OXFORD

Briefings in Bioinformatics, 00(00), 2021, 1–11

doi: 10.1093/bib/bbab102
Advance Access Publication Date: 00 Month 0000
Case Study

Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants

Rashedul Islam, Rajan Saha Raju, Nazia Tasnim, Istiak Hossain Shihab, Maruf Ahmed Bhuiyan, Yusha Araf and Tofazzal Islam

Corresponding authors: Rashedul Islam, Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V5Z 4S6, Canada. Tel.: +1778-889-4146; E-mail: rashed1@student.ubc.ca; Tofazzal Islam, Institute of Biotechnology and Genetic Engineering, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur 1706, Bangladesh. E-mail: tofazzalislam@yahoo.com

Abstract

Background: Coronavirus Disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has become a global pandemic following its initial emergence in China. SARS-CoV-2 has a positive-sense single-stranded RNA virus genome of around 30Kb. Using next-generation sequencing technologies, a large number of SARS-CoV-2 genomes are being sequenced at an unprecedented rate and being deposited in public repositories. For the de novo assembly of the SARS-CoV-2 genomes, a myriad of assemblers is being used, although their impact on the assembly quality has not been characterized for this virus. In this study, we aim to understand the variabilities on assembly qualities due to the choice of the assemblers.

Results: We performed 6648 de novo assemblies of 416 SARS-CoV-2 samples using eight different assemblers with different *k*-mer lengths. We used Illumina paired-end sequencing reads and compared the assembly quality of those assemblers. We showed that the choice of assembler plays a significant role in reconstructing the SARS-CoV-2 genome. Two metagenomic assemblers, e.g. MEGAHIT and metaSPAdes, performed better compared with others in most of the assembly quality metrics including, recovery of a larger fraction of the genome, constructing larger contigs and higher N50, NA50 values, etc. We showed that at least 09% (259/2873) of the variants present in the assemblies between MEGAHIT and metaSPAdes are unique to one of the assembly methods.

Conclusion: Our analyses indicate the critical role of assembly methods for assembling SARS-CoV-2 genome using short reads and their impact on variant characterization. This study could help guide future studies to determine the best-suited assembler for the de novo assembly of virus genomes.

Rashedul Islam is a PhD candidate in Bioinformatics at the University of British Columbia, Canada. He is the recipient of the Canadian Institutes of Health Research Training Award to attend graduate school.

Rajan Saha Raju is a Machine Learning Engineer at REVE Chat. B.Sc. on Computer Science and Engineering from the Shahjalal University of Science and Technology.

Nazia Tasnim is an undergraduate final year student, majoring in Computer Science and Engineering at the Shahjalal University of Science and Technology. Her research interests include bioinformatics, DeepLearning and NLP.

Md. Istiak Hossain Shihab is an undergraduate student at the Department of Computer Science and Engineering, Shahjalal University of Science and Technology. His primary research areas include bioinformatics, deep learning and data mining.

Maruf Ahmed Bhuiyan is a physician and currently pursuing his Doctor of Medicine (MD) in Virology at Bangabandhu Sheikh Mujib Medical University. He has a vested interest in interdisciplinary research.

Yusha Araf is an undergraduate final year student, majoring in Genetic Engineering and Biotechnology at the Shahjalal University of Science and Technology. His areas of interests are public health, immunoinformatics, molecular biology and medical genetics.

Tofazzal Islam is a Professor and the Director of the Institute of Biotechnology and Genetic Engineering, Bangabandhu Sheikh Mujibur Rahman Agricultural University.

Submitted: 15 January 2021; **Received (in revised form):** 6 March 2021

Highlights

- Assemblers showed marked differences in de novo assembly of SARS-CoV-2 genome.
- Two metagenomic assemblers, e.g. MEGAHIT and metaSPAdes, constructed a larger fraction of the genome compared with other assemblers.
- At least 09% (259/2873) of the variants present in the assemblies between MEGAHIT and metaSPAdes are unique to one of the assemblers.

Key words: ARS-CoV-2; COVID-19; de novo assembly; benchmarking; short-read; variant

INTRODUCTION

SARS-CoV-2 is the seventh member of the Coronaviridae family to infect humans, which is responsible for the current COVID-19 pandemic [1]. This virus is ravaging the world with more than 1.8 million deaths in the year 2020 [2]. To understand its pathophysiological mechanism, mutation pattern, epidemiological tracing and transmission pathways, the single-stranded RNA genome of SARS-CoV-2 has been sequenced in different countries. Around 50 000 SARS-CoV-2 genomic sequences have been submitted to NCBI Nucleotide records and Nextstrain database since the first whole-genome was sequenced in January 2020 [3, 4]. This unprecedented speed of genome sequencing was possible due to the advancement in sequencing technologies and the availability of open-source bioinformatics tools.

By the end of 2020, 84% (140,837/168,547) of SARS-CoV-2 sequencing runs deposited on NCBI's Sequence Read Archive (SRA) were generated using Illumina short-read sequencing technology [5]. Along with short-read sequencing technologies, long-read sequencing technologies were also used in combination with short reads or alone to decipher SARS-CoV-2 genome [6]. Many assembly tools (assemblers) are publicly available to assemble the genome from short reads. These assemblers use a combination of, or solely, these methods: De Bruijn graph, Overlay Layout Consensus and greedy graph method. The quality of the virus genome assembly varies depending on the assembler of choice, genome composition, depth of sequencing, sample preparation, etc. [7]. Due to the rapid pace of SARS-CoV-2 genome sequencing, use of different sequencing assays and availability of multiple assemblers, the assemblers need to be benchmarked and updated for the de novo assembly of the SARS-CoV-2 genome.

RNA viruses naturally accumulate random genetic variations during the course of infection [8]. Mutations in the genomes are used to track the transmission of SARS-CoV-2 virus in which closely related genomes are anticipated to be closely related infections [3]. Phylogeny of genomes is used to cluster similar clades where the genetic diversity solely depends on the variants present on the genomes [9]. However, assemblers could introduce erroneous base(s) in the sequence due to their error correction method, quality filtering or parameter selections [10–12]. Assemblers have their unique error profiles, and therefore, genomic variants also vary by assemblers [13]. Here, we sought to find out the instances of genomic variants that were solely driven by the choice of assemblers.

To date, the degree of variation in assembly qualities among different assemblers has not been reported for SARS-CoV-2. In this study, we present a comprehensive investigation on the performance of assemblers for SARS-CoV-2 genome assembly with publicly available Illumina paired-end datasets. To compare assemblers, we used different assembly quality matrices, e.g. percentage of genome recovery, largest contig, total assembly length, N50, NA50, L50, LA50, etc. We further called the variants from the assembled contigs. We showed that the number of variants occurring in the contigs varies significantly among different assemblers and there are discordances in variants between assemblers for the same sample.

Table 1. Summary of the viral RNA dataset of SARS-CoV-2 available by 14th June 2020

Assay type	Library source	Library layout	# Libraries	# Selected
AMPLICON	VIRAL RNA	PAIRED	5254	82
OTHER	VIRAL RNA	PAIRED	68	66
RNA-Seq	VIRAL RNA	PAIRED	682	75
Targeted-Capture	VIRAL RNA	PAIRED	194	93
WGA	VIRAL RNA	PAIRED	1061	100

RESULTS

De novo assemblers showed marked differences in assembly quality

Illumina paired-end data for the SARS-CoV-2 genomes have been collected from different assay types, e.g. amplicon, whole genome amplification (WGA), RNA-Seq, targeted-capture and other (Figure 1A). We randomly selected 100 libraries for each assay type and then removed the libraries that did not pass the sanity check (Figure 1B, Table 1). To evaluate assembly quality among different assemblers, we selected four de novo metagenome assemblers (e.g. metaSPAdes, MetaVelvet, MEGAHIT and Ray Meta) and four de novo genome/transcriptome assemblers (e.g. ABySS, Velvet, SPAdes and Trinity) (Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). To rule out the influence of the choice of *k*-mer lengths on assembly quality, we used three *k*-mer lengths (21, 63 and 99) for the assemblers requiring a fixed *k*-mer length.

Genome fraction recovery was highly variable across the different assemblers (assembly quality terminologies [14]; Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). Most of the assemblers (e.g. ABySS-K21,63,99, Ray Meta-K21,63,99, SPAdes, Trinity) recovered a larger fraction (median > 90%) of the genome,

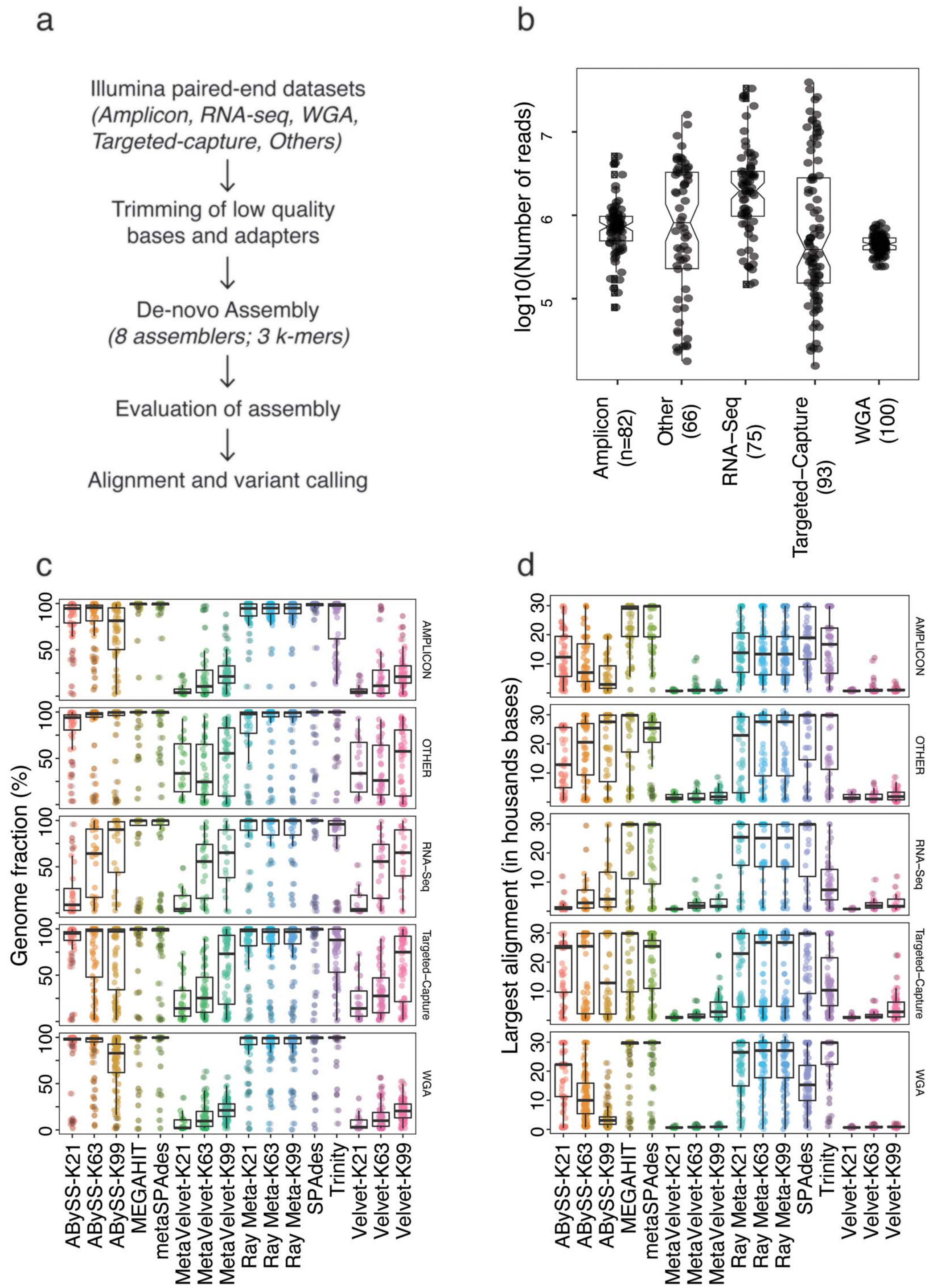

Figure 1. Comparison of assemblers for different sequencing assays. (A) Experimental strategies to assemble SARS-CoV-2 genomes using different assemblers and calling variants from the assembled contigs. From different sequencing assay types, samples were randomly selected for assembly and subsequent analysis. (B) Total number of reads for each sample across different assay types. (C) Fraction of SARS-CoV-2 genome assembled by different assemblers. (D) The largest continuous alignment of the assemblies produced by different assemblers.

although there was high variability in recovering the genome across the samples (Figure 1C). In contrast, MEGAHIT and metaSPAdes recovered almost the entire genome (median of ≥99.7%) with low variabilities. Velvet and Metavelvet recovered a lower fraction (median < 30%) of the genome compared with other assemblers invariably across assay types, despite an increase in genome fraction recovery with increasing *k*-mer length (Figure 1C). We confirmed that the recovery of the fraction of the genome by different assemblers was not affected by the sequencing depth of the libraries in different assay types (Figure 2A). For all assemblers, genome fractions were invariably recovered across the span of the sequencing depth except for Targeted-Capture assay which showed a negative correlation (Spearman: −0.24). Consistent with the recovery of the genome, the larger uninterrupted alignment of assemblies to the reference genome were obtained by MEGAHIT (median of >29 000 bp) and metaSPAdes (median of >27 000 bp) across all assay types (Figure 1D). In contrast, Velvet and MetaVelvet showed relatively lower contiguous alignments (median of <1100 bp). MEGAHIT and metaSPAdes also generated the highest number of assemblies (∼40% of the samples) with 90% of the genome covered by a single contig, whereas ABySS was unable to generate longer contiguous sequences at different *k*-mer lengths (Figure 2B).

About 87% (4291/4912) of the assemblies did not produce any misassembled events. The duplication ratio was higher in Trinity (median of 1.1005) compared with other assemblers (Supplementary Figure S1A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/). The median duplication ratio for all assemblies is 1.005 where 25% (1247/4912) of the assemblies had duplication ratio equal to 1.

MEGAHIT and metaSPAdes performed better in most of the assembly quality matrices

N50 is the minimum contig length needed to construct 50% of the genome, where contigs are sorted by their lengths. There were large variations among the assemblers for N50. All assemblers performed poorly on the RNA-seq assay type (Figure 3A). MEGAHIT and metaSPAdes had the highest N50 values (median of >21 000 bp) compared with other assemblers (median of ≤10 000 bp) across all samples. NA50 is an improved matrix of the N50 contig length which breaks contigs into aligned blocks at misassembly events and removes all unaligned bases. MEGAHIT and metaSPAdes were able to generate larger NA50 values in all assay types compared with other assemblers (Figure 3B). Metavelvet and Velvet were unable to produce large N50 as well as NA50 values across different assay types. MEGAHIT and metaSPAdes produced lower L50 (median = 1) and LA50 (median = 1) values compared with other assemblers across different assay types (Figure 3C and D). We observed MEGAHIT and metaSPAdes outperformed other assemblers in several other quality matrices, e.g. contig lengths, number of Ns, percent overlap with genes and N75, NA75, L75, LA75 values (Supplementary file 1 and 2, see Supplementary Data available online at http://bib.oxfordjournals.org/).

SARS-CoV-2 genome assembly contiguity breaks at the repeat region

We investigated the simple tandem repeat of 585 bp in the SARS-CoV-2 genome located at the 3′-end (MN908947.3:29 870–29 903). Most of the assemblers failed to assemble the repeat region and more than 100 bp gap was created at the 3′-end in different assemblies. To identify the presence of similar assembly gaps in the assemblies, we binned the genome into 50 bp non-overlapping windows and counted the number of bases assembled in each bin throughout the genome (Figure 4A). We compared the assemblies of top-performing four assemblers, e.g. MEGAHIT, metaSPAdes, Trinity and ABySS-K63 for 392 samples across assay types. Gaps in the assemblies are shown in red color and assembled regions are in gray in the heatmap. We found that in addition to gaps in the 3′-end, there were assembly gaps (around 50 bp size) at the 5′-end of the genome across the assemblies. Besides gaps in the 5′ and 3′-ends, we did not observe other consistent gaps in those assemblies.

Some samples showed higher or lower recovery of the genome fraction independent of the assemblers we used. To investigate this, we analyzed four samples consisting of two better and two worse quality assemblies in terms of the fraction of genome recovered across four assemblers. We visualized the read densities on the genome along with gaps (red) and assembled regions (grey) (Figure 4B). The samples (SRR11903415 and SRR12182155) with better quality assembly had uniform read coverage throughout the genome, whereas other two samples (SRR11783612 and SRR11954291) showed gapped assembly independent of assemblers due to lack of read coverage at the gapped regions.

Influence of *k*-mer length on assembly quality

All eight assemblers we tested use graph-based methods where the choice of *k*-mer length affects the contiguity of an assembly [15, 16]. Four of the de novo assemblers (e.g. ABySS, Velvet, MetaVelvet and Ray Meta) we used require a single *k*-mer length. To analyze the variability in assembly quality across 416 samples, we used three different *k*-mer lengths (i.e. 21, 63 and 99). With different *k*-mer lengths ABySS, Velvet and MetaVelvet showed variabilities in assembly quality but Ray Meta did not show larger variabilities on average (Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). ABySS performed relatively better with smaller *k*-mer lengths in recovering fraction of genome, largest alignment, N50, NA50, LA50 values with lower misassemblies and duplication ratio. Velvet and MetaVelvet performed better with higher *k*-mer lengths in recovering fraction of genome, largest alignment, N50, NA50 values at the cost of higher misassemblies and duplication ratios. However, L50 values improved overall with the increasing *k*-mer lengths for all four assemblers.

Variant calling from de novo assemblies varies between assemblers

We aligned the assembled contigs to the reference genome and called variants (see Methods). For 416 samples, the assemblers produced different number of variants which are expectedly correlated (Pearson = 0.73) to the fraction of genome recovered (Supplementary Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). To investigate the occurrences of variants exclusively present in an assembler, we compared the variants identified from the two best-performing assemblers, e.g. MEGAHIT and metaSPAdes. Among all the variants identified in the common genomic regions assembled by both MEGAHIT and metaSPAdes, 92% of the variants overlapped between two assemblers and 25% (95/385) of the samples had at least one assembler specific variant (Figure 5A). To further understand the consequence of assembler specific variants in the biologically important genomic features, we analyzed the Spike (S) gene locus and found that 06% (23/385) of the samples contain variants that are unique to MEGAHIT or metaSPAdes (Figure 5B). Here, we showed the example of variants in the Spike locus which are concordant (ERR4208998) or discordant (SRR11783589) between MEGAHIT and metaSPAdes at the common assembly regions. In addition, there are variant differences between MEGAHIT and metaSPAdes due to assembly gaps in one of the assemblies. For example, in SRR12182180, a variant occurs in the metaSPAdes assembly but not in the MEGAHIT assembly

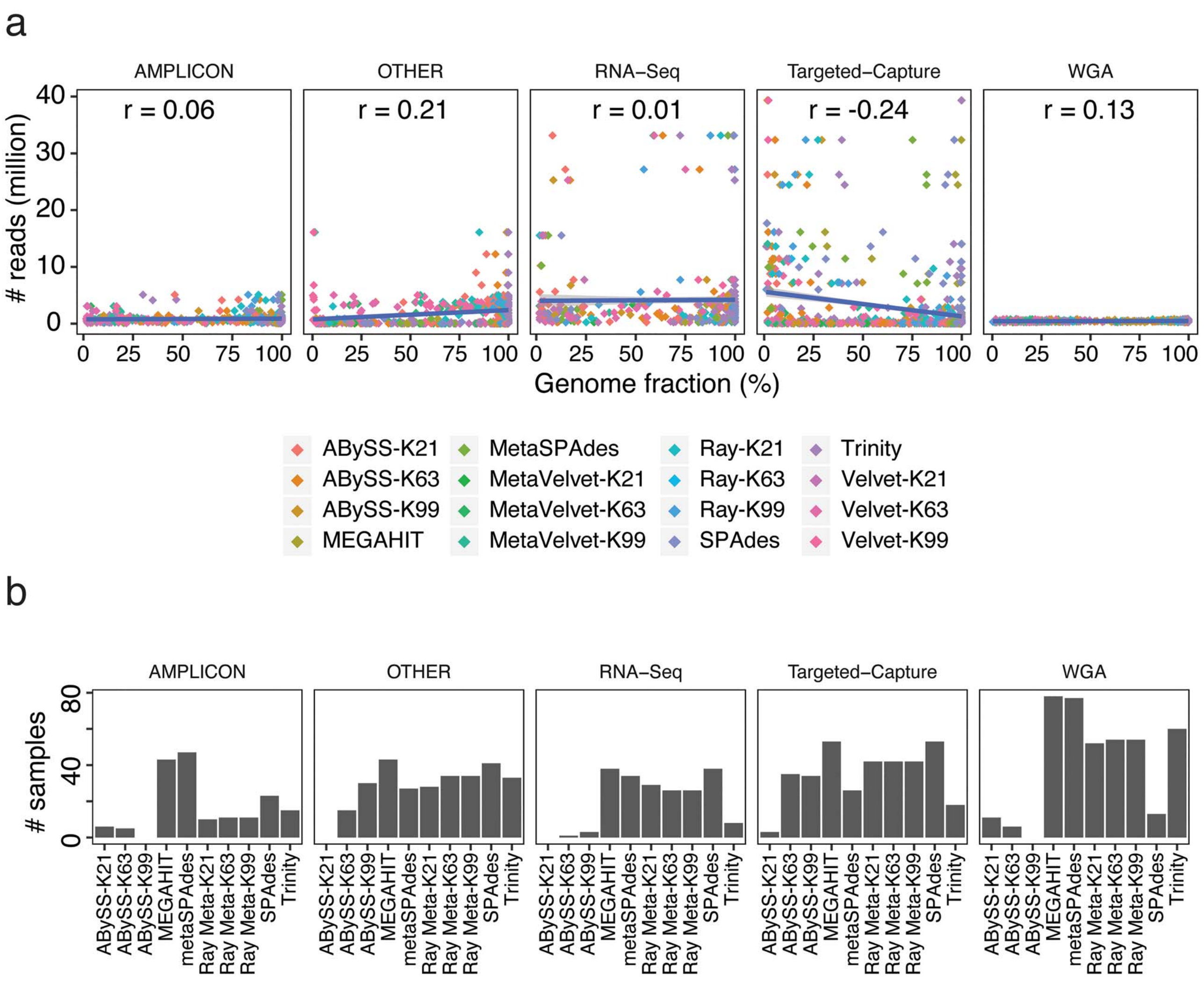

Figure 2. Assessment of sequencing and assembly quality. (a) Pearson correlation coefficient between the number of reads and fraction of genome assembled in respect to assay types. (b) Assemblies with 90% of the genome covered by a single contig.

because of the assembly gap at this location (Figure 5B). Variants present in both raw reads and in assemblies (Figure 5C) but in many instances raw reads do not contain those variants and suggesting that spurious variants arise due to assembly errors (Supplementary Figure S4A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/). This highlights the importance of correcting assembler specific spurious variants before functional characterization of the variants and using de novo assemblies for phylogeny construction or other pan-genome analyses.

Computational performance by different assemblers

Scarcity of computational resources also forces us to pay special attention to space and time complexity during de novo assembly of genomes. For calculating the Central Processing Unit (CPU) time consumption and Random Access Memory (RAM) usage, we randomly selected 17 amplicon libraries that have around 1 million paired-end reads. We utilized 4 cores and 8 threads on a dedicated computer for all the assemblers. Assembly completion time has been adopted as time consumed in wallclock CPU seconds (Figure 6A). Trinity and MetaVelvet-K99 consumed the lowest time among all assemblers with the median values of 0.73 and 4.1 s, respectively (Figure 6A). Ray Meta and metaSPAdes took the longest time. With regards to RAM usage, we observed that Trinity and MEGAHIT required the least amount of RAM with the median values of 12.5 and 13.0%, respectively (Figure 6B). metaSPAdes required the highest amount (31.9%) of RAM compared with other assemblers. Notably, the choice of *k*-mer length for the same assembler had little to no effect on the CPU time and RAM usage measurements.

DISCUSSION

In this study, we compared 16 assembler variations using eight de novo assemblers for the benchmarking of the genome assembly quality of the SARS-CoV-2 virus. We observed two metagenomic assemblers, e.g. MEGAHIT and metaSPAdes outperformed other assemblers in regards to the genome fraction recovery, largest contig length, N50 length, NA50 length, L50 and LA50 contig number. The fraction of genome recovery could be 10-folds different between assemblers, e.g. MEGAHIT (99%)

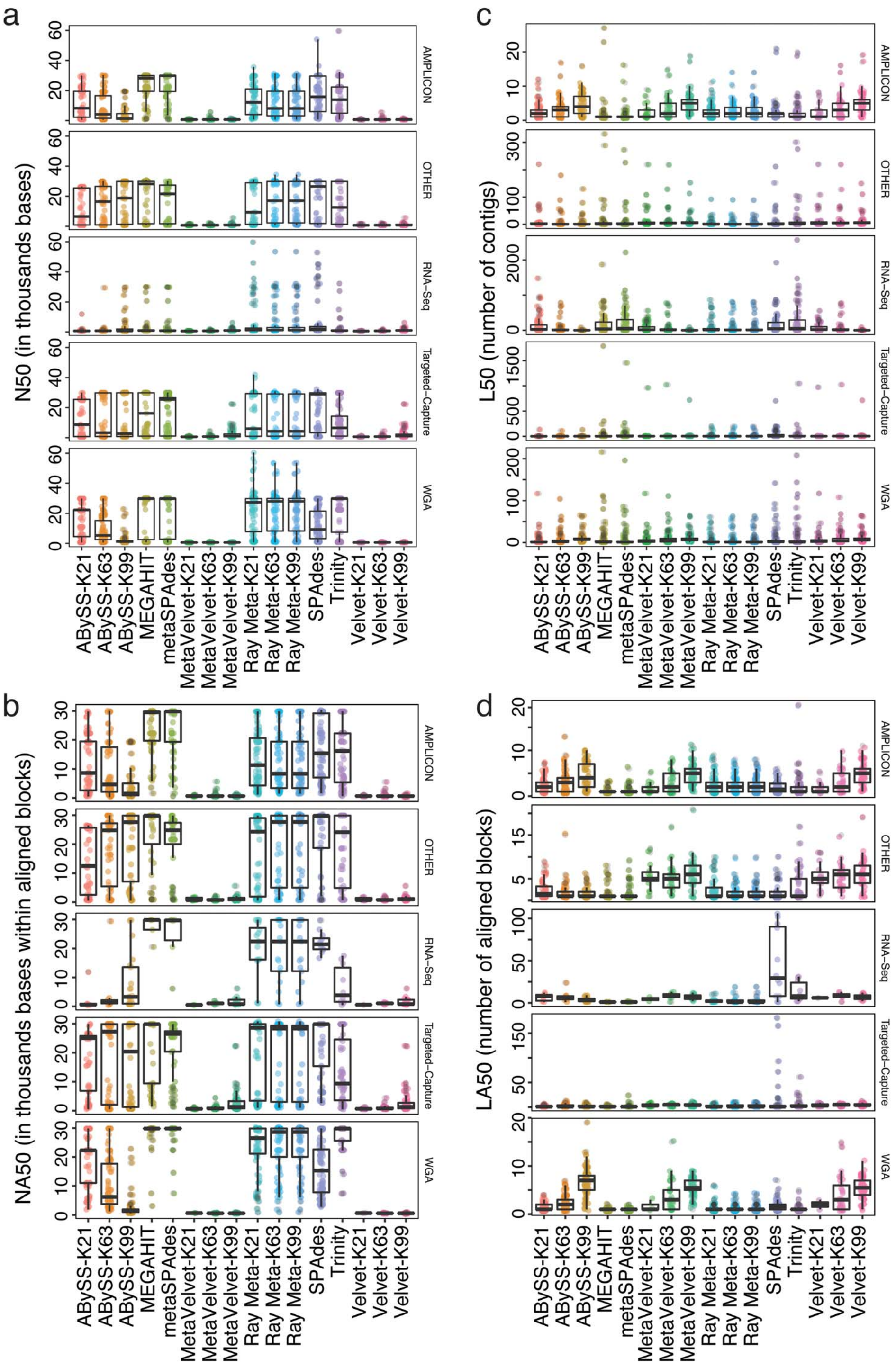

Figure 3. Comparison of contigs and aligned genomic blocks. (a, b) Length of the smallest contig (N50) and aligned block (NA50) at 50% of the total genome length. (c, d) Minimum number of contigs (L50) and aligned blocks (LA50) at 50% of the total genome length.

versus MetaVelvet-K21 (10%). Although all eight assemblers used the graph-based method for de novo assembly, the differences we observed are due to the variations in their implementation, error correction, quality thresholds and choice of other parameters. Despite better performances by the two metagenomic assemblers, the entire viral genome was not assembled in most cases, especially at the termini of the genome. Therefore, there is a need to develop newer assembly methods specially designed to assemble complete viral genomes. The SARS-CoV-2 virus genome could also be assembled by aligning the reads to the reference genome or using a reference guided assembly.

Single nucleotide variants and short insertions and deletions vary by the assemblers, possibly correlated to the

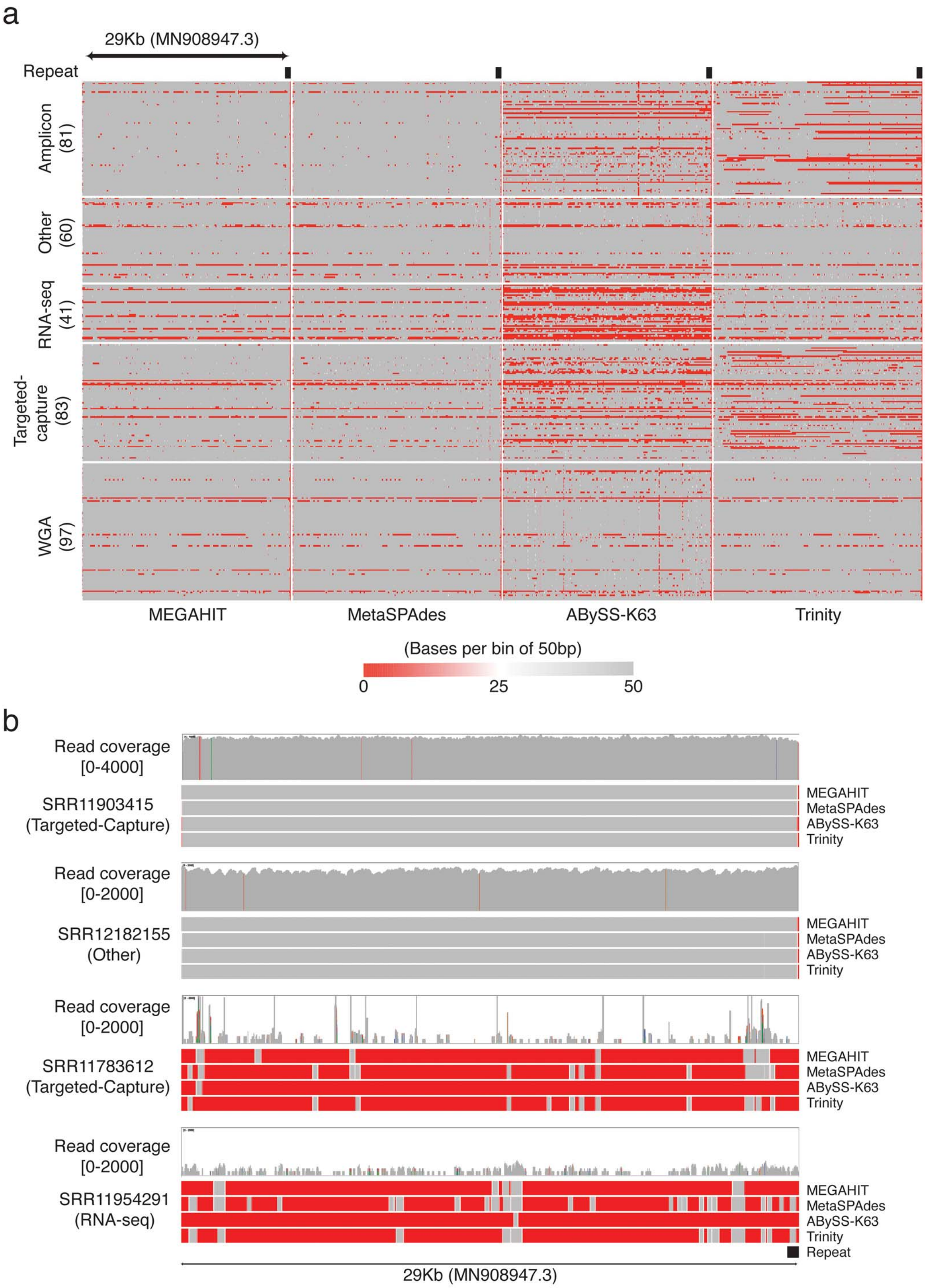

Figure 4. Analysis of assembly contiguity and gap. (A) SARS-CoV-2 genome was binned into 50 bp non-overlapping windows. For each bin, a number of bases assembled were plotted in the heatmap using a continuous color scale. For each bin, 50 bp assembly is shown in gray and an assembly gap in red. Samples with successful assemblies for all four assemblers were included here. (B) Using a similar binning approach as in (A), the contiguity (gray) or gaps (red) of four assemblers are shown. Two samples (SRR11903415 and SRR12182155) with most contiguous assembly and two samples (SRR11783612 and SRR11954291) with gapped assembly were plotted with sequencing read coverage across the SARS-CoV-2 genome.

'aggressiveness' of the assembler [13]. Differences in variants introduced by assemblers may have an impact on downstream comparative genomic applications, such as pan-genome comparison or constructing phylogenetic tree using de novo genome assemblies. Often assembler specific variants are the result of assembly errors. Such assembly errors could be resolved by using post-assembly genome improvement pipelines that use local assembly and/or raw read alignment to the erroneous variants [11].

To discover novel viruses, the sequence of complete viral genomes is inevitable rather than fragmented viral contigs. Low read coverage and genomic repeats resulted in assemblies

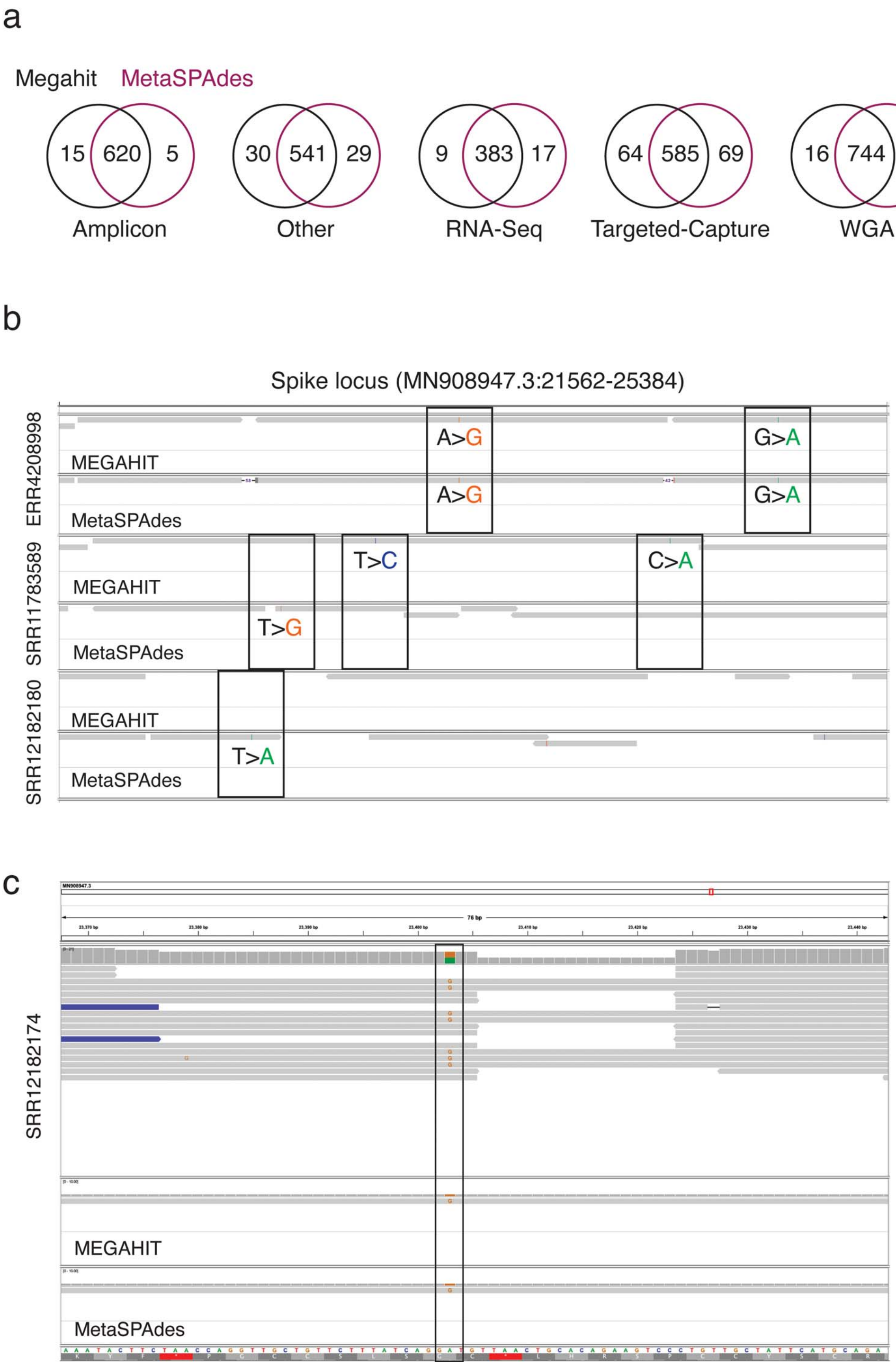

Figure 5. Differences in genomic variants between two assemblers. (A) Overlap of variants within the common genomic regions between MEGAHIT and metaSPAdes assemblies for different assay types. Single nucleotide variants and short insertions/deletions are included here. (B) Examples of concordant (ERR4208998) and discordant (SRR11783589, SRR12182180) variants between MEGAHIT and metaSPAdes are shown in genome browser. Contigs are represented in gray bars and variant nucleotides are highlighted. (C) Example of variant presents in both assembly and raw reads.

with poor genome recovery independent of assemblers. Recent benchmarking studies reported metagenomic assemblers resulted in the relatively higher contiguous viral assemblies using viral metagenomic data [7, 17]. Our analysis for SARS-CoV-2 data, using different sequencing assay types, identified two metagenome assemblers, e.g. MEGAHIT and MetaSPAdes addressed the challenges of virome data better than other assemblers. Our benchmarking data for SARS-CoV-2 genome can be used to choose suitable de novo assemblers for similar genomes.

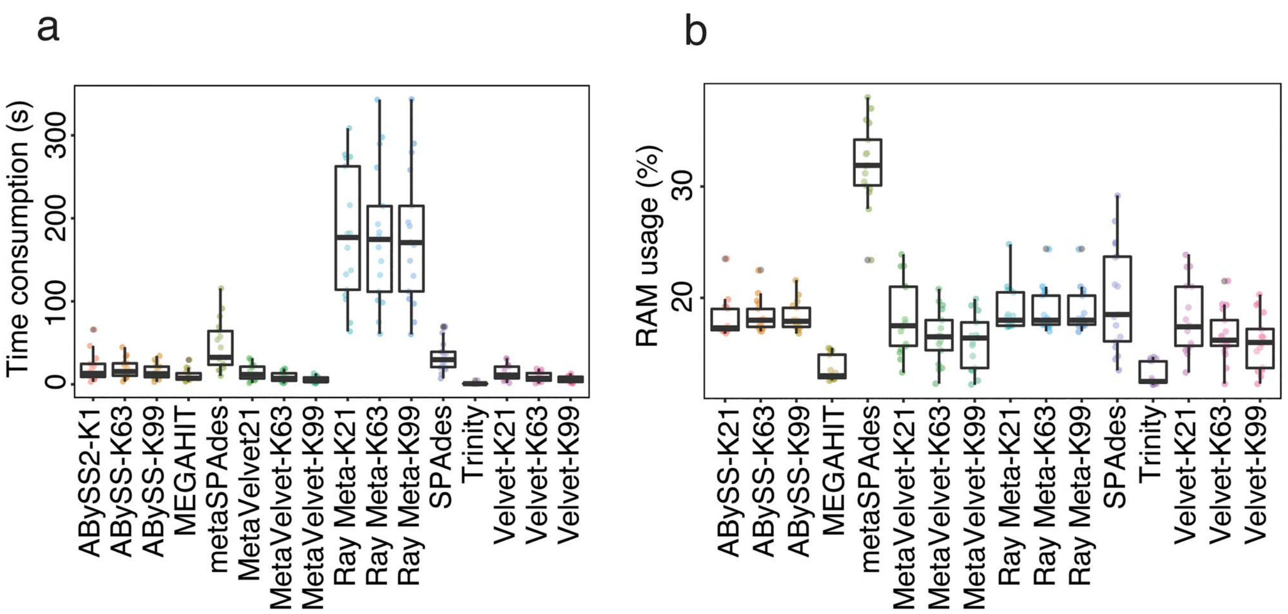

Figure 6. Computational resources required for different assemblers. (A) Time consumed by different assemblers in wallclock CPU seconds. (B) RAM percentage usage by different assemblers.

METHODS

Data source and annotation

Publicly available raw sequencing data of SARS-CoV-2 genome were acquired from SRA. In this study, we used paired-end Illumina sequencing libraries of viral RNA. We randomly selected 100 paired-end Illumina libraries from six different assay types, e.g. amplicon, RNA-seq, targeted-capture, WGA and other categories (Table 1, Figure 1B). The samples that did not pass the quality check were removed from subsequent analysis. As a reference of SARS-CoV-2, 'MN908947.3' genome version was used. For the annotation of genomics features, 'Sars_cov_2.ASM985889v3.100.gff3' was downloaded from the Ensembl database.

Read pre-processing

Adapter and low-quality bases were trimmed using Trimmomatic [18] with default parameters. Raw reads were quality checked using FastQC and multiQC [19, 20] (Figure 1A).

Assemblers tested

In this study, de novo assembly of paired-end reads was performed using the current versions of eight different short-read assemblers. We used ABySS assembler which is optimized for short reads. The parallel version of ABySS is capable of assembling large genomes [21]. MEGAHIT is an ultra-fast and memory-efficient short-read assembler, optimized for metagenomes, also works well on generic single genome assembly of small or mammalian size [22]. Ray Meta is used for metagenome assembly and profiling [23]. SPAdes can assemble sequences from single-cell and multi-cell data types [24]. The Velvet assembler was designed for short-read sequencing data [25]; metaSPAdes is a metagenomic assembler and MetaVelvet is an extension of Velvet for metagenome assembly from short reads [26, 27]. Trinity performs de novo transcriptome assembly [28]. For every assembler mentioned above, we have used default parameters unless otherwise mentioned. *K*-mer lengths 21, 63, 99 were used for ABySS, Velvet, MetaVelvet and Ray Meta. For the rest of the assemblers, default *k*-mer length was applied (Figure 1A, Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Generation of assembly quality matrix

To generate an assembly quality matrix using metaQUAST [29], we removed the contigs with <500 bp and compared all the assemblies to SARS-CoV-2 'MN908947.3' reference genome.

Alignment

We aligned the assembled contigs to the SARS-CoV-2 'MN908947.3' reference genome using Minimap2 (version 2.17 r941) [30]. In Minimap2, we used 5% divergence between reference and assembly sequences to ensure inclusion. To align Illumina paired-end reads, we used BWA tools (version 0.7.17 r1188) [31] with its mem feature enabled and default parameters.

Variant calling from assembled contigs

After aligning the assembled contigs to the reference genome using Minimap2, we sorted the contigs by coordinates and indexed using SAMtools (version 1.11) [32]. BCFtools (version 1.1) mpileup utility was used to generate genotype likelihoods at each genomic position with coverage, from the sorted BAM files to raw VCF formats. To extract the variant calls from the VCF file, we used BCFtools' 'call' command, with the default definition of the 'ploidy' parameter. Other parameters were also unchanged, as suggested.

CPU and RAM usage

To check the computational performance of the assemblers, we randomly took 17 samples from our dataset and used 4 cores (8 threads) to perform assembly. Time to accomplish the assembly by an assembler has been adopted as CPU time for the particular assembler. Mathematically,

$$\begin{aligned}\text{Wallclock CPU time} &= \text{End timestamp of an assembly}\\ &\quad -\text{Start timestamp of an assembly.}\end{aligned}$$

For finding RAM usage, we tracked the percentage usage of RAM every 0.5 s during assembly. We used a dedicated computer with 8 GB of RAM and accepted the maximum RAM usage among all values as final RAM usage. Mathematically,

$$\text{MAX}\left(\text{RAM usage}_{t=\text{start timestamp to end timestamp}}\right).$$

Authors' contributions

Conceptualization, R.I.; Methodology, R.I., R.S.R., N.T. and M.I.H.S.; Software, R.S.R., N.T., M.I.H.S. and R.I.; Formal analysis, R.S.R., N.T., M.I.H.S. and R.I.; Investigation, R.I., R.S.R., N.T. and M.I.H.S.; Writing–Original Draft, R.I., M.A.B., R.S.R., N.T., M.I.H.S. and Y.A.; Writing–Review and Editing, R.I., M.A.B., Y.A. and T.I.; Visualization, R.I.; Supervision, R.I.; Project administration, Y.A.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

We acknowledge Omics-Lab (https://omics-lab.com/) for providing the research platform. We also thank the authors of all datasets used in this study for the availability of their data.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Zhu N, Zhang D, Wang W, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727–33.
2. World Health Organization. Weekly epidemiological update –5 January 2021. *WHO COVID-19 Epidemiol Update* 2021.
3. Hadfield J, Megill C, Bell SM, *et al.* Next strain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;**34**: 4121–3.
4. Wu F, Zhao S, Yu B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**: 265–9.
5. Leinonen R, Sugawara H, Shumway M, *et al.* The sequence read archive. *Nucleic Acids Res* 2011;**39**:D19–21.
6. Lu R, Zhao X, Li J, *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;**395**:565–74.
7. Sutton TDS, Clooney AG, Ryan FJ, *et al.* Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 2019;**7**(12).
8. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. *Cell Mol Life Sci* 2016;**73**:4433–48.
9. Volz EM, Koelle K, Bedford T. Viral Phylodynamics. *PLoS Comput Biol* 2013;**9**:e1002947.
10. Baker M. De novo genome assembly: what every biologist should know. *Nat Methods* 2012;**9**:333–7.
11. Swain MT, Tsai IJ, Assefa SA, *et al.* A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc* 2012;**7**:1260–84.
12. Olson ND, Lund SP, Colman RE, *et al.* Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 2015;**6**:235.
13. Salzberg SL, Phillippy AM, Zimin A, *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res* 2012;**22**:557–67.
14. Gurevich A, Saveliev V, Vyahhi N, *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5.
15. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2014;**30**: 31–7.
16. Vollmers J, Wiegand S, Kaster AK. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters. *PLoS One* 2017;**12**: e0169662.
17. Roux S, Emerson JB, Eloe-Fadrosh EA, *et al.* Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 2017;**e3817**:2017.
18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
19. Andrews S. Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data. *Soil* 1973;**5**:47–81.
20. Ewels P, Magnusson M, Lundin S, *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.
21. Simpson JT, Wong K, Jackman SD, *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009;**19**:1117–23.
22. Li D, Liu CM, Luo R, *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**: 1674–6.
23. Boisvert S, Raymond F, Godzaridis É, *et al.* Ray meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 2012;**13**:R122.
24. Bankevich A, Nurk S, Antipov D, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.
25. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.
26. Nurk S, Meleshko D, Korobeynikov A, *et al.* MetaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34.
27. Namiki T, Hachiya T, Tanaka H, *et al.* MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 2012; **40**:e155.

28. Grabherr MG, Haas BJ, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. *Australas Biotechnol* 2011;**29**:644–52.
29. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;**32**: 1088–90.
30. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.
31. Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*, 2013.
32. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/ map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.