

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

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

Н.А. Шмаков<sup>1, 2</sup> 

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

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

 [shmakov@bionet.nsc.ru](mailto:shmakov@bionet.nsc.ru)

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

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

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

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

N.A. Shmakov<sup>1, 2</sup> 

<sup>1</sup> Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

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

 [shmakov@bionet.nsc.ru](mailto:shmakov@bionet.nsc.ru)

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

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

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

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

## Введение

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

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

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

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

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

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

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

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

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

## Материалы и методы

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

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

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

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

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

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

**Таблица 1.** Характеристики использованных библиотек коротких прочтений

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

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

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

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

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

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

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

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

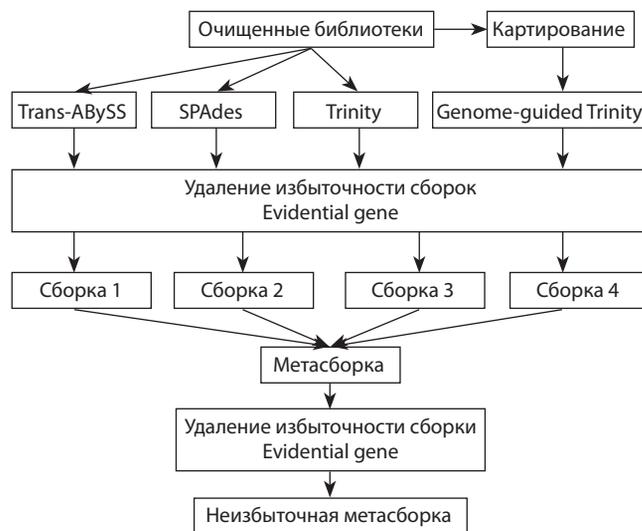


Рис. 1. Схема получения индивидуальных сборок *de novo* и метасборки транскриптома ячменя.

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

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

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

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

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

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

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

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

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

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

## Результаты

### Эксперимент alm

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

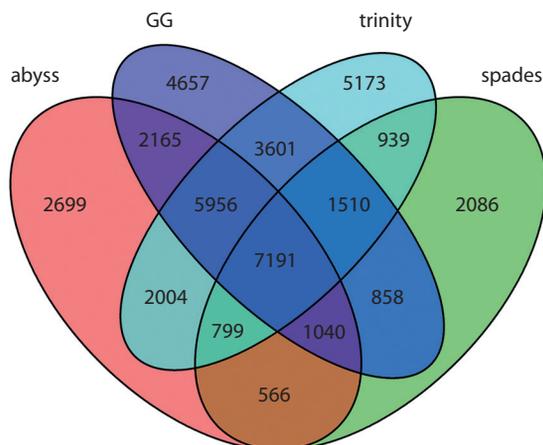
Метасборка транскриптома ячменя линий i:BwAlm и Bowman, полученная из сборок *de novo*, созданных с помощью rnaSPAdes, Trans-ABYSS и Trinity, и геном-ориентированной сборки trinity, до удаления избыточности состоит из 169232 контигов. Избыточность метасборки включает 68414 контигов суммарной длиной 46440750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

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

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

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

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



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

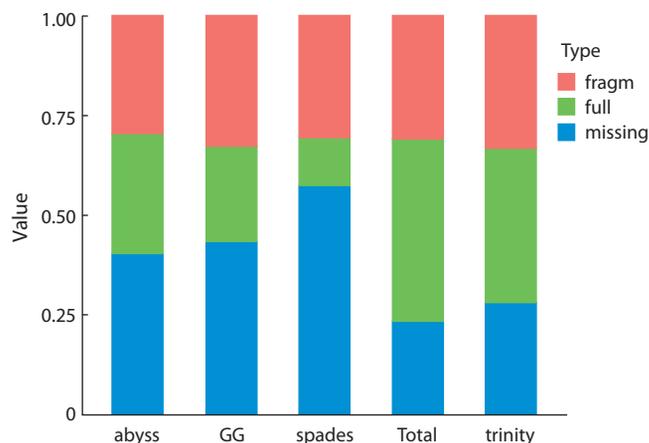
Проведена оценка покрытия контигов прочтениями библиотек в индивидуальных сборках и метасборке транскриптома с помощью технологии псевдокартирования. Установлено, что наибольшая доля прочтений была выровнена на сборку транскриптома abyss, тогда как наименьшая – на сборку spades. На метасборку транскриптома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

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

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

Далее для оценки вклада каждого из сборщиков в структуру метасборки транскриптома была проведена оценка перекрывания множеств CDS ячменя, встреченных в каждой из индивидуальныхборок (рис. 2). Как можно видеть, 7191 CDS ячменя был обнаружен во всех четырех индивидуальных сборках транскриптома, еще 9305 CDS найдены в трех сборках из четырех. 14615 CDS были обнаружены только в одной из четырехборок, из которых наибольшее количество (5173) выявлено только в сборке trinity, наименьшее (2086) – только в сборке spades. Максимальное перекрывание множеств, обнаруженных CDS, наблюдалось между индивидуальными сборками trinity и GG – 18258 CDS.

В контигах каждой изборок были предсказаны открытые рамки считывания (ОРС). Найденные в контигах общей сборки ОРС кодируют 58636 белковых продуктов длинами не менее 30 аминокислотных остатков. Эти белковые продукты были использованы для того, чтобы оценить полноту сборки при помощи программы BUSCO (рис. 3). В метасборке транскриптома количество выяв-



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

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

#### Эксперимент blp

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

Исходная избыточная метасборка транскриптома ячменя линий Bowman и BLP состоит из 133 070 контигов. После удаления избыточности в метасборке осталось 32466 контигов суммарной длиной 25 184 753 основания. Таким образом, в ходе удаления избыточности количество контигов было уменьшено до 24.4 % от исходного. Отметим также, что метасборка транскриптома в эксперименте blp имеет более высокое значение длин контигов N50, чем индивидуальные сборки, из которых она составлена. 72.1 % всех прочтений из библиотек эксперимента blp было картировано на метасборку транскриптома. По этому показателю метасборка уступает сборке GG (77.6 %), но опережает три другие индивидуальные сборки.

В сборке транскриптома *de novo* исследуемых линий был проведен поиск известных CDS с помощью программы TransRate (табл. 5). Гомологию к известным CDS ячменя показывают от 19848 контигов в сборке spades до 29412 контигов в сборке GG. При этом наибольшее количество CDS ячменя обнаружено в сборке trinity, а максимальное количество CDS ячменя, покрытых контигами сборки не менее чем на 95 % своей длины, – в метасборке транскриптома (1825). Доля контигов из сборки, для которых была установлена гомология к известным CDS ячменя, в метасборке составляет 74.5 %, что ниже, чем у всех индивидуальныхборок, кроме trinity.

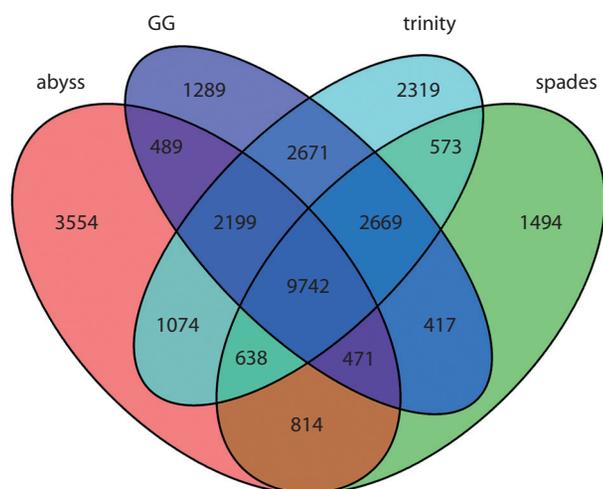
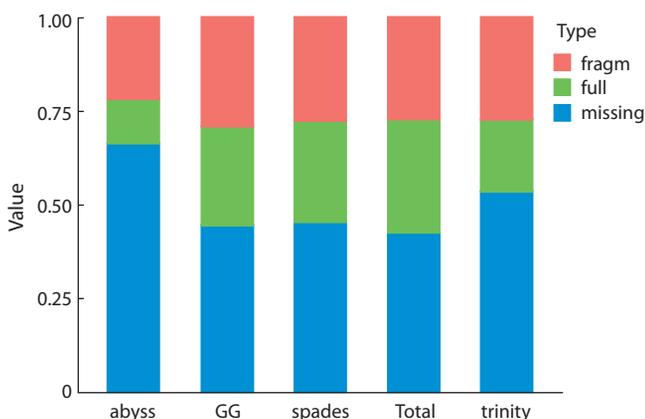
Далее был проведен поиск перекрывания полученных для индивидуальныхборок транскриптома списков CDS и оценен вклад каждой индивидуальной сборки в общую структуру (рис. 4). Во всех четырех индивидуальных сборках транскриптома *de novo* были обнаружены

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

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

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

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

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

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

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

#### Сравнение качества сборок *de novo*

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

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

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

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

## Обсуждение

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

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

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

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

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

## Заключение

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

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

## Список литературы / References

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

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

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

**Конфликт интересов.** Автор заявляет об отсутствии конфликта интересов.

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