Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(2):320-329

doi 10.18699/vjgb-25-35

CropGene: программный комплекс анализа геномных и транскриптомных данных сельскохозяйственных растений

А.Ю. Пронозин [1, 2] ✉, Д.И. Каретников [1, 2], Н.А. Шмаков [1, 2], М.Е. Бочарникова [1, 2], С.Д. Афонникова [1, 2], Д.А. Афонников [1, 2], Н.А. Колчанов [1, 2]

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

Аннотация. В настоящее время селекция сельскохозяйственных растений все больше опирается на использование молекулярно-биологических данных о генетических последовательностях, что позволяет существенно ускорить селекционный процесс создания новых сортов растений за счет геномного редактирования. Эти данные имеют большой объем, разнообразны и требуют для анализа затрат большого количества ресурсов, как трудовых, так и вычислительных. Анализ данных с такими объемом и сложностью может быть эффективным лишь с применением современных методов биоинформатики, включающих алгоритмы идентификации генов, предсказания их функции, оценку влияния эффекта мутации на фенотип растений. Такой анализ в последнее время стал невозможным без использования интегрированных программных комплексов, решающих задачи разного уровня за счет выполнения вычислительных конвейеров. В статье описан программный комплекс CropGene, разработанный для комплексного анализа геномных и транскриптомных данных сельскохозяйственных растений. Система включает в себя несколько блоков биоинформатического анализа, таких как анализ вариаций генов, сборка геномов и транскриптомов, а также аннотация генов и белков. В комплексе реализованы новые методы анализа длинных некодирующих РНК, белковых доменов, поиска и анализа полиморфизмов и полногеномного исследования ассоциаций. В работе представлены примеры применения CropGene для анализа сельскохозяйственных организмов, таких как *Solanum tuberosum*, *Zea mays*. С помощью данного программного пакета найдены: генетические маркеры, объясняющие до 50 % изменчивости параметров окраски семян; потенциальные гены, которые могут стать перспективным материалом для получения сортов картофеля; более 100 тыс. новых длинных некодирующих РНК. Также обнаружены ортогруппы, доменная структура которых проявляет заметное сходство с доменной архитектурой характерных секретируемых фосфолипаз А2. Таким образом, CropGene представляет собой важный инструмент для ученых и практиков, работающих в области агробиотехнологий и генетики растений.

Ключевые слова: биоинформатический конвейер; программный пакет; SNP; анализ полиморфизмов; идентификация генов

Для цитирования: Пронозин А.Ю., Каретников Д.И., Шмаков Н.А., Бочарникова М.Е., Афонникова С.Д., Афонников Д.А., Колчанов Н.А. CropGene: программный комплекс анализа геномных и транскриптомных данных сельскохозяйственных растений. *Вавиловский журнал генетики и селекции*. 2025;29(2):320-329. doi 10.18699/vjgb-25-35

Финансирование. Работа по созданию программного комплекса CropGene выполнена при поддержке бюджетного проекта № FWNR-2022-0020.

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

CropGene: a software package for the analysis of genomic and transcriptomic data of agricultural plants

A.Yu. Pronozin [1, 2] ✉, D.I. Karetnikov [1, 2], N.A. Shmakov [1, 2], M.E. Bocharnikova [1, 2], S.D. Afonnikova [1, 2], D.A. Afonnikov [1, 2], N.A. Kolchanov [1, 2]

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

Abstract. Currently, the breeding of agricultural plants is increasingly based on the use of molecular biological data on genetic sequences, which makes it possible to significantly accelerate the breeding process, create new plant varieties through genomic editing. These data have a large volume, variety and require a large amount of resources, both labor and computing, to analyze the costs. Data analysis of such volume and complexity can be effective only when using modern bioinformatics methods, which include algorithms for identifying genes, predicting their function, and evaluating the effect of mutation on plant phenotype. Such an analysis has recently become impossible without the use of integrated software systems that solve problems of different levels by executing computational pipelines. The paper de-

scribes the CropGene software package developed for the comprehensive analysis of genomic and transcriptomic data of agricultural plants. CropGene includes several blocks of bioinformatic analysis, such as analysis of gene variations, assembly of genomes and transcriptomes, as well as annotation of genes and proteins. CropGene implements new methods for analyzing long non-coding RNAs, protein domains, searching and analyzing polymorphisms, and genome-wide association research. CropGene has a user-friendly interface and supports working with various types of data, which greatly simplifies its use for researchers who do not have deep knowledge in the field of bioinformatics. The paper provides examples of the use of CropGene for the analysis of agricultural organisms such as *Solanum tuberosum* and *Zea mays*. With CropGene, genetic markers have been identified that explain up to 50 % of the variability in seed color parameters; potential genes that may become promising material for producing potato varieties; more than 100 thousand new long non-coding RNAs. Orthogroups were also found, the domain structure of which shows a marked similarity with the domain architecture of characteristic secreted A2 phospholipases. Thus, CropGene is an important tool for scientists and practitioners working in the field of agrobiotechnology and plant genetics.

Key words: bioinformatics pipeline; software package; SNP; analyzing polymorphisms; identification of genes

For citation: Pronozin A.Yu., Karetnikov D.I., Shmakov N.A., Bocharnikova M.E., Afonnikova S.D., Afonnikov D.A., Kolchanov N.A. CropGene: a software package for the analysis of genomic and transcriptomic data of agricultural plants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(2):320-329. doi 10.18699/vjgb-25-35

Введение

В настоящее время селекция сельскохозяйственных растений все больше опирается на использование молекулярно-биологических данных о генетических последовательностях, что позволяет существенно ускорить селекционный процесс (Хлесткина, 2013) создания новых сортов растений за счет геномного редактирования. Эти данные имеют большой объем, разнообразны и требуют для анализа затрат большого количества ресурсов, как трудовых, так и вычислительных. Анализ данных такого объема и сложности может быть эффективным лишь с применением современных методов биоинформатики, которые включают алгоритмы идентификации генов, предсказания их функции, оценку влияния эффекта мутации на фенотип растений. Такой анализ в последнее время стал невозможным без использования компьютерного моделирования и алгоритмов глубокого машинного обучения. Для автоматизации обработки данных в биоинформатике разрабатываются технологии вычислительных конвейеров.

При анализе геномных и транскриптомных данных сельскохозяйственных растений можно выделить несколько важных задач. Одна из них – изучение генетического разнообразия, которое является важнейшей основой для поиска генов устойчивости растений к биотическим и абиотическим стрессам и создания новых высокоадаптивных и урожайных сортов сельскохозяйственных культур. Изучение генетического разнообразия осуществляется с применением различных методов генетического анализа. В частности, используются генетические маркеры (Хлесткина, 2013). Среди них важное место занимают маркеры, основанные на однонуклеотидных полиморфизмах (single-nucleotide polymorphism, SNP), это замены единичных нуклеотидов в геноме, которые в популяции растений встречаются с различными частотами (Сухарева, Кулуев, 2018). Анализ SNP широко применяют для изучения аллельного полиморфизма, анализа гаплотипа и родословных, а также для генотипирования и построения генетических карт.

Помимо анализа SNP, для исследования генетического разнообразия используется изучение вариаций числа копий (copy number variation, CNV). Это тип генетического полиморфизма, при котором определенные участки генома у разных особей демонстрируют различие в количестве копий. Изменения включают делеции или дупликации генов или групп сцепленных генов. Такие вариации могут иметь протяженность до нескольких миллионов пар нуклеотидов.

Благодаря технологиям высокопроизводительного секвенирования нового поколения можно получить информацию об однонуклеотидных заменах в масштабе генома для популяции из сотен образцов. Идентификация SNP возможна с помощью стратегий полногеномного секвенирования (WGS) и генотипирования путем секвенирования (GBS) (Scheben et al., 2017). Метод GBS – более быстрый и экономически эффективный по сравнению с WGS. Протокол GBS позволяет секвенировать фрагменты геномной ДНК лишь вблизи сайтов рестрикции. За счет этого процесс секвенирования существенно удешевляется. При этом покрытие генома фрагментарно, а количество SNP оказывается меньше, чем при полногеномном секвени-

Ключевые понятия

Интронные днРНК – перекрываются с интроном гена

Антисмысловые днРНК – ориентированы против направления транскрипции гена, кодирующего белок

Межгенные днРНК – расположены между двумя локусами генов

Полногеномные ассоциации (Genome-Wide Association Studies, GWAS) – метод исследования геномов, целью которого является обнаружение статистических связей между генетическими вариациями и определенными фенотипическими признаками

Транскриптом – совокупность всех транскриптов, присутствующих в клетке на определенной стадии развития или в определенных физиологических условиях

Генная сеть – группа координированно функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования генных сетей

ровании. Тем не менее данных, полученных при помощи протокола GBS, оказывается вполне достаточно, чтобы с приемлемой точностью характеризовать генетическое разнообразие популяций сельскохозяйственных растений. Данные, полученные методом GBS, применяются также для полногеномных исследований ассоциаций (GWAS). Этот инструмент разработан для выявления генов, влияющих на сложные количественные признаки (Burghardt et al., 2017).

Помимо фундаментальных результатов о генетических механизмах формирования интересующих признаков, полногеномное исследование ассоциаций также позволяет найти генетические маркеры, которые в дальнейшем можно использовать непосредственно в селекционных программах (Tsai et al., 2010; Zatybekov et al., 2017; Larkin et al., 2019; Muqaddasi et al., 2020).

Еще один блок биоинформатических задач для сельскохозяйственных растений – это сборка последовательностей геномов, транскриптомов. Реконструкция генома представляет собой первый и основной этап при геномном анализе. Сборки геномов позволяют получить информацию о белок-кодирующих генах, мобильных генетических элементах. Транскриптом, в свою очередь, выполняет роль связующего звена между геномом организма и его фенотипическими признаками (Velculescu et al., 1997). На сегодняшний день самым популярным методом транскриптомных исследований является технология RNA-seq – массовое высокопроизводительное секвенирование транскриптома с помощью платформ для секвенирования нового поколения (Shendure, 2008).

Наиболее распространенное применение RNA-seq – идентификация дифференциальной экспрессии генов в экспериментах типа «опыт-контроль» (Drewe et al., 2013). Однако, помимо этого, технология имеет и другие важные применения: реконструкция транскриптома *de novo* (Cardoso-Silva et al., 2014), обнаружение полиморфизмов (Piskol et al., 2013) и поиск неизвестных вариантов сплайсинга. Секвенирование и реконструкция геномов немодельных организмов чаще всего сопровождается также секвенированием транскриптома, что облегчает аннотацию генома, предсказание и функциональную аннотацию белок-кодирующих генов. Однако за счет высокого геномного и морфологического разнообразия у видов, вызванного структурными вариациями, один референсный геном не способен охватить все изоформы гена одного вида. Для решения данной проблемы используется концепция пангенома и пантранскриптома.

Реконструкция геномов и транскриптомов для популяции позволяет получить и исследовать пангеномы или пантранскриптомы растений (Пронозин и др., 2021). Концепция пангенома подразумевает охват последовательностей, подверженных структурным вариациям и, возможно, отсутствующих в референсной последовательности каждого представителя вида (Vernikos et al., 2015). Во многих работах отмечается, что анализ пангеномов и пантранскриптомов повышает эффективность исследования и общее количество предсказанных генов по сравнению с использованием генома одного представителя вида (Jin et al., 2016). Это позволило повысить точность и полноту исследуемого набора генов.

Еще одно направление биоинформатического анализа – аннотация генома и трансриптома. Для генов, кодирующих белки, при аннотации их важную часть составляет идентификация белковых доменов, структурный фрагмент белка, выступающий в качестве независимой функциональной единицы. Он может образовывать уникальную структуру или быть частью мультидоменных белков, функционируя как самостоятельно, так и в сочетании с другими доменами. Для функциональной идентификации белков также важно производить поиск ортологов в уже известных геномах, белков, которые в разных организмах выполняют одинаковые функции.

Отметим также, что более 90 % всех транскриптов не транслируются в белки (Carninci et al., 2005) и являются некодирующими последовательностями. Некодирующие РНК (нкРНК) выполняют в геномах растений ряд важнейших функций, связанных с регуляцией экспрессии генов, гомеостазом физиологических параметров растений. Один из важных классов нкРНК – длинные некодирующие РНК (днРНК) (Назипова, 2021). ДнРНК представляют собой класс линейных или кольцевых молекул РНК длиной от 200 нуклеотидов, не кодирующих белок (Kim, Sung, 2012). Участие днРНК обнаружено в регуляции экспрессии генов, формировании структуры макромолекулярных комплексов, во взаимодействии с белками, в патогенезе. К настоящему времени идентифицировано более полумиллиона последовательностей днРНК для различных организмов.

Данные об уровнях экспрессии генов, полученные из транскриптомных экспериментов, широко используются для реконструкции генных сетей (Johnson, Krishnan, 2022). Генные сети, в свою очередь, позволяют моделировать динамику конкретных процессов в организме и прогнозировать его поведение в различных условиях.

В настоящей работе представлена система CropGene для комплексного анализа геномных, транскриптомных данных, особенностей молекулярной эволюции генов сельскохозяйственных растений. В систему входят следующие блоки биоинформатического анализа данных: анализ вариаций генов, сборка геномов и транскриптомов, аннотация генов и белков.

Материалы и методы

Программный комплекс CropGene включает программные пакеты, представленные на рис. 1.

В структуру программного комплекса входят блоки для решения задач:

Программный модуль анализа полногеномных ассоциаций реализует следующие этапы анализа:

- анализ данных фенотипирования. Обработка данных фенотипирования производится с использованием пакетов R, pastecs, psych (Grosjean et al., 2018),
- обработка данных генотипирования. Этот этап направлен на процессинг данных генотипирования, полученных методом генотипирования на микрочипах и методом GBS. Обработка включает в себя проверку качества сырых прочтений, картирование на референсный геном с помощью BWA-MEM (Li, 2013) и поиск полиморфизмов с использованием vcftools (Danecek et al., 2011). Варианты, определенные вышеуказанными мето-

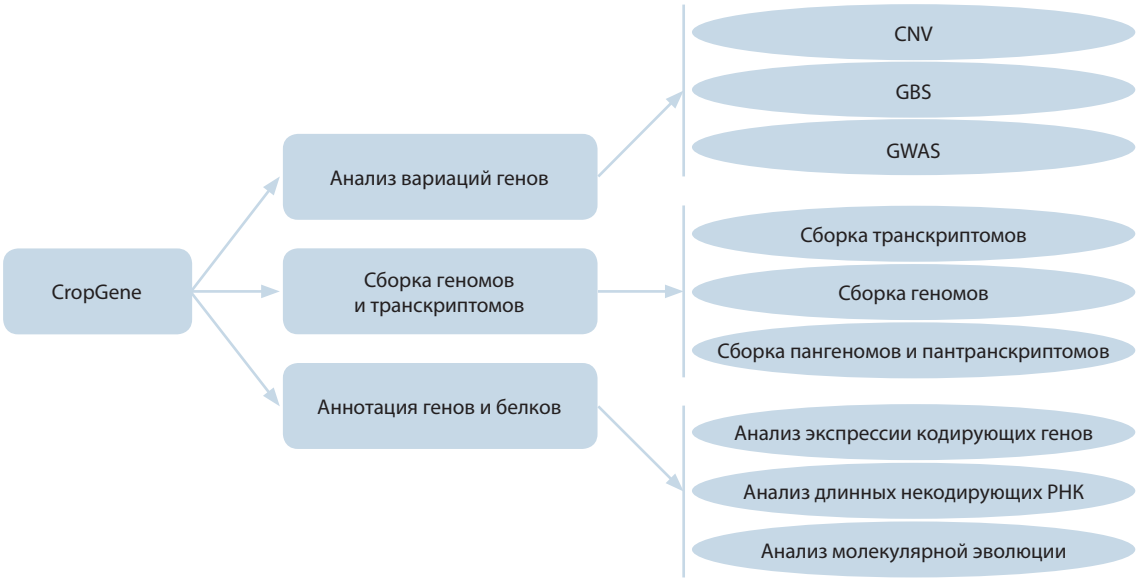

Рис. 1. Схема программного комплекса CropGene с указанием основных блоков анализа (скругленные прямоугольники в центре) и конкретных решаемых задач (овалы справа).

дами генотипирования, фильтруют по качеству, частоте минорного аллеля, гетерозиготности и количеству пропущенных данных. Этот этап осуществляется инструментом bcftools (Danecek et al., 2021). Для восстановления пропущенных данных генотипирования используют BEAGLE 5.2 (Browning et al., 2018),

- полногеномный анализ ассоциаций. На данном этапе осуществляется непосредственно полногеномный анализ ассоциаций, реализуемый на языке программирования R при помощи функций пакета "GAPIT3" (Wang, Zhang, 2021),
- приоритизация генов в обнаруженных локусах. Модуль полногеномного анализа ассоциаций направлен на выявление генов-кандидатов, связанных с интересующими признаками. Прежде всего с использованием функций пакета R "genetics" определяются границы локусов, которые включают в себя значимо ассоциированные с фенотипом варианты. Далее, основываясь на опубликованных данных по экспрессии генов у исследуемого организма и на ресурсах платформы Knetminer (Hassani-Pak et al., 2021), проводят приоритизацию генов среди обнаруженных локусов.

Программный модуль анализа CNV направлен на решение задач по оценке и анализу вариаций количества копий в геноме. Он реализует несколько этапов анализа:

- наборы сырых прочтений фильтруются по качеству и длине с помощью программы fastp (Chen et al., 2018). Далее фильтрованные и обработанные наборы прочтений картируются на референсный геном картофеля с помощью программы BWA (Li, Durbin, 2009). Дупликаты в картированных прочтениях маркируются, удаляются, после чего происходят сортировка и индексирование прочтений с помощью программы SAMtools (Li et al., 2009),
- полученные файлы формата BAM используются как входные данные в программе CNVpytor (Suvakov et al., 2021). Вариации количества копий выявлялись на всех хромосомах референсного генома. Найденные CNV фильтруются следующим образом: длина более 1 т. п. н., p-value < 0.01, q0 < 50 % и pN < 50 %. Для сопоставления обнаруженных CNV с генами референсного генома используется R пакет intansv (Jia et al., 2020),
- для последующей обработки список CNV представлен в виде матрицы, в которой строки соответствуют конкретному генотипу, а столбцы – гену референсного генома. Каждый элемент матрицы представлен в трех вариантах: +1 (потенциальная дупликация), –1 (потенциальная делеция) и 0 (отсутствие значимого CNV). Далее проводится анализ главных компонент (PCA) с помощью пакета Scikit-learn v1.1.2, что позволяет оценить генетическое разнообразие (Pedregosa et al., 2011).

Биоинформатический конвейер GBS-DP направлен на анализ данных, полученных методом GBS, состоит из трех основных этапов (Pronozin et al., 2023):

- предобработка данных включает проверку качества сырых прочтений FastQC, удаление адаптеров fastp (Chen et al., 2018) и построение индекса референсного генома,
- поиск полиморфизмов состоит из картирования предобработанных прочтений на референсный геном BwaMem2 (Li, Durbin, 2009), сортировки картированных прочтений Samtools (Li et al., 2009) и поиска однонуклеотидных полиморфизмов Bcftools (Li, 2011),
- анализ генетического разнообразия разделяется на два варианта обработки данных: если полученные данные превышают занимаемый объем памяти в 1 Тб и если полученные данные не превышают занимаемый объем памяти в 1 Тб. Выбор соответствующей опции осуществляется автоматически и связан с увеличенной нагрузкой на оперативную память компьютера при работе с большими данными. Для анализа главных компонент и построения филогенетического дерева на основе фильтрованных SNP применяется пакет R – SNPrelate (Zheng, 2013).

Программный модуль по реконструкции транскриптома реализует следующие этапы анализа:

- непосредственно сборка последовательностей контигов из прочтений библиотек RNA-seq. На этой стадии используются программы: Trinity (Grabherr et al., 2011), Trans-ABySS (Robertson et al., 2010), rnaSpades (Bushmanova et al., 2019),
- объединение полученных наборов контигов и удаление избыточности программой tr2aacds.pl из конвейера EvidentialGene,
- оценка качества полученных последовательностей; программа BUSCO (Simão et al., 2015) используется для определения полноты транскриптома; программа kallisto (Bray et al., 2016) показывает, насколько полно исходные библиотеки прочтений были задействованы для реконструкции транскриптома; rnaQUAST (Bushmanova et al., 2016) оценивает различные метрики качества полученного транскриптома, в том числе наличие гомологии с последовательностью генома организма, или генома близкородственного организма в случае работы с немодельным видом.

Программный модуль реконструкции и анализа пангенома реализует следующие шаги анализа:

- реконструкция каждого генома на основе парных коротких прочтений с помощью геномного сборщика MaSuRCA (Zimin et al., 2013),
- маскирование мобильных генетических элементов с использованием RepeatMasker и дальнейшая *de novo* аннотация реконструированных маскированных геномов с трансляцией открытых рамок считывания с помощью программы AUGUSTUS (Stanke et al., 2004),
- выявление ортологических групп в наборе аминокислотных последовательностей, полученных на основе открытых рамок считывания, с помощью OrthoFinder (Emms, Kelly, 2019).

Программный модуль оценки экспрессии генов. В этом модуле оценка экспрессии генов может проводиться на основе как референсного генома, так и транскриптома, реконструированного *de novo*:

- для подсчета экспрессии генов референсного генома выполняется выравнивание прочтений библиотек RNA-seq на последовательность генома с помощью программы Dart (Lin, Hsu, 2018). Далее применяется разметка генома с позициями известных генов для подсчета количества прочтений, картированных на каждый ген, с помощью программы featureCounts (Liao et al., 2014),
- для оценки экспрессии транскриптов из реконструированного ранее транскриптома используется программа kallisto, которая проводит так называемые псевдовыравнивания прочтений для определения, к какому транскрипту они принадлежат, на основании чего далее подсчитываются уровни экспрессии.

Биоинформатический конвейер ICAnnoLncRNA направлен на выявление и аннотацию днРНК, реализует три этапа обработки транскриптомных последовательностей (Pronozin, Afonnikov, 2023):

1) контроль качества. Данный этап включает две операции: построение индексного файла для геномной последовательности программой gmap (Wu, Watanabe, 2005) и обучение модели распознавания днРНК программой LncFinder v1.1.4 (Han et al., 2019),
2) идентификация днРНК. Этот блок состоит из трех этапов: предсказание кандидатов в днРНК из входного набора транскриптов с помощью метода LncFinder; фильтрация полученных последовательностей-кандидатов на основе идентификации трансмембранных сегментов в ОРС; выравнивание фильтрованных последовательностей-кандидатов днРНК на референсный геном,
3) анализ пантранскриптомов. Аннотация включает определение типов последовательностей днРНК по выравниванию на гены, кодирующие белок, выявление консервативных днРНК, анализ структурных особенностей днРНК и их экспрессии.

Программный модуль анализа эволюции белков OrthoDOM реализует четыре ключевых этапа анализа белковых последовательностей:

1) проводятся валидация входных данных и проверка наличия функциональных доменов, заданных пользователем у референсных белков,
2) проверяется наличие ключевых доменов в референсных последовательностях,
3) выполняется работа программы Orthofinder для исследуемых протеомов,
4) производится проверка выявленных ортологов по наличию в их последовательности наборов заданных доменов.

Результаты и обсуждение

Модули программного комплекса CropGene были применены при решении различных задач биоинформатического анализа геномов и транскриптомов сельскохозяйственных растений.

Программный конвейер, выявляющий CNV на основе полногеномных данных, был использован ранее в работе по анализу структуры геномов картофеля отечественных сортов (Karetnikov et al., 2023). Он позволил найти все вариации количества копий в геномах картофеля и провести сравнительный анализ количества копий генов с южноамериканским картофелем. Анализ дал возможность обнаружить, что частота встречаемости CNV в 4 из 48 известных генов, связанных с формированием клубней и реакцией на фотопериод, различается между геномами российских сортов, адаптированных к длинному световому дню в северных широтах, и местных андских сортов, приспособленных к короткому световому дню.

С помощью GBS-DP в настоящей работе было проанализировано 219 сортов ячменя. Обнаружено 61 620 SNP. На основе найденных полиморфизмов построены кластеризация – методом главных компонент (рис. 2) и дендрограмма – методом иерархической кластеризации (рис. 3).

Модуль полногеномного анализа ассоциаций был использован в работе по поиску генов-кандидатов мягкой озимой пшеницы, связанных с предуборочным прорастанием и красной окраской зерна (Afonnikova et al., 2024). Помимо обнаружения генетических маркеров, которые объясняют до 50 % изменчивости красноты зерна, в работе удалось выявить два гена-кандидата, которые связаны с формированием окраски зерна. Первый ген,

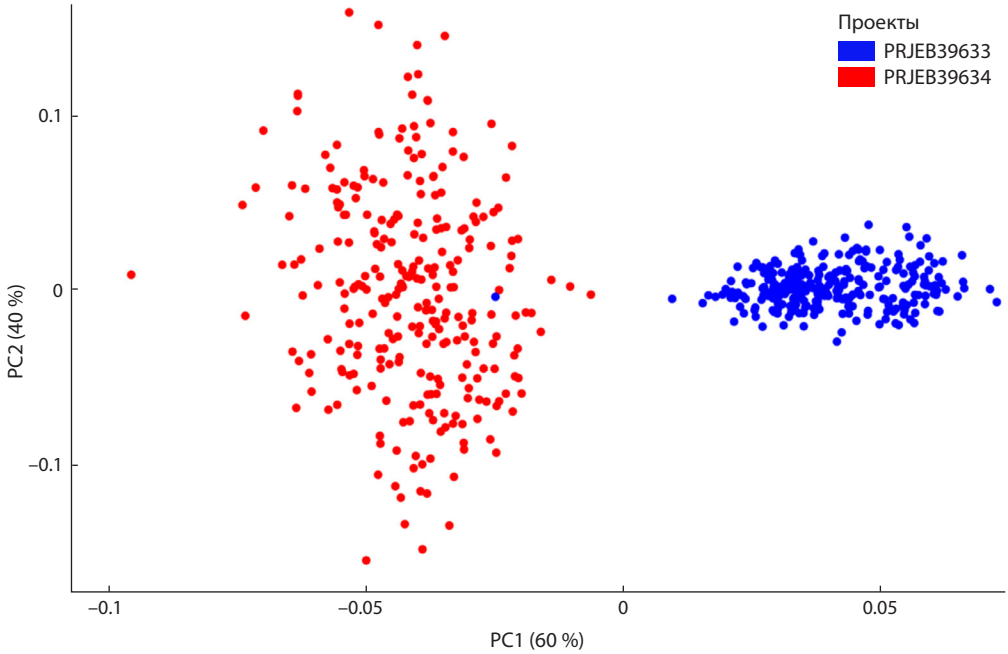

Рис. 2. Визуализация генетического разнообразия 219 библиотек ячменя методом PCA на основе выявленных однонуклеотидных полиморфизмов.

По осям *X* и *Y* направлены первая и вторая главные компоненты соответственно.

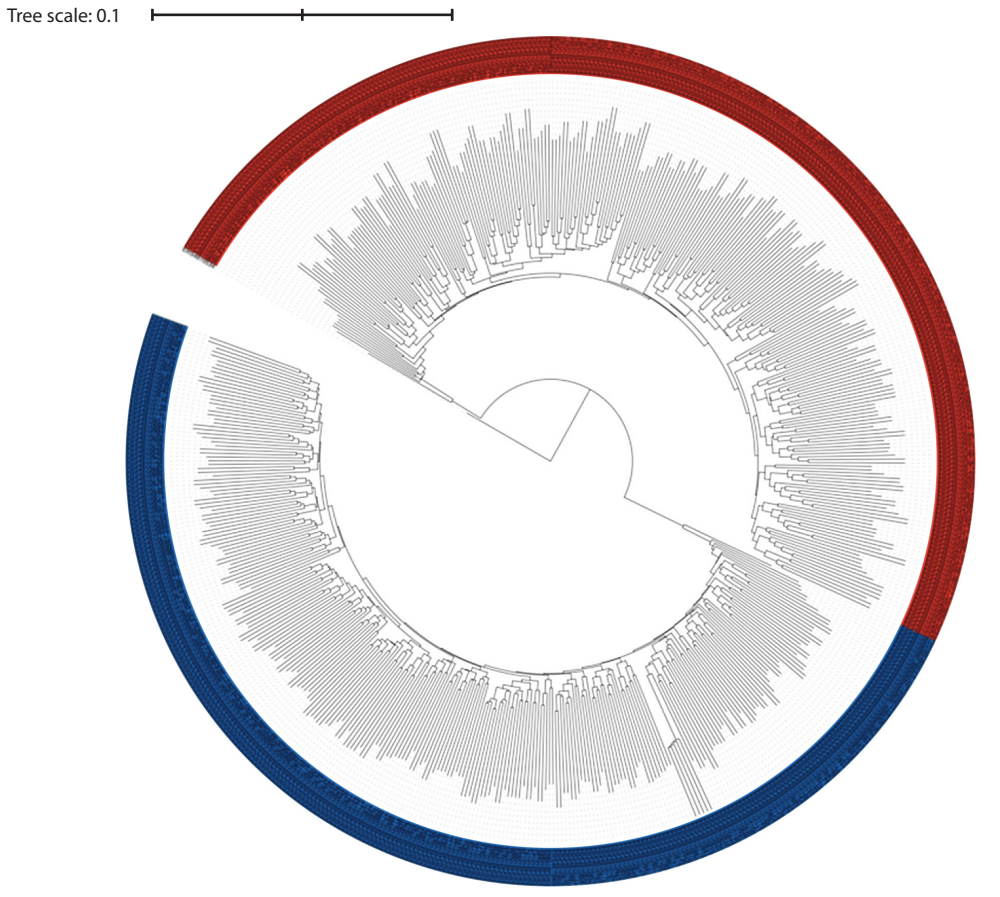

Рис. 3. Дендрограмма, характеризующая генетическое разнообразие 219 библиотек ячменя, построенная методом иерархической кластеризации по данным GBS.

Дендрограмма построена на основе найденных однонуклеотидных полиморфизмов.

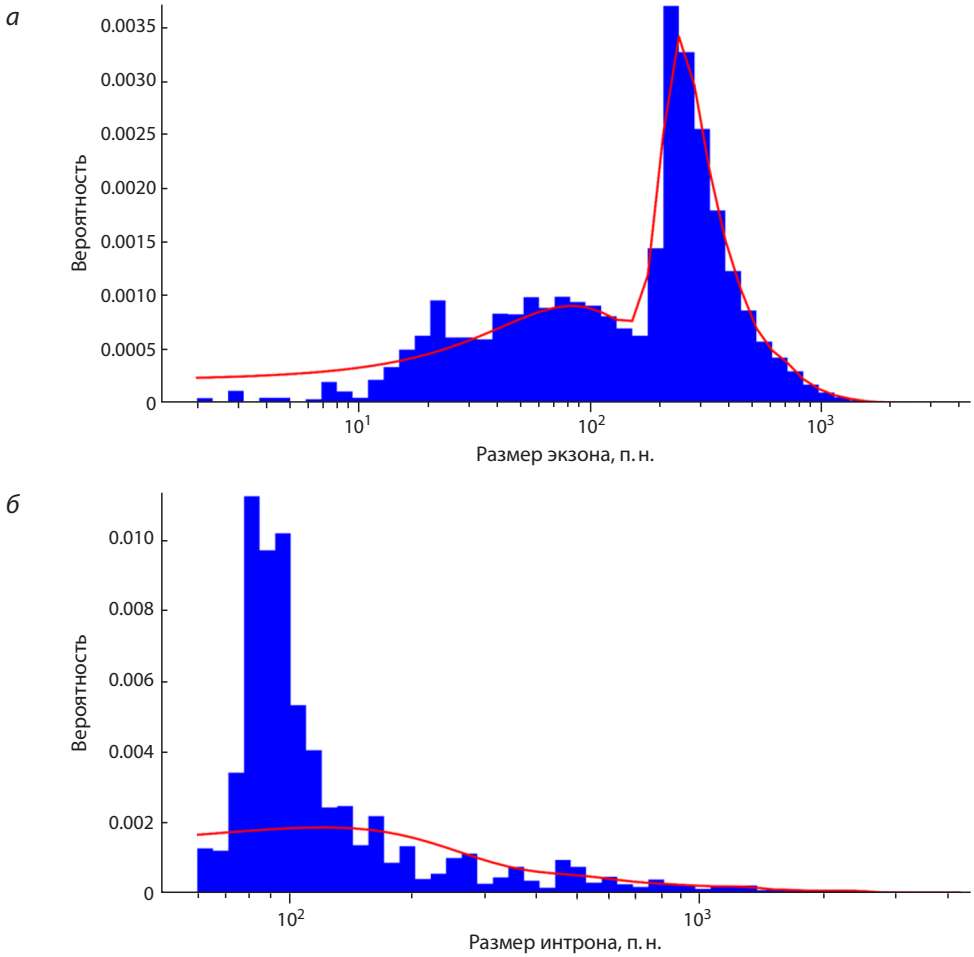

Рис. 4. Отношение количества экзонов, приходящихся на одну днРНК (*а*), и распределение размера интронов относительно днРНК (*б*).

TraesCS1D02G319700, расположен на хромосоме 1D, участвует в синтезе флавонолов и биосинтезе флавоноидов. Другой ген, TraesCS7B02G482000, локализируется на хромосоме 7B и кодирует фитоен синтазу, вовлеченную в один из начальных этапов синтеза каротиноидов. Для устойчивости мягкой пшеницы к предуборочному прорастанию главным геном-кандидатом является ген TraesCS6B02G147900, кодирующий белок морфогенеза алейронового слоя. Также были обнаружены генетические маркеры, которые объясняют до 50 % изменчивости параметров окраски «светлота», «краснота» и «синева» и до 25.3 % изменчивости оценки предуборочного прорастания – индекса прорастания на молочной/восковой стадии развития зерна.

На основе модуля транскриптомного анализа ранее была проведена сборка транскриптома четырех сортов картофеля *Solanum tuberosum* group *phureja* (Бинтье, Сиверский, Сударыня, Евразия) и дикорастущего *S. stoloniferum* L. Обнаружены гены, кодирующие белки семейства Nucleotide-binding site – Leucine rich repeats (NBS-LRR), участвующие в формировании иммунного ответа растений (Kochetov et al., 2021). Установлено, что репертуары этих генов у исследованных сортов картофеля и у дикорастущего пасленового существенно различаются, что согласуется с имеющими данными о быстрой эволюции этих генов. Некоторые из генов семейства NBS-LRR, наблюдаемых в этой работе, ранее не были обнаружены у Solanaceae и у картофеля в частности. Эти гены могут стать перспективным материалом для получения сортов картофеля, более устойчивого к воздействию различных патогенов и паразитов.

С помощью конвейера ICAnnoLncRNA проанализировано 54 транскриптома ячменя. Выявлено 143279 новых днРНК, из них 29987 принадлежат к классу интронных днРНК, 48369 – межгенных, 64923 – антисмысловых днРНК. Анализ структуры днРНК показал, что большинство из них (60 %) содержат лишь один экзон. При этом средняя длина экзона составляет 371 нуклеотид, небольшая доля экзонов имеет длину до 10 п. н., основная часть имеет длины от 10 до 1000 п. н., и их распределение имеет два характерных пика, один, широкий, с максимумом в области 100 п. н., а другой, узкий, в области 250–300 п. н. (рис. 4). Анализ тканеспецифичности показал, что большинство днРНК экспрессируется в тканях ростков (seedling) ячменя (рис. 5, *а*). Это наблюдается как для консервативных, так и для неконсервативных днРНК, а также характерно для мРНК (см. рис. 5, *а*).

Применение конвейера OrthoDOM для выявления белков семейства фосфолипаз А2 в ячмене и пшенице позволило подтвердить наличие их в геномах этих растений.

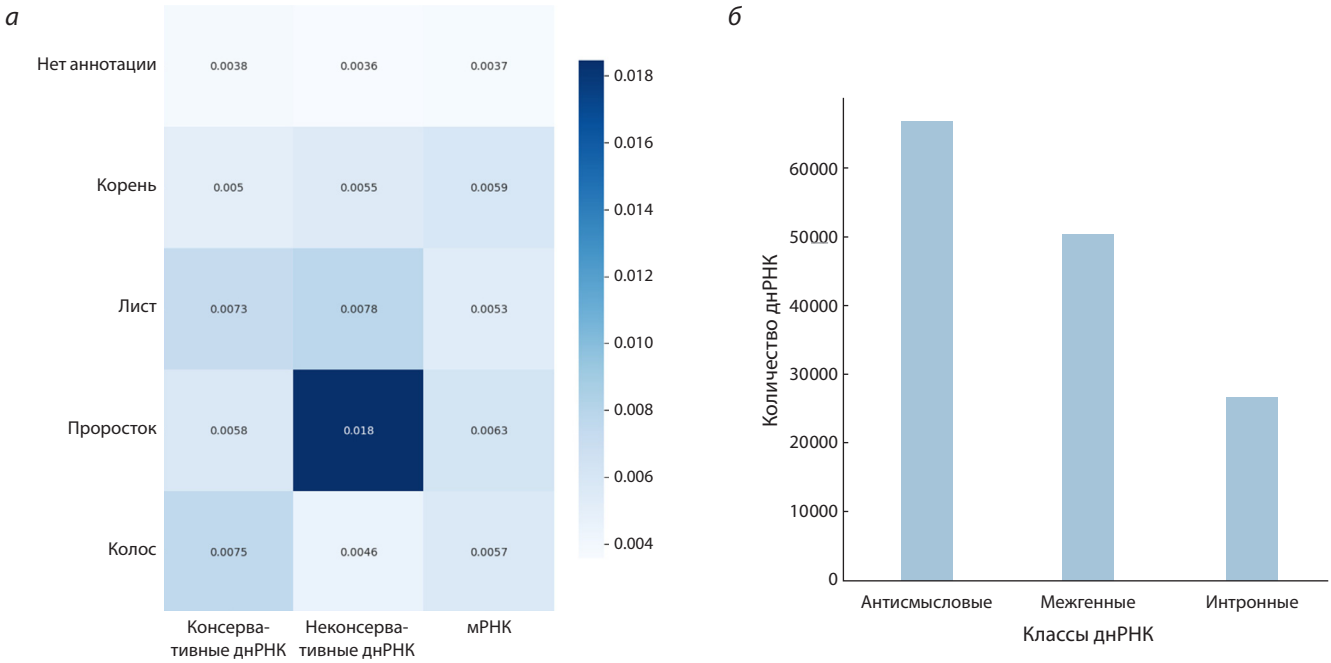

Рис. 5. Специфичность экспрессии днРНК по отношению к различным тканям ячменя, представленная в виде тепловой диаграммы.

По оси *X* приведены данные для двух классов днРНК (консервативные и неконсервативные) и мРНК. Соответствие цвета ячейки и величины специфичности показано шкалой справа от диаграммы. По оси *X* направлены классы днРНК (консервативные и неконсервативные) и мРНК, по оси *Y* – ткани, специфичные данному классу транскриптов (чем выше значение в ячейке, тем большему количеству транскриптов специфична данная ткань) (*а*), и распределение классов днРНК ячменя (*б*).

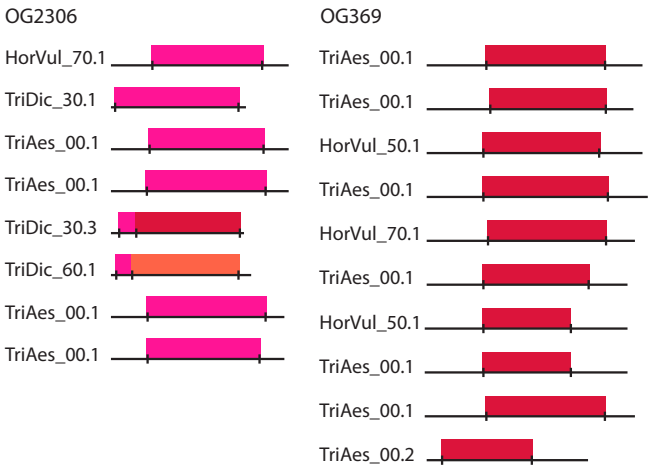

Рис. 6. Доменная структура последовательностей из ортогрупп 2306, 369, слева направо.

Красным цветом отмечен домен ФА2 бета, розовым – ФА2 альфа, оранжевым – ФА2 G12.

В ходе исследования было обнаружено две ортогруппы. Доменная структура (рис. 6) этих групп демонстрирует заметное сходство с доменной архитектурой характерных секретируемых фосфолипаз А2 (Larkin et al., 2019). Длину последовательностей фосфолипаз А2 в ортогруппах можно оценить приблизительно в 150 аминокислот, при этом преобладающая часть последовательностей – домен ФА2, что соответствует известной структуре секретируемых форм ФА2.

Заключение

Разработанный программный комплекс CropGene включает основные блоки программ, необходимых для анализа геномных и транскриптомных данных сельскохозяйственных растений. Эти блоки связаны со сборкой и анализом генома и транскриптома, включают формирование пангенома и пантранскриптома, анализ данных GBS, анализ экспрессии генов, распознавание длинных некодирующих РНК в транскриптомах растений, а также они позволяют производить анализ геномных, транскриптомных данных, особенностей молекулярной эволюции генов сельскохозяйственных растений. Использование указанных модулей позволило решить ряд важных задач по анализу геномных и транскриптомных данных для таких культур, как картофель, пшеница и ячмень.

Список литературы / References

Назипова Н.Н. Разнообразие некодирующих РНК в геномах эукариот. *Математическая биология и биоинформатика*. 2021; 16(2):256-298. doi 10.17537/2021.16.256
[Nazipova N.N. Variety of non-coding RNAs in eukaryotic genomes. *Matematicheskaya Biologiya i Bioinformatika = Mathematical Biology Bioinformatics*. 2021;16(2):256-298. doi 10.17537/2021.16.256 (in Russian)]

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
[Pronozin A.Yu., Bragina M.K., Salina E.A. Crop pangenomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2021;25(1):57-63. DOI 10.18699/VJ21.007]

Сухарева А.С., Кулуев Б.Р. ДНК-маркеры для генетического анализа сортов культурных растений. *Биомика*. 2018;10(1):69-84. doi 10.31301/2221-6197.bmcs.2018-15

[Sukhareva A.S., Kuluev B.R. DNA markers for genetic analysis of crops. *Biomika = Biomics*. 2018;10(1):69-84. doi 10.31301/2221-6197.bmcs.2018-15 (in Russian)]

Хлесткина Е.К. Молекулярные маркеры в генетических исследованиях и в селекции. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):1044-1054
[Khlestkina E.K. Molecular markers in genetic studies and breeding. *Russ J Genet Appl Res.* 2014;4;236-244. doi 10.1134/S2079059714030022]

Afonnikova S.D., Kiseleva A.A., Fedyaeva A.V., Komyshev E.G., Koval V.S., Afonnikov D.A., Salina E.A. Identification of novel loci precisely modulating pre-harvest sprouting resistance and red color components of the seed coat in *T. aestivum* L. *Plants*. 2024;13(10):1309. doi 10.3390/plants13101309

Bray N.L., Pimentel H., Melsted P., Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016;34(5):525-527. doi 10.1038/nbt.3519

Browning B.L., Zhou Y., Browning S.R. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet.* 2018;103(3):338-348. doi 10.1016/j.ajhg.2018.07.015

Burghardt L.T., Young N.D., Tiffin P. A guide to genome-wide association mapping in plants. *Curr Protoc Plant Biol.* 2017;2(1):22-38. doi 10.1002/cppb.20041

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

Bushmanova E., Antipov D., Lapidus A., Prjibelski A.D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience*. 2019;8(9):giz100. doi 10.1093/gigascience/giz100

Cardoso-Silva C.B., Costa E.A., Mancini M.C., Balsalobre T.W.A., Canesin L.E.C., Pinto L.R., Carneiro M.S., Garcia A.A.F., de Souza A.P., Vicentini R. *De novo* assembly and transcriptome analysis of contrasting sugarcane varieties. *PloS One*. 2014;9(2):e88462. doi 10.1371/journal.pone.0088462

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N., Oyama R., … Watahiki A., Okamura-Oho Y., Suzuki H., Kawai J., Hayashizaki Y. The transcriptional landscape of the mammalian genome. *Science*. 2005;309(5740):1559-1563. doi 10.1126/science.1112014

Chen S., Zhou Y., Chen Y., Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi 10.1093/bioinformatics/bty560

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R.; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi 10.1093/bioinformatics/btr330

Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., Li H. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. doi 10.1093/gigascience/giab008

Drewe P., Stegle O., Hartmann L., Kahles A., Bohnert R., Wachter A., Borgwardt K., Rätsch G. Accurate detection of differential RNA processing. *Nucleic Acids Res.* 2013;41(10):5189-5198. doi 10.1093/nar/gkt211

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

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., … Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29(7):644-652. doi 10.1038/nbt.1883

Grosjean P., Ibanez F., Etienne M., Grosjean M.P. Package 'Pastecs'. 2018. Available online: http://masterdistfiles.gentoo.org/pub/cran/web/packages/pastecs/pastecs.pdf

Han S., Liang Y., Ma Q., Xu Y., Zhang Y., Du W., Wang C., Li Y. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief Bioinform.* 2019;20(6):2009-2027. doi 10.1093/bib/bby065

Hassani-Pak K., Singh A., Brandizi M., Hearnshaw J., Parsons J.D., Amberkar S., Phillips A.L., Doonan J.H., Rawlings C. KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol J.* 2021;19(8):1670-1678. doi 10.1111/pbi.13583

Jia L., Liu N., Huang F., Zhou Z., He X., Li H., Wang Z., Yao W. intansv: an R package for integrative analysis of structural variations. *PeerJ*. 2020;8:e8867. doi 10.7717/peerj.8867

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(1):18936. doi 10.1038/srep18936

Johnson K.A., Krishnan A. Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome Biol.* 2022;23(1):1. doi 10.1186/s13059-021-02568-9

Karetnikov D.I., Vasiliev G.V., Toshchakov S.V., Shmakov N.A., Genaev M.A., Nesterov M.A., Ibragimova S.M., Rybakov D.A., Gavrilenko T.A., Salina E.A., Patrushev M.V., Kochetov A.V., Afonnikov D.A. Analysis of genome structure and its variations in potato cultivars grown in Russia. *Int J Mol Sci.* 2023;24(6):5713. doi 10.3390/ijms24065713

Kim E.-D., Sung S. Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.* 2012;17(1):16-21. doi 10.1016/j.tplants.2011.10.008

Kochetov A.V., Afonnikov D.A., Shmakov N., Vasiliev G.V., Antonova O.Y., Shatskaya N.V., Glagoleva A.Y., Ibragimova S.M., Khiutti A., Afanasenko O.S., Gavrilenko T.A. NLR genes related transcript sets in potato cultivars bearing genetic material of wild Mexican Solanum species. *Agronomy*. 2021;11(12):2426. doi 10.3390/agronomy11122426

Larkin D.L., Lozada D.N., Mason R.E. Genomic selection – considerations for successful implementation in wheat breeding programs. *Agronomy*. 2019;9(9):479. doi 10.3390/agronomy9090479

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi 10.1093/bioinformatics/btr509

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv.* 2013;1303.3997

Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi 10.1093/bioinformatics/btp324

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi 10.1093/bioinformatics/btp352

Liao Y., Smyth G.K., Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi 10.1093/bioinformatics/btt656

Lin H.-N., Hsu W.-L. DART: a fast and accurate RNA-seq mapper with a partitioning strategy. *Bioinformatics*. 2018;34(2):190-197. doi 10.1093/bioinformatics/btx558

Muqaddasi Q.H., Brassac J., Ebmeyer E., Kollers S., Korzun V., Argillier O., Stiewe G., Plieske J., Ganal M.W., Röder M.S. Prospects of GWAS and predictive breeding for European winter wheat's grain protein content, grain starch content, and grain hardness. *Sci Rep.* 2020;10(1):12541. doi 10.1038/s41598-020-69381-5

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., … Passos A., Cournapeau D., Brucher M.,

Perrot M., Duchesnay E. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830

Piskol R., Ramaswami G., Li J.B. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet.* 2013;93(4):641-651. doi 10.1016/j.ajhg.2013.08.008

Pronozin A.Yu., Afonnikov D.A. ICAnnoLncRNA: A Snakemake pipeline for a long non-coding-RNA search and annotation in transcriptomic sequences. *Genes*. 2023;14(7):1331. doi 10.3390/genes 14071331

Pronozin A.Yu., Salina E.A., Afonnikov D.A. GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing. *Vavilov J Genet Breed.* 2023;27(7):737-745. doi 10.18699/VJGB-23-86

Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S.D., Mungall K., … Hirst M., Marra M.A., Jones S.J., 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

Scheben A., Batley J., Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol J.* 2017;15(2):149-161. doi 10.1111/pbi.12645

Shendure J. The beginning of the end for microarrays? *Nat Methods*. 2008;5(7):585-587. doi 10.1038/nmeth0708-585

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(19):3210-3212. doi 10.1093/bioinformatics/btv351

Stanke M., Steinkamp R., Waack S., Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 2004;32(Suppl. 2):W309-W312. doi 10.1093/nar/gkh379

Suvakov M., Panda A., Diesh C., Holmes I., Abyzov A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *GigaScience*. 2021;10(11):giab074. doi 10.1093/gigascience/giab074

Tsai M.-C., Manor O., Wan Y., Mosammaparast N., Wang J.K., Lan F., Shi Y., Segal E., Chang H.Y. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*. 2010;329(5992):689-693. doi 10.1126/science.1192002

Velculescu V.E., Zhang L., Zhou W., Vogelstein J., Basrai M.A., Bassett D.E., Hieter P., Vogelstein B., Kinzler K.W. Characterization of the yeast transcriptome. *Cell*. 1997;88(2):243-251. doi 10.1016/S0092-8674(00)81845-0

Vernikos G., Medini D., Riley D.R., Tettelin H. Ten years of pangenome analyses. *Curr Opin Microbiol.* 2015;23:148-154. doi 10.1016/j.mib.2014.11.016

Wang J., Zhang Z. GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinformatics*. 2021;19(4):629-640. doi 10.1016/j.gpb.2021.08.005

Wu T.D., Watanabe C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005;21(9):1859-1875. doi 10.1093/bioinformatics/bti310

Zatybekov A., Abugalieva S., Didorenko S., Gerasimova Y., Sidorik I., Anuarbek S., Turuspekov Y. GWAS of agronomic traits in soybean collection included in breeding pool in Kazakhstan. *BMC Plant Biol.* 2017;17(S1):179. doi 10.1186/s12870-017-1125-0

Zheng X. A tutorial for the R Package SNPRelate. Washington, USA: University of Washington, 2013

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

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 27.11.2024. После доработки 15.01.2025. Принята к публикации 15.01.2025.