличие положительных и отрицательных обратных связей во взаимной регуляции оттока ауксина из клетки посредством функционирования белков PIN и количества этих транспортеров, контролируемого ауксином. Регуляция экспрессии чувствительных к ауксину генов опосредована двумя семействами белков. Первое семейство – транскрипционные факторы ARF, которые в промоторе чувствительного к ауксину гена связываются с сайтом AuxRe и выступают в роли активатора или репрессора экспрессии (Ulmasov et al., 1997). В некоторых источниках предположительно активаторами экспрессии считаются только ARF5-ARF8, ARF19, однако экспериментального подтверждения этому нет (Guilfoyle, Hagen, 2007). Второе – корепрессоры Aux/IAA, которые в отсутствие ауксина связаны с ARF.

Ранее были сообщения о подавлении экспрессии PIN1–4, PIN7 в мутантах *axr3/iaa17* и *solitary-root-1(slr-1)/iaa14* (Vieten et al., 2005) и о регуляции экспрессии PIN1 транскрипционным фактором ARF5 (Wenzel et al., 2007), который взаимодействует с IAA12 (Hamann, 2002). В настоящей работе, используя компьютерные методы метаанализа полногеномных данных и реконструкции генных сетей, мы предсказали детали сигнального пути ауксина к его транспортерам PIN. Результаты говорят о том, что для регуляции ауксином транскрипции *PIN1*, *PIN3*, *PIN7* и *PIN1*, *PIN7* существуют общие механизмы, а также есть специфичные механизмы регуляции ауксином экспрессии *PIN*. Общим механизмом для *PIN1*, *PIN3*, *PIN7* мы предсказываем активацию их экспрессии через ARF4-IAA12, ARF4-IAA18, а для *PIN1* и *PIN7* – дополнительно через ARF4-IAA17. Специфичные механизмы осуществляются через ARF4-IAA4 и ARF10-IAA32 для *PIN1* и *PIN4* соответственно. Взаимодействия между указанными ARF и IAA имеют экспериментальные подтверждения (Рапонов et al., 2008). Недавно было показано, что засоленность снижает экспрессию генов *PIN* и приводит к стабилизации IAA17 (Liu et al., 2015). Причем этот вид стресса вызывает уменьшение размера апикальной меристемы корня из-за снижения накопления ауксина, опосредованного падением уровня экспрессии *PIN1*, *PIN3*, *PIN7*. В наших данных в транскриптомах, индуцированных ауксином, повышению экспрессии *PIN1* и *PIN7* сопутствует увеличение экспрессии *IAA17*.

Для сигнальных пептидов семейства RGF/GLV/CLEL ранее было отмечено, что при гравитропизме они меняют градиент ауксина в гипокотиле и корне (Whitford et al., 2012). В корне это происходит за счет регуляции пептидами этого семейства локализации белков PIN2. При этом показано, что пептиды GLV3 и, возможно, GLV6 и GLV9 секретируются из кортекса и эндодермиса и проходят во внешние слои для регуляции локализации PIN2. Пептид GLV1 не экспрессируется в корне, но есть в гипокотиле, где также меняет градиент ауксина при гравитропизме, как при сверхэкспрессии, так и при потере функции при мутации (Whitford et al., 2012). И, судя по нашим данным, пептиды RGF/GLV/CLEL участвуют в сигнальном пути, регулирующем локализацию на мембране PIN1 и PIN7 транспортеров, и, возможно, опосредованно влияют на увеличение экспрессии этих генов *PIN*. Сверхэкспрессия

или обработка GLV1 приводят к удлинению корня и его апикальной меристемы за счет того, что увеличивается зона клеточных делений в корне, т. е. клетки позже переходят к дифференцировке (Fernandez et al., 2013). Этот переход также связан с изменением распределения ауксина, которое формируется его транспортерами.

Заключение

Таким образом, разработанный алгоритм метаанализа полногеномных данных был применен к задаче поиска участников и реконструкции сигнального пути ауксина к его транспортерам PIN. Нам удалось выявить, что ауксин контролирует экспрессию *PIN1*, *PIN3* и *PIN7* как через общие регуляторы, так и специфично, в то время как для *PIN4* были определены только специфичные регуляторы. Мы нашли опубликованные экспериментальные данные, которые частично подтверждают наши предположения. В результате проведенного компьютерного исследования нами выдвинуты новые кандидаты для экспериментальной проверки.

Список литературы / References

- Calderon-Villalobos L.L., Tan X., Zheng N., Estelle M. Auxin perception – structural insights. *Cold Spring Harb. Perspect. Biol.* 2010;2: a005546-a005546. DOI 10.1101/cshperspect.a005546.
- Campanoni P., Nick P. Auxin-dependent cell division and cell elongation. 1-Naphthaleneacetic acid and 2,4-dichlorophenoxyacetic acid activate different pathways. *Plant Physiol.* 2005;137:939-948. DOI 10.1104/pp.104.053843.
- Cherenkov P., Novikova D., Omelyanchuk N., Levitsky V., Grosse I., Weijers D., Mironova V. Diversity of cis-regulatory elements associated with auxin response in *Arabidopsis thaliana*. *J. Exp. Bot.* 2018; 69:329-339. DOI 10.1093/jxb/erx254.
- Dharmasiri N., Dharmasiri S., Estelle M. The F-box protein TIR1 is an auxin receptor. *Nature.* 2005;435:441-445. DOI 10.1038/nature 03543.
- Dhonukshe P. PIN polarity regulation by AGC-3 kinases and ARF-GEF. *Plant Signal. Behav.* 2011;6:1333-1337. DOI 10.4161/psb.6.9. 16611.
- Feraru E., Friml J. PIN polar targeting. *Plant Physiol.* 2008;147:1553-1559. DOI 10.1104/pp.108.121756.
- Fernandez A., Hilson P., Beeckman T. GOLVEN peptides as important regulatory signalling molecules of plant development. *J. Exp. Bot.* 2013;64:5263-5268. DOI 10.1093/jxb/ert248.
- Friml J. A PINOID-dependent binary switch in apical-basal PIN polar targeting directs auxin efflux. *Science.* 2004;306:862-865. DOI 10.1126/science.1100618.
- Geldner N., Friml J., Stierhof Y.-D., Jürgens G., Palme K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature.* 2001;413:425-428. DOI 10.1038/35096571.
- Grieneisen V.A., Xu J., Marée A.F.M., Hogeweg P., Scheres B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature.* 2007;449:1008-1013. DOI 10.1038/nature 06215.
- Guilfoyle T.J., Hagen G. Auxin response factors. *Curr. Opin. Plant Biol.* 2007;10:453-460. DOI 10.1016/j.pbi.2007.08.014.
- Habets M.E.J., Offringa R. PIN-driven polar auxin transport in plant developmental plasticity: a key target for environmental and endogenous signals. *New Phytol.* 2014;203:362-377. DOI 10.1111/nph. 12831.
- Hamann T. The *Arabidopsis* *BODENLOS* gene encodes an auxin response protein inhibiting MONOPTEROS-mediated embryo patterning. *Genes Dev.* 2002;16:1610-1615. DOI 10.1101/gad.229402.
- Hayashi K. The interaction and integration of auxin signaling components. *Plant Cell Physiol.* 2012;53:965-975. DOI 10.1093/pcp/ pcs035.

- Kepinski S., Leyser O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature*. 2005;435:446-451. DOI 10.1038/nature03542.
- Krizek B.A. Auxin regulation of *Arabidopsis* flower development involves members of the AINTEGUMENTA-LIKE/PLETHORA (AIL/PLT) family. *J. Exp. Bot.* 2011;62:3311-3319. DOI 10.1093/jxb/err127.
- Liu W., Li R.-J., Han T.-T., Cai W., Fu Z.-W., Lu Y.-T. Salt stress reduces root meristem size by nitric oxide-mediated modulation of auxin accumulation and signaling in *Arabidopsis*. *Plant Physiol.* 2015;168:343-356. DOI 10.1104/pp.15.00030.
- Mironova V.V., Omelyanchuk N.A., Yosiphon G., Fadeev S.I., Kolchanov N.A., Mjolsness E., Likhoshvai V.A. A plausible mechanism for auxin patterning along the developing root. *BMC Syst. Biol.* 2010; 4:98. DOI 10.1186/1752-0509-4-98.
- Mroue S., Simeunovic A., Robert H.S. Auxin production as an integrator of environmental cues for developmental growth regulation. *J. Exp. Bot.* 2018;69:201-212. DOI 10.1093/jxb/erx259.
- Omelyanchuk N.A., Kovrizhnykh V.V., Oshchepkova E.A., Pasternak T., Palme K., Mironova V.V. A detailed expression map of the PIN1 auxin transporter in *Arabidopsis thaliana* root. *BMC Plant Biol.* 2016;16:5. DOI 10.1186/s12870-015-0685-0.
- Omelyanchuk N.A., Wiebe D.S., Novikova D.D., Levitsky V.G., Klimova N., Gorelova V., Weinholdt C., Vasiliev G.V., Zemlyanskaya E.V., Kolchanov N.A., Kochetov A.V., Grosse I., Mironova V.V. Auxin regulates functional gene groups in a fold-change-specific manner in *Arabidopsis thaliana* roots. *Sci. Rep.* 2017;7:2489. DOI 10.1038/s41598-017-02476-8.
- Pan X., Chen J., Yang Z. Auxin regulation of cell polarity in plants. *Curr. Opin. Plant Biol.* 2015;28:144-153. DOI 10.1016/j.pbi.2015.10.009.
- Paponov I.A., Paponov M., Teale W., Menges M., Chakrabortee S., Murray J.A.H., Palme K. Comprehensive transcriptome analysis of auxin responses in *Arabidopsis*. *Mol. Plant.* 2008;1:321-337. DOI 10.1093/mp/ssm021.
- Petrasek J. PIN proteins perform a rate-limiting function in cellular auxin efflux. *Science*. 2006;312:914-918. DOI 10.1126/science.1123542.
- Remington D.L., Vision T.J., Guilfoyle T.J., Reed J.W. Contrasting modes of diversification in the *Aux/IAA* and *ARF* gene families. *Plant Physiol.* 2004;135:1738-1752. DOI 10.1104/pp.104.039669.
- Sauer M., Balla J., Luschnig C., Wisniewska J., Reinohl V., Friml J., Benkova E. Canalization of auxin flow by Aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes Dev.* 2006;20:2902-2911. DOI 10.1101/gad.390806.
- Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J., von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47: D607-D613. DOI 10.1093/nar/gky1131.
- Teale W.D., Paponov I.A., Palme K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* 2006;7:847-859. DOI 10.1038/nrm2020.
- Ulmasov T., Murfett J., Hagen G., Guilfoyle T.J. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell.* 1997;9:1963-1971. DOI 10.1105/tpc.9.11.1963.
- Vanneste S., Friml J. Auxin: a trigger for change in plant development. *Cell.* 2009;136:1005-1016. DOI 10.1016/j.cell.2009.03.001.
- Vieten A., Sauer M., Brewer P.B., Friml J. Molecular and cellular aspects of auxin-transport-mediated development. *Trends Plant Sci.* 2007;12:160-168. DOI 10.1016/j.tplants.2007.03.006.
- Vieten A., Vanneste S., Wisniewska J., Benkova E., Benjamins R., Beeckman T., Luschnig C., Friml J. Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression. *Development.* 2005;132:4521-4531. DOI 10.1242/dev.02027.
- Weijers D., Franke-van Dijk M., Vencken R.J., Quint A., Hooykaas P., Offringa R. An *Arabidopsis* Minute-like phenotype caused by a semi-dominant mutation in a RIBOSOMAL PROTEIN S5 gene. *Development.* 2001;128:4289-4299.
- Wenzel C.L., Schuetz M., Yu Q., Mattsson J. Dynamics of MONOPTEROS and PIN-FORMED1 expression during leaf vein pattern formation in *Arabidopsis thaliana*. *Plant J.* 2007;49:387-398. DOI 10.1111/j.1365-3113X.2006.02977.x.
- Whitford R., Fernandez A., Tejos R., Pérez A.C., Kleine-Vehn J., Vanneste S., Drozdzecki A., Leitner J., Abas L., Aerts M., Hoogewijs K., Baster P., De Groot R., Lin Y.-C., Storme V., Van de Peer Y., Beeckman T., Madder A., Devreese B., Luschnig C., Friml J., Hilson P. GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. *Dev. Cell.* 2012;22:678-685. DOI 10.1016/j.devcel.2012.02.002.
- Xu J., Hofhuis H., Heidstra R., Sauer M., Friml J., Scheres B. A molecular framework for plant regeneration. *Science*. 2006;311:385-388. DOI 10.1126/science.1121790.

ORCID ID

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

Благодарности. Работа поддержана бюджетным проектом № 0259-2021-0009 и грантом Президента РФ МК-3470.2021.1.4.

Прозрачность финансовой деятельности. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 24.10.2020. После доработки 12.01.2021. Принята к публикации 14.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

Филостратиграфический анализ генных сетей заболеваний человека

З.С. Мустафин¹✉, С.А. Лашин^{1,2}, Ю.Г. Матушкин¹

¹ Федеральное исследовательское учреждение Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ mustafinzs@bionet.nsc.ru

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

Ключевые слова: эволюция; филостратиграфия; ортолог; генная сеть; возраст гена.

Для цитирования: Мустафин З.С., Лашин С.А., Матушкин Ю.Г. Филостратиграфический анализ генных сетей заболеваний человека. *Вавиловский журнал генетики и селекции*. 2021;25(1):46-56. DOI 10.18699/VJ21.006

Phylostratigraphic analysis of gene networks of human diseases

Z.S. Mustafin¹✉, S.A. Lashin^{1,2}, Yu.G. Matushkin¹

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

² Novosibirsk State University, Novosibirsk, Russia

✉ mustafinzs@bionet.nsc.ru

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

Key words: evolution; phylostratigraphic analysis; ortholog; gene network; gene age.

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

Введение

Исследование ключевых факторов, влияющих на развитие и протекание заболеваний, – одно из важнейших направлений как для медицины, так и для биологии (Степанов, 2016). Как известно, формирование фенотипических признаков, обеспечивающих адаптацию организмов к условиям окружающей среды, контролируется не отдельными генами, а генными сетями – группами координированно функционирующих генов и продуктов их работы (РНК, белками, метаболитами и др.) (Колчанов и др., 2013). Возникает задача выделения ключевых структурных особенностей сетей, элементов сетей, а также их численного описания. Одной из важных характеристик является возраст гена. Возраст гена, принадлежащего к ортологической группе, определяется как этап возникновения наиболее недавнего предка всех видов, представленных в этой группе (Liebeskind et al., 2016).

Современные методы анализа дают возможность оценить эволюционные характеристики генов, в частности филостратиграфический анализ – методология, предложенная в 2007 г. Т. Domazet-Lošo, – позволяет определить возраст гена с помощью специального индекса, получаемого в результате анализа ортологичных генов и сравнения положения организмов, чьи гены рассматриваются в анализе на филогенетическом дереве (Domazet-Lošo et al., 2007).

Для работы с генными сетями существует множество программных средств. Одни сконцентрированы на реконструкции сетей на основании данных из биологических баз, например String (Szklarczyk et al., 2019), GeneMANIA (Montejo et al., 2010). Другие имеют обширный функционал по визуализации элементов сети, выявлению ее структурных особенностей: Cytoscape (Shannon et al., 2003), yEd (<https://www.yworks.com/products/yed>). Программный комплекс Cytoscape выгодно отличается от других средств тем, что, помимо обширных возможностей по построению сети, компоновке и покраске ее элементов, анализу структурных особенностей, он позволяет пользователям писать собственные приложения на языке Java и встраивать их в Cytoscape в качестве плагинов. Это открывает сообществу возможность реализовывать весь интересующий функционал и добавлять его в Cytoscape. Например, такие известные средства, как String и GeneMANIA, способные реконструировать сеть по списку генов на основании извлечения взаимодействий из баз биологических данных, имеют свои собственные плагины в Cytoscape и позволяют пользоваться своей функциональностью, сочетая ее с возможностями Cytoscape и других его плагинов. Пользователю также становится доступным импорт готовых сетей, например из баз Pathway Commons (Cerami et al., 2011) или KEGG Pathway (Kanehisa et al., 2017), без необходимости разбора форматов представления сети в этих базах. Наконец, с учетом всех имеющихся возможностей любой пользователь может написать собственное приложение под свои задачи и поделиться им с сообществом.

В настоящей работе представлены результаты анализа генных сетей одним из таких плагинов, Orthoscape (Mustafin et al., 2017), способным проанализировать эволюционные особенности генов в генной сети. Продемонстрировано, что большинство генов, описанных в генных

сетях, подвержено стабилизирующему отбору, и обнаружена высокая достоверная корреляция между временем возникновения гена и наблюдаемым уровнем генетической изменчивости – чем моложе ген, тем выше уровень генетической изменчивости. Показано, что среди проанализированных генных сетей наибольшая доля эволюционно молодых генов выявлена в сетях, связанных с заболеваниями иммунной системы (65 %), а эволюционно древних – в сетях, ответственных за формирование зависимостей человека от веществ, вызывающих привыкание к химическим соединениям (88 %); генные сети, связанные с развитием инфекционных заболеваний, вызванных паразитами, достоверно обогащены эволюционно молодыми генами, а генные сети, ответственные за развитие специфических типов рака, – эволюционно древними генами.

Материалы и методы

Исходные данные для анализа. В работе использовали генные сети, представленные в базе KEGG Pathway, раздел Human Diseases. Сети в этом разделе разбиты на категории, всего таких категорий 11 (суммарно включающих 80 сетей): нейродегенеративные заболевания (neurodegenerative diseases, 5 сетей), сердечно-сосудистые заболевания (cardiovascular diseases, 5 сетей), заболевания, связанные с иммунной системой (immune diseases, 8 сетей), эндокринные заболевания и нарушения метаболизма (endocrine and metabolic diseases, 6 сетей), инфекционные заболевания, вызванные бактериями (infectious diseases: bacterial, 10 сетей), инфекционные заболевания, вызванные вирусами (infectious diseases: viral, 9 сетей), инфекционные заболевания, вызванные паразитами (infectious diseases: parasitic, 6 сетей), лекарственная устойчивость к противоопухолевым препаратам (drug resistance: antineoplastic, 4 сети), рак: обобщение (cancers: overview, 7 сетей), специфические типы рака (cancers: specific types, 15 сетей), зависимость от химических соединений, вызывающих привыкание (substance dependence, 5 сетей).

Необходимые данные для проведения анализа: списки ортологичных генов, нуклеотидные последовательности генов и аминокислотные последовательности кодируемых ими белков, доменный состав, информация о таксономических рядах организмов, чьи гены рассматривали в анализе, – также были взяты из базы KEGG.

Используемое программное обеспечение. Анализ проводили на базе программного комплекса Cytoscape (Shannon et al., 2003) – многофункционального средства для визуализации и анализа сетей. Для импорта сетей из KEGG Pathway в работе использовали плагин CyKEGGParser (Nersisyan et al., 2014). Для выполнения филостратиграфического анализа и анализа индекса эволюционной изменчивости был взят плагин Orthoscape (Mustafin et al., 2017).

Методы оценки эволюционных характеристик генов. Orthoscape позволяет оценить две эволюционные характеристики генов. Первая характеристика, вычисляемая с помощью Orthoscape, – *филостратиграфический индекс гена* (phylostratigraphic age index, PAI). Он показывает, в какой степени отдален от корня филогенетического дерева таксон, отражающий возраст гена, т. е. такой таксон, на

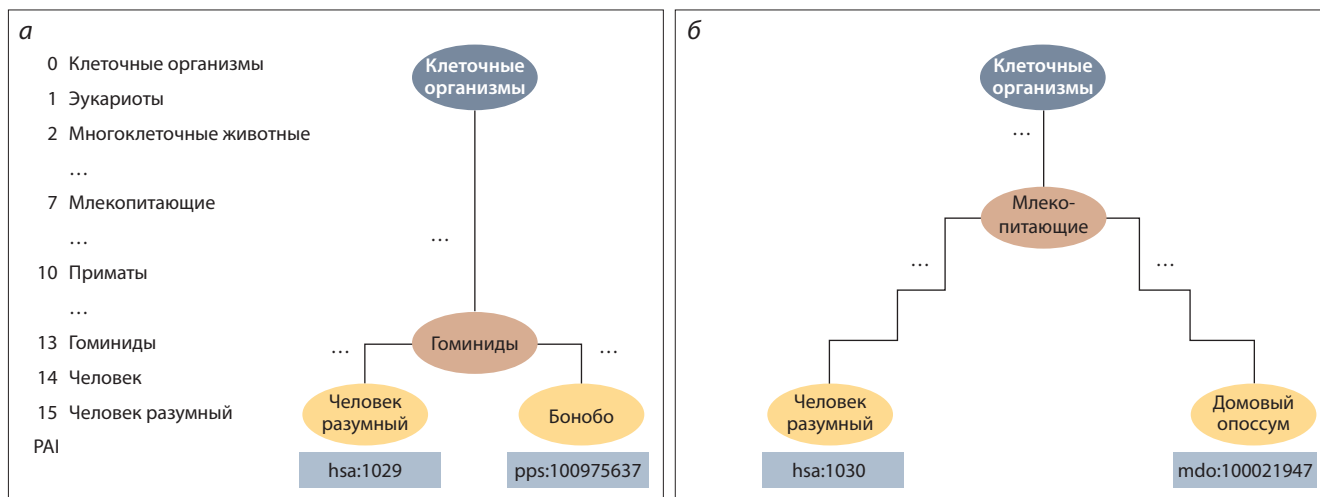


Рис. 1. Пример определения PAI для двух генов *Homo sapiens* (человек).

а – пример эволюционно молодого гена hsa:1029, наиболее отдаленным от исследуемого организмом, у которого был найден ортолог этого гена, является *Pan paniscus* (шимпанзе бонобо); *б* – пример эволюционно более древнего гена hsa:1030, наиболее отдаленный от исследуемого организмом, у которого был найден ортолог этого гена, – *Monodelphis domestica* (домовый опоссум). Можно заключить, что ген на примере (*а*) эволюционно моложе гена на примере (*б*). Шкала слева показывает индекс PAI, который соответствует глубине узла таксономического дерева (подробнее см. табл. 1).

котором произошло расхождение исследуемого вида с наиболее отдаленным родственником таксоном, в котором обнаружен ортолог рассматриваемого гена. Таким образом, чем больше PAI исследуемого гена, тем он моложе (рис. 1). Для расчета PAI в Orthoscape применяется сервис KEGG Orthology, что дает возможность учитывать среди всех гомологов гена именно ортологичные.

Таблица 1. Список таксонов, выделенных для филостратиграфического анализа генов *H. sapiens*

PAI	Таксон	Возраст, млн лет
0	Cellular organism (клеточные организмы, корень дерева)	4100 (Bell et al., 2015)
1	Eukaryota (эукариоты)	1850 (Leander, 2020)
2	Metazoa (многоклеточные животные)	665 (Maloof et al., 2010a)
3	Chordata (хордовые)	541 (Maloof et al., 2010b)
4	Craniata (плеченогие)	535 (Maloof et al., 2010b)
5	Vertebrata (позвоночные)	525 (Shu et al., 1999)
6	Euteleostomi (костные позвоночные)	420 (Diogo, 2007)
7	Mammalia (млекопитающие)	225 (Datta, 2005)
8	Eutheria (плацентарные)	160 (Luo et al., 2011)
9	Euarchontoglires (грызунообразные + зуархонты)	65 (Kumar et al., 2013)
10	Primates (приматы)	55 (Chatterjee et al., 2009)
11	Haplorrhini (обезьяны)	50 (Dunn et al., 2016)
12	Catarrhini (узконосые обезьяны)	44 (Harrison, 2013)
13	Hominidae (гоминиды)	17 (Hey, 2005)
14	Homo (люди)	2.8 (Schrenk et al., 2014)
15	Homo sapiens (человек разумный)	0.35 (Scerri et al., 2018)

Важная характеристика для филостратиграфического анализа – список таксономических единиц, описывающих этапы расхождения на эволюционном дереве организма, чьи гены исследуются с другими организмами, ортологи которых могут быть найдены. Полный список таксонов, использованный в анализе для определения филостратиграфического индекса генов *H. sapiens*, а также примерный эволюционный возраст этих таксонов в млн лет от нашего времени приведены в табл. 1. Следует отметить, что дискуссии на эту тему ведутся, в разных источниках указаны разные показатели образования того или иного таксона; значения в табл. 1 отражают примерные оценки.

Программа Orthoscape также позволяет оценить индекс эволюционной изменчивости гена (divergence index, DI). Он показывает тип отбора, которому подвержен ген. Индекс DI вычисляется на основании отношения dN/dS , где dN – доля несинонимичных замен в последовательностях исследуемого гена и его ортолога, т.е. таких замен, которые приводят к смене кодируемой данным триплетом аминокислоты; dS – доля синонимичных замен, т.е. не приводящих к замене кодируемой аминокислоты. Значение индекса в диапазоне от 0 до 1 свидетельствует о том, что ген подвержен стабилизирующему отбору, 1 – нейтральной эволюции, а больше 1 – движущему отбору. Анализ данного индекса имеет смысл только при сравнении близкородственных организмов, поскольку методика не дает учесть многократные замены в одной и той же позиции, которые неизбежно будут накоплены при сравнении с организмами, эволюционно отдаленными от исследуемого. Вычисление dN/dS проходит в два этапа: 1. Выравнивание исходных последовательностей рассматриваемого гена и ортологичного гена. Оно осуществляется с помощью алгоритма Нидлмана–Вунша, выравниваются аминокислотные последовательности с сохранением нуклеотидных триплетов, кодирующих аминокислоты. Затем позиции с разрывами удаляются.

2. Выравненные последовательности подаются на вход средству PAML (phylogenetic analysis by maximum likelihood) (Yang, 2007). Для вычисления dN/dS применяются методы, по-разному учитывающие позиции триплетов, их частоту встречаемости и прочие факторы. В PAML реализованы методы: Nei–Gojobori (Nei, Gojobori, 1986), Yang & Nielsen (Yang, Nielsen, 2000), LWL85 (Li, 1985), LWLm (Li, 1993), LPB93 (Pamilo, Bianchi, 1993). Для расчета DI мы использовали значение dN/dS , вычисленное по методу LPB93. Значение dN/dS рассчитывается для каждой пары ген-ортолог, итоговое значение DI определяется по формуле

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

где $dnds_i$ – значение dN/dS отношения для последовательности гена и ортолога i ; n – число ортологов, попавших в анализ.

Результаты и обсуждение

Анализ эволюционных характеристик генных сетей

С помощью Orthoscape были посчитаны индексы PAI и DI для всех генов, представленных в 80 проанализированных генных сетях из KEGG Pathway, Human Diseases. На основании этих данных были вычислены значения PAI для каждой генной сети (табл. 2) как среднее значение PAI всех генов, задействованных в сети, и PAI категории как среднее значение PAI всех сетей из этой категории. Аналогичным образом для каждой генной сети был определен индекс DI.

Среди проанализированных 80 сетей наблюдается варьирование PAI от 0.44 (т.е. большая часть генов эволюционно древняя, генная сеть «Никотиновая зависимость») до 6.38 (т.е. большая часть генов эволюционно молодая, генная сеть «Астма»). Изменение DI гена, как правило, происходит в пределах $DI < 1$, т.е. в пределах стабилизирующего отбора, тем не менее уровень изменчивости генов, задействованных в разных сетях, также сильно варьирует: от 0.16 до 0.64. Наиболее выделяются по индексам PAI и DI сети «Астма» и «Никотиновая зависимость». В сети «Астма» преобладают эволюционно молодые и изменчивые гены, а в сети «Никотиновая зависимость» – эволюционно древние и консервативные. Результат анализа PAI для сетей «Астма» и «Никотиновая зависимость» приведен на рис. 2; результаты анализа DI этих же сетей – на рис. 3.

Большинство генов в сети «Астма» – эволюционно молодые, появившиеся на уровне позвоночных (см. рис. 2, а, окрашены зеленым и желтым цветами). Напротив, в сети «Никотиновая зависимость» (см. рис. 2, а) все гены были определены как эволюционно древние, возникшие на этапах образования клеточной формы жизни (Cellular organisms) до многоклеточных животных (Metazoa).

Анализ индекса DI свидетельствует о том, что практически все гены в сети «Астма» (см. рис. 3, а) являются более эволюционно изменчивыми, чем гены, вовлеченные в сеть «Никотиновая зависимость» (см. рис. 3, б), гены которой очень консервативны.

Рассмотрим полученные оценки величин PAI для 11 категорий заболеваний (см. табл. 2). Наиболее выделяются из них 4: по высокому показателю PAI и DI – это болезни, связанные с иммунной системой (immune diseases, 8 сетей), и инфекционные заболевания, вызванные паразитами (infectious diseases: parasitic, 6 сетей). Низкий показатель PAI и DI характерен для специфических типов рака (cancers: specific types, 15 сетей) и зависимостей от химических соединений, вызывающих привыкание (substance dependence, 5 сетей).

Гены из рассмотренных выше категорий, а также полный набор 1436 генов были разбиты на две группы: 1) группа эволюционно древних генов с $PAI < 5$ (возраст генов соответствует периоду эволюции от формирования одноклеточных организмов (Cellular organisms) до хордовых (Chordata)); 2) группа эволюционно молодых генов с $PAI \geq 5$ (возраст генов соответствует периоду эволюции от плеченогих (Staniata) до современного человека). Далее были составлены таблицы сопряженности и с помощью точного теста Фишера проведена оценка, является ли достоверным отличие в разбиении генов на группы в категории от разбиения в полном списке генов (табл. 3).

Среднее значение PAI всех 1436 исследованных генов составило 2.49. По результатам табл. 3 видно, что генные сети, связанные с заболеваниями иммунной системы, обладают не только самым высоким значением филостратиграфического индекса (5.21), но и достоверно отличным распределением доли молодых и древних генов от аналогичной доли среди всех проанализированных генов (см. в последней строке табл. 3).

Доля молодых генов в категории заболеваний, связанных с иммунной системой (immune diseases), составила 65 %. При этом наибольшая доля генов приходится на позвоночных (Vertebrata) и костных позвоночных (Euteleostomi), что соответствует современным представлениям о развитии специфического иммунитета: он существует у хрящевых рыб (акул и скатов) и, следовательно, появился по крайней мере 400–500 млн лет назад. У этих рыб есть гены, родственные генам варибельной области Ig (IgV) или генам рецепторов Т-клеток (TkP). При этом еще более примитивные позвоночные – круглоротые (миксины и миноги) – не имеют системы приобретенного иммунитета, у них нет ни IgV , ни TkP -генов (Галактионов, 2004). Анализ выявил также и некоторую долю эволюционно древних генов в категории заболеваний, связанных с иммунной системой. Это соответствует сложившемуся представлению о том, что некоторые функции иммунной системы возникали еще у одноклеточных, например способность к фагоцитозу; клетки, имеющие маркер Т-лимфоцита, впервые обнаруженные у кольчатых червей, система гистосовместимости, – у губок (Хаитов, 2016). С другой стороны, наибольшая доля эволюционно древних генов характерна для категории зависимостей от химических соединений, вызывающих привыкание (substance dependence), а именно 88 %. Большинство рассмотренных генов вовлечены в функционирование нервной системы, включая нейротрансмиттерные функции.

Достоверным отличием доли эволюционно древних и эволюционно молодых генов от аналогичной среди всех проанализированных генов обладает категория инфек-

Таблица 2. Средние значения индексов PAI и DI для генов, вовлеченных в генные сети заболеваний человека из базы данных KEGG Pathway, Human Diseases

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

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

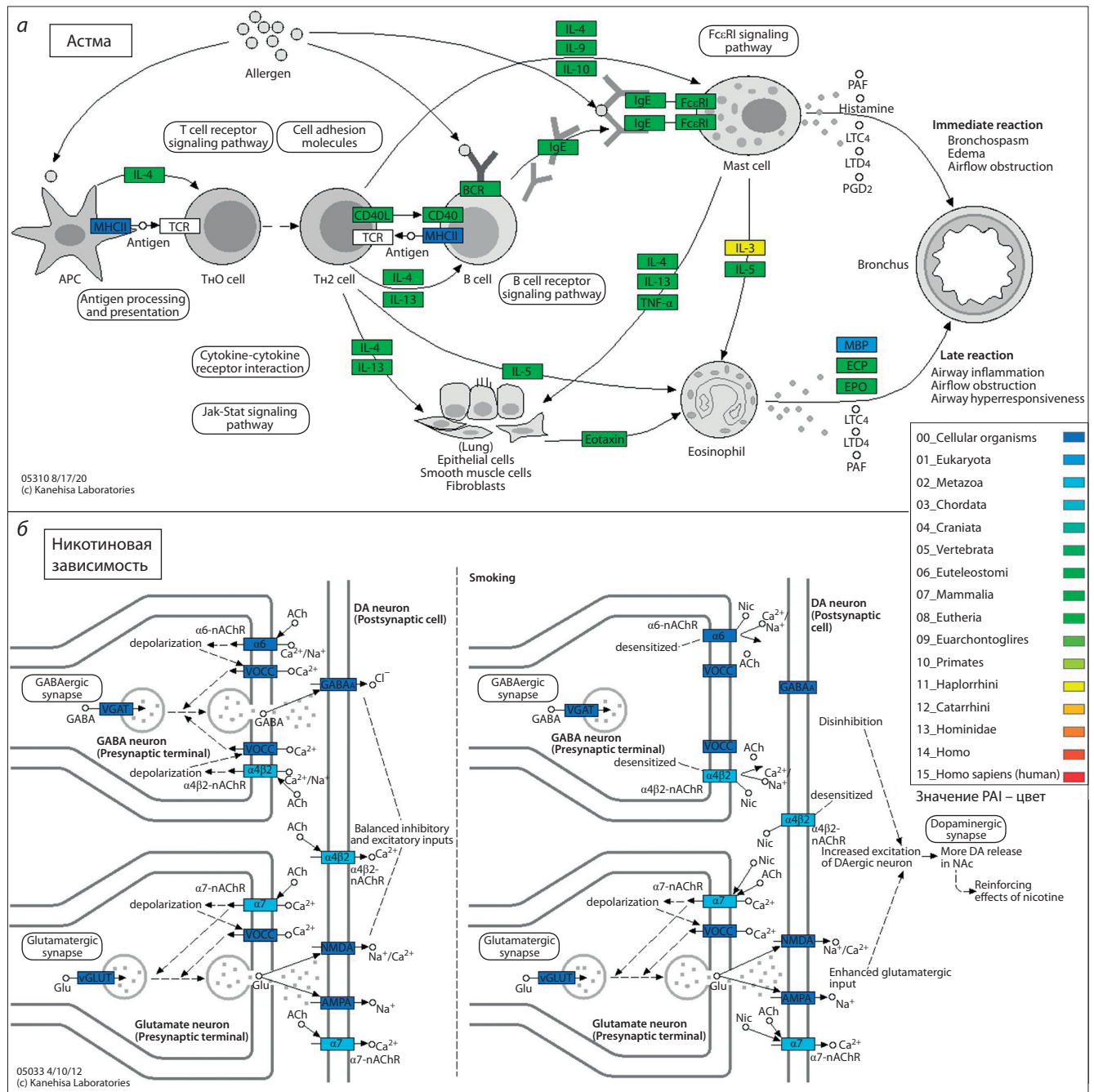


Рис. 2. Схемы генных сетей заболеваний «Астма» (а) и «Никотиновая зависимость» (б) из базы данных KEGG Pathway, Human Diseases с вычисленными значениями PAI.

Гены, кодирующие белки в этих сетях, показаны прямоугольниками с названием гена; цвет прямоугольника соответствует возрасту гена. Схема соответствия цвета и возраста гена приведена справа. Окрашенные в синий и голубой цвета гены относятся к наиболее эволюционно древним таксонам, в зеленый и желтый – к более эволюционно молодым относительно обозначенных голубым.

ционных заболеваний, вызванных паразитами (infectious diseases parasitic), 53 % эволюционно молодых генов. В этом случае высокая доля эволюционно молодых генов может быть напрямую связана с высокой долей эволюционно молодых генов и высокой эволюционной изменчивостью генов, найденной в категории заболеваний, связанных с иммунной системой. Именно инфекционные заболевания – один из важнейших движущих факторов эволюции иммунной системы. При этом инфекционные заболевания различной природы и иммунная система ко-

эволюционируют в процессе формирования механизмов борьбы друг с другом (Sasaki et al., 2000; Khakoo, 2004; Zheleznikova, 2014).

Отметим также категорию специфических типов рака (cancers specific types), включающую гены, ассоциированные с канцерогенезом. Для нее наблюдается достоверное превышение доли древних генов над молодыми в сравнении с их распределением (древние/молодые) в полной выборке проанализированных генов. Этот результат соответствует современным представлениям о том,

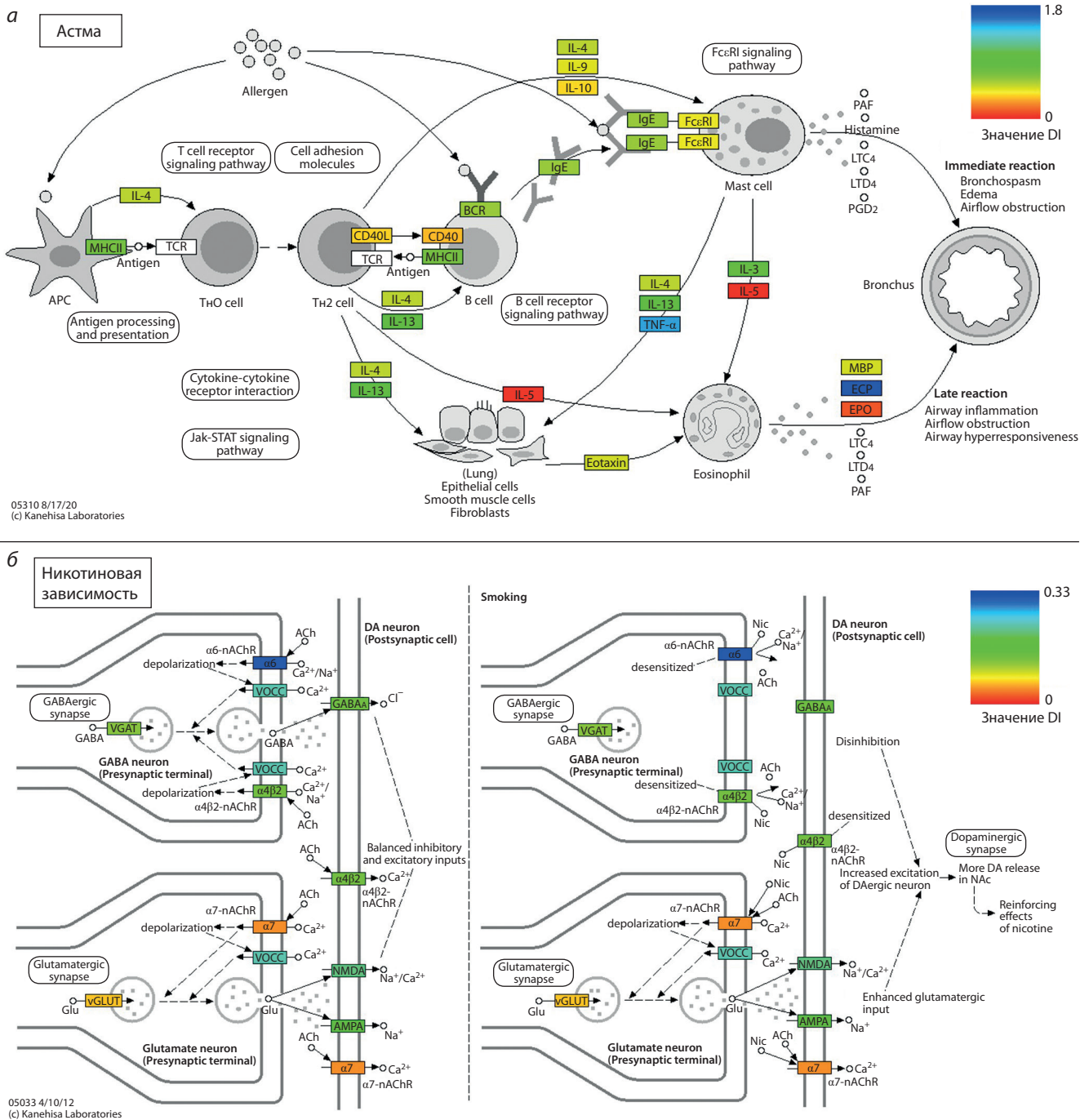


Рис. 3. Схемы генных сетей заболеваний «Астма» (а) и «Никотиновая зависимость» (б) из базы данных KEGG Pathway, Human Diseases с вычисленными значениями DI.

Гены, кодирующие белки в этих сетях, показаны прямоугольниками с названием гена; цвет прямоугольника соответствует уровню изменчивости гена. В правой верхней части графика для каждой сети приведена цветовая схема соотношения цветов и индекса DI. Шкала для каждой сети индивидуальна, и даже наиболее изменчивые гены, задействованные в сети «Никотиновая зависимость», обладают минимальной изменчивостью по сравнению с генами, вовлеченными в сеть «Астма».

что генные сети, вовлеченные в процессы развития рака, формировались на этапах возникновения многоклеточных организмов (Domazet-Lošo, Tautz, 2010).

Рассмотрим более подробно две категории заболеваний: 1) связанных с иммунной системой и обладающих наибольшей долей эволюционно молодых генов и 2) связанных с формированием зависимостей от химических

веществ, вызывающих привыкание, обладающих наибольшей долей эволюционно древних генов (рис. 4). Нижняя и верхняя точки каждого графика показывают минимальное и максимальное значения PAI, оранжевая звезда – медиану значений PAI, ширина графика для каждой позиции по оси ординат (т.е. для каждого PAI) – долю генов с этим конкретным PAI (см. рис. 4). Можно видеть, что в случае

Таблица 3. Результаты точного теста Фишера по сравнению распределения по группам эволюционно древних и эволюционно молодых генов среди всех генов, описанных в генных сетях заболеваний человека из KEGG Pathway, Human Diseases, и среди генов в рамках одной категории

Категория KEGG Pathway, Human Diseases	Гены		PAI	p-value-теста
	эволюционно древние	эволюционно молодые		
Заболевания, связанные с иммунной системой	56	106	5.21	8.84×10^{-15}
Инфекционные заболевания, вызванные паразитами	74	84	4.08	2.79×10^{-6}
Специфические типы рака	187	54	1.77	4.41×10^{-4}
Зависимость от химических соединений, вызывающих привыкание	69	9	1.03	1.75×10^{-5}
Всего из 1436 генов	952	484	2.49	–

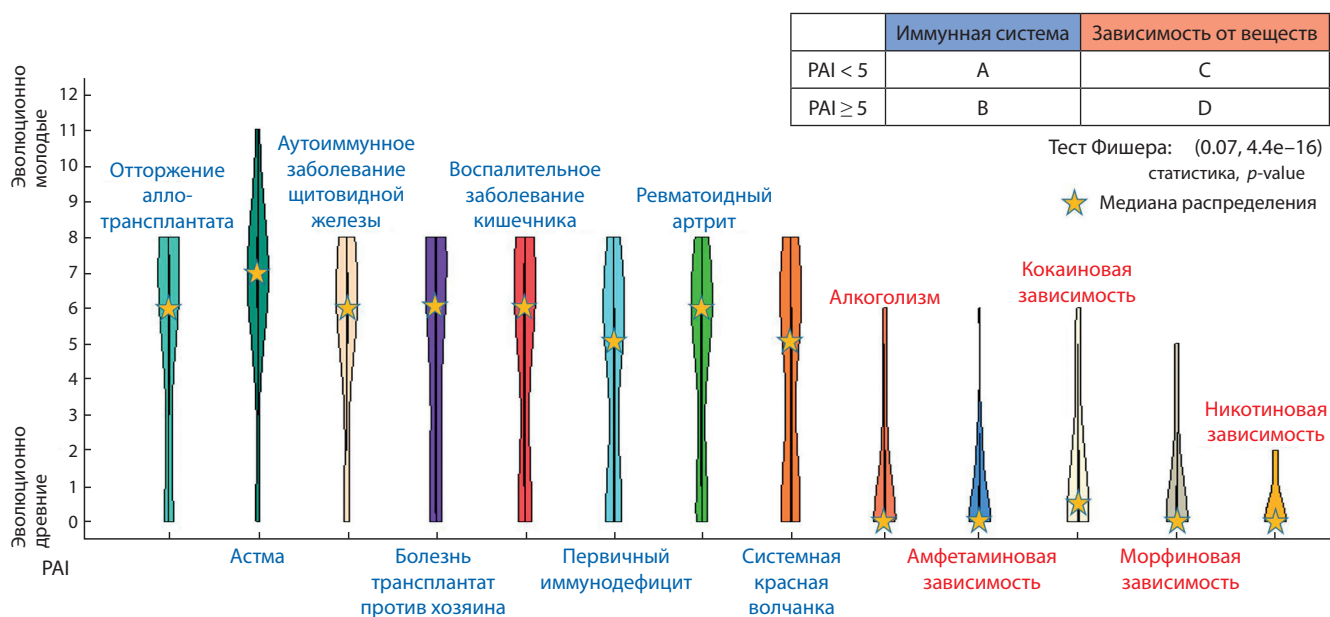


Рис. 4. Распределение PAI среди восьми сетей заболеваний, связанных с иммунной системой (подписаны синим) и пяти сетей заболеваний, связанных с зависимостями от веществ, вызывающих привыкание к химическим соединениям (подписаны красным).

Графики визуализированы с помощью R пакета violinplot, скрипт подготовлен Orthoscape.

заболеваний, связанных с иммунной системой, медиана распределений PAI колеблется в диапазоне (5, 7) (от позвоночных (Vertebrata) до млекопитающих (Mammalia)), а сами распределения имеют характер, выражающийся в уменьшении числа генов с соответствующим значением PAI при уменьшении PAI. В случае заболеваний, связанных с зависимостями от веществ, вызывающих привыкание к химическим соединениям, медиана находится в диапазоне (0, 1) – клеточные организмы (Cellular organisms) и эукариоты (Eukaryota), сами распределения имеют характер, выражающийся в увеличении числа генов с соответствующим значением PAI при уменьшении PAI. Распределения носят принципиально разный характер, если сравнивать в них долю эволюционно древних и эволюционно молодых генов, что показал также и точный тест Фишера с достоверностью $p\text{-value} = 4.4 \times 10^{-16}$.

Распределение PAI среди всех генов, задействованных в 80 рассмотренных генных сетях из KEGG Pathway, Human Diseases, представлено на рис. 5. Это распределение

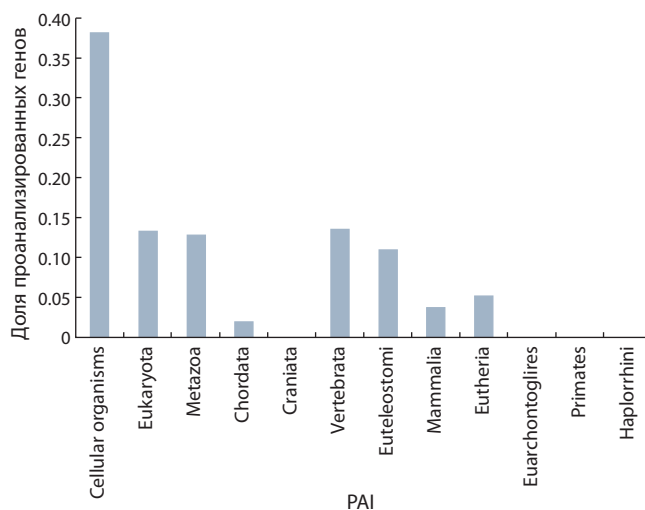


Рис. 5. Распределение PAI среди всех генов, задействованных в генных сетях из KEGG Pathway, Human Diseases.

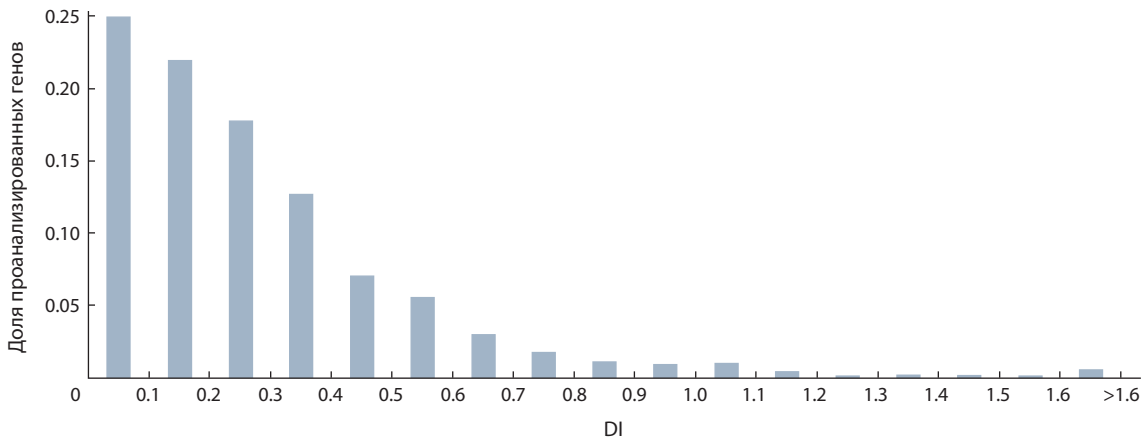


Рис. 6. Распределение DI среди всех генов, задействованных в генных сетях из KEGG Pathway, Human Diseases.

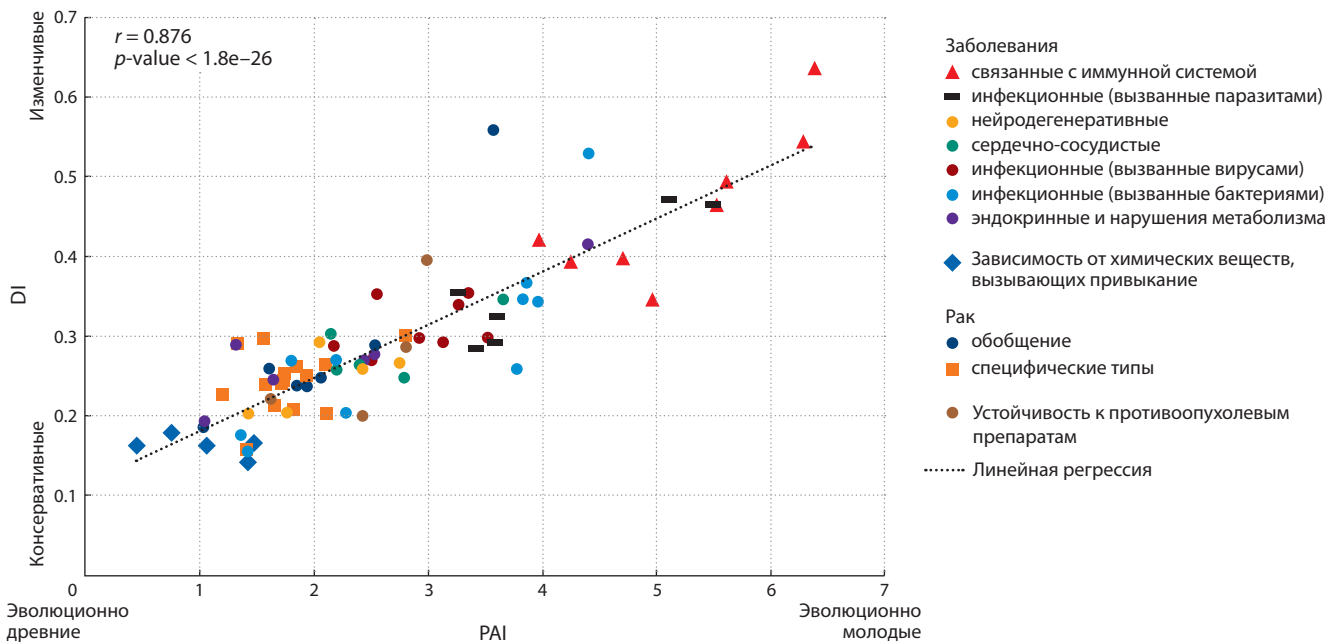


Рис. 7. Диаграмма рассеяния для средних значений индексов PAI и DI для 80 генных сетей заболеваний человека, описанных в базе KEGG Pathway, Human Diseases.

Фигурами разных цветов и размеров отмечены различные категории заболеваний.

имеет два пика. Левый пик включает гены, сформировавшиеся на раннем этапе эволюции (от возникновения клеточной организации жизни до хордовых), а правый – гены, сформировавшиеся на последующих этапах эволюции (от позвоночных до плацентарных). При этом эволюционно древних генов оказалось больше, чем эволюционно молодых.

Распределение DI среди всех генов, задействованных в рассмотренных генных сетях из KEGG Pathway, Human Diseases, приведено на рис. 6. Анализ DI позволяет оценить, какому типу отбора подвержены гены. При этом он корректно интерпретируется только в случае сравнения последовательностей анализируемых генов с ортологичными генами эволюционно близких организмов. Для вычисления dN/dS последовательности генов человека сравнивали с последовательностями ортологичных генов

у других гоминид; если ортологов было несколько, то в качестве DI использовали среднее значение dN/dS . Лишь для 38 из 1436 изученных нами генов были получены значения $DI > 1$ (девять из них приходятся на одну категорию – заболеваний, связанных с иммунной системой). Из данного распределения следует, что большинство генов, входящих в состав исследованных генных сетей, эволюционировало в режиме стабилизирующего отбора ($DI < 1$).

Представлялось интересным изучить взаимоотношение между PAI и DI для исследованных нами 80 генных сетей. Результаты этого анализа показаны на рис. 7 на одном графике, с учетом разбиения заболеваний по категориям.

Анализ показал, что между PAI и DI имеется большая и высокодостоверная корреляция ($r = 0.876$, $p\text{-value} < 1.8 \times 10^{-26}$), т.е. наблюдается зависимость между средним эволюционным возрастом генов в генных сетях

и уровнем их генетической изменчивости: чем меньше эволюционный возраст генов, тем больше уровень их генетической изменчивости. Это хорошо согласуется с тем, что эволюционно древние гены вовлечены в ключевые для функционирования организма процессы, на них наложено множество ограничений со стороны других генов, особенностей организации молекулярно-генетических систем и им не свойственна высокая изменчивость. Эволюционно молодые гены, напротив, обеспечивают адаптацию к современным условиям жизни, и у них более высокая изменчивость.

Заключение

Филостратиграфический анализ – современная методология, позволяющая на основании данных о сходстве генетических последовательностей и происхождении организмов оценить возраст генов в масштабе всего генома. Вместе с информацией о том, какому типу отбора подвержен ген как единица наследственности, результаты анализа дают возможность судить о роли тех или иных генов в эволюции генных сетей организма.

При анализе генных сетей из базы данных KEGG Pathway, Human Diseases выявлено несколько тенденций. Большинство генов, задействованных в исследованных генных сетях, эволюционировали в режиме стабилизирующего отбора ($DI < 1$). Обнаружена достоверная зависимость ($r = 0.876$, $p\text{-value} < 1.8 \times 10^{-26}$) между средним эволюционным возрастом генов в генных сетях и уровнем их генетической изменчивости: чем меньше эволюционный возраст генов, тем больше их уровень генетической изменчивости. Некоторые категории генных сетей значительно выделяются по доле эволюционно молодых и эволюционно древних генов. Наибольшая доля эволюционно молодых генов (65 %) отмечена в генных сетях, связанных с заболеваниями иммунной системы. Наибольшая доля эволюционно древних генов (88 %) обнаружена в генных сетях, описывающих формирование зависимостей человека от химических соединений, вызывающих привыкание.

Показано, что генные сети, ответственные за развитие инфекционных заболеваний, вызванных паразитами, достоверно обогащены эволюционно молодыми генами, а генные сети, ответственные за развитие специфических типов рака, – эволюционно древними генами. Такие результаты говорят об активном процессе адаптации иммунной системы человека к возникающим угрозам. Кроме того, гены, задействованные в заболеваниях, вызывающих привыкание к химическим соединениям, обладают минимальным числом замен, т. е. такие гены максимально консервативны. В этом направлении можно провести отдельную работу с расширением исходных сетей с помощью доступных на сегодняшний день классификаторов и баз данных.

Список литературы / References

Галактионов В.Г. Иммунология: учебник для студентов вузов, обучающихся по направлению 510600 «Биология» и биол. специальностям. М.: Академия, 2004.
[Galaktionov V.G. Immunology: a Guide for University Students Studying in Track 510600 “Biology” and Biological Specialties. Moscow: Academia Publ., 2004. (in Russian)]

- Колчанов Н.А., Игнатъева Е.В., Подколodная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):833-850.
[Kolchanov N.A., Ignat'eva E.V., Podkolodnaya O.A., Likhoshvay V.A., Matushkin Yu.G. Gene Networks. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850. (in Russian)]
- Степанов В.А. Эволюция генетического разнообразия и болезни человека. *Генетика*. 2016;52(7):852-864.
[Stepanov V.A. Evolution of genetic diversity and human diseases. *Russ. J. Genet.* 2016;52(7):746-756.]
- Хайтов Р.М. Иммунология: учебник для студентов медицинских вузов. М.: ГЭОТАР-Медиа, 2016.
[Khaitov R.M. Immunology: a Guide for Students of Medical Universities. Moscow, 2016. (in Russian)]
- Bell E.A., Boehnke P., Harrison T.M., Mao W.L. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. USA*. 2015;112:14518-14521. DOI 10.1073/pnas.1517557112.
- Cerami E.G., Gross B.E., Demir E., Rodchenkov I., Babur Ö., Anwar N., Schultz N., Bader G.D., Sander C. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39:685-690. DOI 10.1093/nar/gkq1039.
- Chatterjee H.J., Ho S.Y., Barnes I., Groves C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.* 2009;9:259. DOI 10.1186/1471-2148-9-259.
- Datta P.M. Earliest mammal with transversely expanded upper molar from the Late Triassic (Carnian) Tiki Formation, South Rewa Gondwana Basin, India. *J. Vertebr. Paleontol.* 2005;25:200-207. DOI 10.1671/0272-4634(2005)025(0200:EMWTEU)2.0.CO;2.
- Diogo R. The Origin of Higher Clades: Osteology, Myology, Phylogeny and Evolution of Bony Fishes and the Rise of Tetrapods. New York: CRC Press, 2007.
- Domazet-Lošo T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 2007;23:533-539. DOI 10.1016/j.tig.2007.08.014.
- Domazet-Lošo T., Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* 2010;8:66.
- Dunn R.H., Rose K.D., Rana R.S., Kumar K., Sahni A., Smith T. New euprimate postcrania from the early Eocene of Gujarat, India, and the strepsirrhine-haplorhine divergence. *J. Hum. Evol.* 2016;99:25-51.
- Harrison T. Catarrhine origins. In: *A Companion to Paleoanthropology*. New York: Blackwell Publ. Ltd., 2013;376-396.
- Hey J. The ancestor's tale A pilgrimage to the dawn of evolution. *J. Clin. Invest.* 2005;115:1680-1680.
- Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45:D353-D361.
- Khakoo S.I. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*. 2004;305(5685):872-874.
- Kumar V., Hallström B.M., Janke A. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PLoS One*. 2013;8(4):e60019.
- Leander B.S. Predatory protists. *Curr. Biol.* 2020;30:R510-R516.
- Li W.-H. Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J. Mol. Evol.* 1993;36(1):96-99.
- Li W.H., Wu C.I., Luo C.C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 1985;2(2):150-174.
- Liebeskind B.J., McWhite C.D., Marcotte E.M. Towards consensus gene ages. *Genome Biol. Evol.* 2016;8(6):1812-1823.
- Luo Z.-X., Yuan C.-X., Meng Q.-J., Ji Q. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature*. 2011;476:442-445.

- Maloof A.C., Porter S.M., Moore J.L., Dudas F.O., Bowring S.A., Higgins J.A., Fike D.A., Eddy M.P. The earliest Cambrian record of animals and ocean geochemical change. *Geol. Soc. Am. Bull.* 2010a; 122:1731-1774.
- Maloof A.C., Rose C.V., Beach R., Samuels B.M., Calmet C.C., Erwin D.H., Poirier G.R., Yao N., Simons F.J. Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nat. Geosci.* 2010b;3:653-659.
- Montojo J., Zuberi K., Rodriguez H., Kazi F., Wrig G., Donaldson S.L., Morris Q., Bader G.D. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop. *Bioinformatics.* 2010;26:2927-2928.
- Mustafin Z.S., Lashin S.A., Matushkin Y.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics.* 2017;18(S1):1-9.
- Nei M., Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 1986;3:418-426.
- Nersisyan L., Samsyan R., Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. *FI1000Res.* 2014;3:145.
- Pamilo P., Bianchi N.O. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 1993;10(2): 271-281.
- Sasaki K., Tsutsumi A., Wakamiya N. Mannose-binding lectin polymorphisms in patients with hepatitis C virus infection. *Scand. J. Gastroenterol.* 2000;35(9):960-965.
- Scerri E.M.L., Thomas M.G., Manica A., Gunz P., Stock J.T., Stringer C., Grove M., Groucutt H.S., Timmermann A., Rightmire G.P., D'Errico F., Tryon C.A., Drake N.A., Brooks A.S., Dennell R.W., Durbin R., Henn B.M., Lee-Thorp J., DeMenocal P., Petraglia M.D., Thompson J.C., Scally A., Chikhi L. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol. Evol.* 2018;33(8):582-594.
- Schrenk F., Kullmer O., Bromage T. The earliest putative homo fossils. In: *Handbook of Paleoanthropology.* Berlin; Heidelberg: Springer, 2014;1-19.
- Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.
- Shu D.-G., Luo H.-L., Conway Morris S., Zhang X.-L., Hu S.-X., Chen L., Han J., Zhu M., Li Y., Chen L.-Z. Lower Cambrian vertebrates from south China. *Nature.* 1999;402(6757):42-46.
- Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J., von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019; 47(D1):D607-D613.
- Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591.
- Yang Z., Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 2000;17(1):32-43.
- Zheleznikova G.F. Infection and immunity: strategies from both sides. *Med. Immunol.* 2014;8(5-6):597-614. <https://doi.org/10.15789/1563-0625-2006-5-6-597-614>. (in Russian)

ORCID ID

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

Благодарности. Работа поддержана грантом РФФИ № 20-04-00885 А и бюджетным проектом № 0259-2021-0009.

Прозрачность финансовой деятельности. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 14.01.2021. После доработки 20.01.2021. Принята к публикации 20.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

Пангеномы сельскохозяйственных растений

А.Ю. Пронозин¹✉, М.К. Брагина^{1, 2}, Е.А. Салина^{1, 2}

¹ Федеральное исследовательское учреждение Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

✉ pronozinartem95@gmail.com

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

Ключевые слова: сельскохозяйственные растения; геномы; пангеномы; гены; эволюция; биоинформатический анализ; вычислительные конвейеры.

Для цитирования: Пронозин А.Ю., Брагина М.К., Салина Е.А. Пангеномы сельскохозяйственных растений. *Вавиловский журнал генетики и селекции*. 2021;25(1):57-63. DOI 10.18699/VJ21.007

Crop pangenomes

A.Yu. Pronozin¹✉, M.K. Bragina^{1, 2}, E.A. Salina^{1, 2}

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

² Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

✉ pronozinartem95@gmail.com

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

Key words: agricultural plants; genomes; pangenomes; genes; evolution; bioinformatics analysis; computational pipelines.

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

Введение

Секвенирование генома организма – важный этап в генетических исследованиях генома. Расшифровка геномной последовательности открывает широкие возможности для исследования строения структуры хромосом, распределения повторенных и кодирующих последовательностей, идентификации и аннотации генов (Брагина и др., 2019). Информация о последовательностях геномов разных видов позволяет проводить сравнительный филогенетический анализ для изучения отношений между видами, их происхождения и особенностей эволюции (Marchant et al., 2016; Wendel et al., 2016). У сельскохозяйственных растений все это дает возможность оценить влияние генетической изменчивости на функцию генов, определить гены, ответственные за наиболее ценные признаки сельскохозяйственных культур (Schnable et al., 2009; Wing et al., 2018).

При решении этих задач геном вида представляется последовательностью одного организма (так называемый референсный геном). Первичная структура референсного генома улучшается в результате целого ряда последовательных экспериментальных и биоинформатических исследований, ее аннотация служит отправной точкой для генетиков, исследующих данную культуру. Количество секвенированных, собранных и аннотированных референсных геномов растений увеличивается с каждым годом (Брагина и др., 2019). В версии 48 базы данных Ensembl plants (сентябрь 2020 г.) содержится 93 собранных и аннотированных генома растений (Howe et al., 2020). На основе референсной геномной последовательности и повторно секвенирования геномных последовательностей представителей одного вида (как правило, с использованием технологии коротких прочтений) производят анализ генетической изменчивости, изучение однонуклеотидных полиморфизмов (single-nucleotide polymorphisms, SNPs) и крупных структурных вариаций (structural variations, SVs) генома. Последний тип вариаций наиболее труден для идентификации на основе секвенирования короткими прочтениями, однако с созданием технологий третьего поколения, позволяющих читать последовательности ДНК длиной до сотен тысяч нуклеотидов (Li et al., 2018), идентификация больших структурных перестроек становится более доступной и надежной. Появляется больше свидетельств в пользу того, что структурные изменения, включая вариации числа копий генов (copy number variations, CNVs) и вариации присутствия/отсутствия генов (presence/absence variations, PAVs), преобладают в сельскохозяйственных культурах и приводят к значительным вариациям функционального набора генов и геномного состава у особей одного вида (Springer et al., 2009; Hirsch et al., 2014; Li et al., 2014; Lu et al., 2015; Zhao Q. et al., 2018).

Геномы и пангеном

Для более эффективного анализа и описания разнообразия геномного состава была предложена концепция пангенома (Tettelin et al., 2005). Пангеном – это информация о полной выборке генов в биологическом кластере (таксоне), например виде, среди которых можно выделить набор универсальных (основных) генов, общих для всех образцов, и набор уникальных (вариабельных) генов, частично

общих или индивидуально специфичных (Tettelin et al., 2005). Исследования пангенома до настоящего времени были сосредоточены на поиске наличия или отсутствия генов у объектов для определения универсального или уникального набора генов.

Термин «пангеном» был изначально сформулирован в работе (Tettelin et al., 2005) для бактериальных видов *Streptococcus agalactiae*. На сегодняшний день существует несколько определений этого термина, которые базируются на двух концепциях: структурной и функциональной (Tranchant-Dubreuil et al., 2018). Структурная концепция рассматривает пангеном как совокупность всех геномных последовательностей таксона. В рамках этой концепции нуклеотидные последовательности геномов-представителей таксона (одного вида или рода) сравниваются между собой, и на этой основе определяется их общий уникальный (не избыточный) набор фрагментов ДНК одинаковой длины (100 п. н. или больше, в зависимости от вида). Эти последовательности и описывают структуру пангенома (Snipen et al., 2009; Alcaraz et al., 2010).

Вторая концепция основана на его функциональном представлении. В качестве функциональной компоненты рассматриваются все кодируемые в нем гены. В этом случае пангеном может быть описан как объединение всех генов для представителей определенного таксона (Plissonneau et al., 2018). Однако для большого количества родственных организмов такой набор является вырожденным, поскольку они содержат много генов с высоким уровнем сходства первичной структуры и, соответственно, функций. Исключить избыточность пангенома можно за счет объединения сходных последовательностей генов в функциональные семейства (Sun et al., 2016). При этом гены-представители одного функционального семейства в разных организмах рассматриваются с точки зрения функции как одна последовательность.

Что касается таксономической принадлежности организмов, которые формируют пангеном, то, как правило, их набор ограничивается отдельным видом. Однако некоторые исследователи используют более широкую трактовку пангенома. Например, в работе В.В. Тец (2003) пангеном рассматривается как полный набор генов живых организмов, вирусов и мобильных элементов.

Структурные особенности пангенома

Гены в пангеноме можно разделить на две группы по их представленности в разных организмах (Golicz et al., 2016). К первой группе относятся гены, которые встречаются у всех представителей таксона. Такая группа генов называется универсальным набором (англ. core gene set). Вторую группу составляют гены, имеющиеся у части представителей таксона. Эту группу генов называют необязательными (indispensable), второстепенными (accessory) или вариабельными генами. Среди генов второй группы особо выделяют уникальные, представленные лишь у одного индивида в таксоне гены. Универсальные и вариабельные гены отражают функциональную основу и разнообразие представителей вида соответственно.

С точки зрения эволюции, универсальные гены в большинстве случаев являются генами, которые выполняют жизненно важные функции и они, как правило, сохра-

няются в пределах вида. Напротив, вариабельные гены и их особая фракция, уникальные гены, вносят вклад в разнообразие видов, что позволяет им адаптироваться к различным условиям окружающей среды. Доля уникальных генов в пангеноме изученных культур варьирует от 8 до 61 % (Тао et al., 2019). Однако полученный размер уникального генома, вероятно, будет недооценен из-за неспособности современных стратегий и технологий определять все функциональные изменения в генах.

На основании последовательности одного генома невозможно определить, какие гены – общие для всех представителей вида, а какие – только для некоторых. Тем не менее для каждой новой последовательности существует возможность идентифицировать, к какой части пангенома она относится: к универсальной или вариабельной. Чем больше геномов-представителей таксона секвенировано, тем больше обнаруживается уникальных генов. Это приводит к росту размера пангенома при увеличении количества геномов. Однако для набора универсальных генов увеличение количества геномов вызывает обратный процесс: часть генов, которые являются универсальными, у новых представителей вида может отсутствовать. В результате размер пангенома – совокупности всех различных генов вида – увеличивается, а предполагаемый размер универсального набора генов, как правило, уменьшается (Golicz et al., 2016; Wang et al., 2018). Схематически эта зависимость показана на рис. 1. Каждая точка на этом графике соответствует оценке количества генов в пангеноме для набора из k -геномов (взятых случайным образом из полной выборки N исследуемых геномов). При этом с увеличением k оценка общего количества генов в пангеноме растет (сплошная красная линия), а количество уникальных генов уменьшается (синяя штриховая линия). Примеры зависимостей для реальных пангеномов можно найти на сайте <https://pangp.zhaopage.com>. Таким образом, на оценку размера пангенома и долю универсальных генов в нем существенно влияет размер выборки организмов.

На размер и долю уникальных генов пангенома, помимо количества секвенированных геномов, также влияют: 1) выбор образцов для анализа – объединение диких и культурных видов даст пангеном с более высокой долей уникальных генов, чем использование только культурных растений (Montenegro et al., 2017; Zhao Q. et al., 2018); 2) уровень ploидности, способ размножения, эффект «бутылочного горлышка» в процессе доместикации и др. Виды растений с более высоким уровнем ploидности и аутбридинга и сокращением разнообразия в результате доместикации, как правило, имеют большую долю уникальных генов (Тао et al., 2019).

Можно предположить, что добавление неограниченного количества новых геномов в пангеном приводит к его неограниченному росту. Однако исследования разнообразия генов у видов сельскохозяйственных культур показали, что для них количество идентифицированных уникальных генов имеет тенденцию к уменьшению по мере увеличения числа секвенированных образцов. Это позволяет считать, что при определенном количестве представителей таксона включение дополнительных геномов в пангеном уже не приведет к дальнейшему увеличению количества его генов. Такие пангеномы называют закрытыми. У томата

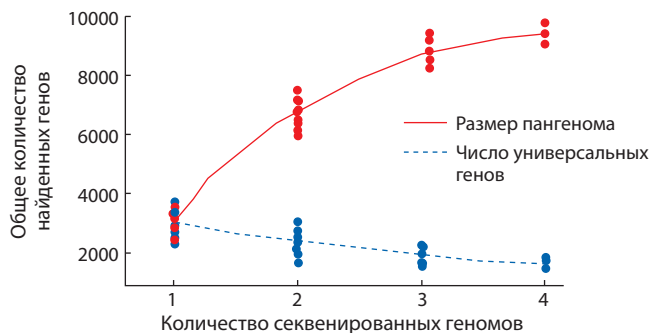


Рис. 1. Зависимость размера пангенома и числа универсальных генов в нем от числа секвенированных геномов-представителей таксона.

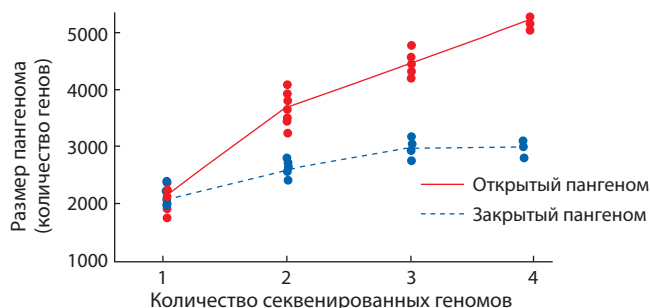


Рис. 2. Зависимость количества генов в пангеноме (ось Y) от количества секвенированных представителей таксона (ось X) для двух типов пангеномов: открытых и закрытых.

Для открытых геномов количество генов растет монотонно, для закрытых – выходит на плато.

(Gao et al., 2019), кукурузы (Hirsch et al., 2014), риса (Wang et al., 2018), сои (Li et al., 2014), подсолнечника (Hübner et al., 2019), *Brachypodium distachyon* (Gordon et al., 2017), *Brassica napus* (Hurgobin et al., 2018) и *B. oleracea* (Golicz et al., 2016) обнаружен закрытый пангеном.

Однако существуют также пангеномы, в которых общее количество генов растет при добавлении каждого нового образца. Такие пангеномы называют открытыми, они характерны для микроорганизмов. Например, результаты анализа пангенома грибного возбудителя септориоза листьев пшеницы *Zyoseptoria tritici* показали, что он относится к открытому типу (Plissonneau et al., 2018). Пангеном бактерии *Paenibacillus polymyxa*, обитающей в ризосфере растений и защищающей их от фитопатогенов (Zhou et al., 2020), также принадлежит к открытому типу.

При условии, что организмы из популяции отбираются случайным образом, тип пангенома можно оценить путем построения графика количества генов, обнаруженных в каждой новой геномной последовательности (рис. 2). Если после анализа определенного количества геномных последовательностей число генов в пангеноме выходит на плато, это считается характеристикой «закрытых» пангеномов. Такая зависимость схематически показана на рис. 2 (синяя штриховая линия). Если в зависимости размера пангенома от количества геномов нет признаков выхода на плато, – это характеристика «открытых» пангеномов. Зависимость числа генов от количества геномов для открытого пангенома схематически показана на рис. 2 (красная сплошная линия).

Сравнение размеров пангеномов и доли универсальной и варибельной части для ряда растительных видов представлено в Приложении 1¹. Данные в Приложении 1 демонстрируют, что количество представителей, включенных в анализ пангенома в растительных проектах, варьирует от 3 (репа, *Brassica rapa*) до 3000 (рис, *Oryza sativa*). Количество генов в пангеноме изменяется от 35 тыс. у риса – диплоида, до 128 тыс. у мягкой пшеницы – гексаплоида. Доля универсальных генов изменяется от 41 % у люцерны до 84 % у репы.

Функциональные особенности пангенома

По отношению функциональных особенностей генов из универсального и варибельного наборов пангеномов исследования показывают, что универсальные гены отвечают за фундаментальные клеточные процессы, в то время как варибельные гены ассоциированы, прежде всего, с функциями, которые могут дать преимущество в различных условиях окружающей среды. Так, при анализе пангенома трахинии двухколосковой *Brachypodium distachyon* (Gordon et al., 2017) было выявлено, что аннотации универсального набора генов обогащены такими терминами, как «гликолиз», «стероид», «гликозилирование», «кофермент». Аннотации генов варибельного набора были более всего обогащены терминами «защитная функция», «развитие». В этой же работе показано, что отношение доли несинонимических замен к синонимическим у варибельных генов выше, чем у универсальных. Кроме того, ортологи универсальных генов у риса и сорго оказались более консервативными, чем ортологи варибельного набора генов. Универсальные гены также имеют более высокий уровень экспрессии, по сравнению с варибельными (Gordon et al., 2017). Сходные результаты были получены при анализе пангенома сои *Glycine max* (Li et al., 2014; Liu et al., 2020), капусты (Golicz et al., 2016) и пшеницы (Montenegro et al., 2017).

Анализ этих и ряда других пангеномов сельскохозяйственных растений показал, что для них присуще следующее (Tao et al., 2019): последовательности варибельных генов более изменчивы по сравнению с универсальными; скорость накопления несинонимических замен у варибельных генов выше; варибельные гены отличаются большим разнообразием функций; функциональные характеристики варибельных и универсальных генов различаются, первые в большей степени связаны с ответом на факторы внешней среды, активностью рецепторов и передачей сигнала, вторые – с выполнением базовых клеточных функций. Таким образом, универсальные гены представляют собой консервативное ядро пангенома (и вида, соответственно), в то время как варибельные гены – это мобильная его часть (как в качестве функций, так и в отношении первичной структуры и паттернов экспрессии).

Пангеномы и пантранскриптомы

Еще один из методов анализа генного состава у нескольких представителей какого-либо таксона – это анализ его транскриптомов. Нуклеотидные последовательности

транскриптов (преимущественно мРНК), оценка уровня их экспрессии и наличие изоформ могут быть получены в результате высокопроизводительного секвенирования (RNA-seq), которое существенно дешевле, чем секвенирование генома. Транскриптомные данные позволяют оценить присутствие генов в геноме только в том случае, если они экспрессируются в какой-либо ткани или органе растения. Таким образом, по набору транскриптов нельзя представить полный состав генов в геноме, но получить приближенную оценку вполне возможно (особенно, если анализируется набор транскриптов из разных тканей на разных стадиях развития). При этом сборка транскриптома требует значительно меньше вычислительных ресурсов, а современные методы дают возможность получить ее с высоким качеством.

Исследование пантранскриптома 503 инбредных линий кукурузы дало возможность выявить генетическое разнообразие в белок-кодирующих генах: обнаружено более полутора миллиона однонуклеотидных вариаций, найдены мутации, ассоциированные с признаками развития растений (время прохождения ряда фаз роста) (Hirsch et al., 2014).

М. Jin с коллегами (2016) также изучали пантранскриптом 368 инбредных линий кукурузы. Они обнаружили более двух тысяч последовательностей, которые не были представлены в референсном геноме кукурузы, среди них гены, ответственные за ответ на биотический стресс. Рассмотрены вариации, ассоциированные с уровнем экспрессии генов (eQTL). Результаты были спроецированы на метаболические сети, что позволило уточнить механизмы их функционирования.

В работе (Ma et al., 2019) проанализировано 288 экспериментов по секвенированию транскриптома ячменя. Среди собранных транскриптов около 30 % не показали сходства с референсным геномом. Данные исследования пантранскриптома показали, что гены устойчивости к патогенам более многочисленны в дикорастущем ячмене. Такие гены в процессе доместикации были подвержены более сильному давлению отбора по сравнению с генами в других видах.

Методы сборки пангенома

В биоинформатическом анализе пангенома можно выделить основные этапы:

1. Сборка последовательностей пангенома.
2. Выделение консервативных и варибельных участков геномных последовательностей.
3. Идентификация/предсказание и функциональная аннотация генов.
4. Идентификация полиморфизмов.
5. Хранение, обеспечение быстрого доступа и визуализация пангеномных данных.

Для сборки пангеномов существуют стратегии: сборка-выравнивание; метагеномный подход; выравнивание-сборка (Golicz et al., 2016; Hurgobin, Edwards, 2017; Tranchant-Dubreuil et al., 2018).

Сборка-выравнивание. Метод основан на сборке *de novo* последовательностей каждого представителя таксона отдельно, с последующим выравниванием последовательностей между собой, а также относительно

¹ Приложения 1–3 см. по адресу:
<http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx2.pdf>

референсного генома, для того чтобы уменьшить избыточность и определить набор общих и варьируемых участков последовательностей. Для сборки генома разработано несколько программных пакетов: Velvet (Zerbino, Birney, 2008), SOAPdenovo (Xie et al., 2014), ALLPATHS (Butler et al., 2008) и MaSuRCA (Zimin et al., 2013). Такой подход требует много времени и вычислительных ресурсов. Стратегия сборки *de novo* использована для анализа пангенома культивируемой сои (Li et al., 2010), дикой сои (Li et al., 2014), риса (Wang et al., 2018), капусты (Golicz et al., 2016), люцерны (Zhou et al., 2020).

Метагеномный подход заключается в объединении всех секвенированных прочтений от разных представителей таксона в один пул и последующей сборке *de novo* контигов пангенома на основе этих данных. Затем каждый собранный контиг относится к определенному геному путем выравнивания исходных прочтений этого представителя на метагеномную сборку и последующего оценивания покрытий контигов. Метагеномный подход позволяет работать с результатами секвенирования с низким уровнем покрытия. Его применяли для анализа геномов риса (Yao et al., 2015), томата (Gao et al., 2019).

Выравнивание-сборка. Эта стратегия использует сборку одного полного генома (референсной последовательности) в качестве основы для сборки геномов остальных представителей таксона (guide assembly). Прочтения из одного представителя вида выравниваются относительно референсного генома, те прочтения, что не совпали, отсеиваются и собираются отдельно. Последовательность референсного генома дополняется новыми собранными последовательностями, далее образцы сравниваются с данным референсным геномом. Выравнивание-сборка дает возможность сократить время построения пангенома. В случае, если геномный фрагмент присутствует сразу у нескольких представителей таксона, его последовательность будет собрана лишь один раз, в то время как при независимой сборке *de novo* этот фрагмент будет собираться столько раз, сколько представителей таксона было исследовано. Такой подход применен при анализе пангенома подсолнечника (Hübner et al., 2019).

Следует также отметить, что в ряде работ исследователи не использовали сборку геномных последовательностей, а выравнивали короткие прочтения на референсный геном. Это позволяет оценить связь однонуклеотидного полиморфизма с фенотипическими характеристиками растений. Существуют также методы, которые на основе выравнивания коротких прочтений дают возможность оценить структурные перестройки, дубликации и потери генов (Zhao et al., 2013). Метод выравнивания использовали при анализе пантранскриптома кукурузы (Hirsch et al., 2014), оценке изменения количества копий генов при анализе пангенома картофеля (Żmieńko et al., 2014).

Методы аннотации и анализа пангенома

С помощью аннотации пангенома можно идентифицировать последовательности генов в геномах представителей таксона, на основе сравнения их последовательностей определить ортологичные гены, а также семейства универсальных и варьируемых генов. Для автоматической аннотации пангеномов разработан ряд программных

пакетов, выполненных в виде вычислительных конвейеров. Они проводят основные этапы анализа пангеномных последовательностей и их аннотации. Ниже – краткое описание возможностей ряда таких программ.

Программа PGAP (Zhao Y. et al., 2012) осуществляет масштабный поиск генов, проводит функциональную аннотацию, обогащение кластеров ортологичных генов терминами онтологии, анализ эволюции видов, выполняет структурный анализ пангенома, идентификацию универсальной и варьируемой части пангенома. В обновленной версии этой программы, PGAP-X (Zhao Y. et al., 2018), дальнейшее развитие получили методы представления и визуализации результатов анализа пангеномов.

Пакет программ PpsPCP (Tahir Ul Qamar et al., 2019) разработан для идентификации вариаций наличия/отсутствия генов (PAVs) в пангеномах. Анализ основан на полногеномном сравнении последовательностей представителей таксона и референсного генома в несколько раундов с последовательной коррекцией как набора генов, так и участков их выравнивания в референсном геноме. В результате создается набор генов пангенома путем объединения последовательностей отдельных геномов с референсным геномом и их аннотации.

Программа BPGA (Chaudhari et al., 2019) реализует широкие возможности по анализу пангеномов: кластеризация генов на основе сходства последовательностей, анализ наличия/отсутствия ортологов, построение графика зависимости размеров пангенома и его универсальной части от количества геномов, реконструкция филогенетического дерева между представителями таксона, анализ метаболических путей и функциональной аннотации, оценка отклонений GC состава, расчет различных статистических характеристик пангенома и др.

Программа panX (Ding et al., 2018) направлена на идентификацию кластеров ортологичных генов. Для этого используются кластеризация на основе сравнения последовательностей, верификация и уточнение состава кластеров на базе анализа эволюционных расстояний и филогенетической реконструкции; программа оценивает ассоциацию между геномным составом индивидуальных представителей таксона и их фенотипов.

Программа Pan4Draft (Veras et al., 2018) разработана для получения улучшенной аннотации пангеномов за счет добавления к ней информации о последовательностях незавершенных геномов (unfinished genomes). Это геномы, у которых аннотация и сборка до уровня хромосом не завершены, но их последовательности содержат фрагменты геномной ДНК и представляют ценную информацию о разнообразии геномов вида. Методы анализа ряда пангеномов растений описаны в Приложениях 2 и 3.

Перспективы использования пангеномных данных

В настоящее время исследование в направлении секвенирования и анализа пангеномов сельскохозяйственных растений активно продолжают и дают возможность получить все больше сведений о генетических вариациях и новых генах.

Одна из фундаментальных задач в изучении пангеномов сельскохозяйственных растений – оценка генетического

разнообразия их культурных представителей, а также диких сородичей. Такой анализ позволяет установить происхождение и эволюцию культурных растений, оценить влияние процесса селекции на генетическую структуру сортов. Анализ пангеномов, таким образом, отвечает на ряд важных вопросов о закономерностях эволюции геномов на уровне вида, механизмах возникновения новых генов, разнообразии функций генов и их ассоциациях с фенотипическими признаками растений.

Важным направлением исследования пангеномов сельскохозяйственных растений являются секвенирование и анализ геномов их диких сородичей. Предполагают, что дикие сородичи культурных растений могут содержать пул генов, связанных с адаптацией организмов к условиям окружающей среды, ответом на биотический и биотический стрессы, т.е. гены, которые могли быть утрачены представителями культурных растений в результате искусственного отбора (эффект «бутылочного горлышка») (Гончаров, Кондратенко, 2008; Гончаров, 2013; Purugganan, 2019). Обнаруженные гены могут быть в дальнейшем использованы для создания новых генотипов, более устойчивых к патогенам, вредителям и абиотическому стрессу. Таким образом, изучение пангеномов сельскохозяйственных растений не только имеет фундаментальный аспект, но также важно с точки зрения практической селекции.

Заключение

Более точное понимание генетического разнообразия генофонда в сочетании с передовыми технологиями секвенирования и высокопроизводительным фенотипированием может облегчить анализ признаков для выявления полезных генетических мутаций, позволить программам селекции получить доступ к более широкому спектру генетических ресурсов, помочь отбору лучших стратегий в селекционных программах и ускорить селекцию культурных растений для создания сортов со стабильно высокой урожайностью в стрессовых условиях.

Пангеномные исследования предлагают гораздо более широкое понимание генетического разнообразия генофондов сельскохозяйственных культур, чем анализ по ресеквенированию геномов, и, таким образом, могут быть чрезвычайно полезны для улучшения культурных растений. Тем не менее знания, полученные с помощью пангеномных исследований, требуют интеграции с QTL/GWAS и исследованиями по ресеквенированию геномов для определения важных генов и аллелей, которые будут использоваться в эффективной стратегии селекции.

Список литературы / References

Брагина М.К., Афонников Д.А., Салина Е.А. Прогресс в секвенировании геномов растений – направления исследований. *Вавиловский журнал генетики и селекции*. 2019;23(1):38-48. DOI 10.18699/VJ19.459.
[Bragina M.K., Afonnikov D.A., Salina E.A. Progress in plant genome sequencing: research directions. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2019;23(1):38-48. DOI 10.18699/VJ19.459. (in Russian)]
Гончаров Н.П. Доместикация растений. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):884-899.
[Goncharov N.P. Plants domestication. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):884-899. 2013;17(4/2):884-899. (in Russian)]

Гончаров Н.П., Кондратенко Е.Я. Происхождение, доместикация и эволюция пшеницы. *Информационный вестник ВОГиС*. 2008;12(1-2):159-179.
[Goncharov N.P., Kondratenko E.Ja. Wheat origin, domestication and evolution. *Informatcionniy Vestnik VOGiS* = *The Herald of Vavilov Society for Geneticists and Breeders*. 2008;12(1-2):159-179. (in Russian)]
Тец В.В. Пангеном. *Цитология*. 2003;45(5):526-531.
[Tets V.V. Pangenome. *Citologiya* = *Cytology*. 2003;45(5):526-531. (in Russian)]
Alcaraz L.D., Moreno-Hagelsieb G., Eguiarte L.E., Souza V., Herrera-Estrella L., Olmedo G. Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics*. 2010;11(1):332.
Butler J., MacCallum I., Kleber M., Shlyakhter I.A., Belmonte M.K., Lander E.S., Nusbaum C., Jaffe D.B. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res*. 2008;18(5):810-820. DOI 10.1101/gr.7337908.
Chaudhari N.M., Gupta V.K., Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* 2019;6(1):1-10. DOI 10.1038/srep24373.
Ding W., Baumdicker F., Neher R.A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* 2018;46(1):e5-e5. DOI 10.1093/nar/gkx977.
Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D.M., Thannhauser T.W., Burzynski-Chang E.A., Fish T.L., Stromberg K.A., Sacks G.L., Foolad M.R., Diez M.J., Blanca J., Canizares J., Xu Y., Knaap E., Huang S., Klee H.J., Giovannoni J.J., Fei Z. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 2019;51(6). DOI 10.1038/s41588-019-0410-2.
Golicz A.A., Batley J., Edwards D. Towards plant pangenomics. *Plant Biotechnol. J.* 2016;14(4):1099-1105. DOI 10.1111/pbi.12499.
Gordon S.P., Contreras-Moreira B., Woods D.P., Des Marais D.L., Burgess D., Shu S., Stritt C., Roulin A.C., Schackwitz W., Tyler L., Martin J., Lipzen A., Dochy N., Phillips J., Barry K., Geuten K., Budak H., Juenger T.E., Amasino R., Caicedo A.L., Goodstein D., Davidson P., Mur L.A.J., Figueroa M., Freeling M., Catalan P., Vogel J.P. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 2017;8(1):2184. DOI 10.1038/s41467-017-02292-8.
Hirsch C.N., Foerster J.M., Johnson J.M., Sekhon R.S., Muttoni G., Vaillancourt B., Peñagaricano F., Lindquist E., Pedraza M., Barry K., Leon N., Kaepler S.H., Buell R.C. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 2014;26(1):121-135. <https://doi.org/10.1105/tpc.113.119982>.
Howe K.L., Contreras-Moreira B., De Silva N., Maslen G., Akanni W., Allen J., Carbajo M. Ensembl Genomes 2020 – enabling non-vertebrate genomic research. *Nucleic Acids Res.* 2020;48(D1):D689-D695. DOI 10.1093/nar/gkz890.
Hübner S., Bercovich N., Todesco M., Mandel J.R., Odenheimer J., Ziegler E., Lee J.S., Baute G.J., Owens G.L., Grassa C.J., Ebert D.P., Ostevik K.L., Moyers B.T., Yakimowski S., Masalia R.R., Gao L., Čalić I., Bowers J.E., Kane N.C., Swanevelter D.Z.H., Kubach T., Muñoz S., Langlade N.B., Burke J.M., Rieseberg L.H. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*. 2019;5(1):54-69. DOI 10.1038/s41477-018-0329-0.
Hurgobin B., Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete. *Biology*. 2017;6(1):21. DOI 10.3390/biology6010021.
Hurgobin B., Golicz A.A., Bayer P.E., Chan C.K., Tirnaz S., Dolatabadian A., Schiessl S.V., Samans B., Montenegro J.D., Parkin I.A., Pires J.C. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* 2018;16(7):1265-1274. DOI 10.1111/pbi.12867.
Jin M., Liu H., He C., Fu J., Xiao Y., Wang Y., Xie W., Wang G., Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci. Rep.* 2016;6:18936. DOI 10.1038/srep18936.

- Li C., Lin F., An D., Wang W., Huang R. Genome sequencing and assembly by long reads in plants. *Genes*. 2018;9(1):6. DOI 10.3390/genes9010006.
- Li R., Zhu H., Ruan J., Qian W., Fang W., Shi Z., Li Y., Li Sh., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wang J. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20(2):265-272. DOI 10.1101/gr.097261.109.
- Li Y.H., Zhou G., Ma J., Jiang W., Jin L.G., Zhang Z., Guo Y., Zhang J., Sui Y., Zheng L., Zhang S.S. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol*. 2014;32(10):1045. DOI 10.1038/nbt.2979.
- Liu Y., Du H., Li P., Shen Y., Peng H., Liu S., Zhou G., Zhang H., Liu Z., Shi M., Huang X., Li Y., Zhang M., Wang Z., Zhu B., Han B., Liang C., Tian Z. Pan-genome of wild and cultivated soybeans. *Cell*. 2020;182(1):162-176. DOI 10.1016/j.cell.2020.05.023.
- Lu F., Romay M.C., Glaubitz J.C., Bradbury P.J., Elshire R.J., Wang T., Li Y., Li Y., Semagn K., Zhang X., Hernandez A.G. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun*. 2015;6:6914. DOI 10.1038/ncomms7914.
- Ma Y., Liu M., Stiller J., Liu Ch. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. *BMC Genomics*. 2019;20(1):12. <https://doi.org/10.1186/s12864-018-5357-7>.
- Marchant D.B., Soltis D.E., Soltis P.S. Genome evolution in plants. *eLS*. 2016;1-8. DOI 10.1002/9780470015902.a0026814.
- Montenegro J.D., Golicz A.A., Bayer P.E., Hurgobin B., Lee H., Chan C.K., Visendi P., Lai K., Doležel J., Batley J., Edwards D. The pangenome of hexaploid bread wheat. *Plant J*. 2017;90(5):1007-1013. DOI 10.1111/tj.13515.
- Plissonneau C., Hartmann F.E., Croll D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol*. 2018;16(1):5. DOI 10.1186/s12915-017-0457-4.
- Purugganan M.D. Evolutionary insights into the nature of plant domestication. *Curr. Biol*. 2019;29(14):R705-R714. DOI 10.1016/j.cub.2019.05.053.
- Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Minx P. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326(5956):1112-1115. DOI 10.1126/science.1178534.
- Snipen L., Almøy T., Ussery D.W. Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*. 2009;10(1):385. DOI 10.1186/1471-2164-10-385.
- Springer N.M., Ying K., Fu Y., Ji T., Yeh C.T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A.L., Barbazuk W.B., Jeddalah J.A., Nettleton D., Schnable P.S. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*. 2009;5(11):e1000734. DOI 10.1371/journal.pgen.1000734.
- Sun C., Hu Z., Zheng T., Lu K., Zhao Y., Wang W., Shi J., Wang C., Lu J., Zhang D., Li Z., Wei C. RPA: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res*. 2016;45(2):597-605. DOI 10.1093/nar/gkw958.
- Tahir Ul Qamar M., Zhu X., Xing F., Chen L.L. ppsPCP: a plant presence/absence variants scanner and pan-genome construction pipeline. *Bioinformatics*. 2019;35(20):4156-4158. DOI 10.1093/bioinformatics/btz168.
- Tao Y., Zhao X., Mace E., Henry R., Jordan D. Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant*. 2019;12(2):156-169. DOI 10.1016/j.molp.2018.12.016.
- Tettelin H., Massignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., DeBoy R.T. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. USA*. 2005;102(39):13950-13955. DOI 10.1073/pnas.0506758102.
- Tranchant-Dubreuil C., Rouard M., Sabot F. Plant pangenome: impacts on phenotypes and evolution. *Ann. Plant Rev. Online*. 2018;453-478. DOI 10.1002/9781119312994.apr0664.
- Veras A., Araujo F., Pinheiro K., Guimarães L., Azevedo V., Soares S., Costa da Silva A., Ramos R. Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes. *Sci. Rep*. 2018;8(1):1-8. DOI 10.1038/s41598-018-27800-8.
- Wang W., Mauleon R., Hu Z., Chebotarov D., Tai S., Wu Z., Li M., Zheng T., Fuentes R.R., Zhang F., Mansueto L. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018;557(7703):43. DOI 10.1038/s41586-018-0063-9.
- Wendel J.F., Jackson S.A., Meyers B.C., Wing R.A. Evolution of plant genome architecture. *Genome Biol*. 2016;17:37. DOI 10.1186/s13059-016-0908-1.
- Wing R.A., Purugganan M.D., Zhang Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet*. 2018;19:505-517. DOI 10.1038/s41576-018-0024-z.
- Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Zhou X., Lam T., Li Y., Xu X., Wong G.K., Wang J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2014;30(12):1660-1666. DOI 10.1093/bioinformatics/btu077.
- Yao W., Li G., Zhao H., Wang G., Lian X., Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol*. 2015;16:187. DOI 10.1186/s13059-015-0757-3.
- Zerbino D.R., Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. DOI 10.1101/gr.074492.107.
- Zhao M., Wang Q., Wang Q., Jia P., Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(1). DOI 10.1186/1471-2105-14-S11-S1.
- Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet*. 2018;50(2):278-284. DOI 10.1038/s41588-018-0041-z.
- Zhao Y., Sun C., Zhao D., Zhang Y., You Y., Jia X., Yang J., Wang L., Wang J., Fu H., Kang Y., Chen F., Yu J., Wu J., Xiao J. PGAP-X: extension on pan-genome analysis pipeline. *BMC Genomics*. 2018;19(1):115-124. DOI 10.1186/s12864-017-4337-7.
- Zhao Y., Wu J., Yang J., Sun S., Xiao J., Yu J. PGAP: pan-genomes analysis pipeline. *Bioinformatics*. 2012;28(3):416-418. DOI 10.1093/bioinformatics/btr655.
- Zhou L., Zhang T., Tang S., Fu X., Yu Sh. Pan-genome analysis of *Paenibacillus polymyxa* strains reveals the mechanism of plant growth promotion and biocontrol. *Antonie van Leeuwenhoek*. 2020;113:1539-1558. DOI 10.1007/s10482-020-01461-y.
- Zimin A.V., Marçais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. The MaSuRCA genome assembler. *Bioinformatics*. 2013; 29(21):2669-2677. DOI 10.1093/bioinformatics/btt476.
- Żmieńko A., Samelak A., Kozłowski P., Figlerowicz M. Copy number polymorphism in plant genomes. *Theor. Appl. Genet*. 2014;127:1-18. DOI 10.1007/s00122-013-2177-7.

ORCID ID

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

Благодарности. Работа выполнена при поддержке Российского научного фонда, грант № 18-14-00293.

Авторы благодарны Н.А. Шмакову и Д.А. Афонникову за помощь в работе над текстом статьи. Считаем своим приятным долгом поблагодарить анонимных рецензентов за ценные замечания.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 04.11.2020. После доработки 27.12.2020. Принята к публикации 03.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

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


Е.А. Урбанович¹ , Д.А. Афонников^{2, 3}, С.В. Николаев^{2, 4}

¹ Новосибирский государственный технический университет, Новосибирск, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

⁴ Московская государственная академия ветеринарной медицины и биотехнологии – МВА им. К.И. Скрябина, Москва, Россия

 e.urbanovich98@gmail.com

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

Для цитирования: Урбанович Е.А., Афонников Д.А., Николаев С.В. Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса. *Вавиловский журнал генетики и селекции*. 2021;25(1):64-70. DOI 10.18699/VJ21.008

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

Е.А. Urbanovich¹ , D.A. Afonnikov^{2, 3}, S.V. Nikolaev^{2, 4}

¹ Novosibirsk State Technical University, Novosibirsk, Russia

² Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Novosibirsk State University, Novosibirsk, Russia

⁴ Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K.I. Skryabin, Moscow, Russia

 e.urbanovich98@gmail.com

Abstract. Determining the quantitative content of chlorophylls in plant leaves by their reflection spectra is an important task both in monitoring the state of natural and industrial phytocenoses, and in laboratory studies of normal and pathological processes during plant growth. The use of machine learning methods for these purposes is promising, since these methods allow inferring the relationships between input and output variables (prediction model), and in order to improve the quality of the prediction, a researcher may modify predictors and selects a set of method

parameters. Here, we present the results of the implementation and evaluation of the random forest algorithm for predicting the total concentration of chlorophylls *a* and *b* from the reflection spectra of plant leaves in the visible and infrared wavelengths. We used the reflection spectra for 276 leaf samples from 39 plant species obtained from open sources. 181 samples were from the sycamore maple (*Acer pseudoplatanus* L.). The reflection spectrum represented wavelengths from 400 to 2500 nm with a step of 1 nm. The training set consisted of the 85 % of *A. pseudoplatanus* L. samples, and the performance was evaluated on the remaining 15 % samples of this species (validation sample). Six models based on the random forest algorithm with different predictors were evaluated. The selection of control parameters was performed by cross-checking on five partitions. For the first model, the intensity of the reflection spectra without any transformation was used. Based on the analysis of this model, the optimal ranges of wavelengths for the remaining five models were selected. The best results were obtained by models that used a two-point estimation of the derivative of the reflection spectrum in the visible wavelength range as input data. We compared one of these models (the two-point estimation of the derivative of the reflection spectrum in the range of 400–800 nm with a step of 1 nm) with the model by other authors (which is based on the functional dependence between two unknown parameters selected by the least squares method and two reflection coefficients, the choice of which is described in the article). The comparison of the results of predictions of the model based on the random forest algorithm with the model of other authors was carried out both on the validation sample of maple and on the sample from other plant species. In the first case, the predictions of the method based on a random forest had a lower estimate of the standard deviation. In the second case, the predictions of this method had a large error for small values of chlorophyll, while the third-party method had acceptable predictions. The article provides the analysis of the results, as well as recommendations for using this machine learning method to assess the quantitative content of chlorophylls in leaves. Key words: random forest; remote methods; leaf optics; pigments.

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

Введение

Пигменты – низкомолекулярные соединения, которые придают окрашивание органам растений и играют в их жизни важную роль, выполняя фотосинтетические, защитные и метаболические функции. У наземных растений наиболее известными пигментами являются хлорофиллы (обеспечивают зеленую окраску органов растений и играют важнейшую роль в фотосинтезе), каротиноиды (придают красную и желтую окраску, также участвуют в фотосинтезе), антоцианы (обеспечивают фиолетовую окраску, выполняют защитные функции), а также ряд других соединений (Croft, Chen, 2018). Фотосинтетические пигменты, хлорофиллы и каротиноиды, привлекают наибольшее внимание исследователей, они имеют разные спектры поглощения и выполняют в процессе фотосинтеза разные функции, что обуславливается структурными различиями между молекулами этих веществ.

Хлорофилл в растениях представлен молекулами двух типов, *a* и *b*, которые имеют структурные отличия и различаются по своим светопоглощающим свойствам (Du et al., 1998). Это позволяет фотосинтезирующим организмам собирать солнечный свет на различных длинах волн, чтобы максимизировать энергию света, доступную для фотосинтеза. Изменение концентраций фотосинтетических пигментов тесно связано с физиологическим состоянием растений. Например, при увядании листьев растений происходит быстрое снижение концентрации хлорофиллов по сравнению с каротиноидами, тем самым увеличивается отношение содержания каротиноидов к хлорофиллам, что вызывает появление у листьев окраски красных и желтых оттенков (Croft, Chen, 2018). Содержание пигментов, в частности хлорофиллов *a* и *b*, таким образом, может служить индикатором состояния растений в ходе нормального роста и при развитии инфекций, а также стресса, фотосинтетической активности, нарушения метаболизма и т. д. (Młodzińska, 2009). Потребности в определении физио-

логического состояния растений часто возникают в ходе решения многих научных и практических задач, поэтому методы оценки содержания пигментов в органах и тканях растений постоянно развиваются и совершенствуются.

Количественную и качественную информацию о пигментах можно получить с использованием химических методов (Lichtenthaler, 1987; Porra et al., 1989; Wellburn, 1994). Однако для многих задач более удобный подход – применение дистанционных методов на основе спектров отражения света от листа растения (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003). Отражательная способность листа в оптическом и инфракрасном (ИК) диапазонах волн (400–2500 нм) зависит от различных биохимических и физических факторов, включая содержание хлорофилла и других пигментов листьев, азота, воды, а также от внутренней структуры листьев и особенностей их поверхности (Croft, Chen, 2018). Для растительных пигментов характерно поглощение электромагнитного излучения в видимом (400–700 нм) и ближнем ИК (1300–2500 нм) диапазонах длин волн. Поглощение компонентами листа в ближней инфракрасной области в диапазоне 750–1300 нм низкое, так как в этом интервале длин волн происходит интенсивное отражение от компонентов внутренней структуры листьев. Таким образом, коэффициент отражения в ближнем ИК-диапазоне зависит и от концентрации ферментов, и от структуры листа. Все это позволяет применять методы дистанционного наблюдения как в видимом, так и ближнем ИК-диапазоне длин волн для мониторинга физиологического состояния растений (Merzlyak et al., 2003; Alt et al., 2020).

Один из подходов к оценке содержания хлорофиллов по спектру отражения заключается в подборе эмпирических зависимостей (индексов) между коэффициентами отражения на определенных длинах волн, выбор которых – также важная часть метода, и содержанием хлорофиллов (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001,

2003; Suo et al., 2010; Nikolaev et al., 2018). Успех такого «классического» подхода прямо зависит от глубины нашего понимания физики процесса.

В настоящее время в задачах предсказания характеристик биологических объектов часто применяются методы машинного обучения (Doktor et al., 2014; Feng et al., 2020). Их достоинство в том, что обычно сложную нелинейную зависимость от многих переменных можно аппроксимировать с необходимой точностью методами машинного обучения. В простых случаях на вход программы данные подаются без какой-либо обработки, тем не менее точность предсказываемого параметра будет достаточно высокой. Для каждого метода машинного обучения имеются свои способы улучшения точности предсказания, например при помощи варьирования управляющих воздействий. Существуют также способы преобразования входных данных, позволяющие улучшить результат. Так, при анализе спектров расчет производной дает возможность устранить аддитивные компоненты и выделить такие характерные особенности спектра, как положения максимумов, минимумов и точек.

Целью нашего исследования была разработка метода машинного обучения с использованием алгоритма случайного леса для предсказания суммарной концентрации хлорофиллов a и b в листьях растений по значениям спектров отражения в видимом и инфракрасном диапазонах длин волн. Проведена оценка точности предсказания в сравнении с результатами, полученными по аналитической функциональной зависимости, определены преимущества и недостатки обоих подходов.

Материалы и методы

Экспериментальные данные. Характеристики спектров отражения листьев при различных концентрациях в них хлорофиллов a и b были загружены из базы данных EcoSIS (ecosis.org), набор *angers2003* (Jacquemoud et al., 2003; Féret et al., 2008). Рассматривали 276 образцов листьев 39 видов растений. При этом 181 образец был получен при анализе листьев белого клена (*Acer pseudo-platanus* L.). Данные по спектру отражения представлены в диапазоне 400–2500 нм с шагом 1 нм. Для этого использован спектрорадиометр ASD FieldSpec; концентрации пигментов определены по методу Лихтенхелера и представлены в единицах измерения $\text{мкг}/\text{см}^2$ (см. детали в (Jacquemoud et al., 2003; Féret et al., 2008)).

Математическая постановка задачи. Пусть есть генеральная совокупность $R_\lambda^{\text{ген}}$ всех возможных коэффициентов отражения листьев растений для заданных длин волн λ и $Chl^{\text{ген}}$ – значения суммы концентрации хлорофиллов a и b , соответствующие $R_\lambda^{\text{ген}}$. Мы имеем R_λ – подвыборку из $R_\lambda^{\text{ген}}$, и Chl – значения суммы концентрации хлорофиллов a и b , соответствующие R_λ . Требуется по набору (R_λ, Chl) построить функционал $f: R_\lambda^{\text{ген}} \rightarrow Chl^{\text{ген}}$. Причем, так как этот идеализированный функционал невозможно реализовать, то получится аппроксимирующий функционал: $\tilde{f}: R_\lambda \rightarrow \widehat{Chl}$.

Построение модели предсказания методом случайного леса. Для построения функционала был выбран метод случайного леса (random forest, RF) (Breiman, 2001; Hastie et al., 2009). Он позволяет получить точность пред-

сказания целевой функции, как правило, выше, чем в случае методов линейной регрессии. Идея алгоритма заключается в применении ансамбля решающих деревьев. Каждое дерево решений в этом ансамбле задает кусочно-постоянную функцию, которая получается при минимизации функции потерь (например, среднего квадрата отклонения). Алгоритм сочетает в себе две основные идеи: метод бэггинга Бреймана (Breiman, 1996) и метод случайных подпространств, предложенный Т.К. Но (1998). В его работе использована реализация метода случайного леса из библиотеки *sklearn* (scikit-learn.org) языка Python.

Для предсказания концентраций хлорофилла методом случайного леса были взяты несколько моделей, которые отличались наборами входных данных. Каждый набор характеризовался, во-первых, интервалом длин волн, интенсивность отражения на которых принималась во внимание. Всего было рассмотрено несколько наборов интервалов: 400–2450, 400–800 нм и комбинированный набор из двух интервалов 500–600 и 680–740 нм. Во-вторых, модели отличались типом входных данных. К ним относились значения интенсивности спектров отражения на определенных длинах волн (тип данных *base*), значения первых производных спектральных кривых для этих же длин волн (тип данных *der*), значения вторых производных (тип данных *der2*). Ряд моделей базировался лишь на одном типе данных, в других были совместно несколько типов данных. Такие комбинации отмечали знаком суммирования (например, *base+der*).

Было рассмотрено шесть моделей, они обозначены как RF-(X-Y)-Z, где (X-Y) – интервалы длин волн, Z – тип модели данных: RF-(400–2450)-*base* (интенсивности спектра в интервалах длин волн 400–2450 нм); RF-(400–800)-*base* (интенсивности спектра в интервалах длин волн 400–800 нм); RF-(400–800)-*base+der* (интенсивности спектра и первые производные в интервалах длин волн 400–800 нм); RF-(400–800)-*der* (первые производные в интервалах длин волн 400–800 нм); RF-(400–800)-*der+der2* (первые и вторые производные в интервалах длин волн 400–800 нм); RF-(500–600; 680–740)-*base+der+der2* (интенсивности, первые и вторые производные в интервалах длин волн 500–600 и 680–740 нм).

В качестве аппроксимации производной спектральных кривых выступала разностная производная первого порядка с единичным приращением, которую вычисляли по формуле: $D_i = R_i - R_{i-1}$. При таком расчете для первого значения нет производной. Для упрощения во всем тексте разностная производная именуется просто как производная. Вторую производную рассчитывали как производную от производной спектральной кривой.

При настройке алгоритма случайного леса выбраны следующие управляющие параметры:

- *max_depth*: [2, 3, 4, 5, 6] – максимальная глубина дерева;
- *max_features*: [2, 7, sqrt, log2, auto] – число признаков, по которым ищется разбиение (auto – все признаки);
- *n_estimators*: [5, 10, 15, 30, 40] – число деревьев в ансамбле случайного леса;
- *random_state*: 20200605.

Указанные параметры алгоритма подбирали при помощи перекрестной проверки на пяти выборках одинакового размера, полученных из предварительно пере-

мешанной случайным образом исходной тренировочной выборки. Четыре подвыборки служили для обучения модели, а пятая – для ее тестирования. Для определения наилучших управляющих параметров результаты тестирования (средний квадрат отклонения целевого показателя – mse) были усреднены между моделями с одинаковыми управляющими параметрами (т.е. полученными во время перекрестной проверки) и отсортированы. Управляющие параметры, для которых усредненное mse – минимальное, являются наилучшими. В качестве итоговой модели выбирается одна из пяти моделей с лучшими управляющими параметрами, имеющая минимальное mse при тестировании среди моделей, полученных по методу перекрестной проверки.

Максимальная глубина деревьев выбрана равной 6, что дает $2^6 = 64$ интервала разбиения пространства параметров, при том, что длина выборки, используемая для построения модели, равна 123. Увеличение глубины могло привести к переобучению. Количество деревьев в лесу (до 40) может показаться избыточным для 123 значений выборки, но параметры каждого из решающих деревьев подбирали на разных подпространствах (так как применяется метод случайных подпространств), а размерность признаков всегда была больше количества элементов в выборке.

Следует отметить, что алгоритм, реализованный в библиотеке `sklearn`, позволяет получить информативность каждого из признаков модели и отобрать из них наиболее информативные для полученных решающих правил (Breiman, 2001; Hastie et al., 2009; Louppe et al., 2013).

Построение эмпирических функциональных зависимостей. В качестве функционала $\tilde{f}: R_\lambda \rightarrow \widehat{Chl}$ мы дополнительно выбрали эмпирическую зависимость из работы (Gitelson et al., 2003) (метод GGM, названный нами по фамилиям авторов), представленную выражением

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

где \widehat{Chl} – суммарная концентрация хлорофиллов a и b ; R_λ – коэффициент отражения на длине волны λ ; R_{NIR} – коэффициент отражения в ближнем инфракрасном диапазоне (например, на длине волны 800 нм); α и β подбираются таким образом, чтобы минимизировать выбранную функцию потерь. А.А. Gitelson с коллегами (2003) рекомендуют выбирать в качестве предикторов длины волн из диапазона $\lambda \in [525; 555] \cup [695; 725]$. По мнению авторов, достоинство этого алгоритма в том, что коэффициент R_{NIR} «корректирует» влияние структуры ткани растения на спектр отражения и позволяет распространить найденную функцию на растения с различающимся строением листа.

Сравнение методов предсказания концентрации хлорофилла. Выборка белого клена из набора данных `angers2003` была поделена случайным образом на обучающую и валидационную в соотношении 85:15. Для примененных в настоящей работе методов предсказания алгоритмом случайного леса (RF) и функциональной зависимости (GGM) оптимальные параметры подбираются на обучающей выборке. Проверка качества алгоритмов проводится на валидационной выборке, представленной белым кленом, и на выборке образцов, не относящихся к

клену. В качестве метрик для оценки точности предсказания концентраций хлорофилла были: mse , средняя абсолютная ошибка (mae) и коэффициент детерминации R^2 . Формулы для расчета метрик следующие:

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

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

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

где x – истинные значения; \hat{x} – предсказанные значения; n – количество образцов; \bar{x} – математическое ожидание для истинных значений. С точки зрения оптимизации, mae и R^2 эквивалентны. Коэффициент детерминации R^2 удобен тем, что это безразмерная величина обычно в интервале $[0; 1]$, значение $R^2 < 0$ показывает, что среднее арифметическое \bar{x} имеет лучший результат, чем предсказания построенной модели.

Результаты

Подбор параметров для метода функциональной зависимости. Для предсказания методом GGM на обучающей выборке образцов мы подбирали коэффициенты α и β уравнения (1), а также значения λ так, чтобы максимизировать значение R^2 . В качестве длины волны в ближнем инфракрасном диапазоне выбрано значение $\lambda_{NIR} = 800$ нм. Для получения коэффициентов α и β взяли линейную модель на основе метода наименьших квадратов (класс `LinearRegression` из пакета `sklearn.linear_model`). Для каждого $\lambda \in [400; 800]$ с шагом 1 нм был найден конкретный вид кривой GGM. Коэффициенты детерминации R^2 для предсказаний полученных моделей представлены на рис. 1. Наибольший коэффициент детерминации достигался на длине волны $\lambda = 705$ нм. Результат согласуется с рекомендованным диапазоном $\lambda \in [525; 555] \cup [695; 725]$ (Gitelson et al., 2003). Метод RF сравнивают с полученной на этой длине волны ($\lambda = 705$ нм) моделью GGM.

Результаты построения алгоритма на основе метода случайного леса. Характеристики точности предсказания концентраций хлорофилла (значения параметров mse , mae , R^2) для всех шести моделей на тестовой выборке образцов приведены в таблице. Методы RF-(400–800)-der и RF-(400–800)-der+der2 продемонстрировали высокую

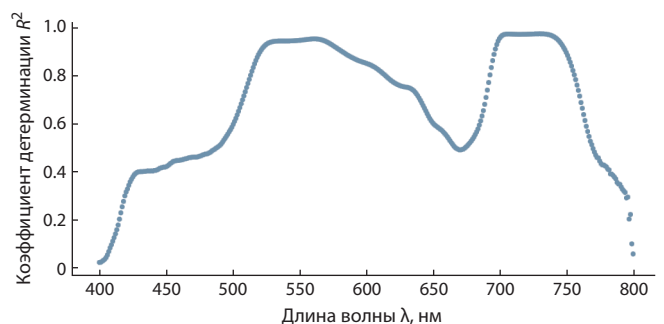


Рис. 1. Коэффициенты детерминации полученных моделей GGM при $\lambda \in [400; 800]$, которые рассчитывали на обучающей выборке.

Результаты работы модели случайного леса, обученной на различных наборах входных признаков

№ п/п	Модель случайного леса	Кол-во входных признаков	<i>mse</i>	<i>mae</i>	<i>R</i> ²
1	RF-(400–2450)-base	2051	30.5	3.7	0.945
2	RF-(400–800)-base	401	26.6	3.8	0.952
3	RF-(400–800)-base+der	401 + 400 = 801	10.1	2.4	0.981
4	RF-(400–800)-der	400	<u>9.1</u>	<u>2.4</u>	<u>0.984</u>
5	RF-(400–800)-der+der2	400 + 399 = 799	<u>8.9</u>	<u>2.3</u>	<u>0.984</u>
6	RF-(500–600; 680–740)-base+der+der2	101 + 100 + 99 + 61 + 60 + 59 = 380	10.5	2.7	0.981

Примечание. Цифрами при описании признака указан диапазон длин волн. Дополнительные характеристики признаков: base – спектр отражения; der – значения первой производной спектра; der2 – значения второй производной спектра. Курсивом выделены значения, которые имеют наилучшую точность, подчеркнутым полужирным шрифтом – наилучшую.

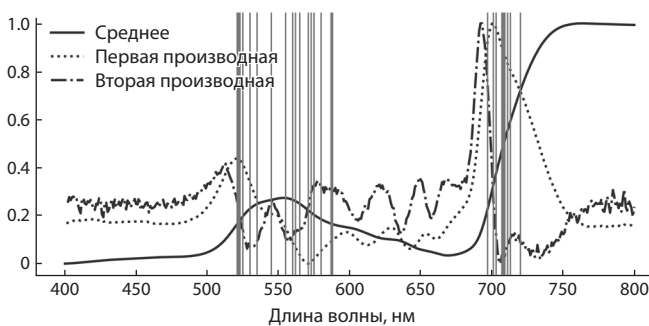


Рис. 2. Характеристики спектра отражения образцов пигментов белого клена, на которых производилось обучение моделей.

Линиями показаны: среднее значение интенсивности спектра отражения R_λ (ось Y) для различных длин волн (ось X); значение первой производной от средней интенсивности; значение второй производной. Значения производных нормированы на интервал [0; 1]. Вертикальными линиями отмечены длины волн, интенсивности спектра для которых вносят наибольший вклад в точность предсказания модели RF-(400–2450)-base.

точность предсказаний. В качестве наилучшего из них был отобран метод RF-(400–800)-der как имеющий меньшее количество входных параметров.

Отбор длин волн, коэффициенты отражения для которых брали в качестве входных признаков для предсказания концентраций хлорофилла методом случайного леса, осуществляли на основе первой модели (RF-(400–2450)-base). Это связано с тем, что сначала не было известно, нужен ли весь спектр, или только его часть, и какая именно. Как было указано ранее, алгоритм RF позволяет оценить информативность признаков, на которых происходило обучение. После настройки управляющих параметров модели RF-(400–2450)-base мы брали полученные параметры, чтобы заново обучить модели на пяти тренировочных выборках (из перекрестной проверки). Для этих пяти моделей мы выделили по 10 признаков с наибольшим вкладом в предсказание. Результаты показаны на рис. 2: вертикальными линиями представлен объединенный набор длин волн, интенсивности спектра для которых вносят наиболее значимый вклад в точность предсказания (26 длин волн из $10 \cdot 5 = 50$ возможных, если бы значения не пересекались). Интересно, что наиболее значимые признаки лежат в видимом диапазоне, большинство из этих признаков находится в диапазоне длин волн 500–600 и

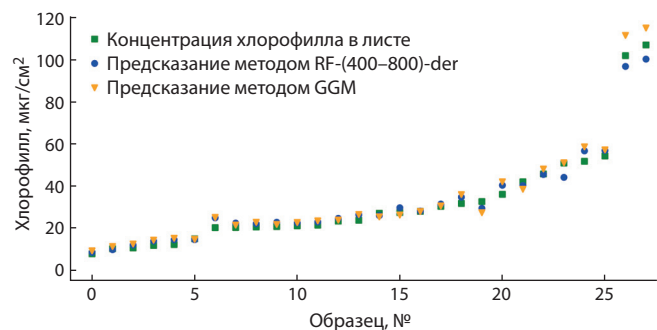


Рис. 3. Сравнение истинных и предсказанных значений концентрации хлорофилла в тканях листьев белого клена для верификационной выборки образцов.

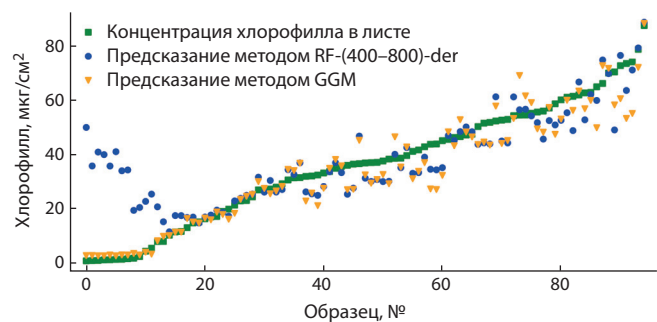


Рис. 4. Сравнение истинных и предсказанных значений концентрации хлорофилла в тканях листьев выборки образцов, не относящихся к белому клену.

680–740 нм. На основании этого нами были сформированы длины волн входных признаков для оставшихся пяти моделей предсказания методом случайного леса (см. выше).

Сравнение точности методов RF и GGM. Результаты сравнения методов предсказания концентраций хлорофилла методами RF-(400–800)-der и GGM и их экспериментально измеренные значения при разных значениях концентраций представлены на рис. 3 и 4. Для образцов белого клена (вида, взятого для подгонки параметров) метод RF-(400–800)-der показывает лучший результат по сравнению с методом GGM: $\sqrt{mse_{RF}} = 3.01$ $\mu\text{г}/\text{см}^2$ против $\sqrt{mse_{GGM}} = 3.21$ $\mu\text{г}/\text{см}^2$. При тестировании методов

на выборке листьев растений из других видов преимущество у метода функциональной зависимости GGM: $\sqrt{mse_{GGM}} = 6.31$ мкг/см² против $\sqrt{mse_{RF}} = 12.97$ мкг/см². Метод GGM демонстрирует высокую точность при малых концентрациях хлорофилла, в то время как метод RF на этих значениях показывает большую ошибку. Однако на интервале концентраций хлорофилла выше 20 мкг/см² алгоритм RF-(400–800)-deg имеет лучший результат: $\sqrt{mse_{RF}} = 5.91$ мкг/см² против $\sqrt{mse_{GGM}} = 7.01$ мкг/см².

При дальнейшем анализе выяснилось, что для образцов, у которых концентрация хлорофилла меньше 7 мкг/см², коэффициенты отражения R_{550} (максимум спектра отражения) и R_{680} (минимум спектра отражения) визуально значительно отличны от всех остальных (рис. 5, точки в верхней правой четверти). Предсказания для этих образцов имеют значительную ошибку. Тем не менее не удалось выяснить, с чем связаны различия в спектре отражения: данные образцы не отличаются от остальных ни поверхностной плотностью листа, ни эквивалентной толщиной воды для листа (leaf equivalent water thickness) (Jacquemound et al., 2003). Шесть из десяти видов растений из этих образцов имеют также образцы с нормально предсказанными значениями. Дальнейший анализ причин аномального спектра затруднен, так как данные взяты из открытых источников, а сами измерения проводили более 17 лет назад.

Обсуждение

Во многих работах по применению спектров отражения для оценки концентраций пигментов задействуют нейронные сети (Golhani et al., 2018), в то же время в исследовательских задачах по машинному обучению также распространены методы, основанные на деревьях решений. Мы задействовали метод деревьев решений для предсказания концентраций хлорофилла в листьях растений и сравнили результаты с методом функциональной зависимости. Нами обнаружены диапазоны спектра, интенсивность отражения в которых наиболее сильно влияет на точность предсказания методом случайного леса.

Диапазон 690–750 нм в литературе называется красным краем фотосинтеза (Curran et al., 1990; Gitelson et al., 2003; Croft, Chen, 2018), а окрестность 550 нм, где находится максимум спектра отражения хлорофилла, известна как зеленый край (green edge) (Gitelson et al., 2003). Как видно из рис. 2, в нашем исследовании эти области содержат наиболее важные предикторы для метода случайного леса. Выбор в качестве входных признаков более узкого диапазона длин волн видимого спектра (400–800 нм) по сравнению с полными исходными данными (400–2450 нм) повысил качество модели. Объяснением является то, что после разделения выборки на подпространства некоторые из них оказываются менее пригодными для обучения, и обученные на этих значениях деревья вносят ошибку в суммарный результат. Наибольшего эффекта удалось добиться с применением производных спектральных зависимостей.

Метод случайного леса RF хорошо проявил себя при работе с образцами белого клена, в то время как функциональная зависимость GGM отлично показала себя при работе с разными видами растений. Это связано с боль-

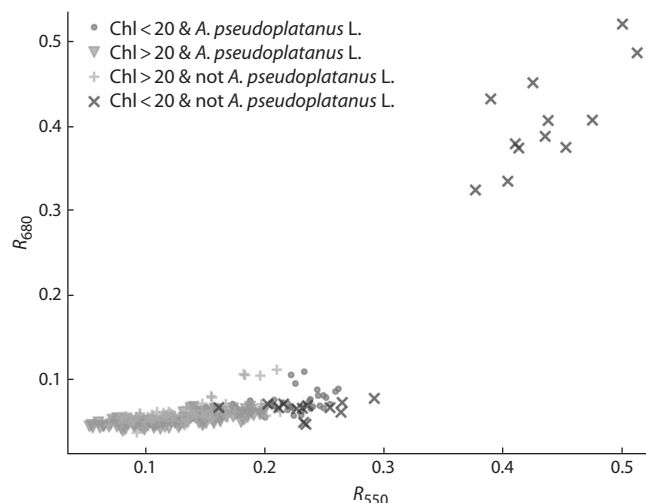


Рис. 5. Диаграмма рассеяния коэффициентов отражения R_{680} от R_{550} с выделенными категориями по концентрации хлорофилла (менее/более 20 мкг/см²) и по виду растения (*A. pseudoplatanus* L. или др.).

шей обобщающей способностью метода GGM, так как он имеет меньшее количество настраиваемых параметров. Вместе с тем более низкая точность методов RF на образцах из других видов растений частично объясняется с небольшим размером обучающей выборки и тем, что в ней представлен лишь один вид. Так, например, лучшие результаты метода случайного леса достигались при глубине деревьев, равной 5 или 6, а для этого требуется минимум 32 или 64 объекта обучающей выборки, в то время как для функционального метода (1) требуется минимум две точки (желательно, точку при малых значениях хлорофилла и точку – при больших значениях). По-видимому, эту особенность метода RF можно будет устранить с помощью большего количества обучающих данных с образцами из разных видов растений.

Тем не менее процедура отбора параметров для метода RF показала, что наиболее значимые для предсказания признаки лежат в видимой области, однако влияние структуры растения в этом методе не принималось во внимание. Наряду с этим в функциональной зависимости (1) структура ткани растения учитывается членом R_{NIR} . Если эксперимент проводится с разными видами растений (см. рис. 4), то при малых значениях хлорофилла структура растения начинает играть значительную роль.

Интересно, что оба метода работают в диапазоне $\lambda \in [525; 555] \cup [695; 725]$. Они работают на спаде производной спектра отражения, что демонстрирует рис. 2.

Слово «случайный» в названии метода «случайный лес» может привести к мысли, что при смене случайного параметра, используемого алгоритмом, можно получить кардинально другие результаты. Полагаем, что при обоснованно выбранных управляющих параметрах, разумном разбиении на обучающую и проверочную выборки такая вероятность невелика. В нашем случае для каждого набора входных признаков строили по 625 моделей (перебор из множества 125 сочетаний управляющих параметров, и по 5 моделей на перекрестной проверке для каждого сочетания). К тому же из приведенной выше

таблицы следует, что методы RF-(400–800)-base+der, RF-(400–800)-der, RF-(400–800)-der+der2 имеют близкие результаты (и, что важно, имеют *mse* меньше, по сравнению с методом GGM), это косвенно подтверждает, что результаты радикально не изменятся.

Заключение

Метод случайного леса – один из алгоритмов построения функциональных зависимостей методами машинного обучения. Поэтому его можно применять для массового автоматического построения функций, связывающих наблюдаемые признаки с искомым в задачах мониторинга. Результаты настоящей работы показали, что использовать алгоритм случайного леса (и ему подобные) в задаче определения содержания хлорофилла в листе растения целесообразно, если имеется большая выборка, минимум 32 элемента, представленная широким диапазоном концентрации хлорофилла, при этом структура ткани листа меняется слабо (к примеру, применение алгоритма только на тех растениях, на которых он был обучен). В остальных случаях лучше отдать предпочтение методам, основанным на эмпирических зависимостях (как рассмотренный здесь метод GGM).

Список литературы / References

- Alt V.V., Gurova T.A., Elkin O.V., Klimenko D.N., Maximov L.V., Pestunov I.A., Dubrovskaya O.A., Genaev M.A., Erst T.V., Genaev K.A., Komyshev E.G., Khlestkin V.K., Afonnikov D.A. The use of Specim IQ, a hyperspectral camera, for plant analysis. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2020;24(3):259-266. DOI 10.18699/VJ19.587. (in Russian)
- Breiman L. Bagging predictors. *Mach. Learn.* 1996;24:123-140. DOI 10.1023/A:1018054314350.
- Breiman L. Random forests. *Mach. Learn.* 2001;45(1):5-32. DOI 10.1023/A:1010933404324.
- Croft H., Chen J. Leaf pigment content. In: Liang S. (Ed.). *Comprehensive Remote Sensing*. Oxford, UK: Elsevier, 2018;117-142. DOI 10.1016/B978-0-12-409548-9.10547-0.
- Curran P.J., Dungan J.L., Gholz H.L. Exploring the relationship between reflectance red edge and chlorophyll content in slash pine. *Tree Physiol.* 1990;7:33-48. DOI 10.1093/treephys/7.1-2-3-4.33.
- Doktor D., Lausch A., Spengler D., Thurner M. Extraction of plant physiological status from hyperspectral signatures using machine learning methods. *Remote Sens.* 2014;6(12):12247-12274. DOI 10.3390/rs61212247.
- Du H., Fuh R.-C. A., Li J., Corkan L.A., Lindsey J.S. PhotochemCAD: A computer-aided design and research tool in photochemistry. *Photochem. Photobiol.* 1998;68:141-142. DOI 10.1111/j.1751-1097.1998.tb02480.x.
- Feng X., Zhan Y., Wang Q., Yang X., Yu C., Wang H., He Y. Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *Plant J.* 2020;101(6):1448-1461. DOI 10.1111/tpj.14597.
- Féret J.-B., François C., Asner G.P., Gitelson A.A., Martin R.E., Bidal L.P.R., Ustin S.L., le Maire G., Jacquemoud S. PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sens. Environ.* 2008;112:3030-3043. DOI 10.1016/j.rse.2008.02.012.
- Gitelson A.A., Gritz Y., Merzlyak M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 2003;160(3):271-282. DOI 10.1078/0176-1617-00887.
- Gitelson A.A., Merzlyak M.N., Chivkunova O.B. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* 2001;74(1):38-45. DOI 10.1562/0031-8655(2001)074<0038:OPANEO>2.0.CO;2.
- Golhani K., Balasundram S.K., Vadmalai G., Pradhan B. A review of neural networks in plant disease detection using hyperspectral data. *Inf. Process. Agric.* 2018;5:354-371. DOI 10.1016/j.inpa.2018.05.002.
- Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2009. DOI 10.1007/978-0-387-84858-7.
- Ho T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998;20(8):832-844. DOI 10.1109/34.709601.
- Horler D.N.H., Dockray M., Barber J. The red edge of plant leaf reflectance. *Int. J. Remote Sens.* 1983;4:273-288. DOI 10.1080/01431168308948546.
- Jacquemoud S., Bidal L., Francois C., Pavan G. ANGERS Leaf Optical Properties Database. 2003. Data set. Available online [ecosis.org] from the Ecological Spectral Information System (EcoSIS), 2003.
- Keskitalo J., Bergquist G., Gardeström P., Jansson S. A cellular timetable of autumn senescence. *Plant Physiol.* 2005;139:1635-1648. DOI 10.1104/pp.105.066845.
- Lichtenthaler H.K. Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. *Methods Enzymol.* 1987;148:350-382. DOI 10.1016/0076-6879(87)48036-1.
- Loupe G., Wehenkel L., Suter A., Geurts P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* 2013;26:431-439.
- Merzlyak M.N., Gitelson A.A., Chivkunova O.B., Solovchenko A.E., Pogosyan S.I. Application of reflectance spectroscopy for analysis of higher plant pigments. *Rus. J. Plant Physiol.* 2003;50(5):704-710. DOI 10.1023/A:1025608728405.
- Młodzińska E. Survey of plant pigments: molecular and environmental determinants of plant colors. *Acta Biol. Crac. Ser. Bot.* 2009;51(1):7-16.
- Nikolaev S.V., Urbanovich E.A., Shayapov V.R., Orlova E.A., Afonnikov D.A. A method of evaluating the absorption spectrum of wheat leaf by the spectrum of diffuse reflection. *Sibirskii Vestnik Sel'skokhozyaistvennoi Nauki = Siberian Herald of Agricultural Science*. 2018;48(5):68-76. DOI 10.26898/0370-8799-2018-5-9. (in Russian)
- Porra R.J., Thompson W.A., Kriedemann P.E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: Verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *BBA – Bioenergetics*. 1989;975:384-394. DOI 10.1016/S0005-2728(89)80347-0.
- Suo X.-M., Jang Y.-T., Yang M., Li S.-K., Wang K.-R., Wang C.-T. Artificial neural network to predict leaf population chlorophyll content from cotton plant images. *Agric. Sci. China*. 2010;9(1):38-45.
- Wellburn A.R. The spectral determination of chlorophylls *a* and *b*, as well as total carotenoids, using various solvents with spectrophotometers of different resolution. *J. Plant Physiol.* 1994;144:307-313. DOI 10.1016/S0176-1617(11)81192-2.

ORCID ID

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

Благодарности. Работа поддержана грантом РФФИ № 17-29-08028 и бюджетным проектом № 0259-2021-0009.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 15.10.2020. После доработки 14.12.2020. Принята к публикации 15.12.2020.

Английский текст <https://vavilov.elpub.ru/jour>

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

А.Ю. Пронозин¹, А.А. Паулиш², Е.А. Заварзин², А.Ю. Приходько², Н.М. Прохошин², Ю.В. Кручинина^{1, 3},
Н.П. Гончаров^{1, 4}, Е.Г. Комышев^{1, 2, 3}, М.А. Генаев^{1, 2, 3} 

¹ Федеральное исследовательское учреждение Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

³ Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия


⁴ Новосибирский государственный аграрный университет, Новосибирск, Россия

 mag@bionet.nsc.ru

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

Для цитирования: Пронозин А.Ю., Паулиш А.А., Заварзин Е.А., Приходько А.Ю., Прохошин Н.М., Кручинина Ю.В., Гончаров Н.П., Комышев Е.Г., Генаев М.А. Автоматическое фенотипирование морфологии колоса тетра- и гексаплоидных видов пшеницы методами компьютерного зрения. *Вавиловский журнал генетики и селекции*. 2021; 25(1):71-81. DOI 10.18699/VJ21.009

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


A.Yu. Pronozin¹, A.A. Paulish², E.A. Zavarzin², A.Yu. Prikhodko², N.M. Prokhoshin², Yu.V. Kruchinina^{1, 3},
N.P. Goncharov^{1, 4}, E.G. Komyshev^{1, 2, 3}, M.A. Genayev^{1, 2, 3} 

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

² Novosibirsk State University, Novosibirsk, Russia

³ Kurchatov Genomics Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

⁴ Novosibirsk State Agrarian University, Novosibirsk, Russia

 mag@bionet.nsc.ru

Abstract. Intraspecific classification of cultivated plants is necessary for the conservation of biological diversity, study of their origin and their phylogeny. The modern cultivated wheat species originated from three wild diploid ancestors as a result of several rounds of genome doubling and are represented by di-, tetra- and hexaploid species. The

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

Key words: wheat spike morphology; wheat; phenomics; image processing; computer vision; machine learning; biotechnology.

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

Введение

Ряд важных вопросов, включая аспекты эффективного сохранения биологического разнообразия видов возделываемых растений, изучение их происхождения, определение их филогении, предполагает детальную разработку внутривидовых классификаций (Дорофеев и др., 1979; Goncharov, 2011). Создание таких классификаций, отражающих филогенез и генетическую структуру видов, следует считать основной целью современной таксономии (Hammer et al., 2011). При разработке классификации культурных растений предполагается максимальное полное описание всех существующих крупных и мелких форм (таксонов) (Синская, 1969). Это определяется удобством применения такого деления, с одной стороны, в экспериментальной работе, с другой, при селекции и апробации сельскохозяйственных растений.

Успех и эффективность исследовательской работы часто связаны с детальностью и полнотой экспериментальной проработки, которая зависит от того, каков материал и насколько подробно его следует изучать. В связи с этим исключительно важно, чтобы естественная дифференциация того или иного рода, взаимосвязи между видами с высокой точностью были отражены детально разработанной таксономией (Дорофеев, 1985). Следует заметить, что у большей части растений, важных для сельского хозяйства, до настоящего времени однозначно не определены объемы рода и видов (Родионов и др., 2019).

Серьезная проблема систематики культурных растений – аспект укрупнения–дробления таксонов, причем в случаях возделывания она проявляется особенно контрастно (Головнина и др., 2009; Goncharov, 2011). В то же время эффективное использование классификаций (систем родов) возделываемых растений в работе исследователей вызывает определенные сложности. И дихотомические таблицы (Дорофеев и др., 1979; Гончаров, 2009), и идеографические определители (Зуев и др., 2019) требуют определенных навыков, поэтому создание баз данных и

программного обеспечения, позволяющее по цифровым изображениям определять видовую принадлежность, – очень перспективное направление. Разработка данных методов основывается преимущественно на технологиях анализа цифровых изображений органов растений в рамках компьютерной феномики (Афонников и др., 2016; Zhang et al., 2019; Демидчик и др., 2020; Yang et al., 2020).

Пшеница – одна из важнейших мировых продовольственных культур. Современные возделываемые виды пшениц произошли от трех диких диплоидных предков в результате их гибридизации и нескольких раундов удвоения генома (полиплоидизации). В настоящее время культивируемые пшеницы представлены ди- ($2n = 2x = 14$, геном A^bA^b), тетра- ($2n = 4x = 28$, геном BBA^uA^uDD) и гексаплоидными ($2n = 6x = 42$, геном BBA^uA^uDD) видами (Гончаров, Кондратенко, 2008). Основной возделываемый вид, мягкая пшеница (*Triticum aestivum* L.), является гексаплоидом (геномная формула BBA^uA^uDD). Уровень плоидности служит одним из основных таксономических и классифицирующих признаков видов пшениц (Международный классификатор..., 1984; van Slageren, Payne, 2013). Его можно устанавливать цитогенетическими (Родионов и др., 2020), молекулярными методами, а также на основе сравнения морфологических характеристик растений (Международный классификатор..., 1984). В нашей работе проведено изучение морфологических характеристик колосьев растений тетра- и гексаплоидных видов пшеницы с применением метода высокопроизводительного фенотипирования.

Целью исследования было изучение распределения морфологических характеристик колосьев у тетра- и гексаплоидных видов пшениц и сравнение их распределений.

Материалы и методы

Биологический материал. В работе изучено 17 видов полиплоидных пшениц: девять гексаплоидных (*Triticum compactum* Host, *T. aestivum* L., *T. antiquorum* Heer ex

Таблица 1. Описание используемых видов пшениц

Вид	Всего					Плоидность	Список вегетаций	
	фотографий	растений	образцов	на столе	на прищепке			
<i>T. compactum</i> Host	472	101	10	177	295	Hexaploid	II18, IX16	
<i>T. aestivum</i> L.	456	80	8	166	290		II19, IX16, IX18, X14	
<i>T. antiquorum</i> Heer ex Udacz.	184	37	4	116	68		II18, X14	
<i>T. spelta</i> L.	164	49	5	40	124		II18	
<i>T. petropavlovskiyi</i> Udacz. et Migusch.	374	75	6	74	300		II17, IX17, IX18	
i:АНК-23	50	10	1	14	36		IX16	
<i>T. yunnanense</i> King ex S.L. Chen	191	43	3	43	148		IX17, IX18	
<i>T. spelta</i> cv. Rother Sommer Kolben	45	9	1	9	36		IX16, II18	
<i>T. macha</i> Dekapr. et Menabde	46	10	1	10	36		IX17, IX18	
<i>T. sphaerococcum</i> Perciv.	100	20	2	20	80		IX17	
<i>T. vavilovii</i> (Thum.) Jakubz.	15	3	1	3	12		II18	
<i>T. aethiopicum</i> Jakubz.	595	119	12	119	476		Tetraploid	X14
<i>T. dicoccoides</i> (Körn. ex Aschers. et Graebn.) Schweinf.	40	8	1	8	32			II16
<i>T. dicoccum</i> (Schränk) Schuebl.	41	9	1	9	32	II17		
<i>T. durum</i> Desf.	275	56	5	55	220	II16, II17, II19, IX18		
<i>T. militinae</i> Zhuk. et Migusch.	40	8	1	8	32	IX17		
<i>T. polonicum</i> L.	95	19	2	19	76	II16, II19		
<i>T. timopheevii</i> (Zhuk.) Zhuk.	125	25	3	25	100	II16, IX18		
<i>T. turgidum</i> L.	40	8	1	8	32	II15		

Udacz., *T. spelta* L. (включая стародавний сорт *T. spelta* Rother Sommer Kolben), *T. petropavlovskiyi* Udacz. et Migusch., *T. yunnanense* King ex S.L. Chen, *T. macha* Dekapr. et Menabde, *T. sphaerococcum* Perciv., *T. vavilovii* (Thum.) Jakubz.), изогенная линия сорта мягкой пшеницы Новосибирская 67 АНК-23 и восемь тетраплоидных (*T. aethiopicum* Jakubz., *T. dicoccoides* (Körn. ex Aschers. et Graebn.) Schweinf., *T. dicoccum* (Schränk) Schuebl., *T. durum* Desf., *T. militinae* Zhuk. et Migusch., *T. polonicum* L., *T. timopheevii* (Zhuk.) Zhuk., *T. turgidum* L.); выборка представлена колосьями 595 уникальных и выращенных в девяти вегетациях растений. Растения были выращены в 2014–2019 гг. в гидропонной теплице центра коллективного пользования «Лаборатория искусственного выращивания растений» ИЦиГ СО РАН (г. Новосибирск). Материал, использованный в работе, описан в табл. 1.

Нужно отметить, что ни в одном из крупных генетических банков мира не существует типовых коллекций по пшенице, поэтому выборки, как правило, либо отражают взгляд исследователей на проблему создания таких коллекций (Пальмова, 1935), либо определяются репрезентативностью доступного для исследования материала (Гончаров, Шумный, 2008). Стандартное таксономическое описание образцов есть в публичных базах данных на сайтах генбанков (<http://db.vir.nw.ru/virdb/maindb>).

Получение цифровых изображений. В работе использовали два протокола получения фотографий зрелых колосьев. Первый – колос располагается на стекле просветного столика, который находится на столе с поверхностью синего цвета (фон). Фотокамера фиксируется на

стойке над стеклом. С помощью данного метода можно производить съемку лицевой проекции колоса. Второй – колос располагается вертикально перед синим фоном. Опорой колоса служат прищепки, которые помещаются на штатив. Применяя этот метод, при вращении колоса относительно его оси, можно производить съемку колоса в четырех или более проекциях (Генаев и др., 2018). Согласно протоколам, на фотографиях должна присутствовать цветовая палитра (ColorChecker). Она нужна для нормализации цветов и определения масштаба. Одному растению в нашем наборе данных может соответствовать до пяти фотографий его колоса, снятых по разным протоколам и в разных проекциях. Примеры изображений колосьев (по одному для каждого вида) представлены на рис. 1. Всего с использованием двух протоколов было получено 3348 изображений колосьев в разных проекциях, 2097 из них относились к гексаплоидным видам, а 1251 – к тетраплоидным, из них 915 изображений получены по протоколу «на столе» и 2433 – «на прищепке».

Оценка количественных характеристик колосьев. Для всесторонней оценки таких характеристик на основе анализа изображений использовали программу WERecognizer (Genaev et al., 2019). Эта программа описывает колос пшеницы в виде модели двух четырехугольников (рис. 2). Геометрия данной модели описывается девятью независимыми параметрами. Для верхнего четырехугольника – это параметры x_{u1} , x_{u2} , y_{u1} , y_{u2} ; для нижнего четырехугольника – x_{b1} , x_{b2} , y_{b1} , y_{b2} ; общий параметр для двух четырехугольников – длина колоса. Дополнительно программа рассчитывает ряд общих ха-

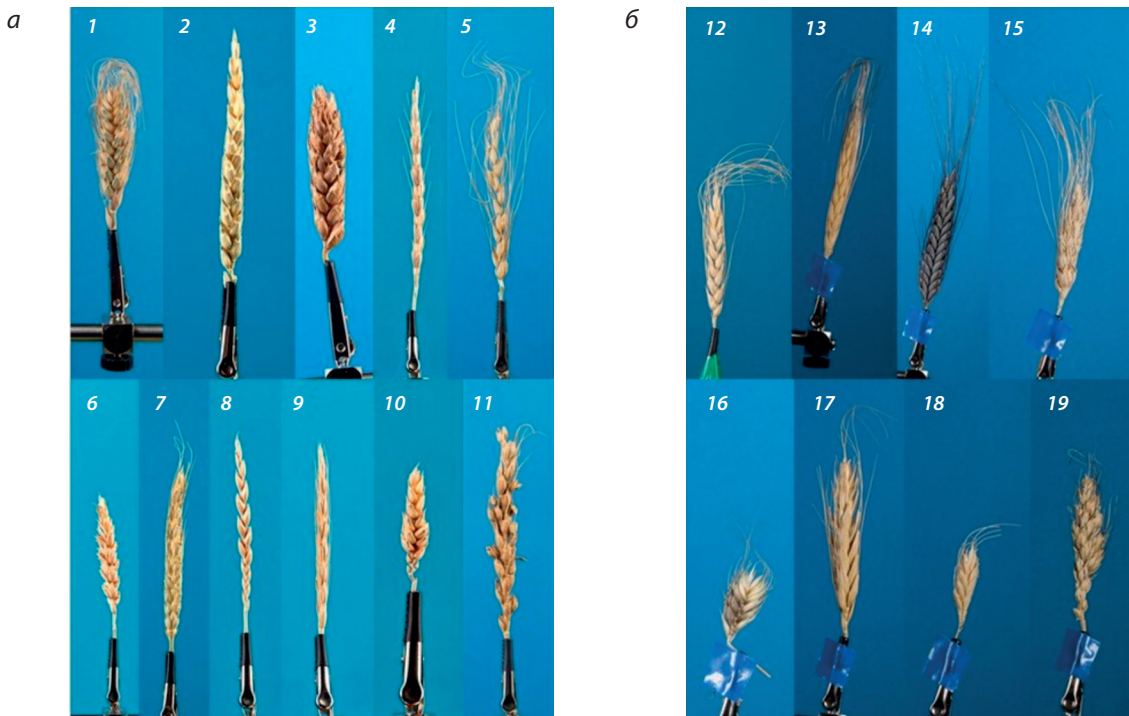


Рис. 1. Изображения колосьев различных видов гексаплоидных (а) и тетраплоидных (б) пшениц.

1 – *T. compactum*; 2 – *T. aestivum*; 3 – *T. antiquorum*; 4 – *T. spelta*; 5 – *T. petropavlovskiyi*; 6 – i:АНК-23; 7 – *T. yunnanense*; 8 – *T. spelta* cv. Rother Sommer Kolben; 9 – *T. macha*; 10 – *T. sphaerococcum*; 11 – *T. vavilovii*; 12 – *T. aethiopicum*; 13 – *T. dicoccoides*; 14 – *T. dicoccum*; 15 – *T. durum*; 16 – *T. militinae*; 17 – *T. polonicum*; 18 – *T. timopheevii*; 19 – *T. turgidum*.

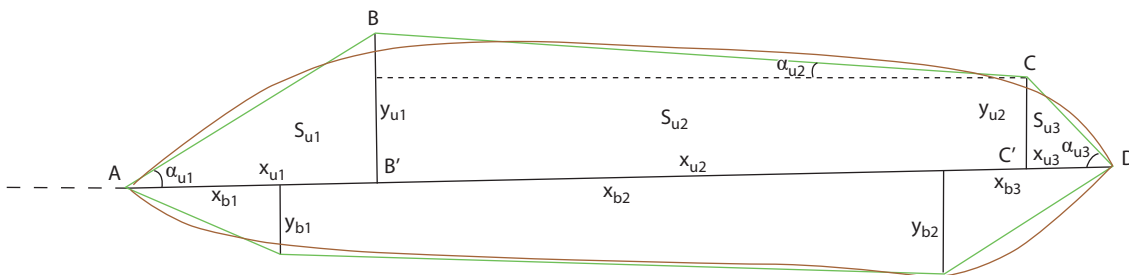


Рис. 2. Представление формы колоса в виде двух четырехугольников (Genaev et al., 2019).

Черной горизонтальной линией обозначена осяевая линия колоса. Его контур показан коричневой линией. Четырехугольники, аппроксимирующие контур колоса, отмечены зелеными линиями. Штриховая линия слева – его основание. Для верхнего четырехугольника указаны основные параметры, характеризующие его геометрию. Аналогичные параметры определяются и для нижнего четырехугольника.

рактических характеристик формы и размера колоса, а также параметры его остистости. Детали алгоритма извлечения признаков приведены в статье (Genaev et al., 2019).

Мы использовали признаки модели, которые были отобраны как наиболее информативные для предсказания индекса плотности колоса в нашем предыдущем исследовании (Genaev et al., 2019), а также общие признаки формы и остей. Эти признаки характеризуют комплексное представление о морфологии (фенотипе) колоса, описывая его форму (Circularity, Roundness), физические размеры тела колоса (Perimeter, Rachis length) и площадь остей (Awns area). Признаки, полученные в результате аппроксимации колоса двумя четырехугольниками, связаны с шириной (x_{u2} , y_{bm}) и длиной (x_{b2} , y_{u2}) отдельных сегментов колоса (табл. 2).

Анализ данных. Для того чтобы оценить распределение колосьев в пространстве изучаемых признаков, мы воспользовались нелинейным алгоритмом снижения размерности t-SNE (t-distributed stochastic neighbor embedding; Maaten, Hinton, 2008). Этот метод позволяет визуализировать многомерные данные путем отображения объектов в многомерном пространстве в пространство меньшей размерности (двух- или трехмерное). Основная идея t-SNE заключается в уменьшении размерности пространства при сохранении относительных попарных расстояний между объектами. Преимуществом метода t-SNE является склонность к локализации изолированных плотных пространственных структур произвольной геометрии. Метод t-SNE был применен для ординации изображений колосьев; при этом изображения каждой из

Таблица 2. Описание признаков колоса

Название признака	Описание	Размерность
Awns area	Площадь остей	мм ²
Circularity	Индекс округлости равен отношению периметра окружности с площадью, равной площади контура, к периметру контура. Индекс отражает, насколько форма контура близка к форме окружности. Значение варьирует от 0 до 1	Безразмерная
Roundness	Индекс закругленности равен отношению площади контура к площади окружности с диаметром, равным осевой линии колоса	
Perimeter	Периметр контура колоса без остей	мм
Rachis length	Длина ломаной линии вдоль оси сложного колоса (линии оси колоса)	
X _{u2}	Параметр модели четырехугольников, связанный с длиной левой центральной части колоса (вверху на рис. 2)	
X _{b2}	Параметр модели четырехугольников, связанный с длиной правой центральной части колоса (внизу на рис. 2)	
Y _{u2}	Расстояние от вершины С до ее проекции С' на основание AD (см. рис. 2)	
Y _{bm}	Параметр модели четырехугольников. Среднее значение высоты правого (нижнего) четырехугольника	

проекций одного колоса рассматривались как отдельные объекты.

Для оценки сходства количественных характеристик колосьев для разных видов мы применили иерархическую кластеризацию (Johnson, 1967) 17 видов пшеницы по признакам, полученным в результате усреднения по всем колосьям одного и того же вида. При этом каждый вид был охарактеризован вектором признаков длины 9. В качестве метрики расстояния между видами было значение $1 - r$, где r – величина коэффициента корреляции Пирсона между значениями признаков (Müllner, 2011). Для кластеризации и построения дендрограммы использованы соответственно функции linkage (алгоритм UPGMA) и dendrogram из библиотеки SciPy (Virtanen et al., 2020).

Для сравнения дисперсий признаков у растений, относящихся к разным типам плоидности, мы применили F -статистику (Snedecor, Cochran, 1989), которая оценивает значимость различий дисперсий двух распределений. Данные нормированы функцией StandardScaler библиотеки scikit-learn (Pedregosa et al., 2011). Тест проводили независимо для каждого из девяти признаков, описанных в табл. 2. При выполнении этого теста для каждого растения взято одно изображение колоса, полученное в проекции по протоколу «на столе».

Результаты и обсуждение

Средние значения, медиана, среднее квадратическое отклонение и дисперсия девяти признаков, рассчитанных для 17 видов пшениц, представлены в Приложении 1¹.

Рассмотрим распределение колосьев нашей выборки растений по признаку «площадь остей». Чем выше этот параметр, тем больше остей было идентифицировано для колоса на изображении. По этой характеристике колосья гексаплоидных пшениц условно можно разделить на три класса: остистые (значение параметра выше 90), умеренно остистые (значение параметра от 30 до 90) и безостые

(значение параметра менее 30). К остистым по такому критерию относятся представители видов *T. compactum*, *T. spelta*, *T. petropavlovskiyi*, *T. vavilovii*; к умеренно остистым – *T. aestivum*, *T. yunnanense*, *T. macha*; к безостым – *T. antiquorum*, i:АНК-23, *T. spelta* cv. Rother Sommer Kolben, *T. sphaerococcum* (см. Приложение 1). Эти данные хорошо согласуются с внешним видом колосьев (см. рис. 1, а). Таким образом, образцы гексаплоидных пшениц демонстрируют значительное разнообразие по наличию/отсутствию остей.

Если применить указанную классификацию для тетраплоидных пшениц, то к категории безостых относится лишь вид *T. militinae* (среднее значение параметра 24.09 мм²). Четыре вида можно причислить по этому признаку к умеренно остистым: *T. dicoccoides*, *T. polonicum*, *T. timopheevii*, *T. turgidum*; три вида – к остистым: *T. aethiopicum*, *T. dicoccum*, *T. durum*. В целом представленность остистых видов (образцов) у тетраплоидных пшениц значительно выше, чем у гексаплоидных.

Анализ такой характеристики, как длина колоса, показывает, что колосья можно также разделить на три класса: длина менее 60 мм (короткие), от 60 до 90 мм (средние) и более 90 мм (длинные). По данной классификации виды гексаплоидных пшениц *T. spelta*, *T. petropavlovskiyi* и *T. vavilovii* можно отнести к длинноколосым; *T. aestivum*, *T. yunnanense*, *T. spelta* cv. Rother Sommer Kolben и *T. macha* – к среднеколосым, а *T. compactum*, *T. antiquorum*, *T. sphaerococcum* и изогенную линию АНК-23 – к короткоколосым. При этом граница между видами, характеризующимися длинными и средними колосьями, оказывается достаточно условной. У тетраплоидных видов мы не обнаружили ни одного вида, который попал бы по этому параметру в категорию длинноколосых. К категории среднеколосых можно отнести *T. aethiopicum*, *T. dicoccoides*, *T. polonicum* и *T. turgidum*, к категории короткоколосых – *T. dicoccum*, *T. durum*, *T. timopheevii* и *T. militinae*.

На рис. 3, а приведено распределение изученных образцов по длине колосьев для гекса- и тетраплоидных видов; на рис. 3, б показано распределение параметра, также

¹ Приложения 1–4 см. по адресу: <http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx3.pdf>

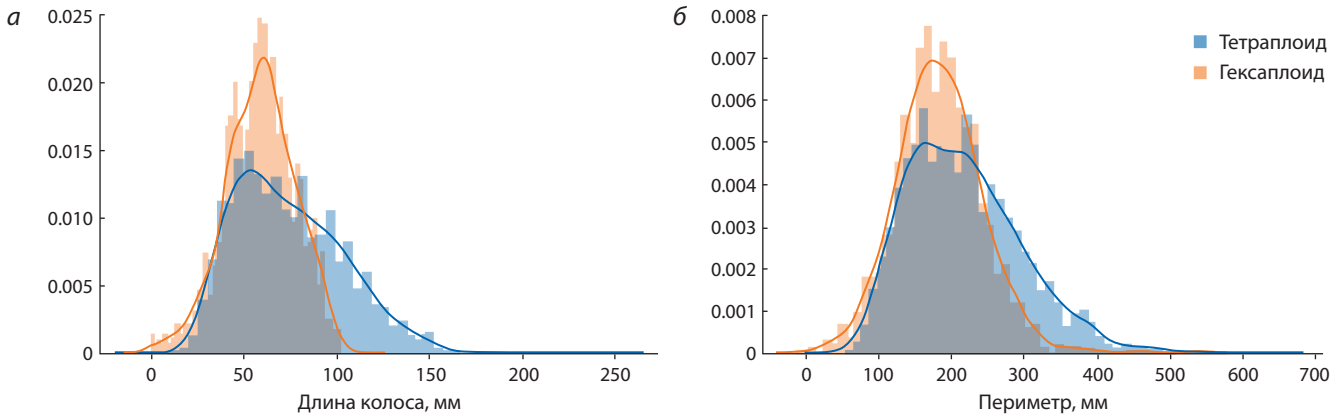


Рис. 3. Распределение длины (а) и периметра (б) колоса у тетраплоидных (синий цвет) и гексаплоидных (оранжевый цвет) видов пшениц.

характеризующего размер колоса, – периметра контура тела колоса на изображении.

Согласно рис. 3, распределения обоих параметров у гексаплоидных пшениц имеют более высокий разброс, при этом изменчивость этих признаков у гексаплоидных пшениц выше в основном за счет большей частоты встречаемости колосьев с большими значениями этих признаков.

Распределение проанализированных изображений колосьев в пространстве девяти характеристик мы визуализировали при помощи метода t-SNE, получив на его основе двумерное пространство параметров (компоненты 1 и 2). Результаты преобразования приведены на рис. 4. На диаграмме каждая точка представляет одно из проанализированных изображений колоса. На рис. 4, а точки окрашены в соответствии с типом плоидности растения (синий цвет – тетраплоидные виды пшеницы, оранжевый – гексаплоидные); на рис. 4, б цвет и форма каждой точки соответствуют определенному виду пшеницы.

На диаграмме рис. 4, а показано, что области, которые занимают образцы гекса- и тетраплоидных видов пшеницы, сильно перекрываются. Следовательно, по своим характеристикам колосья этих двух групп достаточно близки. Это согласуется с результатами, представленными в Приложениях 1 и 2, а также на рис. 3. Однако нужно отметить, что на диаграмме рис. 4, а образцы гексаплоидных видов занимают большую площадь, прежде всего, за счет преобладания соответствующих точек в правой части диаграммы. Видно, что в области при значениях компоненты 1 более -20 преобладают оранжевые точки (гексаплоидные пшеницы), такое превалирование еще более заметно в верхнем правом углу диаграммы (значения компоненты 1 менее 0, а компоненты 2 – больше 20). Это означает, что ряд характеристик колоса наблюдается лишь у гексаплоидных видов, но не у тетраплоидных, что хорошо согласуется с результатом, приведенным на рис. 3. В частности, такие области могут соответствовать большим значениям параметров «периметр» и «длина колоса».

На диаграмме рис. 4, б продемонстрировано, что области, занятые образцами разных видов, в большой степени перекрываются. Например, видам *T. aestivum* и *T. durum* соответствует значительная площадь на графике (см. рис. 4, б, штриховая линия). Однако следует отметить, что

изображения колосьев, принадлежащих к одному виду пшеницы, занимают преимущественно компактные области на графике. В то же время существуют виды, для которых образцы колосьев распадаются по своим характеристикам на несколько хорошо заметных кластеров. К таким видам можно отнести *T. compactum* (метки: малый синий круг, компонента 1 от -60 до 0, компонента 2 от -60 до 0) и *T. petropavlovskiyi* (фиолетовый треугольник, компонента 1 от -20 до 0, компонента 2 от 40 до 80).

На рис. 1 гексаплоиды представлены растениями с двумя характерными типами колосьев: длинными и тонкими (*T. aestivum*, *T. spelta*, *T. petropavlovskiyi*, *T. yunnanense*, *T. spelta* cv. Rother Sommer Kolben); короткими и округлыми (*T. compactum*, *T. antiquorum*, i:АНК-23, *T. sphaerococcum*, *T. macha*, *T. vavilovii*). На рис. 4, б группа растений с короткими и округлыми колосьями располагается в диапазоне значений компоненты 2 от -80 до 0 (нижняя часть графика). Растения с длинными и тонкими колосьями имеют значения компоненты 2 в пределах 0 до 80 (верхняя часть графика). На рис. 4, а две этих группы растений грубо соответствуют двум облакам точек у гексаплоидных пшениц в верхней и нижней частях графика, которые слабо перекрываются в его центральной части. Таким образом, продемонстрированные на рис. 4 диаграммы позволяют наглядно показать разнообразие колосьев по своим характеристикам как внутри, так и между видами.

Чтобы охарактеризовать более детально сходство морфометрических параметров колосьев у разных видов пшеницы, мы провели иерархический кластерный анализ для них на основе сравнения средних значений изучаемых признаков (рис. 5).

На рис. 5 виды пшеницы разделены по характеристикам колосьев на два кластера (они обозначены красным и зеленым цветом). К первому кластеру (красный цвет) преимущественно относятся тетраплоидные виды (показаны синими прямоугольниками у конечных вершин дерева). Однако в него не включен дикий вид тетраплоидных пшениц *T. diccocooides*, а из гексаплоидных видов в него входят *T. compactum*, *T. antiquorum*, *T. sphaerococcum*, отличающиеся от всех остальных видов компактной формой колоса, т. е. из всех изученных видов гексаплоидных пшениц у них самый короткий колос. Отметим, что в работе (Zatybekov et al., 2020) при использовании хозяйственно

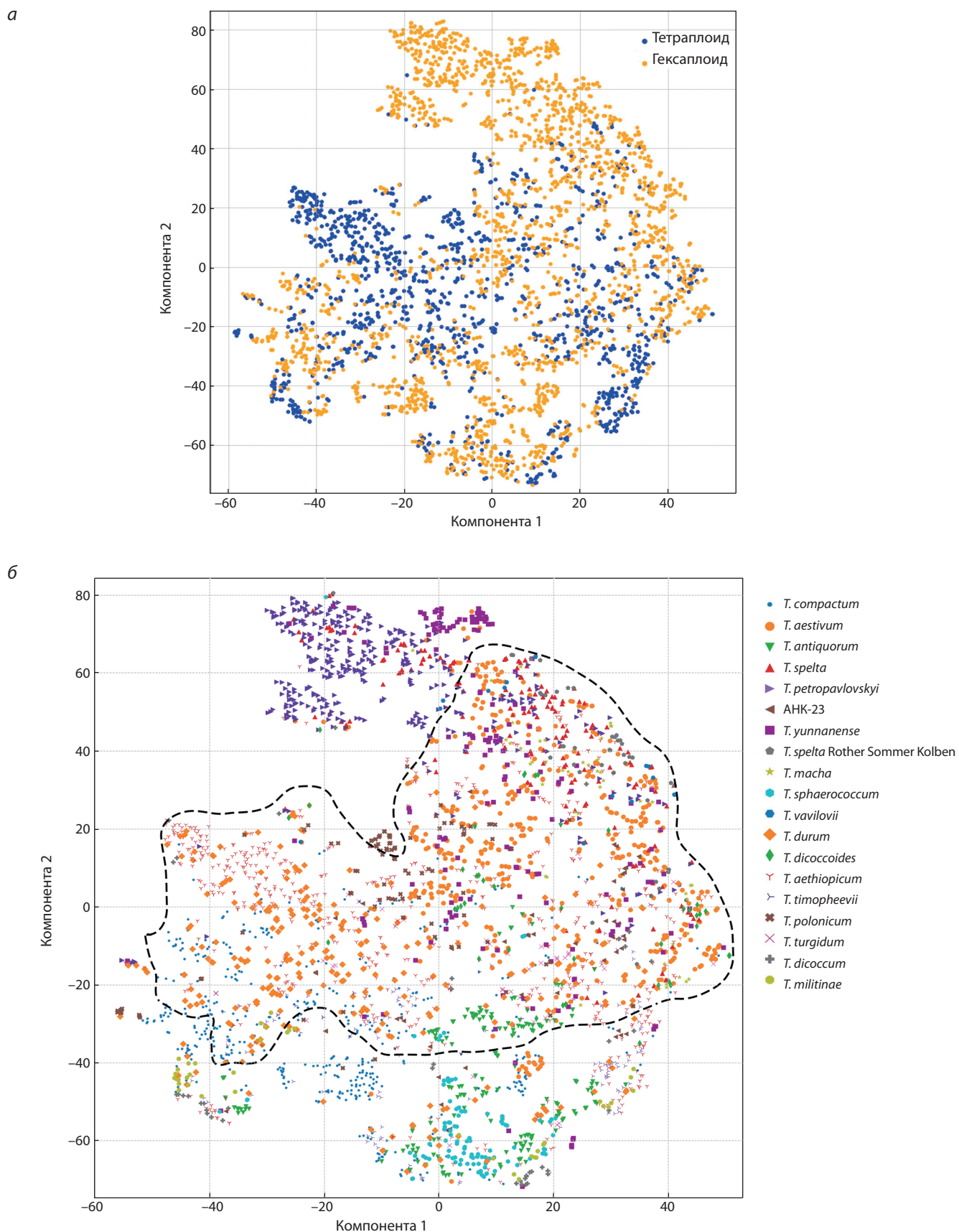


Рис. 4. Ординация цифровых изображений колосьев отдельных генотипов методом t-SNE, полученная на основании количественных признаков из табл. 2.

а – тетраплоидные виды пшеницы – синий цвет, гексаплоидные – оранжевый; *б* – цвет и форма каждой точки соответствуют определенному виду. Штриховой линией отмечена область, занимаемая видами *T. aestivum* и *T. durum*.

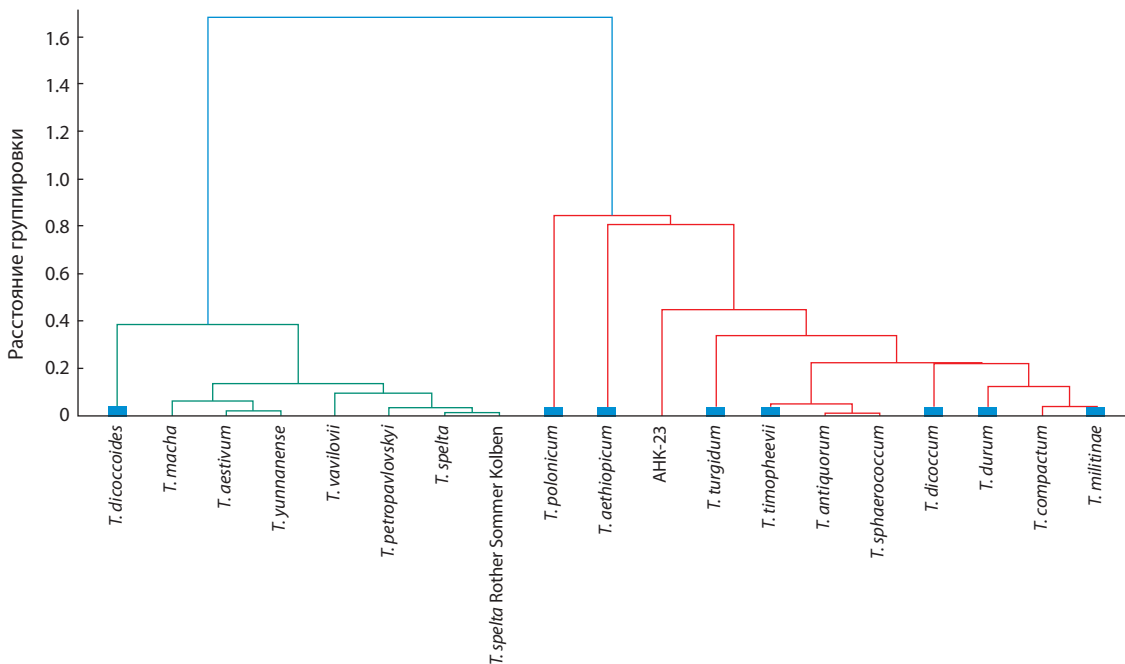


Рис. 5. Результаты иерархического кластерного анализа для девяти признаков колоса пшеницы. Синие квадраты соответствуют тетраплоидам.

важных признаков образцы шести тетраплоидных видов кластеризовались произвольно, т.е. вне зависимости от их видовой принадлежности. Интересно, что остальные гексаплоидные виды четко разделились на два кластера – среднеколосые (*T. macha*, *T. aestivum*, *T. yunnanense*) и длинноколосые (*T. vavilovii*, *T. petropavlovskiy*, *T. spelta*).

Включение в один кластер образцов *T. spelta* и *T. spelta* cv. Rother Sommer Kolben (стародавнего немецкого сорта) позволило сделать вывод, что «видовая» форма колоса в процессе многолетней селекции этого вида пшениц не менялась продолжительное время (в данном случае более 50 лет) и может быть успешно использована при классификации видов.

Следует отметить, что к гексаплоидам попал единственный в роде дикий тетраплоидный рыхлоколосый вид *T. dicoccoides*. В то время как в тетраплоидные виды попали гексаплоидные пшеницы с компактным типом колоса, *T. compactum*, *T. antiquorum*, *T. sphaerococcum*, и рукотворная изогенная линия АНК-23, созданная на сорте яровой мягкой пшеницы Новосибирская 67 (Коваль, 1997). Последнее допускает сделать заключение, что, несмотря на то, что изогенные линии создаются на определенном (конкретном) виде, тем не менее к их видовой принадлежности следует относиться с осторожностью.

Чуть подробнее остановимся на *T. petropavlovskiy*. Вид приручен к китайскому Припамирию – маршруту Великого шелкового пути. По результатам изучения запасных белков, все образцы этого вида были очень похожи на потомков гибридной комбинации, полученной от скрещивания мягкой пшеницы с *T. polonicum* (Watanabe et al., 2004). Авторы монографии «Культурная флора СССР» также считали возможным гибридогенное происхождение вида (Дорофеев и др., 1979). Кроме того, по ряду таксономических признаков *T. petropavlovskiy* также похожа

на мягкую пшеницу (Goncharov, 2005). К *T. petropavlovskiy* ssp. *mexicana* Bogusl. P.Л. Богуславским (1982) были отнесены межвидовые гибриды, полученные селекционерами СИММУТ. Исходя из вышеизложенного, мы посчитали целесообразным перевод *T. petropavlovskiy* в подвид *T. aestivum*:

***Triticum aestivum* ssp. *petropavlovskiy* comb. et stat. nov. (Udacz. et Migusch.) N.P. Gontsch.** – *T. turanicum* Jakubz. convar. *montanostepposum* Jakubz. f. *aristiforme* Jakubz. *Бот. журн.* 1959;10:1428, nom. illig. – *T. petropavlovskiy* Udacz. et Migusch. *Вестник с.-х. науки.* 1970;9:20. – **II. Петропавловского.**

Тип: описан по образцу из Китая «Происхождение: Китай, провинция Синцзян, сел. Курля, К-48376, 1957. Эксп. А.М. Горского. Репродукция Средняя Азия, Ташкент, САС ВИР. 8 VII 1969. Собрал/определил: Р.А. Удачин и Э.Ф. Мигушова» в С.-Петербурге (WIR!) (Гербарные экземпляры типа и паратипа приведены в Приложениях 3 и 4).

***Triticum aestivum* ssp. *petropavlovskiy* comb. et stat. nov. (Udacz. et Migusch.) N.P. Gontsch.** – *T. turanicum* Jakubz. convar. *montanostepposum* Jakubz. f. *aristiforme* Jakubz. 1959. *Bot. Zhur.* 10:1428, nom. illig. – *T. petropavlovskiy* Udacz. et Migusch. 1970. *Vestn. Sel'skokhoz. Nauki.* 9:20.

Типус: described by an accession from China “Origin: China, Xinjiang Province, village Kurlia, K-48376, 1957. A.M. Gorsky exp[edition]. Reproduction of Central Asia, Tashkent, Central Asian Station of VIR. 08. VII. 1969, Collected/defined: R.A. Udachin & E.F. Migushova” in St. Petersburg (WIR!).

Отметим, что результаты, представленные на рис. 3 и 4, а, свидетельствуют, что у гексаплоидных видов характеристики формы, размеров и остистости колосьев оказываются более разнообразными. Поэтому мы пред-

Таблица 3. Результаты применения *F*-статистики для подтверждения гипотезы о значимом различии дисперсии двух распределений

Название признака	<i>F</i> -статистика	<i>p</i> -value	Дисперсия		Среднее	
			гексаплоидов	тетраплоидов	гексаплоидов	тетраплоидов
Awns area	0.376	1.000	1.415	3.763	84.875	160.643
Circularity index	1.188	0.065	0.959	0.807	0.178	0.181
Roundness	1.828	1.110e-07	1.312	0.718	0.141	0.172
Perimeter	1.570	4.710e-05	1.080	0.688	218.124	185.015
Rachis length	3.500	< 1e-15	1.320	0.377	74.136	59.280
<i>X</i> _{u2}	3.928	< 1e-15	1.336	0.340	53.837	36.853
<i>X</i> _{b2}	4.437	< 1e-15	1.331	0.300	54.004	36.726
<i>Y</i> _{u2}	4.275	< 1e-15	2.491	0.583	3.844	4.171
<i>Y</i> _{bm}	1.081	0.248	0.695	0.643	0.225	0.246

Примечание. Значимые различия дисперсии выделены полужирным шрифтом.

положили, что характеристики колосьев у гексаплоидных видов могут иметь более высокий разброс значений, по сравнению с тетраплоидными видами. Для проверки этого предположения мы провели сравнение дисперсий по оцененным параметрам при помощи *F*-распределения (табл. 3).

Из результатов, приведенных в табл. 3, следует, что дисперсия большинства характеристик для гекса- и тетраплоидов имеет значимые различия ($p < 0.05$). В то же время значимые различия дисперсий не были нами обнаружены для таких признаков, как *y*_{bm} (параметр модели четырехугольников), Awns area и Circularity index. Интересно, что для всех значимых различий мы наблюдаем значение дисперсии у гексаплоидов выше, чем у тетраплоидов. Таким образом, проведенный анализ показал, что гексаплоидные виды демонстрируют более высокое разнообразие по морфометрическим характеристикам колоса, по сравнению с тетраплоидными.

В данных представлены растения 17 видов: 9 гексаплоидных: *T. compactum*, *T. aestivum* (в том числе изогенная линия сорта Новосибирская 67 АНК-23), *T. antiquorum*, *T. spelta* (включая стародавний сорт *T. spelta* Rother Sommer Kolben), *T. petropavlovskiyi*, *T. yunnanense*, *T. macha*, *T. sphaerococcum*, *T. vavilovii* и 8 тетраплоидных: *T. aethiopicum*, *T. dicoccoides*, *T. dicoccum*, *T. durum*, *T. militinae*, *T. polonicum*, *T. timopheevii*, *T. turgidum*. Результаты их кластеризации даны в таком варианте, чтобы цвет и форма каждой точки соответствовали определенному виду (см. рис. 5).

Хорошо известно, что удвоение генома в результате дупликаций (автополиплоидия) или гибридизации и последующей полиплоидизации (аллополиплоидия) приводит к заметным изменениям фенотипа растений (Finigan et al., 2012; Романов, Пимонов, 2018; Родионов и др., 2019). Эти изменения у растений происходят как на клеточном уровне (Liu et al., 2018), так и на уровне органов (Robinson et al., 2018). Во многих случаях у растений увеличение плоидности ведет к увеличению размеров клеток и органов (Comai, 2005; Williams Oliveira, 2020), повышению устойчивости к стрессам (Tan et al., 2015). В настоящее

время исследователи предполагают, что существует четыре типа молекулярных механизмов такой изменчивости: 1) увеличение дозы гена/аллеля, 2) увеличение генетического разнообразия, 3) изменение генетической регуляции и 4) эпигенетические перестройки генома (Chen, 2007; Finigan et al., 2012).

В нашей работе на примере анализа морфологических характеристик колосьев пшеницы гекса- ($2n = 6x = 42$) и тетраплоидных ($2n = 4x = 28$) видов мы показали, что для большинства признаков колоса их вариации у пшениц с большей плоидностью значимо выше. Полученные нами результаты находятся в согласии с представлениями о влиянии плоидности на изменчивость фенотипа растений.

Закключение

В работе проведен массовый анализ цифровых изображений колосьев для 595 растений восьми тетра- и девяти гексаплоидных видов пшениц. Рассмотрено девять количественных признаков, описывающих форму, размер и остистость колоса. Изучена изменчивость генотипов по указанным выше признакам и показано, что в пространстве характеристик колоса формируются два кластера. К первому относятся преимущественно гексаплоидные виды (за исключением дикого тетраплоидного вида *T. dicoccoides*). Ко второму – тетраплоидные (за исключением трех гексаплоидных с компактной формой колоса видов, *T. antiquorum*, *T. sphaerococcum* и изогенной линии АНК-23). Анализ дисперсий этих признаков у гекса- и тетраплоидных растений показал значимое увеличение дисперсии для шести из девяти признаков в выборке гексаплоидных растений, т. е. большая плоидность дает более варибельные значения количественных признаков морфологии колоса.

Таким образом, морфологические характеристики колосьев для гекса- и тетраплоидных видов, полученные на основе компьютерного анализа изображений, демонстрируют различия, которые в дальнейшем могут быть использованы для разработки методики классификации растений по плоидности и их видовой принадлежности в автоматическом режиме.

Список литературы / References

- Афонников Д.А., Генаев М.А., Дорошков А.В., Комышев Е.Г., Пшеничникова Т.А. Методы высокопроизводительного фенотипирования растений для массовых селекционно-генетических экспериментов. *Генетика*. 2016;52(7):788-803. DOI 10.7868/S001667581607002X.
[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.]
- Богуславский Р.Л. Новая ботаническая форма гексаплоидной пшеницы. *Науч.-техн. бюл. ВИР*. 1982;119:73-74.
[Boguslavsky R.L. A new botanical form of hexaploid wheat. *Nauchno-Tekhnicheskii Byulleten VIR = Scientific and Technological Bulletin of the Vavilov Institute of Plant Industry*. 1982;119:73-74. (in Russian)]
- Генаев М.А., Комышев Е.Г., Фу Хао, Коваль В.С., Гончаров Н.П., Афонников Д.А. SpikeDroidDB – информационная система для аннотации морфометрических характеристик колоса пшеницы. *Вавиловский журнал генетики и селекции*. 2018;22(1):132-140. DOI 10.18699/VJ18.340.
[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 Journal of Genetics and Breeding*. 2018;22(1):132-140. DOI 10.18699/VJ18.340. (in Russian)]
- Головнина К.А., Кондратенко Е.Я., Блинов А.Г., Гончаров Н.П. Филология А геномов диких и возделываемых видов пшениц. *Генетика*. 2009;45(11):1540-1547.
[Golovnina K.A., Kondratenko E.Ya., Blinov A.G., Goncharov N.P. Phylogeny of the A genome of wild and cultivated wheat species. *Russ. J. Genet.* 2009;45(11):1360-1367. DOI 10.1134/S1022795409110106.]
- Гончаров Н.П. Определитель разновидностей мягкой и твердой пшениц. Новосибирск: Изд-во СО РАН, 2009.
[Goncharov N.P. Manual Book of Common and Hard Wheat Varieties. Novosibirsk: SO RAN Publ., 2009. (in Russian)]
- Гончаров Н.П., Кондратенко Е.Я. Происхождение, доместикация и эволюция пшениц. *Информ. вестник ВОГИС*. 2008;12(1/2):159-179.
[Goncharov N.P., Kondratenko E.Ya. Wheat origin, domestication and evolution. *Informatsionnyy vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2008;12(1/2):159-179. (in Russian)]
- Гончаров Н.П., Шумный В.К. От сохранения генетических коллекций к созданию национальной системы хранения генофондов растений в вечной мерзлоте. *Информ. вестник ВОГИС*. 2008;12(4):509-523.
[Goncharov N.P., Shumny V.K. From preservation of genetic collections to organization of National project of plant gene pools conservation in permafrost. *Informatsionnyy vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2008;12(4):509-523. (in Russian)]
- Демидчик В.В., Шашко А.Ю., Бондаренко В.Ю., Смоликова Г.Н., Пржевальская Д.А., Черныш М.А., Пожванов Г.А., Барковский А.В., Смолич И.И., Соколик А.И., Ю М., Медведев С.С. Феномика растений: фундаментальные основы, программно-аппаратные платформы и методы машинного обучения. *Физиол. растений*. 2020;67(3):227-245. DOI 10.31857/S0015330320030069.
[Demidchik V.V., Shashko A.Y., Bandarenka V.Y., Smolikova G.N., Przhevalskaya D.A., Charnysh M.A., Pozhvanov G.A., Barkovskiy A.V., Smolich I.I., Sokolik A.I., Yu M., Medvedev S.S., Plant phenomics: fundamental bases, software and hardware platforms, and machine learning. *Russ. J. Plant Physiol.* 2020;67:397-412. DOI 10.1134/S1021443720030061.]
- Дорофеев В.Ф. Внутривидовая классификация пшеницы. *Докл. ВАСХНИЛ*. 1985;9:3-5.
[Dorofeev V.F. Intraspecific taxonomy of wheat. *Doklady VASKhNIL = Reports of the Academy of Agricultural Sciences*. 1985;9:1-4. (in Russian)]
- Дорофеев В.Ф., Филатенко А.А., Мигушова Э.Ф., Удачин Р.А., Якубцинер М.М. Культурная флора СССР. Т. 1. Пшеница. Л.: Колос, 1979.
[Dorofeev V.F., Filatenko A.A., Migushova E.F., Udachin R.A., Yakubtsiner M.M. Flora of Cultivated Plants of USSR. Vol. 1. Wheat. Leningrad: Kolos Publ., 1979. (in Russian)]
- Зуев Е.В., Амри А., Брыкова А.Н., Пюккенен В.П., Митрофанова О.П. Атлас разнообразия мягкой пшеницы (*Triticum aestivum* L.) по признакам колоса и зерновки. СПб.: Копи-Р, 2019.
[Zuev E.V., Amri A., Brykova A.N., Pyukkenen V.P., Mitrofanova O.P. Atlas of the Diversity of Soft Wheat (*Triticum aestivum* L.) by Ear and Grain Characteristics. St. Petersburg: Kopi-R Publ., 2019. (in Russian)]
- Коваль С.Ф. Каталог изогенных линий яровой мягкой пшеницы Новосибирская 67 и принципы их использования в эксперименте. *Генетика*. 1997;33(8):1168-1173.
[Koval S.F. The catalog of near-isogenic lines of Novosibirskaya-67 common wheat and principles of their use in experiments. *Russ. J. Genet.* 1997;33(8):995-1000.]
- Международный классификатор СЭВ рода *Triticum* L./ Сост. Дорофеев В.Ф., Руденко М.И., Филатенко А.А., Бареш И., Сегналова Я., Леманн Х. Л.: ВИР, 1984.
[Dorofeev V.F., Rudenko M.I., Filatenko A.A., Baras J., Segnalova J., Lemann H. (Compilers). The International Comecon List of Descriptors for the Genus *Triticum* L. Leningrad: VIR Publ., 1984. (in Russian)]
- Пальмова Е.Ф. Введение в экологию пшениц. М.; Л.: Сельхозгиз, 1935.
[Palmova E.F. Introduction to Wheat Ecology. Moscow; Leningrad: Selkhozgiz Publ., 1935. (in Russian)]
- Родионов А.В., Амосова А.В., Беляков Е.А., Журбенко П.М., Михайлова Ю.В., Пунина Е.О., Шнеер В.С., Лоскутов И.Г., Муравенко О.В. Генетические последствия межвидовой гибридизации, ее роль в видообразовании и фенотипическом разнообразии растений. *Генетика*. 2019;55(3):255-272. DOI 10.1134/S0016675819030159.
[Rodionov A.V., Amosova A.V., Belyakov E.A., Zhurbenko P.M., Mikhailova Y.V., Punina E.O., Shneyer V.S., Loskutov I.G., Muravenko O.V. Genetic consequences of interspecific hybridization, its role in speciation and phenotypic diversity of plants. *Russ. J. Genet.* 2019;55(3):278-294. DOI 10.1134/S1022795419030141.]
- Родионов А.В., Шнеер В.С., Гнутиков А.А., Носов Н.Н., Пунина Е.О., Журбенко П.М., Лоскутов И.Г., Муравенко О.В. Диалектика видов: от исходного единообразия, через максимально возможное разнообразие к конечному единообразию. *Бот. журн.* 2020;105(9):835-853. DOI 10.31857/S0006813620070091.
[Rodionov A.V., Shneyer V.S., Gnutikov A.A., Nosov N.N., Punina E.O., Zhurbenko P.M., Loskutov I.G., Muravenko O.V. Species dialectics: from initial uniformity, through the greatest possible diversity to ultimate uniformity. *Botanicheskii Zhurnal = Botanical Journal*. 2020;105(9):835-853. DOI 10.31857/S0006813620070091. (in Russian)]
- Романов Б.В., Пимонов К.И. Феномономика продукционных признаков видов пшеницы. пос. Персиановский: Донской ГАУ, 2018.
[Romanov B.V., Pimonov K.I. Phenomogenomics of Production Traits of Wheat Species. Persianovsky: Donskoy GAU Publ., 2018. (in Russian)]
- Синская Е.Н. Историческая география культурной флоры (На заре земледелия). Ленинград: Колос, 1969.
[Sinskaya E.N. Historical Geography of Cultural Flora (At the Dawn of Agriculture). Leningrad: Kolos Publ., 1969. (in Russian)]

- Удачин Р.А., Мигушова Э.Ф. Новое в познании рода *Triticum*. *Вестн. с.-х. науки*. 1970(9):20-24.
[Udachin R.A., Migushova E.F. New in the knowledge of the genus *Triticum*. *Vestnik Selskokhozyaystvennoy Nauki = Herald of Agricultural Sciences*. 1970;9:20-24. (in Russian)]
- Якубцинер М.М. К познанию пшениц Китая. *Бот. журн.* 1959; 44(10):1425-1436.
[Yakubtsiner M.M. More on Chinese wheats. *Botanicheskiy Zhurnal = Botanical Journal*. 1959;44(10):1425-1436. (in Russian)]
- Chen Z.J. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* 2007;58:377-406. DOI 10.1146/annurev.arplant.58.032806.103835.
- Comai L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 2005;6(11):836-846. DOI 10.1038/nrg1711.
- Finigan P., Tanurdzic M., Martienssen R.A. Origins of novel phenotypic variation in polyploids. In: *Polyploidy and Genome Evolution*. Berlin; Heidelberg: Springer Press, 2012;57-76. DOI 10.1007/978-3-642-31442-1_4.
- Genaeв M.A., Komyshev E.G., Smirnov N.V., Kruchinina Y.V., Goncharov N.P., Afonnikov D.A. Morphometry of the wheat spike by analyzing 2D images. *Agronomy*. 2019;9(7):390. DOI 10.3390/agronomy9070390.
- Goncharov N.P. Comparative-genetic analysis – a base for wheat taxonomy revision. *Czech J. Genet. Plant Breed.* 2005;41:52-55.
- Goncharov N.P. Genus *Triticum* L. taxonomy: the present and the future. *Plant Syst. Evol.* 2011;295(1-4):1-11. DOI 10.1007/s00606-011-0480-9.
- Hammer K., Filatenko A.A., Pistrick K. Taxonomic remarks on *Triticum* L. and \times *Triticosecale* Wittm. *Genet. Resour. Crop Evol.* 2011; 58(1):3-10. DOI 10.1007/s10722-010-9590-4.
- Johnson S.C. Hierarchical clustering schemes. *Psychometrika*. 1967; 32(3):241-254.
- Liu W., Zheng Y., Song S., Huo B., Li D., Wang J. *In vitro* induction of allohexaploid and resulting phenotypic variation in *Populus*. *Plant Cell Tiss. Organ Cult.* 2018;134(2):183-192. DOI 10.1007/s11240-018-1411-z.
- Müllner D. Modern hierarchical, agglomerative clustering algorithms. *arXiv*. 2011;1109.2378.
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 2011;12:2825-2830.
- Robinson D.O., Coate J.E., Singh A., Hong L., Bush M., Doyle J.J., Roeder A.H. Ploidy and size at multiple scales in the *Arabidopsis* sepal. *Plant Cell*. 2018;30(10):2308-2329. DOI 10.1105/tpc.18.00344.
- Snedecor G.W., Cochran W.G. *Statistical Methods*. Ames, Iowa: Iowa State University Press, 1989.
- Tan F., Tu H., Liang W., Long J.M., Wu X.M., Zhang H.Y., Guo W.W. Comparative metabolic and transcriptional analysis of a doubled diploid and its diploid citrus rootstock (*C. junos* cv. Ziyang xiangcheng) suggests its potential value for stress resistance improvement. *BMC Plant Biol.* 2015;15:89. DOI 10.1186/s12870-015-0450-4.
- van der Maaten L., Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 2008;9:2579-2605.
- van Slageren M., Payne T. Concepts and nomenclature of the Farro wheats, with special reference to Emmer, *Triticum turgidum* subsp. *dicoccum* (Poaceae). *Kew Bull.* 2013;68:477-494. DOI 10.1007/S12225-013-9459-8.
- Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S.J., Brett M., Jones E., Kern R., Larson E., Carey C.J., Polat I., Feng Yu, Moore E.W., VanderPlas J., Laxalde D., Perktold J., Cimrman R., Henriksen I., Quintero E.A., Harris C.R., Archibald A.M., Riberio A.H., Pedregosa F., van Mulbregt P. SciPy 1.0 Contributors. SciPy 1.0 – fundamental algorithms for scientific computing in Python. *Nat. Meth.* 2020;17(3):261-272. DOI 10.1038/s41592-019-0686-2.
- Watanabe N., Bannikova S.V., Goncharov N.P. Inheritance and chromosomal location of the genes for long glume phenotype found in Portuguese landraces of hexaploid wheat, 'Arrancada'. *J. Genet. Breed.* 2004;58:273-278.
- Williams J.H., Oliveira P.E. For things to stay the same, things must change: polyploidy and pollen tube growth rates. *Ann. Bot.* 2020; 125(6):925-935. DOI 10.1093/aob/mcaa007.
- Yang W., Feng H., Zhang X., Zhang J., Doonan J.H., Batchelor W.D., Xiong L., Yan J. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant*. 2020;13(2):187-214. DOI 10.1016/j.molp.2020.01.008.
- Zatybekov A., Anuarbek S., Abugaliev S., Turuspekov Y. Phenotypic and genetic variability of a tetraploid wheat collection grown in Kazakhstan. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2020;24(6):605-612. DOI 10.18699/VJ20.654.
- Zhang Y., Zhao C., Du J., Guo X., Wen W., Gu S., Wang J., Fan J. Crop phenomics: current status and perspectives. *Front. Plant Sci.* 2019; 10:714. DOI 10.3389/fpls.2019.00714.

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
Y.V. Kruchinina orcid.org/0000-0002-1084-9521

Благодарности. Подготовка образцов колосьев, фенотипирование, разработка алгоритмов анализа формы и классификации выполнены за счет финансирования Курчатковского геномного центра Федерального исследовательского центра ИЦиГ СО РАН, соглашение с Министерством образования и науки РФ № 075-15-2019-1662. Выращивание экспериментальных растений и определение *de visu* их видовой принадлежности по признакам, определяющим архитектуру колоса, поддержано грантом РНФ 16-16-10021. Обработка данных проведена с использованием ресурсов ЦКП «Биоинформатика» в рамках бюджетного проекта № 0259-2021-0009. Работа авторов А.А.П., Е.А.З., А.Ю.П., Н.М.П. осуществлена при поддержке Математического центра в Академгородке, соглашение с Министерством науки и высшего образования Российской Федерации № 075-15-2019-1675. Авторы благодарны Д.А. Афонникову (ИЦиГ СО РАН, г. Новосибирск) за замечания и рекомендации в процессе работы и И.Г. Чухиной (ВИР, г. Санкт-Петербург) за предоставление фотографий гербарных экземпляров пшеницы Петропавловского.

Прозрачность финансовой деятельности. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 27.10.2020. После доработки 31.12.2020. Принята к публикации 02.01.2021.

Английский текст <https://vavilov.elpub.ru/jour>

Анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19

О.И. Криворотко^{1,2}✉, С.И. Кабанихин^{1,2}, М.И. Сосновская², Д.В. Андорная²

¹ Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ krivorotko.olya@mail.ru

Аннотация. Разработан алгоритм анализа чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19 в Новосибирской области, основанных на системах дифференциальных уравнений и законе действующих масс. Основу алгоритма составляет анализ матрицы чувствительности методами дифференциальной и линейной алгебры, показывающей степень зависимости неизвестных параметров моделей от заданных измерений. В результате работы алгоритма выявляются наименее и наиболее чувствительные к измерениям параметры, что способствует построению регуляризирующего алгоритма решения задачи идентификации параметров для построения более точных сценариев развития эпидемии в регионе. Анализ чувствительности математических моделей распространения коронавирусной инфекции COVID-19 показал, что параметр contagiозности вируса устойчиво определяется по количеству ежедневно выявляемых заболевших, критических и вылечившихся больных. С другой стороны, прогнозируемая доля госпитализированных больных, находящихся в критическом состоянии и требующих подключения аппарата ИВЛ, а также коэффициент смертности определяются гораздо менее устойчиво. Для построения более реалистичного прогноза необходимо добавить дополнительную информацию о процессе (например, о количестве ежедневных случаев госпитализации). Задачи уточнения идентифицируемых параметров по дополнительной информации о количестве выявленных, критических и смертельных случаев в Новосибирской области были сведены к задачам минимизации соответствующих целевых функционалов. Задача минимизации была решена с помощью метода дифференциальной эволюции, широко используемого в задачах стохастической глобальной оптимизации. Показано, что более общая камерная модель, состоящая из семи обыкновенных дифференциальных уравнений, описывает основную тенденцию распространения коронавирусной инфекции, чувствительна к пикам выявленных случаев, однако некачественно описывает небольшие статистики (количество ежедневных критических, смертельных случаев), что может приводить к ошибочным выводам. Более подробная агентно-ориентированная математическая модель, учитывающая поведение отдельных агентов, позволяет улавливать небольшие шумы в данных и строить сценарии развития распространения эпидемии в регионе.

Ключевые слова: чувствительность параметров; идентифицируемость; обыкновенные дифференциальные уравнения; обратные задачи; эпидемиология; COVID-19; прогнозирование; Новосибирская область.

Для цитирования: Криворотко О.И., Кабанихин С.И., Сосновская М.И., Андорная Д.В. Анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19. *Вавиловский журнал генетики и селекции*. 2021;25(1):82-91. DOI 10.18699/VJ21.010

Sensitivity and identifiability analysis of COVID-19 pandemic models

O.I. Krivorotko^{1,2}✉, S.I. Kabanikhin^{1,2}, M.I. Sosnovskaya², D.V. Andornaya²

¹ Institute of Computational Mathematics and Mathematical Geophysics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

✉ krivorotko.olya@mail.ru

Abstract. The paper presents the results of sensitivity-based identifiability analysis of the COVID-19 pandemic spread models in the Novosibirsk region using the systems of differential equations and mass balance law. The algorithm is built on the sensitivity matrix analysis using the methods of differential and linear algebra. It allows one to determine the parameters that are the least and most sensitive to data changes to build a regularization for solving an identification problem of the most accurate pandemic spread scenarios in the region. The performed analysis has demonstrated that the virus contagiousness is identifiable from the number of daily confirmed, critical and recovery cases. On the other hand, the predicted proportion of the admitted patients who require a ventila-

tor and the mortality rate are determined much less consistently. It has been shown that building a more realistic forecast requires adding additional information about the process such as the number of daily hospital admissions. In our study, the problems of parameter identification using additional information about the number of daily confirmed, critical and mortality cases in the region were reduced to minimizing the corresponding misfit functions. The minimization problem was solved through the differential evolution method that is widely applied for stochastic global optimization. It has been demonstrated that a more general COVID-19 spread compartmental model consisting of seven ordinary differential equations describes the main trend of the spread and is sensitive to the peaks of confirmed cases but does not qualitatively describe small statistical datasets such as the number of daily critical cases or mortality that can lead to errors in forecasting. A more detailed agent-oriented model has been able to capture statistical data with additional noise to build scenarios of COVID-19 spread in the region.

Key words: parameter sensitivity; identifiability; ordinary differential equations; inverse problems; epidemiology; COVID-19; forecasting; Novosibirsk region.

For citation: Krivorotko O.I., Kabanikhin S.I., Sosnovskaya M.I., Andornaya D.V. Sensitivity and identifiability analysis of COVID-19 pandemic models. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):82-91. DOI 10.18699/VJ21.010

Введение

Многие математические модели в биологии (эпидемиология, иммунология, фармакокинетика, системная биология), медицине (томография), физике и химии (метеорология, химическая кинетика), социологии описываются системами дифференциальных уравнений: обыкновенных (Kermack, McKendrick, 1927), в частных производных (Habtemariam et al., 2008), стохастических (Lee et al., 2020). Коэффициенты этих уравнений характеризуют особенности протекания процессов в конкретных условиях. Для построения адекватных математических моделей необходимо уточнять коэффициенты уравнений по некоторой дополнительной информации о процессе или его известных характеристиках. Так, например, в задачах эпидемиологии скорость передачи инфекции в регионе, вероятности возникновения критических случаев в зависимости от сопутствующих заболеваний, возрастных и демографических характеристик, количество бессимптомных/латентных инфицированных и пр. неизвестны или заданы приближенно из статистической информации. Нередко эти характеристики чувствительны к измерениям, заданным с погрешностью (ошибки округления, погрешность прибора, человеческий фактор), что влечет за собой неустойчивость решения задачи идентификации параметров моделей.

Анализ идентифицируемости систем дифференциальных уравнений, возникающих при моделировании процессов в биологии, медицине, физике, является важным шагом перед разработкой вычислительных алгоритмов их решения (Bellu et al., 2007; Raue et al., 2010, 2014; Miao et al., 2011; Kabanikhin et al., 2016; Voropaeva, Tsgoev, 2019). В статье (Krivorotko et al., 2020a) приведена классификация методов идентифицируемости: структурная, практическая и анализ чувствительности, а также рассмотрены системы обыкновенных дифференциальных уравнений, описывающие процессы в эпидемиологии и иммунологии, на предмет чувствительности параметров к ошибкам в измерениях и практическую идентифицируемость.

Подробный обзор методов и примеров анализа структурной идентифицируемости в задачах биологии, описываемых системами обыкновенных дифференциальных уравнений (ОДУ), можно найти в работах (Miao et al.,

2011; Kabanikhin et al., 2016). Основываясь на структуре модели

$$\begin{cases} \frac{d\bar{x}}{dt} = g(t, \bar{x}, q), & t \in (0, T), x(t) \in C^1(\mathbb{R}^N), q \in \mathbb{R}^L, \\ \bar{x}(0) = \bar{x}_0, \\ x_i(t_k) = f_{ik}, & i \in \{1, \dots, M\}, k = 1, \dots, K, \end{cases} \quad (1)$$

можно установить единственность решения задачи определения параметров q и начальных условий (или их части) \bar{x}_0 модели по имеющимся измерениям f_{ik} , а также дать рекомендации по добавлению информации или изменению условий задачи поиска параметров для получения единственного решения.

Нами выполнен анализ полуотносительной чувствительности математических моделей эпидемиологии и социальных процессов, предложенный Adams et al. (2004) для анализа систем ОДУ, который позволяет установить степень чувствительности параметров к измерениям и недостающие/излишние измерения относительно некоторого «эталонного» набора для решения поставленной задачи идентификации неизвестных параметров. В качестве примера рассмотрены описываемые системой ОДУ две математические модели распространения новой коронавирусной инфекции, вызванной вирусом SARS-CoV-2. Построен алгоритм регуляризации численного решения задачи идентификации параметров математической модели SEIR-типа и камерной модели на основе состояний агентов в агентно-ориентированной модели по статистическим данным из открытых источников для Новосибирской области. Приведены результаты моделирования, а также сценарий развития распространения заболевания COVID-19 в Новосибирской области.

Анализ чувствительности параметров в системах обыкновенных дифференциальных уравнений

Анализ чувствительности используется для оценки идентифицируемости неизвестных параметров модели для системы ОДУ (1) до построения алгоритмов численного решения задач идентификации параметров. Для методов анализа чувствительности реальные экспериментальные данные не требуются, но количество измерений и момен-

ты времени, в которые выполнены измерения, могут быть необходимы. Исследование чувствительности для математической модели проводится относительно некоторых номинальных параметров q^* , значения которых берутся из литературных источников или доступной статистической информации.

Методы анализа чувствительности основаны на построении матрицы чувствительности. Предположим, что $t_1 \leq t_2 \leq \dots \leq t_K$ – фиксированные моменты времени, в которые проводятся измерения f_{ik} . Тогда коэффициенты матрицы чувствительности для вектора параметров q^* вычисляются по формуле

$$s_{ij}(t) = \frac{\partial f_i(t, q^*)}{\partial q_j} \cdot q_j^*, \quad (2)$$

где f_i ($i = 1, \dots, M$) – i -я компонента вектора измеряемых функций, а q_j ($j = 1, \dots, L$) – j -я компонента вектора параметров.

Таким образом, матрица чувствительности определяется следующим образом:

$$S_{M \cdot K \times L} = \begin{pmatrix} s_{11}(t_1) & \dots & s_{1L}(t_1) \\ \vdots & \ddots & \vdots \\ s_{M1}(t_1) & \dots & s_{ML}(t_1) \\ \vdots & \ddots & \vdots \\ s_{11}(t_K) & \dots & s_{1L}(t_K) \\ \vdots & \ddots & \vdots \\ s_{M1}(t_K) & \dots & s_{ML}(t_K) \end{pmatrix}.$$

Для расчета матрицы чувствительности рассматривается традиционная функция чувствительности:

$$s_{q_j}(t) = \frac{\partial x}{\partial q_j}(t), \quad j = 1, \dots, L.$$

Дифференцируя первое уравнение системы (1) по q_j , получаем, что каждая вектор-функция s_{q_j} удовлетворяет следующей задаче Коши:

$$\begin{cases} \dot{s}_{q_j}(t) = \frac{\partial g}{\partial \bar{x}}(t, \bar{x}(t; q), q) \cdot s_{q_j}(t) + \frac{\partial g}{\partial q_j}(t, \bar{x}(t; q), q), \\ s_{q_j}(t_0) = \frac{\partial \bar{x}_0}{\partial q_j}. \end{cases} \quad (3)$$

Численно решая систему дифференциальных уравнений (3), получаем $s_{q_j}(t)$.

В качестве первого шага оцениваются те параметры q , к которым решение модели наиболее чувствительно. Они, в свою очередь, определяются с помощью расчета полуотносительной чувствительности. Представим информацию о чувствительности как функцию времени на интересующем интервале. Мы хотим иметь некоторую общую меру чувствительности решения к параметрам, поэтому для каждой комбинации состояния/параметр берем норму (в пространстве L_2) по времени, а затем ранжируем полученные скаляры, чтобы определить наиболее чувствительные параметры. Чем меньше значение

$\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$, тем меньше влияние параметра q_k на переменную f_i . Данную общую меру назовем полуотносительной чувствительностью.

Далее для анализа чувствительности в работе использован ортогональный метод (Yao et al., 2003). Его основная идея состоит в том, чтобы исследовать линейные зависимости столбцов матрицы чувствительности S . Таким образом, одновременно можно оценить как чувствительность параметров к входным данным, так и зависимость между параметрами.

Анализ чувствительности математических моделей распространения COVID-19

Особенность разработанных к настоящему времени математических моделей распространения COVID-19 состоит в анализе поведения бессимптомного протекания болезни и влияния инкубационного периода заболевания на характер эпидемиологической ситуации в регионах. Разработан ряд пакетов с открытым кодом (Gomez et al., 2020; Tuomisto et al., 2020; Wolfram, 2020), а также web-сервисов для моделирования сценариев развития COVID-19:

- в странах мира: <https://covid19-scenarios.org/> (Базельский университет, Швейцария);
- в Москве, Новосибирской области и ряде европейских стран: <https://covid19.biouml.org/> (Федеральный исследовательский центр Институт вычислительных технологий СО РАН, Новосибирск);
- в Алматы, Казахстан: <http://covid19.mmay.info/almaty/?fbclid=IwAR20yx7F4MdWRqwUDzripUK29IWAvoyCSkDPafgpj25ummay23e7oFHBdXg>.

Основные подходы в моделировании распространения эпидемий можно разделить на две группы.

1. *Камерный подход* (моделирование «сверху вниз»). Взаимодействие агентов в популяции, распределенных в характерные группы со схожими признаками (восприимчивые, (бес)симптомные инфицированные, госпитализированные агенты, критические случаи и т.п.), строится на основе закона действующих масс в рамках камерного моделирования, впервые предложенного в 1927 г. (Kermack, McKendrick, 1927). Распределение агентов во времени происходит в зависимости от заданных вероятностей перехода между группами: вероятность инфицирования, параметр контагиозности вируса, смертность и др.
2. *Агентно-ориентированный подход* (моделирование «снизу вверх»), базирующийся на исследовании взаимодействий отдельных индивидуумов и их влияния на глобальные показатели (параметр контагиозности, смертность, вероятность тяжелого протекания болезни и др.). Агентно-ориентированные модели характеризуются случайными графами, в которых длинам ребер соответствуют вероятности перемещения агента в различные состояния.

Часто параметры перехода между группами и состояниями агентов неизвестны или заданы в широком диапазоне (например, инкубационный период заболевания составляет 2–14 дней по оценкам ВОЗ), что затрудняет анализ модели и построение качественных сценариев развития заболевания.

Рассмотрим два варианта разделения популяции конкретного возраста (20–29 лет) на группы, в которых агенты переходят в различные состояния в ходе прогрессирования

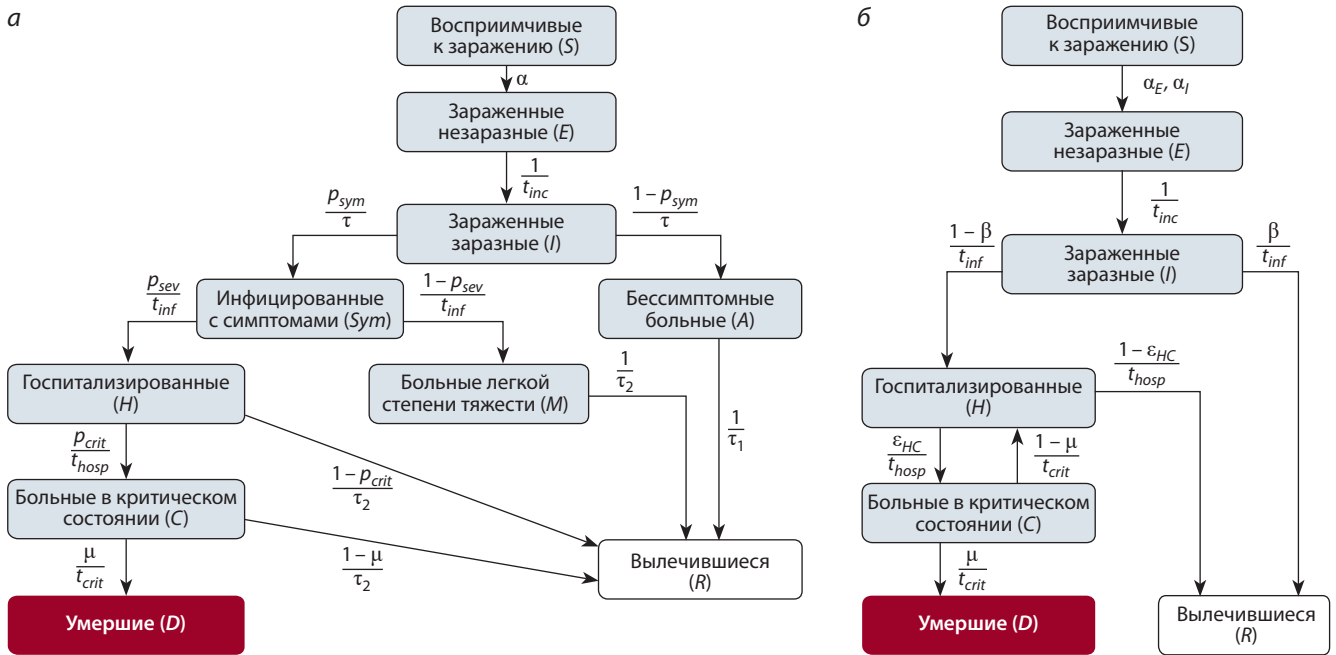


Рис. 1. Диаграмма состояний агентов: а – в пакете COVASIM (Kerr et al., 2020); б – в рамках SEIR-HCD модели (Unlu et al., 2020).

заболевания, вызванного вирусом SARS-CoV-2 (рис. 1). В моделях не учитываются: разделение по половому признаку; ежегодные рождаемость и смертность (поскольку промежуток моделирования составляет менее одного года); вакцинация; пассажиропотоки; сопутствующие заболевания агентов, влияющие на вероятности перехода между группами. Цель анализа двух математических моделей – продемонстрировать корреляцию зависимостей схожих параметров к одним и тем же измерениям, а также дать рекомендации, какие параметры по каким измерениям удастся определить устойчиво.

Система ОДУ типа (1), описывающая распространение COVID-19 в популяции, разделенной на десять характерных групп (Kerr et al., 2020), записывается на основе баланса масс следующим образом:

$$\begin{cases}
 \frac{dS}{dt} = -\alpha S(t), \\
 \frac{dE}{dt} = \alpha S(t) - \frac{1}{t_{inc}} E(t), \\
 \frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{\tau} I(t), \\
 \frac{dA}{dt} = \frac{1-p_{sym}}{\tau} I(t) - \frac{1}{\tau_1} A(t), \\
 \frac{dSym}{dt} = \frac{p_{sym}}{\tau} I(t) - \frac{p_{sev}}{t_{inf}} Sym(t) - \frac{1-p_{sev}}{\tau} Sym(t), \\
 \frac{dR}{dt} = \frac{1}{\tau_1} A(t) + \frac{1-p_{crit}}{\tau_2} H(t) + \frac{1}{\tau_1} M(t) + \frac{1-\mu}{\tau_2} C(t), \\
 \frac{dH}{dt} = \frac{p_{sev}}{t_{inf}} Sym(t) - \frac{p_{crit}}{t_{hosp}} H(t) - \frac{1-p_{crit}}{\tau_2} H(t), \\
 \frac{dM}{dt} = \frac{1-p_{sev}}{\tau} Sym(t) - \frac{1}{\tau_2} M(t), \\
 \frac{dC}{dt} = \frac{p_{crit}}{t_{hosp}} H(t) - \frac{\mu}{t_{crit}} C(t) - \frac{1-\mu}{\tau_2} C(t), \\
 \frac{dD}{dt} = \frac{\mu}{t_{crit}} C(t),
 \end{cases} \quad (4)$$

с начальными условиями:

$$S(0) = S_0, E(0) = E_0, I(0) = I_0, A(0) = A_0, Sym(0) = Sym_0, \\
 R(0) = R_0, H(0) = H_0, M(0) = M_0, C(0) = C_0, D(0) = D_0.$$

Модель (4) характеризует один из классов состояний агентов одной возрастной группы в агентно-ориентированной модели (см. рис. 1, а).

Аналогично записывается система уравнений SEIR-HCD модели, в которой популяция разделена на семь групп (Krivorotko et al., 2020b; Unlu et al., 2020):

$$\begin{cases}
 \frac{dS}{dt} = -\frac{5-a(t-\tau)}{5} \left[\frac{\alpha_I S(t)I(t)}{N(t)} + \frac{\alpha_E S(t)E(t)}{N(t)} \right], \\
 \frac{dE}{dt} = \frac{5-a(t-\tau)}{5} \left[\frac{\alpha_I S(t)I(t)}{N(t)} + \frac{\alpha_E S(t)E(t)}{N(t)} \right] - \frac{1}{t_{inc}} E(t), \\
 \frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{t_{inf}} I(t), \\
 \frac{dR}{dt} = \frac{\beta}{t_{inf}} I(t) + \frac{1-\epsilon_{HC}}{t_{hosp}} H(t), \\
 \frac{dH}{dt} = \frac{1-\beta}{t_{inf}} I(t) + \frac{1-\mu}{t_{crit}} C(t) - \frac{1}{t_{hosp}} H(t), \\
 \frac{dC}{dt} = \frac{\epsilon_{HC}}{t_{hosp}} H(t) - \frac{1}{t_{crit}} C(t), \\
 \frac{dD}{dt} = \frac{\mu}{t_{crit}} C(t),
 \end{cases} \quad (5)$$

с начальными условиями:

$$S(0) = S_0, E(0) = E_0, I(0) = I_0, R(0) = R_0, \\
 H(0) = H_0, C(0) = C_0, D(0) = D_0.$$

Здесь $S(t)$ – восприимчивый агент в момент времени t ; $E(t)$ – зараженный незаразный (не передающий вирус); $I(t)$ – зараженный заразный (передающий вирус); $A(t)$ – больной без симптомов; $Sym(t)$ – больной с симптомами;

Таблица 1. Параметры моделей (4), (5) и их усредненные значения для Новосибирской области (Kerr et al., 2020; Unlu et al., 2020)

Параметр	Описание	Значение/ интервал
τ_1	Количество дней до выздоровления бессимптомных и легких больных	6–11
τ_2	Количество дней до выздоровления зараженных пациентов в критическом и тяжелом состоянии	12–17
p_{sym}	Вероятность проявления симптомов после заражения	0.6
p_{sev}	Вероятность перехода больного с симптомами в тяжелое состояние (нуждается в госпитализации)	0.0072
p_{crit}	Вероятность перехода больного из тяжелого состояния в критическое (нуждается в ИВЛ)	0.00036
α	Вероятность передачи вируса для агентов, имевших контакт	(0.01, 0.025)
$E(0)$	Начальное количество инфицированных людей в популяции	(1, 100)
$a(t)$	Индекс самоизоляции от Яндекс, который описывает степень изоляции населения по шкале от 0 (отсутствие изоляции) до 5 (полная изоляция)	(0, 5)
α_E	Вероятность заражения между бессимптомной и восприимчивой группами населения ($\alpha_E \gg \alpha_i$)	(0, 1)
α_i	Вероятность заражения между инфицированным и восприимчивым населением, которая связана с контагиозностью вируса и социальными факторами	(0, 1)
β	Вероятность выздоровления зараженных пациентов, которые перенесли болезнь без осложнений	(0, 1)
ε_{HC}	Доля госпитализированных больных, находящихся в критическом состоянии и требующих подключения аппарата ИВЛ	(0, 1)
μ	Вероятность смерти в результате COVID-19	(0, 0.5)
τ	Продолжительность латентного периода (характеризует запаздывание выделения вирионов), дней	2
t_{inc}	Продолжительность инкубационного периода, дней	2–14
t_{inf}	Продолжительность периода заражения, дней	2.5–14
t_{hosp}	Продолжительность периода госпитализации, дней	4–5
t_{crit}	Продолжительность использования аппарата ИВЛ, дней	10–20

$H(t)$ – тяжелобольной; $C(t)$ – критически больной (требующий подключения аппарата ИВЛ); $M(t)$ – больной легкой степени тяжести; $R(t)$ – выздоровевший; $D(t)$ – умерший. Усредненные параметры моделей (4) и (5) для Новосибирской области приведены в табл. 1 (Lauer et al., 2020; Verity et al., 2020; Wölfel et al., 2020).

Отметим, что коэффициенты t_{inc}^{-1} , t_{inf}^{-1} , t_{hosp}^{-1} , t_{crit}^{-1} , τ^{-1} , τ_1^{-1} , τ_2^{-1} перед соответствующими состояниями агентов в моделях (4) и (5) описывают запаздывание перехода между состояниями (Лихошвай и др., 2004). Например, в уравнении (модель (5))

$$\frac{dI}{dt} = \frac{1}{t_{inc}} E(t) - \frac{1}{t_{inf}} I(t)$$

коэффициент t_{inc}^{-1} означает (в линейном приближении) запаздывание в t_{inc} дней перехода из группы зараженных незаразных индивидуумов $E(t)$ в группу инфицированных заразных $I(t)$, а коэффициент $-t_{inf}^{-1}$ – задержку агента в группе инфицированных заразных индивидуумов в течение периода инфицирования t_{inf} , дней.

Математическая модель 1 (схема приведена на рис. 1, а). Предположим, что известна дополнительная информация о количестве вылеченных и умерших больных в фиксированные дни в случае математической модели (4):

$$R(t_k) = R_k, D(t_k) = D_k, k = 1, \dots, 225. \quad (6)$$

Здесь R_k – количество выздоровевших агентов в день k ; D_k – количество умерших в результате заболевания в день k .

Проанализируем полутносительную чувствительность двух неизвестных параметров заразности α и начальное

Таблица 2. Полуотносительные чувствительности различных состояний модели (4) к параметрам, отсортированные по убыванию

Переменная f_i	Параметр q_k	$\left\ \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\ _2$
$R(t)$	$E(0)$	$8.9 \cdot 10^6$
	α	$7.6 \cdot 10^{14}$
$D(t)$	$E(0)$	$6.7 \cdot 10^{-14}$
	α	$6.07 \cdot 10^{-6}$

го количества бессимптомных больных $E(0)$ в модели (4) к измерениям (6). Это позволит установить возможность устойчивого определения неизвестных параметров по имеющимся данным для построения адекватной эпидемиологической картины в регионе. В табл. 2 приведены отсортированные по убыванию значения функции чувствительности исследуемых параметров $q_1 = \alpha$, $q_2 = E(0)$ на измерения $(f_i) = (R, D)$, $i = 1, 2$, представленной в виде нормы $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$. Чем меньше значение $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$, тем меньше влияние параметра q_k на измерения f_i .

На рис. 2 представлены графики изменения функции чувствительности $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ от времени в зависимости от варьируемого параметра. Таким образом, в модели (4)

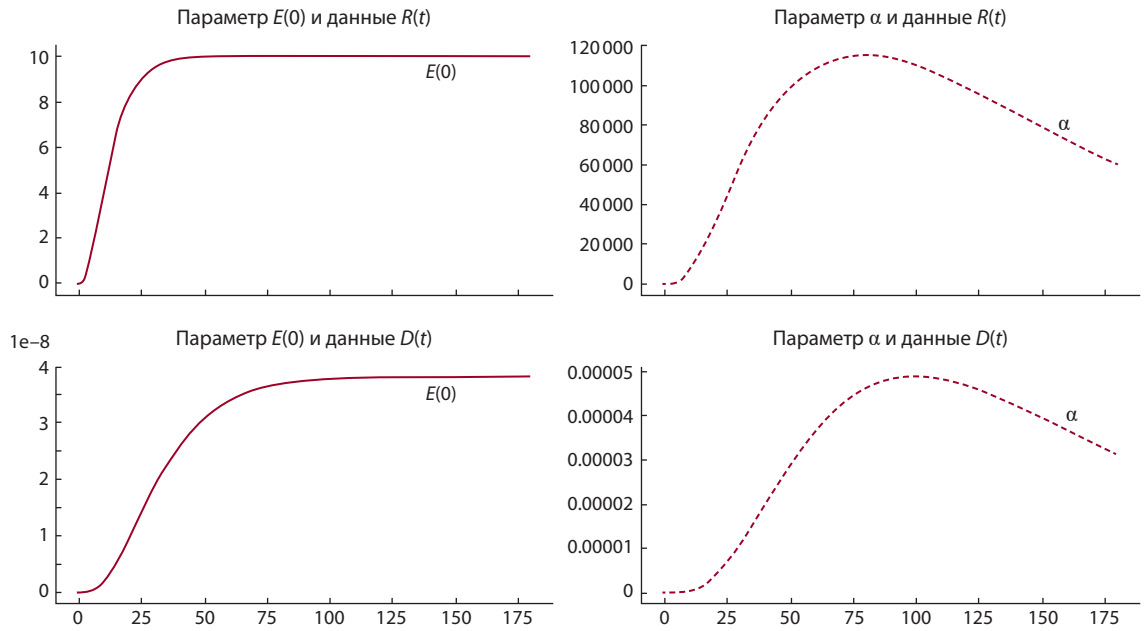


Рис. 2. Функция чувствительности $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ для модели (4) для периода моделирования 182 дня (с 12.03.2020 по 09.09.2020).

параметры α и $E(0)$ наименее чувствительны к переменной $D(t)$ и не могут быть определены только по данным о смертельных случаях. В свою очередь эти параметры чувствительны к функции $R(t)$, а следовательно, по измерениям о случаях выздоровления восстанавливаются более устойчиво.

Математическая модель 2 (схема приведена на рис. 1, б). Теперь исследуем математическую модель SEIR-HCD (5). Предположим, что известны дополнительные измерения о количестве выявленных, критических и смертельных случаев в фиксированные моменты времени:

$$I(t_k) = (1 - b_k) f_k, \quad C(t_k) = C_k, \quad D(t_k) = D_k, \quad (7)$$

$$t_k \in (t_0, T), \quad k = 1, \dots, 205,$$

где $b(t) \in [0, 1]$ – доля бессимптомных больных в выявленных случаях; f_k – количество выявленных больных в день k ; C_k – количество пациентов в критическом состоянии в день k .

Пусть параметры $q = (\alpha_E, \alpha_I, \beta, \varepsilon_{HC}, \mu, E_0)^T \in \mathbb{R}^6$ неизвестны. Проанализируем полурелятивную чувствительность вектора параметров q к измерениям (7) для математической модели (5). Для этого построим $\frac{\partial f_i(t)}{\partial q_k} q_k^*$,

$(f_i) = (I, C, D)$, $i = 1, 2, 3$, и проанализируем значение $\left\| \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\|_2$ (табл. 3). Качество определения параметров

β, ε_{HC} и μ при решении обратной задачи практически не зависит от имеющихся измерений количества инфицированных $I(t)$, в отличие, например, от более чувствительных к этим данным коэффициентов α_E, α_I, E_0 .

На рис. 3 представлены графики изменения от времени функции чувствительности $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ в зависимости от варьируемого параметра. Чем более изменчив параметр

в динамике, тем чувствительность к данным измерениям выше, а значит, определяться он будет более устойчиво.

Результаты анализа чувствительности параметров математической модели (5) при различных итерациях ортогонального алгоритма (описание алгоритма см. (Krivorotko

Таблица 3. Полуотносительные чувствительности различных состояний модели (5) к параметрам, отсортированные по убыванию

Переменная f_i	Параметр q_k	$\left\ \frac{\partial f_i(t)}{\partial q_k} q_k^* \right\ _2$
I	α_E	$2.865 \cdot 10^{14}$
I	α_I	$2.396 \cdot 10^{14}$
I	E_0	$1.854 \cdot 10^{14}$
C	α_E	$2.386 \cdot 10^{12}$
C	α_I	$1.996 \cdot 10^{12}$
C	E_0	$1.544 \cdot 10^{12}$
D	α_E	$7.110 \cdot 10^{11}$
D	α_I	$5.947 \cdot 10^{11}$
D	E_0	$4.601 \cdot 10^{11}$
C	ε_{HC}	$4.833 \cdot 10^4$
C	β	$3.428 \cdot 10^4$
D	ε_{HC}	$3.041 \cdot 10^4$
D	μ	$2.982 \cdot 10^4$
D	β	$2.164 \cdot 10^4$
C	μ	$3.695 \cdot 10^2$
I	β	$2.03 \cdot 10^{-6}$
I	ε_{HC}	$1.6 \cdot 10^{-7}$
I	μ	$3 \cdot 10^{-8}$

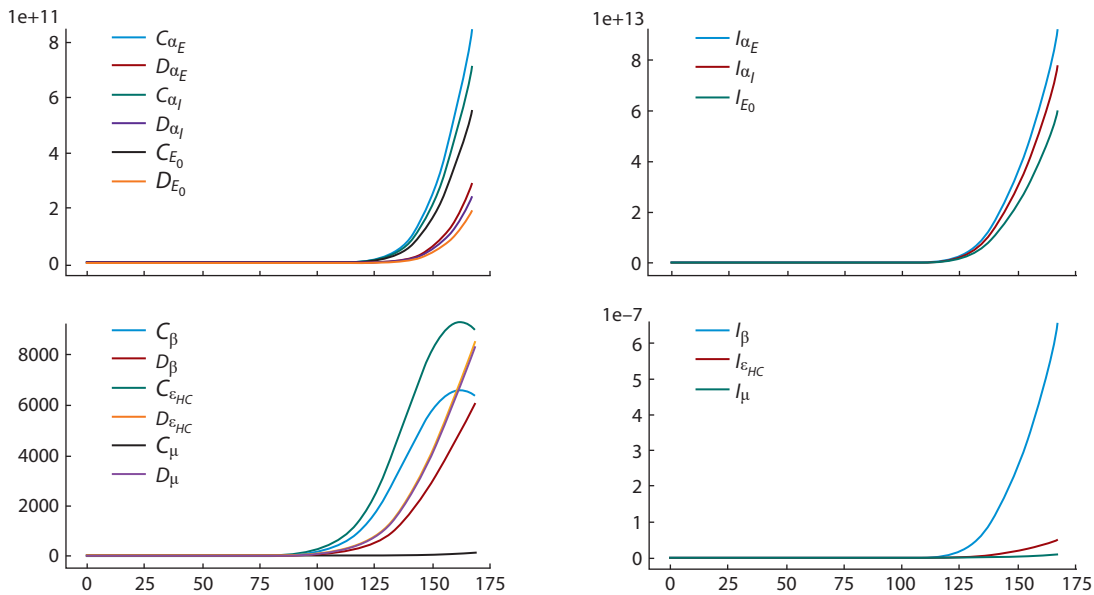


Рис. 3. Функция полуотносительной чувствительности $\frac{\partial f_i(t)}{\partial q_k} q_k^*$ для временного интервала 170 дней (с 15.04.2020 по 01.10.2020).

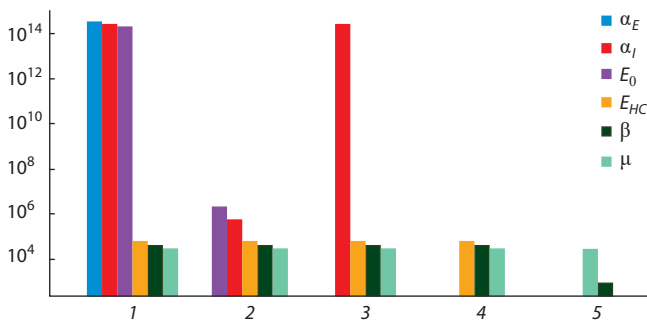


Рис. 4. Величины норм перпендикуляров для каждого параметра на различных итерациях (1–5) ортогонального алгоритма чувствительности параметров математической модели (5).

et al., 2020a) приведены на рис. 4. По горизонтальной оси отложены итерации ортогонального алгоритма, количество которых на единицу меньше размерности вектора неизвестных параметров (количества столбцов матрицы чувствительности), а по вертикальной – значения норм перпендикуляров для полученных преобразований матриц чувствительности. Показано, что наиболее идентифицируемыми также оказались параметры заражения между бессимптомной и восприимчивой группами населения α_E , между инфицированным и восприимчивым населением α_I (связан с контагиозностью вируса и социальными факторами), а также начальное значение зараженных или находящихся в инкубационном периоде индивидуумов E_0 . С помощью метода анализа чувствительности для математической модели (5) получена последовательность параметров (от наиболее до наименее чувствительного): $\alpha_E, E_0, \alpha_I, \epsilon_{HC}, \mu, \beta$.

После анализа идентифицируемости можно заключить, что наименее чувствительными (более идентифицируе-

мыми) параметрами модели к вариациям в данных (погрешностям) являются α_E, E_0 и α_I , иными словами, эти параметры более устойчиво определяются при решении обратной задачи (5), (7). Наиболее чувствительными (менее идентифицируемыми) к ошибкам в измерениях являются параметры ϵ_{HC}, μ и β (с наименьшими значениями норм перпендикуляров в матрице чувствительности), т. е. необходимо разработать алгоритм регуляризации, позволяющий контролировать качество определения чувствительных параметров.

Математическое моделирование распространения COVID-19 в Новосибирской области

Для математического моделирования COVID-19 в Новосибирской области были использованы данные открытых источников:

- количество тестированных (в том числе выявленных f и процента $b(t)$ бессимптомных из них), вылеченных (R) и умерших (D) от COVID-19;
- промежутки длительности инкубационного t_{inc} , латентного τ периодов болезни, периодов инфицирования t_{inf} , госпитализации t_{hosp} , длительности использования аппарата ИВЛ t_{crit} ;
- скорости выздоровления легких τ_1 и тяжелых τ_2 случаев;
- демографические показатели (возрастные распределения в регионе, размер популяции);
- сведения о среднем размере семьи (2.6 человека) в Российской Федерации по данным ООН в 2019 г. (<https://population.un.org/Household/#/countries/840>).
Исследуемая информация периодически обновлялась на сайтах:
 - Минздрава Новосибирской области: <https://zdrav.nso.ru/> (данные пункта (г));

Таблица 4. Меры по сдерживанию коронавирусной инфекции, принятые в Новосибирской области в 2020 г., которые учитываются в моделях (4) и (5)

Дата	Мера
18 марта	Начало удаленных занятий в школах и университетах Новосибирской области
28 марта	Приостановлены все массовые, развлекательные, общественные мероприятия на территории региона
27 апреля	Указ губернатора об обязательном ношении масок в магазинах
6 июля	Открытие летних веранд в кафе и ресторанах
1 сентября	Начало очных занятий в школах и университетах
28 сентября	Введение обязательного ношения масок в любых помещениях, ужесточение мер в образовательных учреждениях

- Федеральной службы государственной статистики, Новосибирская область: <https://novosibstat.gks.ru/folder/31729> (данные пункта (в));
- Стопкоронавирус: <https://стопкоронавирус.рф> (данные пункта (а));
- Всемирной организации здравоохранения: <https://www.who.int> (данные пункта (б)).

Моделирование проводилось с учетом введенных в Новосибирской области сдерживающих мер в отношении коронавирусной инфекции COVID-19 (табл. 4).

Решения обратных задач (4), (6) и (5), (7) сводятся к решению задачи минимизации целевого функционала (Kabanikhin, 2008):

$$J(q) = \sum_s \sum_{i=1}^T w_s \cdot G(c_d^{i,s}, c_m^{i,s}(q)).$$

Здесь s – статистики, по которым сравнивались данные (кумулятивные выявленные, критические, умершие); w_s – весовой коэффициент; $c_d^{i,s}, c_m^{i,s}$ значения данных (с индексом d) и модели (с индексом m); T – количество дней моделирования; q – вектор неизвестных параметров: $q_1 = (\beta, E_0)^T$ в случае обратной задачи (4), (6) и $q_2 = (\alpha_E(t), \alpha_I(t), \beta, \varepsilon_{HC}, \mu, E_0)^T$ в случае обратной задачи (5), (7). В вычислительных экспериментах использовалась абсолютная норма

$$G_1 = \frac{|c_d^{i,s} - c_m^{i,s}|}{M}, \text{ где } M = \max_i \{c_d^{i,s}\} - \text{нормирующий член,}$$

и квадратичное отклонение в виде $G_2 = (c_d^{i,s} - c_m^{i,s})^2 / T$.

Поиск минимума функционала $J(q)$ реализуется с помощью метода дифференциальной эволюции библиотеки SciPy.Optimize Python. Общая схема алгоритма нахождения глобального минимума записывается следующим образом:

1. Генерация начального поколения $\{\vec{q}_i\} \in B, i = 1 \dots N$.
2. Генерация нового поколения.

- Мутация:

$\forall \vec{q}_i \in B$ выбираются три случайных вектора

$$\vec{v}_1, \vec{v}_2, \vec{v}_3 \in B (\vec{v}_j \neq \vec{q}_i, j = 1, 2, 3).$$

Мутантный вектор: $\vec{v} = \vec{v}_1 + F(\vec{v}_2 - \vec{v}_3), F \in [0, 2]$.

- Скрещивание: пробный вектор \vec{u} вычисляется следующим образом:

$$u_k = \begin{cases} v_k, & \text{если } rand < p, \\ q_k, & \text{если } rand \geq p, \end{cases} \quad k = 1 \dots N_q.$$

3. Отбор:

$$\vec{q}_i = \begin{cases} \vec{q}_i, & \text{если } J(\vec{x}_i) < J(\vec{u}_i), \\ \vec{u}_i, & \text{иначе.} \end{cases}$$

Результат моделирования распространения коронавирусной инфекции в Новосибирской области с прогнозом до 10 декабря 2020 г. представлен на рис. 5. При конструировании данной модели мы использовали агентно-ориентированный подход, основанный на исследовании взаимодействий отдельных индивидуумов и их влияния на глобальные показатели. Моделирование производили с помощью пакета Covasim – инструмента для создания стохастических агентных моделей. Подробнее со структурой модели можно ознакомиться в работе (Kerr et al., 2020). Использовались также статистические данные по выявленным случаям и смертям в период с 12 марта по 23 октября 2020 г. Минимизировался следующий функционал с учетом проведенного анализа идентифицируемости моделей (4), (6):

$$J(q_1) = \frac{1}{T} \sum_{i=1}^T (f_d^i - f_m^i)^2 + 100 \cdot (D_d^i - D_m^i)^2.$$

Здесь f_d^i, f_m^i – кумулятивные выявленные случаи; D_d^i, D_m^i – кумулятивные смерти.

На рис. 5, а, б представлены результаты моделирования f_m^i и статистики f_d^i по выявленным случаям – кумулятивным и ежедневным соответственно, на рис. 5, в – результаты моделирования D_m^i и статистики D_d^i по кумулятивной смертности в результате COVID-19 в Новосибирской области. Отметим, что в статистических данных и в результатах моделирования наблюдается вторая волна эпидемии в середине сентября, которая после введения с 28 октября более жестких мер растет незначительно (не более 215 ежедневных выявленных случаев к середине декабря 2020 г.).

Обратная задача (5), (7) была сведена к минимизации целевого функционала (Krivorotko et al., 2020b):

$$J(q_2) = \sum_{k=1}^K (w_1 |t_{inc}^{-1} E(t_{k-1}; q_2) - (1 - b_k) f_k| + w_2 |C(t_k; q_2) - C_k| + w_3 |D(t_k; q_2) - D_k|).$$

Параметры скорости распространения инфекции $\alpha_E(t), \alpha_I(t)$, связанные с контагиозностью вируса и зависящие от времени, представлялись в виде кусочно-постоянных функций в зависимости от введенных карантинных мер (см. табл. 4).

Учитывая анализ идентифицируемости математической модели (5), (7), были введены более жесткие ограничения на поиск слабоидентифицируемых параметров (см. табл. 1). Результат моделирования решения обратной задачи (5), (7) для SEIR-HCD модели в Новосибирской

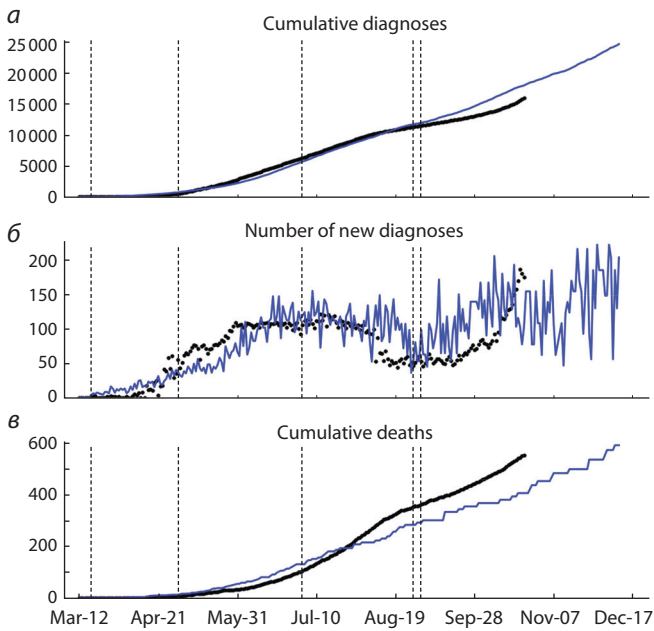


Рис. 5. Моделирование распространения COVID-19 в Новосибирской области (синяя линия) с помощью агентного подхода и статистические данные (черные точки) с указанием введенных мер, влияющих на результаты моделирования и статистику (вертикальные штриховые линии).

области на период с 15 апреля по 3 октября 2020 г. приведен на рис. 6.

Отметим, что более грубая математическая модель (семь уравнений в системе ОДУ) улавливает основную тенденцию по выявленным случаям (пик заболевания в регионе, см. рис. 6, а), однако в некоторых моментах не согласуется с общей статистикой (критические случаи, требующие аппаратов ИВЛ, см. рис. 6, б). Негладкие решения на рис. 6 являются результатом применения индекса самоизоляции от Яндекс, характеризующегося недельной сезонностью, сглаживание которого нарушит суть использования в математическом моделировании. Более подробный анализ моделирования и прогнозирования распространения коронавирусной инфекции в Москве и

Новосибирской области приведен в работе (Krivorotko et al., 2020b). В этом случае необходимо применять агентно-ориентированную модель, подробно описывающую небольшие статистики.

Заключение

В работе проведен анализ чувствительности и идентифицируемости математических моделей распространения эпидемии COVID-19, основанных на системах дифференциальных уравнений. Основу алгоритма составляет анализ матрицы чувствительности методами дифференциальной и линейной алгебры, показывающей степень зависимости неизвестных параметров моделей от заданных измерений.

Анализ чувствительности математической модели распространения коронавирусной инфекции COVID-19 показал, что параметр контагиозности вируса устойчиво определяется по количеству ежедневно выявляемых заболевших, критических больных и вылечившихся. С другой стороны, прогнозируемая доля госпитализированных пациентов, больных, находящихся в критическом состоянии и требующих подключения аппарата ИВЛ, а также коэффициент смертности определяются гораздо менее устойчиво. Для построения более реалистичного прогноза необходимо добавлять дополнительную информацию о процессе (например, о количестве ежедневных случаев госпитализации).

Задачи уточнения идентифицируемых параметров были сведены к задачам минимизации соответствующих целевых функционалов, описывающих близость данных моделирования к статистикам по выявленным, критическим и смертельным случаям в Новосибирской области. Использование абсолютной и квадратичной нормы отклонения данных от результатов моделирования при минимизации целевых функционалов не показало существенных различий при анализе результатов моделирования. Более общая камерная модель, состоящая из семи ОДУ, описывает основную тенденцию распространения коронавирусной инфекции, чувствительна к пикам выявленных случаев, однако некачественно описывает небольшие статистики (количество критических случаев в день t_k , смертельных

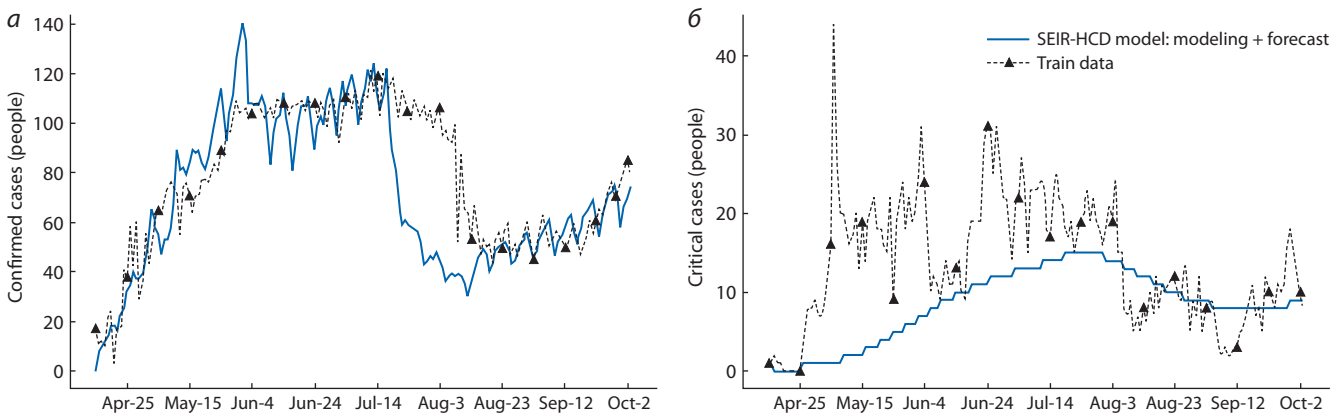


Рис. 6. Моделирование распространения COVID-19 в Новосибирской области (синяя линия) с 15 апреля по 3 октября 2020 г. и статистические данные (штриховая черная линия):

а – ежедневные выявленные случаи f_k ; б – критические случаи C_k , требующие подключения ИВЛ.

случаев), что может приводить к ошибочным выводам. Более подробная агентно-ориентированная математическая модель, в которой один из классов состояний агентов описывается системой из десяти ОДУ, позволяет детальнее улавливать небольшие изменения в статистике данных и строить сценарии развития распространения эпидемии.

Список литературы/ References

- Лихошвай В.А., Фадеев С.И., Демиденко Г.В., Матушкин Ю.Г. Моделирование уравнением с запаздывающим аргументом многостадийного синтеза без ветвления. *Сиб. журн. индустр. математики*. 2004;7(1):73-94.
[Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. Modeling nonbranching multistage synthesis by an equation with retarded argument. *Sibirskiy Zhurnal Industrialnoy Matematiki = Journal of Applied and Industrial Mathematics*. 2004;7(1):73-94. (in Russian)]
- Adams B.M., Banks H.T., Davidiana M., Kwona H.D., Trana H.T., Wynnea S.N., Rosenbergs E.S. HIV dynamics: modeling, data analysis, and optimal treatment protocols. *J. Comput. Appl. Math.* 2004; 184:10-49. DOI 10.1016/j.cam.2005.02.004.
- Bellu G., Saccomani M.P., Audoly S., D'Angiò L. DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput. Methods Programs Biomed.* 2007;88(1):52-61. DOI 10.1016/j.cmpb.2007.07.002.
- Gomez J., Prieto J., Leon E., Rodriguez A. INFEKTA: a general agent-based model for transmission of infectious diseases: studying the COVID-19 propagation in Bogotá – Colombia. *MedRxiv*. 2020. DOI 10.1101/2020.04.06.20056119.
- Habtemariam T., Tameru B., Nganwa D., Beyene G., Ayanwale L., Robnett V. Epidemiologic modeling of HIV/AIDS: use of computational models to study the population dynamics of the disease to assess effective intervention strategies for decision-making. *Adv. Syst. Sci. Appl.* 2008;8(1):35-39.
- Kabanikhin S.I. Definitions and examples of inverse and ill-posed problems. *J. Inverse Ill-Posed Probl.* 2008;16(4):317-357. DOI 10.1515/JIIP.2008.019.
- Kabanikhin S.I., Voronov D.A., Grodz A.A., Krivorotko O.I. Identifiability of mathematical models in medical biology. *Russ. J. Genet. Appl. Res.* 2016;6(8):838-844. DOI 10.1134/S2079059716070054.
- Kermack W.O., McKendrick A.G. A contribution of the mathematical theory of epidemics. *Proc. R. Soc. Lond. A.* 1927;115:700-721. DOI 10.1098/rspa.1927.0118.
- Kerr C., Stuart R., Mistry D., Abeysuriya R., Hart G., Rosenfeld K., Selvaraj P., Nunez R., Hagedorn B., George L., Izzo A., Palmer A., Delpont D., Bennette C., Wagner B., Chang S., Cohen J., Panovska-Griffiths J., Jastrzebski M., Oron A., Wenger E., Famulare M., Klein D. Covasim: an agent-based model of COVID-19 dynamics and interventions. *MedRxiv*. 2020. DOI 10.1101/2020.05.10.20097469.
- Krivorotko O.I., Andornaya D.V., Kabanikhin S.I. Sensitivity analysis and practical identifiability of some mathematical models in biology. *J. Appl. Ind. Math.* 2020a;14:115-130. DOI 10.1134/S1990478920010123.
- Krivorotko O.I., Kabanikhin S.I., Zyat'kov N.Yu., Prikhod'ko A.Yu., Prokhoshin N.M., Shishlenin M.A. Mathematical modeling and forecasting of COVID-19 in Moscow and Novosibirsk region. *Numer. Analysis Applications*. 2020b;13(4):332-348. DOI 10.1134/S1995423920040047.
- Lauer S.A., Grantz K.H., Bi Q., Jones F.K., Zheng Q., Meredith H., Azman A.S., Reich N.G., Lessler J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* 2020;172:577-582. DOI 10.7326/m20-0504.
- Lee W., Liu S., Tembine H., Li W., Osher S. Controlling propagation of epidemics via mean-field games. *ArXiv*. 2020;arXiv:2006.01249.
- Miao H., Xia X., Perelson A.S., Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.* 2011;53(1):3-39. DOI 10.1137/090757009.
- Raue A., Becker V., Klingmüller U., Timmer J. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*. 2010;20(4):045105. DOI 10.1063/1.3528102.
- Raue A., Karlsson J., Saccomani M.P., Jirstrand M., Timmer J. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*. 2014;30(10):1440-1448. DOI 10.1093/bioinformatics/btu006.
- Tuomisto J.T., Yrjölä J., Kolehmainen M., Bonsdorff J., Pekkanen J., Tikkanen T. An agent-based epidemic model REINA for COVID-19 to identify destructive policies. *MedRxiv*. 2020. DOI 10.1101/2020.04.09.20047498.
- Unlu E., Leger H., Motorny O., Rukubayihunga A., Ishacian T., Chouiten M. Epidemic analysis of COVID-19 outbreak and counter-measures in France. *MedRxiv*. 2020. DOI 10.1101/2020.04.27.20079962.
- Verity R., Okell L., Dorigatti I., Winskill P., Whittaker C., Imai N., Cuomo-Dannenburg G., Thompson H., Walker P., Fu H., Dighe A., Griffin J., Baguelin M., Bhatia S., Boonyasiri S., Cori A., Cucunubá Z., FitzJohn R., Gaythorpe K., Green W., Hamlet A., Hinsley W., Laydon D., Nedjati-Gilani G., Riley S., Elstrand S., Volz E., Wang H., Wang Y., Xi X., Donnelly C., Ghani A., Ferguson N.M. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* 2020;20(6):669-677. DOI 10.1016/S1473-3099(20)30243-7.
- Voropaeva O.F., Tsgoev Ch.A. A numerical model of inflammation dynamics in the core of myocardial infarction. *J. Appl. Ind. Math.* 2019;13(2):372-383. DOI 10.1134/S1990478919020182.
- Wolfram C. An agent-based model of COVID-19. *Complex Syst.* 2020; 29(1):87-105. DOI 10.25088/ComplexSystems.29.1.87.
- Wölfel R., Corman V.M., Guggemos W., Seilmaier M., Zange S., Müller M.A., Niemeyer D., Jones T.C., Vollmar P.V., Rothe C., Hoelscher M., Bleicker T., Brünink S., Schneider J., Ehmann R., Zwirgmaier K., Drosten C., Wendtner C. Virological assessment of hospitalized patients with COVID-2019. *Nature*. 2020;581:465-469. DOI 10.1038/s41586-020-2196-x.
- Yao K.Z., Shaw B.M., Kou B., McAuley K.B., Bacon D.W. Modeling ethylene/butene copoly-merization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Engineering*. 2003;11(3):563-588. DOI 10.1081/PRE-120024426.

ORCID ID

O.I. Krivorotko orcid.org/0000-0003-0125-4988

Благодарности. Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 18-31-20019) и Совета по грантам Президента Российской Федерации (соглашение № 075-15-2019-1078, МК-814.2019.1).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 25.10.2020. После доработки 17.12.2020. Принята к публикации 18.12.2020.

Английский текст <https://vavilov.elpub.ru/jour>

Механический стресс клеток мозга, локальная трансляция и нейродегенеративные заболевания: молекулярно-генетические аспекты

Т.М. Хлебодарова^{1, 2}

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

² Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

✉ tamara@bionet.nsc.ru

Аннотация. Идея о том, что хронический механический стресс, который испытывают клетки мозга при повышенном внутричерепном давлении, артериальной гипертензии или вследствие травмы, может быть одним из факторов риска в развитии нейродегенеративных заболеваний, появилась еще в 90-е годы прошлого столетия и поддерживается в настоящее время. Однако молекулярно-генетические механизмы реализации событий, ведущих от механического воздействия на клетки к нарушению пластичности синапсов и последующему изменению поведения, когнитивных способностей и памяти, не ясны. В настоящем обзоре рассмотрены существующие данные о молекулярно-генетических механизмах регуляции локальной трансляции и актинового цитоскелета в активированном синапсе, играющих центральную роль в формировании различных видов пластичности синапса и долговременной памяти, и возможных путей влияния механического стресса на их состояние. Обсуждается роль mTOR сигнального каскада, РНК-связывающего белка FMRP, белка CYFIP1, взаимодействующего с FMRP, семейства малых ГТФаз и WAVE регуляторного комплекса в регуляции инициации локальной трансляции и перестройки актинового цитоскелета в дендритных шипиках активированного синапса. Приводятся факты, свидетельствующие о том, что в условиях хронического механического стресса возможна aberrantная активация mTOR сигнального каскада и WAVE регуляторного комплекса через сенсор механических сигналов – регуляторный фактор YAP/TAZ, следствием которой могут быть нарушения активности системы локальной трансляции, а также связанных с ними механизмов регуляции формирования F-актиновых филаментов и структуры дендритных шипиков. Это может быть одной из причин развития различных неврологических патологий, включая аутистические расстройства и эпилептическую энцефалопатию. Высказывается оригинальная гипотеза, согласно которой одной из возможных причин синаптопатий может быть нарушение стабильности протеома, связанное с гиперактивностью mTOR и формированием сложных динамических режимов синтеза белков *de novo* в ответ на стимуляцию синапса, в том числе и в условиях хронического механического стресса.

Ключевые слова: синапс; механосенсор YAP/TAZ; mTOR; FMRP-зависимая трансляция; сложная динамика; F-актин; WAVE регуляторный комплекс; расстройства аутистического спектра; эпилептическая энцефалопатия.

Для цитирования: Хлебодарова Т.М. Механический стресс клеток мозга, локальная трансляция и нейродегенеративные заболевания: молекулярно-генетические аспекты. *Вавиловский журнал генетики и селекции*. 2021;25(1):92-100. DOI 10.18699/VJ21.011

The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases

Т.М. Khlebodarova^{1, 2}

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

² Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

✉ tamara@bionet.nsc.ru

Abstract. The assumption that chronic mechanical stress in brain cells stemming from intracranial hypertension, arterial hypertension, or mechanical injury is a risk factor for neurodegenerative diseases was put forward in the 1990s and has since been supported. However, the molecular mechanisms that underlie the way from cell exposure to mechanical stress to disturbances in synaptic plasticity followed by changes in behavior, cognition, and memory are still poorly understood. Here we review (1) the current knowledge of molecular mechanisms regulating local translation and the actin cytoskeleton state at an activated synapse, where they play a key role in the formation of various sorts of synaptic plasticity and long-term memory, and (2) possible pathways of mechanical stress intervention. The roles of the mTOR (mammalian target of rapamycin) signaling pathway; the RNA-binding FMRP protein; the CYFIP1 protein, interacting with FMRP; the family of small GTPases; and the WAVE regulatory complex in the regulation of translation initiation and actin cytoskeleton rearrangements in dendritic spines of the activated synapse are discussed. Evidence is provided that chronic mechanical stress may result in aberrant activation of mTOR signaling and the WAVE regulatory complex via the YAP/TAZ system, the key sensor of mechanical signals, and influence the associated pathways regulating the formation of F actin filaments and the dendritic

spine structure. These consequences may be a risk factor for various neurological conditions, including autistic spectrum disorders and epileptic encephalopathy. In further consideration of the role of the local translation system in the development of neuropsychic and neurodegenerative diseases, an original hypothesis was put forward that one of the possible causes of synaptopathies is impaired proteome stability associated with mTOR hyperactivity and formation of complex dynamic modes of *de novo* protein synthesis in response to synapse-stimulating factors, including chronic mechanical stress. Key words: synapse; YAP/TAZ mechanosensor; mTOR; FMRP-dependent translation; complex dynamics; F-actin; WAVE regulatory complex; autism spectrum disorders; epileptic encephalopathy.

For citation: Khlebodarova T.M. The molecular view of mechanical stress of brain cells, local translation, and neurodegenerative diseases. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):92-100. DOI 10.18699/VJ21.011

Механический стресс и нейродегенеративные заболевания

Механические сигналы играют важную роль в принятии решений о судьбе клеток, касающихся пролиферации, выживания, дифференцировки, а также процессов регенерации тканей и заживления ран. Механотрансдукция включает в себя восприятие этих сил и их трансляцию в биохимические и молекулярно-генетические сигналы, в том числе активацию сигнальных каскадов и экспрессии определенных генов, что позволяет клеткам адаптироваться к своей физической среде. Многочисленные исследования выявили центральную роль транскрипционного регулятора YAP (yes-associated protein 1) и его паралога TAZ (transcriptional co-activator with PDZ-binding motif) (YAP/TAZ) как сенсоров и медиаторов механических сигналов (Dupont et al., 2011; Totaro et al., 2018; Dasgupta, McCollum, 2019). Нарушение взаимодействия клетки с окружающей средой приводит к aberrантной активации YAP/TAZ, что способствует множественным заболеваниям, таким как атеросклероз, фиброз, легочная гипертензия, воспаление, мышечная дистрофия и рак (Levy Nogueira et al., 2015, 2018; Yu et al., 2015; Panciera et al., 2017; Hong et al., 2019; Zhu et al., 2020). В последние годы появились данные о том, что механический стресс может быть также одной из причин развития нейродегенеративных процессов в мозге, в частности при болезни Альцгеймера (Levy Nogueira et al., 2015, 2016a, b, 2018).

Гипотеза о том, что хронический механический стресс, который испытывают клетки мозга при повышенном внутричерепном давлении, артериальной гипертензии или вследствие травмы, может быть одним из факторов риска в развитии нейродегенеративных заболеваний, высказывалась достаточно давно (Wostyn, 1994) и поддерживается в настоящее время (Levy Nogueira et al., 2018).

Какие данные могут свидетельствовать о том, что существуют механизмы воздействия механического стресса на функции нервных клеток? Во-первых, оказалось, что YAP/TAZ – ключевой сенсор и медиатор механических сигналов – активирует mTOR (mammalian target of rapamycin) сигнальный каскад (Tumaneng et al., 2012; McCarthy, 2013; Hu et al., 2017). Он играет основную роль в регуляции локальной сар-зависимой трансляции в синапсе, обеспечивающей динамическую пластичность синапса в ответ на внешние стимулы, лежащую в основе процессов обучения и памяти (Costa-Mattioli et al., 2009; Buffington et al., 2014; Rosenberg et al., 2014; Santini et al., 2014), нарушения которых ведет к синаптической дисфункции и развитию различных форм нейропсихических заболеваний (Trifonova et al., 2017). YAP/TAZ активирует mTOR

через два механизма (рис. 1): стимулируя транскрипцию ГТФазы Rheb (Ras homologue enriched in brain) (Hu et al., 2017), которая является активатором mTORC1 киназы, и ингибируя трансляцию PTEN (phosphatase and tensin homolog) микроРНК miR29, способствуя таким образом aberrантной PI3K-опосредованной активации mTORC1 и mTORC2 киназ (Tumaneng et al., 2012; McCarthy, 2013).

Во-вторых, ключевой медиатор механических сигналов – это актиновый цитоскелет клетки (Seo, Kim, 2018). Его перестройки в дендритных шипиках нервных клеток играют существенную роль в процессах обучения и формирования долговременной памяти (Basu, Lamprecht, 2018; Borovac et al., 2018) и контролируются активностью Rho ГТФаз (Tapon, Hall, 1997), чрезмерная или недостаточная активность которых приводит к нарушению структуры дендритных шипиков, снижению памяти и способности к обучению и может быть причиной множественных расстройств нервного развития с разной этиологией (Ba et al., 2013; Rygonneau et al., 2017; Zamboni et al., 2018; Nishiyama, 2019). В активированном синапсе функционирование Rho ГТФаз в существующей степени зависит от их *de novo* синтеза, и следовательно, от mTOR (Briz et al., 2015).

Активация mTOR сигнального каскада в условиях механического стресса через механосенсор YAP/TAZ (Tumaneng et al., 2012; McCarthy, 2013; Hu et al., 2017) создает также условия для индукции через киназу S6K и ГТФазу RAC1 (Derivery et al., 2009) распада гетеропентамерного WAVE регуляторного комплекса (WASP family verprolin homologue) на субкомплексы, что способствует взаимодействию WAVE1 с Arp2/3 (Cory, Ridley, 2002; Millard et al., 2004; Abekhoukh, Bardoni, 2014; Molinie, Gautreau, 2018) и aberrантной полимеризации актина, нарушающей структуру дендритных шипиков (см. рис. 1).

Таким образом, пути влияния механических воздействий на функционирование нервных клеток могут быть связаны с активацией mTOR сигнального каскада и перестройками актинового цитоскелета в дендритных шипиках, которые, в свою очередь, зависят от активности системы локальной трансляции в синапсе, контролируемой mTOR. Именно нарушения функционирования системы локальной трансляции в синапсе, в том числе вызванные повышенной активностью mTOR, характеризующиеся нарушением пластичности синапса в виде дисбаланса процессов его возбуждения и торможения (Gobert et al., 2020), связывают с различными нейропсихическими заболеваниями, в том числе с аутистическими расстройствами, эпилепсией, болезнями Паркинсона и Альцгеймера (Gkogkas, Sonenberg, 2013; Meng et al., 2013; Won et

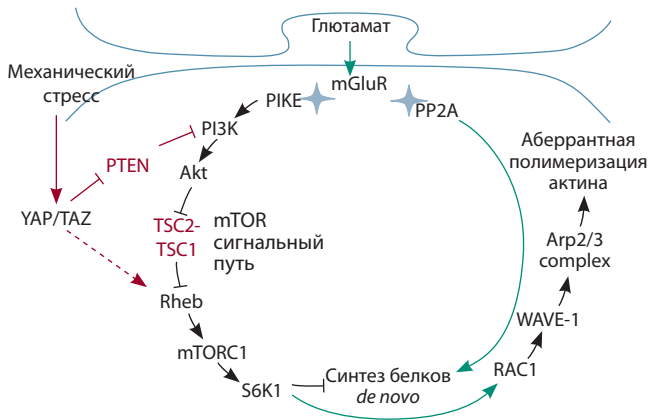


Рис. 1. Возможные пути влияния механического стресса через mTOR сигнальный путь на активность локальной трансляции и формирование актинового цитоскелета в дендритных шипиках глутаматергических синапсов пирамидальных клеток гиппокампа.

mGluR – белок-рецептор; PIKE (PI3-kinase enhancer); Rheb (Ras homologue enriched in brain), Rac1 – ГТФазы; PI3K (phosphatidylinositol-3-kinase), Akt (protein kinase B), S6K1 – киназы; TSC1/2 (tuberous sclerosis complex 1/2) – комплекс туберозного склероза; mTOR (mechanistic target of rapamycin) – серин/треонин киназа; PTEN (phosphatase and tensin homolog), PP2A (protein phosphatase 2A) – фосфатазы; YAP/TAZ – механосенсор; WAVE-1 – компонент WAVE (WASP family verprolin homologue) регуляторного комплекса; Arp2/3 – актин-связывающие белки. Красным цветом показаны белки, мутации генов которых связаны с неврологическими заболеваниями. Зелеными стрелками отмечены процессы активации трансляции через PP2A фосфатазу и полимеризации актина через S6 киназу и Rac1 ГТФазу в ответ на стимуляцию синапса глутаматом. Красными стрелками обозначены возможные механизмы влияния механического стресса на активность mTOR сигнального пути.

al., 2013; Cai et al., 2015; Huber et al., 2015; Pramparo et al., 2015; Klein et al., 2016; Martin, 2016; Onore et al., 2017).

В связи с этим молекулярно-генетические механизмы регуляции локальной трансляции и динамических перестроек актинового цитоскелета в дендритных шипиках, которые контролируются ее активностью (Bramham, 2008), привлекают особое внимание.

Локальная трансляция и нейродегенеративные заболевания

В настоящее время существуют убедительные доказательства, что локальная сар-зависимая трансляция в постсинаптическом пространстве дендритного шипика обеспечивает его динамическую пластичность в ответ на внешние стимулы, лежащую в основе процессов обучения и памяти (Huber et al., 2000; Costa-Mattioli et al., 2009; Rosenberg et al., 2014; Santini et al., 2014; Louros, Osterweil, 2016).

Существует достаточно много примеров того, что нарушение механизмов контроля локальной трансляции в синапсе приводит к различным нейропсихическим заболеваниям, таким как аутизм, эпилепсия, болезнь Паркинсона и др. (Gkogkas, Sonenberg, 2013; Buffington et al., 2014; Klein et al., 2016; Martin, 2016; Trifonova et al., 2017). Основные регуляторные события, обеспечивающие активацию локального синтеза белков в дендритных шипиках глутаматергических синапсов пирамидальных клеток гиппокампа в ответ на стимуляцию глутаматом mGluR (metabotropic glutamate receptor) рецепторов на

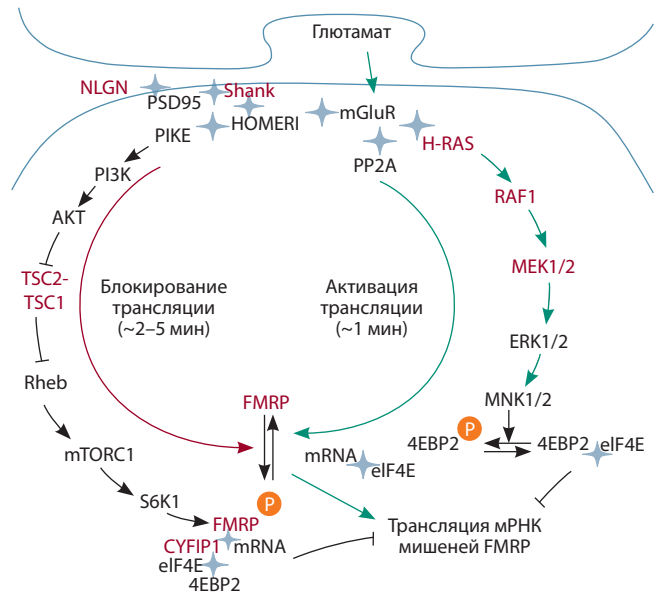


Рис. 2. Простейшая схема регуляции локальной трансляции в дендритных шипиках глутаматергических синапсов пирамидальных клеток гиппокампа в ответ на стимуляцию синапса.

mGluR – белок-рецептор; NLGN, Shank, PSD95, HOMER1 – белки, формирующие структуру постсинаптической мембраны; PIKE (PI3-kinase enhancer); Rheb (Ras homologue enriched in brain) – ГТФазы; PI3K (phosphatidylinositol-3-kinase), Akt (protein kinase B), S6K1 – киназы; TSC1/2 (tuberous sclerosis complex 1/2) – комплекс туберозного склероза; mTOR (mechanistic target of rapamycin) – серин/треонин киназа; FMRP (fragile X mental retardation protein) – РНК связывающий белок, негативный регулятор трансляции; PP2A (protein phosphatase 2A) – фосфатаза; H-RAS – ГТФаза; RAF1, MEK1/2, ERK1/2, MNK1/2 – киназы; eIF4E – фактор инициации трансляции; 4EBP2 – 4E-связывающий белок; CYFIP1 (cytoplasmic FMRP interacting protein 1) – FMRP-взаимодействующий белок. Красным цветом отмечены названия белков, мутации генов которых связаны с неврологическими заболеваниями. Зелеными стрелками показаны пути активации локальной трансляции через PP2A фосфатазу и RAS/ERK сигнальный путь, красной стрелкой – блокирование через mTOR сигнальный каскад.

постсинаптической мембране возбуждающих синапсов, приведены на рис. 2. Регуляция активности локальной трансляции обеспечивается через mTOR и RAS/ERK сигнальные каскады (Huber et al., 2000; Darnell, Klann, 2013; Beggs et al., 2015; Chen, Joseph, 2015).

Центральным звеном регуляции локальной сар-зависимой трансляции в синапсе является РНК-связывающий белок FMRP (fragile X mental retardation protein) (Feng et al., 1997). Он служит мишенью S6 киназы и PP2A фосфатазы, которые активируются в ответ на стимуляцию mGluR рецепторов (Narayanan et al., 2007, 2008), и блокирует трансляцию в фосфорилированном состоянии, связываясь с мРНК, рибосомами и eIF4E фактором инициации трансляции (Brown et al., 1998; Napoli et al., 2008; Chen et al., 2014). Дефосфорилирование нарушает связь FMRP со своими мишенями, что ведет, с одной стороны, к активации трансляции мРНК, а с другой – к быстрой деградации самого FMRP (Nalavadi et al., 2012). Белок FMRP контролирует эффективность трансляции через сайты связывания с РНК (Chen, Joseph, 2015). Он напрямую связывается с кодирующей и 3'-UTR последовательностью мРНК (Brown et al., 1998; Darnell et al., 2011) и L5 белком 80S рибосомы (Chen et al., 2014), контролируя таким об-

разом элонгацию и терминацию транскрипции. Причем репрессия трансляции в 3'-UTR может осуществляться также через физическое взаимодействие FMRP с TDP-43 (TAR DNA binding protein, 43 kDa) белком (Majumder et al., 2016).

Белок FMRP участвует также в регуляции трансляции на стадии ее инициации через взаимодействие с белком CYFIP1 (cytoplasmic FMRP interacting protein 1) (Napoli et al., 2008). Существующие гипотезы о механизмах регуляции трансляции через FMRP (Napoli et al., 2008; Majumder et al., 2016) предполагают участие в этом процессе одной молекулы белка, взаимодействующей с 3'-UTR через TDP-43 и с фактором инициации трансляции eIF4E через CYFIP1, т.е. FMRP и CYFIP1 – ключевые регуляторы инициации трансляции в активированном синапсе.

Мишенями FMRP являются мРНК белков компонентов mTOR сигнального пути (PI3K киназа, PTEN фосфатаза, TSC2 – tuberous sclerosis complex-2, mTOR, PP2A фосфатаза), белков-рецепторов (mGluR, NMDAR, AMPAR), белков, формирующих структуру постсинаптической мембраны (NLGN, SHANK, PSD95), системы убиквитин-зависимой деградации белков (E3 убиквитин-лигаза) и собственная мРНК – FMR1 (Brown et al., 1998; Mudashetty et al., 2007; Gross et al., 2010; Sharma et al., 2010; Darnell et al., 2011; Ascano et al., 2012), что свидетельствует о ключевой роли FMRP в динамической регуляции протеома в активированном синапсе (Zukin et al., 2009; Iacoangeli, Tiedge, 2013).

Известно, что мутации генов, кодирующих большинство из этих белков, ведут к нарушению функций синапса и различным патологиям. Так, мутации гена, кодирующего белок постсинаптической мембраны SHANK3, вызывают синдром Фелана–МакДемиды, фосфатазы PTEN – синдром Каудена, NF1 – нейрофиброматоз 1-го типа, ГТФазы H-RAS и RAF1 и киназы MEK1 – синдром Костелло–Нунана, TSC2-TSC1 – туберозный склероз, FMRP – синдром ломкой X-хромосомы, убиквитин-лигазы UBE3A – синдром Энгельмана, нейролигины NLGN3/4 и нейрексин NRXN1 – типичный аутизм (Трифорова и др., 2016). Мутации гена *Shank3* и нарушение экспрессии его мРНК связывают также с аутизмом, шизофренией и эпилепсией (Peça et al., 2011; Mei et al., 2016; de Sena Cortabitarte et al., 2017; Monteiro, Feng, 2017; Fu et al., 2020). Мутации гена фосфатазы PTEN часто сопровождаются такими неврологическими проявлениями, как макроцефалия, эпилепсия, снижение интеллекта и аутизм (Zhou, Parada, 2012; Трифорова и др., 2016).

Эти данные подтверждают ключевую роль системы локальной трансляции в функционировании синапса и позволяют предположить, что одной из возможных причин синаптопатий, наблюдаемых при аутизме и некоторых других нейropsychических заболеваниях, может быть нарушение стабильности протеома (Klein et al., 2016; Louros, Osterweil, 2016), важного с точки зрения формирования пластичности синапса и долговременной памяти (Cajigas et al., 2010).

Здесь необходимо отметить, что в структурно-функциональной организации системы регуляции активности FMRP присутствуют петли негативной и позитивной регуляции, которые служат одним из факторов нестабильности

в молекулярно-генетических системах (Mackey, Glass, 1977; Decroly, Goldbeter 1982; Goldbeter et al., 2001; Bastos de Figueiredo et al., 2002; Likhoshvai et al., 2013, 2015, 2016, 2020; Когай и др., 2015; Suzuki et al., 2016; Khlebodarova et al., 2017; Kogai et al., 2017).

Эти регуляторные петли функционируют в различных временных диапазонах и связаны с быстрой (~1 мин) активацией трансляции FMRP-зависимых мРНК через PP2A фосфатазу и достаточно быстрым ее блокированием через активацию S6 киназы (2–5 мин) (Narayanan et al., 2007, 2008). То есть нормальное функционирование синапса обеспечивается тонкими динамическими взаимоотношениями между компонентами этих сигнальных путей в активированном синапсе (см. рис. 2).

Теоретический анализ динамических особенностей функционирования системы локальной трансляции показал, что увеличение скорости и эффективности FMRP-зависимой трансляции может быть фактором возникновения нестабильности в системе локальной трансляции, причем в физиологической области ее функционирования (Хлебодарова и др., 2018; Лихошвай, Хлебодарова, 2019). То есть в основе известных фактов аутистических расстройств, сопровождаемых повышенной активностью аппарата трансляции в синапсе (Pramparo et al., 2015; Onore et al., 2017), могут лежать нарушения стабильности протеома, возникающие в результате формирования сложных динамических режимов синтеза рецепторных белков в ответ на стимуляцию синапса (Khlebodarova et al., 2018, 2020). И это совершенно новый взгляд на возможные причины возникновения синаптопатий.

Следует добавить, что повышенная активность mTOR сигнального пути является общей характеристикой не только для аутистических расстройств, но также для таких психических и неврологических заболеваний, как болезнь Альцгеймера (Pei, Hugon, 2008), эпилепсия (Wong, 2010) и даже синдром Дауна (Troca-Marin et al., 2012). Раннее старение и возрастные нейродегенеративные патологии у людей также ассоциируют с повышенной активностью mTOR (Johnson et al., 2013).

Гипотеза о том, что высокая копияность генов рибосомной РНК у отдельных индивидуумов может быть фактором риска в развитии аутистических расстройств, шизофрении и умственной отсталости, кажется вполне достоверной (Chestkov et al., 2018; Porokhovnik, 2019; Porokhovnik, Lyapunova, 2019), если при этом предположить, что вариации числа копий генов рРНК у особей коррелируют с концентрацией рибосом в клетке и активностью аппарата трансляции.

Актиновый цитоскелет и нейродегенеративные заболевания

Структура актинового цитоскелета определяет морфологию дендритных шипиков нервных клеток, а его перестройки, обеспечиваемые быстрой сборкой или разборкой мономеров актина (G-актин) в филаменты (F-актин), играют важную роль в формировании синаптической пластичности и долговременной памяти (Penzes, Rafalovich, 2012; Basu, Lamprecht, 2018). Ряд нейродегенеративных заболеваний, таких как болезнь Альцгеймера, шизофрения и аутизм, связывают с нарушениями механизмов ре-

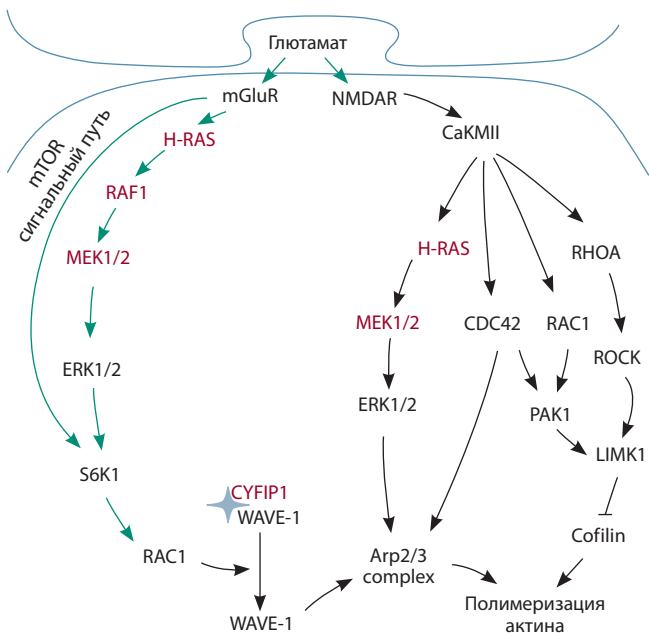


Рис. 3. Схема регуляции формирования актинового цитоскелета в дендритных шипиках глутаматергических синапсов пирамидальных клеток гиппокампа в ответ на стимуляцию синапса.

mGluR, NMDAR – белки-рецепторы; H-RAS, RHOA, RAC1, CDC42 – RAS семейство малых ГТФаз; CaMKII, MEK1/2, RAF1, ERK1/2, S6K1, PAK1, ROCK, LIMK1 – киназы; CYFIP1 – FMRP-взаимодействующий белок; WAVE-1 – компонент WAVE регуляторного комплекса; Arp2/3 – актин-связывающие белки; Cofilin – кофилин, фактор деполимеризации актина. Красным цветом отмечены белки, мутации генов которых связаны с неврологическими заболеваниями. Зелеными стрелками показан путь регуляции полимеризации актина через mTOR сигнальный путь, черными – через CaMKII киназу.

гуляции формирования F-актиновых филаментов и структуры дендритных шипиков (Bamburg, Bernstein, 2016; Borovac et al., 2018; Forrest et al., 2018; Ben Zablah et al., 2020; Lauterborn et al., 2020). Основные регуляторные события, обеспечивающие перестройку актинового цитоскелета в дендритных шипиках глутаматергических синапсов пирамидальных клеток гиппокампа, которые активируются в ответ на стимуляцию глутаматом mGluR (metabotropic glutamate receptor) и NMDAR (N-methyl-D-aspartate receptor, ionotropic glutamate receptor) рецепторов на постсинаптической мембране возбуждающих синапсов, представлены на рис. 3.

В активированном синапсе индукция формирования актиновых филаментов и их стабилизация в существенной степени зависят от активности кофилина и WAVE регуляторного комплекса, которая контролируется S6K, LIMK1 и PAK1 киназами и осуществляется через сигнальные пути, опосредуемые RAS семейством малых ГТФаз – H-RAS, RhoA, Rac1, Cdc42 (Tapon, Hall, 1997; Rex et al., 2009; Ip et al., 2011; Chen et al., 2017; Schaks et al., 2018). Активность этих сигнальных каскадов в значительной степени зависит от быстрого *de novo* синтеза Rho ГТФаз (Briz et al., 2015). Так, блокирование синтеза белка в дендритных шипиках клеток гиппокампа полностью подавляло стимуляцию активности RhoA ГТФазы, фосфорилирование кофилина и полимеризацию актина (Briz et al., 2015), а мутация гена *Frm1*, кодирующего РНК-связывающий белок FMRP,

полностью подавляла физиологическую стимуляцию активности ГТФазы Rac1 и ее эффектора PAK1 киназы и нарушала стабилизацию актиновых филаментов в синапсах клеток гиппокампа (Chen et al., 2010).

Таким образом, активность RAS ГТФаз, контролирующей формирование и стабилизацию актиновых филаментов в дендритных шипиках, напрямую зависит от их *de novo* синтеза, т.е. от активности mTOR и FMRP-зависимой локальной трансляции, возможная нестабильность функционирования которой (Khlebodarova et al., 2018, 2020; Лихошвай, Хлебодарова, 2019) также может быть причиной недостаточной или чрезмерной активности RAS, ведущей к нарушению структуры дендритных шипиков и возникновению связанных с этим неврологических расстройств (Ba et al., 2013; Puygonneau et al., 2017; Zamboni et al., 2018; Nishiyama, 2019).

В регуляции гетеропентамерного WAVE регуляторного комплекса ключевую роль играет Rac1 ГТФаза, активность которой в существенной степени зависит от S6K и mTOR1 киназ. Этот комплекс в обычном состоянии неактивен, но взаимодействие с Rac-GTPазой ведет к его диссоциации на два субкомплекса, CYFIP1- и WAVE1-содержащий (Derivery et al., 2009). Последний взаимодействует с Arp2/3 (actin-related proteins) комплексом и индуцирует полимеризацию актина (см. рис. 3) (Cory, Ridley, 2002; Millard et al., 2004; Abekhoukh, Bardoni, 2014; Molinie, Gautreau, 2018).

Распад WAVE регуляторного комплекса и aberrantная активация WAVE1 приводят к эпилептической энцефалопатии (Nakashima et al., 2018; Zhang et al., 2019; Zweier et al., 2019; Schaks et al., 2020). Эта возможность существует при нарушении стехиометрического контроля синтеза компонентов WAVE (Abekhoukh et al., 2017) и мутациях, влияющих на интерфейс взаимодействия WAVE1 и CYFIP2 белка (Nakashima et al., 2018; Zhang et al., 2019; Zweier et al., 2019; Schaks et al., 2020).

Необходимо также отметить, что CYFIP1, являясь одним из основных компонентов WAVE регуляторного комплекса, также участвует в регуляции трансляции на стадии ее инициации через взаимодействие с РНК-связывающим белком FMRP (Napoli et al., 2008). То есть механизмы регуляции локальной трансляции и перестроек актинового цитоскелета в дендритных шипиках нервных клеток оказываются взаимосвязанными и через белок CYFIP1 (De Rubeis et al., 2013).

Заключение

Анализ существующих к настоящему времени данных показал, что механизмы регуляции системы локальной трансляции в синапсе и динамических перестроек актинового цитоскелета в дендритных шипиках нервных клеток, играющих центральную роль в формировании различных видов пластичности синапса и долговременной памяти, тесно связаны между собой и с активностью механосенсора YAP/TAZ, который опосредованно, через mTOR и S6K киназу, может влиять и на активность трансляции, и на состояние актиновых филаментов в дендритных шипиках (Tapon, Hall, 1997; Tumaneng et al., 2012; McCarthy, 2013; Reddy et al., 2013; Briz et al., 2015; Hu et al., 2017; Seo, Kim, 2018).

Тот факт, что гиперактивность mTOR и нарушение функций практически каждого компонента системы локальной трансляции и систем, контролирующих перестройки актинового цитоскелета в дендритных шипиках, могут быть причиной множественных расстройств нервного развития с разной этиологией, достаточно обоснован (Pei, Hugon, 2008; Wong, 2010; Johnson et al., 2013; Prampero et al., 2015; Onore et al., 2017; Pyronneau et al., 2017; Trifonova et al., 2017; Nakashima et al., 2018; Nishiyama, 2019; Zhang et al., 2019).

Теоретические исследования динамических особенностей функционирования системы локальной трансляции (Khlebodarova et al., 2018, 2020; Лихошвай, Хлебодарова, 2019) позволяют предположить, что одним из возможных механизмов неврологических расстройств, возникающих при хроническом механическом стрессе, может быть aberrantная гиперактивация mTOR, провоцирующая динамическую нестабильность синтеза белков *de novo* в активированном синапсе.

Таким образом, в настоящее время ясно, что хронический механический стресс может быть одним из факторов риска возникновения синаптопатий и развития нейродегенеративных заболеваний вследствие гиперактивации mTOR, ведущей к нарушению стабильности протеома, столь необходимого для формирования пластичности синапса и долговременной памяти (Klein et al., 2016; Louros, Osterweil, 2016).

Список литературы / References

- Когай В.В., Хлебодарова Т.М., Фадеев С.И., Лихошвай В.А. Сложная динамика в системах альтернативного сплайсинга мРНК: математическая модель. *Вычисл. технологии*. 2015;20(1): 38-52.
[Kogai V.V., Khlebodarova T.M., Fadeev S.I., Likhoshvai V.A. Complex dynamics in alternative mRNA splicing: mathematical model. *Vychislitelnye Tekhnologii = Computational Technologies*. 2015;20(1):38-52. (in Russian)]
- Лихошвай В.А., Хлебодарова Т.М. О стационарных решениях уравнения с запаздывающими аргументами: модель локальной трансляции в синапсе. *Матем. биол. биоинформ.* 2019;14(2): 554-569. DOI 10.17537/2019.14.554.
[Likhoshvai V.A., Khlebodarova T.M. On stationary solutions of delay differential equations: a model of local translation in synapses. *Matematicheskaya Biologiya i Bioinformatika = Mathematical Biology and Bioinformatics*. 2019;14(2):554-569. DOI 10.17537/2019.14.554. (in Russian)]
- Трифонова Е.А., Хлебодарова Т.М., Груntenко Н.Е. Аутизм как проявление нарушения молекулярных механизмов регуляции развития и функций синапсов. *Вавиловский журнал генетики и селекции*. 2016;20(6):959-967. DOI 10.18699/VJ16.217.
[Trifonova E.A., Khlebodarova T.M., Gruntenko N.E. Molecular mechanisms of autism as a form of synaptic dysfunction. *Vavilovskii Zhurnal Genetiki i Selektcii = Vavilov Journal of Genetics and Breeding*. 2016;20(6):959-967. DOI 10.18699/VJ16.217. (in Russian)]
- Хлебодарова Т.М., Когай В.В., Лихошвай В.А. О хаотическом потенциале системы локальной трансляции в активированном синапсе. В: *Математическая биология и биоинформатика*. Пушкино: ИМПБ РАН, 2018;7:e68.1-e68.6. DOI 10.17537/icmbb18.6.
[Khlebodarova T.M., Likhoshvai V.A., Kogai V.V. On the chaotic potential of local translation at activated synapses. In: *Mathematical Biology and Bioinformatics*. Pushchino: IMPB RAS, 2018;7:e68.1-e68.6. DOI 10.17537/icmbb18.6. (in Russian)]
- Abekhouk S., Bardoni B. CYFIP family proteins between autism and intellectual disability: links with Fragile X syndrome. *Front. Cell. Neurosci.* 2014;8:81. DOI 10.3389/fncel.2014.00081.
- Abekhouk S., Sahin H.B., Grossi M., Zongaro S., Maurin T., Madrigal I., Kazue-Sugioka D., Raas-Rothschild A., Doulazmi M., Carrera P., Stachon A., Scherer S., Drula Do Nascimento M.R., Trembleau A., Arroyo I., Szatmari P., Smith I.M., Milà M., Smith A.C., Giangrande A., Caillé I., Bardoni B. New insights into the regulatory function of CYFIP1 in the context of WAVE- and FMRP-containing complexes. *Dis. Model. Mech.* 2017;10(4):463-474. DOI 10.1242/dmm.025809.
- Ascano M. Jr., Mukherjee N., Bandaru P., Miller J.B., Nusbaum J.D., Corcoran D.L., Langlois C., Munschauer M., Dewell S., Hafner M., Williams Z., Ohler U., Tuschl T. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*. 2012;492: 382-386. DOI 10.1038/nature11737.
- Ba W., van der Raadt J., Nadif Kasri N. Rho GTPase signaling at the synapse: implications for intellectual disability. *Exp. Cell Res.* 2013; 319(15):2368-2374. DOI 10.1016/j.yexcr.2013.05.033.
- Bamburg J.R., Bernstein B.W. Actin dynamics and cofilin-actin rods in Alzheimer disease. *Cytoskeleton (Hoboken)*. 2016;73(9):477-497. DOI 10.1002/cm.21282.
- Bastos de Figueiredo J.C., Diambra L., Glass L., Malta C.P. Chaos in two-looped negative feedback systems. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 2002;65:051905.
- Basu S., Lamprecht R. The role of actin cytoskeleton in dendritic spines in the maintenance of long-term memory. *Front. Mol. Neurosci.* 2018;11:143. DOI 10.3389/fnmol.2018.00143.
- Beggs J.E., Tian S., Jones G.G., Xie J., Iadevaia V., Jenei V., Thomas G., Proud C.G. The MAP kinase-interacting kinases regulate cell migration, vimentin expression and eIF4E/CYFIP1 binding. *Biochem J.* 2015;467(1):63-76. DOI 10.1042/BJ20141066.
- Ben Zablah Y., Merovitch N., Jia Z. The role of ADF/cofilin in synaptic physiology and Alzheimer's disease. *Front. Cell Dev. Biol.* 2020; 8:594998. DOI 10.3389/fcell.2020.594998.
- Borovac J., Bosch M., Okamoto K. Regulation of actin dynamics during structural plasticity of dendritic spines: Signaling messengers and actin-binding proteins. *Mol. Cell. Neurosci.* 2018;91:122-130. DOI 10.1016/j.mcn.2018.07.001.
- Bramham C.R. Local protein synthesis, actin dynamics, and LTP consolidation. *Curr. Opin. Neurobiol.* 2008;18(5):524-531. DOI 10.1016/j.conb.2008.09.013.
- Briz V., Zhu G., Wang Y., Liu Y., Avetisyan M., Bi X., Baudry M. Activity-dependent rapid local RhoA synthesis is required for hippocampal synaptic plasticity. *J. Neurosci.* 2015;35(5):2269-2282. DOI 10.1523/JNEUROSCI.2302-14.2015.
- Brown V., Small K., Lakkis L., Feng Y., Gunter C., Wilkinson K.D., Warren S.T. Purified recombinant Fmrp exhibits selective RNA binding as an intrinsic property of the fragile X mental retardation protein. *J. Biol. Chem.* 1998;273(25):15521-15527. DOI 10.1074/jbc.273.25.15521.
- Buffington S.A., Huang W., Costa-Mattioli M. Translational control in synaptic plasticity and cognitive dysfunction. *Annu. Rev. Neurosci.* 2014;37:17-38. DOI 10.1146/annurev-neuro-071013-014100.
- Cai Z., Chen G., He W., Xiao M., Yan L.J. Activation of mTOR: a culprit of Alzheimer's disease? *Neuropsychiatr. Dis. Treat.* 2015;11: 1015-1030. DOI 10.2147/NDT.S75717.
- Cajigas I.J., Will T., Schuman E.M. Protein homeostasis and synaptic plasticity. *EMBO J.* 2010;29:2746-2752. DOI 10.1038/emboj.2010.173.
- Chen B., Chou H.T., Brautigam C.A., Xing W., Yang S., Henry L., Doolittle L.K., Walz T., Rosen M.K. Rac1 GTPase activates the WAVE regulatory complex through two distinct binding sites. *Elife*. 2017;6:e29795. DOI 10.7554/eLife.29795.
- Chen E., Joseph S. Fragile X mental retardation protein: A paradigm for translational control by RNA-binding proteins. *Biochimie*. 2015; 114:147-154. DOI 10.1016/j.biochi.2015.02.005.

- Chen E., Sharma M.R., Shi X., Agrawal R.K., Joseph S. Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Mol. Cell.* 2014;54:407-417. DOI 10.1016/j.molcel.2014.03.023.
- Chen L.Y., Rex C.S., Babayan A.H., Kramár E.A., Lynch G., Gall C.M., Lauterborn J.C. Physiological activation of synaptic Rac>PAK (p-21 activated kinase) signaling is defective in a mouse model of fragile X syndrome. *J. Neurosci.* 2010;30(33):10977-10984. DOI 10.1523/JNEUROSCI.1077-10.2010.
- Chestkov I.V., Jestkova E.M., Ershova E.S., Golimbet V.E., Lezheiko T.V., Kolesina N.Y., Porokhovnik L.N., Lyapunova N.A., Izhevskaya V.L., Kutsev S.I., Veiko N., Kostyuk S.V. Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophr. Res.* 2018;197:305-314. DOI 10.1016/j.schres.2018.01.001.
- Cory G.O.C., Ridley A.J. Cell motility: braking WAVES. *Nature.* 2002;418:732-733. DOI 10.1038/418732a.
- Costa-Mattioli M., Sossin W.S., Klann E., Sonenberg N. Translational control of long-lasting synaptic plasticity and memory. *Neuron.* 2009;61:10-26. DOI 10.1016/j.neuron.2008.10.055.
- Darnell J.C., Klann E. The translation of translational control by FMRP: therapeutic targets for FXS. *Nat. Neurosci.* 2013;16(11):1530-1536. DOI 10.1038/nn.3379.
- Darnell J.C., Van Driesche S.J., Zhang C., Hung K.Y., Mele A., Fraser C.E., Stone E.F., Chen C., Fak J.J., Chi S.W., Licatalosi D.D., Richter J.D., Darnell R.B. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell.* 2011;146:247-261. DOI 10.1016/j.cell.2011.06.013.
- Dasgupta I., McCollum D. Control of cellular responses to mechanical cues through YAP/TAZ regulation. *J. Biol. Chem.* 2019;294(46):17693-17706. DOI 10.1074/jbc.REV119.007963.
- De Rubeis S., Pasciuto E., Li K.W., Fernández E., Di Marino D., Buzzi A., Ostroff L.E., Klann E., Zwartkruis F.J., Komiyama N.H., Grant S.G., Poujol C., Choquet D., Achsel T., Posthuma D., Smit A.B., Bagni C. CYFIP1 coordinates mRNA translation and cytoskeleton remodeling to ensure proper dendritic spine formation. *Neuron.* 2013;79(6):1169-1182. DOI 10.1016/j.neuron.2013.06.039.
- de Sena Cortabitarte A., Degenhardt F., Strohmaier J., Lang M., Weiss B., Roeth R., Giegling I., Heilmann-Heimbach S., Hofmann A., Rujescu D., Fischer C., Rietschel M., Nöthen M.M., Rapaport G.A., Berkel S. Investigation of SHANK3 in schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 2017;174(4):390-398. DOI 10.1002/ajmg.b.32528.
- Decroly O., Goldbeter A. Bihybricity, chaos, and other patterns of temporal self-organization in a multiply regulated biochemical system. *Proc. Natl. Acad. Sci. USA.* 1982;79:6917-6921. DOI 10.1073/pnas.79.22.6917.
- Derivery E., Lombard B., Loew D., Gautreau A. The Wave complex is intrinsically inactive. *Cell Motil. Cytoskeleton.* 2009;66(10):777-790. DOI 10.1002/cm.20342.
- Dupont S., Morsut L., Aragona M., Enzo E., Giulitti S., Cordenonsi M., Zanconato F., Le Dıgabel J., Forcato M., Bicciato S., Elvassore N., Piccolo S. Role of YAP/TAZ in mechanotransduction. *Nature.* 2011;474(7350):179-183. DOI 10.1038/nature10137.
- Feng Y., Absher D., Eberhart D.E., Brown V., Malter H.E., Warren S.T. FMRP associates with polyribosomes as an mRNP, and the 1304N mutation of severe fragile X syndrome abolishes this association. *Mol. Cell.* 1997;1(1):109-118. DOI 10.1016/s1097-2765(00)80012-x.
- Forrest M.P., Parnell E., Penzes P. Dendritic structural plasticity and neuropsychiatric disease. *Nat. Rev. Neurosci.* 2018;19(4):215-234. DOI 10.1038/nrn.2018.16.
- Fu Y.J., Liu D., Guo J.L., Long H.Y., Xiao W.B., Xiao W., Feng L., Luo Z.H., Xiao B. Dynamic change of shank gene mRNA expression and DNA methylation in epileptic rat model and human patients. *Mol. Neurobiol.* 2020;57(9):3712-3726. DOI 10.1007/s12035-020-01968-5.
- Gkogkas C.G., Sonenberg N. Translational control and autism-like behaviors. *Cell. Logist.* 2013;3:e24551. DOI 10.4161/cl.24551.
- Gobert D., Schohl A., Kutsarova E., Ruthazer E.S. TORC1 selectively regulates synaptic maturation and input convergence in the developing visual system. *Dev. Neurobiol.* 2020. DOI 10.1002/dneu.22782.
- Goldbeter A., Gonze D., Houart G., Leloup J.C., Halloy J., Dupont G. From simple to complex oscillatory behavior in metabolic and genetic control networks. *Chaos.* 2001;11:247-260. DOI 10.1063/1.1345727.
- Gross C., Nakamoto M., Yao X., Chan C.B., Yim S.Y., Ye K., Warren S.T., Bassell G.J. Excess phosphoinositide 3-kinase subunit synthesis and activity as a novel therapeutic target in fragile X syndrome. *J. Neurosci.* 2010;30:10624-10638. DOI 10.1523/JNEUROSCI.0402-10.2010.
- Hong L., Li Y., Liu Q., Chen Q., Chen L., Zhou D. The Hippo signaling pathway in regenerative medicine. *Methods Mol. Biol.* 2019;1893:353-370. DOI 10.1007/978-1-4939-8910-2_26.
- Hu J.K., Du W., Shelton S.J., Oldham M.C., DiPersio C.M., Klein O.D. An FAK-YAP-mTOR signaling axis regulates stem cell-based tissue renewal in mice. *Cell Stem Cell.* 2017;21(1):91-106. DOI 10.1016/j.stem.2017.03.023.
- Huber K.M., Kayser M.S., Bear M.F. Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science.* 2000;288(5469):1254-1257. DOI 10.1126/science.288.5469.1254.
- Huber K.M., Klann E., Costa-Mattioli M., Zukin R.S. Dysregulation of mammalian target of rapamycin signaling in mouse models of autism. *J. Neurosci.* 2015;35(41):13836-13842. DOI 10.1523/JNEUROSCI.2656-15.2015.
- Iacoangeli A., Tiedge H. Translational control at the synapse: role of RNA regulators. *Trends Biochem. Sci.* 2013;38(1):47-55. DOI 10.1016/j.tibs.2012.11.001.
- Ip C.K., Cheung A.N., Ngan H.Y., Wong A.S. p70 S6 kinase in the control of actin cytoskeleton dynamics and directed migration of ovarian cancer cells. *Oncogene.* 2011;30(21):2420-2432. DOI 10.1038/onc.2010.615.
- Johnson S.C., Rabinovitch P.S., Kaeberlein M. mTOR is a key modulator of ageing and age-related disease. *Nature.* 2013;493(7432):338-345. DOI 10.1038/nature11861.
- Khlebodarova T.M., Kogai V.V., Fadeev S.I., Likhoshvai V.A. Chaos and hyperchaos in simple gene network with negative feedback and time delays. *J. Bioinform. Comput. Biol.* 2017;15(2):1650042. DOI 10.1142/S0219720016500426.
- Khlebodarova T.M., Kogai V.V., Likhoshvai V.A. On the dynamical aspects of local translation at the activated synapse. *BMC Bioinformatics.* 2020;21(Suppl. 11):258. DOI 10.1186/s12859-020-03597-0.
- Khlebodarova T.M., Kogai V.V., Trifonova E.A., Likhoshvai V.A. Dynamic landscape of the local translation at activated synapses. *Mol. Psych.* 2018;23(1):107-114. DOI 10.1038/mp.2017.245.
- Klein M.E., Monday H., Jordan B.A. Proteostasis and RNA binding proteins in synaptic plasticity and in the pathogenesis of neuropsychiatric disorders. *Neural Plast.* 2016;2016:3857934. DOI 10.1155/2016/3857934.
- Kogai V.V., Likhoshvai V.A., Fadeev S.I., Khlebodarova T.M. Multiple scenarios of transition to chaos in the alternative splicing model. *Int. J. Bifurcat. Chaos.* 2017;27(2):1730006. DOI 10.1142/S0218127417300063.
- Lauterborn J.C., Cox C.D., Chan S.W., Vanderklish P.W., Lynch G., Gall C.M. Synaptic actin stabilization protein loss in Down syndrome and Alzheimer disease. *Brain Pathol.* 2020;30(2):319-331. DOI 10.1111/bpa.12779.
- Levy Nogueira M., da Veiga Moreira J., Baronzio G.F., Dubois B., Steyaert J.M., Schwartz L. Mechanical stress as the common denominator between chronic inflammation, cancer, and Alzheimer's disease. *Front. Oncol.* 2015;5:197. DOI 10.3389/fonc.2015.00197.
- Levy Nogueira M., Epelbaum S., Steyaert J.M., Dubois B., Schwartz L. Mechanical stress models of Alzheimer's disease pathology. *Alzheimers Dement.* 2016a;12(3):324-333. DOI 10.1016/j.jalz.2015.10.005.

- Levy Nogueira M., Hamraz M., Abolhassani M., Bigan E., Lafitte O., Steyaert J.M., Dubois B., Schwartz L. Mechanical stress increases brain amyloid β , tau, and α -synuclein concentrations in wild-type mice. *Alzheimers Dement.* 2018;14(4):444-453. DOI 10.1016/j.jalz.2017.11.003.
- Levy Nogueira M., Lafitte O., Steyaert J.M., Bakardjian H., Dubois B., Hampel H., Schwartz L. Mechanical stress related to brain atrophy in Alzheimer's disease. *Alzheimers Dement.* 2016b;12(1):11-20. DOI 10.1016/j.jalz.2015.03.005.
- Likhoshvai V.A., Fadeev S.I., Kogai V.V., Khlebodarova T.M. On the chaos in gene networks. *J. Bioinform. Comput. Biol.* 2013;11(1):1340009. DOI 10.1142/S021972001340009X.
- Likhoshvai V.A., Golubyatnikov V.P., Khlebodarova T.M. Limit cycles in models of circular gene networks regulated by negative feedbacks. *BMC Bioinformatics.* 2020;21(Suppl. 11):255. DOI 10.1186/s12859-020-03598-z.
- Likhoshvai V.A., Kogai V.V., Fadeev S.I., Khlebodarova T.M. Alternative splicing can lead to chaos. *J. Bioinform. Comput. Biol.* 2015;13(1):1540003. DOI 10.1142/S021972001540003X.
- Likhoshvai V.A., Kogai V.V., Fadeev S.I., Khlebodarova T.M. Chaos and hyperchaos in a model of ribosome autocatalytic synthesis. *Sci. Rep.* 2016;6:38870. DOI 10.1038/srep38870.
- Louros S.R., Osterweil E.K. Perturbed proteostasis in autism spectrum disorders. *J. Neurochem.* 2016;139(6):1081-1092. DOI 10.1111/jnc.13723.
- Mackey M.C., Glass L. Oscillation and chaos in physiological control systems. *Science.* 1977;197(4300):287-289. DOI 10.1126/science.267326.
- Majumder P., Chu J.F., Chatterjee B., Swamy K.B., Shen C.J. Co-regulation of mRNA translation by TDP-43 and fragile X syndrome protein FMRP. *Acta Neuropathol.* 2016;132(5):721-738. DOI 10.1007/s00401-016-1603-8.
- Martin I. Decoding Parkinson's disease pathogenesis: the role of deregulated mRNA translation. *J. Parkinsons Dis.* 2016;6(1):17-27. DOI 10.3233/JPD-150738.
- McCarthy N. Signalling: YAP, PTEN and miR-29 size each other up. *Nat. Rev. Cancer.* 2013;13(1):4-5. DOI 10.1038/nrc3422.
- Mei Y., Monteiro P., Zhou Y., Kim J.A., Gao X., Fu Z., Feng G. Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. *Nature.* 2016;530(7591):481-484. DOI 10.1038/nature16971.
- Meng X.F., Yu J.T., Song J.H., Chi S., Tan L. Role of the mTOR signaling pathway in epilepsy. *J. Neurol. Sci.* 2013;332(1-2):4-15. DOI 10.1016/j.jns.2013.05.029.
- Millard T.H., Sharp S.J., Machesky L.M. Signalling to actin assembly via the WASP (Wiskott-Aldrich syndrome protein)-family proteins and the Arp2/3 complex. *Biochem. J.* 2004;380(Pt. 1):1-17. DOI 10.1042/BJ20040176.
- Molinie N., Gautreau A. The Arp2/3 regulatory system and its deregulation in cancer. *Physiol. Rev.* 2018;98(1):215-238. DOI 10.1152/physrev.00006.2017.
- Monteiro P., Feng G. SHANK proteins: roles at the synapse and in autism spectrum disorder. *Nat. Rev. Neurosci.* 2017;18(3):147-157. DOI 10.1038/nrn.2016.183.
- Muddashetty R.S., Kelić S., Gross C., Xu M., Bassell G.J. Dysregulated metabotropic glutamate receptor-dependent translation of AMPA receptor and postsynaptic density-95 mRNAs at synapses in a mouse model of fragile X syndrome. *J. Neurosci.* 2007;27(20):5338-5348. DOI 10.1523/JNEUROSCI.0937-07.2007.
- Nakashima M., Kato M., Aoto K., Shiina M., Belal H., Mukaida S., Kumada S., Sato A., Zerem A., Lerman-Sagie T., Lev D., Leong H.Y., Tsurusaki Y., Mizuguchi T., Miyatake S., Miyake N., Ogata K., Saito H., Matsumoto N. De novo hotspot variants in CYFIP2 cause early-onset epileptic encephalopathy. *Ann. Neurol.* 2018;83(4):794-806. DOI 10.1002/ana.25208.
- Nalavadi V.C., Muddashetty R.S., Gross C., Bassell G.J. Dephosphorylation-induced ubiquitination and degradation of FMRP in dendrites: a role in immediate early mGluR-stimulated translation. *J. Neurosci.* 2012;32(8):2582-2587. DOI 10.1523/JNEUROSCI.5057-11.2012.
- Napoli I., Mercaldo V., Boyl P.P., Eleuteri B., Zalfa F., De Rubeis S., Di Marino D., Mohr E., Massimi M., Falconi M., Witke W., Costa-Mattioli M., Sonenberg N., Achsel T., Bagni C. The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell.* 2008;134(6):1042-1054. DOI 10.1016/j.cell.2008.07.031.
- Narayanan U., Nalavadi V., Nakamoto M., Pallas D.C., Ceman S., Bassell G.J., Warren S.T. FMRP phosphorylation reveals an immediate-early signaling pathway triggered by group I mGluR and mediated by PP2A. *J. Neurosci.* 2007;27(52):14349-14357. DOI 10.1523/JNEUROSCI.2969-07.2007.
- Narayanan U., Nalavadi V., Nakamoto M., Thomas G., Ceman S., Bassell G.J., Warren S.T. S6K1 phosphorylates and regulates fragile X mental retardation protein (FMRP) with the neuronal protein synthesis-dependent mammalian target of rapamycin (mTOR) signaling cascade. *J. Biol. Chem.* 2008;283(27):18478-18482. DOI 10.1074/jbc.C800055200.
- Nishiyama J. Plasticity of dendritic spines: Molecular function and dysfunction in neurodevelopmental disorders. *Psychiat. Clin. Neurosci.* 2019;73(9):541-550. DOI 10.1111/pcn.12899.
- Onore C., Yang H., Van de Water J., Ashwood P. Dynamic Akt/mTOR signaling in children with autism spectrum disorder. *Front. Pediatr.* 2017;5:43. DOI 10.3389/fped.2017.00043.
- Panciera T., Azzolin L., Cordenonsi M., Piccolo S. Mechanobiology of YAP and TAZ in physiology and disease. *Nat. Rev. Mol. Cell Biol.* 2017;18(12):758-770. DOI 10.1038/nrm.2017.87.
- Peça J., Feliciano C., Ting J.T., Wang W., Wells M.F., Venkatraman T.N., Lascola C.D., Fu Z., Feng G. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature.* 2011;472(7344):437-442. DOI 10.1038/nature09965.
- Pei J.J., Hugon J. mTOR-dependent signalling in Alzheimer's disease. *J. Cell. Mol. Med.* 2008;12(6b):2525-2532. DOI 10.1111/j.1582-4934.2008.00509.x.
- Penzes P., Rafalovich I. Regulation of the actin cytoskeleton in dendritic spines. *Adv. Exp. Med. Biol.* 2012;970:81-95. DOI 10.1007/978-3-7091-0932-8_4.
- Porokhovnik L. Individual copy number of ribosomal genes as a factor of mental retardation and autism risk and severity. *Cells.* 2019;8(10):1151. DOI 10.3390/cells8101151.
- Porokhovnik L.N., Lyapunova N.A. Dosage effects of human ribosomal genes (rDNA) in health and disease. *Chromosome Res.* 2019;27(1-2):5-17. DOI 10.1007/s10577-018-9587-y.
- Pramparo T., Pierce K., Lombardo M.V., Carter Barnes C., Marinero S., Ahrens-Barbeau C., Murray S.S., Lopez L., Xu R., Courchesne E. Prediction of autism by translation and immune/inflammation co-expressed genes in toddlers from pediatric community practice. *JAMA Psychiatry.* 2015;72:386-394. DOI 10.1001/jamapsychiatry.2014.3008.
- Pyronneau A., He Q., Hwang J.Y., Porch M., Contractor A., Zukin R.S. Aberrant Rac1-cofilin signaling mediates defects in dendritic spines, synaptic function, and sensory perception in fragile X syndrome. *Sci. Signal.* 2017;10(504):eaan0852. DOI 10.1126/scisignal.aan0852.
- Reddy P., Deguchi M., Cheng Y., Hsueh A.J. Actin cytoskeleton regulates Hippo signaling. *PLoS One.* 2013;8(9):e73763. DOI 10.1371/journal.pone.0073763.
- Rex C.S., Chen L.Y., Sharma A., Liu J., Babayan A.H., Gall C.M., Lynch G. Different Rho GTPase-dependent signaling pathways initiate sequential steps in the consolidation of long-term potentiation. *J. Cell Biol.* 2009;186(1):85-97. DOI 10.1083/jcb.200901084.
- Rosenberg T., Gal-Ben-Ari S., Dieterich D.C., Kreutz M.R., Ziv N.E., Gundelfinger E.D., Rosenblum K. The roles of protein expression in synaptic plasticity and memory consolidation. *Front. Mol. Neurosci.* 2014;7:86. DOI 10.3389/fnmol.2014.00086.
- Santini E., Huynh T.N., Klann E. Mechanisms of translation control underlying long-lasting synaptic plasticity and the consolidation of long-term memory. *Prog. Mol. Biol. Transl. Sci.* 2014;122:131-167. DOI 10.1016/B978-0-12-420170-5.00005-2.

- Schaks M., Reinke M., Witke W., Rottner K. Molecular dissection of neurodevelopmental disorder-causing mutations in CYFIP2. *Cells*. 2020;9(6):1355. DOI 10.3390/cells9061355.
- Schaks M., Singh S.P., Kage F., Thomason P., Klünemann T., Steffen A., Blankenfeldt W., Stradal T.E., Insall R.H., Rottner K. Distinct interaction sites of RAC GTPase with WAVE regulatory complex have non-redundant functions *in vivo*. *Curr. Biol.* 2018;28(22):3674-3684.e6. DOI 10.1016/j.cub.2018.10.002.
- Seo J., Kim J. Regulation of Hippo signaling by actin remodeling. *BMB Rep.* 2018;51(3):151-156. DOI 10.5483/bmbrep.2018.51.3.012.
- Sharma A., Hoeffler C.A., Takayasu Y., Miyawaki T., McBride S.M., Klann E., Zukin R.S. Dysregulation of mTOR signaling in fragile X syndrome. *J. Neurosci.* 2010;30(2):694-702. DOI 10.1523/JNEUROSCI.3696-09.2010.
- Suzuki Y., Lu M., Ben-Jacob E., Onuchic J.N. Periodic, quasi-periodic and chaotic dynamics in simple gene elements with time delays. *Sci. Rep.* 2016;6:21037. DOI 10.1038/srep21037.
- Tapon N., Hall A. Rho, Rac and Cdc42 GTPases regulate the organization of the actin cytoskeleton. *Curr. Opin. Cell Biol.* 1997;9(1):86-92. DOI 10.1016/s0955-0674(97)80156-1.
- Totaro A., Panciera T., Piccolo S. YAP/TAZ upstream signals and downstream responses. *Nat. Cell Biol.* 2018;20(8):888-899. DOI 10.1038/s41556-018-0142-z.
- Trifonova E.A., Khlebodarova T.M., Gruntenko N.E. Molecular mechanisms of autism as a form of synaptic dysfunction. *Russ. J. Genet.: Appl. Res.* 2017;7(8):869-877.
- Troca-Marin J.A., Alves-Sampaio A., Montesinos M.L. Deregulated mTOR-mediated translation in intellectual disability. *Prog. Neurobiol.* 2012;96(2):268-282. DOI 10.1016/j.pneurobio.2012.01.005.
- Tumaneng K., Schlegelmilch K., Russell R.C., Yimlamai D., Basnet H., Mahadevan N., Fitamant J., Bardeesy N., Camargo F.D., Guan K.L. YAP mediates crosstalk between the Hippo and PI(3)K-TOR pathways by suppressing PTEN via miR-29. *Nat. Cell Biol.* 2012;14(12):1322-1329. DOI 10.1038/ncb2615.
- Won H., Mah W., Kim E. Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. *Front. Mol. Neurosci.* 2013;6:19. DOI 10.3389/fnmol.2013.00019.
- Wong M. Mammalian target of rapamycin (mTOR) inhibition as a potential antiepileptogenic therapy: From tuberous sclerosis to common acquired epilepsies. *Epilepsia*. 2010;51(1):27-36. DOI 10.1111/j.1528-1167.2009.02341.x.
- Wostyn P. Intracranial pressure and Alzheimer's disease: a hypothesis. *Med. Hypotheses*. 1994;43(4):219-222. DOI 10.1016/0306-9877(94)90069-8.
- Yu F.X., Zhao B., Guan K.L. Hippo pathway in organ size control, tissue homeostasis, and cancer. *Cell*. 2015;163(4):811-828. DOI 10.1016/j.cell.2015.10.044.
- Zamboni V., Jones R., Umbach A., Ammoni A., Passafaro M., Hirsch E., Merlo G.R. Rho GTPases in intellectual disability: from genetics to therapeutic opportunities. *Int. J. Mol. Sci.* 2018;19:1821. DOI 10.3390/ijms19061821.
- Zhang Y., Lee Y., Han K. Neuronal function and dysfunction of CYFIP2: from actin dynamics to early infantile epileptic encephalopathy. *BMB Rep.* 2019;52(5):304-311. DOI 10.5483/BMBRep.2019.52.5.097.
- Zhou J., Parada L.F. PTEN signaling in autism spectrum disorders. *Curr. Opin. Neurobiol.* 2012;22(5):873-879. DOI 10.1016/j.conb.2012.05.004.
- Zhu T., Ma Z., Wang H., Jia X., Wu Y., Fu L., Li Z., Zhang C., Yu G. YAP/TAZ affects the development of pulmonary fibrosis by regulating multiple signaling pathways. *Mol. Cell. Biochem.* 2020;475(1-2):137-149. DOI 10.1007/s11010-020-03866-9.
- Zukin R.S., Richter J.D., Bagni C. Signals, synapses, and synthesis: how new proteins control plasticity. *Front. Neural Circuits*. 2009;3:14. DOI 10.3389/neuro.04.014.2009.
- Zweier M., Begemann A., McWalter K., Cho M.T., Abela L., Banka S., Behring B., Berger A., Brown C.W., Carneiro M., Chen J., Cooper G.M. Deciphering Developmental Disorders (DDD) Study, Finnila C.R., Guillen Sacoto M.J., Henderson A., Hüffmeier U., Jøset P., Kerr B., Lesca G., Leszinski G.S., McDermott J.H., Meltzer M.R., Monaghan K.G., Mostafavi R., Öunap K., Plecko B., Powis Z., Purcarin G., Reimand T., Riedhammer K.M., Schreiber J.M., Sirsi D., Wierenga K.J., Wojcik M.H., Papuc S.M., Steindl K., Sticht H., Rauch A. Spatially clustering *de novo* variants in CYFIP2, encoding the cytoplasmic FMRP interacting protein 2, cause intellectual disability and seizures. *Eur. J. Hum. Genet.* 2019;27(5):747-759. DOI 10.1038/s41431-018-0331-z.

Благодарности. Работа выполнена при поддержке бюджетного проекта № 0259-2021-0009.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию 19.10.2020. После доработки 21.12.2020. Принята к публикации 22.12.2020.

Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes

K.V. Ustyantsev, E.V. Berezikov

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

Abstract. In eukaryotes, trans-splicing is a process of nuclear pre-mRNA maturation where two different RNA molecules are joined together by the spliceosomal machinery utilizing mechanisms similar to cis-splicing. In diverse taxa of lower eukaryotes, spliced leader (SL) trans-splicing is the most frequent type of trans-splicing, when the same sequence derived from short small nuclear RNA molecules, called SL RNAs, is attached to the 5' ends of different non-processed pre-mRNAs. One of the functions of SL trans-splicing is processing polycistronic pre-mRNA molecules transcribed from operons, when several genes are transcribed as one pre-mRNA molecule. However, only a fraction of trans-spliced genes reside in operons, suggesting that SL trans-splicing must also have some other, less understood functions. Regenerative flatworms are informative model organisms which hold the keys to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is fueled by the division and differentiation of the adult somatic stem cell population called neoblasts. *Macrostomum lignano* is a flatworm model organism where substantial technological advances have been achieved in recent years, including the development of transgenesis. Although a large fraction of genes in *M. lignano* were estimated to be SL trans-spliced, SL trans-splicing was not studied in detail in *M. lignano* before. Here, we performed the first comprehensive study of SL trans-splicing in *M. lignano*. By reanalyzing the existing genome and transcriptome data of *M. lignano*, we estimate that 30 % of its genes are SL trans-spliced, 15 % are organized in operons, and almost 40 % are both SL trans-spliced and in operons. We annotated and characterized the sequence of SL RNA and characterized conserved cis- and SL trans-splicing motifs. Finally, we found that a majority of SL trans-spliced genes are evolutionarily conserved and significantly over-represented in neoblast-specific genes. Our findings suggest an important role of SL trans-splicing in the regulation and maintenance of neoblasts in *M. lignano*.

Key words: flatworms; regeneration; splicing; trans-splicing; neoblasts; spliced leader; *Macrostomum lignano*.

For citation: Ustyantsev K.V., Berezikov E.V. Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2021;25(1):101-107. DOI 10.18699/VJ21.012

Биоинформационный анализ сплайс-лидерного транс-сплайсинга у регенерирующего плоского червя *Macrostomum lignano* показал его преобладание среди консервативных генов и генов стволовых клеток

К.В. Устьянцев, Е.В. Березиков

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

Аннотация. Транс-сплайсинг у эукариот – это процесс созревания ядерных пре-мРНК, когда две различные молекулы РНК соединяются с помощью структур сплайсосомы по механизму, схожему с цис-сплайсингом. У различных таксонов низших эукариот наиболее распространенный тип транс-сплайсинга – сплайс-лидерный (СЛ) транс-сплайсинг, при котором одинаковая последовательность, происходящая от коротких малых ядерных РНК молекул, называемых СЛ РНК, присоединяется к 5'-концам различных непроцессированных пре-мРНК. Одна из функций СЛ транс-сплайсинга состоит в процессировании полицистронных молекул пре-мРНК, транскрибируемых с оперонов, когда транскрипция нескольких генов осуществляется как одна молекула пре-мРНК. Однако лишь часть генов, подвергающихся транс-сплайсингу, содержится в оперонах, что говорит о том, что у СЛ транс-сплайсинга должны быть и другие, менее изученные, функции. Регенерирующие плоские черви являются информативными модельными организмами, хранящими ключи к пониманию механизмов регуляции стволовых клеток и их дифференцировки во время регенерации и при гомеостазе. Их способность к регенерации – следствие деления и дифференцировки соматических стволовых клеток, называемых необластами, которые присутствуют у взрослых особей. *Macrostomum lignano* – модельный плоский червь, в исследованиях на котором в

последние годы достигнут существенный технологический прогресс, включая разработку метода трансгенеза. Сплайс-лидерный транс-сплайсинг ранее не был детально изучен у *M. lignano*, хотя известно, что значительная часть генов *M. lignano* подвергается этому типу транс-сплайсинга. В настоящей работе мы осуществили первое обширное исследование СЛ транс-сплайсинга у *M. lignano*. Повторно проанализировав геномные и транскриптомные данные *M. lignano*, мы оцениваем, что 30 % его генов подвергаются СЛ транс-сплайсингу, 15 % расположены в оперонах, а почти 40 % находятся в оперонах и проходят через СЛ транс-сплайсинг. Мы провели аннотацию и охарактеризовали последовательность СЛ РНК и консервативных мотивов цис- и транс-сплайсинга. Обнаружено, что большинство генов, подвергающихся СЛ транс-сплайсингу, эволюционно консервативны и значительно перепредставлены в генах, специфичных для необластов. Наши результаты предполагают важную роль СЛ транс-сплайсинга в регуляции функционирования необластов у *M. lignano*.

Ключевые слова: плоские черви; регенерация; сплайсинг; транс-сплайсинг; необласты; сплайс-лидер; *Macrostomum lignano*.

Introduction

Before being used as templates for protein production, majority of RNA molecules transcribed in the nucleus (pre-mRNA) undergo three major modifications to become mature and fully functional mRNA. This is called RNA processing and it involves capping of the 5' end, polyadenylation of the 3' end, and splicing. Two types of splicing are distinguished – cis- and trans-splicing. During cis-splicing all the processing happens with the same pre-mRNA molecule, resulting in the removal of introns and merging of its exons. During trans-splicing, on the other hand, two different pre-mRNA molecules expressed from distinct genomic loci are joint into a new chimeric trans-spliced mRNA (Lasda, Blumenthal, 2011).

Trans-splicing was originally discovered in trypanosomes (Euglenozoa), where it was found that a short 39 bp leader sequence was post-transcriptionally attached to the 5' ends of variant surface glycoproteins pre-mRNA (Boothroyd, Cross, 1982). Later, 5' end addition of a 22 bp spliced leader (SL) was also observed in *Caenorhabditis elegans* mRNA of actin gene and some other genes (Krause, Hirsh, 1987). Now this process is well known as SL trans-splicing. A distinct feature of SL trans-splicing is that all such processed transcripts have the same short SL sequence, or its variant, at their 5' ends. The SL sequence is derived from an exon of a non-coding small nuclear RNA molecule called SL RNA, which is ~100 nt in length and has 2,2,7-trimethylguanosine cap at its 5' end instead of 7-methylguanosine cap, which is found in non-spliced mRNAs (Liou, Blumenthal, 1990; Lasda, Blumenthal, 2011). SL RNAs have a splicing donor site at the exon 3' end, while SL trans-spliced pre-mRNAs have a splicing acceptor site at the 5' end of their first exon. SL trans-splicing results in removal of the 5' non-exon pre-mRNA part called outtron (Lasda, Blumenthal, 2011). It is experimentally shown that the only requirement for a gene to be predominantly SL trans-spliced is an acceptor splicing site close to the 5' end of the first exon that is not complemented by a donor splicing site upstream in cis (Conrad et al., 1993). Thus comes another important feature of SL trans-splicing, namely that it allows formation and resolving of operons – adjacent genes transcribed as a single pre-mRNA from the same promoter region (Blumenthal, Gleason, 2003). However, apart from a clear function in polycistronic transcripts resolution, the function of SL trans-splicing for monocistronic transcripts is still in debate (Danks et al., 2015). It is hypothesized that the function may be in equalization of 5' UTRs in length and their clearance from out-of-frame AUG start codons, while at the same time allowing less restricted evolution of 5' upstream regulatory sequences, and in additional control of translation

(Hastings, 2005; Danks, Thompson, 2015). So far, SL trans-splicing was found in several clades of eukaryotes: dinoflagellates, euglenozoans, cnidarians, flatworms, nematodes, and ascidians (Lei et al., 2016). SL trans-splicing is most prominent in trypanosomes (100 % genes are trans-spliced) and in nematodes (70 % genes are trans-spliced) (Allen et al., 2011; Lei et al., 2016).

Regenerative flatworms are informative models to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is driven by the division and differentiation of the adult somatic stem cell population called neoblasts (Wagner et al., 2011; Mouton et al., 2018). *Macrostomum lignano* is the only flatworm species for which a method for stable transgenesis is available so far. The worm also has a number of features allowing for efficient cell lineage tracing and phenotype screening, which makes *M. lignano* an attractive model to study a wide range of biological processes (Grudniewska et al., 2016; Wudarski et al., 2017, 2019, 2020). Well-annotated *M. lignano* genome and transcriptome assemblies were recently published (Wudarski et al., 2017; Grudniewska et al., 2018). It was estimated that almost 21 % of its genes are SL trans-spliced to the same 35 bp SL sequence (Grudniewska et al., 2018). However, trans-splicing was not studied in details in *M. lignano*, and its impact on the genome functioning and maintenance is still unknown. Here, we present the first comprehensive study of SL trans-splicing in *M. lignano* and show that it is strongly connected with genes specific for the neoblasts of the worm.

Materials and methods

Data. The published *M. lignano* genome Mlig_3_7 (Wudarski et al., 2017) and transcriptome Mlig_RNA_3_7_DV1_v3 (Grudniewska et al., 2018) assemblies and the corresponding annotation tracks were obtained from (http://gb.macgenome.org/downloads/Mlig_3_7/).

Genome deduplication. Mlig_3_7 genome assembly was deduplicated using `purge_dups` software (v1.0.1) with default settings (Guan et al., 2020) and utilizing published PacBio genome sequencing data (Wasik et al., 2015) for the calculation of contig coverages. Contig names from the deduplicated genome assembly were used to extract respective gene annotations from the full Mlig_3_7 genome annotation.

Motif discovery and SL RNA annotation. Presence of the SL sequence at the 5' end of the *M. lignano* transcripts was established in the previous studies (Wasik et al., 2015; Grudniewska et al., 2016). For the annotation of trans-spliced genes, SL-containing RNA-seq reads were mapped to the

Mlig_3_7 genome assembly and the presence of such reads at the beginning of transcripts was used as an evidence of SL trans-splicing (Wudarski et al., 2017; Grudniewska et al., 2018). Therefore, all the SL trans-spliced transcripts have the corresponding annotation in the Mlig_3_7 genome assembly, and the sequences upstream of their first exon were considered as outons. Using the deduplicated annotation track of gene coordinates, we retrieved nucleotide sequences of genomic regions corresponding to exon-intron and exon-outron (for the trans-spliced genes) junctions with 50 bp flanks in both directions. All the sequences were converted to forward orientation and split into three groups corresponding to cis-donor, cis-acceptor, and trans-spliced acceptor sites. The sequences then were analysed for the presence of enriched motif using a stand-alone version of the DREME tool (Bailey, 2011).

To determine the SL RNA gene sequence in the genome assembly, we used the 35 bp *M. lignano* SL sequence (CGG TCTTCTACTGCGAAGACTCAATTTA TTGCATG) as a seed for a BLASTn (Altschul et al., 1990) search requiring only 100 % matching hits. Next, we manually investigated genomic sequences surrounding the BLAST hits by matching the SL sequence track in the genome browser to the expected size of SL RNA (~100 bp). The corresponding sequences were then checked for folding into secondary structure canonical for SL RNA folding using Mfold web server (Zuker, 2003), and conserved motifs were then manually identified.

Prediction of operons. Intergenic distances were retrieved from the deduplicated genome annotation file. We only considered distances between immediately adjacent transcripts with the same transcriptional orientation and not interrupted by transcripts in opposite orientation. Distances were split into three categories: between SL trans-spliced genes, between a non-SL trans-spliced gene and an SL trans-spliced gene, and between non-SL trans-spliced genes. To adjust for repetitive element insertions, we retrieved the corresponding coordinates from the genome browser RepeatMasker and TRF tracks (<http://gb.macgenome.org/>) and subtracted them from the previously identified intergenic distances. Distribution of the distances was visualized as density plots using ggplot2 library in R.

After the analysis of the graphical data of the distances distributions, we selected the threshold value of 1000 bp, below which a pair of adjacent and SL trans-spliced genes were considered as belonging to the same operon. The same applies if the first gene is non-trans-spliced, but the second is SL trans-spliced. The distributions of lengths of operons of various sizes was visualized as violin plots using ggplot2 library in R.

Estimation of gene conservation. Gene annotation and data classifying genes as being specific to neoblasts or germline were retrieved from the previous study (Grudniewska et al., 2018). A gene was considered to be conserved if it has an open reading frame with a detectable homology to a human gene, which is indicated in its annotation, and non-conserved if lacking the homology to human, but has a predicted open reading frame with homology to proteins in other organisms. Otherwise, a gene was considered non-coding.

Results

Deduplication of genome assembly. The published Mlig_3_7 genome assembly is based on the sequencing data from DV1

M. lignano line. This line has a $2n = 10$ karyotype (four large and six small chromosomes) and was demonstrated to have undergone a duplication of its large chromosome (Zadesenets et al., 2017), while the karyotype of the basal wild type population is $2n = 8$ (two large and six small chromosomes) (Wudarski et al., 2017). The size of Mlig_3_7 assembly is 764 Mb, which corresponds to the experimental measurement of the genome size in the DV1 line (Wudarski et al., 2017), and the assembly contains the duplicated large chromosome sequences. To avoid gene overcounting due to the presence of these duplicated sequences in the Mlig_3_7 assembly, we removed the most redundant scaffolds by deduplicating Mlig_3_7 assembly using purge_dups software (Guan et al., 2020). This resulted in approximately 46 % drop in the number of scaffolds (from 5270 to 2841) and decreased the genome size to 580 Mb, which is close to the genome size measurements for the NL10 line of *M. lignano*, which does not have the chromosomal duplication (Wudarski et al., 2017). Next, we removed the records from transcriptome annotation which corresponded to the redundant scaffolds.

Motif discovery and SL RNA gene mapping. Investigation of the deduplicated part of the transcriptome shows that a significant fraction of genes, 21 754 out of 71 499 (30 %), are SL trans-spliced in *M. lignano*. This means that they all have the same 35 bp SL sequence (CGGTCTTCTACTGCGAAGACTCAATTTA TTGCATG) at the 5' end of their processed transcripts (Wudarski et al., 2017; Grudniewska et al., 2018). Despite this, SL trans-splicing was not characterized in more detail in *M. lignano*. First, we retrieved genomic DNA sequences near the cis-splicing and SL trans-splicing exon-intron/exon-outron junction sites and checked if they are enriched for some motifs using DREME (Fig. 1, a) (Bailey, 2011). In total, we obtained 187 627 regions around 5' donor and 3' acceptor cis-splicing sites and 21 754 regions around SL trans-splicing sites. The first most enriched motifs near cis-splicing 5' donor and 3' acceptor sites were GT[G/A]AG (found in 122 399 regions, p -value: $8.8e^{-23468}$) and CAG (found in 112 174 regions, p -value: $1.7e^{-12459}$), respectively, corresponding to canonical cis-splicing motifs. A motif [T/C]TNCAG (found in 9551 regions, p -value: $1.3e^{-1631}$) was the top enriched motif near SL trans-splicing 3' acceptor sites. All the motifs were positioned right at the exon-intron/exon-outron junctions of the corresponding sites (see Fig. 1, a).

Next, to confirm the presence of the SL RNA gene in the genome assembly, we analysed the secondary structure of the previously published sequence of *M. lignano* SL RNA from the ML2 version of the genome (Wasik et al., 2015). However, we found that the reported sequence was erroneously assigned as SL RNA, since it clearly maps to the 5' end of an SL trans-spliced protein-coding gene (Mlig013257.g1, scaf577:45663-48770) in the Mlig_3_7 assembly, and also does not fold into canonical structure with three hairpin loops (data not shown) (Xie, Hirsh, 1998). Therefore, we decided to identify the actual SL RNA gene in the newer Mlig_3_7 assembly. Using SL sequence as a seed for the genomic BLASTn search, we mapped a 109 bp sequence, which is repeated eight times in the deduplicated genome and has the canonical SL RNA secondary structure predicted by Mfold web server (see Fig. 1, b) (Zuker, 2003). Subsequent sequence analysis showed clear signatures of an SL RNA: the SL sequence

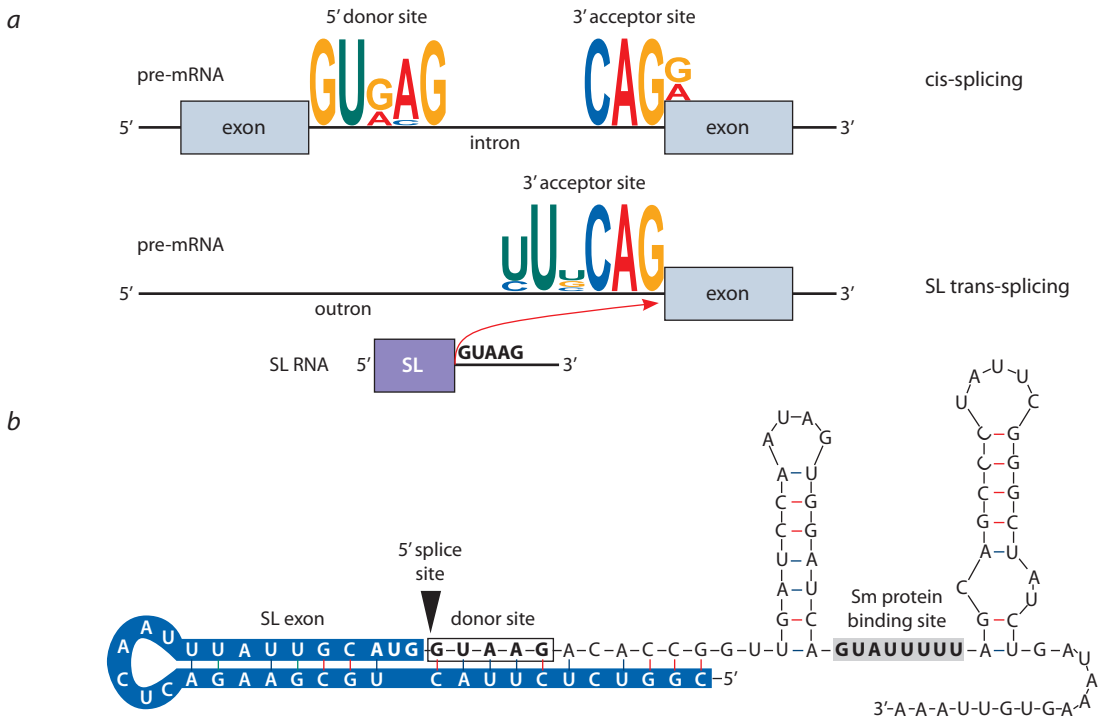


Fig. 1. Features of cis-splicing and SL trans-splicing in *M. lignano*.

a – conserved motifs enriched at the splicing junction sites in cis-spliced and SL trans-spliced genes; *b* – sequence and predicted secondary structure of *M. lignano* SL RNA gene.

is at the 5' end of the gene and forms the first hairpin loop, immediately after the SL sequence there is a clear 5' donor splicing site (GTAAG), and between two other hairpin loops there is a motif similar to the binding site of Sm spliceosomal protein (see Fig. 1, *b*) (Ganot et al., 2004; Stover et al., 2006).

Operon analysis. The important feature of SL trans-splicing is that it allows for processing of long polycistronic pre-mRNA molecules expressed from a single promoter region in a way similar to prokaryote operons. In principle, genome-guided transcriptome assembly using RNA-seq data allows identification of such operons and their corresponding pre-mRNA sequences, which we previously annotated as transcriptional units (Wudarski et al., 2017; Grudniewska et al., 2018). However, it is not always possible to fully reconstruct an operon from RNA-seq data alone, since transcriptional units predicted from RNA-seq data tend to split in the repeat-rich intergenic regions of operons, where read coverage depends on both operon expression level and the frequency of repeats in the genome. Instead, to estimate what fraction of *M. lignano* genes are organized in operons based on their genomic organization, we first explored how intergenic distances between trans- and non-trans-spliced genes are distributed in *M. lignano* genome (Fig. 2, *a*). We found that distribution of distances between trans-spliced genes has multimodal distribution, while it is unimodal distribution for non-trans-spliced/trans-spliced and non-trans-spliced/non-trans-spliced intergenic distances (see Fig. 2, *a*). SL trans-splicing is an ancient evolutionary mechanism (Lei et al., 2016), which is mostly abundant in the genomes of simply organized organisms, which have low repetitive content and relatively small genomes (Gregory et al., 2007). We hypothesized that neutral

accumulation of repeats could have influenced the distances between genes in the operons. Interestingly, after we adjusted the intergenic distances by subtracting the fraction occupied by repetitive sequences (simple repeats and transposable elements), it had the most impact on the distances between trans-spliced genes, revealing a clear bimodal distribution with the most prominent peak at around 100 bp (see Fig. 2, *a*). This observation indicates that repeats have a substantial contribution to intergenic distances in operons. To classify genes as belonging to the same operon, we decided to use the repeat-adjusted distances with a threshold value of 1 Kb, which separates the two modes of the intergenic distances between trans-spliced genes (see Fig. 2, *a*).

Using these criteria for defining operons, we found that 10458 genes (approx. 15 % of all genes and 40 % of SL trans-spliced genes) can be assigned to operons, of which 1854 (18 %) start from a non-trans-spliced gene (see Fig. 2, *b*, Fig. 3). The vast majority of them are comprised of two and three genes (75 and 18 %), with the maximum operon size reaching nine genes (two operons) (see Fig. 2, *b*). An example of an operon defined in this way is provided in Fig. 2, *c*.

SL trans-splicing is enriched in evolutionary conserved and stem cell genes. We know from a previous study (Grudniewska et al., 2018) that evolutionary conserved protein-coding genes, which still have detectable homology between *M. lignano* and human, are enriched in somatic stem cells – neoblasts (85 % compared to overall 47 %) (see Fig. 3). On the contrary, only 38 % of germline-specific genes in *M. lignano* are conserved in human, suggesting their relatively recent appearance in evolution of flatworms. We investigated whether there is a correlation between gene conservation and

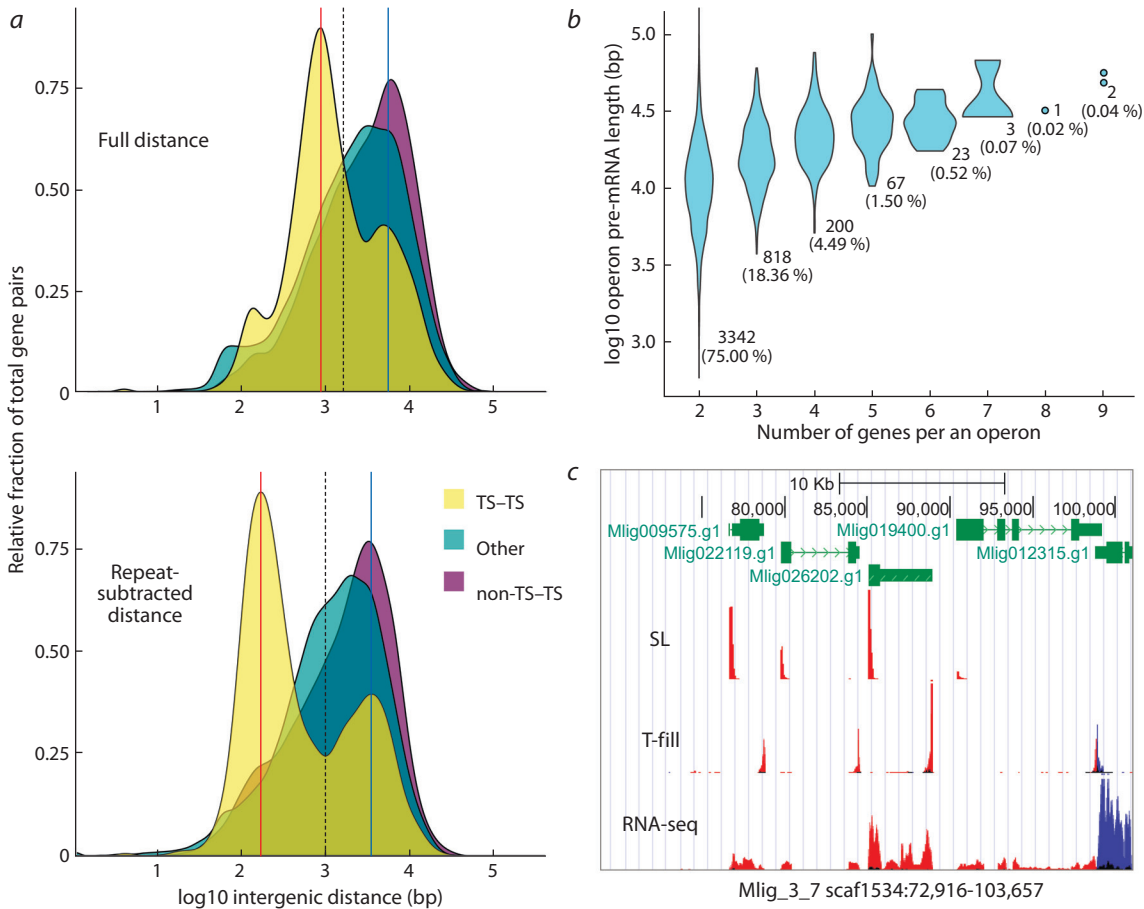


Fig. 2. Identification and characteristics of genes in operons in *M. lignano* genome.

a – distribution of intergenic distances between various gene types. TS – SL trans-spliced, non-TS – non-SL trans-spliced. Red and blue vertical lines indicate modes of the distributions. Vertical black dashed line indicates distance threshold value selected to separate genes in operons; *b* – putative pre-mRNA length and abundance of different operon sizes; *c* – an example of an operon with four genes as depicted in the *M. lignano* genome browser (<http://gb.macgenome.org>). Genes are in green, with exons as blocks and introns as dashed lines. Non-protein-coding part of the exons are narrower. SL – RNA-seq reads mapped which contained the SL sequence at their 5' ends (trimmed). T-fill – RNA-seq reads mapped containing mRNA 3' poly-A ends. Reads mapped in forward orientation are in red, and the reversed reads are in blue.

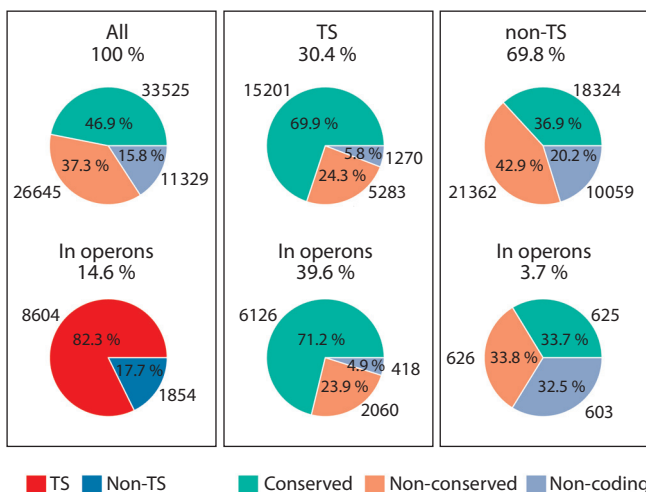


Fig. 3. Evolutionary conservation of *M. lignano* genes.

TS – SL trans-spliced; non-TS – non-SL trans-spliced; conserved – protein-coding genes with a homology to human; non-conserved – protein-coding genes lacking the homology to human; non-coding – genes do not code for a protein.

SL trans-splicing in *M. lignano*. We found that 69.9 % of the SL trans-spliced genes are conserved between *M. lignano* and human, while 24.3 % are not conserved and 5.8 % are non-coding (see Fig. 3). Trans-spliced genes that are located in operons have a very similar distribution of conserved, non-conserved and non-coding genes (see Fig. 3). In contrast, among non-trans-spliced genes only 36.9 % are conserved in human, while 42.9 % are non-conserved and 20.2 % are non-coding (see Fig. 3). Thus, SL trans-spliced genes are strongly enriched for conserved genes but there is no dependence on whether these genes are in operons or not.

Next, we calculated the fraction of SL trans-spliced genes among genes enriched in neoblasts (stem cells) and germline – the only proliferation capable cell types in the worm (Grudniewska et al., 2018). Intriguingly, 85 % of the stem cell genes (746) are SL trans-spliced, and almost 86 % (752) are conserved in human (see the Table), and 728 genes are both conserved in human and SL trans-spliced, which is 96.8 % of all the conserved genes in neoblasts. Given that out of 33525 conserved genes present in the Mlig_3_7 genome annotation 15201 (45.3 %) are trans-spliced (see Fig. 3), this represents

Summary of transcripts from *M. lignano* proliferation-capable cell types

Cell type	Total transcripts	Trans-spliced (%)	In operons (%)	Conserved in human (%)	Non-conserved in human (%)	Non-coding (%)
Neoblasts	878	746 (85.0)	343 (39.1)	752 (85.6)	19 (2.2)	107 (12.2)
Germline	1985	362 (18.2)	192 (9.7)	736 (37.1)	248 (12.5)	1001 (50.4)

a 2.13-fold enrichment for conserved SL trans-spliced genes among neoblast genes relative to the expected from the random distribution (p -value: $1.98e-7$, chi-square test). On the contrary, only 18 % of the germline genes are SL trans-spliced and 37 % are conserved in human. Taken all together, this suggests that SL trans-splicing plays an important role in stem cell regulation in *M. lignano*.

Discussion

SL trans-splicing is widespread in diverse flatworm taxa, including both parasitic and free-living species (Zayas et al., 2005; Protasio et al., 2012; Wudarski et al., 2017; Ershov et al., 2019). However, most of the studies of SL trans-splicing were focused on nematodes and trypanosomes (Lasda, Blumenthal, 2011; Lei et al., 2016). Here, we performed the first study which focuses on SL trans-splicing in the free-living regenerative flatworm model *M. lignano*. By reanalyzing the available genome and transcriptome data, we found that 30 % of the worm genes are SL trans-spliced, and 15 % are estimated to be organized in operons (see Fig. 3). For a comparison, in *C. elegans* 70 % of genes are SL trans-spliced and 17 % are in operons, in ascidian chordate *Ciona intestinalis* it is 58 and 20 %, respectively, and in the parasitic liver fluke *Schistosoma mansoni* 11 % are SL trans-spliced with a few genes in operons (Blumenthal, Gleason, 2003; Satou et al., 2008; Matsumoto et al., 2010; Protasio et al., 2012). Among free-living flatworms, trans-splicing was studied before (Zayas et al., 2005; Rossi et al., 2014), but there is no firm estimation of its abundance and prevalence of genes in operons. The size of operons in *M. lignano* also varies similarly to *C. elegans*, where it ranges from two to eight genes, with the most frequent intergenic distance around 100 bp, and the majority of operons comprised of two genes (see Fig. 2) (Allen et al., 2011).

The most striking finding of our study is that most of *M. lignano* SL trans-spliced genes are evolutionary conserved (see Fig. 3) and, most importantly, that overwhelming majority of neoblast-specific genes (85 %) are SL trans-spliced (see the Table). Interestingly, 39 % of neoblast genes are also clustered in operons (see the Table), suggesting their early evolutionary origin and the necessity for synchronized expression and similar transcriptional regulation. Neoblasts are the key players of outstanding regeneration capacity in free-living flatworms, and thus they are the primary subject of the studies on flatworm regeneration. All the tissue renewal and growth in adult flatworms is due to neoblast proliferation and differentiation (Egger et al., 2006; Ladurner et al., 2008; Wagner et al., 2011). Our data clearly indicates importance of SL trans-splicing for the gene regulation of neoblasts in *M. lignano* and lay ground for further studies of how exactly SL trans-splicing machinery contributes to different stages of neoblast activity.

Conclusion

Spliced leader trans-splicing affects a substantial fraction of *M. lignano* genes. We annotated and characterized the sequence of SL RNA, identified the conserved motifs at the exon-intron/exon-outtron junction sites in cis- and SL trans-spliced genes, and provided the first comprehensive analysis of genes comprising operons in *M. lignano*. Most importantly, we found that the SL trans-spliced fraction is over-represented by evolutionary conserved protein-coding genes, in contrast to the non-trans-spliced part of the genome, and that the stem cell-specific genes are predominantly SL trans-spliced. Our findings suggest an important and evolutionary conserved role of SL trans-splicing in regulation and maintenance of neoblasts in *M. lignano*. Thus, a thorough investigation of the molecular mechanism of SL trans-splicing is required to fully understand the regulation of regeneration and stem cell differentiation in flatworms.

References

- Allen M.A., Hillier L.W., Waterston R.H., Blumenthal T. A global analysis of *C. elegans* trans-splicing. *Genome Res.* 2011;21(2):255-264. DOI 10.1101/gr.113811.110.
- Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990;215(3):403-410. DOI 10.1016/S0022-2836(05)80360-2.
- Bailey T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics.* 2011;27(12):1653-1659. DOI 10.1093/bioinformatics/btr261.
- Blumenthal T., Gleason K.S. *Caenorhabditis elegans* operons: form and function. *Nat. Rev. Genet.* 2003;4(2):110-118. DOI 10.1038/nrg995.
- Boothroyd J.C., Cross G.A. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end. *Gene.* 1982;20(2):281-289. DOI 10.1016/0378-1119(82)90046-4.
- Conrad R., Liou R.F., Blumenthal T. Conversion of a trans-spliced *C. elegans* gene into a conventional gene by introduction of a splice donor site. *EMBO J.* 1993;12(3):1249-1255.
- Danks G.B., Raasholm M., Campsteijn C., Long A.M., Manak J.R., Lenhard B., Thompson E.M. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol. Biol. Evol.* 2015;32(3):585-599. DOI 10.1093/molbev/msu336.
- Danks G., Thompson E.M. Trans-splicing in metazoans: A link to translational control? *Worm.* 2015;4(3):e1046030. DOI 10.1080/21624054.2015.1046030.
- Egger B., Ladurner P., Nimeth K., Gschwentner R., Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. *Dev. Genes Evol.* 2006;216(10):565-577. DOI 10.1007/s00427-006-0069-4.
- Ershov N.I., Mordvinov V.A., Prokhortchouk E.B., Pakharukova M.Y., Gunbin K.V., Ustyantsev K., Genaev M.A., Blinov A.G., Mazur A., Boulygina E., Tsygankova S., Khrameeva E., Chekanov N., Fan G., Xiao A., Zhang H., Xu X., Yang H., Solovyev V., Lee S.M.-Y.,

- Liu X., Afonnikov D.A., Skryabin K.G. New insights from *Opisthorchis felineus* genome: update on genomics of the epidemiologically important liver flukes. *BMC Genomics*. 2019;20(1):399. DOI 10.1186/s12864-019-5752-8.
- Ganot P., Kallesøe T., Reinhardt R., Chourrout D., Thompson E.M. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol. Cell. Biol.* 2004;24(17):7795-7805. DOI 10.1128/MCB.24.17.7795-7805.2004.
- Gregory T.R., Nicol J.A., Tamm H., Kullman B., Kullman K., Leitch I.J., Murray B.G., Kapraun D.F., Greilhuber J., Bennett M.D. Eukaryotic genome size databases. *Nucleic Acids Res.* 2007;35(Suppl. 1):D332-D338. DOI 10.1093/nar/gkl828.
- Grudniewska M., Mouton S., Grelling M., Wolters A.H.G., Kuipers J., Giepmans B.N.G., Berezikov E. A novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. *Sci. Rep.* 2018; 8(1):1-10. DOI 10.1038/s41598-018-21107-4.
- Grudniewska M., Mouton S., Simanov D., Beltman F., Grelling M., de Mulder K., Arindarto W., Weissert P.M., van der Elst S., Berezikov E. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife*. 2016;5:e20607. DOI 10.7554/eLife.20607.
- Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36(9):2896-2898. DOI 10.1093/bioinformatics/btaa025.
- Hastings K.E.M. SL trans-splicing: easy come or easy go? *Trends Genet.* 2005;21(4):240-247. DOI 10.1016/j.tig.2005.02.005.
- Krause M., Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell*. 1987;49(6):753-761. DOI 10.1016/0092-8674(87)90613-1.
- Ladurner P., Egger B., De Mulder K., Pfister D., Kuaes G., Salvenmoser W., Schärer L. The stem cell system of the basal flatworm *Macrostomum lignano*. In: Bosch T.C.G. (Ed.). *Stem Cells: From Hydra to Man*. Dordrecht: Springer, Netherlands, 2008;75-94. DOI 10.1007/978-1-4020-8274-0_5.
- Lasda E.L., Blumenthal T. Trans-splicing. *Wiley Interdiscip. Rev. RNA*. 2011;2(3):417-434. DOI 10.1002/wrna.71.
- Lei Q., Li C., Zuo Z., Huang C., Cheng H., Zhou R. Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol.* 2016;8(3):562-577. DOI 10.1093/gbe/evw025.
- Liou R.F., Blumenthal T. trans-spliced *Caenorhabditis elegans* mRNAs retain trimethylguanosine caps. *Mol. Cell. Biol.* 1990;10(4):1764-1768.
- Matsumoto J., Dewar K., Wasserscheid J., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Satou Y., Hastings K.E.M. High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. *Genome Res.* 2010;20(5):636-645. DOI 10.1101/gr.100271.109.
- Mouton S., Grudniewska M., Glazenburg L., Guryev V., Berezikov E. Resilience to aging in the regeneration-capable flatworm *Macrostomum lignano*. *Aging Cell*. 2018;17(3):e12739. DOI 10.1111/accel.12739.
- Protasio A.V., Tsai I.J., Babbage A., Nichol S., Hunt M., Aslett M.A., Silva N.D., Velarde G.S., Anderson T.J.C., Clark R.C., Davidson C., Dillon G.P., Holroyd N.E., LoVerde P.T., Lloyd C., McQuillan J., Oliveira G., Otto T.D., Parker-Manuel S.J., Quail M.A., Wilson R.A., Zerlotini A., Dunne D.W., Berriman M. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* 2012;6(1):e1455. DOI 10.1371/journal.pntd.0001455.
- Rossi A., Ross E.J., Jack A., Sánchez Alvarado A. Molecular cloning and characterization of SL3: A stem cell-specific SL RNA from the planarian *Schmidtea mediterranea*. *Gene*. 2014;533(1):156-167. DOI 10.1016/j.gene.2013.09.101.
- Satou Y., Mineta K., Ogasawara M., Sasakura Y., Shoguchi E., Ueno K., Yamada L., Matsumoto J., Wasserscheid J., Dewar K., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Hastings K.E.M., Lemaire P., Lindquist E., Endo T., Hotta K., Inaba K. Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol.* 2008;9(10):R152. DOI 10.1186/gb-2008-9-10-r152.
- Stover N.A., Kaye M.S., Cavalcanti A.R.O. Spliced leader trans-splicing. *Curr. Biol.* 2006;16(1):R8-R9. DOI 10.1016/j.cub.2005.12.019.
- Wagner D.E., Wang I.E., Reddien P.W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*. 2011;332(6031):811-816. DOI 10.1126/science.1203983.
- Wasik K., Gurtowski J., Zhou X., Ramos O.M., Delás M.J., Battistoni G., Demerdash O.E., Falcatori I., Vizoso D.B., Smith A.D., Ladurner P., Schärer L., McCombie W.R., Hannon G.J., Schatz M. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl. Acad. Sci. USA*. 2015;112(40):12462-12467. DOI 10.1073/pnas.1516718112.
- Wudarski J., Egger B., Ramm S.A., Schärer L., Ladurner P., Zadesenets K.S., Rubtsov N.B., Mouton S., Berezikov E. The free-living flatworm *Macrostomum lignano*. *EvoDevo*. 2020;11(1):5. DOI 10.1186/s13227-020-00150-1.
- Wudarski J., Simanov D., Ustyantsev K., de Mulder K., Grelling M., Grudniewska M., Beltman F., Glazenburg L., Demircan T., Wunderer J., Qi W., Vizoso D.B., Weissert P.M., Olivieri D., Mouton S., Guryev V., Aboobaker A., Schärer L., Ladurner P., Berezikov E. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nat. Commun.* 2017; 8(1):2120. DOI 10.1038/s41467-017-02214-8.
- Wudarski J., Ustyantsev K., Glazenburg L., Berezikov E. Influence of temperature on development, reproduction and regeneration in the flatworm model organism, *Macrostomum lignano*. *Zool. Lett.* 2019; 5(1):7. DOI 10.1186/s40851-019-0122-6.
- Xie H., Hirsh D. *In vivo* function of mutated spliced leader RNAs in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*. 1998;95(8):4235-4240.
- Zadesenets K.S., Schärer L., Rubtsov N.B. New insights into the karyotype evolution of the free-living flatworm *Macrostomum lignano* (Platyhelminthes, Turbellaria). *Sci. Rep.* 2017;7(1):6066. DOI 10.1038/s41598-017-06498-0.
- Zayas R.M., Bold T.D., Newmark P.A. Spliced-leader trans-splicing in freshwater planarians. *Mol. Biol. Evol.* 2005;22(10):2048-2054. DOI 10.1093/molbev/msi200.
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 2003;31(13):3406-3415. DOI 10.1093/nar/gkg595.

ORCID ID

K.V. Ustyantsev orcid.org/0000-0003-4346-3868
E.V. Berezikov orcid.org/0000-0002-1145-2884


Acknowledgements. A part of work on SL motifs discovery and SL RNA gene mapping was done by K. Ustyantsev at the Institute of Cytology and Genetics SB RAS and supported by the Russian State Budget project No. 0259-2021-0009. The rest of the study was performed by K. Ustyantsev and E. Berezikov at the Institute of Cytology and Genetics SB RAS and supported by the Russian Science Foundation grant No. 20-14-00147 to E. Berezikov.


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

Received October 17, 2020. Revised December 3, 2020. Accepted December 8, 2020.

Английский текст <https://vavilov.elpub.ru/jour>

Macrostomum lignano как модельный объект для исследования генетики и геномики паразитических плоских червей


К.В. Устьянцев, В.Ю. Вавилова, А.Г. Блинов, Е.В. Березиков 


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

Аннотация. Инфекциям различных видов паразитических плоских червей подвержены сотни миллионов человек по всему миру. Как острые, так и хронические инфекции в отсутствие лечения с высокой частотой приводят к развитию тяжелых патологий и даже к смерти. Данные о снижении эффективности некоторых важных противогельминтных лекарственных препаратов и развитии резистентности к ним вынуждают исследователей искать альтернативные соединения. Паразитические плоские черви обладают сложным жизненным циклом, трудоемки и дорогостоящи в разведении, а также имеют ряд приспособлений, осложняющих работу с ними стандартными молекулярно-биологическими методами. Напротив, эволюционно близкородственные паразитическим плоским червям свободноживущие виды плоских червей лишены вышеописанных трудностей, что делает их перспективными альтернативными модельными объектами для поиска и исследования гомологичных генов. В этом обзоре мы описываем применение базального свободноживущего плоского червя *Macrostomum lignano* в качестве такой модели. *M. lignano* обладает большим набором удобных биологических и экспериментальных особенностей, таких как быстрое время репродукции, дешевизна и легкость в лабораторном разведении, оптическая прозрачность тела, облигатное половое размножение, аннотированные геномные и транскриптомные сборки, а также доступность современных молекулярных методов исследования, включая трансгенез, геномный нокадаун с помощью РНК-интерференции и гибридизацию *in situ*. Все это делает *M. lignano* пригодным для применения самых современных подходов «прямой» и «обратной» генетики, таких как транспозонный инсерционный мутагенез и методы направленного редактирования генома с использованием системы CRISPR/Cas9. Благодаря растущему количеству доступных сборок геномов и транскриптомов различных видов паразитических плоских червей новые знания, полученные в исследованиях на *M. lignano*, могут быть легко транслированы на паразитических плоских червей с применением современных биоинформационных подходов сравнительной геномики и транскриптомики. В подтверждение этому мы приводим результаты нашего биоинформационного поиска и анализа гомологичных генов *M. lignano* и паразитических плоских червей, которые позволили определить список перспективных генов-мишеней для дальнейшего исследования. Ключевые слова: плоские черви; паразитические черви; модельный организм.

Для цитирования: Устьянцев К.В., Вавилова В.Ю., Блинов А.Г., Березиков Е.В. *Macrostomum lignano* как модельный объект для исследования генетики и геномики паразитических плоских червей. *Вавиловский журнал генетики и селекции*. 2021;25(1):108-116. DOI 10.18699/VJ21.013

Macrostomum lignano as a model to study the genetics and genomics of parasitic flatworms

K.V. Ustyantsev, V.Yu. Vavilova, A.G. Blinov, E.V. Berezikov 

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

Abstract. Hundreds of millions of people worldwide are infected by various species of parasitic flatworms. Without treatment, acute and chronic infections frequently lead to the development of severe pathologies and even death. Emerging data on a decreasing efficiency of some important anthelmintic compounds and the emergence of resistance to them force the search for alternative drugs. Parasitic flatworms have complex life cycles, are laborious and expensive in culturing, and have a range of anatomic and physiological adaptations that complicate the application of standard molecular-biological methods. On the other hand, free-living flatworm species, evolutionarily close to parasitic flatworms, do not have the abovementioned difficulties, which makes them potential alternative models to search for and study homologous genes. In this review, we describe the use of the basal free-living flatworm *Macrostomum lignano* as such a model. *M. lignano* has a number of convenient biological and experimental properties, such as fast reproduction, easy and non-expensive laboratory culturing, optical body transparency, obligatory sexual reproduction, annotated genome and transcriptome assemblies, and the availability of modern molecular methods, including transgenesis, gene knockdown by RNA interference and *in situ* hybridization. All this makes *M. lignano* amenable to the most modern approaches of forward and reverse genetics, such as transposon insertional mutagenesis and

methods of targeted genome editing by the CRISPR/Cas9 system. Due to the availability of an increasing number of genome and transcriptome assemblies of different parasitic flatworm species, new knowledge generated by studying *M. lignano* can be easily translated to parasitic flatworms with the help of modern bioinformatic methods of comparative genomics and transcriptomics. In support of this, we provide the results of our bioinformatics search and analysis of genes homologous between *M. lignano* and parasitic flatworms, which predicts a list of promising gene targets for subsequent research.

Key words: flatworms; parasitic flatworms; model organism.

For citation: Ustyantsev K.V., Vavilova V.Yu., Blinov A.G., Berezikov E.V. *Macrostomum lignano* as a model to study the genetics and genomics of parasitic flatworms. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):108-116. DOI 10.18699/VJ21.013

Введение

Ежегодно сотни миллионов человек по всему миру подвержены инфекциям различных видов паразитических плоских червей (ППЧ) (Waikagul et al., 2018). Наибольшая частота заражений, а также самые тяжелые патологии приходится на виды класса трематод, или дигенетических сосальщиков, вызывающих такие известные паразитарные заболевания, как шистосомозы, клонорхозы и описторхозы. К характерным тяжелым последствиям от заражений печеночными сосальщиками можно отнести острые и хронические воспаления печени и желчевыводящих путей, которые могут переходить в фиброз печени и холангиокарциному соответственно (Wongratanacheewin et al., 2003; Kaewpitoon et al., 2008; Andrade, 2009; Pomaznoy et al., 2016; Schwartz, Fallon, 2018). Инфекции представителей другого класса ППЧ – цестод, или ленточных червей, – часто не приводят к столь тяжелым патологиям и смертельному исходу, но в долгосрочной перспективе и в отсутствие лечения могут вызывать существенные нарушения процессов жизнедеятельности и закономерное снижение качества жизни больных людей (Budke et al., 2009; Waikagul et al., 2018).

Во всем мире «средством номер один» по борьбе с гельминтными инфекциями уже более 40 лет выступают препарат «Празиквантел» и различные соединения на его основе (Chai, 2013; Pakharukova et al., 2015). Однако длительное и повсеместное применение Празиквантела уже привело к тому, что все чаще встречаются сообщения о возникновении резистентности к нему у разных видов гельминтов (Botros, Bennett, 2007; Wang et al., 2012; Mwangi et al., 2014; Jesudoss Chelladurai et al., 2018). Экспериментально индуцированную резистентность к Празиквантелу удалось продемонстрировать для некоторых видов шистосом (Mwangi et al., 2014). Изначальные успехи применения Празиквантела задержали инвестиции в разработку новых противогельминтных препаратов, что лишь осложняет ситуацию. Тем не менее созданные альтернативы Празиквантелу демонстрируют аналогичную, а чаще даже меньшую эффективность, еще большее количество побочных эффектов, а также оказываются действенными лишь для отдельных видов трематод (Siqueira et al., 2017). Все это подтверждает чрезвычайную необходимость поиска новых и более эффективных противогельминтных препаратов.

Паразитические плоские черви имеют сложный жизненный цикл со сменой нескольких хозяев (Mogand et al., 1995; Poulin, Cribb, 2002), трудоемки и дорогостоящи в лабораторном разведении, а также обладают рядом приспособлений, осложняющих работу с ними стан-

дартными молекулярно-биологическими методами. Эти особенности, несомненно, препятствуют быстрому поиску новых противогельминтных лекарственных средств. Наши знания о широком круге биологических вопросов получены в работах на удобных модельных организмах, таких как нематоды, плодовые мушки, мыши, дрожжи и др. Изучение свободноживущих представителей также помогает в получении новой информации об их паразитических родственниках. Например, при исследовании модельного свободноживущего круглого червя (нематоды) *Caenorhabditis elegans* были получены данные, которые позволили более детально определить механизм действия уже созданных противонематодных лекарственных препаратов, а также помогли поиску новых потенциальных генов, регулирующих жизненный цикл паразитических нематод. В дальнейшем эти гены могут стать мишенью для еще не разработанных лекарств (Cully et al., 1994; Couthier et al., 2004; Guest et al., 2007; Laing et al., 2010). Свободноживущие плоские черви могут быть использованы в качестве моделей для скрининга новых лекарственных препаратов, направленных против их паразитических родственников (Collins, Newmark, 2013). Несмотря на принципиальные различия в организации жизненного цикла, они обладают рядом эволюционно консервативных свойств, общих с ППЧ, касающихся их внутренней физиологии и репродукции.

В настоящей работе мы описываем свойства, преимущества и потенциальное применение свободноживущего плоского червя *Macrostomum lignano* как удобного модельного объекта для эффективного поиска консервативных генов, гомологичных генам ППЧ, которые могут стать мишенями для разработки новых противогельминтных лекарственных препаратов.

Общие характеристики *M. lignano* как модельного объекта

M. lignano – морской свободноживущий плоский червь (тип Platyhelminthes, класс Rhabditophora), относящийся к базальной (наиболее рано отделившейся от основной ветви) клade Macrostomorpha (Ladurner et al., 2005; Egger et al., 2015). Он неприхотлив, обладает широкой нормой реакции на различные условия среды обитания, такие как температура, соленость и концентрация кислорода (Rivera-Ingraham et al., 2013, 2016; Wudarski et al., 2019). Экспериментально показано, что черви способны выживать при температуре в диапазоне от 4 до 37 °C (Wudarski et al., 2019). *M. lignano* удобно разводить в лабораторных условиях (Wudarski et al., 2020). Размер взрослых особей варьирует от 1 до 3 мм в длину и 0.3 мм в ширину. Их мож-

но содержать в чашках Петри с искусственной морской водой. В качестве источника питания *M. lignano* служит вид одноклеточных диатомовых водорослей *Nitzschia curvilineata*, которые также удобно разводить непосредственно в лаборатории при искусственном освещении. В одной стандартной (9 см) чашке Петри можно одновременно содержать до 500–600 особей. Стандартными условиями разведения считаются температура 20 °C и 14/10-часовой цикл смены дня и ночи.

Свободноживущие плоские черви знамениты своей высокой способностью к регенерации (Egger et al., 2006; Mouton et al., 2018; Ivankovic et al., 2019). Известными чемпионами в этом являются планарии, способные восстановить полноценную особь из нескольких клеток (Wagner et al., 2011). *M. lignano* немногим уступает планариям и может полностью регенерировать тело постериально уровню глотки и антериально головному нервному ганглию (Egger et al., 2006). Способность к регенерации плоских червей обеспечивается делением и дифференцировкой популяции плюрипотентных соматических стволовых клеток, называемых необластами (Wagner et al., 2011). Необласты и их потомки, делящиеся клетки-предшественники дифференцированных тканей, – это единственные делящиеся клетки в организме плоских червей, и, кроме регенерации, они также ответственны за естественное обновление тканей при гомеостазе (Nimeth et al., 2002; Ladurner et al., 2008). Следует отметить, что у ППЧ также выделяют неопластоподобные клетки, морфологически схожие с хорошо описанными необластами свободноживущих видов (Brehm, 2010; Collins, Newmark, 2013; Collins et al., 2013; McCusker et al., 2016). Неопластоподобные клетки способны дифференцироваться в другие типы клеток и ответственны за регенерацию потерянных частей тела у ППЧ, а также обладают сходным транскрипционным профилем с необластами свободноживущих видов. Таким образом, прослеживается очевидная гомология в устройстве центральной системы поддержания гомеостаза и регенерации между ППЧ и свободноживущими плоскими червями.

Важное преимущество *M. lignano* по сравнению с другими популярными свободноживущими модельными плоскими червями, планариями, – оптическая прозрачность его тела (Ivankovic et al., 2019; Wudarski et al., 2020). Это позволяет легко производить морфологические наблюдения структур его внутренних органов и происходящих в них процессов с использованием световой микроскопии. *M. lignano* – гермафродит с облигатным перекрестным оплодотворением, что тоже выгодно отличает его от планарий, которые в лаборатории преимущественно размножаются бесполом путем через прямое деление тела, а также проявляют сильный генетический мозаицизм даже внутри одной особи (Schärer, Ladurner, 2003; Leria et al., 2019). Облигатное половое размножение дает возможность использовать *M. lignano* в контролируемых генетических исследованиях.

Уникальная особенность *M. lignano* среди всех видов плоских червей – наличие простого и эффективного метода получения трансгенных особей – трансгенеза (Wudarski et al., 2017). *M. lignano* откладывает 1–2 одноклеточных яйца в день. Яйца достаточно крупные по размеру (~100 мкм), с относительно твердой оболочкой, и ими

можно легко манипулировать с помощью самодельных пластиковых микроинструментов. Эти особенности позволили разработать удачный протокол для доставки различных генетических конструкций (ДНК, мРНК и белков) в яйца посредством микроинъекций (Wudarski et al., 2017; Вударски и др., 2020). К настоящему времени уже доступен ряд трансгенных линий *M. lignano*, экспрессирующих ряд репортерных зеленых и красных флуоресцентных белков в различных органах и тканях, что способствует более детально исследовать место и динамику экспрессии гена в различных условиях *in vivo* (Wudarski et al., 2017, 2019).

Помимо трансгенеза, в работе с *M. lignano* успешно применяются другие классические молекулярно-биологические и цитологические методы исследования. Так, локализацию экспрессии интересующего гена в теле червя можно изучать с применением метода гибридизации *in situ* (Pfister et al., 2007; Grudniewska et al., 2016; Wudarski et al., 2017; Lengerer et al., 2018). Для исследования функции гена существует очень простой и эффективный протокол «нокдаун» экспрессии с помощью РНК-интерференции, причем не требуется специальной доставки двухцепочечных РНК (дцРНК) конструкций. Червей помещают в раствор дцРНК, и через 1–3 недели благодаря прозрачности *M. lignano* можно наблюдать произошедшие морфологические либо физиологические изменения, а также перемены в поведении (Grudniewska et al., 2016, 2018; Lengerer et al., 2018; Wudarski et al., 2019). Таким образом, совокупное применение доступных экспериментальных методов позволяет проводить комплексные исследования экспрессии и функции генов у *M. lignano*.

У каждого современного модельного организма должна быть качественная сборка генома и транскриптома с аннотацией генов и повторенных последовательностей, транспозонов и простых/тандемных повторов, *M. lignano* – не исключение (Wasik et al., 2015; Grudniewska et al., 2016, 2018; Wudarski et al., 2017; Biryukov et al., 2020). Он обладает достаточно компактным геномом: ~500 млн п. н. Транскриптомная и геномные сборки находятся в открытом доступе и могут быть просмотрены на удобном веб-ресурсе <http://gb.macgenome.org/> (Wudarski et al., 2017; Grudniewska et al., 2018). Уже известны гены, которые дифференциально экспрессируются только в необластах, а также в органах репродуктивной системы червя (Grudniewska et al., 2018, 2016). Поэтому *M. lignano* открыт для методов биоинформационного анализа эволюции, сравнительной геномики и транскриптомики, что является принципиальным для поиска консервативных последовательностей генов, гомологичных генам паразитических плоских червей (см. таблицу).

Частные характеристики *M. lignano* как модели для поиска генов-мишеней, ответственных за развитие и функционирование репродуктивной системы паразитических плоских червей

Важную роль в развитии патологий при трематодных инфекциях играет развитие острого и хронического воспаления, вызванных постоянной откладкой новых яиц паразитами, активирующих иммунный ответ, что особенно

Сравнение основных свойств свободноживущих плоских червей *M. lignano*, планарий и паразитических плоских червей как модельных организмов

Свойство	<i>M. lignano</i>	Планарии	Паразитические плоские черви
Общие свойства			
Стоимость содержания и разведения в культуре	Дешево	Дешево	Дорого
Трудоемкость содержания	Легко	Легко	Сложно
Содержание <i>in vitro</i>	Да	Да	Возможно, но затруднительно
Жизненный цикл	Простой, без метаморфозов	Простой, без метаморфозов	Сложный, смена нескольких организмов-хозяев и различных личиночных стадий
Тип размножения	Только половое, перекрестное	Бесполое и половое	Бесполое и половое
Возможность контролируемых генетических исследований	Да	Нет, лабораторные линии размножаются бесполом путем	Нет, половое размножение происходит внутри организма-хозяина и неконтролируемо (Richards, 1975)
Прозрачность тела	Да	Нет, сильная пигментация	Варьирует между разными видами и стадиями жизненного цикла
Наличие аннотированных геномной и транскриптомной сборок	Да (Wudarski et al., 2017; Grudniewska et al., 2018)	Да (Grohme et al., 2018)	Да (Berriman et al., 2009; Zheng et al., 2013; Ershov et al., 2019)
Доступные методики работы			
Трансгенез	Да, микроинъекции в одноклеточные яйца (Wudarski et al., 2017; Вударски и др., 2020)	Нет	Затруднен и неэффективен, наследование трансгена не показано: электропорация и микроинъекция во взрослых особей (Beckmann, Grevelding, 2012; Moguel et al., 2015)
РНК-интерференция	Да, помещение в раствор дцРНК (Wudarski et al., 2020)	Да, инъекция дцРНК, кормление смеси дцРНК с едой (Rouhana et al., 2013)	Да, эффективная доставка дцРНК с помощью электропорации, бомбардировки с микрочастицами, липофекции на всех стадиях развития (McGonigle et al., 2008; Pierson et al., 2010; Da'dara, Skelly, 2015)
<i>In situ</i> гибридизация	Да (Wudarski et al., 2020)	Да (Rouhana et al., 2013)	Да (Cogswell et al., 2011)

актуально для шистосомозов (Wongratanacheewin et al., 2003; Kaewpitoon et al., 2008; Collins, Newmark, 2013; Schwartz, Fallon, 2018). Таким образом, репродуктивная система гельминтов и гены, контролирующие ее развитие и гомеостаз, представляются перспективными мишенями для разработки новых лекарственных препаратов, направленных на подавление их экспрессии.

В недавней работе на *M. lignano* (Grudniewska et al., 2018) показано, что значительная часть его генов, классифицированных как гены репродуктивной системы, состоит из генов, специфичных для плоских червей (как свободноживущих, так и паразитических), для которых отсутствует ближайший гомолог у человека и других модельных организмов. Исследование генов, характерных именно для плоских червей, может оказаться ключом к поиску новых противогельминтных препаратов, обладающих малым числом побочных эффектов, благодаря их целенаправленному действию на продукты генов, не встречающихся у человека. *M. lignano* – очень удобный объект для поиска таких генов-мишеней. Как уже отмечено, все органы его репродуктивной системы отчетливо различимы под обычным световым диссекционным микроскопом. Это существенно облегчает скрининг фенотипов, связанных с нарушением работы генов гонад

и/или копулятивных органов (Grudniewska et al., 2018). Причем гермафродитизм червя позволит сохранять в популяциях генетические нарушения, связанные с работой либо мужской, либо женской половых систем. Нарушения в фертильности будут заметны уже в течение одной недели при 25 °C (Wudarski et al., 2019), что позволит не пропустить мутации в отсутствие явного морфологического фенотипа.

Основные методы и применение *M. lignano* для решения задач сравнительной геномики

Мы живем в начале эры целенаправленного геномного редактирования, которая наступила при повсеместном распространении технологии CRISPR/Cas9 (Anzalone et al., 2020). При наличии хорошо аннотированной геномной сборки возможно внесение мутаций в конкретный ген, которые привели бы к полному нарушению его функции («нокауту») (Chen et al., 2014). Особый интерес представляет встраивание маркерных последовательностей (например, флуоресцентных белков) непосредственно в рамку считывания целевого гена («нокин», англ. knock-in), что позволит визуализировать паттерн его экспрессии по непосредственной локализации кодируемого белка (Albadri et al., 2017; Artegiani et al., 2020). При сочетании марки-

рования нескольких белков с разными флуоресцентными белками возможно, например, проведение интерактивных исследований.

Работа системы CRISPR/Cas9 зависит всего лишь от двух (в случае с получением «нокаутов») или трех (в случае с получением «нокинов») или внесения целенаправленных замен и делеций) компонент: направляющей РНК, белка-нуклеазы Cas9 и матрицы для гомологичной рекомбинации. В простейшем случае это два плазмидных вектора, один из которых кодирует гидовую РНК и Cas9, а второй представляет матрицу для гомологичной рекомбинации (Hsu et al., 2014). То есть это может быть сочетание непосредственно *in vitro* синтезированной гидовой РНК и Cas9 в форме мРНК либо белка Cas9 уже в комплексе с гидовой РНК, что исключает вероятность встройки ненужного впоследствии плазмидного вектора (Hsu et al., 2014; Kim et al., 2014). Успешное и воспроизводимое применение технологии CRISPR/Cas9 невозможно без эффективной доставки генетических конструкций (ДНК, мРНК или белки). В настоящее время *M. lignano* – единственный плоский червь, для которого это осуществимо с использованием метода микроинъекций в одноклеточные яйца червя (Wudarski et al., 2017). Такой метод доставки является, несомненно, наиболее эффективным, так как все компоненты системы доставляются одновременно в нужном молярном соотношении на стадии одной клетки, что снижает вероятность получения мозаичного потомства. Хотя на сегодняшний день нет опубликованных данных о применении методов CRISPR/Cas9 в *M. lignano*, наши предварительные эксперименты свидетельствуют о том, что этот подход можно эффективно использовать для внесения «нокинов» в геном *M. lignano*.

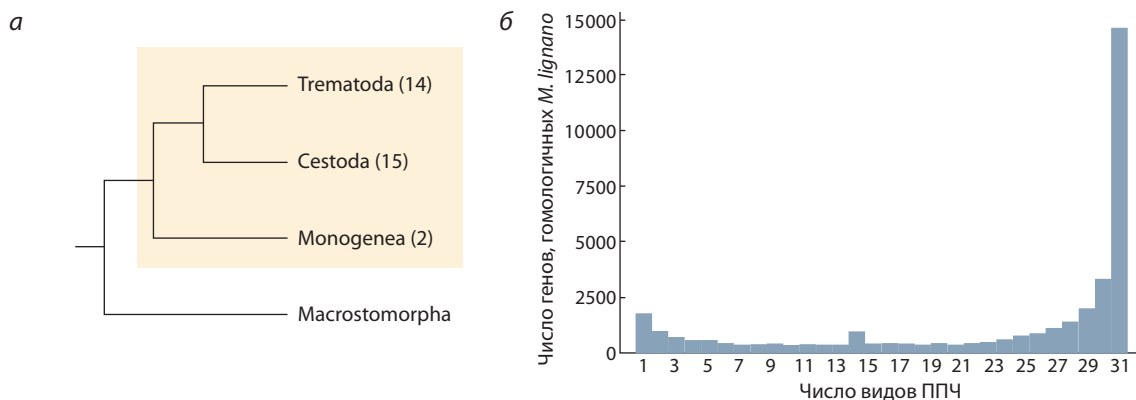
Исследование фенотипа по целенаправленному нарушению/маркированию конкретного гена относится к методам «обратной» генетики (Pareek et al., 2018). Основной недостаток такого подхода – высокое требование к качеству аннотации генома, необходимое для корректной подборки места модификации и предварительной оценки функции гена по его гомологии с уже известными белками (Skromne, Prince, 2008). Геномное редактирование с помощью CRISPR/Cas9 зависит также от частоты встречаемости паттерна GG в геноме, так как белку Cas9 необходимо сначала обнаружить PAM-сайт (Protospacer Adjacent Motif) NGG в целевой последовательности (Hsu et al., 2014). Отдельной проблемой остается существенный разброс в эффективности внесения двуникового разрыва различными гидовыми РНК, который далеко не всегда удается точно предсказать на стадии дизайна *in silico* (Chuai et al., 2017). И если классические модели – клеточные линии человека, мышь, дрозофила, нематода *C. elegans* и дрожжи – изучены досконально, имеется достаточно информации о функции их генов, чтобы ориентировочно предсказать фенотип, и их геномы обладают подходящим GC-составом, то с альтернативными объектами иначе.

Не всегда известно функциональное назначение того или иного гена, так как этот ген может быть консервативен только внутри определенного эволюционного таксона (как, например, с генами репродуктивной системы плоских червей). Геном может также быть с низким GC-

составом, менее 40 %, что снижает вероятность встречи последовательности GG в необходимых районах, мутации в которых могли бы привести к нокауту целевого гена (Casandra et al., 2018). В таких случаях следует обратиться к исторически более раннему подходу «прямой» генетики: от фенотипа к гену (Pareek et al., 2018).

Транспозонный инсерционный мутагенез наиболее развит среди методов прямой генетики. В отличие от химических мутагенов, которые эффективно производят мутации по всему геному, но требуют затем длительного времени на картирование мутации, перемещение транспозона и место его встраивания можно легко обнаружить с применением современных подходов в течение одного-двух дней (Potter, Luo, 2010; Frøkjær-Jensen et al., 2012; Stefano et al., 2016; Kalendar et al., 2019). Это достигается за счет того, что последовательности транспозона исходно нет в исследуемом геноме, а также потому что внутри транспозона возможно поместить различные репортерные конструкции, промоторные, энхансерные и генные ловушки, которые будут дополнительно сообщать о встраивании транспозона в какой-либо локус (Bonin, Mann, 2004; Song et al., 2012; Chang et al., 2019). В недавней работе на возбудителе малярии именно с помощью транспозонного мутагенеза, при использовании ДНК-транспозона *piggyBac*, создано 38000 мутантов малярийного плазмодия, среди которых было выявлено 2680 генов, ответственных за размножение этого паразита в клетках крови (Casandra et al., 2018). Авторы работы отмечают, что применение метода CRISPR/Cas9 не представлялось эффективным из-за аномально низкого GC-состава (<20 %) генома плазмодия. *M. lignano* и другие плоские черви, включая ППЧ, еще не относятся к классическим и повсеместно используемым модельным объектам. Как отмечено выше, гены репродуктивной системы плоских червей практически не имеют гомологов у других животных, что исключает предсказательную силу в рамках методов «обратной» генетики. Таким образом, транспозонный мутагенез представляется наиболее перспективным методом для поиска генов, ответственных за работу репродуктивной системы плоских червей, а также генов, специфичных для плоских червей и ответственных за ряд других процессов, а разработка эффективного протокола транспозонного мутагенеза у *M. lignano* является актуальной задачей.

Немаловажно, что перенос знаний, полученных в рамках экспериментальных подходов на *M. lignano*, на ППЧ возможен благодаря наличию большого количества геномных и транскриптомных сборок наиболее значимых видов паразитических плоских червей, доступных в базе данных WormBase ParaSite (<https://parasite.wormbase.org/index.html>) (Berriman et al., 2009; Zheng et al., 2013; Cwiklinski et al., 2015; Ershov et al., 2019). С применением современных биоинформационных методов сравнительной геномики и транскриптомики можно сразу найти последовательности потенциальных генов-мишеней, обнаруженных у *M. lignano*, гомологичные у разных видов ППЧ, провести их сравнительный и филогенетические анализы *in silico*. Это позволит отобрать гены-кандидаты, которые будут наиболее консервативными среди всех геномов ППЧ, а также (желательно) будут обладать низкой степенью родства с генами человека.



Категории гомологов с <i>M. lignano</i>	Trematoda	Cestoda	Monogenea	Все таксоны
Всего гомологов	34 844	34 035	29 307	27 889
Консервативные у ПЧ	5895	5527	3525	2887
В необластах	37	34	19	18
В гонадах	129	130	70	56
Консервативные у человека	28 949	28 508	25 782	25 002
В необластах	693	695	644	637
В гонадах	554	561	502	486

Гомология генов *M. lignano* и видов ППЧ.

a – филогенетические взаимоотношения между *M. lignano* (Macrostomorpha) и классами ППЧ, по (Park et al., 2007). В скобках после названий таксонов указано число видов из базы данных WormBase ParaSite, использованных в анализе; *б* – распределение количества генов-гомологов *M. lignano* по числу видов ППЧ; *в* – распределение различных категорий гомологичных генов *M. lignano* между классами ППЧ. ПЧ – плоские черви. В столбце «Все таксоны» указано количество гомологов, найденных хотя бы у одного вида каждого класса.

Биоинформационный анализ консервативных генов между ППЧ и *M. lignano*

Из базы данных WormBase ParaSite были взяты аминокислотные последовательности белок-кодирующих генов 31 вида ППЧ: 14 видов класса Trematoda, 15 – Cestoda и 2 вида – Monogenea (см. рисунок, *a*, Приложение 1)¹.

Среди 60 170 белок-кодирующих последовательностей *M. lignano* обнаружено 37 113 гомологов как минимум к одному из видов ППЧ, причем 14 576 гомологов было найдено у 31 вида (медиана – 29 видов) (см. рисунок, *б*, Приложения 1 и 2). Основные количественные показатели распределения гомологов *M. lignano* по видам классов ППЧ представлены на рисунке (*в*) и в Приложении 2. Мы обнаружили 2887 белок-кодирующих генов, консервативных среди видов всех трех классов ППЧ, но не имеющих гомолога у человека, среди них 18 генов специфичны для необластов *M. lignano* и 56 – для генов репродуктивной системы червя (см. рисунок (*в*), Приложение 2) (Grudniewska et al., 2018). Эти гены представляются наиболее перспективными кандидатами для исследования экспериментальными методами обратной генетики.

Заключение

В настоящей работе мы осветили основные свойства свободноживущего плоского червя *M. lignano* как модельного организма в целом и те его особенности, которые делают

его перспективным объектом для быстрого и эффективного поиска потенциальных генов-мишеней новых противогельминтных препаратов. Так, доступность легкого метода трансгенеза у *M. lignano* открывает путь ко всему арсеналу современных молекулярно-биологических методов для исследования функции генов, а прозрачность строения позволяет без дополнительных манипуляций наблюдать *in vivo* фенотипические изменения, вызванные нарушением или маркированием последовательности гена с применением методов «прямой» и «обратной» генетики. Гены регуляции развития и функционирования репродуктивной системы плоских червей представляются наиболее перспективными мишенями ввиду консервативности среди плоских червей и отсутствия гомологов у человека.

Список литературы / References

Вударски Я., Устьянцев К.В., Березиков Е.В. Подходы к эффективному редактированию генома регенерирующего свободноживущего плоского червя *Macrostomum lignano*. В: Методы редактирования генов и геномов. Новосибирск, 2020;101-115. [Wudarski J., Ustyantsev K.V., Berezikov E.V. Approaches to efficient genome editing in the regenerating free-living flatworm *Macrostomum lignano*. In: Methods for Editing Genes and Genomes. Novosibirsk, 2020;101-115. (in Russian)]

Albadri S., Del Bene F., Revenu C. Genome editing using CRISPR/Cas9-based knock-in approaches in zebrafish. *Methods*. 2017;121-122:77-85. DOI 10.1016/j.ymeth.2017.03.005.

Andrade Z.A. Schistosomiasis and liver fibrosis. *Parasite Immunol*. 2009;31:656-663. DOI 10.1111/j.1365-3024.2009.01157.x.

¹ Приложения 1 и 2 см. по адресу: http://www.macgenome.org/download/pdf/Ustyantsev_2021/

- Anzalone A.V., Koblan L.W., Liu D.R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* 2020;38:824-844. DOI 10.1038/s41587-020-0561-9.
- Artegiani B., Hendriks D., Beumer J., Kok R., Zheng X., Joore I., Chuva de Sousa Lopes S., van Zon J., Tans S., Clevers H. Fast and efficient generation of knock-in human organoids using homology-independent CRISPR-Cas9 precision genome editing. *Nat. Cell Biol.* 2020;22:321-331. DOI 10.1038/s41556-020-0472-5.
- Beckmann S., Grevelding C.G. Paving the way for transgenic schistosomes. *Parasitology.* 2012;139:651-668. DOI 10.1017/S0031182011001466.
- Berriman M., Haas B.J., LoVerde P.T., Wilson R.A., Dillon G.P., Cerqueira G.C., Mashiyama S.T., Al-Lazikani B., Andrade L.F., Ashton P.D., Aslett M.A., Bartholomeu D.C., Blandin G., Caffrey C.R., Coghlan A., Coulson R., Day T.A., Delcher A., DeMarco R., Djikeng A., Eyre T., Gamble J.A., Ghedin E., Gu Y., Hertz-Fowler C., Hirai H., Hirai Y., Houston R., Ivens A., Johnston D.A., Lacerda D., Macedo C.D., McVeigh P., Ning Z., Oliveira G., Overington J.P., Parkhill J., Perteua M., Pierce R.J., Protasio A.V., Quail M.A., Rajandream M.-A., Rogers J., Sajid M., Salzberg S.L., Stanke M., Tivey A.R., White O., Williams D.L., Wortman J., Wu W., Zamanian M., Zerlotini A., Fraser-Liggett C.M., Barrell B.G., El-Sayed N.M. The genome of the blood fluke *Schistosoma mansoni*. *Nature.* 2009;460:352-358. DOI 10.1038/nature08160.
- Biryukov M., Berezikov E., Ustyantsev K. Classification of LTR retrotransposons in the flatworm *Macrostomum lignano*. *Pisma v Vavilovskii Zhurnal Genetiki i Seleksii = Letters to Vavilov Journal of Genetics and Breeding.* 2020;6(2):54-59. DOI 10.18699/Letters 2020-6-12.
- Bonin C.P., Mann R.S. A piggyBac transposon gene trap for the analysis of gene expression and function in *Drosophila*. *Genetics.* 2004;167:1801-1811. DOI 10.1534/genetics.104.027557.
- Botros S.S., Bennett J.L. Praziquantel resistance. *Expert Opin. Drug Discov.* 2007;S35-S40. DOI 10.1517/17460441.2.S1.S35.
- Brehm K. *Echinococcus multilocularis* as an experimental model in stem cell research and molecular host-parasite interaction. *Parasitology.* 2010;137:537-555. DOI 10.1017/S0031182009991727.
- Budke C.M., White A.C., Jr., Garcia H.H. Zoonotic Larval cestode infections: neglected, neglected tropical diseases? *PLoS Negl. Trop. Dis.* 2009;3:e319. DOI 10.1371/journal.pntd.0000319.
- Cassandra D., Oberstaller J., Jiang R.H.Y., Bronner I.F., Adams J.H., Rayner J.C., Brown J., Mayho M., Swanson J., Otto T.D., Li S., Zhang M., Liao X., Wang C., Udenze K., Adapa S.R. Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. *Science.* 2018;360:eap7847. DOI 10.1126/science.aap7847.
- Chai J.-Y. Praziquantel treatment in trematode and cestode infections: an update. *Infect. Chemother.* 2013;45:32-43. DOI 10.3947/ic.2013.45.1.32.
- Chang H., Pan Y., Landrette S., Ding S., Yang D., Liu L., Tian L., Chai H., Li P., Li D.-M., Xu T. Efficient genome-wide first-generation phenotypic screening system in mice using the piggyBac transposon. *Proc. Natl. Acad. Sci. USA.* 2019;116:18507-18516. DOI 10.1073/pnas.1906354116.
- Chen X., Xu F., Zhu C., Ji J., Zhou X., Feng X., Guang S. Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in *Caenorhabditis elegans*. *Sci. Rep.* 2014;4:7581. DOI 10.1038/srep 07581.
- Chuai G., Wang Q.-L., Liu Q. *In silico* meets *in vivo*: towards computational CRISPR-based sgRNA design. *Trends Biotechnol.* 2017;35:12-21. DOI 10.1016/j.tibtech.2016.06.008.
- Cogswell A.A., Collins J.J., Newmark P.A., Williams D.L. Whole mount *in situ* hybridization methodology for *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* 2011;178:46-50. DOI 10.1016/j.molbiopara.2011.03.001.
- Collins J.J., Newmark P.A. It's no fluke: the planarian as a model for understanding schistosomes. *PLoS Pathog.* 2013;9:e1003396. DOI 10.1371/journal.ppat.1003396.
- Collins J.J., Wang B., Lambrus B.G., Tharp M., Iyer H., Newmark P.A. Adult somatic stem cells in the human parasite, *Schistosoma mansoni*. *Nature.* 2013;494:476-479. DOI 10.1038/nature11924.
- Couthier A., Smith J., McGarr P., Craig B., Gilleard J.S. Ectopic expression of a *Haemonchus contortus* GATA transcription factor in *Caenorhabditis elegans* reveals conserved function in spite of extensive sequence divergence. *Mol. Biochem. Parasitol.* 2004;133:241-253.
- Cully D.F., Vassilatis D.K., Liu K.K., Paress P.S., Van der Ploeg L.H.T., Schaeffer J.M., Arena J.P. Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature.* 1994;371:707-711. DOI 10.1038/371707a0.
- Cwiklinski K., Dalton J.P., Dufresne P.J., La Course J., Williams D.J., Hodgkinson J., Paterson S. The *Fasciola hepatica* genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biol.* 2015;16:71. DOI 10.1186/s13059-015-0632-2.
- Da'dara A.A., Skelly P.J. Gene suppression in schistosomes using RNAi. In: Peacock C. (Ed.) *Parasitic Genomics Protocols, Methods in Molecular Biology*. New York: Springer, 2015;143-164. DOI 10.1007/978-1-4939-1438-8_8.
- Egger B., Ladurner P., Nimeth K., Gschwentner R., Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. *Dev. Genes Evol.* 2006;216:565-577. DOI 10.1007/s00427-006-0069-4.
- Egger B., Lapraz F., Tomiczek B., Müller S., Dessimoz C., Girstmair J., Škunca N., Rawlinson K.A., Cameron C.B., Beli E., Todaro M.A., Gammoudi M., Noreña C., Telford M.J. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr. Biol.* 2015;25:1347-1353. DOI 10.1016/j.cub.2015.03.034.
- Ershov N.I., Mordvinov V.A., Prokhortchouk E.B., Pakharukova M.Y., Gunbin K.V., Ustyantsev K., Genaev M.A., Blinov A.G., Mazur A., Boulygina E., Tsygankova S., Khrameeva E., Chekanov N., Fan G., Xiao A., Zhang H., Xu X., Yang H., Solovyev V., Lee S.M.-Y., Liu X., Afonnikov D.A., Skryabin K.G. New insights from *Opisthorchis felinus* genome: update on genomics of the epidemiologically important liver flukes. *BMC Genomics.* 2019;20:399. DOI 10.1186/s12864-019-5752-8.
- Frøkjær-Jensen C., Davis M.W., Ailion M., Jorgensen E.M. Improved Mos1-mediated transgenesis in *C. elegans*. *Nat. Methods.* 2012;9:117-118. DOI 10.1038/nmeth.1865.
- Grohme M.A., Schloissnig S., Rozanski A., Pippel M., Young G.R., Winkler S., Brandl H., Henry I., Dahl A., Powell S., Hiller M., Myers E., Rink J.C. The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature.* 2018;554:56-61. DOI 10.1038/nature25473.
- Grudniewska M., Mouton S., Grelling M., Wolters A.H.G., Kuipers J., Giepmans B.N.G., Berezikov E. A novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. *Sci. Rep.* 2018;8:1-10. DOI 10.1038/s41598-018-21107-4.
- Grudniewska M., Mouton S., Simanov D., Beltman F., Grelling M., de Mulder W., Arindrarto W., Weissert P.M., van der Elst S., Berezikov E. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife.* 2016;5:e20607. DOI 10.7554/eLife.20607.
- Guest M., Bull K., Walker R.J., Amliwala K., O'Connor V., Harder A., Holden-Dye L., Hopper N.A. The calcium-activated potassium channel, SLO-1, is required for the action of the novel cyclo-octadepsipeptide anthelmintic, emodepside, in *Caenorhabditis elegans*. *Int. J. Parasitol.* 2007;37:1577-1588. DOI 10.1016/j.ijpara.2007.05.006.
- Hsu P.D., Lander E.S., Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell.* 2014;157:1262-1278. DOI 10.1016/j.cell.2014.05.010.
- Ivankovic M., Haneckova R., Thommen A., Grohme M.A., Vila-Far-ré M., Werner S., Rink J.C. Model systems for regeneration: planarians. *Development.* 2019;146. DOI 10.1242/dev.167684.

- Jesudoss Chelladurai J., Kifleyohannes T., Scott J., Brewer M.T. Praziquantel resistance in the zoonotic cestode *Dipylidium caninum*. *Am. J. Trop. Med. Hyg.* 2018;99:1201-1205. DOI 10.4269/ajtmh.18-0533.
- Kaewpitoon N., Kaewpitoon S.J., Pengsaa P., Sripa B. *Opisthorchis viverrini*: The carcinogenic human liver fluke. *World J. Gastroenterol.* 2008;14:666-674. DOI 10.3748/wjg.14.666.
- Kalendar R., Shustov A.V., Seppänen M.M., Schulman A.H., Stoddard F.L. Palindromic sequence-targeted (PST) PCR: a rapid and efficient method for high-throughput gene characterization and genome walking. *Sci. Rep.* 2019;9:1-11. DOI 10.1038/s41598-019-54168-0.
- Kim S., Kim D., Cho S.W., Kim J., Kim J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* 2014;24:1012-1019. DOI 10.1101/gr.171322.113.
- Ladurner P., Egger B., De Mulder K., Pfister D., Kuales G., Salvenmoser W., Schärer L. The stem cell system of the basal flatworm *Macrostomum lignano*. In: Bosch T.C.G. (Ed.). *Stem Cells: From Hydra to Man*. Dordrecht: Springer, 2008;75-94. DOI 10.1007/978-1-4020-8274-0_5.
- Ladurner P., Schärer L., Salvenmoser W., Rieger R.M. A new model organism among the lower Bilateria and the use of digital microscopy in taxonomy of meiobenthic Platyhelminthes: *Macrostomum lignano*, n. sp. (Rhabditophora, Macrostromorpha). *J. Zool. Syst. Evol. Res.* 2005;43:114-126. DOI 10.1111/j.1439-0469.2005.00299.x.
- Laing S.T., Ivens A., Laing R., Ravikumar S., Butler V., Woods D.J., Gilleard J.S. Characterization of the xenobiotic response of *Caenorhabditis elegans* to the anthelmintic drug albendazole and the identification of novel drug glucoside metabolites. *Biochem. J.* 2010;432:505-516. DOI 10.1042/BJ20101346.
- Lengerer B., Wunderer J., Pjeta R., Carta G., Kao D., Aboobaker A., Beisel C., Berezikov E., Salvenmoser W., Ladurner P. Organ specific gene expression in the regenerating tail of *Macrostomum lignano*. *Dev. Biol.* 2018;433(2):448-460. DOI 10.1016/j.ydbio.2017.07.021.
- Leria L., Vila-Farré M., Solà E., Riutort M. Outstanding intraindividual genetic diversity in fissiparous planarians (*Dugesia*, Platyhelminthes) with facultative sex. *BMC Evol. Biol.* 2019;19:130. DOI 10.1186/s12862-019-1440-1.
- McCusker P., McVeigh P., Rathinasamy V., Toet H., McCammick E., O'Connor A., Marks N.J., Mousley A., Brennan G.P., Halton D.W., Spithill T.W., Maule A.G. Stimulating neoblast-like cell proliferation in juvenile *Fasciola hepatica* supports growth and progression towards the adult phenotype *in vitro*. *PLoS Negl. Trop. Dis.* 2016;10:e0004994. DOI 10.1371/journal.pntd.0004994.
- McGonigle L., Mousley A., Marks N.J., Brennan G.P., Dalton J.P., Spithill T.W., Day T.A., Maule A.G. The silencing of cysteine proteases in *Fasciola hepatica* newly excysted juveniles using RNA interference reduces gut penetration. *Int. J. Parasitol.* 2008;38:149-155. DOI 10.1016/j.ijpara.2007.10.007.
- Moguel B., Moreno-Mendoza N., Bobes R.J., Carrero J.C., Chimal-Monroy J., Diaz-Hernández M.E., Herrera-Estrella L., Lacleste J.P. Transient transgenesis of the tapeworm *Taenia crassiceps*. *SpringerPlus.* 2015;4:496. DOI 10.1186/s40064-015-1278-y.
- Morand S., Robert F., Connors V.A. Complexity in parasite life cycles: population biology of cestodes in fish. *J. Anim. Ecol.* 1995;64:256-264. DOI 10.2307/5760.
- Mouton S., Grudniewska M., Glazenburg L., Guryev V., Berezikov E. Resilience to aging in the regeneration-capable flatworm *Macrostomum lignano*. *Aging Cell.* 2018;17:e12739. DOI 10.1111/ace1.12739.
- Mwangi I.N., Sanchez M.C., Mkoji G.M., Agola L.E., Runo S.M., Cupit P.M., Cunningham C. Praziquantel sensitivity of Kenyan *Schistosoma mansoni* isolates and the generation of a laboratory strain with reduced susceptibility to the drug. *Int. J. Parasitol. Drugs Drug Resist.* 2014;4:296-300. DOI 10.1016/j.ijpdr.2014.09.006.
- Nimeth K., Ladurner P., Gschwenter R., Salvenmoser W., Rieger R. Cell renewal and apoptosis in *Macrostomum* sp. [*Lignano*]. *Cell Biol. Int.* 2002;26:801-815. DOI 10.1006/cbir.2002.0950.
- Pakharukova M.Y., Shilov A.G., Pirozhkova D.S., Katokhin A.V., Mordvinov V.A. The first comprehensive study of praziquantel effects *in vivo* and *in vitro* on European liver fluke *Opisthorchis felineus* (Trematoda). *Int. J. Antimicrob. Agents.* 2015;46:94-100. DOI 10.1016/j.ijantimicag.2015.02.012.
- Pareek A., Arora A., Dhankher O.P. Stepping forward and taking reverse as we move ahead in genetics. *Ind. J. Plant Physiol.* 2018;23:609-611. DOI 10.1007/s40502-018-0428-y.
- Park J.-K., Kim K.-H., Kang S., Kim W., Eom K.S., Littlewood D. A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). *BMC Evol. Biol.* 2007;7:11. DOI 10.1186/1471-2148-7-11.
- Pfister D., De Mulder K., Philipp I., Kuales G., Hrouda M., Eichberger P., Borgonie G., Hartenstein V., Ladurner P. The exceptional stem cell system of *Macrostomum lignano*: Screening for gene expression and studying cell proliferation by hydroxyurea treatment and irradiation. *Front. Zool.* 2007;4:9. DOI 10.1186/1742-9994-4-9.
- Pierson L., Mousley A., Devine L., Marks N.J., Day T.A., Maule A.G. RNA interference in a cestode reveals specific silencing of selected highly expressed gene transcripts. *Int. J. Parasitol.* 2010;40:605-615. DOI 10.1016/j.ijpara.2009.10.012.
- Pomaznoy M.Y., Logacheva M.D., Young N.D., Penin A.A., Ershov N.I., Katokhin A.V., Mordvinov V.A. Whole transcriptome profiling of adult and infective stages of the trematode *Opisthorchis felineus*. *Parasitol. Int.* 2016;65:12-19. DOI 10.1016/j.parint.2015.09.002.
- Potter C.J., Luo L. Splinkerette PCR for mapping transposable elements in *Drosophila*. *PLoS One.* 2010;5:e10168. DOI 10.1371/journal.pone.0010168.
- Poulin R., Cribb T.H. Trematode life cycles: Short is sweet? *Trends Parasitol.* 2002;18:176-183. DOI 10.1016/S1471-4922(02)02262-6.
- Richards C.S. Genetic studies on variation in infectivity of *Schistosoma mansoni*. *J. Parasitol.* 1975;61:233-236. DOI 10.2307/3278999.
- Rivera-Ingraham G.A., Bickmeyer U., Abele D. The physiological response of the marine platyhelminth *Macrostomum lignano* to different environmental oxygen concentrations. *J. Exp. Biol.* 2013;216:2741-2751. DOI 10.1242/jeb.081984.
- Rivera-Ingraham G.A., Nommick A., Blondeau-Bidet E., Ladurner P., Lignot J.-H. Salinity stress from the perspective of the energy-redox axis: Lessons from a marine intertidal flatworm. *Redox Biol.* 2016;10:53-64. DOI 10.1016/j.redox.2016.09.012.
- Rouhana L., Weiss J.A., Forsthoefel D.J., Lee H., King R.S., Inoue T., Shibata N., Agata K., Newmark P.A. RNA interference by feeding *in vitro*-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev. Dyn.* 2013;242:718-730. DOI 10.1002/dvdy.23950.
- Schärer L., Ladurner P. Phenotypically plastic adjustment of sex allocation in a simultaneous hermaphrodite. *Proc. Biol. Sci.* 2003;270:935-941. DOI 10.1098/rspb.2002.2323.
- Schwartz C., Fallon P.G. *Schistosoma* "eggs-iting" the host: granuloma formation and egg excretion. *Front. Immunol.* 2018;9. DOI 10.3389/fimmu.2018.02492.
- Siqueira L.D.P., Fontes D.A.F., Aguilera C.S.B., Timoteo T.R.R., Angelos M.A., Silva L.C.P.B.B., de Melo C.G., Rolim L.A., da Silva R.M.F., Neto P.J.R. Schistosomiasis: Drugs used and treatment strategies. *Acta Trop.* 2017;176:179-187. DOI 10.1016/j.actatropica.2017.08.002.
- Skromne I., Prince V.E. Current perspectives in zebrafish reverse genetics: moving forward. *Dev. Dyn.* 2008;237:861-882. DOI 10.1002/dvdy.21484.
- Song G., Li Q., Long Y., Gu Q., Hackett P.B., Cui Z. Effective gene trapping mediated by sleeping beauty transposon. *PLoS One.* 2012;7:e44123. DOI 10.1371/journal.pone.0044123.
- Stefano B., Patrizia B., Matteo C., Massimo G. Inverse PCR and quantitative PCR as alternative methods to southern blotting analysis to assess transgene copy number and characterize the integration site

- in transgenic woody plants. *Biochem. Genet.* 2016;54:291-305. DOI 10.1007/s10528-016-9719-z.
- Wagner D.E., Wang I.E., Reddien P.W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science.* 2011;332:811-816. DOI 10.1126/science.1203983.
- Waikagul J., Kobayashi J., Pongvongsa T., Sato M.O., Adsakwattana P., Fontanilla I.K.C., Sato M., Fornillos R.J.C. Odds, challenges and new approaches in the control of helminthiasis, an Asian study. *Parasite Epidemiol. Control.* 2018;4:e00083. DOI 10.1016/j.parepi.2018.e00083.
- Wang W., Wang L., Liang Y.-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. *Parasitol. Res.* 2012;111:1871-1877. DOI 10.1007/s00436-012-3151-z.
- Wasik K., Gurtowski J., Zhou X., Ramos O.M., Delás M.J., Battistoni G., Demerdash O.E., Falciatori L., Vizoso D.B., Smith A.D., Ladurner P., Schärer L., McCombie W.R., Hannon G.J., Schatz M. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl. Acad. Sci. USA.* 2015;112:12462-12467. DOI 10.1073/pnas.1516718112.
- Wongratanchewin S., Sermwan R.W., Sirisinha S. Immunology and molecular biology of *Opisthorchis viverrini* infection. *Acta Trop.* 2003;88:195-207. DOI 10.1016/j.actatropica.2003.02.002.
- Wudarski J., Egger B., Ramm S.A., Schärer L., Ladurner P., Zadesnets K.S., Rubtsov N.B., Mouton S., Berezikov E. The free-living flatworm *Macrostomum lignano*. *EvoDevo.* 2020;11:5. DOI 10.1186/s13227-020-00150-1.
- Wudarski J., Simanov D., Ustyantsev K., de Mulder K., Grelling M., Grudniewska M., Beltman F., Glazenburg L., Demircan T., Wunderer J., Qi W., Vizoso D.B., Weissert P.M., Olivieri D., Mouton S., Guryev V., Aboobaker A., Schärer L., Ladurner P., Berezikov E. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nat. Commun.* 2017;8:2120. DOI 10.1038/s41467-017-02214-8.
- Wudarski J., Ustyantsev K., Glazenburg L., Berezikov E. Influence of temperature on development, reproduction and regeneration in the flatworm model organism, *Macrostomum lignano*. *Zool. Lett.* 2019;5:7. DOI 10.1186/s40851-019-0122-6.
- Zheng H., Zhang W., Zhang L., Zhang Z., Li J., Lu G., Zhu Y., Wang Y., Huang Y., Liu J., Kang H., Chen J., Wang L., Chen A., Yu S., Gao Z., Jin L., Gu W., Wang Z., Zhao L., Shi B., Wen H., Lin R., Jones M.K., Brejova B., Vinar T., Zhao G., McManus D.P., Chen Z., Zhou Y., Wang S. The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat. Genet.* 2013;45:1168-1175. DOI 10.1038/ng.2757.

ORCID ID

K.V. Ustyantsev orcid.org/0000-0003-4346-3868
E.V. Berezikov orcid.org/0000-0002-1145-2884


Благодарности. Работа по сравнительному анализу характеристик *M. lignano*, планарий и ППЧ выполнена В.Ю.В., А.Г.Б. и Е.В.Б. при поддержке бюджетного проекта № 0259-2021-0009. Поиск и анализ гомологичных генов *M. lignano* и ППЧ, а также определение методов и перспективных генов-мишеней проведены К.В.У. в ИЦиГ СО РАН при финансовой поддержке гранта РНФ № 19-74-00029.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 17.10.2020. После доработки 03.12.2020. Принята к публикации 08.12.2020.

Английский текст <https://vavilov.elpub.ru/jour>


Трансгенная клеточная линия с индуцируемой транскрипцией для исследования механизмов экспансии (CGG)_n повторов

И.В. Грищенко¹, А.А. Тулупов^{2, 3}, Ю.М. Рымарева², Е.Д. Петровский², А.А. Савелов², А.М. Коростышевская², Ю.В. Максимова^{4, 5}, А.Р. Шорина⁵, Е.М. Шитик¹, Д.В. Юдкин¹ ¹ Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, р.п. Кольцово, Новосибирская область, Россия² Институт «Международный томографический центр» Сибирского отделения Российской академии наук, Новосибирск, Россия³ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия⁴ Новосибирский государственный медицинский университет, Новосибирск, Россия⁵ Городская клиническая больница № 1, Новосибирск, Россия yudkin_dv@vector.nsc.ru

Аннотация. Существует ряд наследственных заболеваний человека, причиной которых является экспансия тандемных повторов. К ним относятся миотоническая дистрофия первого типа, болезнь Хантингтона, заболевания, ассоциированные с ломкой X-хромосомой. Синдром ломкой X-хромосомы – наиболее распространенная причина наследственной умственной отсталости у человека. На сегодняшний день причины развития экспансии остаются неисследованными. Важная особенность протяженных повторов – их способность формировать альтернативные вторичные структуры ДНК. Существуют гипотезы, объясняющие природу нестабильности повторов, однако все они предполагают возникновение устойчивых вторичных структур ДНК на различных этапах клеточного цикла. Источником нестабильности считаются нарушения в различных процессах метаболизма ДНК (репликация, репарация и рекомбинация), вызванные образованием вторичных структур. Однако ни одна из гипотез до конца не подтверждена и, видимо, не является единственно верной. Вероятно, в различных типах клеток и на определенных стадиях клеточного цикла источником нестабильности выступает множество процессов. В настоящей работе мы предлагаем экспериментальную систему для изучения вклада транскрипции и ассоциированной с ней репарации в нестабильность повтора (CGG)_n, поскольку это наименее изученный механизм возникновения нестабильности. Однако предложенные модели могут учитывать вклад и других процессов метаболизма ДНК, например репликации, что делает полученные системы универсальными и применимыми в разных исследованиях. Нами были созданы трансгенные клеточные линии, несущие повтор нормальной и премутантной длины под тетрациклин-индуцируемым промотором. Один тип линий содержит плазмиду с экзогенным повтором, интегрированным в геном посредством транспозона Sleeping Beauty, в другой клеточной линии вектор поддерживается в виде эписомы благодаря оридгину репликации SV40. Такие трансгенные клеточные линии могут служить экспериментальной системой для поиска причин нестабильности и создания терапевтических средств. Кроме того, был разработан критерий для оценки нестабильности экзогенного (CGG)_n повтора в геноме трансгенных клеточных линий, расчет которого не зависит от эффективности синтеза протяженных повторов. Ключевые слова: наследственная умственная отсталость; синдром ломкой X-хромосомы; экспансия повторов; транскрипция; репликация; трансгенная клеточная линия; соматическая нестабильность.

Для цитирования: Грищенко И.В., Тулупов А.А., Рымарева Ю.М., Петровский Е.Д., Савелов А.А., Коростышевская А.М., Максимова Ю.В., Шорина А.Р., Шитик Е.М., Юдкин Д.В. Трансгенная клеточная линия с индуцируемой транскрипцией для исследования механизмов экспансии (CGG)_n повторов. *Вавиловский журнал генетики и селекции*. 2021;25(1): 117-124. DOI 10.18699/VJ21.014

A transgenic cell line with inducible transcription for studying (CGG)_n repeat expansion mechanisms

I.V. Grishchenko¹, A.A. Tulupov^{2, 3}, Y.M. Rymareva², E.D. Petrovskiy², A.A. Savelov², A.M. Korostyshevskaya², Y.V. Maksimova^{4, 5}, A.R. Shorina⁵, E.M. Shitik¹, D.V. Yudkin¹ ¹ State Research Center of Virology and Biotechnology "Vector", Rospotrebnadzor, Koltsovo, Novosibirsk region, Russia² International Tomography Center of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia³ Novosibirsk State University, Novosibirsk, Russia⁴ Novosibirsk State Medical University, Novosibirsk, Russia⁵ Novosibirsk City Clinical Hospital No. 1, Novosibirsk, Russia yudkin_dv@vector.nsc.ru

Abstract. There are more than 30 inherited human disorders connected with repeat expansion (myotonic dystrophy type I, Huntington's disease, Fragile X syndrome). Fragile X syndrome is the most common reason for inherited intellectual disability in the human population. The ways of the expansion development remain unclear. An important feature

of expanded repeats is the ability to form stable alternative DNA secondary structures. There are hypotheses about the nature of repeat instability. It is proposed that these DNA secondary structures can block various stages of DNA metabolism processes, such as replication, repair and recombination and it is considered as the source of repeat instability. However, none of the hypotheses is fully confirmed or is the only valid one. Here, an experimental system for studying (CGG)_n repeat expansion associated with transcription and TCR-NER is proposed. It is noteworthy that the aberrations of transcription are a quiet mechanism of (CGG)_n instability. However, the proposed systems take into account the contribution of other processes of DNA metabolism and, therefore, the developed systems are universal and applicable for various studies. Transgenic cell lines carrying a repeat of normal or premutant length under the control of an inducible promoter were established and a method for repeat instability quantification was developed. One type of the cell lines contains an exogenous repeat integrated into the genome by the Sleeping Beauty transposon; in another cell line, the vector is maintained as an episome due to the SV40 origin of replication. These experimental systems can serve for finding the causes of instability and the development of therapeutic agents. In addition, a criterion was developed for the quantification of exogenous (CGG)_n repeat instability in the transgenic cell lines' genome.

Key words: hereditary intellectual disability; fragile X syndrome; repeat expansion; transcription; replication; transgenic cell lines; somatic instability.

For citation: Grishchenko I.V., Tulupov A.A., Rymareva Y.M., Petrovskiy E.D., Savelov A.A., Korostyshevskaya A.M., Maksimova Y.V., Shorina A.R., Shitik E.M., Yudkin D.V. A transgenic cell line with inducible transcription for studying (CGG)_n repeat expansion mechanisms. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2021;25(1): 117-124. DOI 10.18699/VJ21.014

Введение

Экспансия повторов – это особый тип мутаций, для которого характерно быстрое увеличение количества тандемных повторов в ДНК. В большей степени к экспансии склонны триплетные повторы, на сегодняшний день известно более 30 заболеваний, ассоциированных с их нестабильностью (Grishchenko et al., 2020). Одной из таких патологий является синдром ломкой X-хромосомы – наиболее распространенная форма наследственной умственной отсталости. Причиной заболевания выступает экспансия CGG-повтора, расположенного в 5'-нетранслируемой области гена *FMR1*. В норме количество повторов относительно стабильно и не превышает 54 триплетов, при экспансии повторов до 200 триплетов аллель становится премутантным и развиваются синдромы атаксии/тремора и синдром первичной овариальной недостаточности, ассоциированные с ломкой X-хромосомой. Премутантный аллель встречается в популяции с частотой 1:100. Несмотря на то что клинических проявлений зачастую не наблюдается, удлиненная версия повтора способна передаваться в ряду поколений. Размер повтора более 200 триплетов считается полной мутацией: промотор гена *FMR1* метилируется, происходят гетерохроматинизация локуса и полная потеря экспрессии белка FMRP, что приводит к развитию синдрома ломкой X-хромосомы. Этот белок необходим для нормального функционирования нейронов, и его отсутствие вызывает ярко выраженные фенотипические проявления: макроорхизм, эндокринные патологии, морфологические изменения мозжечка, умственную отсталость, проблемы с поведением, обучением (Roberts et al., 2003; Martin et al., 2012; Heulens et al., 2013). Частота полной мутации варьирует от 1:4000 у мужчин и до 1:6000 у женщин.

Несмотря на понимание деталей патогенеза синдрома, механизм возникновения экспансии до сих пор не изучен. Вероятно, различные процессы метаболизма ДНК в клетке способны усиливать нестабильность триплетных повторов. Так, установлен вклад репликации в экспансию повторов: образовавшаяся шпилька на вновь синтезированной цепи ДНК приводит к повторной репликации участка, содержащего последовательность (CGG)_n и, следовательно, ее увеличению (Fouche et al., 2006). Однако

у людей, страдающих заболеваниями, связанными с экспансией повторов, и у модельных линий мышей экспансия часто наблюдается в тканях с низкой пролиферативной активностью, включая мозг, ооциты, печень и мышцы (Lokanga et al., 2013), что подтверждает теорию о зависимости экспансии от других процессов, затрагивающих ДНК. Действительно, для многих белков, участвующих в репарации и рекомбинации ДНК, показано их вероятное участие в экспансии повторов. Есть экспериментальные данные, свидетельствующие об участии белков MMR репарации в процессе увеличения триплетных повторов (Kovalenko et al., 2012; Zhao et al., 2016). Другим возможным источником нестабильности могут служить транскрипция и ассоциированная с ней репарация, поскольку для многих трактов повторов характерны образование R-петель – РНК: ДНК-устойчивых дуплексов во время синтеза РНК, а также нарушение инициации транскрипции PolII (Krasilnikova et al., 2007). Возникающие повреждения иницируют TCR – вариант эксцизионной репарации нуклеотидов. Для некоторых белков этого каскада выявлены корреляции с уровнем нестабильности повтора (CGG)_n. Нужно отметить, что для премутантных аллелей гена *FMR1*, быстро накапливающих повторенные единицы, обнаружено значительное увеличение уровня транскрипции, что, вероятно, указывает на вовлеченность системы TCR в развитие экспансии и увеличение геномной нестабильности, однако однозначных подтверждений этой гипотезы пока нет.

Для изучения взаимосвязи всех описанных каскадов необходимо иметь модель, в которой можно отслеживать все изменения, происходящие с повтором и окружающими регионами в ответ на индукцию определенного процесса метаболизма ДНК. Подобные модели уже были предложены (Gorbunova et al., 2003; Kononenko et al., 2020), однако ни в одной из них нельзя непосредственно оценить вклад транскрипции в нестабильность (CGG)_n. В настоящей статье мы описываем разработанные трансгенные клеточные линии для изучения механизма экспансии на основе двух типов плазмид: интегрированных и не интегрированных в геном. Предложенные модели позволят учесть вклад репликации, транскрипции, TCR-NER и расположение в геноме в нестабильность CGG-повтора.

Кроме того, полученные клеточные линии можно использовать для изучения повтор-индуцируемого мутагенеза, наблюдающегося в клетках с увеличенным размером повтора (Shah, Mirkin, 2015).

Материалы и методы

Информированное согласие на участие в исследовании и соответствие этическим нормам. Процедура включения пациентов в исследование разработана в строгом соответствии с международными стандартами, включающими осведомленность субъекта, его согласие на участие в эксперименте и гарантии конфиденциальности. Все исследования проводили с учетом этических стандартов, предусмотренных Хельсинкской декларацией Всемирной медицинской ассоциации с поправками, внесенными в 2000 г. Было получено письменное согласие участников исследования.

Выделение ДНК и определение размера CGG-повтора. Периферическая венозная кровь от каждого пациента была собрана в пробирки с ЭДТА в Новосибирской городской клинической больнице № 1 и заморожена при температуре -70°C . Тотальную ДНК очищали набором Wizard® Genomic DNA Purification Kit (Promega, США). Аналогично была получена ДНК из культур клеток.

Последовательности с высоким ГЦ-составом амплифицировали по методике, описанной в статье В.Е. Наувард с коллегами (2016). Для амплификации использовали праймеры NewFraxC (5'-d6RG-tgctttctagactcagctccgtttcggtttctactccggt-3') и NewFraxR4 (5'-taagcagaatccctgtagaaagcgcattggagcccccga-3') и 0.02 ед. акт./мкл Q5-ДНК-полимеразы. Размер полученных продуктов определяли с помощью электрофореза в агарозном геле. Для оценки точного размера повтора проводили капиллярный электрофорез со стандартом длин 1200LIZ (Applied Biosystems, США). Поскольку область, фланкирующая повтор, в ПЦР-продукте составляет в общей сложности 269 п. н., длину повтора рассчитывали по формуле

$$N = \frac{\text{Размер полученного ПЦР-продукта} - 269}{3},$$

где N – количество триплетов.

Клонирование CGG-повтора различной длины в векторные системы. Контрольная плаزمида pCDH, не содержащая повтора (CGG)_n, состояла из следующих элементов: (1) индуцируемый доксициклином промотор Tet-O-minimal CMV, последовательность IRES, открытая рамка считывания белка GFP, (2) конститутивный промотор EF1alpha, открытая рамка считывания трансактиватор для Tet-O-элемента rTta, T2A пептид, открытая рамка считывания белка DsRedExpress, (3) промотор бета-лактамазы, открытая рамка считывания белка бета-лактамазы для селекции трансформированных бактериальных клеток, ориджин репликации, (4) ориджин репликации вируса SV40. ПЦР-продукт, несущий CGG-повтор, был клонирован в плазмиду pCDH по сайтам эндонуклеаз рестрикции XbaI и EcoRI («СибЭнзим», Россия) между минимальным промотором CMV и последовательностью IRES.

Плазмиды pSBi для клонирования CGG-повтора были собраны из компонентов: (1) промотор бета-лактамазы, открытая рамка считывания белка бета-лактамазы для селекции трансформированных бактериальных клеток,

ориджин репликации, (2) концевые повторы транспозона Sleeping Beauty, (3) кассета, содержащая промотор PGK и ORF пуриноцилин-N-ацетил трансферазы, (4) промотор hPGK, ORF rTta, (5) индуцируемый промотор TRE3GS, ORF mGFP.

ПЦР-продукт с повтором под контроль индуцируемого промотора клонировали по сайтам эндонуклеаз рестрикции XbaI и EcoRI («СибЭнзим»). Для наработки плазмид осуществляли трансформацию электрокомпетентных клеток *E. coli* штамма NebStable (NEB, США). Показано, что протяженный повтор при трансформации бактериальных клеток и культивировании склонен к резкому сокращению, что согласуется с литературными данными (Bontekoe, 2001), поэтому клетки культивировали в течение суток при температуре 20°C , чтобы избежать уменьшения размера повтора. Для трансфекции клеток HEK293A и HEK293T плазмиды выделяли и очищали с помощью набора QIAGEN® Plasmid Plus Maxi Kit (QIAGEN).

Трансфекция клеток эукариот. Клетки линий HEK293A и HEK293T трансфицировали с помощью реагента Lipofectamine 3000 (Thermo Fisher Scientific, США). Индукция промотора Tet-O-minimal CMV и TRE3GS происходила при добавлении в культуральную среду антибиотика доксициклина до концентрации 1 мкг/мл.

Результаты

Сборка модельных плазмидных конструкций, несущих повтор нормальной и премутантной длины

Получены плазмиды на основе векторов эукариотической экспрессии с индуцируемым промотором, который регулирует уровень транскрипции CGG-повтора нормальной или премутантной длины и ORF зеленого флуоресцентного белка. Эти плазмиды служат основой модельной системы для изучения нестабильности повтора (CGG)_n. В качестве вектора для транзientной экспрессии и поддержания в неинтегрированном в геном состоянии использована плазмиды pCDH (рис. 1, а). Для интеграции в геном была собрана конструкция на основе системы транспозон/транспозаза Sleeping Beauty pSBi (см. рис. 1, б).

Вектор pCDH кодирует два репортерных белка: DsRedExpress под контролем промотора EF1 и EGFP, экспрессия которого регулируется индуцируемым промотором Tet-O-CMV. После промотора Tet-O-CMV расположен множественный сайт клонирования, в который встраивается фрагмент, содержащий повтор. За счет такого взаимного расположения индуцируемого промотора и места встройки повтора можно установить, что транскрипция идет через встроенный повтор, детектируя синтез EGFP. Спустя несколько раундов транскрипции можно судить о влиянии процесса на экспансию повтора. Кроме того, в плазмиде pCDH находится ориджин репликации вируса SV40, таким образом, конструкция способна реплицироваться в клетках HEK293T, продуцирующих большой Т-антиген вируса SV40. В этом случае можно оценить вклад не только транскрипции, но и репликации во время поддержания плазмиды в клетке в виде эписомы.

Вектор pSBi кодирует белок mGFP, находящийся под контролем индуцируемого промотора TRE3GS. Перед ORF mGFP расположены сайты для клонирования CGG-

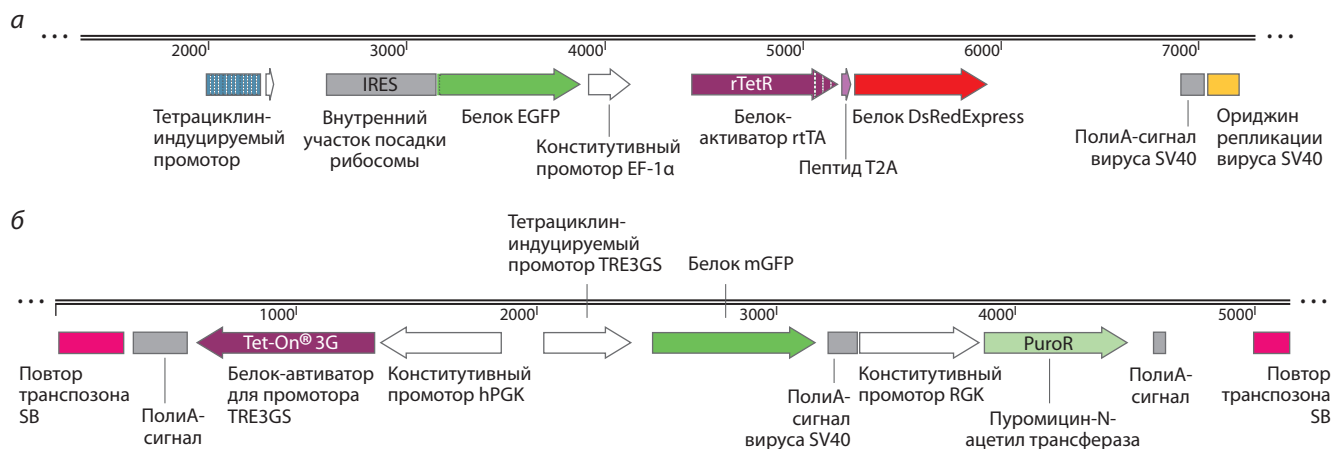


Рис. 1. Карты векторов, использованных для получения модельных клеточных линий.

a – карта плазмиды pCDH. IRES – внутренний участок посадки рибосомы; EGFP – зеленый флуоресцентный белок; EF1 – конститутивный промотор; rtTA – трансаkтиватор, взаимодействующий с тетрациклином/доксикациклином; DsRedExpress – красный флуоресцентный белок; *б* – карта плазмиды pSBi. Повтор транспозона SB – последовательность для взаимодействия с транспозазой Sleeping Beauty; Tet-On® 3G – трансаkтиватор, взаимодействующий с тетрациклином/доксикациклином.

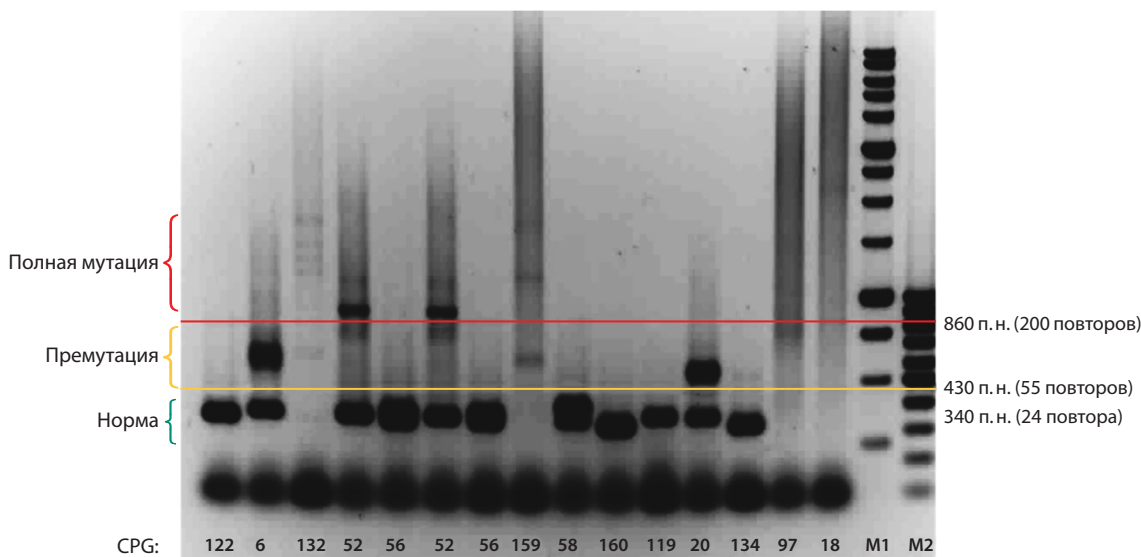


Рис. 2. Пример результатов амплификации ГЦ-богатых матриц.

Номера под рисунком – образцы ДНК, полученные от пациентов CPG; M1 – маркер молекулярного веса 1 т.п.н.; M2 – маркер молекулярного веса 100 п. н.

повтора. Таким образом, можно провести анализ влияния транскрипции на изменения CGG-повтора. Поскольку вектор pSBi основан на транспозоне, часть плазмиды, ограниченная специфическими повторенными последовательностями, может быть встроена в различные области генома. При определении места встройки можно оценить влияние мест интеграции на нестабильность CGG-повтора.

Для синтеза фрагментов, содержащих повтор, мы использовали образцы ДНК, выделенные из постоянных культур В-лимфоцитов и цельной крови пациентов с синдромом ломкой X-хромосомы (рис. 2).

Для создания модельных конструкций было решено использовать только повторы нормальной и премутантной длины. Ожидается, что нестабильность этих видов повто-

ров в создаваемой молекулярной модели будет резко отличаться, поскольку премутантный аллель – наиболее нестабильный, при этом нормальный аллель склонен лишь к незначительному полиморфизму (Lokanga et al., 2013).

Было собрано пять вариантов плазмид, несущих 5 (pCDH-5), 25 (pSBi-25), 59 (pCDH59), 85 (pCDH85) и 160 повторов (pSBi-160). Структуры всех плазмид подтверждены секвенированием по Сэнгеру (рис. 3).

Исследование работоспособности полученных модельных плазмид

Проведена оценка эффективности трансфекции эукариотических клеток собранными конструкциями для подтверждения отсутствия нарушений экспрессии репортерных белков в присутствии протяженного повтора (CGG)n.

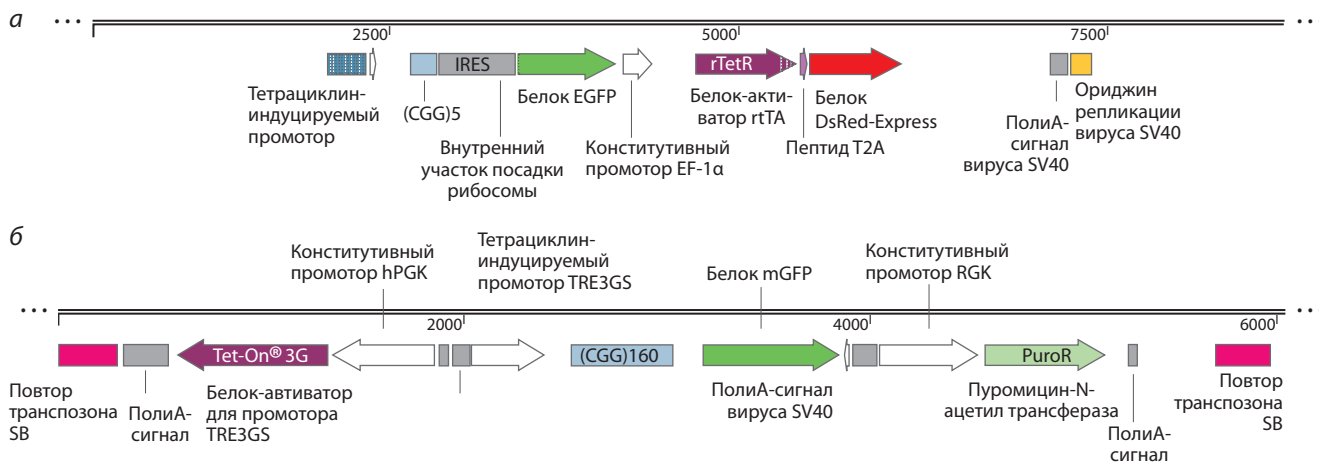


Рис. 3. Карты векторов rCDH и pSBi после клонирования повтора (CGG)_n.
 а – пример схемы плазмиды rCDH; б – pSBi с клонированным CGG-повтором.

Показано, что трансфекция плазмидами, несущими повтор нормальной или премутантной длины, происходит с такой же эффективностью, как и трансфекция контрольными плазмидами, не содержащими повторенную последовательность.

При трансфекции плазмидами rCDH отмечена экспрессия DsRedExpress, после трансфекции векторами pSBi была подтверждена возможность проведения селекции на среде, содержащей пурамицин. После проверки эффективности трансфекции исследовали способность векторов к спонтанной индукции промоторов с элементом Tet-O, регулирующих экспрессию зеленого флуоресцентного белка. Важно, чтобы в трансфицированных клетках не было высокого уровня фоновой экспрессии зеленого белка, поскольку она мешает оценить точное влияние транскрипции и репарации, ассоциированной с транскрипцией, на увеличение размера повтора. Без индукции

промотора в клетках, трансфицированных векторами на основе rCDH, отмечены активная экспрессия красного белка и отсутствие зеленого белка. Для индукции к клеткам ежедневно добавляли доксициклин, в результате чего был обнаружен высокий уровень флуоресценции зеленого белка (рис. 4, а). При использовании плазмиды pSBi, содержащей промотор TRE3GS, фоновая индукция не выявлена, что позволяет проводить селекцию стабильных трансформантов на пурамицине и избежать влияния фонового уровня транскрипции на встроенный повтор (см. рис. 4, б).

Разработка метода анализа нестабильности повтора в модели

Ожидается, что экспансия в клетках трансгенной клеточной линии будет происходить с разной скоростью, что при длительном культивировании может привести к появлению

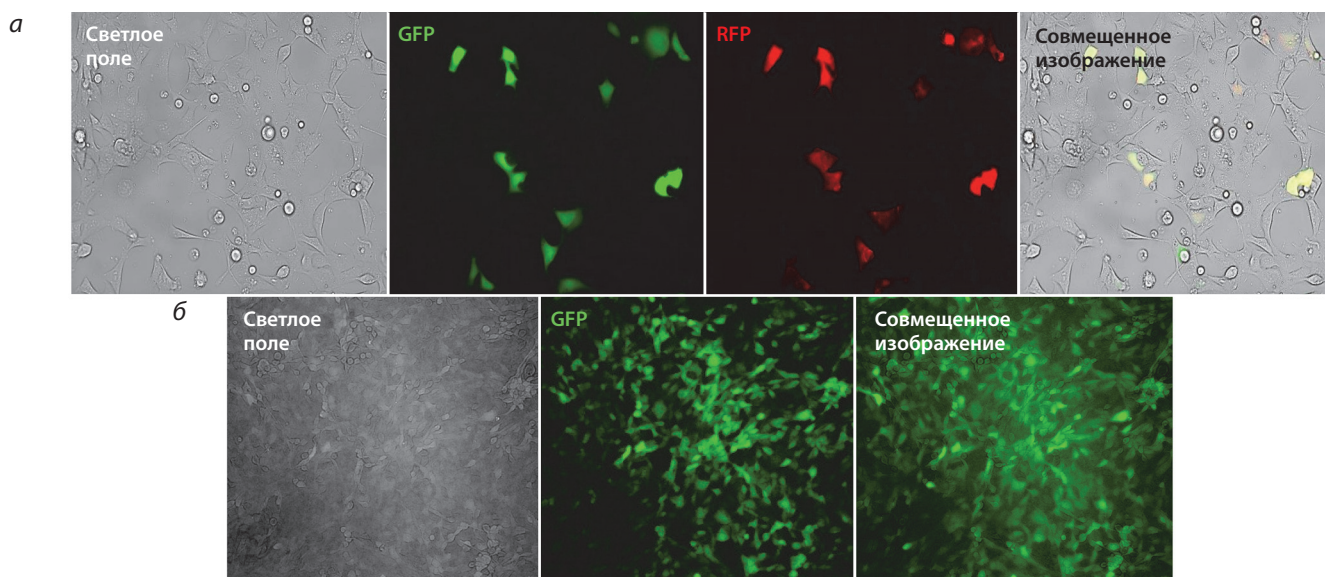


Рис. 4. Индукция тетрациклин-зависимых промоторов в разработанных плазмидах.
 а – результат индукции промотора с Tet-O-элементом в векторе rCDH с помощью доксициклина в клетках HEK293T; б – результат индукции промотора с TRE3GS в векторе pSBi с помощью доксициклина после селекции клеток HEK293A на пурамицине.

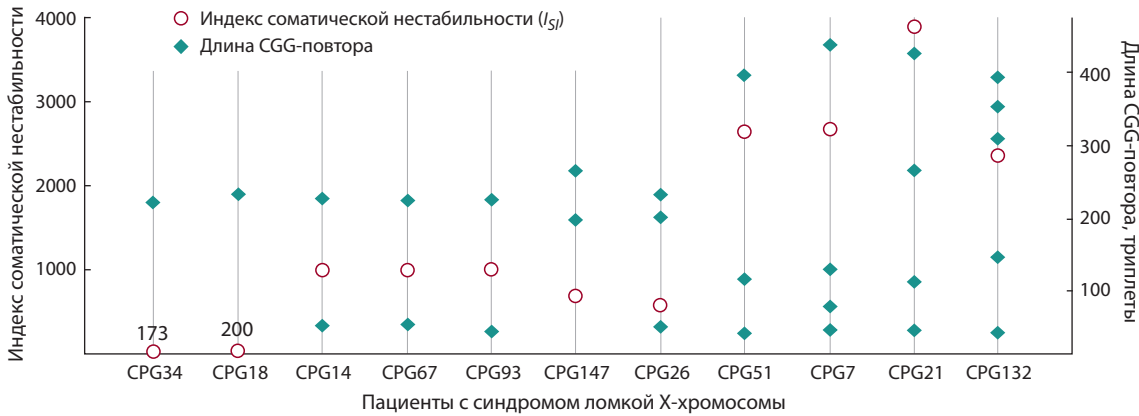


Рис. 5. Длина повтора и индексы соматической нестабильности у пациентов с синдромом ломкой X-хромосомы. Значения над маркерами – индексы соматической нестабильности.

нию мозаицизма по размеру экзогенного CGG-повтора. В связи с этим необходимо использовать критерий оценки, который позволит проводить сравнение мозаичных клеточных линий. Ранее были предложены подходы для оценки соматической нестабильности тринуклеотидных повторов у пациентов с экспансионными заболеваниями. Например, метод оценки нестабильности (CAG)n повторов при болезни Гентингтона основан на определении главного аллеля по максимальному и дополнительным пикам, детектируемым во фрагментном анализе, с последующей нормализацией на суммарные значения высот всех пиков (Lee et al., 2010). Другой метод оценки нестабильности повторов предполагает получение серийных разведений матрицы с последующей ПЦР (Monckton et al., 1995; Morales et al., 2012).

Амплификация при разведении позволяет обнаружить мозаицизм, который невозможно детектировать при обычной ПЦР из-за низкой эффективности синтеза менее представленных или очень крупных аллелей. Эти методы плохо применимы для оценки нестабильности (CGG)n повторов, так как амплификация крупного аллеля происходит с гораздо меньшей эффективностью, чем амплификация короткого аллеля (Usdin, Woodford, 1995; Woodford et al., 1995; Jensen et al., 2010). Для количественной оценки нестабильности (CGG)n повторов нами предложен метод анализа, использующий расчет индекса соматической нестабильности (I_{SI}) после проведения амплификации ГЦ-богатых матриц, по В.Е.Науward с коллегами (2016). Это значение позволяет учитывать не только размер повтора, но и разброс значений между аллелями, вне зависимости от эффективности их синтеза. Для (CGG)n повторов, расположенных в гене *FMRI*, мы предлагаем расчет I_{SI} по формуле

$$I_{SI} = Me \cdot (N_{max} - Me),$$

где Me – медиана, N_{max} – максимальная длина (CGG)n повтора в образце.

Медиана – это значение, учитывающее множество аллелей и разделяющее данные на две половины. Оно отражает неоднородность выборки и не чувствительно к слишком длинной или слишком короткой длине повтора, в отличие от среднего арифметического. При использовании среднего арифметического в расчете индекса вклад

более крупных аллелей будет иметь больший вес, чем вклад более коротких. Вследствие этого клеточные линии, имеющие разную степень нестабильности экзогенного повтора, будут иметь близкие значения величины соматической нестабильности, что приведет к ложной интерпретации результатов.

Значение ($N_{max} - Me$) учитывает разнообразие и разброс значений в образцах, где большие значения Me указывают на большую медианную длину повтора. В расчете I_{SI} не используется количество ПЦР продукта (высота пика) для каждого аллеля, т. е. эффективность ПЦР не влияет на конечное значение. Для определения индекса соматической нестабильности была использована ДНК 11 пациентов с синдромом ломкой X-хромосомы, выделенная из цельной крови, которая служила исходным материалом для синтеза протяженных повторов (CGG)n (рис. 5).

Как видно из расчета, значение индекса I_{SI} растет при увеличении количества и разброса значений повторов. Необходимо отметить, что метод анализа нестабильности работает для двух или более аллелей у пациентов с мозаицизмом. В случае одного аллеля индекс соматической нестабильности принимается равным размеру CGG-повтора, поскольку у пациента с одним аллелем $N_{max} - Me = 0$. Мы не можем принять $I_{SI} = 0$, потому что повтор (CGG)n нестабилен по своей природе.

Обсуждение

Синдром ломкой X-хромосомы – одна из самых распространенных причин наследственной умственной отсталости (Yudkin et al., 2015). Частота полной мутации в популяции изменяется от 1:6000 у женщин до 1:4000 у мужчин, при этом премутантный аллель, самый нестабильный вариант промоторной области гена *FMRI*, встречается в 1:100 случаев. Нестабильность CGG-повтора выражается в его склонности к экспансии – многократному и быстрому увеличению длины тракта этой повторенной последовательности. Кроме экспансии, в клетках пациентов и тканях модельных животных наблюдается сокращение повтора, что приводит к соматическому мозаицизму, степень которого коррелирует с тяжестью симптомов (Mailick et al., 2018). Однако вероятность экспансии в 10 раз выше (Bontekoe, 2001; DeJesus-Hernandez et al., 2011), что может

служить причиной усиления проявления симптомов заболевания при передаче в ряду поколений.

Существует ряд гипотез, объясняющих механизм экспансии, которые пока не подкреплены достаточным количеством экспериментальных данных. Все гипотезы предполагают формирование альтернативных вторичных структур ДНК на определенном участке во время процессов метаболизма ДНК и нарушение этих процессов. В экспериментах *in vitro* и *in vivo* показано формирование альтернативных вторичных структур ДНК, таких как шпильки, R-петли, G-квадруплексы (Usdin, Woodford, 1995; Groh et al., 2014; Lam et al., 2014). Возникновение таких структур может значительно нарушить указанные процессы метаболизма ДНК, что в свою очередь влияет на нестабильность повторов. Одна из возможных причин связана с проскальзыванием нити ДНК во время репликации (Pearson, Sinden, 1996; Fouche et al., 2006). Установлено, что проскальзывание нитей ДНК может происходить в различных случаях: при репликации ДНК в делящихся клетках, как было предположено ранее, а также при процессах репарации. Однако эта модель не может объяснить, почему не все повторы подвергаются экспансии и почему пороговое значение длины повторенной последовательности сходно для различных заболеваний. Есть свидетельства вклада некоторых белков репарационных каскадов, рекомбинации и транскрипции в нестабильность повторов. Однако все гипотезы имеют определенные недостатки и противоречия, поэтому необходимо продолжать поиски молекулярного механизма экспансии повтора.

Удобной экспериментальной системой служит модель экспансии, основанная на трансгенной клеточной линии с экзогенным повтором (CGG)_n, в которой можно отследить изменения повтора в ответ на индукцию всех каскадов метаболизма ДНК. Разработанные нами конструкции являются такой моделью и позволяют оценить вклад репликации, транскрипции и репарации в клетке. Нами собраны плазмиды двух типов: плазмиды с ориджином вируса SV40, способные к репликации в культурах клеток, экспрессирующих SV40 Т антиген, а также векторные системы, представляющие собой модифицированный транспозон Sleeping Beauty для интеграции кассеты с повтором и репортерными белками в различные локусы генома. Эффективность трансфекции и уровень начальной экспрессии белков были сопоставимы с такими же показателями контрольной плазмиды, не содержащей CGG-повтор. Методом сортирования, предельных разведений или селективным отбором можно получить культуру клеток, обладающих одним генотипом. Изменения в длине экзогенного повтора и, следовательно, мозаицизм, который будет формироваться в культуре со временем, можно оценить с помощью разработанного индекса I_{SY} . Этот метод оценки нестабильности удобен в использовании и отражает взаимосвязь нестабильности повтора и фенотипических проявлений, наблюдаемых в головном мозге пациентов.

В созданной системе можно оценивать непосредственно уровень экспансии, а также изменения, вызванные нестабильностью. Конструкция дает возможность детектировать изменение длины экзогенного повтора (CGG)_n в

разных локусах генома как при длительном культивировании, так и при небольшом времени поддержания повтора в культуре. Кроме того, можно отслеживать уровень нестабильности повтора во время активной транскрипции или без ее индукции. Измерение уровней экспрессии флуоресцентных белков может помочь отследить увеличение нестабильности и накопление мутаций, опосредованных повтор-индуцируемым мутагенезом. Проведение иммунопреципитации хроматина трансформированных клеток со специфическими антителами позволило бы определить вклад конкретных белков из различных каскадов в развитие нестабильности. Кроме того, уровень нестабильности в созданных клеточных моделях экспансии можно оценивать предложенным индексом соматической нестабильности. Мы предполагаем, что этот индекс должен также иметь биологический смысл, т. е. отражать степень фенотипических изменений у пациентов с заболеваниями, ассоциированными с ломкой X-хромосомой. Для проверки этой гипотезы начато исследование зависимостей значений I_{SY} у пациентов с изменениями в головном мозге по данным функциональной магнитно-резонансной томографии. Предварительные результаты указывают на наличие определенных корреляций, однако требуются дополнительные исследования.

Заключение

В настоящее время механизм нестабильности тринуклеотидных повторов остается до конца не изученным. При этом эта область исследований очень актуальна в силу того, что заболевания, вызванные этой мутацией, являются социально значимыми. Для поиска причин нестабильности повторов необходимо разрабатывать такие клеточные модели, в которых есть возможность отслеживать все изменения, вызванные экспансией, а также обнаружить влияние различных белков и путей метаболизма ДНК на этот процесс. Созданные в нашей работе конструкции для оценки нестабильности могут быть использованы в таких исследованиях.

Различные клеточные линии могут быть трансфицированы собранными векторами. Нами проверена работоспособность конструкций в двух клеточных линиях, HEK293A и HEK293T. После трансфекции клеток и индукции экспрессии на различных пассажах можно провести точное определение размера повтора (CGG)_n, а также других параметров и показать наличие или отсутствие экспансии повтора, в зависимости от его исходной длины и числа пассажей. Наша модель в дальнейшем послужит комплексному изучению всех аспектов нестабильности повторов в геноме человека и поможет сформировать более полное понимание механизмов этой мутации.

Список литературы / References

- Bontekoe C.J.M. Instability of a (CGG)₉₈ repeat in the Fmr1 promoter. *Hum. Mol. Genet.* 2001;10(16):1693-1699. DOI 10.1093/hmg/10.16.1693.
- DeJesus-Hernandez M., Mackenzie I.R., Boeve B.F., Boxer A.L., Baker M., Rutherford N.J., Nicholson A.M., Finch N.A., Flynn H., Adamson J., Kouri N., Wojtas A., Sengdy P., Hsiung G.Y.R., Karydas A., Seeley W.W., Josephs K.A., Coppola G., Geschwind D.H., Wszolek Z.K., Feldman H., Knopman D.S., Petersen R.C., Miller B.L., Dickson D.W., Boylan K.B., Graff-Radford N.R., Rade-

- makers R. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*. 2011;72(2):245-256. DOI 10.1016/j.neuron.2011.09.011.
- Fouche N., Ozgur S., Roy D., Griffith J.D. Replication fork regression in repetitive DNAs. *Nucleic Acids Res.* 2006;34(20):6044-6050. DOI 10.1093/nar/gkl1757.
- Gorbunova V., Seluanov A., Dion V., Sandor Z., Meservy J.L., Wilson J.H. Selectable system for monitoring the instability of CTG/CAG triplet repeats in mammalian cells. *Mol. Cell. Biol.* 2003; 23(13):4485-4493. DOI 10.1128/mcb.23.13.4485-4493.2003.
- Grishchenko I.V., Purvinsh Y.V., Yudkin D.V. Mystery of expansion: DNA metabolism and unstable repeats. In: Zharkov D.O. (Ed.). *Mechanisms of Genome Protection and Repair*. Cham: Springer International Publishing, 2020;101-124. DOI 10.1007/978-3-030-41283-8_7.
- Groh M., Lufino M.M.P., Wade-Martins R., Gromak N. R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. *PLoS Genet.* 2014;10(5): e1004318. DOI 10.1371/journal.pgen.1004318.
- Hayward B.E., Zhou Y., Kumari D., Usdin K. A Set of assays for the comprehensive analysis of fMRI alleles in the Fragile X-related disorders. *J. Mol. Diagn.* 2016;18(5):762-774. DOI 10.1016/j.jmoldx.2016.06.001.
- Heulens I., Suttie M., Postnov A., De Clerck N., Perrotta C.S., Mattina T., Faravelli F., Forzano F., Kooy R.F., Hammond P. Craniofacial characteristics of fragile X syndrome in mouse and man. *Eur. J. Hum. Genet.* 2013;21(8):816-823. DOI 10.1038/ejhg.2012.265.
- Jensen M.A., Fukushima M., Davis R.W. DMSO and betaine greatly improve amplification of GC-rich constructs in *de novo* synthesis. *PLoS One.* 2010;5:e11024. DOI 10.1371/journal.pone.0011024.
- Kononenko A.V., Ebersole T., Mirkin S.M. Experimental system to study instability of (CGG)_n repeats in cultured mammalian cells. In: Richard G.-F. (Ed.). *Trinucleotide Repeats: Methods and Protocols*. New York: Springer, 2020;137-150. DOI 10.1007/978-1-4939-9784-8_9.
- Kovalenko M., Dragileva E., St Claire J., Gillis T., Guide J.R., New J., Dong H., Kucherlapati R., Kucherlapati M.H., Ehrlich M.E., Lee J.M., Wheeler V.C. *Msh2* acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. *PLoS One.* 2012;7(9): e44273. DOI 10.1371/journal.pone.0044273.
- Krasilnikova M.M., Kireeva M.L., Petrovic V., Knijnikova N., Kshlev M., Mirkin S.M. Effects of Friedreich's ataxia (GAA)_n*(TTC)_n repeats on RNA synthesis and stability. *Nucleic Acids Res.* 2007; 35(4):1075-1084. DOI 10.1093/nar/gkl1140.
- Lam E.Y.N., Beraldi D., Tannahill D., Balasubramanian S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* 2014;4(1)1-8. DOI 10.1038/ncomms2792.
- Lee J.M., Zhang J., Su A.I., Walker J.R., Wiltshire T., Kang K., Dragileva E., Gillis T., Lopez E.T., Boily M.J., Cyr M., Kohane I., Gussella J.F., MacDonald M.E., Wheeler V.C. HA novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst. Biol.* 2010;4(1):29. DOI 10.1186/1752-0509-4-29.
- Lokanga R.A., Entezam A., Kumari D., Yudkin D., Qin M., Smith C.B., Usdin K. Somatic expansion in mouse and human carriers of fragile X premutation alleles. *Hum. Mutat.* 2013;34(1):157-166. DOI 10.1002/humu.22177.
- Maillick M.R., Movaghar A., Hong J., Greenberg J.S., DaWalt L.S., Zhou L., Jackson J., Rathouz P.J., Baker M.W., Brilliant M., Page D., Berry-Kravis E. Health profiles of mosaic versus non-mosaic FMR1 premutation carrier mothers of children with fragile X syndrome. *Front. Genet.* 2018;9:173. DOI 10.3389/fgene.2018.00173.
- Martin G.E., Roberts J.E., Helm-Estabrooks N., Sideris J., Vanderbilt J., Moskowitz L. Perseveration in the connected speech of boys with Fragile X syndrome with and without autism spectrum disorder. *Am. J. Intellect. Dev. Disab.* 2012;117(5):384-399. DOI 10.1352/1944-7558-117.5.384.
- Monckton D.G., Wong L.J.C., Ashizawa T., Caskey C.T. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.* 1995;4(1):1-8. DOI 10.1093/hmg/4.1.1.
- Morales F., Couto J.M., Higham C.F., Hogg G., Cuenca P., Braidia C., Wilson R.H., Adam B., Del Valle G., Brian R., Sittenfeld M., Ashizawa T., Wilcox A., Wilcox D.E., Monckton D.G. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet.* 2012;21(16):3558-3567. DOI 10.1093/hmg/dds185.
- Pearson C.E., Sinden R.R. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry.* 1996;35(15):5041-5053. DOI 10.1021/bi9601013.
- Roberts J., Hennon E.A., Anderson K. Fragile X syndrome and speech and language. *ASHA Leader.* 2003;8(19):6-27. DOI 10.1044/leader.FTR2.08192003.6.
- Shah K.A., Mirkin S.M. The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair.* 2015;32:106-112. DOI 10.1016/j.dnarep.2015.04.020.
- Usdin K., Woodford K.J. CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucleic Acids Res.* 1995;23(20):4202-4209.
- Woodford K., Weitzmann M.N., Usdin K. The use of K(+)-free buffers eliminates a common cause of premature chain termination in PCR and PCR sequencing. *Nucleic Acids Res.* 1995;23(3):539. DOI 10.1093/nar/23.3.539.
- Yudkin D.V., Lemskaya N.A., Grishchenko I.V., Dolskiy A.A. Chromatin changes caused by expansion of CGG repeats in *fmr1* gene. *Mol. Biol.* 2015;49(2):179-184.
- Zhao X.-N., Lokanga R., Allette K., Gazy I., Wu D., Usdin K. A MutSbeta-dependent contribution of MutSalpha to repeat expansions in fragile X premutation mice? *PLoS Genet.* 2016;12(7): e1006190. DOI 10.1371/journal.pgen.1006190.

ORCID ID

I.V. Grishchenko orcid.org/0000-0002-2227-8500
A.A. Tulupov orcid.org/0000-0002-1277-4113
E.D. Petrovskiy orcid.org/0000-0003-4325-4062

A.A. Savelov orcid.org/0000-0002-5332-2607
A.M. Korostyshevskaya orcid.org/0000-0002-0095-8994
E.M. Shitik orcid.org/0000-0001-8529-9176
D.V. Yudkin orcid.org/0000-0002-8940-9173

Благодарности. Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 18-15-00099 в части молекулярно-биологических исследований и 19-75-20093 – в теоретической части. Авторы выражают благодарность к.б.н. В.С. Филшману (ФИЦ ИЦиГ СО РАН, сектор геномных механизмов онтогенеза) за предоставленные плазмиды.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 23.10.2020. После доработки 16.12.2020. Принята к публикации 17.12.2020.

Английский текст <https://vavilov.elpub.ru/jour>

Продукция субтилизиновых протеаз в бактериях и дрожжах

А.С. Розанов^{1, 2}✉, С.В. Шеховцов^{1, 2}, Н.В. Богачева^{1, 2}, Е.Г. Першина^{1, 2}, А.В. Ряполова³, Д.С. Бытак³, С.Е. Пельтек^{1, 2}

¹ Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, лаборатория молекулярных биотехнологий, Новосибирск, Россия

³ Инновационный центр «Бирюч-НТ», с. Малобыково, Белгородская область, Россия

✉ rozanov@bionet.nsc.ru

Аннотация. В настоящей работе мы рассматриваем прогресс в изучении и модификации субтилизиновых протеаз. Несмотря на длительное время применения микробных протеаз и значительное число работ, посвященных их исследованию, поиск новых генов протеаз, создание продуцентов и развитие методов их применения остаются актуальными, о чем говорит высокий уровень цитирования публикаций, описывающих протеазы и их продуценты. На данный класс ферментов приходится максимальный объем производства промышленных белков в мире, что объясняет большой интерес к нему. Это говорит о чрезвычайно высокой важности получения собственных технологий их производства. В статье представлены сведения о классификации субтилизинов, истории их открытия и дальнейших работ по оптимизации их свойств. Дан обзор классов субтилизиновых протеаз и родственных им ферментов. Проанализированы проблемы поиска и отбора субтилиз из природных штаммов различных микроорганизмов, пути и особенности их модификации и используемые при этом методы генетической инженерии. Детально изучены методы оптимизации продукции промышленных субтилиз у различных штаммов, касающихся важнейших аспектов культивирования: состава среды, времени культивирования, влияния температуры и pH. Приводятся результаты последних исследований по техникам культивирования – глубинному и твердофазному культивированию. На основании рассмотренных литературных данных можно заключить, что в настоящее время практически не применяются нативные, т.е. обнаруженные в природе ферменты, в связи с решающими преимуществами, предоставляемыми белками, модифицированными при помощи генной инженерии и обладающими улучшенными свойствами: термостабильностью, общей устойчивостью к детергентам и специфической – к различным окислителям, высокой активностью в разных диапазонах температур, независимостью от ионов, стабильностью в отсутствие кальция и т.д. Большинство субтилизиновых протеаз синтезируется в штаммах-продуцентах, относящихся к разным видам рода *Bacillus*. В то же время ведутся работы по адаптации синтеза этих ферментов в других микроорганизмах, в частности дрожжей *Pichia pastoris*.

Ключевые слова: субтилизин; субтилаза; протеаза; щелочная сериновая протеаза; *Pichia pastoris*; *Bacillus subtilis*; биотехнология; генетическая инженерия; культивирование.

Для цитирования: Розанов А.С., Шеховцов С.В., Богачева Н.В., Першина Е.Г., Ряполова А.В., Бытак Д.С., Пельтек С.Е. Продукция субтилизиновых протеаз в бактериях и дрожжах. *Вавиловский журнал генетики и селекции*. 2021;25(1):125-134. DOI 10.18699/VJ21.015

Production of subtilisin proteases in bacteria and yeast

A.S. Rozanov^{1, 2}✉, S.V. Shekhovtsov^{1, 2}, N.V. Bogacheva^{1, 2}, E.G. Pershina^{1, 2}, A.V. Ryapolova³, D.S. Bytyak³, S.E. Peltek^{1, 2}

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

² Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Laboratory of Molecular Biotechnologies, Novosibirsk, Russia

³ Innovation Centre "Biruch-NT", Malobykovo village, Belgorod region, Russia

✉ rozanov@bionet.nsc.ru

Abstract. In this review, we discuss the progress in the study and modification of subtilisin proteases. Despite long-standing applications of microbial proteases and a large number of research papers, the search for new protease genes, the construction of producer strains, and the development of methods for their practical application are still relevant and important, judging by the number of citations of the research articles on proteases and their microbial producers. This enzyme class represents the largest share of the industrial production of proteins worldwide. This situation can explain the high level of interest in these enzymes and points to the high importance of designing domestic technologies for their manufacture. The review covers subtilisin classification, the history of their discovery, and subsequent research on the optimization of their properties. An overview of the classes of subtilisin proteases and related enzymes is provided too. There is a discussion about the problems with the search for (and selection of) subtilases from natural strains of various microorganisms, approaches to (and specifics of) their modification, as well as the relevant genetic engineering techniques. Details are provided on the methods for expression optimization of industrial subtilases of various strains: the details of the most important parameters of cultivation, i.e., composition of the media, culture duration, and the influence of temperature and pH. Also presented are the results

of the latest studies on cultivation techniques: submerged and solid-state fermentation. From the literature data reviewed, we can conclude that native enzymes (i.e., those obtained from natural sources) currently hardly have any practical applications because of the decisive advantages of the enzymes modified by genetic engineering and having better properties: e.g., thermal stability, general resistance to detergents and specific resistance to various oxidants, high activity in various temperature ranges, independence from metal ions, and stability in the absence of calcium. The vast majority of subtilisin proteases are expressed in producer strains belonging to different species of the genus *Bacillus*. Meanwhile, there is an effort to adapt the expression of these enzymes to other microbes, in particular species of the yeast *Pichia pastoris*.

Key words: subtilisin; subtilase; protease; alkaline serine protease; *Pichia pastoris*; *Bacillus subtilis*; biotechnology; genetic engineering; cultivation.

For citation: Rozanov A.S., Shekhovtsov S.V., Bogacheva N.V., Pershina E.G., Ryapolova A.V., Bytyak D.S., Peltek S.E. Production of subtilisin proteases in bacteria and yeast. *Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):125-134. DOI 10.18699/VJ21.015

Введение

Протеазы – ферменты, разрушающие белки путем гидролиза пептидных связей. Протеазы соответствуют общему классу ферментов – EC 3.4.X.X (Garcia-Carreno, Del Toro, 1997). Эндопептидазы преимущественно действуют на интактные белки; они расщепляют пептидные связи не терминальных аминокислотных остатков. Экзопептидазы разрывают пептидные связи между аминокислотными остатками на конце полипептидной цепи. Они делятся на амино- и карбоксипептидазы, в зависимости от того, с какого конца (N- или C-конца) белка они отщепляют аминокислоты (Barrett, McDonald, 1986). Протеазы разделяются на семейства в соответствии с механизмом их действия. Согласно базе данных MEROPS (<http://merops.sanger.ac.uk>) (Rawlings et al., 2014), выделяют следующие семейства: аспарагиновые, цистеиновые, глутаминовые, сериновые и треониновые пептидазы, металлопептидазы, а также смешанные пептидазы и пептидазы с неизвестным механизмом действия.

Протеазы встречаются у всех типов организмов. В настоящее время максимальное распространение приобрели протеазы прокариот, в основном бактерий, что связано с их высоким потенциалом для различных технологических применений. Так как протеазы необходимы в больших количествах, наряду с их свойствами огромное значение имеет их стоимость производства, что привело к тому, что протеазы главным образом производятся с использованием бактерий. Микроорганизмы способны синтезировать ферменты быстрее и дешевле, чем клетки млекопитающих и растений; на производство ферментов не влияют климатические условия или сезонные изменения, а также нормативные или этические проблемы. Кроме того, предпочтение обычно отдается внеклеточным ферментам, продуцируемым микроорганизмами, поскольку это упрощает последующую обработку, еще более снижая затраты (Tufvesson et al., 2010). По сумме характеристик, активности, диапазонам pH и температуры, а также стоимости самым востребованным классом протеаз оказались субтилизины, или субтилазы.

Субтилазы – один из самых больших классов сериновых протеаз, которые встречаются в геномах всех форм жизни, включая вирусы. По аминокислотным последовательностям субтилазы делятся на шесть семейств: субтилизины, термитазы, протеиназы K, антибиотические пептидазы, кексины и пирролизины. Субтилизины, в свою очередь, также подразделяются на несколько подсемейств: настоя-

щие субтилизины, высокощелочные протеазы, внутриклеточные протеазы, промежуточные субтилизины и субтилизины высокой молекулярной массы.

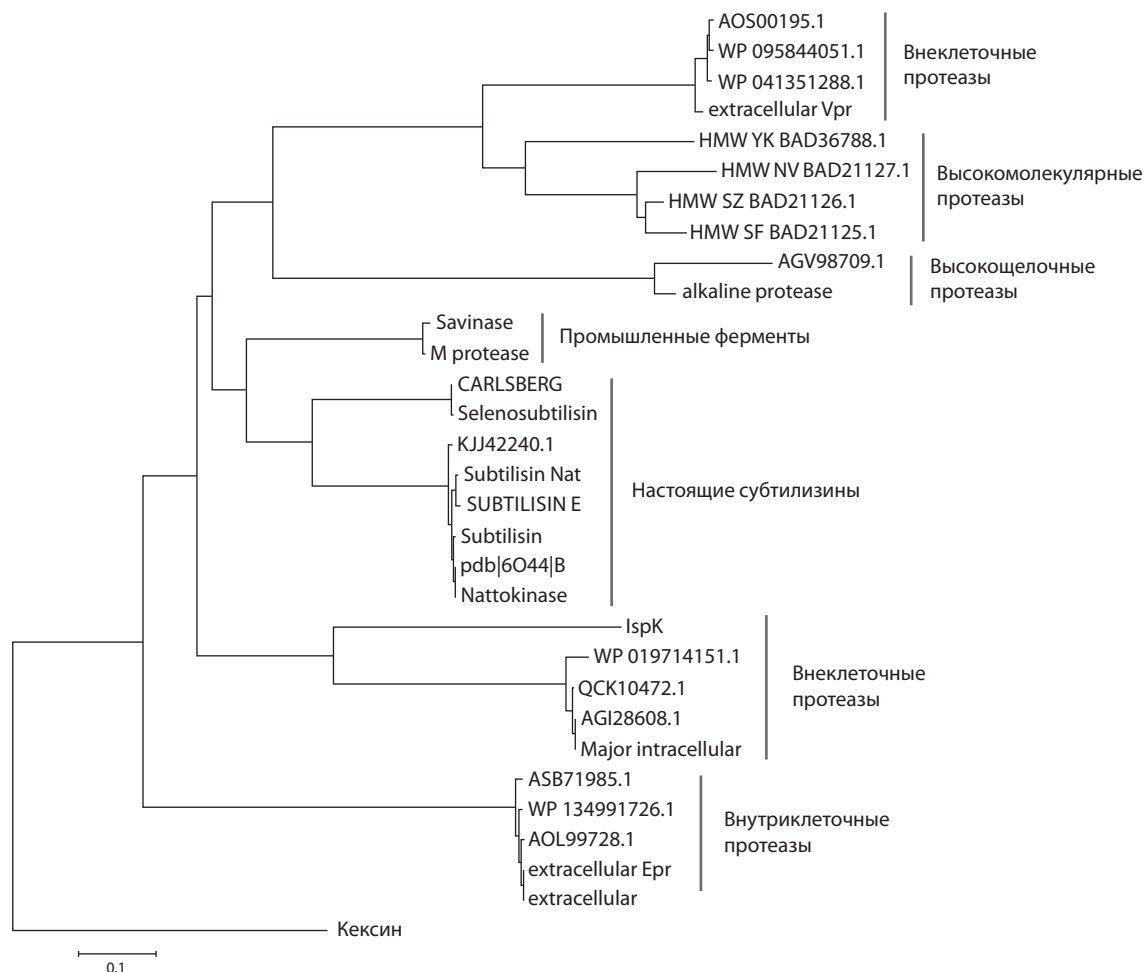
Все подсемейства субтилизинов имеют биотехнологический потенциал. Первой щелочной сериновой протеазой, получившей широкое распространение, стал субтилизин А (E.C. 3.4.21.62) – щелочная сериновая протеаза из *Bacillus subtilis*. Своё название фермент получил по видовому названию продуцента (Ottesen, Svendsen, 1970; Kkemura et al., 1987). История открытия и изучения субтилизинов началась в научно-исследовательском центре пивоваренной компании Carlsberg, и первый описанный фермент носит название «субтилизин Карлсберг» (Smith et al., 1966).

Каталитический центр сериновых протеаз образован тремя аминокислотными остатками: Asp-32, His-64 и Ser-221. Так как аминокислотным остатком, осуществляющим нуклеофильную атаку, является Ser-221, субтилизины и родственные им протеолитические ферменты называются сериновыми протеиназами.

К высокощелочным протеазам принадлежит, например, фермент, выделенный из штамма *Bacillus* sp. KSM-K16 (Kobayashi et al., 1995). Оптимум его активности находится при 55 °C и pH 12.3. Этот фермент используется в промышленности в комплексах с детергентами, как и родственные высокощелочные протеазы Savinase и Maxacal. Промежуточные субтилизины занимают позицию между настоящими субтилизинами и высокощелочными протеазами; к ним также относятся некоторые перспективные ферменты. Так, фермент ALTP, выделенный из *Alkaliphilus transvaalensis* (Kobayashi et al., 2007), имел максимальную активность при очень высоких температурах и pH, а именно при 70 °C и pH более 12.6. При этом ALTP был способен выполнять каталитическую функцию и при меньших температурах и pH. Филогенетическое дерево, построенное по аминокислотным последовательностям субтилизиновых протеаз, представлено на рисунке.

Внутриклеточные протеазы изучены сравнительно слабо по сравнению с вышеперечисленными семействами. Это связано с тем, что они работают при более низком pH, характерном для цитоплазмы. Так, например, внутриклеточная протеаза из *B. megaterium* (Jeong et al., 2018) при 50 °C демонстрировала оптимум активности при pH 6.0–7.0.

Из алкалофильных *Bacillus* spp. был также получен набор субтилизинов высокой молекулярной массы (Okuda



Филогенетическое дерево, построенное по аминокислотным последовательностям субтилизиновых протеаз и родственных им ферментов *B. subtilis*, а также некоторых промышленных ферментов.

et al., 2004) длиной около 650 аминокислот (длина прекурсора – 800 аминокислот). Оптимальный уровень pH 10.5–11.0, температуры 40–45 °C.

Бактерии наиболее широко используют в качестве продуцентов протеаз, а род *Bacillus* – наиболее известный источник среди них. В первую очередь, это связано с высокой способностью к секреции белка, в результате чего удается получать более 20 г белка на 1 литр среды (Harwood, Cranenburgh, 2008). Различные виды рода *Bacillus* продуцируют нейтральные и щелочные протеазы (Anandharaj et al., 2016; Rehman et al., 2017), что важно для промышленности. Протеазы представителей *Bacillus* обладают уникальными характеристиками, позволяющими использовать их во многих отраслях промышленности. В связи с этим примерно 60 % от общего объема продаж ферментов по всему миру приходится на протеазы из различных видов *Bacillus*. Благодаря широкому диапазону pH и температурной активности и стабильности они применяются в индустрии моющих средств (Potges et al., 2002). Для этого ферменты должны быть устойчивыми к щелочной среде и сохранять активность в присутствии ингибиторов, включая окислители и поверхностно-активные вещества. Кроме того, протеазы, выделенные из штаммов рода *Bacillus*, могут применяться в пищевой

промышленности для получения биологически активных пептидов и обработки различных пищевых продуктов (Latiffi et al., 2013; Ke et al., 2018). Другой особенностью этих протеаз является стабильность в органических растворителях и, следовательно, возможность их применения в органическом синтезе (Hu et al., 2013). Ввиду большой коммерческой значимости значительное число патентов основано на использовании штаммов, относящихся к роду *Bacillus* (см. таблицу).

Распространенность производства протеаз с использованием штаммов рода *Bacillus* обуславливает их экономическая эффективность. В качестве среды для них можно использовать побочные продукты агропромышленного производства, включая мелассу сахарного тростника и кукурузного крахмала для глубокой ферментации (Shikha et al., 2007), а также различных типов отрубей и жмыхов для твердофазной ферментации (Shivasharana, Naik, 2012).

Поиск щелочных сериновых протеаз в природе

Протеазы являются важными коммерческими белками, на которые приходится большая часть мирового производства белка. Вариантов их использования множество, и каждый технологический процесс обладает своими особенностями и требованиями к ферментам. Постоянный

Промышленные субтилазы, полученные из видов рода *Bacillus*

Фермент	Класс	Вид
Dispase I VR	Протеазы	<i>B. polymyxa</i>
Dispase II VR		
Proteinase	Субтилизин А	<i>B. licheniformis</i>
Neutrase	Металлопротеазы	<i>B. amyloliquefaciens</i>
Esperase	Сериновые эндопептидазы (субтилизин А)	<i>Bacillus</i> sp.
Everlase	Субтилизин А	
Protamex	Протеазы	
Savinase	Сериновые эндопептидазы (субтилизин А)	
Alcalase		<i>B. licheniformis</i>
Optimase PR	Сериновые эндопептидазы	<i>B. subtilis</i>
GenencorVR Protease 899	Нейтральные металлопептидазы	
ProtexTM 6L	Субтилизины, сериновые эндопептидазы	<i>B. licheniformis</i>
Multifect	Нейтральные сериновые эндопептидазы	<i>B. amyloliquefaciens</i>

интерес к ним обусловлен также поиском ферментов, пусть и не превосходящих по своим свойствам уже известные варианты, но не попадающих под действие патентов. В связи с этим ежегодно публикуется множество новых статей. Наибольшее количество генов щелочных сериновых протеаз обнаружено в геномах бактерий, относящихся к роду *Bacillus*. Вторая по количеству выделяемых протеаз группа – актиномицеты. Значительное количество статей посвящено также поиску щелочных протеаз грибного происхождения (Sharma et al., 2017). В последних работах акцент ставится на поиски ферментов, обладающих кератиназной активностью, что обусловлено возросшим интересом к переработке кератинсодержащих остатков, например перьев.

Источником одного из перспективных генов, кодирующих сериновую протеазу, стал штамм *Bacillus licheniformis* NMS-1, выделенный из почвы вблизи природного термального источника в Шри-Ланке (Mathew, Gunathilaka, 2015). Этот белок применяется в создании моющих средств. Близкородственный штамм *B. licheniformis* K7A, синтезирующий щелочную протеазу, был получен R. Najidj с коллегами (2018). Анализ синтезируемого белка показал, что он обладает наибольшей активностью при температуре 10 и 70 °C. Активность фермента была выше, чем у коммерческих препаратов Alcalase и Thermolysin. Ген еще одной сериновой протеазы был найден в геноме бактерии *Bacillus amyloliquefaciens* FSE-68, выделенной из закваски для ферментации сои в Южной Корее. Ее последовательность была определена при помощи LC/ESI-MS/MS анализа и полногеномного секвенирования. По сравнению с родственным гомологом хорошо изученного субтилизина BPN из *B. amyloliquefaciens*, фермент продемонстрировал немного большую стабильность в присутствии ионов кальция (Cho, 2019). Белок, выделенный из алкалофильного штамма *Bacillus luteus* H11, проявлял протеолитическую активность при концентрации NaCl до 5 М, температуре 45 °C и pH 10.5 (Kalwasińska et al., 2018). В Китае при скрининге бактерий из продуктов ферментации сои удалось выделить штамм *B. subtilis* MX-6,

обладающий высоким уровнем продукции наттокиназоподобного белка (Gulmez et al., 2018).

Список работ, посвященных поиску новых вариантов протеаз как бактериального, так и грибного происхождения, опубликованных в последнее время, можно расширять бесконечно. Это говорит о том, что по разным причинам развитие производств протеолитических ферментов актуально до настоящего времени. Особенно это характерно для развивающихся стран, где велико желание увеличить долю продуктов, в том числе биотехнологических, на внутреннем рынке. Особенно много статей в этом направлении опубликовано научными группами из Индии. В настоящее время в России подобные исследования практически не ведутся.

Генетическая инженерия субтилизина

Субтилизин – это, пожалуй, самый изученный при помощи как статистического, так и направленного мутагенеза промышленный фермент. Применение субтилизина постоянно росло сразу после начала его производства. Для удовлетворения нужд промышленности требовалось улучшение его свойств. В начале 1980-х годов стали активно развиваться методы направленной инженерии белков. В результате применения этих методов к субтилизину до 2000 г. в научной литературе были описаны мутации более чем половины из его 275 аминокислотных остатков. Патентная литература содержит множество примеров, и, несомненно, еще большее их число похоронено в морозильных камерах биотехнологических компаний. Наиболее мутагенизированными являются протеазы *B. amyloliquefaciens* (BPNP), *B. subtilis* (субтилизин E) и *Bacillus lentus* (Savinase).

Белковая инженерия предоставляет несколько эффективных методов, которые включают в себя рациональный дизайн и направленную эволюцию. Рациональный дизайн включает сайт-направленного мутагенеза для замены специфических аминокислотных остатков в структуре белка, что может привести к получению белков с желаемыми свойствами, в том числе повышенную термостабильность

(Jaouadi et al., 2010; Huang et al., 2015), позволяет получать информацию распознавания субстрата и модификации субстратной специфичности (Jaouadi et al., 2014). С другой стороны, направленная эволюция основана на выполнении последовательных циклов мутагенеза и отбора (Liu et al., 2014).

Стабильность субтилизина

Стабильность субтилизина была насущной потребностью его производства, в связи с чем эти работы получили широкое распространение. Интересная особенность субтилизина – то, что его биосинтез требует участия N-концевого продомена (Ikemura et al., 1987). Фолдинг зрелого субтилизина без продомена теоретически возможен, но занимает тысячи лет.

Важное свойство субтилизина – его строгая зависимость от кальция (Voordouw et al., 1976; Genov et al., 1995). Универсальная особенность субтилизинов – наличие одного или нескольких сайтов связывания кальция. Рентгеновские структуры высокого разрешения субтилизина BPNP, а также нескольких гомологов (Bode et al., 1987; Betzel et al., 1992) выявили детали консервативного сайта связывания кальция, названного сайтом А. Кальций в сайте А координируется пятью карбонильными атомами кислорода и остатком аспарагиновой кислоты. Четыре атома кислорода обеспечены петлей, содержащей аминокислотные остатки 75–83. Геометрия лигандов представляет пятиугольную бипирамиду, ось которой проходит через карбонилы аминокислотных остатков 75 и 79. На одной стороне петли находится бидентатный карбоксилат (D41), а на другой – N-конец белка и боковая цепь Q2. Семь координационных расстояний варьируют от 2.3 до 2.6 Å, самое короткое из которых относится к аспартилкарбоксилату.

Второй ион-связывающий сайт (В) расположен в 32 ангстремах от сайта А в неглубокой щели между двумя сегментами полипептидной цепи вблизи поверхности молекулы. Координационная геометрия этого участка очень напоминает искаженную пятиугольную бипирамиду. Три из формальных лигандов являются производными белка и включают атом кислорода карбонильного атома E195 и два кислородных атома карбоксилата боковой цепи D197. Четыре молекулы воды завершают первую координационную сферу.

Так как зависимость от кальция нежелательна, были проведены работы по получению стабильных белков субтилизина, не зависящих от присутствия или отсутствия кальция в растворе. В статье (Strausberg et al., 2005) описана модификация аминокислотной последовательности субтилизина с поврежденным сайтом связывания ионов кальция для повышения его стабильности. В результате был получен вариант, в 15000 раз более стабильный в сравнении с исходным.

Новейшие исследования по модификации щелочных сериновых протеаз

Несмотря на значительный прогресс в развитии свойств щелочных сериновых протеаз, исследования по их модификации продолжают до настоящего времени. Так, Н.У. Zhao и Н. Feng (2018) путем направленной эволюции

получили семь вариантов (P9S, A1G/K27Q, A38V, A116T, T162I, S182R и T243S) протеазы, выделенной из *Bacillus pumilus* BA06. Все они обладали повышенной протеолитической активностью при 15 °С в отношении казеина и синтетических пептидных субстратов. За исключением T243S, термостабильность этих вариантов не снижалась по сравнению с ферментом дикого типа. Комбинированные варианты мутаций продемонстрировали дальнейшее увеличение специфической казеинолитической активности. Комбинированные варианты P9S/K27Q и P9S/T162I показали приблизительно пятикратное увеличение казеинолитической активности при 15 °С почти без потери термостабильности (Zhao, Feng, 2018). В другой работе этой же группы направленному мутагенезу подвергалась щелочная протеаза *B. pumilus* (Zhao et al., 2016). Полученный в результате двойной мутант (W106K/V149I и W106K/M124L) имел в два с половиной раза более высокую активность в сравнении с исходным вариантом при 15 °С, при этом его стабильность при 60 и 70 °С была выше в 2.7 и 5 раз соответственно (Zhao et al., 2016).

При сравнении галотолерантных субтилизинов с устойчивыми были выявлены шесть аминокислотных позиций, в которых полярные аминокислотные остатки были заменены неполярными. Исследователи предположили, что эти замены могут привести к повышению термостабильности. Для проверки этого был выполнен мутагенез алкалазы штамма *B. subtilis* no. 16 и субтилизина Карлсберг. При этом наблюдалось повышение устойчивости ферментов к высоким содержаниям солей (125 г/л) в 1.2 и 1.8 раза соответственно (Takenaka et al., 2018). N.M. Ashraf с коллегами (2019) модифицировали сериновую протеазу из *Pseudomonas aeruginosa* по позициям A29G и V336I, в итоге было достигнуто повышение температуры наблюдаемой остаточной активности на 5 °С и увеличение каталитической активности в 1.4 раза (Ashraf et al., 2019). В другой работе (Gong et al., 2017) статистический мутагенез гена щелочной протеазы, обнаруженного при метагеномном анализе, привел к увеличению активности в 6.6 раз.

Получение протеаз в штаммах *Bacillus spp.*

Основными продуцентами сериновых протеаз на протяжении всего времени их использования остаются бактерии рода *Bacillus*. Условия культивирования и состав используемых сред играют важную роль в производстве ферментов микроорганизмами (Abidi et al., 2008). Чтобы получить высокий и коммерчески значимый уровень продукции протеаз, важно подобрать условия роста и индукции (Sharma et al., 2015). Не существует единой среды, пригодной для всех штаммов-продуцентов. Каждый организм или штамм имеет свои особые условия для максимальной выработки конкретного фермента. Рассмотрим различные аспекты культивирования подробнее.

Состав среды

Углерод и азот – основные компоненты среды, действующие в том числе как важный стимулятор для роста микроорганизмов и синтеза ферментов. Самый распространенный и часто наиболее дешевый (после крахмала) источник углерода – глюкоза, однако при ее использовании может

возникать эффект катаболической репрессии многих биосинтетических процессов в клетке. Максимальная продукция фермента бактериальным штаммом AKS-4 была при использовании глюкозы в концентрации 1 %. При этом уровень продукции протеазы составил 59.10 ед/мл (Sharma et al., 2015). Повышенный уровень продукции протеазы *Bacillus pseudofirmus* AL-89 отмечен при добавлении глюкозы, тогда как для *Nesterenkonia* sp. продукция протеазы AL-20 в присутствии глюкозы подавлялась (Gessesse et al., 2003).

Максимальная продукция щелочной протеазы (2450 ед/мл) для *B. licheniformis* была получена в среде, содержащей 60 г/л глюкозы, дальнейшее увеличение концентрации привело к незначительному снижению продукции фермента. Глюкоза в высокой концентрации ингибировала выработку фермента *Streptomyces* sp., причем концентрация 0.5 % была оптимальной для производства фермента, а рост – при 1 % (Mehta et al., 2006). Продукция протеазы *P. aeruginosa* MCM B-327 в соево-триптоновой среде подавлялась на 95 и 60 % при добавлении глюкозы и фруктозы соответственно (Zambare et al., 2011). K. Sharma с коллегами (2014) использовали для производства протеазы *Bacillus aryabhatai* K3 различные источники углерода, такие как глюкозу, лактозу, галактозу и крахмал. Максимальная выработка протеазы (622.64 ед/мл) была при использовании лактозы (10 г/л) в качестве источника углерода (Sharma et al., 2014). Аналогичным образом M.S. Dodia с коллегами (2006) обнаружили, что для большинства исследованных изолятов секреция фермента была оптимальной при использовании лактозы в качестве источника углерода. *B. licheniformis* BBRC 100053 также продемонстрировал более высокую продуктивность протеазы в культуральных средах, содержащих лактозу как источник углерода (Nejad et al., 2010).

Кроме простых сахаров, для производства протеаз опробовали и другие источники углерода. Использование 5 % крахмала привело к максимальной выработке протеазы *Bacillus* sp. 2–5 (Khosravi-Darani et al., 2008). Штамм *Bacillus clausii* № 58 хорошо рос на различных источниках углерода на основе крахмала (Kumar et al., 2004). Кукурузный крахмал в концентрации 0.5 % способствовал наибольшему выходу протеазы, затем следует пшеничная мука и пшеничные отруби. Однако добавление картофельного крахмала привело к снижению титра протеазы, что, возможно, связано с присутствием ингибиторов протеазы в картофеле (Kumar et al., 2004). Использование пшеничной муки в качестве источника сахаров показало хороший результат при наработке протеаз *Bacillus* sp. (Chu, 2007). *Bacillus lateosporus* продуцировал протеазы при широком спектре источников углерода; лучшими источниками углерода для секреции протеазы были растворимый крахмал, тринатрий цитрат, лимонная кислота и глицерин (Usharani, Muthuraj, 2010).

Источники азота также оказывают значительное влияние на выход целевого белка, при этом оптимальные источники для разных штаммов различаются. Наивысший уровень продукции протеазы штаммом *Bacillus cereus* 146 отмечен в присутствии экстракта говядины в качестве источника азота. Присутствие дрожжевого экстракта, пептона и триптона увеличивало показатели роста культур,

но количество целевого белка при этом было невысоким (Shafee et al., 2005). T. Srinivasan с коллегами (2009) установили, что триптон увеличивает выработку протеазы для штамма *Bacillus* sp. Пептон оказался оптимальным для продукции протеазы *B. licheniformis* BBRC 100053 (Nejad et al., 2010). Дрожжевой экстракт продемонстрировал максимальное увеличение продукции ферментов *Bacillus* sp. (Prakasham et al., 2006). В случае *Bacillus* sp. APP1 среди всех используемых источников органического азота соевый шрот заметно увеличил выработку внеклеточной протеазы (Chu, 2007). R.K. Jaswal с коллегами (2008) также сообщили, что использование соевого шрота показало лучший результат в сравнении с казеином, желатином и пептоном для синтеза протеазы *Bacillus circulans*. При использовании казеина, пептона, дрожжевого экстракта и экстракта говядины в качестве источника азота для продукции протеазы бактериальным штаммом AKS-4 наибольший выход наблюдался в присутствии казеина. Среди различных источников органического азота обезжиренное молоко давало максимальный выход протеазы в случае *Bacillus caseinilyticus*, за которым следовали солодовый экстракт, пептон и дрожжевой экстракт. Хлорид аммония как неорганический источник азота ингибировал наработку протеиназы (Mothe, 2016).

Влияние pH и температуры на уровень продукции протеаз

Влияние pH на скорость синтеза целевого продукта индивидуально для каждого штамма-продуцента. Так, для продукции протеаз в *Bacillus* sp. MIG (Gouda, 2006) и *B. cereus* SIU1 (Singh et al., 2010) был оптимален слабощелочной pH (6.3–6.5). В слабощелочной среде (pH 8.0–8.5) были зафиксированы максимальные уровни продукции для *B. licheniformis* IKBC-17 (Olajuyigbe et al., 2005), *B. subtilis* IKBS 10 (Olajuyigbe et al., 2005), *Bacillus mace-rans* IKBM-11 (Olajuyigbe et al., 2005), *B. amovivorus* (Sharm-in et al., 2005). Для восьми изолятов *Bacillus* M.S. Dodia с коллегами (2006) показали, что наилучшие условия для роста бактерий наблюдаются при pH 9.0, тогда как оптимальное значение pH для секреции фермента варьировало от 8.0 до 10.0. Значение pH 9 было оптимальным для продукции протеаз в *Bacillus* sp. (Prakasham et al., 2006), *Bacillus* sp. APP1 (Chu, 2007), *B. proteolyticus* CFR3001 (Bhaskar et al., 2007). Более высокий начальный pH был установлен для продукции протеазы *B. licheniformis* TISTR 1010 (pH 10.0) (Vaithanomsat et al., 2008), для *B. circulans* (pH 10.5) (Jaswal et al., 2008) и *Bacillus* sp. 2–5 (pH 10.7) (Khosravi-Darani et al., 2008).

Температура также является важным параметром, индивидуальным для каждого штамма. Для *P. aeruginosa* PseA (Gupta, Khare, 2007), *B. licheniformis* (Asokan, Jayanthi, 2010), *Bacillus coagulans* (Asokan, Jayanthi, 2010), *B. cereus* (Kebabci, Cihangir, 2010), *P. aeruginosa* MCM B-327 (Zambare et al., 2011), *P. chrysogenum* IHN5 (Ikram-Ul-Haq et al., 2006) и *A. oryzae* 637 (Srinubabu et al., 2007) для продукции протеаз оптимальна температура 30 °C. Более низкая оптимальная температура (25 °C) характерна для *B. circulans* (Jaswal et al., 2008), *Microbacterium* sp. (Thys et al., 2006), в то время как наибольшая продукция у *B. cinerea* отмечена при 28 °C (Abidi et al., 2008).

При 37 °С максимальная продукция была для штаммов *Bacillus amovivorus* (Sharmin et al., 2005), *B. proteolyticus* CFR3001 (Bhaskar et al., 2007), *Bacillus aquimaris* VITP4 (Shivanand, Jayaraman, 2009), *B. subtilis* Rand (Abusham et al., 2009); при 40 °С – для *Bacillus* sp. 2–5 (Khosravi-Darani et al., 2008), *Vibrio parvohalophilus* (Gupta et al., 2008) и *Streptomyces roseiscleroticus* (Shivanand, Jayaraman, 2009); при 50 °С – для *Bacillus* sp. APP1 (Porres et al., 2002) и *B. subtilis* BS1 (Shaheen et al., 2008).

Продукция щелочных сериновых протеаз в дрожжах

Наработка протеаз возможна не только в штаммах, принадлежащих к роду *Bacillus*, но и в других бактериях, а также в дрожжах, например в штаммах *Pichia pastoris*. Эти штаммы изначально не обладают специфичной активностью, в связи с чем необходима их модификация при помощи генетической инженерии. Таких работ немного, и в основном они нацелены на получение грибных протеаз или протеаз медицинского назначения.

В. Liu с коллегами (2014) провели анализ экспрессии гена кератиназы *B. licheniformis* BBE11-1 в трех гетерологических системах экспрессии: *Escherichia coli*, *B. subtilis* и *P. pastoris*. Наивысший лучший уровень продукции был для *B. subtilis* (3010 ед/мл), что в три раза превышало результат для *P. pastoris*. При этом для культивирования *B. subtilis* не использован метанол, а время культивирования было в два раза меньше. S. Radha и P. Gunasekara (2009) описали сравнительное клонирование гена кератиназы из *B. licheniformis* MKU3 в *Bacillus megaterium* и *P. pastoris*. В результате были получены сравнимые активности конечной культуры с концентрацией целевого белка около 0.35 г/л. Белок из *P. pastoris* был подвергнут гликозилированию. Следует отметить, что культивирование в биореакторе в статье не описано. Схожие данные приведены для продукции кератиназы из *B. licheniformis* PWD-1 (Cheng et al., 1995).

Н.Н. Lin с коллегами (2009) изучали продукцию кератиназы из *Pseudomonas aeruginosa* в *P. pastoris*. Выход составил около 0.5 г белка на 1 литр. В данном случае белок не подвергался гликозилированию. К. Zhou с коллегами (2017) клонировали белок субтилизин QK из *B. subtilis* QK02, имеющий высокое сходство с наттокиназой, в *P. pastoris* GS115. Целью было получение белка, обладающего тромболитическими свойствами. Концентрация общего белка в конечном супернатанте достигала 7.6 г/л. В их исследовании pH поддерживался на уровне 5.0, тогда как в работах (Liu V. et al., 2014) и (Porres et al., 2002) отсутствие контроля pH привело к его повышению, произошли ингибирование роста культуры и снижение содержания кератиназы в растворе. Аналогичная картина наблюдалась в статье (Lin et al., 2009).

Клонирование гена щелочной протеазы из термофильной бактерии *B. stearothermophilus* F1 осуществлено также в *P. pastoris* GS115 (Latiffi et al., 2013). Достигнутая активность составила 4.13 ед/мл; судя по полученной молекулярной массе, белок не был гликозилирован. В исследовании (Ke et al., 2018) в *P. pastoris* был экспрессирован ген щелочной протеазы из гриба *Aspergillus sojae*, полученная конечная активность составила 400 ед/мл.

На уровень продукции целевого белка высокое влияние также оказывает кодонный состав. В работе (Hu et al., 2013) оптимизация кодонного состава гена привела к повышению уровня продукции целевого белка по сравнению с исходным геном, тем не менее не представлено данных по культивированию в контролируемых условиях биореактора. Повышение копииности экспрессионной кассеты также позволяет увеличить выход целевого белка, что было продемонстрировано на примере сериновой протеазы из гриба *Trichoderma koningii* (Shu et al., 2016).

Таким образом, максимальный уровень наработки щелочных сериновых протеаз в экспрессионной системе *P. pastoris* выше по сравнению с *E. coli*, но значительно ниже, чем в стандартных штаммах *B. subtilis*. При этом промышленные штаммы *Bacillus* spp. превосходят как экспрессионные системы *P. pastoris*, так и *B. subtilis* более чем на порядок. В патенте 2005 г. (Shih, 2005) описан штамм *B. licheniformis* T1, обеспечивающий уровень продукции белка на уровне 16 г/л, тогда как максимальная концентрация наработанной кератиназы в *P. pastoris* – около 0.1–0.2 г/л целевого белка.

Заключение

Щелочные сериновые протеазы субтилизинового семейства широко применяются в различных областях промышленности. Примерно 60 % от общего объема продаж ферментов по всему миру приходится на протеазы, выделенные из бактерий рода *Bacillus*.

На сегодняшний день практически не применяются нативные, т. е. обнаруженные в природе ферменты, которые были вытеснены белками, модифицированными при помощи генной инженерии и обладающими улучшенными свойствами: термостабильностью, устойчивостью – общей к детергентам и специфической – к различным окислителям, высокой активностью в разных диапазонах температур, независимостью от ионов, стабильностью при отсутствии кальция и т. д.

В качестве продуцентов щелочных сериновых протеаз в настоящее время используются различные штаммы, относящиеся к роду *Bacillus*. Большинство из них изначально обладали нужной активностью, которая была усилена при помощи мутагенеза или генетической инженерии. Среди штаммов-продуцентов преобладают виды, имеющие статус GRAS (generally regarded as safe, т. е. считающиеся безопасными даже при употреблении в пищу), в первую очередь *B. subtilis* и *B. licheniformis*. Штаммы, изначально не имевшие протеазную активность, пока не удастся довести до уровня бактерий, которые продуцировали протеазы изначально, даже при помощи технологий генетической инженерии.

В литературе описаны попытки получения продуцентов щелочных сериновых протеаз на основе метилотрофного штамма *P. pastoris*. В сравнении с экспрессией тех же генов в ген-инженерных штаммах *B. subtilis* результат оказался заметно хуже. Следовательно, для создания штаммов, эффективно продуцирующих целевые щелочные протеазы, необходимо использовать бациллярные системы экспрессии. Данные штаммы потребуют доработки свойств синтезируемого фермента и уровня его

продукции при помощи методов направленного и статистического мутагенеза. Патентопригодные продуценты щелочной сериновой протеазы (субтилизин А) могут быть получены поиском новых штаммов в природе или при использовании вышедших из-под патентной защиты штаммов.

Список литературы / References

- Abidi F., Limam F., Nejib M.M. Production of alkaline proteases by *Botrytis cinerea* using economic raw materials: Assay as biode detergent. *Process Biochem.* 2008;43(11):1202-1208. DOI 10.1016/j.procbio.2008.06.018.
- Abusham R.A., Rahman R.N.Z.R.A., Salleh A., Basri M. Optimization of physical factors affecting the production of thermo-stable organic solvent-tolerant protease from a newly isolated halo tolerant *Bacillus subtilis* strain Rand. *Microb. Cell Fact.* 2009;8(1):20. DOI 10.1186/1475-2859-8-20.
- Anandharaj M., Sivasankari B., Siddharthan N., Rani R.P., Sivakumar S. Production, purification, and biochemical characterization of thermostable metallo-protease from novel *Bacillus alkalitelluris* TW13 isolated from tannery waste. *Appl. Biochem. Biotechnol.* 2016;178(8):1666-1686. DOI 10.1007/s12010-015-1974-7.
- Ashraf N.M., Krishnagopal A., Hussain A., Kastner D., Sayed A.M.M., Mok Y.-K., Swaminathan K., Zeeshan N. Engineering of serine protease for improved thermostability and catalytic activity using rational design. *Int. J. Biol. Macromol.* 2019;126:229-237. DOI 10.1016/j.ijbiomac.2018.12.218.
- Asokan S., Jayanthi C. Alkaline protease production by *Bacillus licheniformis* and *Bacillus coagulans*. *J. Cell Tissue Res.* 2010;10(1): 2119-2123.
- Barrett A.J., McDonald J.K. Nomenclature: protease, proteinase and peptidase. *Biochem. J.* 1986;237(3):935. DOI 10.1042/bj2370935.
- Betzl C., Klupsch S., Papendorf G., Hastrup S., Branner S., Wilson K.S. Crystal structure of the alkaline proteinase Savinase™ from *Bacillus lentus* at 1.4 Å resolution. *J. Mol. Biol.* 1992;223(2):427-445. DOI 10.1016/0022-2836(92)90662-4.
- Bhaskar N., Sudeepa E.S., Rashmi H.N., Tamil Selvi A. Partial purification and characterization of protease of *Bacillus proteolyticus* CFR3001 isolated from fish processing waste and its antibacterial activities. *Bioresour. Technol.* 2007;98(14):2758-2764. DOI 10.1016/j.biortech.2006.09.033.
- Bode W., Papamokos E., Musil D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. Structural analysis, subtilisin structure and interface geometry. *Eur. J. Biochem.* 1987;166(3):673-692. DOI 10.1111/j.1432-1033.1987.tb13566.x.
- Cheng S.-W., Hu H.-M., Shen S.-W., Takagi H., Asano M., Tsai Y.-C. Production and characterization of keratinase of a feather-degrading *Bacillus licheniformis* PWD-1. *Biosci. Biotechnol. Biochem.* 1995; 59(12):2239-2243. DOI 10.1271/bbb.59.2239.
- Cho S.J. Primary structure and characterization of a protease from *Bacillus amyloliquefaciens* isolated from *meju*, a traditional Korean soybean fermentation starter. *Process Biochem.* 2019;80:52-57. DOI 10.1016/j.procbio.2019.02.011.
- Chu W.H. Optimization of extracellular alkaline protease production from species of *Bacillus*. *J. Ind. Microbiol. Biotechnol.* 2007; 34(3):241-245. DOI 10.1007/s10295-006-0192-2.
- Dodia M.S., Joshi R.H., Patel R.K., Singh S.P. Characterization and stability of extracellular alkaline proteases from halophilic and alkaliphilic bacteria isolated from saline habitat of coastal Gujarat, India. *Braz. J. Microbiol.* 2006;37(3):276-282. DOI 10.1590/S1517-83822006000300015.
- Garcia-Carreño F.L., Navarrete Del Toro M.A. Classification of proteases without tears. *Biochem. Educ.* 1997;25(3):161-167. DOI 10.1016/S0307-4412(97)00005-8.
- Genov N., Filippi B., Dolashka P., Wilson K.S., Betzel C. Stability of subtilisins and related proteinases (subtilases). *Int. J. Pept. Protein Res.* 1995;45(4):391-400. DOI 10.1111/j.1399-3011.1995.tb01054.x.
- Gessesse A., Hatti-Kaul R., Gashe B.A., Mattiasson B. Novel alkaline proteases from alkaliphilic bacteria grown on chicken feather. *Enzyme Microb. Technol.* 2003;32(5):519-524. DOI 10.1016/S0141-0229(02)00324-1.
- Gong B.L., Mao R.Q., Xiao Y., Jia M.L., Zhong X.L., Liu Y., Xu P.-L., Li G. Improvement of enzyme activity and soluble expression of an alkaline protease isolated from oil-polluted mud flat metagenome by random mutagenesis. *Enzyme Microb. Technol.* 2017;106:97-105. DOI 10.1016/j.enzymictec.2017.06.015.
- Gouda M.K. Optimization and purification of alkaline proteases produced by marine *Bacillus* sp. MIG newly isolated from eastern harbour of Alexandria. *Pol. J. Microbiol.* 2006;55(2):119-126.
- Gulmez C., Atakisi O., Dalginli K.Y., Atakisi E. A novel detergent additive: Organic solvent- and thermo-alkaline-stable recombinant subtilisin. *Int. J. Biol. Macromol.* 2019;108:436-443. DOI 2018;108: 436-443. <https://doi.org/10.1016/j.ijbiomac.2017.11.133>.
- Gupta A., Joseph B., Mani A., Thomas G. Biosynthesis and properties of an extracellular thermostable serine alkaline protease from *Virgibacillus pantothenicus*. *World J. Microbiol. Biotechnol.* 2008; 24(2):237-243. <https://doi.org/10.1007/s11274-007-9462-z>.
- Gupta A., Khare S.K. Enhanced production and characterization of a solvent stable protease from solvent tolerant *Pseudomonas aeruginosa* PseA. *Enzyme Microb. Technol.* 2007;42(1):11-16. DOI 10.1016/j.enzymictec.2007.07.019.
- Hadjidj R., Badis A., Mechri S., Eddouaouda K., Khelouia L., Annane R., Hattab M.E., Jaouadi B. Purification, biochemical, and molecular characterization of novel protease from *Bacillus licheniformis* strain K7A. *Int. J. Biol. Macromol.* 2018;114:1033-1048. DOI 10.1016/j.ijbiomac.2018.03.167.
- Harwood C.R., Cranenburgh R. *Bacillus* protein secretion: an unfolding story. *Trends Microbiol.* 2008;16(2):73-79. DOI 10.1016/j.tim.2007.12.001.
- Hu H., Gao J., He J., Yu B., Zheng P., Huang Z., Mau X., Yu J., Han G., Chen D. Codon optimization significantly improves the expression level of a keratinase gene in *Pichia pastoris*. *PLoS One.* 2013; 8(3):e58393. <https://doi.org/10.1371/journal.pone.0058393>.
- Huang R., Yang Q., Feng H. Single amino acid mutation alters thermostability of the alkaline protease from *Bacillus pumilus*: Thermodynamics and temperature dependence. *Acta Biochim. Biophys. Sin.* 2015;47(2):98-105. DOI 10.1093/abbs/gmu120.
- Ikemura H., Takagi H., Inouye M. Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J. Biol. Chem.* 1987;262(16):7859-7864.
- Ikram-Ul-haq H.M., Umber H. Production of protease by *Penicillium chrysogenum* through optimization of environmental conditions. *J. Agric. Soc. Sci.* 2006;2(1):23-25.
- Jaouadi B., Aghajari N., Haser R., Bejar S. Enhancement of the thermostability and the catalytic efficiency of *Bacillus pumilus* CBS protease by site-directed mutagenesis. *Biochimie.* 2010;92(4):360-369. DOI 10.1016/j.biochi.2010.01.008.
- Jaouadi N.Z., Jaouadi B., Hlima H.B., Rekik H., Belhouli M., Hmidi M., Bejar S. Probing the crucial role of Leu31 and Thr33 of the *Bacillus pumilus* CBS alkaline protease in substrate recognition and enzymatic depilation of animal hide. *PLoS One.* 2014;9(9). DOI 10.1371/journal.pone.0108367.
- Jaswal R.K., Kocher G.S., Virk M.S. Production of alkaline protease by *Bacillus circulans* using agricultural residues: A statistical approach. *Ind. J. Biotechnol. (IJBT).* 2008;7(3):356-360.
- Jeong Y.J., Baek S.C., Kim H. Cloning and characterization of a novel intracellular serine protease (IspK) from *Bacillus megaterium* with a potential additive for detergents. *Int. J. Biol. Macromol.* 2018;108: 808-816. DOI 10.1016/j.ijbiomac.2017.10.173.
- Kalwasińska A., Jankiewicz U., Felföldi T., Burkowska-But A., Brzezinska M.S. Alkaline and halophilic protease production by *Bacil-*

- lus luteus* H11 and its potential industrial applications. *Food Technol. Biotechnol.* 2018;56(4):553-561. DOI 10.17113/ftb.56.04.18.5553.
- Ke Y., Yuan X.M., Li J.S., Zhou W., Huang X.H., Wang T. High-level expression, purification, and enzymatic characterization of a recombinant *Aspergillus sojae* alkaline protease in *Pichia pastoris*. *Protein Expr. Purif.* 2018;148:24-29. DOI 10.1016/j.pep.2018.03.009.
- Kebabcı Ö., Cihangir N. Isolation of protease producing novel *Bacillus cereus* and detection of optimal conditions. *Afr. J. Biotechnol.* 2010; 10(7):1160-1164. DOI 10.5897/AJB10.164.
- Khosravi-Darani K., Falahatpishe H.R., Jalali M. Alkaline protease production on date waste by an alkalophilic *Bacillus* sp. 2-5 isolated from soil. *Afr. J. Biotechnol.* 2008;7(10):1536-1542.
- Kobayashi T., Hakamada Y., Adachi S., Hitomi J., Yoshimatsu T., Koike K., Ito S. Purification and properties of an alkaline protease from alkalophilic *Bacillus* sp. KSM-K16. *Appl. Microbiol. Biotechnol.* 1995;43(3):473-481. DOI 10.1007/BF00218452.
- Kobayashi T., Lu J., Li Z., Hung V.S., Kurata A., Hatada Y., Takai K., Ito S., Horikoshi K. Extremely high alkaline protease from a deep-subsurface bacterium, *Alkaliphilus transvaalensis*. *Appl. Microbiol. Biotechnol.* 2007;75(1):71-80. DOI 10.1007/s00253-006-0800-0.
- Kumar C.G., Joo H.S., Koo Y.M., Paik S.R., Chang C.S. Thermostable alkaline protease from a novel marine haloalkalophilic *Bacillus clausii* isolate. *World J. Microbiol. Biotechnol.* 2004;20(4):351-357. DOI 10.1023/B:WIBI.0000033057.28828.a7.
- Latiffi A.A., Salleh A.B., Rahman R.N.Z.R.A., Oslan S.N., Basri M. Secretary expression of the thermostable alkaline protease from *Bacillus stearothermophilus* FI by using native signal peptide and α -factor secretion signal in *Pichia pastoris*. *Genes Genet. Syst.* 2013; 88(2):85-91. DOI 10.1266/ggs.88.85.
- Lin H.H., Yin L.J., Jiang S.T. Functional expression and characterization of keratinase from *Pseudomonas aeruginosa* in *Pichia pastoris*. *J. Agric. Food Chem.* 2009;57(12):5321-5325. DOI 10.1021/jf900417t.
- Liu B., Zhang J., Gu L., Du G., Chen J., Liao X. Comparative analysis of bacterial expression systems for keratinase production. *Appl. Biochem. Biotechnol.* 2014;173(5):1222-1235. DOI 10.1007/s12010-014-0925-z.
- Liu Y., Zhang T., Zhang Z., Sun T., Wang J., Lu F. Improvement of cold adaptation of *Bacillus alcalophilus* alkaline protease by directed evolution. *J. Mol. Catalys. B: Enzymatic.* 2014;106:117-123. DOI 10.1016/j.molcatb.2014.05.005.
- Mathew C.D., Gunathilaka R.M.S. Production, purification and characterization of a thermostable alkaline serine protease from *Bacillus licheniformis* NMS-1. *Int. J. Biotechnol. Mol. Biol. Res.* 2015;6(3): 19-27. DOI 10.5897/IJBMBR2014.0199.
- Mehta V.J., Thumar J.T., Singh S.P. Production of alkaline protease from an alkaliphilic actinomycete. *Bioresour. Technol.* 2006;97(14): 1650-1654. DOI 10.1016/j.biortech.2005.07.023.
- Mothe T., Sultanpuram V.R. Production, purification and characterization of a thermotolerant alkaline serine protease from a novel species *Bacillus caseinilyticus*. *3 Biotech.* 2016;6(1):1-10. DOI 10.1007/s13205-016-0377-y.
- Nejad Z., Yaghmaei S., Hosseini R. Production of extracellular protease and determination of optimal condition by *Bacillus licheniformis* BBRC 100053. *Chem. Eng. Trans.* 2010;1(3):1447-1452. DOI 10.3303/CET1021242.
- Okuda M., Sumitomo N., Takimura Y., Ogawa A., Saeki K., Kawai S., Kobayashi T., Ito S. A new subtilisin family: Nucleotide and deduced amino acid sequences of new high-molecular-mass alkaline proteases from *Bacillus* spp. *Extremophiles.* 2004;8(3):229-235. DOI 10.1007/s00792-004-0381-8.
- Olajuyigbe F.M., Ajele J.O., Ajele J.O. Production dynamics of extracellular protease from *Bacillus* species. *Afr. J. Biotechnol.* 2005; 4(8):776-779.
- Ottesen M., Svendsen I. The Subtilisins. In: *Methods in Enzymology*. Academic Press, 1970;19:199-215. DOI 10.1016/0076-6879(70)19014-8.
- Porres J.M., Benito M.J., Lei X.G. Functional expression of keratinase (kerA) gene from *Bacillus licheniformis* in *Pichia pastoris*. *Biotechnol. Lett.* 2002;24(8):631-636. DOI 10.1023/A:1015083007746.
- Prakasham R.S., Rao C.S., Sarma P.N. Green gram husk-an inexpensive substrate for alkaline protease production by *Bacillus* sp. in solid-state fermentation. *Bioresour. Technol.* 2006;97(13):1449-1454. DOI 10.1016/j.biortech.2005.07.015.
- Radha S., Gunasekaran P. Purification and characterization of keratinase from recombinant *Pichia* and *Bacillus* strains. *Protein Expr. Purif.* 2009;64(1):24-31. DOI 10.1016/j.pep.2008.10.008.
- Rawlings N.D., Waller M., Barrett A.J., Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 2014;42(D1):D503-D509. DOI 10.1093/nar/gkt953.
- Rehman R., Ahmed M., Siddique A., Hasan F., Hameed A., Jamal A. catalytic role of thermostable metalloproteases from *Bacillus subtilis* KT004404 as dehairing and destaining agent. *Appl. Biochem. Biotechnol.* 2017;181(1):434-450. DOI 10.1007/s12010-016-2222-5.
- Shafee N., Aris S., Rahman R., Basri M., Salleh A. Optimization of environmental and nutritional conditions for the production of alkaline protease by a newly isolated bacterium *Bacillus cereus* strain 146. *J. Appl. Sci. Res.* 2005;1(1):1-8.
- Shaheen M., Shah A., Hameed A., Hasan F. Influence of culture conditions on production and activity of protease from *Bacillus subtilis* BS1. *Pak. J. Bot.* 2008;40(5):2161-2169.
- Sharma A., Sharma V., Saxena J., Yadav B., Alam A., Prakash A. Optimization of protease production from bacteria isolated from soil. *Appl. Res. J.* 2015;1(7):388-394.
- Sharma K., Kumar R., Vats S., Gupta A. Production, partial purification and characterization of alkaline protease from *Bacillus aryabhattai* K3. *Int. J. Adv. Pharm. Biol. Chem.* 2014;3(2):290-298.
- Sharma K.M., Kumar R., Panwar S., Kumar A. Microbial alkaline proteases: Optimization of production parameters and their properties. *J. Genet. Eng. Biotechnol.* 2017;15:115-126. DOI 10.1016/j.jgeb.2017.02.001.
- Sharmin S., Hossain T., Anwar M. Isolation and characterization of a protease producing bacteria *Bacillus amovivorus* and optimization of some factors of culture conditions for protease production. *J. Biol. Sci.* 2005;5(3):358-362. DOI 10.3923/jbs.2005.358.362.
- Shih J. Construction of bacillus licheniformis t1 strain and fermentation production of crude enzyme extract therefrom. Patent No. US20050032188A1, 2005.
- Shikha, Sharan A., Darmwal N.S. Improved production of alkaline protease from a mutant of alkalophilic *Bacillus pantotheneticus* using molasses as a substrate. *Bioresour. Technol.* 2007;98(4):881-885. DOI 10.1016/j.biortech.2006.03.023.
- Shivanand P., Jayaraman G. Production of extracellular protease from halotolerant bacterium, *Bacillus aquimaris* strain VITP4 isolated from Kumta coast. *Process Biochem.* 2009;44(10):1088-1094. DOI 10.1016/j.procbio.2009.05.010.
- Shivasharana C.T., Naik G.R. Ecofriendly applications of thermostable alkaline protease produced from a *Bacillus* sp. JB-99 under solid state fermentation. *Int. J. Environ. Sci.* 2012;3(3):956-964. DOI 10.6088/ijes.2012030133003.
- Shu M., Shen W., Yang S., Wang X., Wang F., Wang Y., Ma L. High-level expression and characterization of a novel serine protease in *Pichia pastoris* by multi-copy integration. *Enzyme Microb. Technol.* 2016;92:56-66. DOI.10.1016/j.enzmictec.2016.06.007.
- Singh S.K., Tripathi V.R., Jain R.K., Vikram S., Garg S.K. An antibiotic, heavy metal resistant and halotolerant *Bacillus cereus* SIU1 and its thermoalkaline protease. *Microb. Cell Fact.* 2010;9(1):59. DOI 10.1186/1475-2859-9-59.
- Smith E.L., Markland F.S., Kasper C.B., DeLange R.J., Landon M., Evans W.H. The complete amino acid sequence of two types of subtilisin, BPN' and Carlsberg. *J. Biol. Chem.* 1966;241(24):5974-5976.
- Srinivasan T., Das S., Balakrishnan V., Philip R., Kannan N. Isolation and characterization of thermostable protease producing bacteria from tannery industry effluent. *Recent Res. Sci. Technol.* 2009;1(2): 63-66.

- Srinubabu G., Lokeswari N., Jayaraju K. Screening of nutritional parameters for the production of protease from *Aspergillus oryzae*. *Electr. J. Chem.* 2007;4(2):208-215. DOI 10.1155/2007/915432.
- Strausberg S.L., Ruan B., Fisher K.E., Alexander P.A., Bryan P.N. Directed coevolution of stability and catalytic activity in calcium-free subtilisin. *Biochemistry.* 2005;44(9):3272-3279. DOI 10.1021/bi047806m.
- Takenaka S., Yoshinami J., Kuntiya A., Techapun C., Leksawasdi N., Seesuriyachan P., Chaiyaso I., Watanabe M., Tanaka K., Yoshida K. Characterization and mutation analysis of a halotolerant serine protease from a new isolate of *Bacillus subtilis*. *Biotechnol. Lett.* 2018; 40(1):189-196. DOI 10.1007/s10529-017-2459-2.
- Thys R.C.S., Guzzon S.O., Cladera-Olivera F., Brandelli A. Optimization of protease production by *Microbacterium* sp. in feather meal using response surface methodology. *Process Biochem.* 2006; 41(1):67-73. DOI 10.1016/j.procbio.2005.03.070.
- Tufvesson P., Lima-Ramos J., Nordblad M., Woodley J.M. Guidelines and cost analysis for catalyst production in biocatalytic processes. *Org. Process Res. Dev.* 2010;15(1):266-274. DOI 10.1021/op1002165.
- Usharani B., Muthuraj M. Production and characterization of protease enzyme from *Bacillus laterosporus*. *Afr. J. Microbiol. Res.* 2010; 4(11):1057-1063.
- Vaithanomsat P., Malapant T., Apiwattanapiwat W. Silk degumming solution as substrate for microbial protease production. *Nat. Sci.* 2008;42:543-551.
- Voordouw G., Milo C., Roche R.S. Role of bound calcium ions in thermostable, proteolytic enzymes. Separation of intrinsic and calcium ion contributions to the kinetic thermal stability. *Biochemistry.* 1976;15(17):3716-3724. DOI 10.1021/bi00662a012.
- Zambare V., Nilegaonkar S., Kanekar P. A novel extracellular protease from *Pseudomonas aeruginosa* MCM B-327: enzyme production and its partial characterization. *New Biotechnol.* 2011;28(2): 173-181. DOI 10.1016/j.nbt.2010.10.002.
- Zhao H.Y., Feng H. Engineering *Bacillus pumilus* alkaline serine protease to increase its low-temperature proteolytic activity by directed evolution. *BMC Biotechnol.* 2018;18(1):34. DOI 10.1186/s12896-018-0451-0.
- Zhao H.Y., Wu L.Y., Liu G., Feng H. Single-site substitutions improve cold activity and increase thermostability of the dehairing alkaline protease (DHAP). *Biosci. Biotechnol. Biochem.* 2016;80(12):2480-2485. DOI 10.1080/09168451.2016.1230005.
- Zhou K., Dong Y., Zheng H., Chen B., Mao R., Zhou L., Wang Y. Expression, fermentation, purification and lyophilisation of recombinant Subtilisin QK in *Pichia pastoris*. *Process Biochem.* 2017; 54:1-8. DOI 10.1016/j.procbio.2016.12.028.

ORCID ID

S.V. Shekhovtsov orcid.org/0000-0001-5604-5601
E.G. Pershina orcid.org/0000-0003-2658-7906
S.E. Peltek orcid.org/0000-0002-3524-0456

Благодарности. Работа выполнена при поддержке Курчатковского геномного центра ИЦиГ СО РАН (075-15-2019-1662) и бюджетного проекта ИЦиГ СО РАН № 0259-2019-0005.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 17.11.2020. После доработки 17.12.2020. Принята к публикации 21.12.2020.