Сборка генома *Vitis rotundifolia* Michx. с использованием методов секвенирования третьего поколения (Oxford Nanopore Technologies)

DOI: 10.30901/2227-8834-2021-2-63-71
УДК 634.8; 577.212.3
Поступление/Received: 14.03.2021
Принято/Accepted: 12.05.2021

М. М. АГАХАНОВ[1]*, Е. А. ГРИГОРЬЕВА[2], Е. К. ПОТОКИНА[2], П. С. УЛЬЯНИЧ[3], Ю. В. УХАТОВА[1]

[1] Федеральный исследовательский центр Всероссийский институт генетических ресурсов растений имени Н.И. Вавилова, 190000 Россия, г. Санкт-Петербург, ул. Большая Морская, 42, 44
** m.agahanov@vir.nw.ru*

[2] Санкт-Петербургский государственный лесотехнический университет имени С.М. Кирова, 194021 Россия, г. Санкт-Петербург, Институтский пер., 5

[3] Всероссийский научно-исследовательский институт сельскохозяйственной микробиологии, 196608 Россия, г. Санкт-Петербург, Пушкин, шоссе Подбельского, 10

Genome assembly of *Vitis rotundifolia* Michx. using third-generation sequencing (Oxford Nanopore Technologies)

M. M. AGAKHANOV[1]*, E. A. GRIGOREVA[2], E. K. POTOKINA[2], P. S. ULIANICH[3], Y. V. UKHATOVA[1]

[1] N.I. Vavilov All-Russian Institute of Plant Genetic Resources, 42, 44 Bolshaya Morskaya Street, St. Petersburg 190000, Russia
** m.agahanov@vir.nw.ru*

[2] Saint-Petersburg State Forest Technical University named after S.M. Kirov, 5 Institutsky Lane, St Petersburg, 199034 Russia

[3] All-Russian Research Institute for Agricultural Microbiology, 10 Shosse Podbelskogo, Pushkin, St. Petersburg, 196608 Russia

Североамериканский иммунный к болезням вид винограда *Vitis rotundifolia* Michx. (подрод *Muscadinia* Planch.) рассматривается как потенциальный донор генов устойчивости к опасным болезням винограда – оидиуму и милдью. Сорт 'Dixie' – единственный представитель вида *V. rotundifolia*, сохраняемый в коллекциях *ex situ* на территории России, а именно в коллекции Всероссийского института генетических ресурсов растений имени Н.И. Вавилова (ВИР) в полевых условиях Крымской опытно-селекционной станции – филиала ВИР.

Для получения информации о первичной структуре фрагментов геномной ДНК данного сорта был использован метод секвенирования третьего поколения на платформе MinION, а также привлечены результаты секвенирования на платформе Illumina, имеющиеся в базах данных.

В статье представлено подробное описание последовательности действий для секвенирования генома винограда и полногеномной сборки. Модифицированный метод включает основные этапы оригинальной методики, рекомендованной производителем MinION: 1) выделение ДНК; 2) подготовка библиотек для секвенирования; 3) секвенирование на MinION и биоинформатическая обработка данных; 4) полногеномная сборка методом *de novo* (сборка с использованием данных ONT и сборка с комбинацией данных ONT и Illumina); 5) оценка качества полногеномной сборки. Этап 4 включал не только секвенирование *de novo*, но и анализ имеющихся биоинформатических данных, что позволило уменьшить ошибки и повысить точность при сборке изучаемого генома. ДНК, выделенная из листьев сорта 'Dixie', была секвенирована с использованием двух ячеек MinION типа R9.4.1.

Ключевые слова: виноград, геномная сборка, полногеномное секвенирование, иммунный вид, гены устойчивости.

The immune North American grapevine species *Vitis rotundifolia* Michaux (subgen. *Muscadinia* Planch.) is regarded as a potential donor of disease resistance genes, withstanding such dangerous diseases of grapes as powdery and downy mildews. The cultivar 'Dixie' is the only representative of this species preserved *ex situ* in Russia: it is maintained by the N.I. Vavilov All-Russian Institute of Plant Genetic Resources (VIR) in the orchards of its branch, Krymsk Experiment Breeding Station. Third-generation sequencing on the MinION platform was performed to obtain information on the primary structure of the cultivar's genomic DNA, employing also the results of Illumina sequencing available in databases. A detailed description of the technique with modifications at various stages is presented, as it was used for grapevine genome sequencing and whole-genome sequence assembly. The modified technique included the main stages of the original protocol recommended by the MinION producer: 1) DNA extraction; 2) preparation of libraries for sequencing; 3) MinION sequencing and bioinformatic data processing; 4) *de novo* whole-genome sequence assembly using only MinION data or hybrid assembly (MinION+Illumina data); and 5) functional annotation of the whole-genome assembly. Stage 4 included not only *de novo* sequencing, but also the analysis of the available bioinformatic data, thus minimizing errors and increasing precision during the assembly of the studied genome. The DNA isolated from the leaves of cv. 'Dixie' was sequenced using two MinION flow cells (R9.4.1).

Key words: grapevine, genome assembly, whole-genome sequencing, immune species, resistance genes.

Введение

Виноград (*Vitis vinifera* L.) является экономически значимой для Российской Федерации культурой. По данным международной организации виноградарства и виноделия (OIV) за 2016 г., во всем мире виноградники возделываются на площади ~7,5 млн га, включая 95 тыс. га на территории РФ. Основные проблемы виноградарства связаны с восприимчивостью культуры к фитопатогенам и вредителям. Среди грибковых болезней наиболее опасными считаются оидиум (*Uncinula necator* Burill.), милдью (*Plasmopara viticola* Berl. et Toni.), серая гниль (*Botrytis cinerea* Pers.), фомопсис (*Phomopsis viticola* Sacc.) и антракноз, возбудителем которого является *Elsinoe ampelina* Shear. Получение качественного урожая напрямую зависит от фитосанитарного состояния насаждений.

Наиболее опасным для культуры винограда является оидиум – биотрофный фитопатоген, вызывающий настоящую мучнистую росу виноградной лозы. Патоген поражает не только листья, но и грозди, что приводит к потерям урожая и снижению качества ягод. Ягоды растрескиваются, трещины доходят до семян, что ведет к загниванию ягод и существенному ухудшению их качества из-за повышенной кислотности и пониженного содержания антоцианов и сахара.

Не менее вредоносным является милдью, который поражает почти исключительно молодые зеленые побеги и листья, вследствие чего задерживается и останавливается процесс созревания винограда. Ягоды поражаются значительно реже, чем листья и соцветия; в зависимости от степени зрелости ягод их поражение милдью проявляется различно. Вина из винограда, частично пораженного милдью, имеют специфический горьковатый привкус и неприятный запах. Качество ягод снижается, они содержат мало сахара, много кислот, чрезмерно богаты белковыми и пектиновыми веществами, дают вина, неустойчивые к бактериальным заболеваниям и предрасположенные к окислению.

Использование фунгицидов помогает контролировать развитие грибковых болезней, однако на обработки приходится до 20% от затрат на производство винограда, при этом наносится вред окружающей среде. Кроме того, в последнее время в тренд входит потребление экологически чистых пищевых продуктов, поэтому спрос на органическое, биодинамическое земледелие и переработку экологически чистой продукции растет. Наилучшей альтернативой химической защите является создание новых сортов с генетически обусловленной устойчивостью к патогенам.

Одним из уникальных источников генов устойчивости к оидиуму и милдью для селекции винограда является североамериканский вид *V. rotundifolia* Michx. – виноград круглолистный, мускадин (syn. *V. muscadina* Raf.; *Muscadinia rotundifolia* (Mich.) Smal.), относящийся к подроду *Muscadinia* Planch. Промышленные плантации этого винограда в США являются основой широко распространенного в южных штатах производства мускадинового сока и вина. Этот вид можно считать практически иммунным к возбудителям оидиум и милдью, однако использовать этот ценный источник генов устойчивости для селекции европейских сортов винограда затруднительно. Если виды подрода *Euvitis* Planch. легко скрещиваются между собой, то получить гибриды между *V. vinifera* (подрод *Euvitis*, 38 хромосом) и *V. rotundifolia* (подрод *Muscadinia*, 40 хромосом) ранее удавалось с большим трудом, при этом гибридные сеянцы становились фертильными только после их полиплоидизации (Volynkin et al., 2010). Очевидно, что ни созданный ранее селекционный гибридный материал с участием *V. rotundifolia*, ни материал, полученный вновь с целью поиска генов устойчивости к оидиуму, милдью и филлоксере, не удастся использовать целенаправленно без информации о геноме донора генов устойчивости – иммунного вида *V. rotundifolia*.

До настоящего времени в базе данных Sequence Read Archive (SRA) и European Nucleotide Archive (ENA) были представлены результаты четырех проектов по секвенированию полного генома *V. rotundifolia* с использованием платформы Illumina. Они содержат архивы коротких прочтений ДНК этого вида длиной до 150 пн. Более полную информацию о геноме *V. rotundifolia* можно получить путем прочтения протяженных (до сотен тысяч пар нуклеотидов) последовательностей ДНК с использованием секвенаторов «третьего поколения» – Pacific Biosciences (PacBio) Single Molecule Real-Time (SMRT) и Oxford Nanopore Technologies (ONT) MinION. В отличие от секвенирования коротких (100–300 пн) фрагментов ДНК на платформе Illumina, эти секвенаторы «длинных прочтений» позволяют получить расшифровку фрагмента молекулы ДНК в 100 и более тысяч пар оснований. Кроме того, с появлением таких технологий секвенирования получил широкое распространение так называемый метод «гибридной сборки» полноразмерных геномов (Grigoreva et al., 2019). При гибридной сборке длинные прочтения с секвенаторов PacBio или ONT предоставляют информацию об общей структуре генома, а короткие прочтения с платформы Illumina уточняют сборку в конкретных участках и параллельно корректируют ошибки, которые являются слабым местом технологий секвенирования третьего поколения.

Примечательно, что вариант получения гибридной сборки, по крайней мере бактериальных геномов, по результатам комбинирования данных секвенирования ONT+Illumina оказался эффективнее варианта PacBio+ Illumina, обеспечивая более высокое качество и точность полногеномного прочтения (De Maio et al., 2019).

Совсем недавно с использованием технологий секвенирования PacBio (Pacific Biosciences) и оптического картирования (Bionano's Next Generation Mapping, NGM) был опубликован препринт первой версии полногеномной сборки *Muscadinia rotundifolia* (= *Vitis rotundifolia*) сорта 'Trayshed' с распределением прочтений ДНК по хромосомам (Cochetel et al., 2020).

Настоящая статья посвящена результатам секвенирования геномной ДНК *V. rotundifolia* сорта 'Dixie' с использованием нанопорового секвенатора MinION и созданию на этой основе версии полногеномной сборки этого вида методом *de novo* с использованием исключительно данных нанопорового секвенирования и гибридным методом (с привлечением результатов секвенирования на платформе Illumina, имеющихся в базах данных). Секвенирование *de novo* и анализ имеющихся биоинформационных данных позволили уменьшить ошибки и повысить точность при сборке изучаемого генома

Материал

В качестве объекта исследования был использован сорт 'Dixie' – единственный представитель североамериканского вида *V. rotundifolia*, сохраняемый в живом виде на Крымской опытно-селекционной станции – филиале ВИР (г. Крымск, Краснодарский край), куда он был интро-

дуцирован в 1992 г. из США известным отечественным ампелографом В. А. Носульчаком. С 1992 по 1994 г. на Павловской опытной станции ВИР (г. Санкт-Петербург; ныне научно-производственная база «Пушкинские и Павловские лаборатории ВИР») образец проходил карантинную проверку. После прохождения карантина в 1995 г. саженцы были переданы на Крымскую опытно-селекционную станцию, где их и высадили в коллекцию. Сбор листьев для анализа осуществлялся в октябре 2019 г., затем материал был лиофильно высушен и использован для выделения ДНК.

Методы

Первый этап. ***Выделение ДНК из лиофилизированных листьев винограда***

Для выделения ДНК использовали коммерческий набор DNeasy Plant Mini Kit (Qiagen). Разрушение клеточной стенки в листьях осуществлялось с использованием жидкого азота и гомогенизатора Tissue lyser LT (Qiagen).

Второй этап. ***Подготовка библиотек для секвенирования на платформе MinION***

Секвенирование производилось с использованием прибора MinION (OxfordNanoporeTechnologies). Были задействованы две проточные ячейки типа R9.4.1. Подготовка библиотек и загрузка на ячейки осуществлялась с использованием набора Ligation Sequencing Kit 1D (MinION). Для приготовления каждой из библиотек использовали примерно 2000 нг высокомолекулярной очищенной ДНК. Перед подготовкой библиотек ДНК была фрагментирована с использованием колонок g-TUBE (Covaris), согласно протоколу производителя. Ожидаемый размер фрагментов составил 10 000 пн и 12 000 пн для первой и второй библиотек соответственно.

Третий этап. ***Процедура секвенирования на MinION и биоинформатическая обработка данных***

Технология секвенирования на нанопорах подразумевает получение информации об изменении потенциала силы тока при прохождении молекулы ДНК через белковую пору. В зависимости от структуры азотистого основания (A, G, T, C), сила тока изменяется по-разному. Таким образом, «сырой» сигнал секвенирования представляет собой запись изменения силы тока (fast5) с помощью программ pycoQC. Для последующих биоинформатических операций такой бинарный сигнал необходимо декодировать и перевести в широко используемый и поддерживаемый уже разработанным программным обеспечением (ПО) формат. С этой целью сырой сигнал fast5 проходит процедуру распознавания нуклеотидов (basecalling), которая переводит информацию о нуклеотидах A, G, C, T из бинарной записи fast5 в формат fasta/fastq (de.NBI Nanopore…, 2019). На сегодняшний день доступно несколько программ для такой конвертации, среди них: chiron, flappie, guppy, scrappie (Wick et al., 2019). В данном исследовании был применен один из самых широко используемых и точных алгоритмов декодирования guppy V.0.1.11.

После преобразования сигнала был осуществлен контроль качества секвенирования с помощью программ pycoQC (Leger, Leonardi, 2019) и poretools (Loman, Quinlan, 2014). Контроль качества позволяет получить информацию о базовых статистических показателях проведенного секвенирования: общее количество прочтений (ридов), количество ридов, успешно прошедших процедуру распознавания нуклеотидов, N50 – показатель непрерывности чтения, обозначающий минимальную длину прочтения для более половины (50%) всех полученных ридов, медиана длины ридов и медиана качества, определяемого по шкале Phred (Q) (Ewing, Green, 1998). Для фильтрации возможной контаминации библиотек человеческой ДНК использовали программу minimap2 (Li, 2018). Чтения низкого качества с большой вероятностью ошибок, определяемой по показателю Phred <= 8, и чтения длиной менее 500 пн были отфильтрованы с помощью *NanoFilt* (De Coster et al., 2018).

Четвертый этап. ***Полногеномная сборка V. rotundifolia методом de novo для данных ONT и сборка гибридным методом***

Для сборки генома методом *de novo* применялся алгоритм (pipeline) minimap2/miniasm (Li, 2016), который является одним из наиболее широко используемых при сборке длинных прочтений, и подразумевает последовательное выполнение нескольких этапов (рис. 1). Однако алгоритм minimap2/miniasm не включает в себя шага корректировки возможных ошибок прочтений, полученных с MinION. Кроме того, при использовании только чтений с MinION при сборке *de novo* образуется большое количество пустот в сборке, так называемых «гэпов» (gaps), из-за неравномерности покрытия. Эти проблемы частично решаются процедурой полировки (polishing) сборки, полученной с использованием длинных прочтений, путем ее наложения на множество коротких, но многочисленных прочтений, полученных с Illumina. Поэтому мы дополнили рутину minimap2/miniasm еще одним этапом – коррекцией ошибок с помощью наложения полученной сборки длинными чтениями на сырые прочтения Illumina и ONT, используя программы Racon (Vaser et al., 2017) и Pilon (Walker et al., 2014) (см. рис. 1). Для этого из базы данных сырых прочтений (SRA) геномной ДНК *V. rotundifolia* были использованы риды, полученные с Illumina HiSeq2500, обеспечивающие 30-кратное покрытие генома этого вида (Bioproject SRX2868329: WGS of Muscadine Grape [*Vitis rotundifolia*]). Racon был использован для двух раундов коррекции полученной сборки и Pilon для трех раундов процедуры ее финальной «полировки» (polishing). Это позволило снизить частоту ошибок секвенирования, которые характерны для длинных ридов с MinION, и уменьшить количество пропусков в полученной сборке, тем самым увеличив долю скаффолдов среди всех фрагментов.

Дополнительно был применен альтернативный подход к получению полногеномной сборки *de novo V. rotundifolia* – гибридным методом. Этот метод подразумевает изначальное комбинирование коротких и длинных прочтений, которое выполняет программа SPAdes (hybrid SPAdes) (Antipov et al., 2016). Как и в предыдущем случае, были использованы ранее опубликованные короткие риды с Illumina HiSeq2500 (Bioproject SRX2868329: WGS of Muscadine Grape [*Vitis rotundifolia*]), комбинированные с результатами секвенирования на MinION.

Алгоритм гибридной сборки включает в себя сборку коротких прочтений графами Де Брюина с последующем закрытием «пропусков» сборки с помощью длинных ридов. Отличительной особенностью алгоритма является автоматический подбор наиболее подходящих k-меров (k-mer – короткие фрагменты ридов варьирующей длины, на которые разбиваются прочтения для последующей сборки).

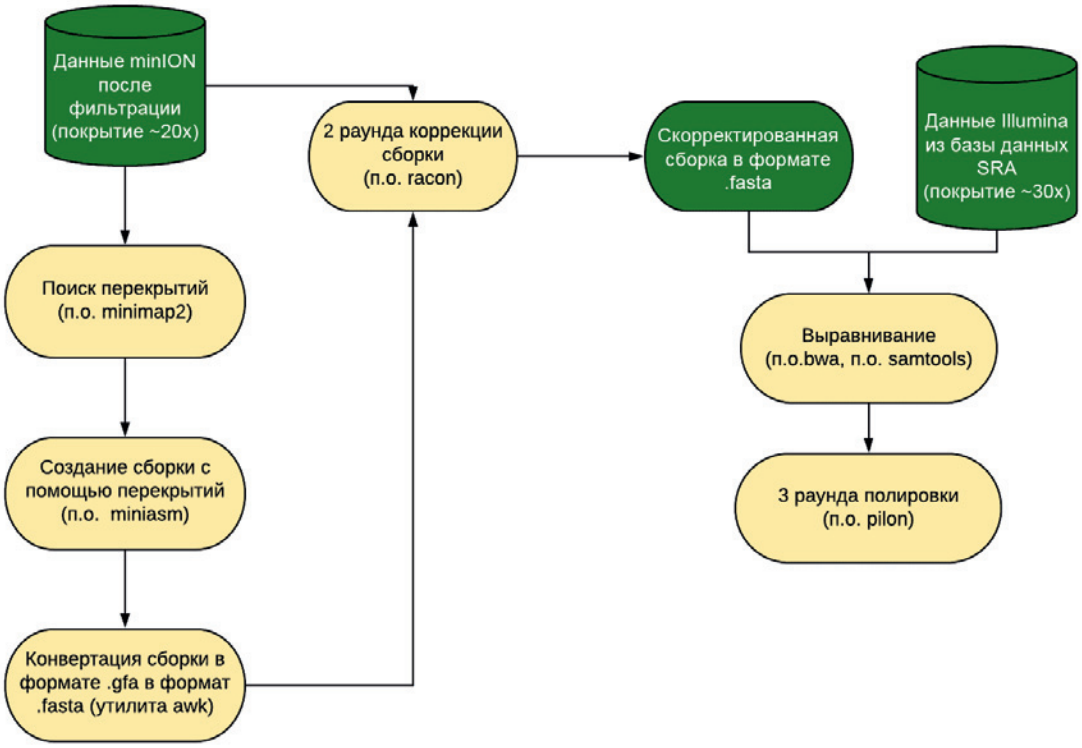

Рис. 1. Схема процесса (pipeline) сборки генома *Vitis rotundifolia* методом *de novo* по результатам секвенирования на MinION с использованием программы minimap2/miniasm

Fig. 1. Flowchart of the *Vitis rotundifolia* genome assembly by the *de novo* method based on the results of sequencing on the MinION platform using the minimap2/miniasm pipeline

Пятый этап. *Оценка качества полногеномной сборки V. rotundifolia*

Для оценки качества сборки и ее фрагментированности полногеномная сборка *V. rotundifolia*, полученная двумя различными способами, была проанализирована с помощью BUSCO V.3.0.2 (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015; Seppey et al., 2019). BUSCO представляет собой биоинформатический инструмент, который позволяет оценить качество полногеномной сборки, полученной из множества коротких прочтений, основываясь не на технических параметрах, как, например, показатель N50 или статистическое распределение длин полученных контигов, а принимая во внимание «смысловой» параметр собранного генома – полноту представленности генов, ортологи которых встречаются, например, более чем у 90% видов Embryophyta. При этом BUSCO принимает во внимание преимущественно уникальные гены (single-copy orthologs).

Алгоритм BUSCO включает три этапа:

1) поиск функциональных последовательностей в анализируемой сборке путем ее выравнивания методом tBlast на одну из доступных баз данных генов-ортологов BUSCO;

2) прогнозирование структуры генов для выявленных функциональных последовательностей с помощью программы Augustus (Stanke et al., 2006);

3) заключительный этап, который определяет, насколько выявленные в сборке функциональные последовательности являются полноразмерными. Если их длина находится в пределах двух стандартных отклонений от длины последовательностей генов-ортологов данной группы в BUSCO, аннотированным генам присваивается статус «полноразмерный» (Complete BUSCOs). Полноразмерные гены, обнаруженные не в единственной копии, относятся к группе «дублированный» (Duplicated). Частично воссозданные гены классифицируются как «фрагментированные» (Fragmented), не обнаруженные гены классифицируются как «отсутствующие» (Missed). Для анализа двух версий полногеномной сборки *V. rotundifolia* использовали 1614 последовательностей ортологичных генов из базы данных BUSCO высших растений (Embryophyta) embryophyta_odb 10.2019-11-20.

Пошаговый протокол представлен в Приложении 1 (Supplementary Materials 1)[1].

Результаты

Секвенирование MinION, распознавание нуклеотидов, контроль качества прочтений

Две ячейки MinION (тип R9.4.1) были использованы для секвенирования двух библиотек геномной ДНК, выделенной из листьев *V. rotundifolia* (сорт 'Dixie'). Перед загрузкой в ячейку первая библиотека содержала 1008 нг ДНК (с концентрацией 84 нг/мкл), вторая библиотека – 1152 нг (96 нг/мкл). Время работы одной ячейки в среднем составило 48 ч. Всего с двух ячеек было получено 1 748 466 прочтений (далее – ридов) (табл. 1).

После декодирования сырого сигнала fast5 в fastq/fasta и коррекции ошибок с помощью gupppy было получено 1 738 535 ридов. Для полученных ридов произвели контроль качества с помощью pyroQC, получив такие статистические показатели, как количество ридов, количество оснований, N50, медианы длин и качества ридов. Показатель N50 (минимальная длина, которую имели более половины полученных ридов) по результатам запуска первой ячейки составил 7310 пн, второй – 14600 пн.

[1] Electronic supplementary material. The online version of this article: (https://doi.org/10.30901/2227-8834-2021-2-63-71).

Таблица 1. Общая статистика «сырых» данных секвенирования двух библиотек геномной ДНК *Vitis rotundifolia*, полученных с двух ячеек MinION с помощью программы NanoStat V.0.8.1из пакета Nanopack

Table 1. General statistics of "raw" sequencing data for two *Vitis rotundifolia* genomic DNA libraries obtainedfrom two MinION cells using the NanoStat V.0.8.1 program from the Nanopack package

Статистический показатель	Значение
Средняя длина рида (пн)	5,987.6
Среднее качество рида (Phred)	12.3
Медиана длины рида (пн)	4,142.0
Медиана качества рида (Phred)	12.5
Общее количество ридов	1,748,466
N50 длины рида (пн)	9,514
Всего прочитано нуклеотидов (пн)	10,469,195,297
>Q5	1744047 (99.7%) 10468.0Mb
>Q7	1740159 (99.5%) 10465.2Mb
>Q10	1478525 (84.6%) 8862.5Mb
>Q12	1040592 (59.5%) 6223.0Mb
>Q15	119125 (6.8%) 497.8Mb
Топ 5 ридов с лучшим качеством по шкале Phred (длина, пн)	
1	20.4 (422)
2	20.1 (334)
3	20.0 (1241)
4	20.0 (525)
5	19.9 (252)
Топ 5 самых длинных ридов – длина (качество по шкале Phred)	
1	122768 (13.1)
2	121913 (9.3)
3	113066 (10.0)
4	112400 (8.7)
5	112314 (11.7)

Медианы длин ридов составили 3500 пн для первой ячейки и 6380 пн – для второй (табл. 2).

На следующем этапе анализа данных осуществили фильтрацию прочтений низкого качества, коротких прочтений и наиболее вероятных прочтений инородной ДНК. Для удаления возможной контаминации биологическим материалом человека процедуру фильтрации произвели с помощью minimap2. Всего был удален 41 501 рид (~2%), содержавший 100% сходство с референсным геномом человека. NanoFilt применили для фильтрации по показателям качества прочтений (вероятности встретить ошибочно распознанный нуклеотид)

Таблица 2. Основные статистические характеристики данных секвенирования ONT после процедуры распознавания нуклеотидов (basecalling с помощью guppy), полученных с помощью pyroQC

Table 2. Main statistical characteristics of the obtained data after nucleotide recognition procedures (basecalling) using pyroQC

Количество активных пор	Номер ячейки	Количество ридов	Количество оснований	Медиана длины	Медиана качества по шкале Phred	N50
481	1	1 220 594	5 732 084 000	3530	11.809	7310
455	2	517 941	4 729 669 000	6380	11.373	14600

и длины ридов. Для последующего анализа были сохранены риды, имеющие показатель качества по шкале Phred не менее 8 и длину прочтения более 500 пн. После фильтрации по этим показателям получили 163 5299 высококачественных ридов, составивших в общей сложности более 10 млрд (10 197 618 064) пн. Отфильтрованные риды низкого качества составили 5,9% от общего числа прочтений.

Принимая во внимание опубликованный размер генома культурного винограда (*V. vinifera*) в ~486 тыс. пн (Canaguier et al., 2017), можно предположить, что полученный объем данных обеспечивает почти 21-кратное «покрытие» секвенируемого генома *V. rotundifolia.* Такая глубина прочтений позволяет осуществлять сборку методом *de novo* без использования референсного генома (NCBI…, 2020).

По результатам комбинирования коротких и длинных прочтений нами были протестированы разные длины k-меров (21, 33, 55, 77 пн). В качестве наиболее результативной алгоритм автоматически выявил длину k-меров 77 пн. Сборка была протестирована на основные статистические показатели с помощью программы QUAST (Gurevich et al., 2013). Общая длина сборки составила 539 Мб (млн пн), что соотносится с аналогичным показателем для референсного генома *V. vinifera*, размер которого был определен в ~486 Мб (Canaguier et al., 2017). В таблице 3 представлены основные статистические показатели качества полногеномной сборки *V. rotundifolia*, полученной с использованием *de novo*, привлекая только данные нанопорового секвенирования (ONT) и используя данные Illumina и ONT (гибридный метод).

По результатам сравнения двух подходов к получению полногеномной сборки *V. rotundifolia* установлено, что оба метода имеют свои преимущества и недостатки. Гибридный метод позволяет получить большее количество скаффолдов, однако при этом сборка получается намного более фрагментированной, чем при сборке с использованием только данных нанопорового секвенирования. Сборка *de novo* с использованием minimap2/miniasm с последующей процедурой полировки (polishing) на сырые риды Illumina дает возможность получить более длинные скаффолды, однако в гораздо меньшем количестве, из-за чего большая часть исследуемого генома остается не покрытой скаффолдами.

Полученные разными способами две версии полногеномной сборки *V. rotundifolia* различаются также по количеству выявленных повторяющихся последовательностей. Для их идентификации в полученных сборках использовали алгоритм «маскинга» (masking), который позволяет найти и скрыть повторы путем сравнения полногеномной сборки и доступных баз данных повторяющихся элементов с помощью программы Repeat Masker (Tarailo-Graovac, Chen, 2009). Для сборки, полученной гибридным методом, выявили повторы общей протяженностью 1 651 434 пн, для сборки с использованием данных ONT – почти в 3,5 раза меньше (484 681 пн).

На рисунке 2 представлены результаты оценки полноты представленности последовательностей ортологичных генов в сборке *de novo*, полученной с использованием данных ONT и гибридным методом.

В целом можно заключить, что обе версии сборки удовлетворяют показателям представленности в них последовательностей генов-ортологов, универсальных для Embryophyta, хотя сборка гибридным методом выглядит более фрагментированной. Результаты ана-

Таблица 3. Основные характеристики полногеномной сборки *Vitis rotundifolia,* полученной методом *de novo* и гибридным методом по результатам оценки с помощью QUAST

Table 3. Main characteristics of the *Vitis rotundifolia* whole-genome assembly by the *de novo* and hybrid methods according to the results of the assessment using QUAST

Показатели качества сборки	Сборка *de novo* ONT+Illumina (гибридный метод) (пн)	Сборка *de novo*, исключительно используя данные ONT (пн)-
Количество контигов (>= 0 пн)	809 308	2039
Количество контигов (>= 1000 пн)	43 425	2037
Количество контигов (>= 5000 пн)	15 430	2020
Количество контигов (>= 10000 пн)	9292	1998
Количество контигов (>= 25000 пн)	4056	1808
Количество контигов (>= 50000 пн)	1516	1493
Общая длина контигов (сборки) (>= 0 пн)	428 439 192	386 122 654
Количество скаффолдов	84 025	2039
Самый длинный скаффолд	319 841	2 353 788
N50	24 761	374 653
N75	6901	173 204
L50	4103	293
L75	12 305	669
GC%	33,31	33,94

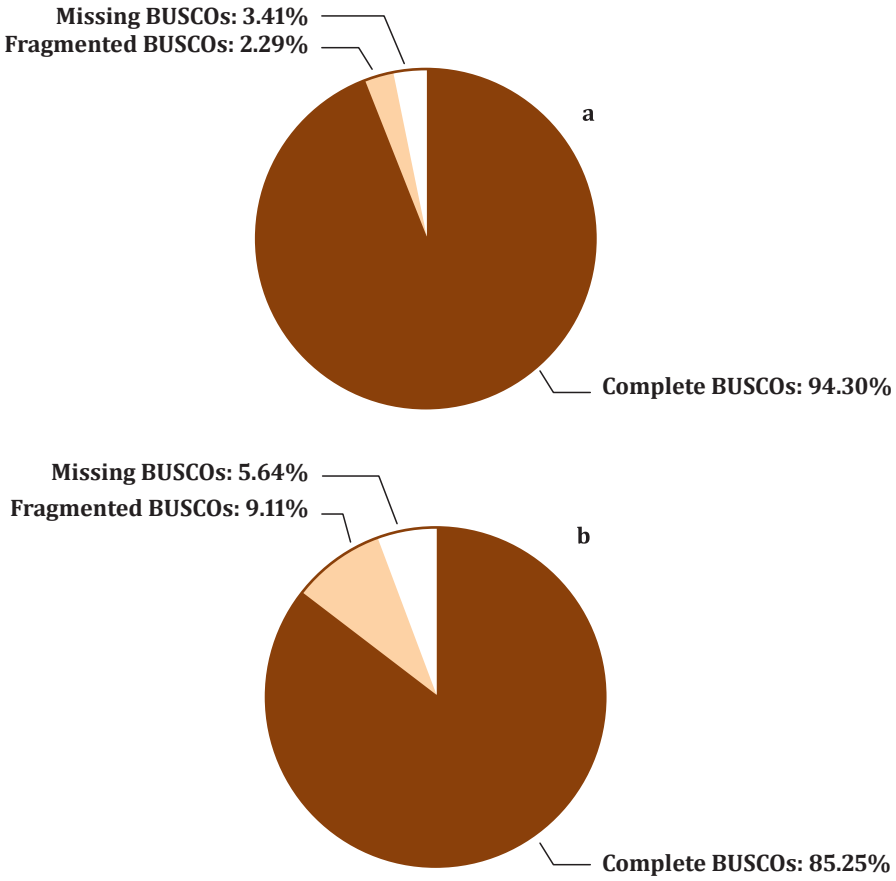

Рис. 2. Представленность последовательностей ортологичных генов из базы данных BUSCO высших растений (Embryophyta) в полногеномной сборке *Vitis rotundifolia*, полученной методом *de novo* с использованием исключительно данных ONT (a) и гибридным методом (ONT+Illumina) (b)

Fig. 2. Representation of orthologous gene sequences from the BUSCO database of higher plants (Embryophyta) in the *de novo* whole-genome assembly of *Vitis rotundifolia* obtained using only ONT data (a) and the hybrid method (ONT+Illumina) (b)

лиза BUSCO выявили несколько больше полных последовательностей генов-ортологов (Complete BUSCOs) для сборки с использованием исключительно данных ONT, что можно объяснить простотой предсказания генов для длинных непрерывных чтений. С другой стороны, судя по показателям общей длины сборки и количества контигов и скаффолдов, можно заключить, что сборка гибридным методом покрыла более протяженную часть уникальных участков генома, чем сборка с использованием minimap2/miniasm.

Предложенная нами полногеномная сборка иммунного вида *V. rotundifolia* может быть также проанализирована с точки зрения идентификации гомологичных участков с опубликованным геномом культурного винограда *V. vinifera* 12X (International Grape Genome Program, GenBank assembly accession: GCA_000003745.2) и оценки степени сходства геномов двух видов. Особый интерес может представлять выравнивание полученной сборки на 12 хромосому в геноме *V. vinifera*, где ранее был картирован локус RUN1/RPV, ассоциированный с устойчивостью к оидиуму (RUN1) и милдью (RPV1).

Исследование Cochetel et al. (2020), опубликовавших первую версию сборки генома *V. rotundifolia*, позволило проанализировать различия в структуре этого локуса RUN1/RPV1 у иммунного *V. rotundifolia* (сорт 'Trayshed') и поражаемого *V. vinifera* (сорт 'Sauvignon blanc'). Локус RUN1/RPV1, фланкированный двумя микросателлитными маркерами VMC4f3.1 и VMC8g9 (Barker et al., 2005), протяженностью 5 млн пар оснований был расшифрован на хромосоме 12 *V. vinifera*. Аналогичный участок на хромосоме 12 у *V. rotundifolia* соответствовал интервалу почти 7,3 млн пн вследствие многочисленных дупликаций внутри этого участка генома. В пределах этого интервала на хромосоме 12 у *V. vinifera* были идентифицированы 33 R-гена из семейства NBS-LRR (Nucleotide binding site leucine-rich repeat), роль которых в формировании устойчивости к патогенам описана для многих видов растений, в том числе и винограда (Zini et al., 2019). Для *V. rotundifolia* в том же локусе выявили 57 генов NBS-LRR, причем один класс этого семейства генов, TIR-X со специфичным доменом, был обнаружен только в геноме у устойчивого *V. rotundifolia* (сорт 'Trayshed'). Присутствие генов TIR-X постулируется в качестве одной из возможных причин устойчивости *V. rotundifolia* к фитопатогенам (Cochetel et al., 2020). Аналогичный анализ структуры локуса RUN1/RPV1 у еще одного сорта *V. rotundifolia* (сорт 'Dixie') с целью выявления вставок, делеций, повторов может быть полезным для уточнения вероятных генов-кандидатов, определяющих устойчивость *V. rotundifolia* к оидиуму и милдью.

Заключение

В настоящей статье представлены методика и результаты полногеномного секвенирования иммунного вида *Vitis rotundifolia* на примере сорта 'Dixie', выполненного с использованием секвенатора «третьего поколения» MinION (Oxford Nanopore Technologies). Более 1,6 млн высококачественных прочтений длиной ~5 тыс. пн, составивших в общей сложности более 10 млрд пн, были депонированы в базы данных NCBI, SRA, ENA и доступны для использования. Помимо депонированных «сырых» прочтений, также была создана и опубликована в NCBI версия полногеномной сборки *V. rotundifolia*, выполненная «гибридным» методом, с комбинированием длинных прочтений, полученных с MinION, и коротких ридов Illumina, доступных из баз данных. Созданный исследовательский ресурс может быть использован для молекулярно-генетической идентификации генов устойчивости к болезням и вредителям винограда, донором которых является этот североамериканский вид.

Доступность данных

Сырые данные секвенирования генома *Vitis rotundifolia* двух ячеек minION были депонированы в базу данных Национального центра биотехнологической информации США (U.S. National Center for Biotechnology Information, NCBI) (NCBI..., 2020) и базу данных SRA (Sequence Reads Archive) (Bioproject: PRJNA649974; Biosample: SAMN15690594; SRA ENA: SRS7124084). Данные сборки гибридным методом (hybrid SPAdes) доступны в базе данных ENA (PRJNA649974).

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-316-90007.

The study was funded by the Russian Foundation for Basic Research in the framework of Research Project No. 19-316-90007

References / Литература

Antipov D., Hartwick N., Shen M., Raiko M., Lapidus A., Pevzner P.A. PlasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016;32(22):3380-3387. DOI: 10.1093/bioinformatics/btw493

Barker C.L., Donald T., Pauquet J., Ratnaparkhe M.B., Bouquet A., Adam-Blondon A.F. et al. Genetic and physical mapping of the grapevine powdery mildew resistance gene, *Run1*, using a bacterial artificial chromosome library. *Theoretical and Applied Genetics*. 2005;111(2):370-377. DOI: 10.1007/s00122-005-2030-8

Canaguier C.A., Grimplet J., Di Gaspero G., Scalabrin S., Duchêne E., Choisne N. et al. A new version of the grapevine reference genome assembly (12X. v2) and of its annotation (VCost. v3). *Genomics Data*. 2017;14:56-62. DOI: 10.1016/j.gdata.2017.09.002

Cochetel N., Minio A., Massonnet M., Vondras A., Figueroa-Balderas R., Cantu D. Diploid chromosome-scale assembly of the *Muscadinia rotundifolia* genome supports chromosome fusion and disease resistance gene expansion during *Vitis* and *Muscadinia* divergence. *G3 (Bethesda)*. 2021;11(4):jkab033. DOI: 10.1093/g3journal/jkab033

De Coster W., D'Hert S., Schultz D.T., Cruts M., Van Broeckhoven Ch. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666-2669. DOI: 10.1093/bioinformatics/bty149

De Maio N., Shaw L.P., Hubbard A., George S., Sanderson N.D., Swann J. et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*. 2019;5(9):e000294. DOI: 10.1099/mgen.0.000294

de.NBI Nanopore Training Course. The Tutorial Data Set. Basecalling. 2019. Available from: https://denbi-nanopore-training-course.readthedocs.io/en/latest/basecalling/index.html [accessed Dec. 10, 2020].

Ewing B., Green P. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Research*. 1998;8(3):186-194. DOI: 10.1101/gr.8.3.186

Grigoreva E., Ulianich P., Ben C., Gentzbittel L., Potokina E. First insights into the guar (*Cyamopsis tetragonoloba* (L.) Taub.) genome of the 'Vavilovskij 130' accession, using second and third-generation sequencing technologies. *Russian Journal of Genetics*. 2019;55(11):1406-1416. DOI: 10.1134/S102279541911005X

Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. DOI: 10.1093/bioinformatics/btt086

Leger A., Leonardi T. PycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*. 2019;4(34):1236. DOI: 10.21105/joss.01236

Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*. 2016;32(14):2103-2110. DOI: 10.1093/bioinformatics/btw152

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100. DOI; 10.1093/bioinformatics/bty191

Loman N.J., Quinlan A.R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*. 2014;30(23):3399-3401. DOI: 10.1093/bioinformatics/btu555

NCBI: National Center for Biotechnology Information. Bioproject 649974. *Vitis rotundifolia* cultivar: Dixie. *Vitis rotundifolia* Michx. whole genome sequencing and assembly using nanopore technology (Oxford Nanopore Technologies). Accession: PRJNA649974. Registration date: Nov. 10, 2020. Available from: https://www.ncbi.nlm.nih.gov/bioproject/649974 [accessed Dec. 07, 2020].

Seppey M., Manni M., Zdobnov E.M. BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology*. 2019;1962:227-245. DOI: 10.1007/978-1-4939-9173-0_14

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., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*. 2006;34 Suppl 2:W435-W439. DOI: 10.1093/nar/gkl200

Tarailo-Graovac M., Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*. 2009;25(1):4.10.1-4.10.14. DOI: 10.1002/0471250953.bi0410s25

Vaser R., Sović I., Nagarajan N., Šikić M. Fast and accurate de novo genome assembly from long uncorrected

reads. *Genome Research*. 2017;27(5):737-746. DOI: 10.1101/gr.214270.116

Volynkin V.A., Zlenko V.A., Poluliakh A.A., Oleinikov N.P., Likhovskoi V.V. Results of experiment research into the formation of genetic diversity in the Vitaceae family during natural evolution. *Magarach. Viticulture and Winemaking*. 2010;40:12-16. [in Russian] (Волынкин В.А. Зленко В.А., Популях А.А., Олейников Н.П., Лиховской В.В. Результаты экспериментальных исследований формирования генетического разнообразия у семейства винограда Vitaceae в процессе естественной эволюции. *Магарач. Виноградарство и виноделие*. 2010;40:12-16).

Walker B.J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar Sh. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9(11):e112963. DOI: 10.1371/journal.pone.0112963

Wick R.R., Judd L.M., Holt K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*. 2019;20(1):129. DOI: 10.1186/s13059-019-1727-y

Zini E., Dolzani Ch., Stefanini M., Gratl V., Bettinelli P., Nicolini D. et al. *R*-loci arrangement versus downy and powdery mildew resistance level: A *Vitis* hybrid survey. *International Journal of Molecular Sciences*. 2019;20(14):3526. DOI: 10.3390/ijms20143526

Прозрачность финансовой деятельности / The transparency of financial activities

Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

The authors declare the absence of any financial interest in the materials or methods presented.

Для цитирования / How to cite this article

Агаханов М.М., Григорьева Е.А., Потокина Е.К., Ульянич П.С., Ухатова Ю.В. Сборка генома *Vitis rotundifolia* Michx. с использованием методов секвенирования третьего поколения (Oxford Nanopore Technologies). Труды по прикладной ботанике, генетике и селекции. 2021;182(2):63-71. DOI: 10.30901/2227-8834-2021-2-63-71

Agakhanov M.M., Grigoreva E.A., Potokina E.K., Ulianich P.S., Ukhatova Y.V. Genome assembly of *Vitis rotundifolia* Michx. using third-generation sequencing (Oxford Nanopore Technologies). Proceedings on Applied Botany, Genetics and Breeding. 2021;182(2):63-71. DOI: 10.30901/2227-8834-2021-2-63-71

Авторы благодарят рецензентов за их вклад в экспертную оценку этой работы / The authors thank the reviewers for their contribution to the peer review of this work

Дополнительная информация / Additional information

Полные данные этой статьи доступны / Extended data is available for this paper at https://doi.org/10.30901/2227-8834-2021-2-63-71

Мнение журнала нейтрально к изложенным материалам, авторам и их месту работы / The journal's opinion is neutral to the presented materials, the authors, and their employer

Авторы одобрили рукопись / The authors approved the manuscript

Конфликт интересов отсутствует / No conflict of interest

ORCID

Agakhanov M.M. https://orcid.org/0000-0003-2438-9156
Grigoreva E.A. https://orcid.org/0000-0002-7285-2291
Potokina E.K. https://orcid.org/0000-0002-2578-6279
Ulianich P.S. https://orcid.org/0000-0002-2768-505X
Ukhatova Y.V. https://orcid.org/0000-0001-9366-0216