Impact of sequencing data filtering on the quality of de novo transcriptome assembly

*Yakov Meger*¹, *Ekaterina Vodiasova*², and *Anastasiya Lantushenko*¹

¹Sevastopol State University, Sevastopol, Russia

²IBSS RAS, Sevastopol, Russia

Abstract. There are many assemblers with different algorithms that are used for de novo transcriptome assembly. At the same time, the filtering stage, which is one of the key stages, also has several approaches and algorithms. However, to date, there are only few studies on the effect of the degree of filtration on the de novo transcriptome assembly, specially for single-end reads. In this paper, we analyzed transcriptomes obtained using two of the most common software (rnaSPADES and Trinity), and also applied various approaches to the stage of filtering reads. The key differences between the two assemblies were shown and the parameters that were sensitive to the degree of filtering and the length of the input reads were identified. An efficient two-stage filtering algorithm was also proposed, which allows one to preserve the volume of input data as much as possible with the required quality of all reads after filtering and trimming.

1 Introduction

Currently, for a deeper understanding of the physiology of processes occurring in organisms under various conditions, it is necessary to study the expression of genes or entire complexes of genes, as one of the aspects of the body's response to a stimulus. The number of studies based on the sequencing of the complete transcriptome of an entire organism, individual tissues or cells gradually increases due to development and reduction in the cost of RNA-Seq technologies [1]. At the same time, bioinformatic methods are also being developed to improve transcriptome assemblies, analyze chimeras and isoforms, annotate contigs, etc.[2], [3], [4]. In the case of working with an unmodeled or insufficiently studied object, there is a need for de novo assembly without using a reference genome. Since most metrics for evaluating the quality of the resulting assemblies are relative (comparative), evaluation of the quality of the resulting transcriptome without a qualitative reference is a difficult task. This situation is complicated by the fact that the choice of an effective and optimal algorithm for pre- and post-processing of the obtained readings is still a subject of discussion. The common pipeline for de novo assembly consists from several steps: quality control, filtering and trimming, assembling (with or without reference genome), quality assessment of assemblies, annotation. The algorithms and their dependence on the input data are different for different programs for transcriptome assembly [5].

The first stage of de novo transcriptome assembly necessarily includes an assessment of the quality of the obtained reads after sequencing and their further filtering. The approaches at this stage differ: either they try to clear the transcripts obtained from low-quality and short reads as much as possible, or they try to preserve all of the sequenced sequences, or they try to find a compromise between reliability and completeness. This stage is of key importance, as the quality of the collected transcriptome directly depends on the input data (the principle of 'garbage in, garbage out'). One of the factors that can affect the quality of de novo transcriptome assembly is the length of reads after filtering, since the length of reads is crucial for choosing the length of k-mers [6]. There is a very little research on this issue [7].

Thus, the aim of this work was to study the effect of different read lengths on the quality and completeness of de novo transcriptome assemblies using different assemblers.

2 Materials and methods

In this work, the assembly was carried out on single-end reads with a length of 150 bp, obtained by NGS-sequencing of total RNA from the gill tissues of *Mytilus galloprovincialis* using the Illumina Hi-Seq platform. The quality of the reads was evaluated using the fastqc program [8]. At the end of all the reads, the quality dropped and there was a high level of duplications, which is associated with the presence of rRNA in the samples. In this case, the rRNA was not removed from the array, and the assembly was performed for the whole data set.

The lengths of the readings were in the range of 35 to 151, more than 90% of which had a length of more than 149. In the remaining range the vast majority fall in the region of 90-140 bp (Figure 1).

Fig. 1. Distribution of read lengths in unfiltered form.

The processing of the readings was carried out using the fastp v. 0. 20. 0 program [9] by various methods:

1) Coarse filtering. Includes two stages of filtering, discarding reads shorter than 100 bp and trimming the beginning and end of the reads with low quality.

2) Medium filtering. Includes two stages of filtering, filtering of discarded reads with a length of less than 85 bp, trimming the beginning and end.

3) Soft filtering. Includes two stages of filtering, filtering of discarded reads with a length of less than 65 bp, trimming the beginning and end.

4) Minimum filtering. Includes filtering in one stage with discarding reads shorter than 50 bp, trimming the beginning and end of reads and reads with low quality (GC % <20).

Filtering was accomplished in two stages according to the following protocol.

Step 1:

Trimming the beginning and end of reads, filtering by quality (phred >20)

```
> -q 20 -b 140 -f 10 -n 35 -y-l 80 --failed_out
```

However, after this procedure most of the reads had been cut too much. Therefore, the readings that had not been filtered, were further analyzed separately. Two clusters differed by length were obtained, viz. 70 and 140 nucleotides. Each cluster contained different errors, so further filtering was accomplished separately for each length range (Stage 2):

```
> -Q -A -l 60 --length_limit 80 -b 70 -f 3 -t 2
```

```
> -Q -A -l 130 -3 -W 3 -M 30 -b 65 -t 5
```

After that, all three files were concatenated and three variants of filtered reads were made according to the minimum length: 65, 85 and 100 nucleotides.

All fastq files of seven transcriptome libraries (according to the minimum read length) were combined into a single file, and then the quality was evaluated again using fastqc.

De novo assembly was performed by three algorithms implemented in the transcriptome assemblers Trinity v.2.1.1. [10] and rnaSPADES v3. 13. 0 [11]. The quality of the assembled transcriptomes was evaluated by the QUAST program v.4.6.3 [12]. The estimation of the percentage of mapping reads per assembly was performed in the Bowtie2 program v. 2.3.4.3 [13]. Protein coding regions were predicted based on the analysis of open reading frames using the Transdecoder program v. 5.5.0 included in the Trinity package. The search for orthologous genes was performed on the basis of the mollusca_odb10 database (https://busco.ezlab.org/list_of_lineages) in the program BUSCO v. 4.0.5 [14]. Clustering of the collected contigs was performed using Usearch v.11.0.667 [15], the identity percentage was chosen as 0.95, and the centroids were chosen by the longest length.

All steps of the de novo assembling and different approaches to the filtering and trimming of raw reads are represented and summarized in the Figure 2.

Fig. 2. The pipeline of the filtering and trimming different approaches used in this study.

3 Results

63.6 million reads were produced by 150 nt single end sequencing from the seven libraries, which have been deposited to the NCBI (accession numbers: SRR13013753, SRR13013754, SRR13013755, SRR13013756, SRR13013757, SRR13013758, SRR13013759). The length of reads before filtering ranged from 35 to 151 bp (more than 90% had a length of more than 149 bp). Application of the coarsest cleaning resulted in exclusion of 11.4% of reads from the dataset (Figure 3).

The characteristics of the resulting assemblies obtained with the help of different assemblers are quite different (Table 1). The length of the largest contigs in assemblies obtained using Trinity was 28175 bp for all degrees of filtration, whereas from rnaSPADES it was 17500±1000 bp. The number of contigs in assemblies obtained using Trinity was ~250 thousands on average, rnaSPADES – 170 thousands. The GC composition of all assemblies was 34.8±0.2 %.

Fig. 3. Reads passed filtering.

Table 1. Characteristics of the obtained QUAST builds.

	Trinity v.2.1.1				
	35-151 bp	50-140 bp	65-140 bp	85-140 bp	100-140 bp
# contigs	272 971	246 069	259 251	256 846	248 355
NGA50	1 790	1 257	1 798	1 799	1 800
Largest contig	28 175	28 171	28 175	28 175	28 175
Total length	209819645	187382525	185382525	187532525	185332525
	rnaSpades v.3.13.0				
	174 049	173 267	173 747	173 517	172 288
# contigs	174 049	173 267	173 747	173 517	172 288
NGA50	1 442	1 436	1 439	1 438	1 438
Largest contig	17 202	17 193	18 672	16 815	16 815
Total length	125592097	114890129	118667182	118488643	117604550

Despite the noted differences between the assemblies based on the standard characteristics, principal differences were also identified in other parameters.

The percentage of reads mapping to assemblies for all variants was $96\pm2\%$, but the number of unique alignments was differed, i.e. for the assemblies obtained using Trinity it was $16.6\pm1\%$, for the assemblies obtained using rnaSPADES it was $40\pm4\%$.

Fig. 4. Alignment of reads to the resulting assembly, performed using Bowtie2.

The search for conservative genes in the mollusca_odb10 database for the assemblies obtained using Trinity allowed to find 4655 ± 30 genes out of 5295 existing in the database, whereas the number of genes found using rnaSPADES was 4452 ± 40 , which is comparable with each other (Figure 5). If we consider the number of unique genes, then again there is a strong difference between the Trinity and rnaSPADES results (~1850 and ~3440 , respectively).

Fig. 5. Results of the search for orthologous genes (Mollusca_odb10) using BUSCO.

The greatest difference between the assemblers was demonstrated by the analysis on the search for open reading frames (ORFs). For assemblies obtained using Trinity, ~110 thousands protein coding transcripts were found, whereas from rnaSPADES we found 78 thousands (Table 2). Contigs with 7 reading frames were also found. By default, the minimum ORF length was set to 100 aa (300 bp). According to the Trinity-derived

assemblies, the number of predicted proteins is much larger, while the distribution of ORF lengths shows greater frequency of short mRNAs. Thus, the number of ORF with a length of 300 n. k. for the Trinity-derived assemblies is twice as big as for the SPADES-derived assemblies.

Table 2. Characteristics of open-frame transcripts.

	Trinity v.2.1.1				
	35-151 bp	50-140 bp	65-140 bp	85-140 bp	100-140 bp
Peptides predicted	132 860	120 327	103 075	102 228	99 150
Proportion of protein-coding transcripts, %	40,07	40,05	39,76	39,80	39,92
Amount of contigs with ≥7 ORF	25	24	0	0	0
Transcripts with c 5'UTR and 3'UTR	38 079	34 362	32 273	32 784	31 860
	rnaSpades v. 3.13.0				
	35-151 bp	50-140 bp	65-140 bp	85-140 bp	100-140 bp
Peptides predicted	86 710	78 481	81 460	81 449	80 636
Proportion of protein-coding transcripts, %	56,02	51,53	46,84	46,94	46,80
Amount of contigs with ≥7 ORF	12	10	15	17	16
Transcripts with c 5'UTR and 3'UTR	21 449	19 078	19 972	19 978	19 749

The number of contigs after clustering for assemblies using Trinity was 78.1±0.7% relative to the initial number, and 94.4±2.5% using rnaSPADES. After clustering, the number of left contigs had become more uniform across all assemblies, ranging from 168330 to 197879 (in the assemblies the the range was from 172288 to 272971). Clustering by the Trinity- and SPADES-derived assemblers was very different (Figure 6). The distribution of centroid lengths is again shifted towards shorter ones in Trinity, with the maximum obtained length being also characteristic for this assembler, which resulted in almost similar average length. There are very large clusters of up to 140 transcripts appear for assembly by Trinity, but the number of such clusters remains small.

Fig. 6. Cluster size distribution for Trinity (left) and SPADES (right). The minimum reading length in the fasta file is 100 aa. The frequency scale is logarithmic.

4 Discussion

The difference between the various assemblers

The analysis revealed a certain pattern in the Trinity- and rnaSPADES-derived assemblies. The longest contig in the Trinity build is 1.6 times longer than that in the SPADES build. The same ratio is observed for the following parameters: the number of contigs, L50, the number of predicted proteins using TransDecoder.

The number of contigs obtained as a result of the Trinity and rnaSPADES builds differs significantly (by 35%). Nevertheless, this indicator should be considered with caution, because such a number of contigs may be caused by the presence of a large quantity of chimeric contigs and non-existing isoforms resulting from the assembly procedure, and may therefore not indicate quality. Probably, the Trinity build is more complete, but at the same time is more 'dirty'.

This assumption is supported by a small percentage of read alignment per build and a huge number of short contigs. Moreover, the percentage of transcripts with open reading frames also differs significantly. This indicates that a large number of contigs obtained during the assembly in Trinity are either chimeras or too short (less than 300 bp).

In addition, a larger number of predicted proteins (for the Trinity-derived assembly) have a small length, but the average length is almost the same, due to the presence of longer

contigs in the Trinity assemblies compared to the SPADeS assemblies (28,175 bp and 17,000 bp, respectively).

However, taking into account the analysis of BUSCO and TransDecoder, the Trinity-derived assembly should not be discarded, as this may lead to the loss of information about some transcripts and genes.

Difference in the degree of filtering of input reads

Analysis of the effect of filtering revealed some features of the resulting assemblies. For both assemblers it was found that filtering in two stages is more efficient. After the first exclusion from the dataset of unsatisfactory reads and rough trimming at the ends, re-analyzing the 'bad' reads we return some of them to the dataset for further assembly. If we compare the number of reads left, it turns out that for a narrower range of lengths (from 65 to 140 bp) the number of reads is 3% greater than for the lengths from 50 to 140 bp. In addition, for the Trinity-derived assembly the number of contigs collected increases with two-stage filtering, and, most importantly, for both assemblers the number of contigs with a longer length increase. The latter is especially noticeable when building using rnaSPADeS. This is probably due to the return on the second stage of filtering to the array of reads, which are the edge sections of the transcripts and therefore may be shorter than the main part of the reads. At the same time, they can carry important information and, when searching for ORF, play a large role in determining the 5' and 3' UTR regions. This assumption is supported by the results of TransDecoder for rnaSPADeS on an increase in the number of predicted proteins, while the relative number of contigs with ORFs decreased. It should be noted that the number of predicted proteins in the Trinity-derived assembly decreases with decreasing length. This difference between the assemblers is a reflection of the different algorithms implemented in them and the possible presence of more chimeras and non-existing isoforms in the Trinity-derived assembly, as already discussed above.

The evaluation of the alignment of reads per assembly revealed a dependence on filtering only for the rnaSPADeS assembly. The highest percentage of unique alignments was found again for the two-stage cleaning of readings, but with the maximum preserved length (range 60-140 bp).

The search for orthologous genes did not show an explicit dependence on filtering for both assemblers.

5 Conclusion

It was found that the de novo assembly of transcriptomes performed using rnaSPADeS is more affected by the degree of filtering and is less complete (but possibly cleaner) than the de novo assembly using Trinity. On the other hand, the assemblies obtained using Trinity have a large number of chimeric contigs and isoforms, which can lead to difficulties at the stage of evaluating the differential expression of transcripts and to incorrect analysis. The study showed that in order to obtain a reference transcriptome we recommend to combine the two assemblies (minimum filtering in two stages using the rnaSPADeS assembler, and maximum filtering using the Trinity assembler), since utilization of only one program for de novo transcriptome assembly can lead to distorted or incomplete data. In addition, the read filtering should be performed twice, aiming to cut off really low-quality reads, but preserving the entire dataset as much as possible.

The work was carried out within the framework of the State Assignment No. AAAA-A18-118020890074-2, with the support of the Ministry of Education and Science of the Russian Federation grant No. 14.W03.31.0015 and the internal grant of SevSU 2020 No. 33/06-31.

References

1. G.K. Marinov, Briefings in functional genomics, **16**, 326 (2017)
2. L. Liu, B. Song, J. Ma, Y. Song, S. Y. Zhang, Y. Tang, X. Wu, Z. Wei, K. Chen, J. Su, R. Rong, Z. Lu, J. P. de Magalhães, D. J. Rigden, L. Zhang, S. W. Zhang, Y. Huang, X. Lei, H. Liu, J. Meng, Computational and structural biotechnology journal, **18**, 1587 (2020)
3. M. Fu, H. Su, Z. Su, Z. Yin, J. Jin, L. Wang, Q. Zhang, X. Xu, Microbial pathogenesis, **34**, 104 (2020)
4. M.T. Seweryn, M. Pietrzak, Q. Computational and structural biotechnology journal, **18**, 1830 (2020)
5. J. Tamames, M. Cobo-Simón, F. Puente-Sánchez, BMC genomics, **20**, 960. (2019)
6. M. Hölzer, M. Manja, GigaScience. **8**, 247 (2019)
7. P. Longone. Physical review. E., **100**, 52 (2019)
8. S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data (2010)
9. S. Chen, Bioinformatics, **34**, 884, 2018
10. M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Nat Biotechnol. **29**, 644 (2011)
11. E. Bushmanova, D. Antipov, A. Lapidus, A.D. Prjibelski, GigaScience. **8**, 103 (2019)
12. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, Bioinformatics, **29(8)**, 1072 (2013)
13. B. Langmead, C. Wilks, V. Antonescu, R. Charles, Bioinformatics, **35**, 421 (2019)
14. M. Seppey, M. Manni, E.M. Zdobnov, Methods in Molecular Biology, **6**, 19 (2019)
15. R.C. Edgar, Bioinformatics, **26**, 2460 (2010)