Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx **1–5**
doi:10.1093/nar/gkn000

WebQUAST: online evaluation of genome assemblies

Alla Mikheenko [1,†], Vladislav Saveliev [2,3,†], Pascal Hirsch [4], and Alexey Gurevich [5,6,*]

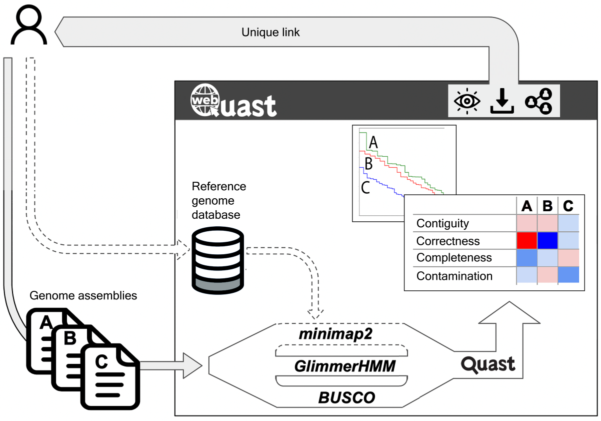
[1]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, University College London, London WC1E 6BT, UK, [2]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, New South Wales 2010, Australia, [3]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria 3052, Australia, [4]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken 66123, Germany, [5]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, Saarbrücken 66123, Germany and [6]Department of Computer Science, Saarland University, Saarbrücken 66123, Germany

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

Selecting proper genome assembly is key for downstream analysis in genomics studies. However, the availability of many genome assembly tools and the huge variety of their running parameters challenge this task. The existing online evaluation tools are limited to specific taxa or provide just a one-sided view on the assembly quality. We present WebQUAST, a web server for multifaceted quality assessment and comparison of genome assemblies based on the state-of-the-art QUAST tool. The server is freely available at `http://cab.cc.spbu.ru/quast/`. WebQUAST can handle an unlimited number of genome assemblies and evaluate them against a user-provided or pre-loaded reference genome or in a completely reference-free fashion. We demonstrate key WebQUAST features in three common evaluation scenarios: assembly of an unknown species, a model organism, and a close variant of it.

GRAPHICAL ABSTRACT



INTRODUCTION

Despite the ongoing long-read sequencing revolution, it is still impossible to read entire chromosomes for most species in a single run (1). Researchers use the so-called genome assembly software that combines the sequencing reads into longer genome fragments commonly referred to as contigs. Dozens of genome assemblers exist nowadays (2). These tools rely on different heuristics that greatly vary their output. Moreover, even different settings of the same tool may result in substantially diverging assemblies. The quality assessment and comparison of multiple genome assemblies are of utmost importance since the assembly choice greatly affects the downstream analysis (3).

The existing assembly evaluation tools comprise two major categories. The reference-based tools, such as GAGE (4), use gold-standard reference genomes to evaluate assemblies on model datasets. The reference-free methods either rely on read mapping back to assemblies to check their consistency

[*]To whom correspondence should be addressed. Email: alexey.gurevich@helmholtz-hips.de [†]These authors contributed equally to this work

with the input data and detect assembly errors, such as REAPR (5) and Inspector (6), or look for conservative genes to estimate the assembly completeness, such as BUSCO (7, 8) and CEGMA (9). Previously, we developed QUAST, an ensemble method that incorporated the best software from both categories, enhanced them with in-house quality metrics and plots, and became the state-of-the-art quality assessment tool for genome assemblies (10, 11). However, QUAST intrinsically inherited the limitations of the embedded tools which are available only for a few platforms (usually Linux) and have a command-line interface making them hardly suitable for researchers with a limited computational background.

Here, we present WebQUAST, a web server complementing QUAST with a user-friendly graphical interface and providing its functionality on any platform. In contrast to a few existing genome assembly evaluation web tools, WebQUAST is not restricted to specific taxa as gEVAL (12) and GenomeQC (13), performs versatile assembly evaluation rather than only completeness estimation as gVolante (14), and supports an unlimited number of assemblies on input. The WebQUAST evaluation reports can be browsed online, downloaded locally, and shared privately with colleagues. We show WebQUAST performance using a sample dataset of four *E. coli* assemblies.

MATERIALS AND METHODS

Web server overview

Workflow. A user uploads genome assemblies in the FASTA format (gzipped files are supported), configures the evaluation parameters, such as the minimal contig length cut-off and the organism type (eukaryote or prokaryote), and optionally selects a reference genome. The user might choose it from the list of pre-loaded genomes or upload a custom FASTA file that will be stored privately and can be reused later. Once the user clicks on the Evaluate button, WebQUAST transfers the input data to the QUAST processing engine.

If a reference genome is provided, the assemblies are aligned against it using minimap2 (15). If the BUSCO checkbox is selected, the assemblies are screened for single-copy orthologues from the corresponding BUSCO database (8). If the gene finding is requested, the assemblies are processed with the GlimmerHMM gene prediction software (16). QUAST combines the outputs of all employed modules to compute numeric quality metrics, create assessment plots and Icarus viewers (17), and generate a single evaluation report. WebQUAST assigns the report a unique web link and renders it for the user. The link enables browsing the results online and sharing them. The user can download the full standalone report to store it permanently. The standalone report also provides additional insights into the analysis, such as the running commands of the embedded tools or the list of identified misassemblies in the GFF format.

Software implementation. The server is built on top of the Python web framework Django. MySQLinstance is used to record users, sessions, and analysis requests. To support long-running analysis, the requests are processed and added into an asynchronous task queue Celery. A queued job represents a simple script that calls the command-line QUAST tool, which allows us to keep the main codebase agnostic to the web implementation. The front-end component is based on the jQuery framework.

Sample data preparation

To demonstrate WebQUAST performance, we generated sample assemblies of a well-studied short-read *E. coli* K-12 MG1655 dataset (SRA accession: ERR008613). The choice of a genome assembler might be influenced by many factors and one popular, yet often suboptimal, strategy is to choose among the most-cited methods (18). We mimicked this behavior by collecting information on short-read genome assemblers (Table 1) and selecting the five most-cited tools. We further excluded SOAPdenovo (19) since the authors discontinued it and recommended using MEGAHIT (20), which was already shortlisted.

Table 1. The most-cited short-read genome assemblers

Assembler	Latest release version (year)	Num citations total	yearly	Key publications with years
SPAdes	3.15.5 (2022)	18833	1847	2020 (21), 2012 (22)
Velvet	1.2.10 (2014)	10633	709	2008 (23)
SOAPdenovo	242 (2018)	7410	630	2012 (19), 2010 (24)
MEGAHIT	1.2.9 (2019)	4519	581	2016 (20), 2015 (25)
ABySS	2.3.5 (2022)	4445	366	2017 (26), 2009 (27)
IDBA	1.1.3 (2016)	2979	266	2012 (28), 2010 (29)
ALLPATHS	52488 (2016)	2868	210	2011 (30), 2008 (31)
MaSuRCA	4.1.0 (2023)	1434	164	2017 (32), 2013 (33)
Ray	2.3.1 (2014)	1232	103	2012 (34), 2010 (35)
SGA	0.10.15 (2016)	909	83	2012 (36)

Version numbers and dates of the latest release were determined from the GitHub repositories of the tools. *Num citations* stands for the number of citations according to Google.Scholar as of 28.03.2022, *yearly* average is the total number of citations divided by the sum of full years past since the publications. At most two key publications per tool are included; if there were more than two publications, we relied on the citation recommendations on the tool webpage (usually the first and the last publication).

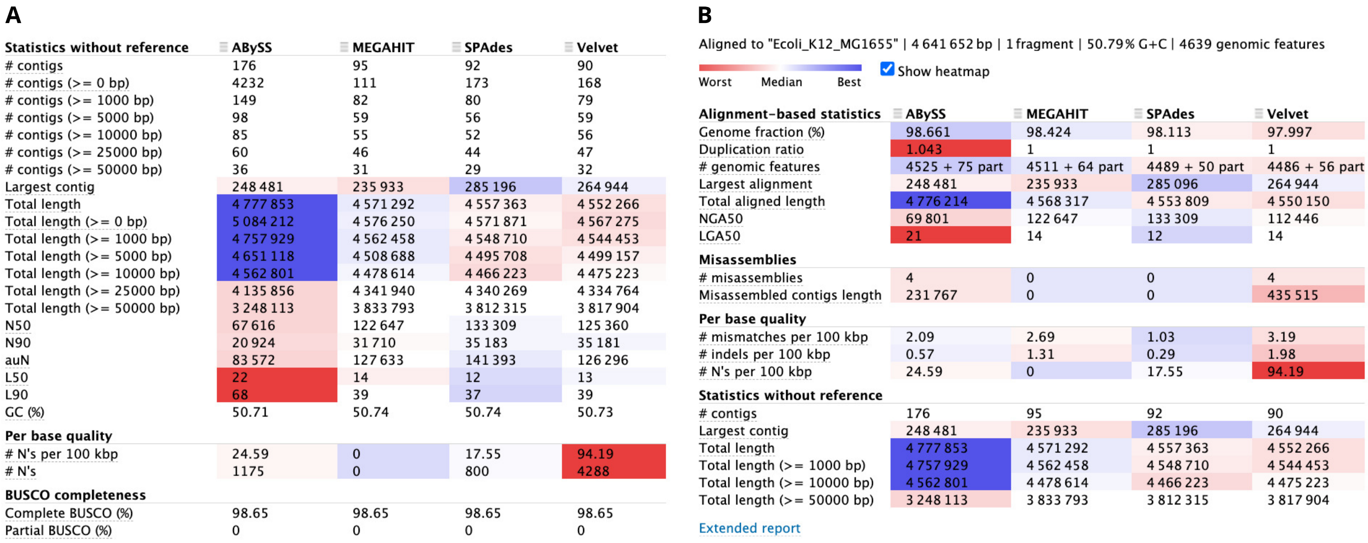
Some of the selected assemblers do not include a read error correction module, so we cleaned the raw sequencing data beforehand to make the comparison fair. We checked the reads with FastQC and trimmed low-quality ends with Trimmomatic (37). All assemblers but ABySS were run with default parameters or based on the recommendations in the documentation wherever available. We used the GAGE-B recipe (38) for ABySS since its default assembly was of very poor quality. All tools were installed via Bioconda (39), the installation and running commands are in the Supplementary Material.

RESULTS

Here we illustrate three typical WebQUAST usage scenarios. In each case, we evaluated the same four assemblies of the *E. coli* K-12 MG1655 dataset but selected the reference genome differently. We assumed the reference was unknown in Case 1, exactly matched the dataset in Case 2, and was closely related to the dataset in Case 3.

Use Case 1: reference-free evaluation

When a reference genome is unavailable, WebQUAST computes 30 quality metrics and draws three assessment plots that mainly address the contiguity and completeness of the



Statistics without reference	ABySS	MEGAHIT	SPAdes	Velvet
# contigs	176	95	92	90
# contigs (>= 0 bp)	4232	111	173	168
# contigs (>= 1000 bp)	149	82	80	79
# contigs (>= 5000 bp)	98	59	56	59
# contigs (>= 10000 bp)	85	55	52	56
# contigs (>= 25000 bp)	60	46	44	47
# contigs (>= 50000 bp)	36	31	29	32
Largest contig	248 481	235 933	285 196	264 944
Total length	4 777 853	4 571 292	4 557 363	4 552 266
Total length (>= 0 bp)	5 084 212	4 576 250	4 571 871	4 567 275
Total length (>= 1000 bp)	4 757 929	4 562 458	4 548 710	4 544 453
Total length (>= 5000 bp)	4 651 118	4 508 688	4 495 708	4 499 157
Total length (>= 10000 bp)	4 562 801	4 478 614	4 466 223	4 475 223
Total length (>= 25000 bp)	4 135 856	4 341 940	4 340 269	4 334 764
Total length (>= 50000 bp)	3 248 113	3 833 793	3 812 315	3 817 904
N50	67 616	122 647	133 309	125 360
N90	20 924	31 710	35 183	35 181
auN	83 572	127 633	141 393	126 296
L50	22	14	12	13
L90	68	39	37	39
GC (%)	50.71	50.74	50.74	50.73
Per base quality				
# N's per 100 kbp	24.59	0	17.55	94.19
# N's	1175	0	800	4288
BUSCO completeness				
Complete BUSCO (%)	98.65	98.65	98.65	98.65
Partial BUSCO (%)	0	0	0	0

Aligned to "Ecoli_K12_MG1655" | 4 641 652 bp | 1 fragment | 50.79% G+C | 4639 genomic features

Alignment-based statistics	ABySS	MEGAHIT	SPAdes	Velvet
Genome fraction (%)	98.661	98.424	98.113	97.997
Duplication ratio	1.043	1	1	1
# genomic features	4525 + 75 part	4511 + 64 part	4489 + 50 part	4486 + 56 part
Largest alignment	248 481	235 933	285 096	264 944
Total aligned length	4 776 214	4 568 317	4 553 809	4 550 150
NGA50	69 801	122 647	133 309	112 446
LGA50	21	14	12	14
Misassemblies				
# misassemblies	4	0	0	4
Misassembled contigs length	231 767	0	0	435 515
Per base quality				
# mismatches per 100 kbp	2.09	2.69	1.03	3.19
# indels per 100 kbp	0.57	1.31	0.29	1.98
# N's per 100 kbp	24.59	0	17.55	94.19
Statistics without reference				
# contigs	176	95	92	90
Largest contig	248 481	235 933	285 196	264 944
Total length	4 777 853	4 571 292	4 557 363	4 552 266
Total length (>= 1000 bp)	4 757 929	4 562 458	4 548 710	4 544 453
Total length (>= 10000 bp)	4 562 801	4 478 614	4 466 223	4 475 223
Total length (>= 50000 bp)	3 248 113	3 833 793	3 812 315	3 817 904

Figure 1. WebQUAST text reports for *E. coli* assemblies in the (**A**) reference-free and (**B**) reference-based evaluation mode. Unless otherwise noted, all statistics are based on contigs of size $\geq$ 500 bp (the default cut-off). Heatmap highlights the best value in each row which could be the largest or the smallest number depending on the quality metric. Heatmap is not used for *# contigs* and *GC (%)* due to the ambiguity of these metrics trends.

provided assemblies (Figure 1A, Supplementary Figure S1). The heatmaps help to detect the best-performing tools in each category.

Figure 1A shows that there is no single winner in all metrics. Compared to three other methods, ABySS produced the largest (4.8 Mbp vs 4.6 Mbp) but also the most fragmented assembly (176 contigs vs 90-95 for Velvet, SPAdes, and MEGAHIT). SPAdes assembled larger contigs on average (the best N50, N90 and auN, the area under the Nx curve, values with Velvet and MEGAHIT being close runner-ups) and has the largest contig overall (285 vs 265, 248, and 236 kbp for Velvet, ABySS, and MEGAHIT). The MEGAHIT assembly does not contain uncalled bases (“N”) while Velvet has the most of them (94 per 100 kbp). All four assemblies are equally complete in terms of fully assembled representative bacterial single-copy orthologs (98.7% of the BUSCO genes). The average G + C content of all assemblies (50.7%) perfectly matches the expected range for *E. coli* (50.4-50.8% (40)) indicating the likely absence of contaminants in the dataset. This hypothesis is further supported by the GC plot (Supplementary Figure S1D), though we cannot exclude a presence of an organism with similar G + C content.

Use Case 2: reference-based evaluation

A reference genome enables accurate and versatile evaluation by WebQUAST in all four quality categories: contiguity, correctness, completeness, and contamination. In this mode, the tool reports more than 60 quality metrics accompanied by eight assessment plots and two Icarus viewers (Figure 1B, Figure 2A, Supplementary Figures S2-S4). By default, WebQUAST displays only 18 key metrics and hides the rest behind the Extended report button (Figure 1B).

As in Use Case 1, there is no undisputed best assembly in Figure 1B. However, we can now investigate some quality categories in more detail. The increased Duplication ratio for ABySS (1.04 vs 1.00 for the rest assemblers) indicates that this method assembled many genomic regions more than once. Still, ABySS assembled the highest percentage of the genome (98.7 vs 98.0-98.4% for Velvet, SPAdes, and MEGAHIT) but its leadership is not as evident as it appeared when we compared the total assembly lengths. SPAdes and ABySS have the best per-base quality with SPAdes being twice better as the runner-up (1.0 vs 2.1 mismatches and 0.3 vs 0.6 indels per 100 kbp). MEGAHIT and SPAdes made no large assembly errors, while Velvet and ABySS have four misassemblies each. Though, the largest contigs in all four assemblies are error-free since their lengths exactly match the largest alignments. The Icarus viewer can be used for deep inspection of the misassembly locations (Figure 2, Supplementary Figures S4).

Use Case 3: evaluation based on a close reference

The true reference genome is rarely known in real studies but a close reference could often be available. Here we used W3110, another *E. coli* K-12 substrain, as an example of a close reference (Figure 2B, Supplementary Figures S5-S7). Naturally, the absolute values of many alignment-based metrics, such as lengths of misassembled and unaligned contigs, substantially deteriorated due to the actual differences between the sequenced organism and the provided reference genome. However, they are still useful for determining the best assembly among available options.

Figure 2 highlights the substantially increased number of misassemblies compared to the evaluation based on the true reference genome (49 vs 8 extensive misassemblies in total). However, a closer look at the misassembly locations, suggests that almost all of them are the same in all assemblies which likely means they are true structural variations rather than assembly errors and can be ignored for evaluation purposes (Figure 2B and Supplementary Figures S7). Though, we cannot exclude the possibility that several assemblers made the same error in a complex genomic region, especially if we

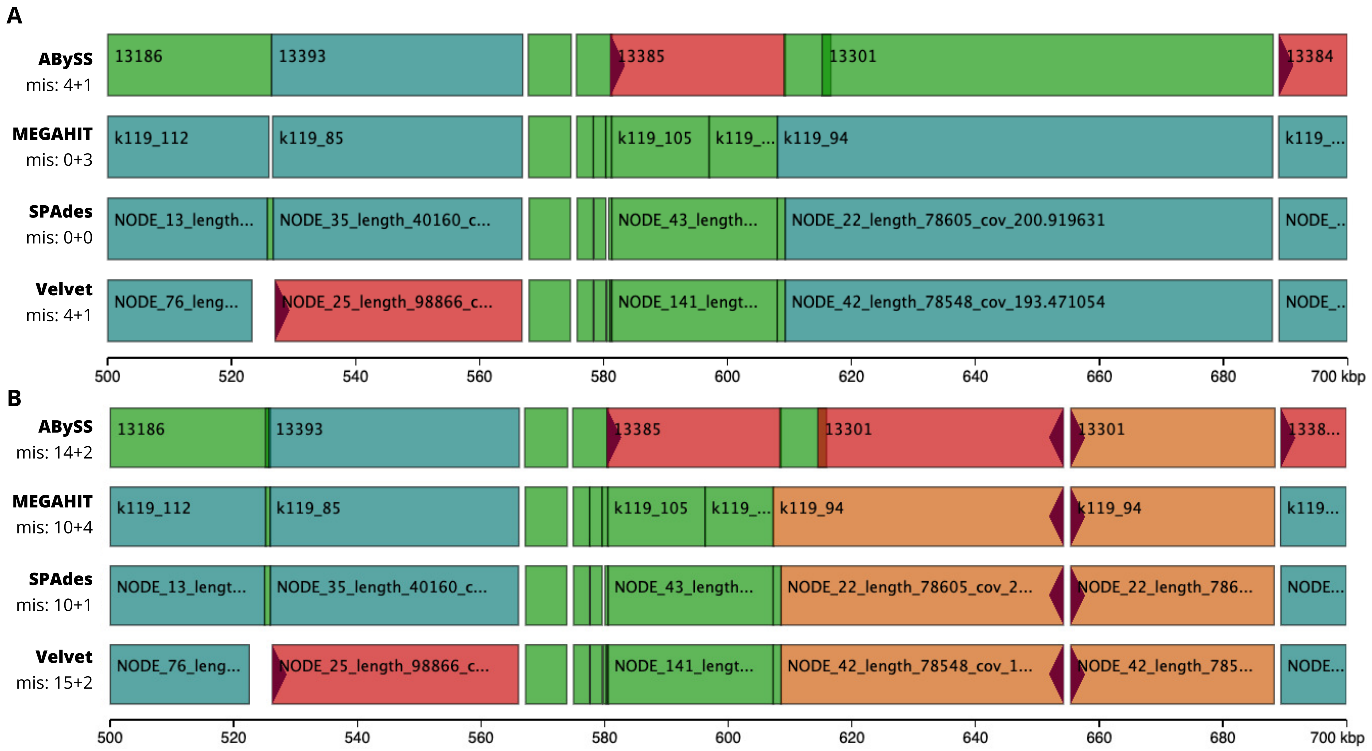


Figure 2. Icarus viewers for *E. coli* assemblies aligned against (**A**) the reference genome matching the dataset and (**B**) a close reference. The reference regions between 0.5 Mbp and 0.7 Mbp are shown. *mis:* $X + Y$ stands for the total number of extensive (X) and local (Y) misassemblies per assembly. Correctly assembled contigs are colored green and aquamarine (if longer than 10 kbp and similar in at least three assemblies), and fragments of misassembled contigs are colored pink and orange (if similar in at least three assemblies). Red triangles designate the sides of alignment breakpoints for misassembled contigs. Contig names are shown for contigs of sufficient size.

compare tools inspired by the same computational approach such as the de Bruijn graph-based assembly (41).

CONCLUSION

Selecting the best – or, more precisely, the most suitable – genome assembly is crucial for downstream analysis. While many post-processing steps, such as structural and functional annotation (42) or genome mining (43), have been available online for years, the assembly validation step is still mainly done with the Linux-based command-line tools. Here we presented WebQUAST, a web server for genome assembly evaluation, that greatly facilitates this task for users with any operating system and computational background and helps them to make an informed choice. Since our tool is suitable for any organism and sequencing technology, we expect it would benefit the broad genomics community. Furthermore, WebQUAST is already incorporated in several bioinformatics massive online open courses (MOOCs), so we hope it would also help to educate the future generation of researchers.

DATA AVAILABILITY

WebQUAST is freely available at http://cab.cc.spbu.ru/quast/. The source code for the server is at https://github.com/ablab/quast-website and for the core QUAST tool is at https://github.com/ablab/quast. The sequencing data for *E. coli* K-12 MG1655 dataset is available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession number ERR008613. The *E. coli* strain K-12 reference genomes and gene annotations are available from NCBI under accession numbers NC_000913.3 and AP009048.1 for substrains MG1655 and W3110, respectively. The ABySS, MEGAHIT, SPAdes, and Velvet assemblies generated in this study and their interactive evaluation reports are available from the WebQUAST front page and in Zenodo at https://doi.org/10.5281/zenodo.7863703.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank St. Petersburg Academic University for hosting WebQUAST in 2012-2020 and the Computer Center of Research Park of St. Petersburg State University for hosting it in 2020-2023.

FUNDING

Funding for open access charge: Helmholtz Centre for Infection Research.

Conflict of interest statement. None declared.

REFERENCES

1. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018) The third revolution in sequencing technology. *Trends in Genetics,* **34**(9), 666–681.
2. Sohn, J.-i. and Nam, J.-W. (2018) The present and future of de novo whole-genome assembly. *Briefings in bioinformatics,* **19**(1), 23–40.
3. Lloret-Villas, A., Bhati, M., Kadri, N. K., Fries, R., and Pausch, H. (2021) Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC genomics,* **22**(1), 1–17.
4. Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research,* **22**(3), 557–567.
5. Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013) REAPR: a universal tool for genome assembly evaluation. *Genome biology,* **14**, 1–10.
6. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., and Chong, Z. (2021) Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology,* **22**(1), 1–21.
7. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* **31**(19), 3210–3212.
8. Seppey, M., Manni, M., and Zdobnov, E. M. (2019) BUSCO: assessing genome assembly and annotation completeness. *Gene prediction: methods and protocols,* pp. 227–245.
9. Parra, G., Bradnam, K., and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,* **23**(9), 1061–1067.
10. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29**(8), 1072–1075.
11. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics,* **34**(13), i142–i150.
12. Chow, W., Brugger, K., Caccamo, M., Sealy, I., Torrance, J., and Howe, K. (2016) gEVAL—a web-based browser for evaluating genome assemblies. *Bioinformatics,* **32**(16), 2508–2510.
13. Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M., and Hufford, M. B. (2020) GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC genomics,* **21**(1), 1–9.
14. Nishimura, O., Hara, Y., and Kuraku, S. (2017) gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics,* **33**(22), 3635–3637.
15. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* **34**(18), 3094–3100.
16. Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics,* **20**(16), 2878–2879.
17. Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V., and Gurevich, A. (2016) Icarus: visualizer for de novo assembly evaluation. *Bioinformatics,* **32**(21), 3321–3323.
18. Gardner, P. P., Paterson, J. M., McGimpsey, S., Ashari-Ghomi, F., Umu, S. U., Pawlik, A., Gavryushkin, A., and Black, M. A. (2022) Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. *Genome biology,* **23**(1), 1–13.
19. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience,* **1**(1), 2047–217X.
20. Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., and Lam, T.-W. (2016) MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods,* **102**, 3–11.
21. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020) Using SPAdes de novo assembler. *Current protocols in bioinformatics,* **70**(1), e102.
22. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology,* **19**(5), 455–477.
23. Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research,* **18**(5), 821–829.
24. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research,* **20**(2), 265–272.
25. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* **31**(10), 1674–1676.
26. Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., et al. (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome research,* **27**(5), 768–777.
27. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome research,* **19**(6), 1117–1123.
28. Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* **28**(11), 1420–1428.
29. Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2010) IDBA–a practical iterative de Bruijn graph de novo assembler. In *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings 14* Springer pp. 426–440.
30. Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences,* **108**(4), 1513–1518.
31. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research,* **18**(5), 810–820.
32. Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., and Salzberg, S. L. (2017) Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research,* **27**(5), 787–792.
33. Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013) The MaSuRCA genome assembler. *Bioinformatics,* **29**(21), 2669–2677.
34. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012) Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology,* **13**(12), 1–13.
35. Boisvert, S., Laviolette, F., and Corbeil, J. (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies.
36. Simpson, J. T. and Durbin, R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome research,* **22**(3), 549–556.
37. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* **30**(15), 2114–2120.
38. Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J., and Salzberg, S. L. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics,* **29**(14), 1718–1725.
39. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., and Team, B. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods,* **15**(7), 475–476.
40. Mann, S. and Chen, Y.-P. P. (2010) Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics,* **95**(1), 7–15.
41. Pevzner, P. A., Tang, H., and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the national academy of sciences,* **98**(17), 9748–9753.
42. Humann, J. L., Lee, T., Ficklin, S., and Main, D. (2019) Structural and functional annotation of eukaryotic genomes with GenSAS. *Gene prediction: methods and protocols,* pp. 29–51.
43. Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Van Wezel, G. P., Medema, M. H., and Weber, T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research,* **49**(W1), W29–W35.