Вавиловский журнал генетики и селекции. 2021;25(1):30-38
DOI 10.18699/VJ21.004

Original Russian text www.bionet.nsc.ru/vogis/

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

N.A. Shmakov[1, 2] ✉

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomics Center, Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ shmakov@bionet.nsc.ru

Abstract. *De novo* transcriptome assembly is an important stage of RNA-seq data computational analysis. It allows the researchers to obtain the sequences of transcripts presented in the biological sample of interest. The availability of accurate and complete transcriptome sequence of the organism of interest is, in turn, an indispensable condition for further analysis of RNA-seq data. Through years of transcriptomic research, the bioinformatics community has developed a number of assembler programs for transcriptome reconstruction from short reads of RNA-seq libraries. Different assemblers makes it possible to conduct a *de novo* transcriptome reconstruction and a genome-guided reconstruction. The majority of the assemblers working with RNA-seq data are based on the De Bruijn graph method of sequence reconstruction. However, specifics of their procedures can vary drastically, as do their results. A number of authors recommend a hybrid approach to transcriptome reconstruction based on combining the results of several assemblers in order to achieve a better transcriptome assembly. The advantage of this approach has been demonstrated in a number of studies, with RNA-seq experiments conducted on the Illumina platform. In this paper, we propose a hybrid approach for creating a transcriptome assembly of the barley *Hordeum vulgare* isogenic line Bowman and two nearly isogenic lines contrasting in spike pigmentation, based on the results of sequencing on the IonTorrent platform. This approach implements several *de novo* assemblers: Trinity, Trans-ABySS and rnaSPAdes. Several assembly metrics were examined: the percentage of reference transcripts observed in the assemblies, the percentage of RNA-seq reads involved, and BUSCO scores. It was shown that, based on the summation of these metrics, transcriptome meta-assembly surpasses individual transcriptome assemblies it consists of.

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

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

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

Н.А. Шмаков[1, 2] ✉

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

Аннотация. Реконструкция транскриптома *de novo* – важная стадия биоинформатического анализа данных RNA-seq, которая позволяет получить последовательности транскриптов, присутствующих в изучаемом биологическом образце. Наличие точной и полной последовательности транскриптома организма, в свою очередь, является необходимым условием для дальнейшей работы с данными RNA-seq. Биоинформатическим сообществом было создано множество программ-сборщиков для реконструкции транскриптома из коротких прочтений RNA-seq. Сборщики позволяют проводить как *de novo* реконструкцию транскриптома, так и реконструкцию, основанную на картировании коротких прочтений RNA-seq на последовательность референсного генома организма. Большинство *de novo* сборщиков, работающих с данными RNA-seq, применяют технологию реконструкции последовательностей методом графов де 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 Bowman и двух почти изогенных линий, полученных на основе Bowman и контрастных по окраске колоса, используя данные, полученные при секвенировании матричной РНК на платформе IonTorrent. В данном подходе применяются несколько индивидуальных сборщиков: Trans-ABySS, rnaSPAdes 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.

Introduction

Next generation massively parallel sequencing technology applied to RNA (RNA-seq) is a method of choice in modern transcriptomics researches. It involves several steps: extraction of total mRNA of a biological sample, fragmentation of mRNA and simultaneous sequencing a large number of obtained short fragments (Engström et al., 2013; Hrdlickova et al., 2017).

De novo transcriptome assembly from sequenced fragments is one of the most important stages of transcriptome profiling experiment (Chang et al., 2014). It allows researcher to obtain sequences of mRNA molecules from the studied sample. Presently there are two main approaches to transcriptome sequences reconstruction from short read libraries – so-called OLC method (Overlap–Layout–Consensus) and de Bruijn graph method (Li et al., 2012; Schliesky et al., 2012). OLC method is based on pairwise alignment of reads and construction of oriented graphs where each node is one read. Overlaps between reads represent edges of the graph. This method is more suitable for contig assembly from a relatively smaller number of long reads with large overlapping regions, and thus is more frequently used to assemble sequences obtained with Saenger sequencing method or third generation sequencing methods (Cui et al., 2020).

The other method is based on construction of de Bruijn graph in which nodes are represented by *k*-mers – nucleotide sequences of given length *k*. Next, all paths on the graph that comprise sequences of short reads in RNA-seq libraries are marked. Then, all paths on the graph that contain continuous sequences of overlapping reads are marked. Thus, sequences of contigs consisting of short reads from the libraries are obtained. This method is implemented in several assemblers, namely Trinity (Grabherr et al., 2013), Trans-ABySS (Robertson et al., 2010), SOAPdenovo-Trans (Xie et al., 2014), Oases (Schulz et al., 2012).

An important parameter for de Bruijn graphs-based assemblers is *k* – length of *k*-mers used in de Bruijn graph construction. *k*-mers are words located in the nodes of de Bruijn graph. This parameter can be set by user prior to starting assembler program. Increasing *k* results in higher precision of assembly, but at the same time it makes it more computationally difficult (Fu et al., 2018). At larger *k*, the assembler might fail to detect a limited intersection between the reads, if its size is smaller than *k*. Often the following strategy is used: several preliminary assemblies are conducted at different values of *k*, then assemblies are combined, and redundancy reduction is performed, which results in a final *de novo* transcriptome assembly (Wang, Gribskov, 2017).

Since a large number of transcriptome *de novo* assemblers have been developed to date, researches were dedicated to answering the question of performance and precision of these programs. Reviews comparing different transcriptome assemblers usually mention Trinity, SOAPdenovo-Trans, Velvet-Oases among the best and most popular tools (Jain et al., 2013; Honaas et al., 2016; Wang, Gribskov, 2017). Trinity distributive, aside from the assembler itself, includes a large number of utilities for assembly quality assessment, removal of poorly represented contigs and other manipulations with the *de novo* assembly. SOAPdenovo-Trans is mentioned as the program fitting for large plant transcriptomes *de novo* assembly (Payá-Milans et al., 2018).

Given the diversity of modern assemblers, none of them are perfect and capable to satisfy all the requirements for completeness and quality of the assembly. Thus, it was proposed that implementing several *de novo* assemblers followed by creating of a single ‘meta-assembly’ may further increase sensitivity and precision of transcriptome sequences reconstruction (Cerveau, Jackson, 2016). Meta-assembly is then defined as a junction of all the *de novo* assemblies obtained with different tools after redundancy reduction. Redundancy reduction is a procedure of removal every contig that is a substring of at least one other contig in a given set. This approach was earlier tested for transcriptome assembly of non-model species using three assemblers – Trinity, Trans-ABySS and rnaSPAdes (Evangelistella et al., 2017). Furthermore, attempts were undertaken to obtain meta-assembly of transcriptome based on genome-guided assemblies (Venturini et al., 2018).

However, to the best of our knowledge, no attempts were made to evaluate performance of this approach on data obtained with IonTorrent sequencing platform. Meanwhile, IonTorrent platform, although being less popular than Illumina, is still in demand in biological researches, including studies of microbial metagenomes (Lee et al., 2019), interspecific diversity of earthworms (Shekhovtsov et al., 2019), transgenic rat lines (Bürckert et al., 2017), sequencing plant genomes (Salina et al., 2018). Furthermore, studies on comparison of Illumina and IonTorrent platforms have been performed that show IonTorrent reads having somewhat lower quality and precision that Illumina reads, and have greater discrepancy of read lengths (Lahens et al., 2017).

This research aims to create a computational pipeline based on transcriptome meta-assembly creation using *de novo* assemblers Trinity, Trans-ABySS and rnaSPAdes, as well as genome-guided version of Trinity based on reference genome. Computational pipeline was tested on transcriptome assembly of *Hordeum vulgare* L. barley isogenic line Bowman and nearly-isogenic lines iBw*Alm* with partial albinism of the spike and BLP with partial melanism of the spike. It was observed that quality of the transcriptome assemblies performed with different tools vary; however, in general they complement each other. Highest quality is observed for the transcriptome meta-assembly, which outstrips individual assemblies based on a number of metrics that characterize overall assembly quality.

Materials and methods

Short read libraries. Libraries of *H. vulgare* isogenic line Bowman and two nearly isogenic lines: iBw*Alm* (characterized by spike partial albinism) and BLP (characterized by partial melanism of the spike) transcriptome were used. The data was obtained from NCBI SRA database BioProjects PRJNA342150 (libraries of NIL i:Bw*Alm* and isogenic line Bowman) and PRJNA399215 (libraries of NIL BLP and isogenic line Bowman).

In PRJNA342150 experiment, transcriptomes of NIL i:Bw*Alm*, based on isogenic line Bowman, plants lemma and line Bowman, taken as a control, plants lemma were compared (Shmakov et al., 2016). For each of the lines three biological replicates were taken. Thus, in this experiment six short read libraries were sequenced. We will refer to this experiment as ‘alm experiment’ in further text.

In PRJNA399215, transcriptomes of NIL BLP, based on Bowman isogenic line, plants lemma and isogenic line Bowman, taken as a control group, plants lemma were compared (Glagoleva et al., 2017). We will refer to this experiment as ‘blp experiment’ in further text.

All libraries were obtained by sequencing using IonTorrent platform. The libraries then underwent filtration procedure, during which adapter sequences were removed using CutAdapt software version 1.9.1 (Martin, 2011), reads with mean quality score below 20 and lengths below 50 or above 270 were removed using Prinseq-lite software version 0.20.4 (Schmieder, Edwards, 2011). Table 1 lists metrics of the libraries used in this research.

Transcriptome reconstruction. In this work, three transcriptome assemblers were used: Trinity (Grabherr et al., 2013) version 2.2.0, Trans-ABySS (Robertson et al., 2010) version 2.0.1 and rnaSPAdes (Bushmanova et al., 2018) version 3.12.0. All three tools were listed among the best in performance and quality in a number of researches dedicated to comparison of transcriptome *de novo* assemblers (Honaas et al., 2016; Lafond-Lapalme et al., 2017; Fu et al., 2018; Hölzer, Marz, 2019).

Libraries from the two experiments were processed independently. Individual transcriptome assemblies obtained with each of the software tools were reconstructed as follows.

Trinity assembler was run with default parameters; all six libraries belonging to the respective experiment were given as input files. While running SPAdes assembler, likewise, all six libraries belonging to the respective experiment were given as input files. When launching SPAdes assembler, options ‘–iontorrent’ and ‘–only-assembler’ were specified.

Trans-ABySS assembly was conducted for each of the libraries separately, with resulting assemblies combined using transabyss-merge tool from Trans-ABySS software package. This assembly was performed with default parameters, with *k*-mer size equal to 32. In the same way, assemblies were conducted with *k*-mer sizes of 48 and 64. Thus, three *de novo* assemblies were obtained with Trans-ABySS, differing by *k*-mer lengths. Next, the three assemblies were combined with transabyss-merge. Resulting assembly was further used as an individual *de novo* transcriptome assembly obtained with Trans-ABySS software.

Additionally, genome-guided transcriptome assembly was performed using Trinity software. First, short read libraries were mapped to barley genome. Mapping files in the SAM (sequence alignment/mapping) format were then concatenated using merge tool from samtools software package version 1.6 into a single alignment file combining mapping of all six libraries belonging to respective experiment. This file, together with the six libraries from the respective experiment, were processed with Trinity tool in genome-guided transcriptome

Table 1. Metrics of the libraries implemented in the work

Experiment	Line	Library	Raw size	Clean size	Read mean length
PRJNA342150	i:Bw*Alm*	Alm_1	4 596 395	3 874 912	166.94
		Alm_2	3 056 413	2 372 255	199.52
		Alm_3	5 794 644	5 332 600	181.47
	Bowman	A_bow_1	4 122 599	2 450 068	175.49
		A_bow_2	4 023 501	2 356 572	126.56
		A_bow_3	6 887 599	6 523 266	201.68
PRJNA399215	BLP	Blp_1	3 583 148	1 311 442	185.39
		Blp_2	4 710 862	1 687 289	156.96
		Blp_3	4 070 591	1 864 073	146.02
	Bowman	B_bow_1	1 769 261	438 702	164.66
		B_bow_2	3 740 926	1 092 191	199.48
		B_bow_3	5 253 524	2 364 034	209.00

assembly mode, with a specified parameter of maximal intron length of 500 000 nucleotides.

In order to remove redundancy of assemblies, tr2aacds.pl tool from software package Evidential Gene (Gilbert, 2019) version 20.05.2020 was implemented. Each of the individual assemblies was processed with this software. Thus, three non-redundant transcriptome *de novo* assemblies and one non-redundant genome-guided transcriptome assembly were obtained. We will further refer to the *de novo* assemblies as short versions of respective software names: abyss, spades and trinity assemblies constructed using Trans-ABySS, rnaSPAdes and Trinity, respectively. We will further refer to genome-guided transcriptome assembly as GG (short of genome-guided).

In order to create an optimal meta-assembly of the transcriptome, individual assemblies were concatenated into one file, which was then processed with tr2aacds.pl tool for redundancy removal. It should be noted that only contigs containing open reading frames are considered, as tr2aacds.pl only uses contigs with predicted open reading frames with length above threshold value for further analysis. Figure 1 illustrates main stages of non-redundant meta-assembly construction.

Thus, for each of the two experiments, four individual assemblies were created: spades and trinity assemblies, consisting of six short libraries belonging to the respective experiment; abyss assembly performed for each of the libraries separately with three different *k*-mer length values, which were later combined into a single abyss transcriptome assembly using transabyss-merge script; genome-guided GG transcriptome assembly performed on six libraries belonging to the respective experiment and alignment file combined from six libraries alignments to the barley genome. Finally, four individual assemblies for each of the experiments were combined into the barley transcriptome meta-assembly.

Transcriptome assemblies quality assessment. In order to analyze qualities of assemblies, each one was processed with the following tools: BUSCO (Simão et al., 2015) version 3.0.2 for completeness assessment based on presence of characteristic sequences for plants; TransRate (Smith-Unna et al., 2016) version 1.0.3 for contigs annotation and completeness of known barley genes presence in the assembly. Then, comparison of CDS lists detected by TransRate in each individual assembly was performed. Based on overlapping of the lists of CDS detected in each assembly, Venn diagrams illustrating the part of each individual assembly in the structure of meta-assembly were drawn.

Next, contigs of two meta-assemblies of barley transcriptome belonging to two experiments were aligned to the *H. vulgare* genome using rnaQUAST software (Bushmanova et al., 2016). rnaQUAST counts several characteristics of assembly mapping to genome, and allows the user to evaluate the assembly's quality based on these characteristics. Specifically, this tool divides the contigs into three groups: contigs mapped to the reference and interlocking with known annotated genes; contigs mapped to the genome but lacking significant overlaps with the known annotated genes; and contigs with no homology to the known genome. We will further refer to this last group of contigs as 'new contigs'.

Transcriptome assemblies' quality comparison. In order to compare the assemblies' quality numerically, an approach suggested in the Hölzner and Marz publication (Hölzner, Marz, 2019) was implemented. This method is to normalize a selected number of parameters that reflect *de novo* transcriptome assembly quality using the following formula:

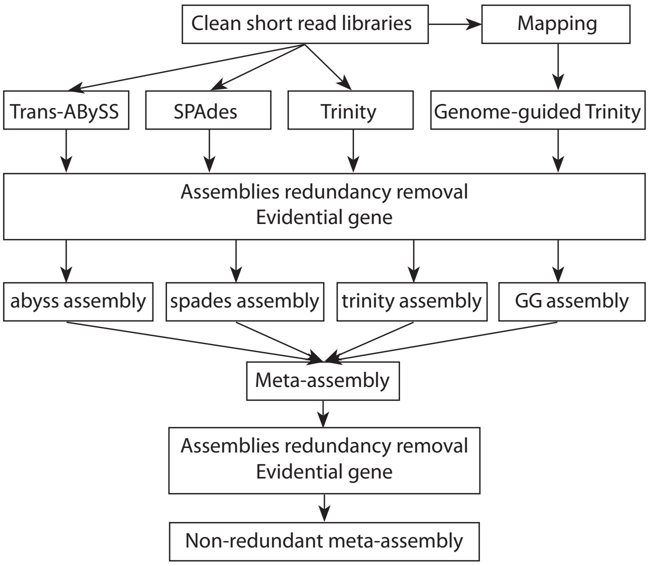

Fig. 1. Pipeline of individual *de novo* barley transcriptome assemblies and barley transcriptome meta-assembly acquisition.

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

where R_j^i is a value of parameter *i* for the transcriptome assembly *j* before normalization, N_j^i is this parameter's value after normalization, V^i is a vector of all values of the parameter *i* for all *k de novo* transcriptome assemblies before normalization: $V^i = (V_1^i, \ldots, V_k^i)$. Thus, after normalization each of the parameters takes a value from 0 to 1 for each of the *de novo* assemblies. Next, for each of the assemblies all the normalized parameters are summed, and assemblies are sorted based on the summed values of normalized parameters. The assembly with the highest value of summed normalized parameters is considered to have the highest quality.

To compare individual assemblies and meta-assemblies of barley transcriptome obtained while working with the short read libraries belonging to two experiments, seven parameters characterizing different aspects of transcriptome assemblies were used: (1) N50; (2) median of contig lengths distribution; (3) number of BUSCO genes detected in the assembly (both complete and fragmentary genes); (4) percentage of contigs with homology to known barley CDS detected using TransRate; (5) number of barley CDS that contigs from *de novo* assembly are homologous with; (6) amount of barley CDS with at least 95 % of the lengths covered with aligned contigs; (7) percentage of short reads from the library that was used in construction of the *de novo* assembly that were mapped back to the assembly using kallisto software.

Parameters 1 and 2 reflect distribution of contig lengths. Parameters 3, 4, 5 and 6 show completeness of the transcriptome assembly. Parameter 7 shows completeness of the transcriptome assembly and how fully were the libraries used in the process of assembly construction.

Results

alm experiment

For barley line i:Bw*Alm* and control isogenic line Bowman four *de novo* assemblies of lemma transcriptome, and one meta-assembly consisting of the four individual assemblies were obtained. Table 2 lists results of *de novo* transcriptome assembly of barley lines i:Bw*Alm* and Bowman, including common for the two lines meta-assembly.

Transcriptome meta-assembly of lines i:Bw*Alm* and Bowman obtained from *de novo* assemblies created with rnaSPAdes, Trans-ABySS and Trinity and genome-guided Trinity assemblies, consists of 169232 contigs before redundancy removal. Non-redundant meta-assembly consists of 68414 contigs with total length of 46440750 bases. Longest contig consists of 9920 nucleotides, mean contig length is 678.8 nucleotides, N50 is 936 nucleotides. Redundancy removal reduced meta-assembly size to 40.4 % of initial.

Coverage of contigs with short reads from the libraries was estimated for individual assemblies and meta-assembly of transcriptome using pseudo-alignment technique. It was observed that the highest percentage of reads was mapped to the abyss transcriptome assembly, while the lowest – to the spades assembly. 61.47 % of all the short reads were mapped to the meta-assembly of the transcriptome (see Table 2).

Search of known annotated barley CDS in transcriptome assemblies was carried out using TransRate software tool. Results of CDS identification for the assemblies are listed in the Table 3.

The highest amount of known CDS (29790) was detected in meta-assembly of transcriptome. Moreover, the highest amount of CDS with coverage no less than 95 % was detected in meta-assembly. However, the highest percentage of contigs that show homology to known barley CDS was detected for the spades assembly – 90.3 %. In meta-assembly this metric is only 62.7 %, which is lower than in any of individual assemblies.

Furthermore, in order to estimate contribution of each of the assemblers into the transcriptome meta-assembly structure, overlapping of CDS lists detected in individual assemblies was counted. Resulting overlaps are illustrated in Figure 2. As seen from Figure 2, 7191 barley CDS were detected in all four individual assemblies; 9305 CDS were detected in three out of four assemblies. 14615 CDS were detected in only single individual assembly, out of which the largest amount (5173) were detected only in trinity assembly, the lowest amount (2086) – only in spades assembly. The biggest intersection of CDS lists were observed between trinity assembly and GG assembly – 18258 CDS.

In contigs of each of the assemblies open reading frames (ORF) were predicted. ORF detected in the contigs of meta-assembly encode 58636 protein products with lengths equal to or greater than 30 amino acid residues. These protein products were used then to evaluate integrity of the assemblies using BUSCO software, which is shown in Figure 3. Transcriptome meta-assembly contains more complete BUSCO sequences than any individual transcriptome assembly, and less fragmented and absent BUSCO sequences. This suggests that meta-assembly has higher quality and integrity.

blp experiment

For RNA-seq libraries from blp experiment, individual transcriptome assemblies and transcriptome meta-assembly were obtained, and quality comparison of the assemblies was performed. Table 4 lists main parameters of the assemblies.

Resulting transcriptome meta-assembly of barley lines Bowman and BLP consists of 133070 contigs. After redundancy removal meta-assembly contains 32466 contigs with total length of 25184753 nucleotides. Thus, redundancy re-

Table 2. Characteristics of barley *de novo* transcriptome assemblies in alm experiment

Assembly	Assembly size, contigs		N50	Mean length	Reads mapped, %
	Redundant	Non-redundant			
abyss	705015	40806	1076	723.6	67.08
spades	22649	19181	1130	1072.65	39.13
trinity	267201	52005	976	741.19	64.97
GG	451309	57240	766	594.82	61.37
Meta-assembly	169232	68414	936	678.82	61.47

Table 3. Numbers of barley CDS detected in *de novo* transcriptome assemblies in alm experiment

Assembly	Contigs		CDS detected	p_95
	detected	%		
abyss	30530	0.748	22420	2542
spades	17323	0.903	14989	644
trinity	35547	0.684	27173	1779
GG	38686	0.676	26978	2240
Meta-assembly	42887	0.627	29790	3073

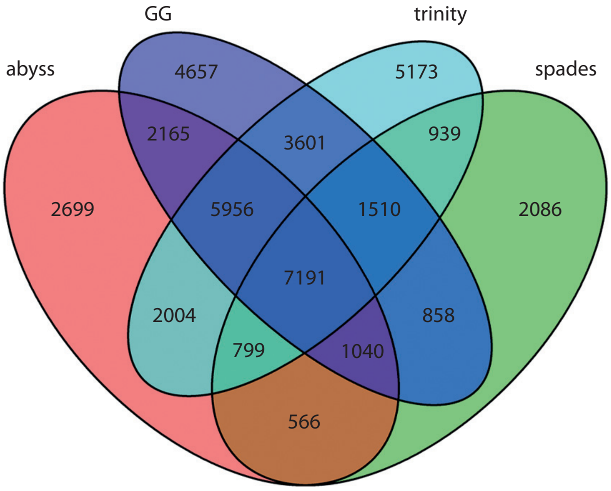

Fig. 2. Venn diagram illustrating overlaps of CDS lists detected in individual transcriptome assemblies in alm experiment.

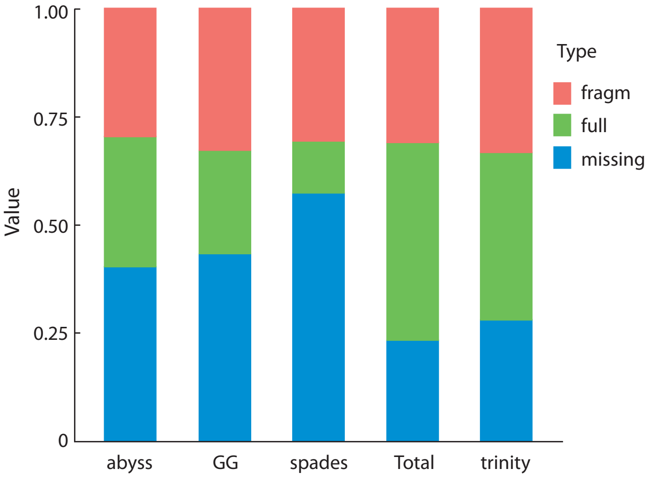

Fig. 3. BUSCO criterion of completeness of transcriptome assembly in alm experiment.

moval reduced assembly size to 24.4 % of initial size. Also, it is worth noting that meta-assembly in blp experiment has a higher N50 value than any of the individual assemblies it consists of. 72.1 % of short reads from blp experiment libraries were mapped back to the transcriptome meta-assembly. For this indicator, meta-assembly is behind GG assembly (77.6 %), but ahead of three other individual assemblies.

Search of known barley CDS was carried out in transcriptome *de novo* assemblies of barley lines under investigation using TransRate software. Results of the search are shown in Table 5. As can be seen from Table 5, from as low as 19 848 contigs in spades assembly to as much as 29 412 contigs in GG assembly show homology to known barley CDS. Meanwhile, the highest amount of barley CDS were detected in trinity assembly, however, the highest amount of barley CDS with no less than 95 % length covered with contigs is detected in transcriptome meta-assembly – 1825 CDS. Percentage of contigs from the assembly for which homology to known CDS was detected is 74.5 % in meta-assembly which is lower that in any individual assembly except for trinity assembly.

Search of overlaps between lists of CDS detected in individual assemblies was performed, and contribution of individual assemblies into meta-assembly structure was evaluated (Fig. 4). 9742 CDS were detected in all four individual transcriptome *de novo* assemblies. 8656 CDS were detected in only one of individual assemblies, of which the largest amount – 3554 were unique for abyss assembly, lowest

Table 4. Characteristics of barley *de novo* transcriptome assemblies in blp experiment

Assembly	Assembly size, contigs		N50	Mean length	Reads mapped, %
	Redundant	Non-redundant			
abyss	214 465	34 987	606	490.32	68.75
spades	31 453	24 401	1046	824.6	58.25
trinity	116 897	34 363	891	661.59	66.55
GG	122 304	39 319	976	707.83	77.55
Meta-assembly	133 070	32 466	1056	775.73	72.07

Table 5. Numbers of barley CDS detected in *de novo* transcriptome assemblies in blp experiment

Assembly	Contigs		CDS detected	p_95
	detected	%		
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
Meta-assembly	24 194	0.745	19 665	1825

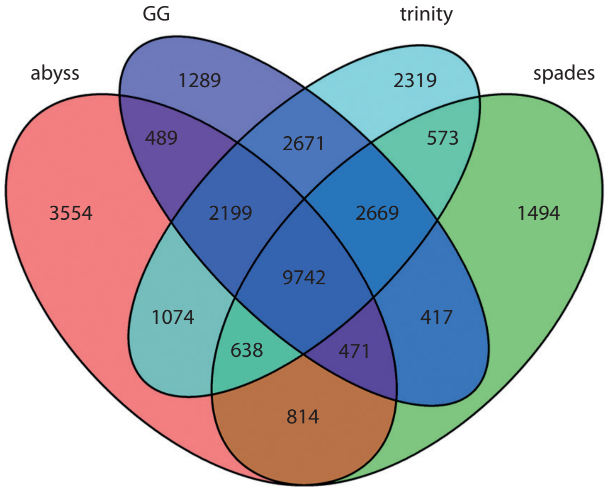

Fig. 4. Overlaps between CDS lists detected in individual transcriptome assemblies in blp experiment.

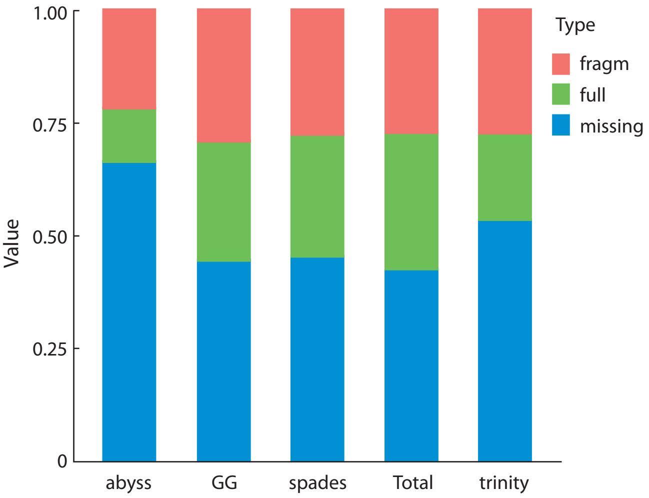

Fig. 5. BUSCO criterion of transcriptome assembly completeness in blp experiment.

amount – 1289 were unique for GG assembly. The highest amount of common CDS is between GG and trinity assemblies – 17281 CDS were detected in both of these assemblies.

Transcriptome assemblies' integrity estimation were carried out using BUSCO tool (Fig. 5). Meta-assembly was shown to have higher completeness than any of the individual assemblies, as it has the highest amount of complete BUSCO sequences detected and lowest amount of BUSCO sequences non-detected. In total 57.6 % of all BUSCO sequences from embryophyte set were detected in non-redundant meta-assembly as completely or partially.

Comparison of *de novo* assemblies' quality

Seven metrics of individual *de novo* assemblies and meta-assembly were evaluated in order to assess quality of the assemblies. These metrics indicate lengths of contigs in *de novo* assemblies (N50 and median of lengths distribution), presence of known barley CDS in the *de novo* assembly (percentage of contigs with homology to known barley CDS, amount of detected CDS and amount of CDS with at least 95 % of length covered) and genes characteristic to vascular plants (BUSCO completeness criterion), and fullness of libraries short reads implementation in the assembly creation (percentage of pseudo-aligned reads). Values of these metrics were normalized and brought into the range of values from 0 to 1 (Hölzner, Marz, 2019), then sums of normalized metrics were taken for each of the individual assemblies and for the meta-assembly. The largest values of the sums show the most optimal transcriptome assembly (Table 6).

As can be seen from the Table 6, highest values of normalized metrics are attributed to the transcriptome meta-assemblies in both experiments. This, together with highest amount of detected genes characteristic to vascular plants detected with BUSCO software, and highest amounts of fully reconstructed barley CDS indicates that meta-assemblies created by combining of individual *de novo* transcriptome assemblies and redundancy removal outstrip individual assemblies by quality.

Table 6. Summarized values of normalized quality metrics for *de novo* transcriptome assemblies in experiments alm and blp

Assembly	alm experiment (lines i:Bw*Alm* and Bowman)	blp experiment (lines BLP and Bowman)
abyss	4.16	1.72
spades	3.00	3.86
trinity	4.07	3.61
GG	2.85	5.22
Meta-assembly	4.32	5.56

Discussion

In this work, an approach to *de novo* transcriptome reconstruction based on creation of meta-assembly from several individual assemblies was tested. It was observed that transcriptome meta-assemblies have higher integrity judging by a number of criteria such as amount of detected BUSCO fragments, amount of barley CDS to which contigs in transcriptome assembly show homology, and percentage of pseudo-aligned to the assembly reads from RNA-seq libraries. Thus, it could be concluded that aforementioned approach to transcriptome *de novo* reconstruction based on creation of several individual assemblies followed by their combining into meta-assembly increases quality of *de novo* reconstructed transcriptome.

Comparison of several aligners showed that rnaSPAdes tool reconstructs fewer contigs, while Trans-ABySS reconstructs the highest amount of contigs. Trinity assembler reconstructs comparable quantities of contigs when run in two modes – *de novo* and genome-guided. At the same time, redundancy removal reduces sizes of Trans-ABySS assemblies most severely – in alm experiment 94.3 % of all contigs reconstructed by Trans-ABySS were removed, in blp experiment – 83.7 %. In the case of spades assembly, 15.3 and 22.4 % of all the contigs were removed, respectively. In trinity assemblies on average 80.5 and 70.6 % of contigs were removed, in genome-guided assemblies – 87.3 and 67.8 % of contigs, respectively. Genome-guided assemblies have the highest sizes after redundancy removal in both experiments, spades assemblies – the lowest.

Spades reconstructs the largest contigs of all individual assemblers, which is indicated by highest N50 values and medians of contig lengths distribution. Lowest N50 value in alm experiment was observed in GG assembly, in blp experiment – in abyss assembly.

The highest completeness of all individual assemblies according to BUSCO criterion in alm experiment is attributed to trinity assembly. In blp experiment it is attributed to GG assembly. The lowest completeness according to BUSCO criterion is attributed to spades assembly in alm experiment and abyss assembly in blp experiment.

Conclusion

To conclude, in the two experiments difference in performance of the *de novo* transcriptome assemblers is observed, despite IonTorrent short read libraries being used in both experiments, and reconstructed transcriptome belonging to the same organism – *H. vulgare* barley. This suggests that implemented assemblers are sensitive to the input data, and their performance can vary depending on the data used.

However, on both accounts transcriptome meta-assemblies created from combined individual assemblies have higher quality than all individual assemblies, which indicates the effectiveness of the approach to *de novo* transcriptome reconstruction as building of meta-assemblies combining results of several individual *de novo* transcriptome assemblers.

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., Prjibelski 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., Rätsch 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., Scalabrin S., Cattonaro F., Morgante M., Mugnozza G.S., Keurentjes 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.Trinity.

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/bfgp/elr035.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal.* 2011;17(1):10-12. PMid: 1000006697. 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., Šafár 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.

Acknowledgements. The work was supported by Russian Science Foundation project No. 18-14-00293 (problem statement, algorithm creation, data analysis). Computational resources of Core Facility ‘Bioinformatics’ supported by budget project No. 0259-2021-0009 were implemented in this work.

Conflict of interest. The author declares no conflict of interest.

Received November 24, 2020. Revised January 15, 2021. Accepted January 15, 2021.