Bioinformatics, 32(7), 2016, 1088–1090
doi: 10.1093/bioinformatics/btv697
Advance Access Publication Date: 26 November 2015
Applications Note

Genome analysis

MetaQUAST: evaluation of metagenome assemblies

Alla Mikheenko, Vladislav Saveliev and Alexey Gurevich*

Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg 199034, Russia

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on 20 August 2015; revised on 10 November 2015; accepted on 23 November 2015

Abstract

Summary: During the past years we have witnessed the rapid development of new metagenome assembly methods. Although there are many benchmark utilities designed for single-genome assemblies, there is no well-recognized evaluation and comparison tool for metagenomic-specific analogues. In this article, we present MetaQUAST, a modification of QUAST, the state-of-the-art tool for genome assembly evaluation based on alignment of contigs to a reference. MetaQUAST addresses such metagenome datasets features as (i) unknown species content by detecting and downloading reference sequences, (ii) huge diversity by giving comprehensive reports for multiple genomes and (iii) presence of highly relative species by detecting chimeric contigs. We demonstrate MetaQUAST performance by comparing several leading assemblers on one simulated and two real datasets.

Availability and implementation: http://bioinf.spbau.ru/metaquast.

Contact: aleksey.gurevich@spbu.ru

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Metagenomics studies genetic material taken directly from environmental samples. NGS technologies allow for extracting short reads even from low abundant organisms without cloning. However, the data produced in such experiments tend to be enormous, noisy, and contain fragments from thousands of species highly varying in abundance and homology. Those challenges led to a new computational problem of metagenome assembly, followed by a diversity of methods (Boisvert *et al.*, 2012; Peng *et al.*, 2012; Haider *et al.*, 2014), which demands a standard benchmark procedure.

Most existing assembly evaluation approaches are not designed to work with metagenomes. However there exist methods which count read likelihoods with respect to the assembly (Clark *et al.*, 2013; Ghodsi *et al.*, 2013), or determine single-copy conservative ubiquitous gene content (Parks *et al.*, 2015; Simao *et al.*, 2015). Unfortunately, none uses contig alignments to a closely related reference genome. In this article we present MetaQUAST, a metagenomic-specific improvement over QUAST (Gurevich *et al.*, 2013). QUAST detects errors based on alignments to a given closely related reference genome, and also reports and plots contig statistics like N50 and gene content which gives an overview of constituent species even without user-supplied reference sequences. To address metagenome assemblies, MetaQUAST adds several new features: (i) ability of using an unlimited number of reference genomes, (ii) automated species content detection, (iii) detection of chimeric contigs (interspecies misassemblies) and (iv) significantly redesigned visualizations.

2 Materials and Methods

2.1 Reference-based evaluation

There are well-studied metagenomic datasets with known species content (Qin *et al.*, 2010) or simulated reads (Boisvert *et al.*, 2012; Namiki *et al.*, 2012). They can be used with MetaQUAST to evaluate assembly methods based on alignments to reference genomes.

The multiple-reference pipeline consists of four major steps (Supplementary Fig. S1):

1. All reference genomes get concatenated into one file (combined reference). QUAST is fed with all input assemblies versus the combined reference. We force QUAST to report all ambiguous alignments instead of only one. For metagenomic datasets containing closely related species, all ambiguous alignments are essential.
2. We partition all contigs into groups, each of which contains sequences mapped to a particular reference genome (based on previously generated alignments). The contigs mapped to several genomes go into every matching group. Unaligned contigs are put into one extra group.
3. Next, QUAST is launched for each input reference separately, feeding it with a corresponding group of contigs. The group of unaligned contigs is processed without any input reference.
4. Finally, the results of all QUAST runs are grouped together into summary reports and visualizations. A user can view both detailed full QUAST outputs for each run, as well as bird-eye overviews of the results from the entire dataset.

In addition to the QUAST standard set of quality statistics (N50, genome fraction, etc.), we added two metrics:

- No. of interspecies translocations: A type of misassembly where the flanking sequences align to distinct references [similarly to translocation introduced in (Gurevich *et al.* 2013) where flanking sequences align to different chromosomes].
- No. of possibly misassembled contigs: The number of contigs that include both large aligned and unaligned fragments, thus possibly contain interspecies translocation with an unknown genome.

In contrast with regular QUAST which uses GeneMarkS, MetaQUAST uses MetaGeneMark (Zhu *et al.*, 2010) for gene prediction, which is developed specially for metagenomes.

2.2 *De novo* evaluation

Most experimental metagenomic studies operate with *de novo* assemblies where the reference information is not available. When MetaQUAST is executed without input reference sequences or species lists, it attempts to identify species content and automatically pull reference sequences. Note that the algorithm works under the assumption that researchers are mostly interested in microbial communities, so the search is restricted to bacteria and archaea.

The workflow (see Supplementary Fig. S2) starts with applying BLASTn (Camacho *et al.*, 2009) to align contigs to the 16S rRNA sequences from the SILVA database (Quast *et al.*, 2012). The 16S subunits, which are present in almost all microbial species, are highly conserved sequences but also include a hypervariable region that can serve to classify organisms into taxonomic groups. For each detected species, one strain with the best score is remained in the assembly.

Top 50 organisms are queried against NCBI, and the least fragmented sequence for each species is downloaded. Due to the known issues with differences in the copy number of rRNA operons between organisms, and intra-genomic heterogeneity of the 16S genes, some of the downloaded genome sequences may not have representation in the assembly under assessment. MetaQUAST attempts to filter false positives by removing genomes with a contig coverage fraction of less than 10% (for all assemblies). In special cases when all sequences have a very low genome fraction, the list remains unfiltered.

As a result, we obtain a set of genomes possibly represented by the assembled sequences. We launch MetaQUAST using these sequences (as in section 2.1) and produce the same output files as in the case of usual reference-based analysis.

Our approach is a compromise between accuracy and time/memory consumption. For more precise results, we would advise using MGTAXA (Williamson *et al.*, 2012), or methods based on exact read alignments, e.g. Kraken (Wood and Salzberg, 2014) or CLARK (Ounit *et al.*, 2015). Very precise results can be obtained by a BLASTx (Altschul *et al.*, 1990) search against the whole NCBI-nr database. The acquired list of species names can be fed to MetaQUAST in a plain text format, making it download the specified sequences from the NCBI database and use them for the reference-based evaluation (see Section 2.1).

2.3 Refining misassemblies based on read mapping

The regular single-genome QUAST algorithm reports structural disagreements between contigs and a reference genome as misassemblies. However, in some cases they may be proof of structural variants (SVs) rather than true assembly errors. This is especially important when analysing a metagenomic community for which no close references are available. MetaQUAST addresses this problem by taking pair reads mapping into consideration (Supplementary Fig. S3). MetaQUAST applies a structural variation finding algorithm to detect breakpoints based on discordant read-pairs, which then used to mute misassemblies that share called breakpoints.

2.3.1 SV detection

MetaQUAST utilizes bowtie2 (Langmead *et al.*, 2009) for performing reads alignment against the combined reference genome. The BAM file (Li *et al.*, 2009) produced by bowtie2 is sorted by coordinate and passed as an input for SV discovery software. We have chosen Manta (Chen *et al.*, 2015) SV caller that outperforms LUMPY (Layer *et al.*, 2014) and Pindel (Ye *et al.*, 2009) by sensitivity and precision on our test datasets.

2.3.2 Misassembly classification

Each misassembly reported by QUAST is compared with breakpoint confidence intervals of all discovered SVs. If both start and end coordinates of the misassembly lie within the SV intervals extended by a small δ, MetaQUAST marks this misassembly fake and does not include into the final report. If no similarity is found between SVs and the misassembly, it is considered true. The default δ value is 100 bp which is based on manual analysis of dozens SVs occurred on real and simulated datasets.

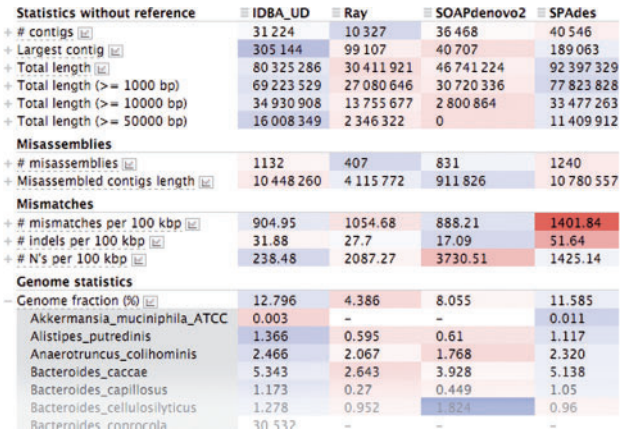
This approach allows us to significantly reduce number of falsely reported misassemblies on all three test datasets. See Supplementary Material for detailed benchmarking results.

2.4 Visualization

MetaQUAST complements the QUAST visualizations with a number of bird-eye overviews. In addition, an interactive summary HTML report combining key statistics for all assemblies and references is generated. Charts and summary HTMLs are demonstrated in Supplementary Material.

We classify summary plots into three groups:

- Misassembly plots: distribution of misassemblies by type (relocations, inversions, translocations and interspecies translocations). They exist in two views: across all assemblies per reference and across all references per assembly.



Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	31 224	10 327	36 468	40 546
Largest contig	305 144	99 107	40 707	189 063
Total length	80 325 286	30 411 921	46 741 224	92 397 329
Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
Total length (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
Total length (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
Misassemblies				
# misassemblies	1132	407	831	1240
Misassembled contigs length	10 448 260	4 115 772	911 826	10 780 557
Mismatches				
# mismatches per 100 kbp	904.95	1054.68	888.21	1401.84
# indels per 100 kbp	31.88	27.7	17.09	51.64
# N's per 100 kbp	238.48	2087.27	3730.51	1425.14
Genome statistics				
Genome fraction (%)	12.796	4.386	8.055	11.585
Akkermansia_muciniphila_ATCC	0.003	-	-	0.011
Alistipes_putredinis	1.366	0.595	0.61	1.117
Anaerotruncus_colihominis	2.466	2.067	1.768	2.320
Bacteroides_caccae	5.343	2.643	3.928	5.138
Bacteroides_capillosus	1.173	0.27	0.449	1.05
Bacteroides_cellulosilyticus	1.278	0.952	1.824	0.96
Bacteroides_coprocola	30 532	-	-	-

Fig. 1. Part of a summary HTML report for the MetaHIT dataset. The cells containing outliers are colored. In this example, the genome fraction per-reference info is expanded

- Metric-level plots: one per metric, for all assemblies versus all references. Genomes are ordered by the average value among all assemblies, starting from the best.
- Krona charts (Ondov *et al.*, 2011): one per assembly, and one for the whole dataset. Round charts show the taxonomic profile. Available only in the *de novo* evaluation mode.

The interactive summary HTML report aggregates tables and plots for all statistics, references and assemblies. Each table row shows a value for the combined reference and can be expanded to show values per each reference (see Fig. 1). The blue/red heatmap emphasizes outliers.

3 Results

We tested MetaQUAST on three datasets: the CAMI (http://cami-challenge.org) simulated toy dataset, the MH0045 sample from MetaHit and the SRS077736 tongue dorsum female sample from HMP (Human Microbiome Project Consortium *et al.*, 2012). We assembled these data using four leading assemblers commonly used in metagenomic studies: IDBA-UD (Peng *et al.*, 2012), SPAdes (Bankevich *et al.*, 2012), Ray Meta (Boisvert *et al.*, 2012) and SOAPdenovo2 (Luo *et al.*, 2012). Comparisons results and MetaQUAST performance on all three datasets are demonstrated in Supplementary Material.

The comparison on these datasets demonstrated that none of the assemblers can be called an undisputed leader in metagenomic assembly. Thus, tools such as MetaQUAST are of great practical importance for the community. It will help scientists to assess different assembly software and choose the best pipeline for their research.

Acknowledgements

We would like to thank A. Prjibelski, S. Nurk, D. Meleshko and D. Antipov from the SPAdes team for substantial feedback on our software; the CAMI team for extremely helpful comments and feature requests; and T. Amariuta for proofreading this text.

Funding

This work was supported by Russian Science Foundation [grant number 14-50-00069].

Conflict of Interest: none declared.

References

Altschul,S. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Boisvert,S. *et al.* (2012) Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122.

Camacho,C. *et al.* (2009) Blast+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chen,X. *et al.* (2015) Manta: rapid detection of structural variants and indels for clinical sequencing applications. *Bioinformatics*, doi:10.1093/bioinformatics/btv710.

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.

Ghodsi,M. *et al.* (2013) De novo likelihood-based measures for comparing genome assemblies. *BMC Res. Notes*, **6**, 334.

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

Haider,B. *et al.* (2014) Omega: an overlap-graph de novo assembler for metagenomics. *Bioinformatics*, **30**, 2717–2722.

Human Microbiome Project Consortium *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Layer,R. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Luo,R. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.

Namiki,T. *et al.* (2012) Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.

Ondov,B. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Ounit,R. *et al.* (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**, 236.

Parks,D. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.

Peng,Y. *et al.* (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1–8.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Quast,C. *et al.* (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

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

Williamson,S. *et al.* (2012) Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS One*, **7**, 18.

Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Zhu,W. *et al.* (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.