Bioinformatics, 32(14), 2016, 2210–2212
doi: 10.1093/bioinformatics/btw218
Advance Access Publication Date: 23 April 2016
Applications Note

Sequence analysis

rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies

Elena Bushmanova[1], Dmitry Antipov[1], Alla Lapidus[1,2], Vladimir Suvorov[3] and Andrey D. Prjibelski[1,2,*]

[1]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia, [2]Algorithmic Biology Lab, St. Petersburg Academic University, St. Petersburg, Russia and [3]Research and Development Department, EMC, St. Petersburg, Russia

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on June 2, 2015; revised on March 2, 2016; accepted on April 18, 2016

Abstract

Summary: Ability to generate large RNA-Seq datasets created a demand for both *de novo* and reference-based transcriptome assemblers. However, while many transcriptome assemblers are now available, there is still no unified quality assessment tool for RNA-Seq assemblies. We present rnaQUAST—a tool for evaluating RNA-Seq assembly quality and benchmarking transcriptome assemblers using reference genome and gene database. rnaQUAST calculates various metrics that demonstrate completeness and correctness levels of the assembled transcripts, and outputs them in a user-friendly report.

Availability and Implementation: rnaQUAST is implemented in Python and is freely available at http://bioinf.spbau.ru/en/rnaquast.

Contact: ap@bioinf.spbau.ru

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Introduction

Next-generation sequencing technologies have raised a challenging problems of *de novo* genome and transcriptome assembly from short reads. As a result, multiple assembly tools were developed in the last decade. Even though most papers describing novel assembly approaches provide benchmarking of various tools, no unified methods for assessing assembly quality were developed until recently.

While multiple tools for both reference-based and reference-free genome assembly evaluation were developed (Clark *et al.*, 2013; Gurevich *et al.*, 2013; Hunt *et al.*, 2013; Salzberg *et al.*, 2011) and later used in various genomic studies (Coil *et al.*, 2014; Howison *et al.*, 2013; Magoc *et al.*, 2013; Zimin *et al.*, 2013), no tool have set a standard for assessing quality of transcriptome assemblies. Although several studies provided evaluation methods (Li *et al.*, 2014; O'Neil and Emrich, 2013; Simão *et al.*, 2015) or performed independent benchmarks of RNA-Seq assemblers (Martin and Wang, 2011; Mundry *et al.*, 2012; Steijger *et al.*, 2013), in papers describing novel transcriptome assembly software the resulting transcripts were evaluated using different in-house methods (Grabherr *et al.*, 2011; Peng *et al.*, 2013; Robertson *et al.*, 2010; Xie *et al.*, 2014), therefore making it difficult to compare the results across various publications.

We present rnaQUAST—a quality assessment tool for transcriptome assemblies, which utilizes reference genome and gene database. rnaQUAST takes assembled transcripts as an input and maps them to the reference genome using either BLAT (Kent, 2002) or GMAP (Wu and Watanabe, 2005). By comparing the resulting alignments with the gene database, rnaQUAST calculates various statistics and generates a summary report. Below we describe methods implemented in rnaQUAST, present an example of summary report and compare rnaQUAST with DETONATE software (Li *et al.*, 2014).

2 Methods

2.1 rnaQUAST pipeline

To evaluate quality of the assembled transcripts, rnaQUAST takes a reference genome in FASTA format and optionally its gene database in GFF/GTF format. A user can provide either a FASTA file with transcripts, which will be aligned to the given reference genome using GMAP (Wu and Watanabe, 2005) or BLAT (Kent, 2002) (can be selected by the user), or alternatively align transcripts to the reference genome using software of his choice (e.g. Splign (Kapustin *et al.*, 2008)) and provide alignments in PSL format. The alignments are analyzed to calculate simple metrics and then are matched against the isoforms from the gene database in order to obtain statistics that represent completeness and correctness levels of the assembly. In addition, rnaQUAST is capable of estimating gene database coverage by raw reads using STAR (Dobin *et al.*, 2013) or TopHat2 (Kim *et al.*, 2013). For *de novo* quality assessment when reference genome and gene database are unavailable, the transcripts are analyzed using BUSCO (Simão *et al.*, 2015) and GeneMarkS-T (Tang *et al.*, 2015). To compare several transcriptome assemblies, rnaQUAST is capable of taking multiple FASTA (or PSL) files as an input.

2.2 Metrics and alignment analysis

rnaQUAST calculates various metrics without using alignment information, e.g. length distribution and N50 of the assembled transcripts. Additionally, rnaQUAST computes the following statistics for the gene database: the total number of genes and isoforms, isoform and exon length distribution, average number of exons per gene, etc.

To analyze transcripts' alignments, rnaQUAST firstly filters out short partial alignments (shorter than a user-defined threshold, default value is 50 bp). Such short alignments are typically caused by genomic repeats and thus are ignored. Afterwards, rnaQUAST selects the best-scored spliced alignment for each transcript. If a transcript has more than one alignment with the highest score, it is reported as *multiply aligned*. Otherwise, it is considered to be *uniquely aligned*. If the best-scored alignment is discordant (e.g. the transcript has partial alignments that are either mapped to different strands or to different chromosomes) the transcript is classified as *misassembled* (see the Supplementary Material for the details). Transcripts without misassemblies are analyzed to calculate such metrics as average transcript alignment fraction and mismatch rate.

For the simplicity of explanation, *transcript* is further referred to as a sequence generated by the assembler and *isoform* denotes a sequence from the gene database. rnaQUAST matches best-scored alignments of non-misassembled transcripts to the isoforms' coordinates and analyzes them to estimate how well the isoforms are covered by the assembly. rnaQUAST computes such metrics as database coverage (the total number of covered bases of all isoforms divided by the total length of all isoforms) and the number of 50%/95%-assembled isoforms. An isoform is considered to be x%-assembled if it has at least x% covered by a single transcript. Vice versa, to evaluate how well the assembled transcripts are covered by the isoforms, rnaQUAST estimates the number of unannotated transcripts (that align to the genome, but do not match to any isoform) and the number of 50%/95%-matched transcripts (that have corresponding fraction mapped to an isoform). Indeed, the thresholds described above (50% and 95%) can be varied by the user. Complete list of metrics reported by rnaQUAST is described in the user manual, available at http://bioinf.spbau.ru/en/rnaquast.

3 Results

In this section we demonstrate assembly quality metrics calculated by rnaQUAST and compare them with scores computed by REF-EVAL and RSEM-EVAL from DETONATE software package version 1.10 (Li *et al.*, 2014). To perform the comparison we assembled *M. musculus* RNA-Seq paired-end library using the following transcriptome assemblers: Trans-ABySS 1.5.3 (Robertson *et al.*, 2010), IDBA-tran 1.1.1 (Peng *et al.*, 2013), SOAPdenovo-Trans 1.03 (Xie *et al.*, 2014) and Trinity 2.1.1 (Grabherr *et al.*, 2011). IDBA-tran and Trinity were run with the default parameters ($k = 20, 30, 40, 50, 60$ for IDBA-tran and $k = 25$ for Trinity). As for Trans-ABySS and SOAPdenovo-Trans, we ran both tools using various k-mer lengths and selected the best assemblies for the comparison. We used the number of 95%-assembled isoforms, number of misassemblies and database coverage reported by rnaQUAST as the main criteria for selecting optimal assemblies. We also included contigs produced by SPAdes 3.6.1 genome assembler (Bankevich *et al.*, 2012; Nurk *et al.*, 2013), which was run in single-cell mode (due to uneven coverage of RNA-Seq data) with the default k-mer lengths ($k = 21$, 33, 55). Although SPAdes was not designed as a transcriptome assembler, it turned out to show decent results on RNA-Seq data.

We included the most important metrics from the reports produced by rnaQUAST and DETONATE and presented them in Table 1. rnaQUAST was launched with the default parameters; REF-EVAL and RSEM-EVAL were run as recommended in the user manual (http://deweylab.biostat.wisc.edu/detonate/vignette.html). Command lines that were used to run all tools are provided in the Supplementary Material.

Although there is no direct connection between rnaQUAST metrics and scores reported by the DETONATE software (see Li *et al.* (2014) for the details), Table 1 shows that they mostly have a good correlation. To ease the comparison with DETONATE, we also added the number of *99%-assembled isoforms*, since REF-EVAL uses the 99% threshold for calculating contig scores.

Table 1 demonstrates that Trans-ABySS assembly has the largest number of 50%-assembled isoforms and the highest database per nucleotide coverage, and at the same the highest contig recall value. On the other hand, Trans-ABySS has the largest fraction of unaligned (10%) and unannotated (25%) transcripts, which correlates with the lowest contig precision reported by REF-EVAL. IDBA-tran assembly, conversely, has the highest fraction of 50%/95%-matched transcripts (76% and 82% respectively), but the lowest database coverage, which correlates with the highest contig and nucleotide precision, and rather low contig, nucleotide and k-mer recall values. SOAPdenovo-Trans generates an accurate assembly in terms of number of misassemblies, mismatch rate and nucleotide precision, but its assembly appears to be rather fragmented (almost 77% of transcripts are shorter than 500 bp, the smallest number of 50%/95%/99%-assembled isoforms). Similarly, Trans-ABySS assembly also contains a lot of short sequences—about 83% of assembled transcripts are shorter than 500 bp.

Trinity and, surprisingly, SPAdes have relatively high database coverage and recall metrics, and the same time assemble the largest number of 95%/99%-assembled isoforms. However, both tools generate rather high number of misassembled contigs with SPAdes having approximately twice more misassemblies than Trinity. Elevated number of misassembled contigs in SPAdes assembly can be explained by the fact that SPAdes is a genome assembler and has no specific algorithm for detecting transcripts in the de Bruijn graph during assembly.

Table 1. rnaQUAST and DETONATE metrics for the transcripts assembled by Trans-**ABySS**, **IDBA**-tran, **SOAP**denovo-Trans, **SPAdes** and **Trinity** on *M. musculus* RNA-Seq non-strand-specific paired-end library with read length 101 bp and average insert size 173 bp (accession number SRX648736)

Assembler	ABySS	IDBA	SOAP	SPAdes	Trinity
k-mer size	32	default	31	default	default
rnaQUAST metrics					
Transcripts	107202	38294	69331	48706	51245
Transcripts $\geq$ 500 bp	17882	***17542***	16021	17512	***21994***
Aligned	95884	***38198***	68591	48027	***51112***
Uniquely aligned	94681	37288	***67878***	45091	49846
Unaligned	11318	***96***	740	679	***133***
50%-matched	66744	***32574***	54581	37447	43039
95%-matched	61633	***29429***	50876	32565	35239
Unannotated	26678	***3905***	12252	7102	5740
Database coverage	***18.5***	***16.9***	***17.2***	***17.6***	***18.1***
50%-assembled isoforms	***7061***	6777	6241	6887	7020
95%-assembled isoforms	1907	1611	1397	***2292***	2053
99%-assembled isoforms	432	431	347	***754***	710
Misassemblies	267	471	***26***	942	465
Mismatches per transcript	***0.50***	***1.04***	***0.58***	***1.13***	***1.28***
REF-EVAL scores					
Nucleotide precision	0.69	***0.86***	0.84	0.81	0.69
Nucleotide recall	0.76	0.75	0.75	***0.79***	0.78
Nucleotide F_1	0.73	***0.80***	0.79	***0.80***	0.73
Contig precision	0.095	***0.17***	0.14	0.14	0.14
Contig recall	***0.096***	0.063	0.089	0.066	0.068
Contig F_1	0.095	0.092	***0.11***	0.090	0.092
k-mer recall	0.84	0.34	0.76	0.67	***0.90***
KC score	0.80	0.31	0.73	0.64	***0.86***
RSEM-EVAL score ($\times 10^9$)	-1.42	-2.31	-1.40	-1.48	***-0.98***

M. musculus gene database (Waterston et al., 2002) contains 38924 genes (22572 protein-coding) and 94545 isoforms (47394 protein-coding). In each row the best values are indicated with ***bold italic***. For the transcript metrics (rows 3–9) we highlighted the best ***relative*** values i.e. divided by the total number of transcripts in the corresponding assembly.

While Trinity fairly has the best RSEM-EVAL score, high *k*-mer recall and *k*-mer compression score, IDBA-tran is the outsider according to these metrics, which may be explained by the small absolute number of transcripts matching to the isoforms and thus, lack of genomic *k*-mers in the assembly. At the same time, both Trans-ABySS and SOAPdenovo-Trans also have decent *k*-mer recall, *k*-mer compression and RSEM-EVAL scores, but produce rather fragmented assemblies comparing to Trinity and SPAdes. Since sequence continuity is an important characteristic for the *de novo* transcript reconstruction, such metrics as *k*-mer recall, KC score and RSEM-EVAL score (which strongly correlates with *k*-mer scores (Li *et al.*, 2014)) do not provide the whole picture of the evaluated assemblies.

Even though rnaQUAST metrics do not have a perfect correlation with all DETONATE scores, it is important that both tools provide similar assembly ranking. While DETONATE provides quantitative assessment of the assembled transcripts by calculating different scores, rnaQUAST computes large variety of metrics leaving the decision for the user, which may prioritize metrics depending on the goals of his particular project. More statistics calculated by DETONATE and rnaQUAST on various RNA-Seq datasets (including strand-specific), as well as examples of plots generated by rnaQUAST are provided in the Supplementary Material.

Funding

This work was supported by EMC Research and Development Department and St. Petersburg State University [grant number 15.61.951.2015].

Conflict of Interest: none declared.

References

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Clark,S. *et al.* (2013) Ale: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, **29**, 435–443.

Coil,D. *et al.* (2014) A5-miseq: an updated pipeline to assemble microbial genomes from illumina miseq data. *Bioinformatics*, **31**, 587–589.

Dobin,A. *et al.* (2013) Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**, 15–21.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Howison,M. *et al.* (2013) Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics*, **29**, 2959–2963.

Hunt,M. *et al.* (2013) Reapr: a universal tool for genome assembly evaluation. *Genome Biol.*, **14**, R47.

Kapustin,Y. *et al.* (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*, **3**, 20.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kim,D. *et al.* (2013) Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Li,b. *et al.* (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.*, **15**, 553.

Magoc,T. *et al.* (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, **29**, 1718–1725.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Mundry,M. *et al.* (2012) Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PLoS ONE*, **7**, e31410.

Nurk,S. *et al.* (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.*, **20**, 1–24.

O'Neil,S.T. and Emrich,S.J. (2013) Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**, 465.

Peng,Y. *et al.* (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, **29**, i326–i334.

Robertson,G. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

Salzberg,S.L. *et al.* (2011) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, **22**, 557–567.

Simão,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Tang,S. *et al.* (2015) Identification of protein coding regions in rna transcripts. *Nucleic Acids Res*, **43**, e78.

Waterston,R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xie,Y. *et al.* (2014) SOAPdenovo-Trans: De novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, **30**, 1660–1666.

Zimin,A.V. *et al.* (2013) The Masurca genome assembler. *Bioinformatics*, **29**, 2669–2677.