Prjibelski *et al. BMC Bioinformatics* 2020, **21**(Suppl 12):302
https://doi.org/10.1186/s12859-020-03614-2

BMC Bioinformatics

SOFTWARE Open Access

Extending rnaSPAdes functionality for hybrid transcriptome assembly

Andrey D. Prjibelski[1*], Giuseppe D. Puglia[2], Dmitry Antipov[1], Elena Bushmanova[1], Daniela Giordano[3], Alla Mikheenko[1], Domenico Vitale[2] and Alla Lapidus[1]

From Bioinformatics: from Algorithms to Applications 2019 (BiATA 2019) conference
Saint Petersburg, Russia. 20-22 June 2019

*Correspondence:
a.przhibelsky@spbu.ru
[1]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia
Full list of author information is available at the end of the article

Abstract

Background: De novo RNA-Seqassembly is a powerful method for analysing transcriptomes when the reference genome is not available or poorly annotated. However, due to the short length of Illumina reads it is usually impossible to reconstruct complete sequences of complex genes and alternative isoforms. Recently emerged possibility to generate long RNA reads, such as PacBio and Oxford Nanopores, may dramatically improve the assembly quality, and thus the consecutive analysis. While reference-based tools for analysing long RNA reads were recently developed, there is no established pipeline for de novo assembly of such data.

Results: In this work we present a novel method that allows to perform high-quality de novo transcriptome assemblies by combining accuracy and reliability of short reads with exon structure information carried out from long error-prone reads. The algorithm is designed by incorporating existing hybridSPAdes approach into rnaSPAdes pipeline and adapting it for transcriptomic data.

Conclusion: To evaluate the benefit of using long RNA reads we selected several datasets containing both Illumina and Iso-seq or Oxford Nanopore Technologies (ONT) reads. Using an existing quality assessment software, we show that hybrid assemblies performed with rnaSPAdes contain more full-length genes and alternative isoforms comparing to the case when only short-read data is used.

Keywords: Transcriptome assembly, Hybrid assembly, De novo assembly, Transcriptomics, RNA-Seq, Iso-seq, Oxford nanopores

Background

While a significant fraction of transcriptome studies are based on reference-assisted RNA-Seq analysis, they are limited to species with high-quality reference genome sequences and gene annotation. De novo assembly of RNA-Seq data allows to analyse transcripts sequences and gene expression without a reference genome, thus presenting a viable alternative to reference-based approaches. However, due to the short length of Illumina reads, recovery of complete transcript sequences originated from complex isoforms appears to be impossible without additional information, e.g. a reference genome. Recent biotechnological advances allowed to apply long-read technologies to transcriptome sequencing [1, 2]. While long RNA reads seem to be extremely promising for transcriptomic studies, there is clearly a lack of software developed for their analysis.

In the past few years, several research projects involving long-read RNA sequencing were carried out [3–7]. All of them, however, were performed for species with relatively well-sequenced genomes. In these studies, researchers used such reference-based tools as Iso-seq pipeline [2], or designed in-house pipelines based on previously developed spliced aligners and tools for short-reads data analysis. However, no studies involving de novo assembly of long RNA reads were reported.

Most of the existing de novo transcriptome assemblers do not support long error-prone reads, since they were designed specifically for short Illumina reads. Among these tools only Trinity [8] supports hybrid assembly using corrected long reads. The only tool specially designed for hybrid transcriptome assembly is IDP-denovo [9], which is capable of improving third-party assemblies using long uncorrected reads. In addition, according to its manual, recently developed RNA-Bloom assembler [10] is capable of performing assembly solely from RNA ONT reads.

In this work we propose an extension for rnaSPAdes de novo transcriptome assembler [11]. Combining rnaSPAdes with previously developed hybridSPAdes approach [12] allows to exploit Iso-seq and ONT RNA reads as additional input and perform hybrid assembly. Since long-read technologies have a beneficial feature of detecting full-length (FL) mRNA sequences using terminal adapters in raw reads, a new version of rnaSPAdes can additionally take FL reads as an input, which further helps to determine complete isoform sequences.

To benchmark the assembly software, we selected several datasets containing both short and long reads. Although a variety of publicly available long-read RNA sequencing data is relatively small compared to conventional RNA-Seq, for this publication we selected three human datasets. The human transcriptome contains complex alternative isoforms, which allows to show the impact of using long reads for the assembly. Additionally, it eases the assembly quality evaluation, since the human genome is comparably well annotated and the ground truth is known. To assess generated assemblies we used rnaQUAST [13], which allows to evaluate their correctness and completeness using reference genome and gene database. Performed benchmarks show that incorporating long reads into the assembly pipeline allows to accurately assemble more complete genes and isoforms.

Implementation

SPAdes assembly pipeline [14] consists of the four major steps: (i) de Bruijn graph construction from short reads, (ii) graph simplification, which removes erroneous edges from the graph and produces a so-called assembly graph, (ii) alignment of paired reads to the assembly graph and (ii) repeat resolution and scaffolding in the exSPAnder module [15, 16].

HybridSPAdes [12] additionally includes mapping long error-prone reads using BWA MEM algorithm [17] and exploiting these alignments during repeat resolution stage.

Since hybridSPAdes is designed for genomic data, it heavily relies on unique (non-repetitive) edges in the assembly graph, which are selected using coverage and length criteria. An edge is considered to be unique if it has coverage close to the average coverage of the dataset and its length exceeds a certain threshold [12]. Indeed, such heuristics is not applicable for transcriptomics data, where the majority of edges are short and coverage is non-uniform.

In rnaSPAdes, graph simplification is modified specifically for RNA-Seq data and the repeat resolution step is substituted with an isoform reconstruction procedure [11]. However, the current version of rnaSPAdes is capable of using only short paired-end and single reads. To extend its functionality for hybrid transcriptome assembly, we combine it with procedures implemented in hybridSPAdes (see Fig. 1). While the read mapping step for transcriptomic data remains unmodified (with the exception of some alignment parameters), alterations were introduced to the isoform reconstruction procedure.

Similarly to genomic SPAdes, in rnaSPAdes isoform reconstruction is based on the concept of path extension implemented in the exSPAnder module. During path prolongation exSPAnder uses all available information simultaneously. In case of hybrid assembly, at every step exSPAnder tries to find correct extension edge using paired-end reads first, and then applies long-read path extension only if paired-end reads do not help (see [12, 15] for details).

Since alternatively spliced isoforms may form very similar paths, e.g. differing only by a single alternative exon, the key modification introduced to the path-extension procedure compared to the genomic pipeline is the possibility to select more than a single extension edge at each step. The same idea can be used for exploiting long-read alignments during the isoform reconstruction stage.

To extend a path $P = (p_1, \ldots, p_n)$ the algorithm considers all long-reads paths matching with P. A path R obtained from a long read alignment is defined as matching with P if there

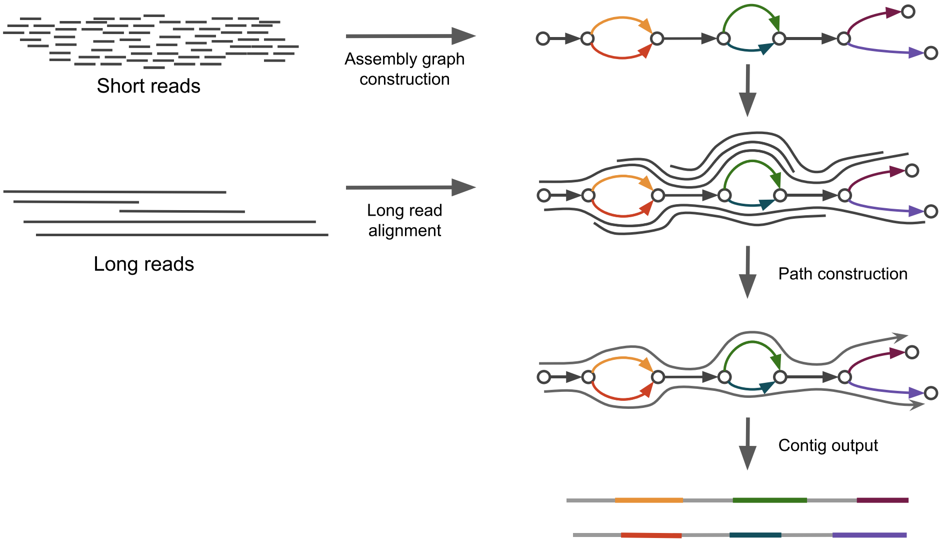

Fig. 1 Hybrid rnaSPAdes pipeline. The assembly graph is constructed and simplified using only accurate short reads. Long error-prone reads are aligned to the graph. The resulting alignments are used in isoform reconstruction procedure. Grey edges represent common exons, colored ones correspond to alternative exons

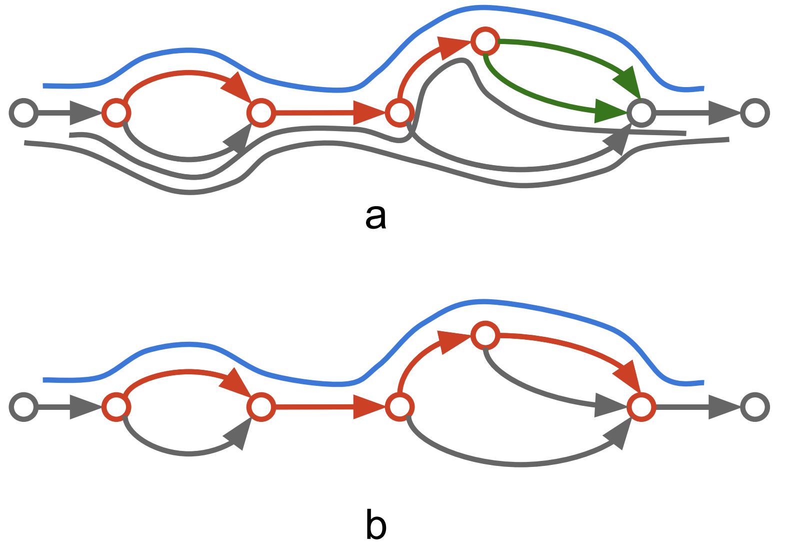

Fig. 2 Example of path extension procedure using long reads. Red edges represent a current path being extended, green — possible extension edges, lines along the graph — long reads aligned to the graph. **a** Among all aligned reads the algorithm select only those that comply with the current path (blue line). **b** Complying paths are being used to extend the current path

exists a suffix of P that is a prefix of R, or P is contained inside R (Fig. 2a). Formally, either (i) $R = (p_i, \ldots, p_n, x_1, \ldots, x_k), i >= 1$ or (ii) $R = (r_1, \ldots, r_l, p_1, \ldots, p_n, x_1, \ldots, x_k)$, where $r_1, \ldots, r_l$ and $x_1, \ldots, x_k$ are arbitrary edges in the graph. Further, from a set of all matching long-read paths the algorithm selects only those, for which the longest common subpath with P is (i) at least L_{min} long and (ii) contains at least N_{min} edges (default parameters are $L_{min} = 200\ bp$ and $N_{min} = 2$). The final set of matching long-read paths is denoted as $\mathbf{R_P}$. Then, among the set of all possible extension edges $\{e1, \ldots, e_m\}$, the algorithm selects all e_i, such that at least one path from $\mathbf{R_P}$ matches $(p_1, \ldots, p_n, e_i)$ (Fig. 2b). Using only paths from $\mathbf{R_P}$ instead of all matching long-read paths prevents from selecting all possible extensions for path P.

Paths in the graph are iteratively extended using paired-end and long reads until every edge is included in at least one path. Finally, to exploit reads capturing full-length transcripts, rnaSPAdes aligns them to the graph and produces FL paths, which are directly added to the set of resulting paths. Identical paths and exact subpaths are removed to avoid duplications, and the resulting set of paths is outputted in FASTA format.

Results and discussion

In this manuscript, we present quality reports for the assemblies of three datasets containing short and long-read sequencing data. Details on the used data are provided in Table 1. All datasets were quality-checked with FastQC [18] and trimmed with Trimmomatic [19] when adapters or significant quality drop were detected.

We ran rnaSPAdes and Trinity with default parameters on Illumina data alone and on combined datasets. Since Trinity supports only corrected long reads, they were corrected using Illumina reads prior to the assembly with Racon [20]. Unfortunately, on hybrid data Trinity pipeline ran for over 4 weeks and did not produce the assembly. To perform hybrid assembly with IDP-denovo, we provided rnaSPAdes contigs and long raw reads

Table 1 Human transcriptome datasets selected for benchmarking

Dataset	Technology	# of reads	Strand-specific	Accession No
Human siNT_48	Illumina	117 M	RF	SRP126849
	Iso-seq	5.6 M	—	SRP126849
Human MCF7	Illumina	84 M	No	SRX426377
	Iso-seq	1.8 M	—	—
	Iso-seq FL	0.5 M	—	—
Human GM12878	Illumina	54 M	No	SRR3103887
	Nanopores	10 M	—	—

as an input. However, either the resulting assembly remained unchanged, or IDP-denovo crashed. To compare hybrid transcriptome assembly against long-read-only assembly, we launched the only available tool RNA-Bloom on long reads from all 3 datasets (although it was designed only for ONT reads). Unfortunately, in all 3 cases RNA-Bloom managed to perform only read error correction step, but never completed the assembly itself. Thus, we present only results for Trinity and rnaSPAdes on short reads, and rnaSPAdes on hybrid data. Additionally, to evaluate the effect of external read error correction we ran rnaSPAdes on the same hybrid datasets, but using long reads corrected by Racon and RNA-Bloom. For a fair comparison, the same minimum contig length threshold was used for all assemblies (200 bp).

To evaluate the assembly quality we used rnaQUAST [13], which was designed specifically for assessing de novo transcriptome assemblies of organisms with high-quality reference genome and gene annotation. Among the large variety of metrics reported by rnaQUAST, we have selected only those that represent the most important characteristics of assembled sequences. In our opinion, one of the most significant statistics is the number of X%-assembled genes/isoforms, i.e. that have at least X% of bases captured by a single assembled contig. This metric shows the ability of de novo assembler to reconstruct complete transcript sequences, i.e. that can be used for further analysis. Presented reports also include database coverage (percentage of reference transcriptome nucleotides covered by all contigs), duplication ratio (the total number of aligned bases in all contigs divided by the total number of covered bases in reference isoforms, 1.0 in an ideal case), number of misassemblies (e.g. chimeric contigs) and average number of mismatch errors per contig. More details on various metrics and methods for assessing quality of de novo transcriptome assembly can be found in [13, 21, 22]. All datasets were assessed using Ensembl *H.sapiens* GRCh38.82 reference genome and gene annotation.

Tables 2, 3 and 4 demonstrate short quality reports for 3 human datasets. Since rnaSPAdes typically generates assemblies with higher quality than Trinity, rnaSPAdes is taken as a baseline in the following comparisons. As quality reports indicate, using raw long reads approach allows to reconstruct 8.4% more 95%-assembled genes on average comparing to short-read assembly. More importantly, the increase in 95%-assembled isoforms is larger (13.5% on average), which emphasizes that long reads are not only capable of reconstructing complex gene sequences, but are also useful for detecting alternatively spliced isoforms. One may also notice that the highest increase in 95%-assembled genes and isoforms (16.9% and 26.9% respectively) is achieved on Human MCF7 dataset, which contains FL reads. This fact suggests that FL reads can, unsurprisingly, be very beneficial for transcriptome assembly. Since long error-prone reads are used only to detect paths in

Table 2 Comparison between Trinity and rnaSPAdes on short-read data, and hybrid assembly performed by rnaSPAdes using raw long reads and long reads corrected with Racon and RNA-Bloom for Human MCF7 dataset

Assembly	Trinity	rnaSPAdes	rnaSPAdes hybrid	rnaSPAdes + Racon	rnaSPAdes + RNA-Bloom
Transcripts	305K	243K	264K	260K	270K
Misassemblies	2719	**2170**	3968	3001	4059
Mismatches per contig	1.85	**1.50**	2.31	1.82	2.26
Database coverage	**0.26**	0.23	0.25	0.24	0.25
Duplication ratio	2.45	**1.45**	2.96	2.28	2.94
50%-assembled genes	13422	13546	**13990**	13728	13974
95%-assembled genes	6091	6310	**7379**	7061	7370
50%-assembled isoforms	**24575**	21048	23885	22427	23795
95%-assembled isoforms	7856	7355	**9333**	8657	9273

Since for this dataset FL reads are available, they were fed into rnaSPAdes using the appropriate option. Best value in each row is highlighted with bold

the graph, per base accuracy decreases only by 0.32 mismatches per assembled transcript on average when long reads are added.

Using pre-corrected Iso-seq reads, however, does not bring any substantial improvement comparing to raw data. Racon seems to remove a significant fraction of input data, which decreases the number of assembled genes and isoforms. Using RNA-Bloom's corrected Iso-seq reads produces the assembly of almost the same quality. However, using ONT reads from Human GM12878 dataset corrected by RNA-Bloom allows to obtain slightly more 95%-assembled genes and isoforms comparing to the original data (1.6% and 3.1% respectively).

Exploiting long reads, however, also increases the number of misassemblies and duplication ratio. Manual analysis of misassembled contigs revealed that they are usually caused by the (i) presence of intronic and intergenic sequences, (ii) potentially unknown isoforms, (iii) merged neighboring genes and (iv) inaccurately assembled contigs that map to different loci, e.g. paralogous genes or intergenic space. However, since typical pipelines for analysis of transcriptomic data involve additional steps, such as transcript validation and annotation, duplicated and chimeric sequences have a less dramatic impact on the final results than, for example, in de novo genome assembly projects.

Table 3 Comparison between Trinity and rnaSPAdes on short-read data, and hybrid assembly performed by rnaSPAdes using raw long reads and long reads corrected with Racon and RNA-Bloom for Human siNT_48 dataset

Assembly	Trinity	rnaSPAdes	rnaSPAdes Hybrid	rnaSPAdes + Racon	rnaSPAdes + RNA-Bloom
Transcripts	463K	426K	444K	436K	443K
Misassemblies	2625	**1978**	2657	2389	2661
Mismatches per contig	1.78	**1.32**	1.43	1.39	1.43
Database coverage	**0.29**	0.25	0.26	0.26	0.26
Duplication ratio	2.43	**1.49**	2.08	1.88	2.07
50%-assembled genes	**15681**	15573	15570	15549	15573
95%-assembled genes	7516	7831	**8186**	8117	8179
50%-assembled isoforms	**29150**	24411	25660	24965	25669
95%-assembled isoforms	9817	9300	**10038**	9886	10023

Best value in each row is highlighted with bold

Table 4 Comparison between Trinity and rnaSPAdes on short-read data, and hybrid assembly performed by rnaSPAdes using raw long reads and long reads corrected with Racon and RNA-Bloom for Human GM12878 dataset

Assembly	Trinity	rnaSPAdes	rnaSPAdes Hybrid	rnaSPAdes + Racon	rnaSPAdes + RNA-Bloom
Transcripts	278K	234K	237K	237K	243K
Misassemblies	2548	**2247**	2350	2332	2763
Mismatches per contig	1.93	**1.66**	1.69	1.69	1.76
Database coverage	**0.21**	0.18	0.19	0.19	0.19
Duplication ratio	1.86	**1.28**	1.41	1.41	1.58
50%-assembled genes	11436	11506	11544	11515	**11598**
95%-assembled genes	4427	5075	5264	5283	**5349**
50%-assembled isoforms	**19396**	16396	16674	16639	17489
95%-assembled isoforms	5463	5713	6038	6039	**6223**

Best value in each row is highlighted with bold

Conclusion

In this work, we present a new rnaSPAdes workflow designed to improve de novo transcriptome assembly using long RNA reads. By using several human datasets containing both short and long reads, we show that the hybrid approach allows to restore more known genes, comparing to short-read only assemblers, Trinity and rnaSPAdes. We also demonstrate that long reads (especially FL-reads) can also be beneficial for discovery of alternatively spliced isoforms, which can be useful in various studies involving transcriptome analysis of previously unsequenced organisms.

Availability and requirements

Project name: rnaSPAdes

Project home page: cab.spbu.ru/software/rnaspades/, github.com/ablab/spades

Operating systems: Linux and MacOS

Programming language: C++, Python

Other requirements: no requirements for pre-compiled binaries; g++ 5.3.1+, cmake 2.8.12+, zlib and libbz2 are required for compiling from source code

License: GPLv2

Restrictions to use by non-academics: None

Abbreviations

FL: Full-length; ONT: Oxford nanopore technologies; PacBio: Pacific biosciences; RNA-Seq: RNA-sequencing

Acknowledgments

For uploading their data to public databases the authors would like to thank Nick Loman and other contributors of Oxford Nanopore RNA sequencing project, as well as the staff of following organizations: Stanford University, Netherlands Cancer Institute, Berlin Institute for Medical Systems Biology, and Pacific Biosciences.

About this supplement

This article has been published as part of [BMC Bioinformatics, Volume 21 Supplement 12, 2020: Selected abstracts and papers of Bioinformatics: from Algorithms to Applications 2019 conference. The full contents of the supplement are available at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-21-supplement-12.

Authors' contributions

Software design and implementation were performed by EB, DA, and AP. EB, AM, and AP were responsible for data curation, assemblers benchmarking and manuscript editing. GP, DG, and DV initiated the project and sequenced data that was used for testing and benchmarking (not presented in this manuscript). AL and DV supervised the project and performed funding acquisition. AP wrote the manuscript and managed the project. All author(s) read and approved the final manuscript.

Funding

Publication of this supplement is funded by Russian Science Foundation (grant number 19-14-00172).

Availability of data and materials

Source code and manual of rnaSPAdes tool is available on http://cab.spbu.ru/software/spades/. Oxford Nanopore reads from Human GM12878 dataset are available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md. Iso-seq data for Human MCF7 dataset was released here http://datasets.pacb.com.s3.amazonaws.com/2015/IsoSeqHumanMCF7Transcriptome/list.html. Other data is available at short read archive (https://www.ncbi.nlm.nih.gov/sra) with the following accession numbers

- Human GM12878: SRR3103887
- Human siNT_48: SRP126849 (siNT_48 sample)
- Human MCF7: SRX426377

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia. [2]Consiglio Nazionale delle Ricerche, Istituto per i Sistemi Agricoli e Forestali del Mediterraneo Catania, Italy. [3]Department of Electrical, Electronics and Computer Engineering, University of Catania Catania, Italy.

Received: 14 June 2020 Accepted: 18 June 2020 Published: 24 July 2020

References

1. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201.
2. Pacific Biosciences. Intro to the Iso-Seq Method: Full-length transcript sequencing. 2014. https://www.pacb.com/blog/intro-to-iso-seq-method-full-leng. Accessed 3 July 2020.
3. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. Nanopore long-read rnaseq reveals widespread transcriptional variation among the surface receptors of individual b cells. Nat Commun. 2017;8:16027.
4. Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. Gigascience. 2017;6(11):086.
5. Herzel L, Straube K, Neugebauer KM. Long-read sequencing of nascent rna reveals coupling among rna processing events. Genome Res. 2018;28(7):1008–19.
6. Zhang G, Sun M, Wang J, Lei M, Li C, Zhao D, Huang J, Li W, Li S, Li J, et al. Pacbio full-length cdna sequencing integrated with rna-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 2019;97(2):296–305.
7. Chen H, Gao F, He M, Ding XF, Wong AM, Sze SC, Yu AC, Sun T, Chan AW-H, Wang X, et al. Long-read RNA sequencing identifies alternative splice variants in hepatocellular carcinoma and tumor-specific isoforms. Hepatology. 2019;70(3):1011–25.
8. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
9. Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J, Au K. IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics. 2018;34(13):2168–76.
10. Nip KM, Chiu R, Yang C, Chu J, Mohamadi H, Warren RL, Birol I, Rna-bloom provides lightweight reference-free transcriptome assembly for single cells. bioRxiv. 2019;701607.
11. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to rna-seq data. GigaScience. 2019;8(9):100.
12. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;32(7):1009–15.
13. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics. 2016;32(14):2210–2.
14. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,. J Comput Biol. 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.
15. Prjibelski AD, Vasilinetc I, Bankevich A, Gurevich A, Krivosheeva T, Nurk S, Pham S, Korobeynikov A, Lapidus A, Pevzner PA. ExSPAnder: a universal repeat resolver for DNA fragment assembly. Bioinformatics. 2014;30(12):293–301.
16. Vasilinetc I, Prjibelski AD, Gurevich A, Korobeynikov A, Pevzner PA. Assembling short reads from jumping libraries with large insert sizes. Bioinformatics. 2015;31(20):3262–8.
17. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:1303.3997. 2013.
18. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. 2010.

19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
20. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
21. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;2(8):1134–44.
22. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read rna-seq assemblers. GigaScience. 2019;8(5):039.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.