Вавиловский журнал генетики и селекции. 2023;27(7):737-745
DOI 10.18699/VJGB-23-86

Перевод на английский язык https://vavilov.elpub.ru/jour

GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования

А.Ю. Пронозин[1, 2] ✉, Е.А. Салина[1, 2, 3], Д.А. Афонников[1, 2, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский государственный аграрный университет, Новосибирск, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ pronozinartem95@gmail.com

Аннотация. Развитие технологий секвенирования нового поколения открыло новые возможности для генотипирования различных организмов, включая растения. Метод генотипирования путем секвенирования (GBS) применяется для идентификации генетической изменчивости и более быстрого генотипирования образцов, а также является более экономически эффективным методом в сравнении с полногеномным секвенированием. GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Этот метод был применен для генетического картирования, выявления молекулярных маркеров, геномной селекции, в исследовании генетического разнообразия, идентификации сортов, а также в исследованиях в области биологии охраны природы и эволюционной экологии. Однако сокращение времени и стоимости секвенирования привело к необходимости разработки качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных. Для этих целей были разработаны биоинформатические конвейеры анализа данных, полученных методом GBS. Вследствие схожести этапов обработки существующие конвейеры в основном различаются комбинацией программных пакетов, специфически подобранных для обработки данных как для определенных, так и для любых организмов. Несмотря на качественно подобранные пакеты программ, конвейеры имеют некоторые недостатки, например отсутствие возможности автоматизации процесса расчета (каждый этап нужно запускать вручную), что значительно снижает скорость исследования. В большинстве конвейеров отсутствует возможность автоматической установки всех необходимых программных пакетов, а также нет возможности отключения ненужного или пройденного этапа. В настоящей работе нами был разработан биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер применим для любых видов организмов. Реализация конвейера на платформе Snakemake позволила полностью автоматизировать процесс расчета и установки необходимых программных пакетов. Конвейер позволяет обрабатывать большие объемы данных (более 400 образцов).
Ключевые слова: генотипирование путем секвенирования; биоинформатический конвейер; ячмень.

Для цитирования: Пронозин А.Ю., Салина Е.А., Афонников Д.А. GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования. *Вавиловский журнал генетики и селекции*. 2023;27(7):737-745. DOI 10.18699/VJGB-23-86

GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing

A.Y. Pronozin[1, 2] ✉, E.A. Salina[1, 2, 3], D.A. Afonnikov[1, 2, 4]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State Agrarian University, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
✉ pronozinartem95@gmail.com

Abstract. The development of next-generation sequencing technologies has provided new opportunities for genotyping various organisms, including plants. Genotyping by sequencing (GBS) is used to identify genetic variability more rapidly, and is more cost-effective than whole-genome sequencing. GBS has demonstrated its reliability and flexibility for a number of plant species and populations. It has been applied to genetic mapping, molecular marker discovery, genomic selection, genetic diversity studies, variety identification, conservation biology and evolutionary studies. However, reduction in sequencing time and cost has led to the need to develop efficient bioinformatics analyses for an ever-expanding amount of sequenced data. Bioinformatics pipelines for GBS data analysis serve the purpose. Due to the similarity of data processing steps, existing pipelines are mainly characterised by a combination of software

packages specifically selected either to process data for certain organisms or to process data from any organisms. However, despite the usage of efficient software packages, these pipelines have some disadvantages. For example, there is a lack of process automation (in some pipelines, each step must be started manually), which significantly reduces the performance of the analysis. In the majority of pipelines, there is no possibility of automatic installation of all necessary software packages; for most of them, it is also impossible to switch off unnecessary or completed steps. In the present work, we have developed a GBS-DP bioinformatics pipeline for GBS data analysis. The pipeline can be applied for various species. The pipeline is implemented using the Snakemake workflow engine. This implementation allows fully automating the process of calculation and installation of the necessary software packages. Our pipeline is able to perform analysis of large datasets (more than 400 samples).
Key words: genotyping by sequencing (GBS); bioinformatic pipeline; hordeum.

For citation: Pronozin A.Y., Salina E.A., Afonnikov D.A. GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023; 27(7):737-745. DOI 10.18699/VJGB-23-86

Введение

Генетическое разнообразие является важнейшей основой для изучения устойчивости растений к биотическим и абиотическим стрессам и создания новых высокоадаптивных и урожайных сортов сельскохозяйственных культур. Изучение генетического разнообразия осуществляется с использованием различных методов генетического анализа. На сегодняшний день один из наиболее перспективных методов – применение молекулярных маркеров (ДНК-маркеров) (Канукова и др., 2019). Это генетические маркеры, анализируемые на уровне ДНК (Хлесткина, 2013). С их помощью можно выявлять генетическое разнообразие популяций, подвидов, видов, эффективно определять хозяйственно ценные признаки еще на начальном этапе селекции на уровне ДНК (Сухарева, Кулуев, 2018).

Для генетического анализа особенно удобны SNP-маркеры (Хлесткина, 2013). SNP (single-nucleotide polymorphism – однонуклеотидный полиморфизм) – это однонуклеотидная позиция в геномной ДНК, для которой в популяции встречаются различные вариации последовательности (аллелей) (Сухарева, Кулуев, 2018). SNP широко используют для изучения аллельного полиморфизма, тестирования чистоты семян, анализа гаплотипа и родословных, а также для генотипирования и построения генетических карт.

Получить информацию об SNP-маркерах в настоящее время можно для любого растения в масштабе полного генома благодаря технологиям высокопроизводительного секвенирования нового поколения. Идентификация SNP возможна с помощью стратегий полногеномного секвенирования (WGS) и генотипирования путем секвенирования (GBS) (Scheben et al., 2017). Цель полногеномного секвенирования – получить короткие фрагменты последовательности полного генома и на этой основе путем их выравнивания на референсный геном или полногеномной сборки оценить вариации ДНК. Это сложная и дорогостоящая задача, цена за один геном превышает 2000$ и зависит от размера и сложности генома, желаемого уровня полноты и вычислительных ресурсов (Narum et al., 2013). Например, секвенирование полного генома ячменя до уровня хромосом обходится примерно в 60,000$ (Monat et al., 2019). Выделяют также методы полногеномного секвенирования с более низкой глубиной прочтения, стоимость которых в разы меньше: 100–400$ за геном. Однако, как утверждают авторы (Bimber et al., 2016), при этом снижается точность получаемых данных о генотипах.

Метод генотипирования путем секвенирования более быстрый и экономически эффективный, чем WGS. Например, стоимость секвенирования фрагментов генома ячменя в эксперименте GBS не превышает 30$ (Monat et al., 2019). В методе GBS выделяют два подхода секвенирования. В первом для фрагментации образцов ДНК используются ферменты рестрикции, специфичные для конкретных сайтов, после чего производится секвенирование полученных фрагментов (Glaubitz et al., 2014). Во втором к обоим концам фрагментов ДНК лигируются уникальные последовательности адаптеров, один из которых содержит уникальную последовательность «штрихкод», после чего производится секвенирование данных маркированных фрагментов ДНК (Elshire et al., 2011). Поскольку при секвенировании фрагменты ДНК прочитываются только вблизи сайтов рестрикции, в методе GBS не происходит прочтения полногеномной последовательности ДНК. За счет этого процесс секвенирования существенно удешевляется, однако количество SNP, которые можно идентифицировать, оказывается меньше, чем при полногеномном секвенировании. Тем не менее данных, полученных при помощи протокола GBS, оказывается вполне достаточно, чтобы с приемлемой точностью характеризовать генетическое разнообразие популяций сельскохозяйственных растений.

Метод GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Он был применен для выявления молекулярных маркеров для генетического картирования и геномной селекции (Poland et al., 2012), в исследовании генетического разнообразия (Lu et al., 2013; Peterson et al., 2014), идентификации сортов (Wang et al., 2020; Rajendran et al., 2022), а также исследованиях в области биологии охраны природы и эволюционной экологии (Narum et al., 2013). GBS существенно сокращает как стоимость, так и время, необходимое для секвенирования исследуемых образцов. Это потребовало разработки качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных. В результате были разработаны биоинформатические конвейеры анализа данных, полученных методом GBS. Существующие конвейеры имеют схожую схему анализа данных, которая включает: проверку качества сырых прочтений, демультиплексирование, картирование на референсный геном и поиск полиморфизмов.

Этап картирования на геном делится на два типа: на основе референсного генома (Glaubitz et al., 2014; Tor-

kamaneh et al., 2017; Wickland et al., 2017) и на основе «имитации» референсного генома (Mock Reference) (Melo et al., 2016). В первом случае после контроля качества сырых прочтений последовательности картируются на референсный геном с целью выявления полиморфизмов (Torkamaneh et al., 2017). Однако если референсный геном отсутствует или имеет низкое качество сборки, применяют метод «имитации» референсного генома. Этот метод производит кластеризацию исследуемых прочтений для выявления консенсусных последовательностей (центроидов), на основе которых осуществляется сборка генома (Melo et al., 2016). Вследствие схожести этапов обработки данных, существующие конвейеры в основном различаются комбинацией программ. Подобные комбинации программ должны учитывать различные геномные характеристики, такие как количество выявленных полиморфизмов, сложность генома, степень гетерозиготности, доля повторяющихся последовательностей во всем геноме. Также более современные конвейеры позволяют подбирать параметры для исследуемых организмов (Torkamaneh et al., 2017; Wickland et al., 2017), тогда как более ранние конвейеры имеют некоторые ограничения. Например, в программе TASSEL надо указывать ограничение длины последовательностей, что приводит к потере значительного количества коротких сырых прочтений (Glaubitz et al., 2014; Melo et al., 2016). Из-за постоянного роста количества секвенированных библиотек конвейеры должны предоставлять возможность обработки большого объема данных за один запуск. Важным аспектом конвейеров является также автоматизация процесса обработки и простота установки программы.

В настоящей работе мы разработали биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер включает схему обработки GBS данных, предложенную в работе (Jayakodi et al., 2020), и применим для любых видов организмов. Конвейер позволяет обрабатывать большие объемы данных (более 400 образцов) и реализован с помощью программного менеджера Snakemake (Köster, Rahmann, 2012).

Материалы и методы

Биоинформатический конвейер GBS-DP анализа данных, полученных методом GBS, представлен на рис. 1.

На вход конвейера подается путь к набору библиотек прочтений и путь к референсному геному. Библиотеки прочтений должны быть в формате FASTQ, референсный геном – в формате FASTA. В случае если библиотеки имеют баркодирование, необходимо предварительно их демультиплицировать.

Конвейер состоит из трех основных этапов: предобработка данных, поиск полиморфизмов, анализ генетического разнообразия. Предобработка данных включает проверку качества сырых прочтений, удаление адаптеров и построение индекса референсного генома. Поиск полиморфизмов состоит из картирования предобработанных прочтений на референсный геном, сортировки картированных прочтений и поиска однонуклеотидных полиморфизмов. 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. Более детальное описание каждого этапа приведено ниже.

Предобработка данных. На этом этапе производится контроль качества, удаление адаптеров сырых прочтений и построение индекса референсного генома. Контроль качества и удаление адаптеров производятся программой cutadapt (Martin, 2011). Для прочтений каждой библиотеки удаляются адаптеры, список которых пользователь должен внести в файл конфигураций.

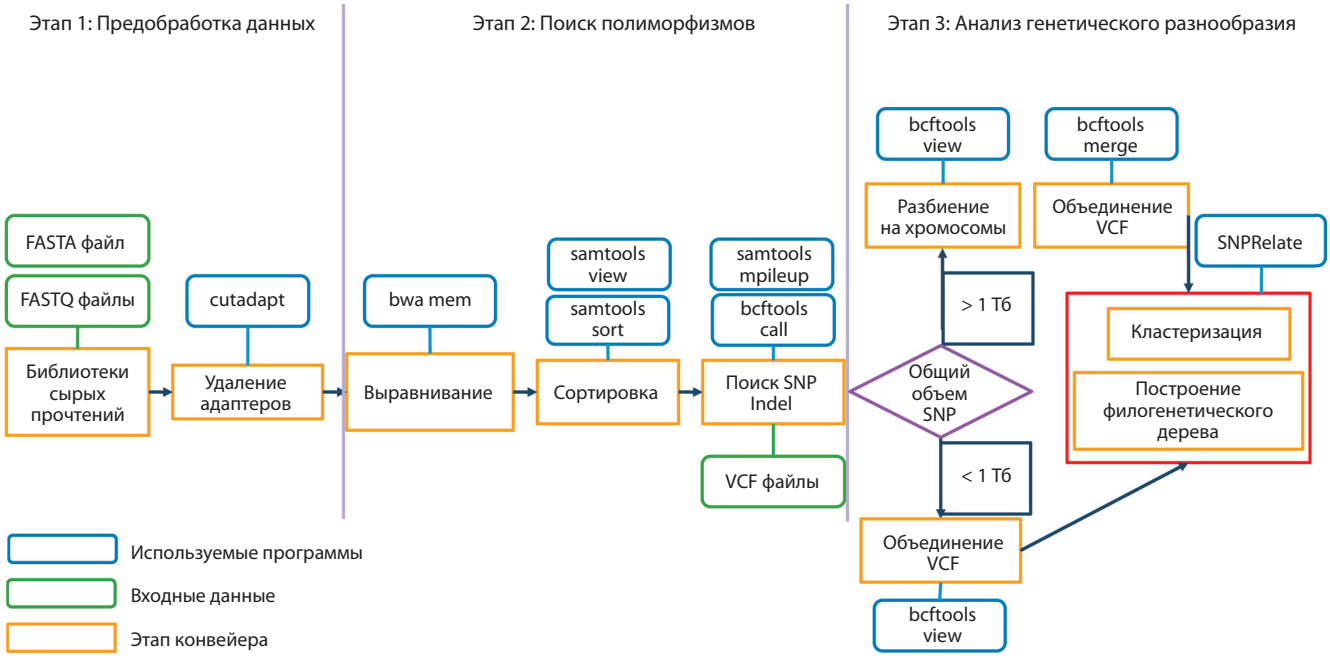

Рис. 1. Блок-схема биоинформатического конвейера GBS-DP обработки данных GBS.

На этом этапе конвейер выполняет построение индекса референсного генома с помощью программы bwa index (Li H., 2013).

Поиск полиморфизмов состоит из картирования предобработанных прочтений на референсный геном, сортировки картированных прочтений и поиска однонуклеотидных полиморфизмов.

Картирование предобработанных прочтений производится программой bwa mem (Li H., 2013) с параметрами "–k 19 –w 100".

Результаты картирования, полученные в формате SAM, переводятся в формат BAM и сортируются комбинацией программ samtools view и samtools sort соответственно (Danecek et al., 2021). В отсортированных файлах производится поиск полиморфизмов (SNP, вставок и делеций (индел)) с помощью комбинации программ samtools mpileup и bcftools call (Danecek et al., 2021). Ранее было показано на примере генома пшеницы (Yao et al., 2020), что комбинация программ "Samtools/mpileup + BWA-mem", которая использована в нашем конвейере, превосходит другие комбинации программ картирования и идентификации полиморфизмов.

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 (если суммарный размер полученных файлов VCF превышает 1 Тб). Вариант обработки для данных с общим объемом меньше 1 Тб включает три этапа:

1) результаты поиска полиморфизмов в формате VCF для каждой библиотеки индексируются с помощью программы bcftools index (Danecek et al., 2021);
2) проиндексированные файлы объединяются в общий файл формата VCF в программе bcftools merge (Danecek et al., 2021). Этот файл содержит данные о полиморфизмах всех исследуемых образцов для всех хромосом;
3) полученный общий файл формата VCF конвертируется в формат GDS (Genomic Data Structure) с помощью пакета R – SeqArray (Zheng et al., 2017). Данный формат позволяет значительно сократить объем оперативной памяти, затрачиваемой на обработку результатов поиска полиморфизмов, за счет перевода табличного формата в бинарный.

Вариант обработки для данных с общим занимаемым объемом больше 1 Тб включает четыре этапа:

1) результаты поиска полиморфизмов в формате VCF для каждой библиотеки разбиваются на хромосомы с помощью программы bcftools view (Danecek et al., 2021);
2) полученные файлы с полиморфизмами для каждой хромосомы индексируются с использованием программы bcftools index (Danecek et al., 2021);
3) далее файлы с полиморфизмами объединяются для каждой хромосомы. В результате получаются файлы, содержащие информацию о полиморфизмах во всех библиотеках для отдельной хромосомы;
4) файлы для отдельных хромосом в формате VCF конвертируются в формат GDS. После этого полученные файлы формата GDS для каждой хромосомы объединяются в общий файл с помощью функции snpgdsCombineGeno пакета SNPRelate (Zheng et al., 2017).

Схема построения филогенетического дерева и кластеризации для обоих вариантов идентичная. Следует отметить, что при оценке функционального значения SNP важно также учитывать функциональное значение полиморфных локусов, находящихся с ним в неравновесии по сцеплению (LD) (Пономаренко, 2018). Два аллеля различных локусов находятся в неравновесии по сцеплению, когда частота состоящего из них гаплотипа значимо отличается от частоты, ожидаемой при случайной сегрегации (Gabriel et al., 2002). Величина LD зависит от ряда факторов: величины и скорости дрейфа генов, генетических примесей в популяции, мутаций и рекомбинаций, размера популяции (Аульченко, Аксенович, 2006). Обычно LD оценивается коэффициентом неравновесности сцепления (D), однако эта мера не всегда удобна, поскольку диапазон его возможных значений зависит от частот аллелей, к которым он относится. Это затрудняет сравнение уровня неравновесия по сцеплению между разными парами аллелей. Таким образом, производится нормировка коэффициента D на основе коэффициента корреляции Пирсона r^2, который варьирует от 0 до 1. Чем ближе значение r^2 к 0, тем больше вероятность, что выявленные SNP случайны.

Для полученного общего файла, содержащего информацию о полиморфизмах для всех библиотек по всем хромосомам в формате GDS, анализируется параметр LD. Расчет производится с помощью пакета R – SNPRelate (Zheng et al., 2017), функция snpgdsLDpruning.

Для анализа главных компонент, отфильтрованных SNP, применяется пакет R – SNPRelate, для построения филогенетического дерева – тоже пакет SNPRelate, но с использованием метода иерархической кластеризации.

Системные требования и установка. Конвейер GBS-DP реализован с применением программного менеджера Snakemake v6.0.0 (Köster, Rahmann, 2012), инструмента для создания конвейеров анализа данных, реализованного на языке Python. Созданные в этой среде конвейеры можно легко масштабировать для серверных, кластерных, сетевых и облачных сред без необходимости изменять определение рабочего процесса. Snakemake совместим с системой Conda, что позволяет без труда устанавливать новые программы, необходимые для конвейера. Конвейер разработан для операционной системы Linux. Для запуска требуется минимум 10 Гб оперативной памяти (чем больше данных, тем больше требуется оперативной памяти). Для запуска конвейера в файле конфигураций, необходимо указать путь к сырым прочтениям и путь к референсному геному, после чего можно запускать программу. Код и пошаговая инструкция запуска конвейера доступны по адресу: https://github.com/artempronozin95/GBS-DP-bioinformatics-pipeline-for-genotyping-by-sequencing-data-processing/tree/main.

Данные для тестового анализа. Для тестового применения конвейера GBS-DP в настоящей работе был использован проект PRJEB39633 из базы данных European Nucleotide Archive (ENA) (Leinonen et al., 2011), который

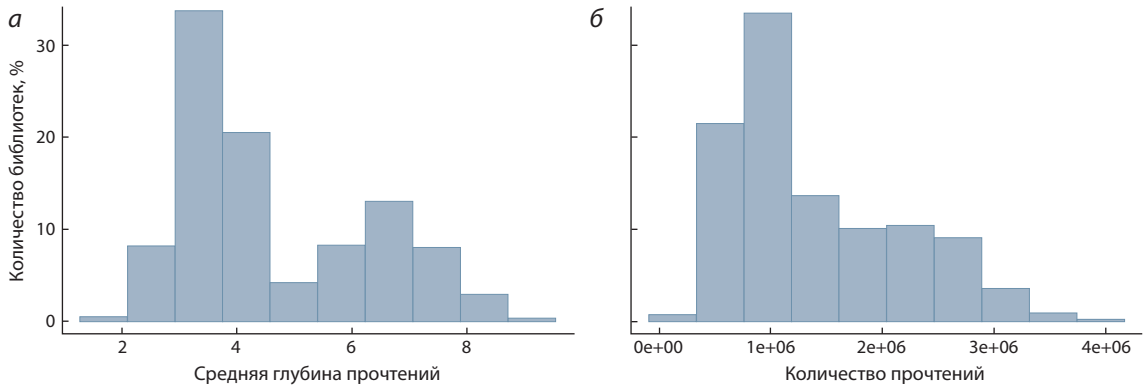

Рис. 2. Распределение средней глубины прочтений (*а*) и количества прочтений (*б*).

содержит библиотеки GBS для популяции ячменя, полученной в результате скрещивания шестирядного ячменя сорта Morex и мутантной линии *luteostrians*-P1 (*lst/LST*) (Li M. et al., 2021). Библиотеки были получены с помощью комбинации ферментов рестрикции *MspI* и *PstI* (Wendler et al., 2015). Всего проект PRJEB39633 содержит 679 библиотек для 272 генотипов; на один генотип в среднем приходится три библиотеки, поэтому перед проведением анализа прочтения библиотеки для одного генотипа объединялись.

Мы использовали референсный геном ячменя 51-й версии (IBSC_v2), загруженный с базы данных Ensembl plants (Bolser et al., 2016).

Результаты

Время, затраченное для обработки данных на различных этапах выполнения конвейера GBS-DP для разного количества библиотек ячменя (10, 50, 100, 150, 200 и 272 шт.), приведено в электронном Приложении[1]. Характеристики вычислительного узла: процессор AMD EPYC 74521, 32 ядра, объем памяти 1 Тб. Для анализа мы использовали 100 Гб оперативной памяти и 20 ядер процессора. Наибольшее время было затрачено на формирование общего файла, содержащего полиморфизмы. Однако можно заметить, что на формирование общего файла для 200 библиотек было затрачено меньше времени, чем для 150 библиотек; это связано с включением режима обработки больших данных, который ускоряет процесс расчета.

Конвейер предоставляет результаты оценки базовых характеристик секвенированных библиотек. Длина прочтения для каждой библиотеки равна 107 нк. Средняя глубина прочтения варьирует в пределах 2–8, что является допустимым значением для метода GBS (рис. 2, *а*). Более 30 % библиотек содержат свыше 1 000 000 прочтений (см. рис. 2, *б*). В среднем для одной библиотеки покрытие референсного генома ячменя (4 225 577 519 нк.) фрагментами ДНК составляет 3 % от общей длины.

Также конвейер предоставляет результаты поиска полиморфизмов между исследуемыми генотипами. Для 272 исследуемых образцов выявлено 447 409 SNP. Общее количество индел 46 557. Медиана значения транзиции/трансверсию = 1.75, что указывает на преобладание транзиций. Параметр LD (r^2) был выбран равным 0.5. После применения фильтра LD осталось 45 402 полиморфных и независимых SNP.

Распределение обнаруженных SNP по хромосомам показало, что больше SNP выявлено для хромосом 3, 6 и 7 (рис. 3). Основными результатами конвейера являются анализ главных компонент генотипов на основе выявленных SNP (рис. 4) и построение филогенетического дерева. Результаты анализа главных компонент на основе 45 402 SNP показывают, что внутри исследованной популяции на диаграмме рассеяния в пространстве двух первых компонент четко выделяется несколько кластеров (см. рис. 4). Однако суммарная доля дисперсии, приходящаяся на две эти компоненты, невелика (20 %), что может свидетельствовать об общем высоком уровне генетического разнообразия в полученной популяции растений.

Филогенетическое дерево, построенное иерархическим методом кластеризации, так же как и при кластеризации методом главных компонент, приведено на рис. 5. На дереве выделяются три больших кластера, что согласуется с данными, представленными на рис. 4.

Обсуждение

Благодаря снижению стоимости и сокращению времени, необходимого для секвенирования методом GBS, появилось множество экспериментов, проведенных этим методом. Например, база данных генетических профилей ячменя IPK Gatersleben (Milner et al., 2019) содержит 22 626 образцов, полученных методом GBS. Такое количество образцов требует быстрого и качественного способа обработки данных. На сегодняшний день уже существуют конвейеры, позволяющие обрабатывать результаты GBS. Однако, несмотря на качественно подобранные пакеты программ и возможность подстраивать параметры под исследуемые организмы, данные конвейеры имеют некоторые недостатки. Так, для GBS-SNP-CROP и TASSEL нельзя автоматизировать процесс расчета (каждый этап нужно запускать вручную), что значительно снижает скорость исследования. GB-eaSy не позволяет одновременно исследовать сразу несколько библиотек сырых прочтений. Во всех существующих конвейерах нет возможности отключения ненужного или пройденного этапа. Например, если нет возможности предоставить данные по бар-кодам для исследуемых библиотек, то ни один из перечислен-

[1] Приложение см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx23.pdf

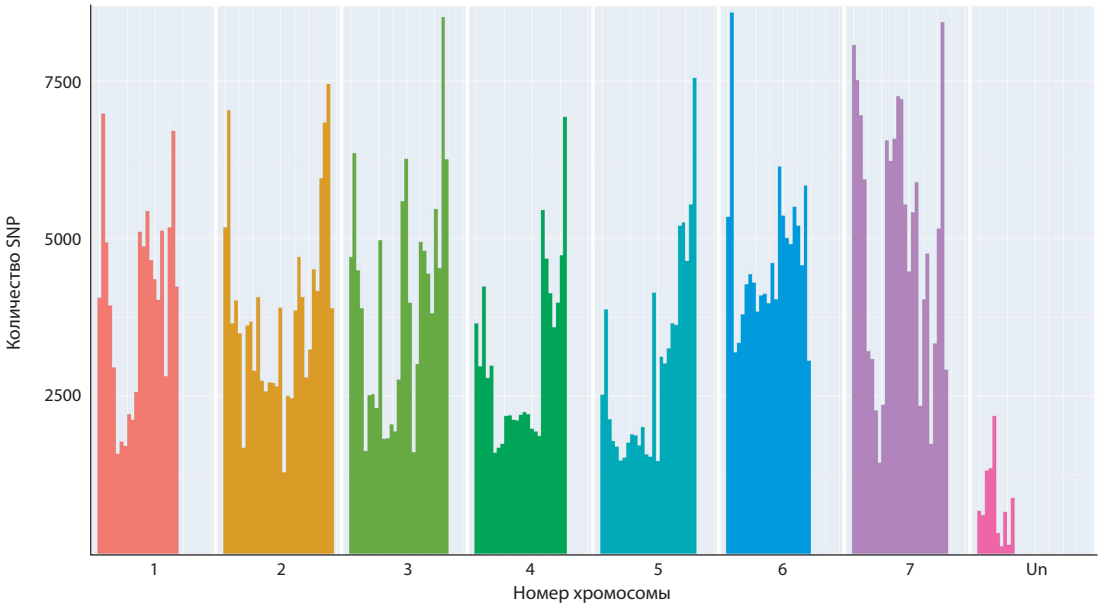

Рис. 3. Распределение выявленных SNP по хромосомам.

Ось *X* – координаты SNP на хромосомах, ось *Y* – количество SNP, соответствующих данным координатам. Шаг 10^8 нк.

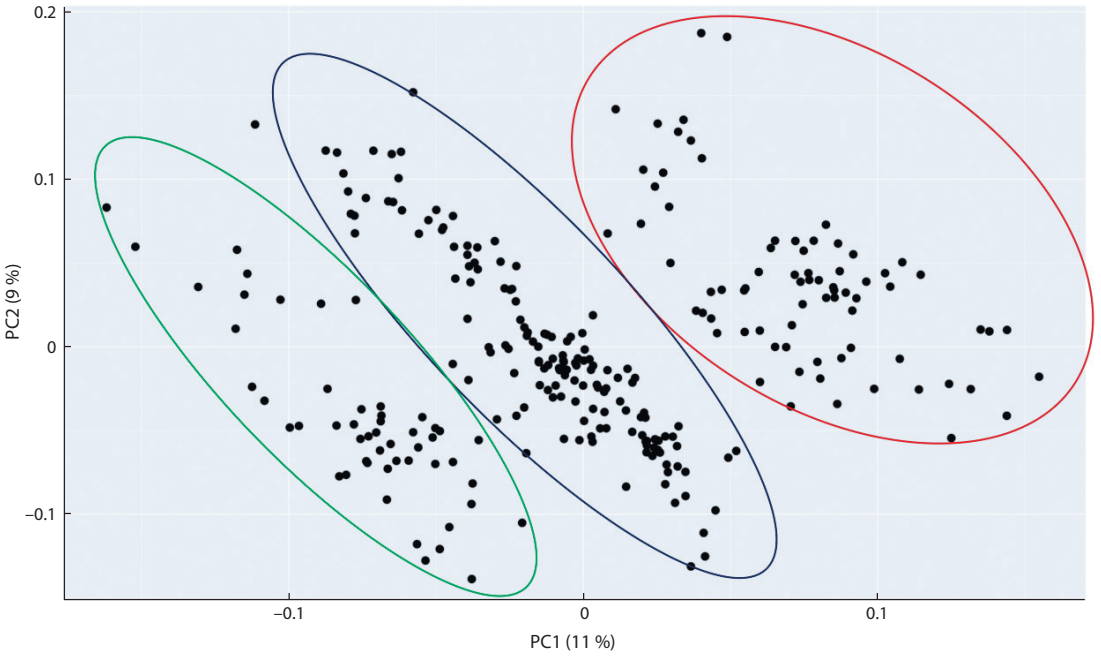

Рис. 4. Диаграмма рассеяния генотипов для популяции ячменя, полученной в результате скрещивания сорта Morex и мутантной линии *luteostrians*-P1 (*lst/LST*) для двух главных компонент, полученных при анализе генетического разнообразия конвейером GBS-DP.

В скобках рядом с названиями компонент указана доля от общей дисперсии.

ных конвейеров работать не будет. Также в большинстве конвейеров отсутствует возможность автоматической установки всех необходимых программных пакетов.

Разработанный нами конвейер основан на методе, предложенном M. Jayakodi с коллегами (Jayakodi et al., 2020). Авторы подобрали программы таким образом, чтобы предоставлять наиболее точный результат по поиску полиморфизмов. Однако этот метод хорошо применим для малых данных – до 50 библиотек. С увеличением количества библиотек увеличивается нагрузка на оперативную память и на занимаемый объем на жестком диске, что приводит к нежелательным ошибкам и прерыванию процесса расчета. Нами был предложен подход для расчета больших данных. Результаты применения данного подхода представлены в электронном Приложении и на рис. 6.

Как видно из рис. 6, предложенный нами подход значительно ускоряет процесс расчета для больших данных, однако для малых данных разница в скорости расчета не-

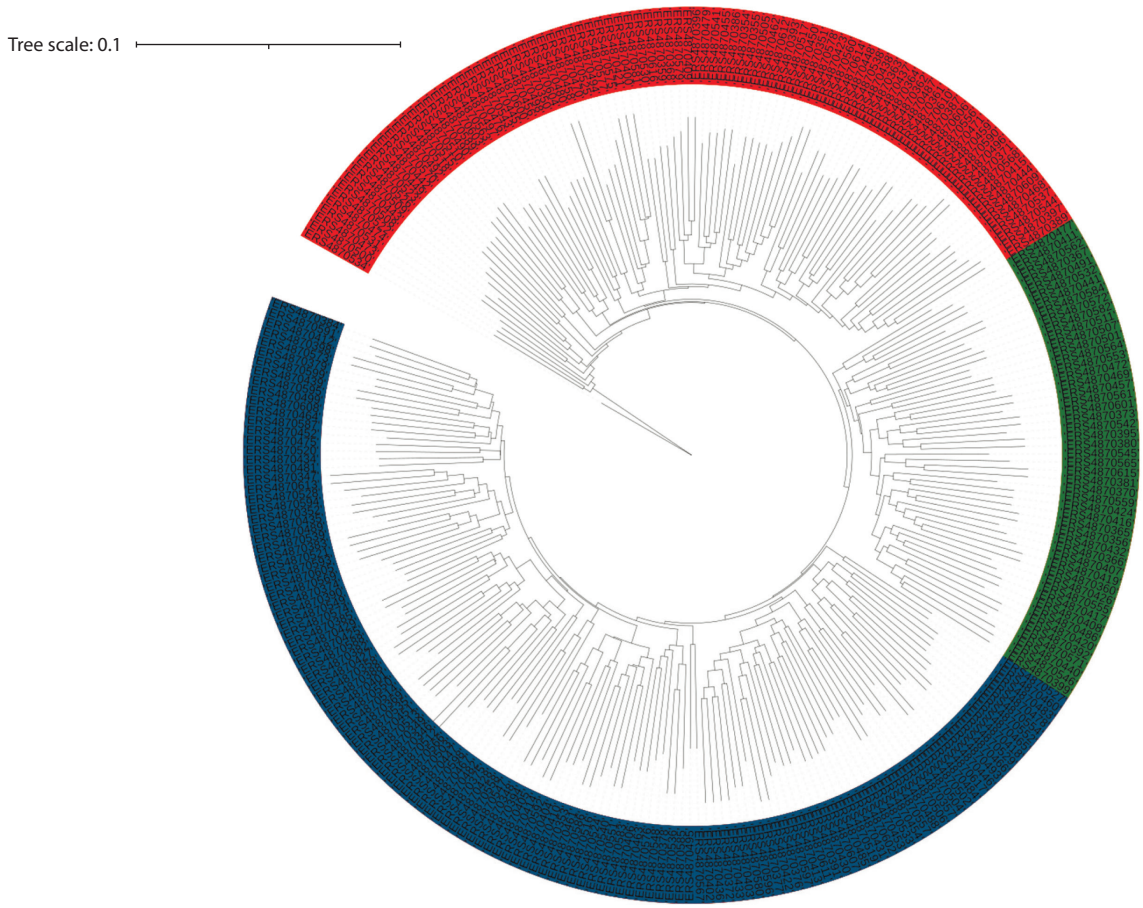

Рис. 5. Филогенетическое дерево 272 библиотек ячменя, построенное методом иерархической кластеризации.

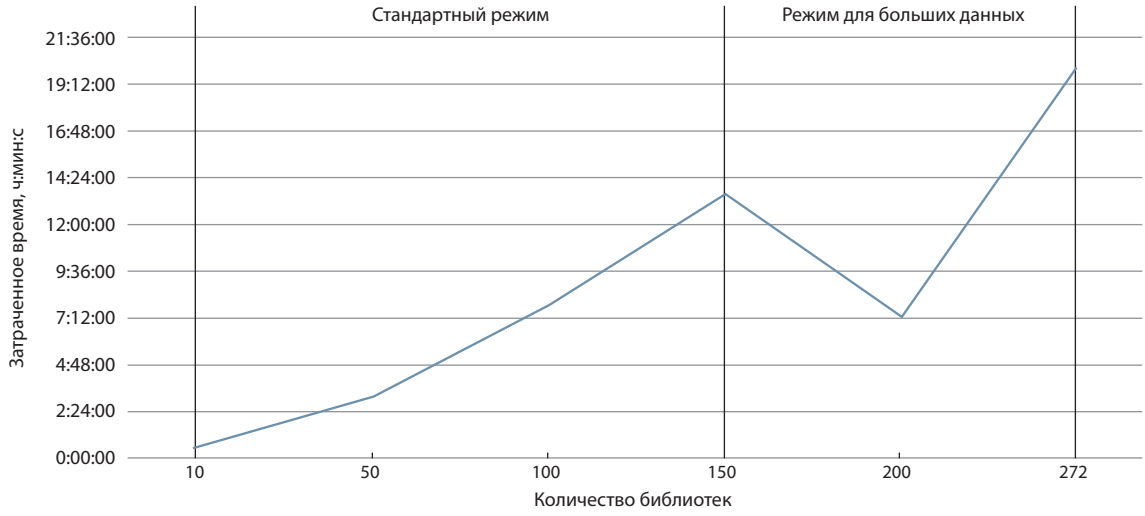

Рис. 6. Зависимость времени, затраченного на работу конвейера, от количества исследуемых библиотек.

велика. Поэтому режим активируется только на данных, общий объем найденных полиморфизмов которых превышает 500 Гб.

Разработанный нами конвейер использует программный менеджер Snakemake. Данный метод реализации автоматически учитывает выполненные задачи для каждого образца, что позволяет исключить дублирование задач, а также дает возможность возобновить процесс расчета с момента его остановки (например, вследствие ошибки). Модульная структура дает более удобный функционал манипуляции этапами конвейера (удаление, добавление, перемещение, отключение). Также Snakemake имеет возможность автоматической установки всех необходимых программ для работы конвейера.

Заключение

Методы генотипирования путем секвенирования продемонстрировали свою надежность и гибкость для ряда видов и популяций растений. Они сократили как стоимость, так и время, необходимое для секвенирования исследуемых образцов, что позволило проводить секвенирование в еще большем объеме. В настоящей работе нами был предложен биоинформатический конвейер GBS-DP, который позволяет обрабатывать данные широкомасштабного секвенирования, проведенного методом GBS. Результаты демонстрируют достаточно высокую скорость работы конвейера как для больших данных (более 400 библиотек), так и для малых (30 библиотек). Конвейер предоставляет также анализ выявленных полиморфизмов.

Список литературы / References

Аульченко Ю.С., Аксенович Т.И. Методологические подходы и стратегии картирования генов, контролирующих комплексные признаки человека. *Информ. вестн. ВОГиС*. 2006;10(1):189-202
[Aulchenko Yu.S., Aksenovich T.I. Methodological approaches and strategies for mapping genes controlling complex human traits. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2006;10(1):189-202 (in Russian)]

Канукова К.Р., Газаев И.Х., Сабанчиева Л.К., Боготова З.И., Аппаев С.П. ДНК-маркеры в растениеводстве. *Изв. Кабардино-Балкарского науч. центра РАН*. 2019;6(92):220-232. DOI 10.35330/1991-6639-2019-6-92-220-232
[Kanukova K.R., Gazaev I.Kh., Sabanchieva L.K., Bogotova Z.I., Appaev S.P. DNA markers in crop production. *Izvestiya Kabardino-Balkarskogo Nauchnogo Tsentra RAN = News of the Kabardin-Balkar Scientific Center of RAS*. 2019;6(92):220-232. DOI 10.35330/1991-6639-2019-6-92-220-232 (in Russian)]

Пономаренко И.В. Отбор полиморфных локусов для анализа ассоциаций при генетико-эпидемиологических исследованиях. *Науч. результаты биомед. исследований*. 2018;4(2):40-54. DOI 10.18413/2313-8955-2018-4-2-0-5
[Ponomarenko I.V. Selection of polymorphic loci for association analysis in genetic-epidemiological studies. *Nauchnye Rezultaty Biomeditsynskikh Issledovaniy = Research Results in Biomedicine*. 2018;4(2):40-54. DOI 10.18413/2313-8955-2018-4-2-0-5 (in Russian)]

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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):1044-1054 (in Russian)]

Bimber B.N., Raboin M.J., Letaw J., Nevonen K.A., Spindel J.E., McCouch S.R., Cervera-Juanes R., Spindel E., Carbone L., Ferguson B., Vinson A. Whole-genome characterization in pedigreed non-human primates using genotyping-by-sequencing (GBS) and imputation. *BMC Genomics*. 2016;17(1):676. DOI 10.1186/s12864-016-2966-x

Bolser D., Staines D.M., Pritchard E., Kersey P. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards D. (Ed.) Plant Bioinformatics. Methods in Molecular Biology. Vol. 1374. New York: Humana Press, 2016;115-140. DOI 10.1007/978-1-4939-3167-5_6

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

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6(5):e19379. DOI 10.1371/journal.pone.0019379

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., Altshuler D. The structure of haplotype blocks in the human genome. *Science*. 2002;296(5576):2225-2229. DOI 10.1126/science.1069424

Glaubitz J.C., Casstevens T.M., Lu F., Harriman J., Elshire R.J., Sun Q., Buckler E.S. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*. 2014;9(2):e90346. DOI 10.1371/journal.pone.0090346

Jayakodi M., Padmarasu S., Haberer G., Bonthala V.S., Gundlach H., Monat C., Lux T., Kamal N., Lang D., Himmelbach A., Ens J., Zhang X.Q., Angessa T.T., Zhou G., Tan C., Hill C., Wang P., Schreiber M., Boston L.B., Plott C., Jenkins J., Guo Y., Fiebig A., Budak H., Xu D., Zhang J., Wang C., Grimwood J., Schmutz J., Guo G., Zhang G., Mochida K., Hirayama T., Sato K., Chalmers K.J., Langridge P., Waugh R., Pozniak C.J., Scholz U., Mayer K.F.X., Spannagl M., Li C., Mascher M., Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. 2020;588(7837):284-289. DOI 10.1038/s41586-020-2947-8

Köster J., Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. DOI 10.1093/bioinformatics/bts480

Leinonen R., Akhtar R., Birney E., Bower L., Cerdeno-Tárraga A., Cheng Y., Cleland I., Faruque N., Goodgame N., Gibson R., Hoad G., Jang M., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Sobhany S., Ten Hoopen P., Vaughan R., Zalunin V., Cochrane G. The European nucleotide archive. *Nucleic Acids Res.* 2011;39(Database issue):D28-D31. DOI 10.1093/nar/gkq967

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. 2013. DOI 10.48550/arXiv.1303.3997

Li M., Guo G., Pidon H., Melzer M., Prina A.R., Börner T., Stein N. ATP-dependent *Clp* protease subunit *C1*, *HvClpC1*, is a strong candidate gene for barley variegation mutant *luteostrians* as revealed by genetic mapping and genomic re-sequencing. *Front. Plant Sci.* 2021;12:664085. DOI 10.3389/fpls.2021.664085

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 2013;9(1):e1003215. DOI 10.1371/journal.pgen.1003215

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

Melo A.T., Bartaula R., Hale I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*. 2016;17(1):29. DOI 10.1186/s12859-016-0879-y

Milner S.G., Jost M., Taketa S., Mazón E.R., Himmelbach A., Oppermann M., Weise S., Knüpffer H., Basterrechea M., König P., Schüler D., Sharma R., Pasam R.K., Rutten T., Guo G., Xu D., Zhang J., Herren G., Müller T., Krattinger S.G., Keller B., Jiang Y., González M.Y., Zhao Y., Habekuß A., Färber S., Ordon F., Lange M., Börner A., Graner A., Reif J.C., Scholz U., Mascher M., Stein N. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 2019;51(2):319-326. DOI 10.1038/s41588-018-0266-x

Monat C., Schreiber M., Stein N., Mascher M. Prospects of pan-genomics in barley. *Theor. Appl. Genet.* 2019;132(3):785-796. DOI 10.1007/s00122-018-3234-z

Narum S.R., Buerkle C.A., Davey J.W., Miller M.R., Hohenlohe P.A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* 2013;22(11):2841-2847. DOI 10.1111/mec.12350

Peterson G.W., Dong Y., Horbach C., Fu Y.-B. Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. *Diversity*. 2014;6(4):665-680. DOI 10.3390/d6040665

Poland J., Endelman J., Dawson J., Rutkoski J., Wu S., Manes Y., Dreisigacker S., Crossa J., Sánchez-Villeda H., Sorrells M., Jannink J.-L. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*. 2012;5(3):103-113. DOI 10.3835/plantgenome2012.06.0006

Rajendran N.R., Qureshi N., Pourkheirandish M. Genotyping by sequencing advancements in barley. *Front. Plant Sci.* 2022;13:931423. DOI 10.3389/fpls.2022.931423

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

Torkamaneh D., Laroche J., Bastien M., Abed A., Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*. 2017;18(1):5. DOI 10.1186/s12859-016-1431-9

Wang N., Yuan Y., Wang H., Yu D., Liu Y., Zhang A., Gowda M., Nair S.K., Hao Z., Lu Y., San Vicente F., Prasanna B.M., Li X., Zhang X. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 2020;10(1):16308. DOI 10.1038/s41598-020-73321-8

Wendler N., Mascher M., Himmelbach A., Johnston P., Pickering R., Stein N. Bulbosum to go: a toolbox to utilize *Hordeum vulgare/bulbosum* introgressions for breeding and beyond. *Mol. Plant*. 2015;8(10):1507-1519. DOI 10.1016/j.molp.2015.05.004

Wickland D.P., Battu G., Hudson K.A., Diers B.W., Hudson M.E. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*. 2017;18:586. DOI 10.1186/s12859-017-2000-6

Yao Z., You F.M., N'Diaye A., Knox R.E., McCartney C., Hiebert C.W., Pozniak C., Xu W. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics.* 2020;21(1):360. DOI 10.1186/s12859-020-03704-1

Zheng X., Gogarten S.M., Lawrence M., Stilp A., Conomos M.P., Weir B.S., Laurie C., Levine D. SeqArray – a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*. 2017;33(15):2251-2257. DOI 10.1093/bioinformatics/btx145

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
E.A. Salina orcid.org/0000-0001-8590-847X
D.A. Afonnikov orcid.org/0000-0001-9738-1409

Благодарности. Работа выполнена при поддержке бюджетного проекта FWNR-2022-0020.

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

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 21.07.2023. После доработки 08.09.2023. Принята к публикации 09.09.2023.