Benchmark datasets for SARS-CoV-2 surveillance bioinformatics

Lingzi Xiaoli[1,*], Jill V. Hagey[1,*], Daniel J. Park[2], Christopher A. Gulvik[1], Erin L. Young[3], Nabil-Fareed Alikhan[4], Adrian Lawsin[1], Norman Hassell[1], Kristen Knipe[1], Kelly F. Oakeson[3], Adam C. Retchless[1], Migun Shakya[5], Chien-Chi Lo[5], Patrick Chain[5], Andrew J. Page[4], Benjamin J. Metcalf[1], Michelle Su[1], Jessica Rowell[6], Eshaw Vidyaprakash[6], Clinton R. Paden[1], Andrew D. Huang[6], Dawn Roellig[1], Ketan Patel[1], Kathryn Winglee[1], Michael R. Weigand[1] and Lee S. Katz[1]

[1] Strain Surveillance and Emerging Variant Team, Centers for Disease Control and Prevention, Atlanta, GA, United States of America

[2] Broad Institute of MIT and Harvard, Cambridge, MA, United States of America

[3] Utah Public Health Laboratory, Salt Lake City, UT, United States of America

[4] Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom

[5] Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States of America

[6] SARS-CoV-2 Emerging Variant Sequencing Project Dry Lab Group Laboratory and Testing Task Force COVID-19 Emergency Response, Centers for Disease Control and Prevention, Atlanta, GA, United States of America

[*] These authors contributed equally to this work.

ABSTRACT

Background. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the cause of coronavirus disease 2019 (COVID-19), has spread globally and is being surveilled with an international genome sequencing effort. Surveillance consists of sample acquisition, library preparation, and whole genome sequencing. This has necessitated a classification scheme detailing Variants of Concern (VOC) and Variants of Interest (VOI), and the rapid expansion of bioinformatics tools for sequence analysis. These bioinformatic tools are means for major actionable results: maintaining quality assurance and checks, defining population structure, performing genomic epidemiology, and inferring lineage to allow reliable and actionable identification and classification. Additionally, the pandemic has required public health laboratories to reach high throughput proficiency in sequencing library preparation and downstream data analysis rapidly. However, both processes can be limited by a lack of a standardized sequence dataset.

Methods. We identified six SARS-CoV-2 sequence datasets from recent publications, public databases and internal resources. In addition, we created a method to mine public databases to identify representative genomes for these datasets. Using this novel method, we identified several genomes as either VOI/VOC representatives or non-VOI/VOC representatives. To describe each dataset, we utilized a previously published datasets format, which describes accession information and whole dataset information. Additionally, a script from the same publication has been enhanced to download and verify all data from this study.

Results. The benchmark datasets focus on the two most widely used sequencing platforms: long read sequencing data from the Oxford Nanopore Technologies platform and short read sequencing data from the Illumina platform. There are six datasets:

Submitted 31 March 2022
Accepted 8 July 2022
Published 5 September 2022

Corresponding authors
Michael R. Weigand, yrh8@cdc.gov
Lee S. Katz, gzu2@cdc.gov

Academic editor
Ravi Kant

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.13821**

How to cite this article Xiaoli L, Hagey JV, Park DJ, Gulvik CA, Young EL, Alikhan N-F, Lawsin A, Hassell N, Knipe K, Oakeson KF, Retchless AC, Shakya M, Lo C-C, Chain P, Page AJ, Metcalf BJ, Su M, Rowell J, Vidyaprakash E, Paden CR, Huang AD, Roellig D, Patel K, Winglee K, Weigand MR, Katz LS. 2022. Benchmark datasets for SARS-CoV-2 surveillance bioinformatics. *PeerJ* **10**:e13821 http://doi.org/10.7717/peerj.13821

three were derived from recent publications; two were derived from data mining public databases to answer common questions not covered by published datasets; one unique dataset representing common sequence failures was obtained by rigorously scrutinizing data that did not pass quality checks. The dataset summary table, data mining script and quality control (QC) values for all sequence data are publicly available on GitHub: https://github.com/CDCgov/datasets-sars-cov-2.

Discussion. The datasets presented here were generated to help public health laboratories build sequencing and bioinformatics capacity, benchmark different workflows and pipelines, and calibrate QC thresholds to ensure sequencing quality. Together, improvements in these areas support accurate and timely outbreak investigation and surveillance, providing actionable data for pandemic management. Furthermore, these publicly available and standardized benchmark data will facilitate the development and adjudication of new pipelines.

Subjects Bioinformatics, Virology, COVID-19
Keywords Standardization, sha256, Benchmarking, WGS, COVID-19

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which caused the coronavirus disease 2019 (COVID-19) pandemic, has infected more than 456 million people globally resulting in over six million deaths as of March 14th, 2022 (https://covid19.who.int/). Since the first genomic sequence of SARS-CoV-2 was made publicly available on January 10th, 2020, whole genome sequencing (WGS) and bioinformatics analyses have been performed extensively to characterize and surveil the virus's evolution. The SARS-CoV-2 virus has undergone rapid evolutionary expansion, leading to the emergence of discrete variants, some of which exhibit altered infectivity, clinical severity, or decreased susceptibility to medical treatments (*Otto et al., 2021*; *Pascall et al., 2021*; *Abdool Karim & De Oliveira, 2021*). The observed waves of novel variants of SARS-CoV-2, with greater transmissibility than the original strain, including B.1.1.7 (Alpha) and B.1.617.2 (Delta), B.1.1.529 (Omicron) has emphasized the need for real-time sequence-based virus surveillance (*Davies et al., 2021*; *Elliott et al., 2021*). Such surveillance has its roots in genomic epidemiology, which had already shown utility and impact for surveillance of other infectious disease agents including but not limited to bacterial foodborne organisms (*PulseNet, 2016*), influenza (*Shu & McCauley, 2017*), ebola (*Quick et al., 2016*), and norovirus (*Vega et al., 2011*).

In 2020, many national viral genomics consortia were established to coordinate SARS-CoV-2 genome sequencing support of public health response efforts including: the Coronavirus Disease 2019 (COVID-19) Genomics UK Consortium (*COGUK, 2020*), SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance (SPHERES) in the USA (*CDC, 2020a*; *CDC, 2020b*; *CDC, 2020c*), the Canadian COVID Genomics Network (*CanCOGeN, 2020*), and the Indian SARS-CoV-2 Genomics Consortium (*INSACOG, 2020*). Such consortia require substantial economic, trained

personnel, equipment, and technical resources, which present considerable barriers to many countries, but are necessary for global control of SARS-CoV-2 outbreaks (*Helmy, Awad & Mosa, 2016*; *Brito et al., 2021*; *Chen et al., 2022*). Leadership from these consortia in their respective countries led many public health laboratories to start building sequencing and bioinformatics capacity for real-time genomic surveillance of SARS-CoV-2. One central challenge to coordinating these efforts, which is not unique to SARS-CoV-2 surveillance, has been the diversity of high-throughput sequencing platforms employed by various laboratories, varying sample preparation methods, and different amplicon strategies. The sequencing platforms include Oxford Nanopore, Illumina, Pacific Biosciences, Ion Torrent; sample preparation methods include amplicon-based, shot-gun, and metagenomics; amplicon strategies include ARTIC V3, ARTIC V4, and SWIFT. Therefore, it is necessary to evaluate whether the sequencing platform that generates data influences downstream bioinformatic processing of consensus sequences in any use-cases, e.g., genomic surveillance or epidemiology. Additionally, the sample preparation methods differ widely, including metagenomic, amplicon and hybrid capture, each of which may impact the creation of assembled sequences or consensus sequences.

Standardization is further complicated by the large number of bioinformatics applications that were either expanded or developed from scratch to rapidly meet the needs of the COVID-19 pandemic. For genomic epidemiology, these platforms include Augur, Auspice, and UShER (*Turakhia et al., 2021*; *Hadfield et al., 2018*). For lineage detection, these platforms include Pangolin and Nextclade (*O'Toole et al., 2021*; *Hadfield et al., 2018*). For QC, many pipelines use different, modular combinations of open-source software including FastQC, NCBI human scrubber, Kraken, Trimmomatic, BBDuk, SeqyClean, SAMtools, and Viral Annotation DefineR (VADR), and Artic field bioinformatics (*Schäffer et al., 2020*; *Li et al., 2009*; *Zhbannikov et al., 2017*; *BBMap, 2021*; *Bolger, Lohse & Usadel, 2014*; *Wood & Salzberg, 2014*; *Katz et al., 2021*; *Andrews, 2010*; *ARTIC, 2020*). A more comprehensive review of the variety of software used was recently published by *Hu et al. (2021)*. There are also many instances of commercial software and/or closed source packages that were rapidly developed, including Illumina's DRAGEN, Clear Dx™ WGS SARS-CoV-2 Bioinformatics Pipeline (BIP), EPISEQ SARS-COV-2 (bioMérieux, Marcy-l'Étoile, France), and CLC Genomics Workbench (QIAGEN, Hilden, Germany).

Critically, a common set of sequence data has not been staged for all available bioinformatics applications to compare their performance. Furthermore, laboratories lack standardized data with which to test their competencies in these analyses, or even train personnel, which is a common requirement for compliance with quality management systems.

To address this gap, we propose specific SARS-CoV-2 sequencing benchmark datasets to aid laboratories in building bioinformatics infrastructure, validating cross-platform sequence analyses, evaluating bioinformatics pipelines, and verifying QC procedures. These datasets may also serve as a training or competency resource for new laboratory staff to understand features of sequence data that either pass or fail common QC metrics.

MATERIALS & METHODS

Datasets

Each dataset listed in Table 1 was designed to address a specific need, following the datasets format described in *Timme et al. (2017)*. Briefly, the tab-separated file format organizes information into two sections: the header and the data. The header contains metadata describing the whole dataset including a unique name, the source of data, and the intended use of the dataset. Directly following the header, the data section includes sample-specific information such as sample name, NCBI accessions, cluster information, and hashsum values of each sequencing read file to be downloaded. This second section may also contain additional, optional columns (such as GISAID accession, lineage, or amplicon strategy) relevant to individual datasets. Essentially, all FASTQ files for sequences in the datasets are stored on NCBI SRA while a simple spreadsheet with accessions is stored in our software repository (*Cock et al., 2010*). A summary of the entire workflow used to identify and validate sequences to be included in the datasets is provided in Fig. 1.

Download script

The script to download each dataset is called GenFSGopher.pl, as described in *Timme et al. (2017)*. GenFSGopher.pl reads a spreadsheet of the user's choosing and downloads the associated data from the specified NCBI accession. Spreadsheets for each dataset are available in the datasets folder of the repository. Each file downloaded by the script is checked against its unique identifier of hashsum, ensuring that the data downloaded on one computer is exactly the same as another computer.

QC metrics and thresholds

Here, we defined specific metrics for both the raw reads and resulting consensus assembly that each sequence must pass for inclusion to our benchmark datasets. There are differences in QC metric thresholds set by pipelines and between the major public repositories of SARS-CoV-2 which cannot be easily accounted for. We summarized the most common metric cutoffs used for our purposes in Table 2.

Sequence quality evaluation

We evaluated the quality of the sequences for the six datasets through three steps. In the first step, we used FastQC to evaluate the basic read quality of the FASTQ files (Fig. 1, Table 2: 1–3). In the second step, we used the genome assembly of Wuhan-1 (accession number: NC_045512.2) as the reference and evaluated the sequencing depth per nucleotide across the total length of the Wuhan reference (Fig. 1, Table 2: 4–8). Metrics 4–6 in Table 2 indicated the average depth per nucleotide, the variation of depth as well as the degree of depth variation. Metrics 7–8 in Table 2 were used to estimate the number of ambiguous nucleotides that would be observed in the downstream consensus assembly. In the third step, we tested all six datasets using the TheiaCoV (formerly 'Titan') workflow (v1.4.4) (*Libuit et al., 2022*) with the default UShER (v0.3.0) as the inference engine (*Turakhia et al., 2021*). TheiaCov is included in our existing bioinformatic infrastructure, is already being used by some public health laboratories, reports our determined metrics of interest, and summarizes

Table 1 The summary description for six datasets. Each dataset is numbered, named, and given a description. The intended use is also listed.

Dataset	Name	Description	Intended use	Reference
1	Boston outbreak	A cohort of 63 samples from a real outbreak with three introductions, metagenomic approach	To understand the features of virus transmission during a real outbreak setting	*Lemieux et al. (2021)*
2	CoronaHiT rapid	A cohort of 39 samples prepared by 18 h wet-lab protocol and sequenced by two platforms (Illumina vs MinION), amplicon-based approach	To verify that a bioinformatics pipeline finds virtually no differences between sequences from the same genome run on different platforms.	*Baker et al. (2021)*
3	CoronaHiT routine	A cohort of 69 samples prepared by 30 h wet-lab protocol and sequenced by two platforms (Illumina vs MinION), amplicon-based approach	To verify that a bioinformatics pipeline finds virtually no differences between sequences from the same genome run on different platforms.	*Baker et al. (2021)*
4	VOI/VOC lineages	A cohort of 16 samples from 11 representative CDC defined VOI/VOC[a] lineages as of 05/30/2021, amplicon-based approach	To benchmark lineage-calling bioinformatics software, especially for VOI/VOCs.	This study
5	Non-VOI/VOC lineages	A cohort of 39 samples from representative non VOI/VOC[a] lineages, amplicon-based approach	To benchmark lineage-calling bioinformatics software, non-specific to VOI/VOCs.	This study
6	Failed QC	A cohort of 24 samples failed basic QC metrics, covering 8 possible failure scenarios, amplicon-based approach	To serve as controls to test bioinformatics QC cutoffs.	This study

Notes.
[a]VOI, variant of interest; VOC, variant of concern

additional QC metrics (Table 2: 9–19) to indicate four aspects: (a) input data size after trimming and human read removal, (b) conflicting read taxonomy, *i.e.*, contamination, (c) amino acid changes especially spike protein mutations, (d) lineage or clade information (Fig. 1). Similar to most SARS-CoV-2 workflows, Pangolin is incorporated into TheiaCov.

Dataset 1—an outbreak

This dataset describes an outbreak resulting from three independent introductions of SARS-CoV-2 in a large metropolitan city (*Lemieux et al., 2021*). The intended use of these two datasets is to evaluate methods for phylogenetic reconstruction, as the resulting phylogenetic tree should accurately delineate three clusters and an outgroup. The expected tree, placement of each sequence and the outgroup are labeled in the "tree" row of the dataset table. It is important to note this study was conducted early in the COVID-19 pandemic, prior to the widespread adoption of amplicon-based sequencing (ARTIC, Swift, etc.), and therefore utilized a shotgun metagenomic approach.

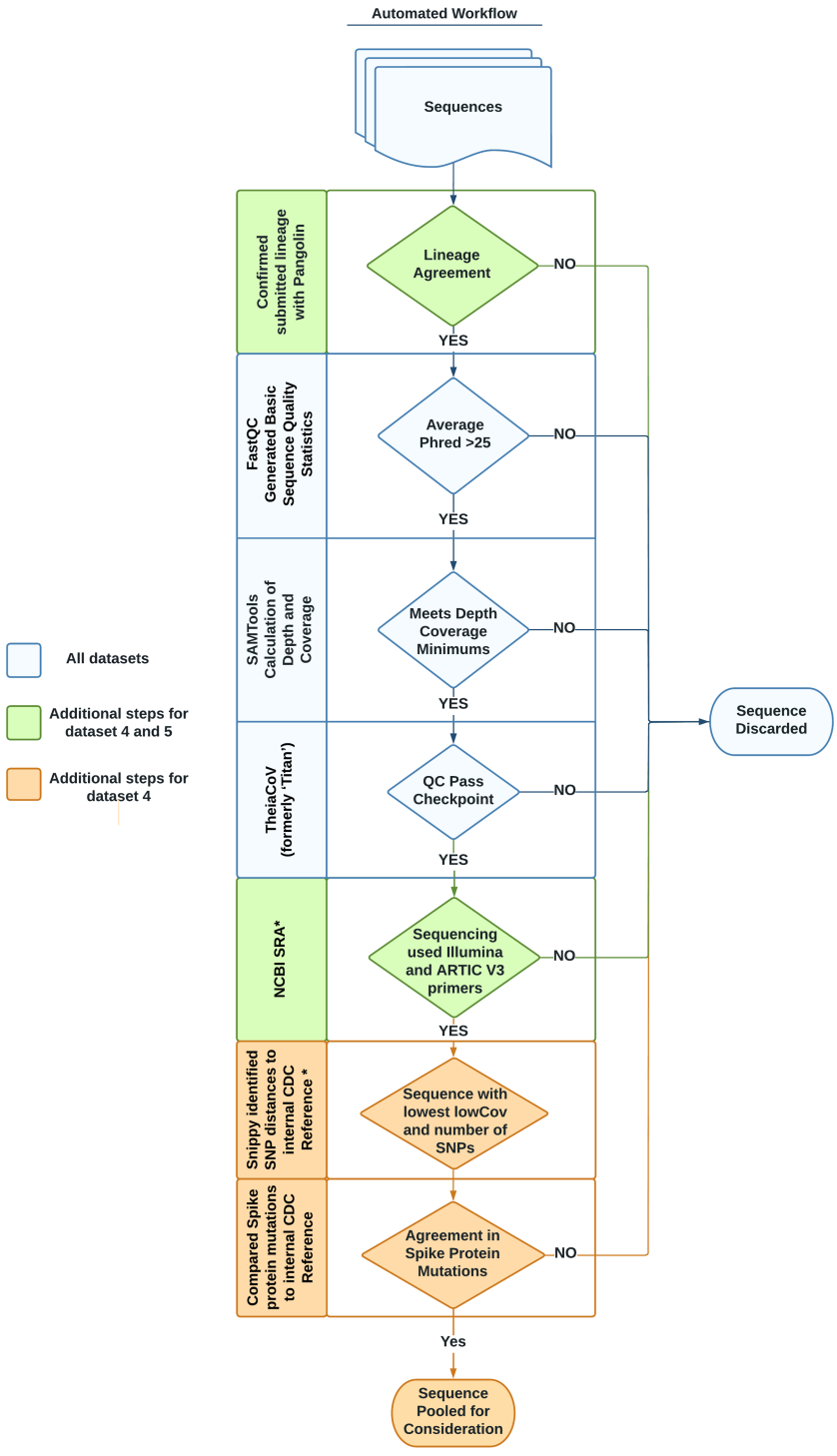

Figure 1 Automated workflow for identifying representative sequences for datasets. Sequences go through several quality checks before being considered as part of a (continued on next page…)

Full-size DOI: 10.7717/peerj.13821/fig-1

Figure 1 (…continued)
benchmark dataset. These checks include lineage agreement with Pangolin, a minimum Phred score, a minimum depth of coverage, a check with the software TheiaCov, a check of the amplicon strategy, a minimization of the count of SNPs in regards to a reference genome, and a check against the spike region's mutations. Asterisks denote steps taken with in-house python scripts.

Table 2 QC metrics. QC metrics are shown with their thresholds, which bioinformatics tool we used, and the QC category.

No.	QC Metrics	Cutoff	Tool (version)	Category
1	total reads	NC	FastQC	Step 1: Fastq quality check
2	read length	NC	FastQC	Step 1: Fastq quality check
3	average phred score	>25	FastQC	Step 1: Fastq quality check
4	mean depth per nucleotide (MDN)	>10	Samtools	Step 2: Depth check
5	standard deviation for MDN	NC	Samtools	Step 2: Depth check
6	coefficient of variation for MDN	NC	Samtools	Step 2: Depth check
7	number of nucleotides with depth <10 (for Illumina)	<3000	Samtools	Step 2: Depth check
8	number of nucleotides with depth <20(for nanopore)	<3000	Samtools	Step 2: Depth check
9	number of paired-end reads	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check
10	assembly total length	>29400	Titan 1.4.4	Step 3: Bioinformatics workflow check
11	ambiguous Ns	<10%	Titan 1.4.4	Step 3: Bioinformatics workflow check
12	assembly mean coverage	>25	Titan 1.4.4	Step 3: Bioinformatics workflow check
13	% mapped to the Wuhan reference	>65%	Titan 1.4.4	Step 3: Bioinformatics workflow check
14	VADR alert number	<=1	Titan 1.4.4	Step 3: Bioinformatics workflow check
15	nextclade_aa_dels	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check
16	nextclade_aa_subs	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check
17	nextclade_version	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check
18	pango_lineage	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check
19	pangolin_version	NC	Titan 1.4.4	Step 3: Bioinformatics workflow check

Notes.
NC, not a criterion.
These values are reported but not used as criteria for passing or failing a sample.

Datasets 2 and 3—multiple platforms

We selected these data from *Baker et al. (2021)*, which describes a novel library preparation method for SARS-CoV-2 for sequencing called Coronavirus High Throughput (CoronaHiT). CoronaHiT can provide flexible throughput using either Illumina or Nanopore technology, which allows sequencing up to 96 samples on Nanopore or 2880 samples on Illumina in a single experiment and generating more even coverage between multiplexed samples. Following the CoronaHiT library preparation, the genomes were sequenced in parallel with both the Illumina and Oxford Nanopore Technologies (ONT) platform. The authors also supply genome sequence data from the ONT platform using the standard library preparation method, known as "LoCost" that can sequence 11-95 samples with one negative control using the Native Barcoding Expansion 96 kit (*Quick, 2020*). Therefore, each genome in datasets 2 and 3 has been sequenced by three approaches (Illumina CoronaHiT, Nanopore CoronaHiT and Nanopore LoCost). *Baker et al. (2021)* further described separate "rapid" and "routine" methods for CoronaHiT, which are also

reflected in datasets 2 and 3, respectively. The difference between ''rapid'' and ''routine'' methods is the run time when using MinION. The ''rapid'' version ran for 18 hrs, while ''routine'' version ran for 30 hrs. The intended use of these two datasets is to verify the consistency of bioinformatics applications for generating consensus sequences from input data produced with various library strategies or sequencing technology.

Dataset 4—lineages

This dataset contains one genome per focal lineage of SARS-CoV-2, named according to the PANGO nomenclature. Important lineages in this paper are defined as a CDC-specified variant of concern (VOC) or variant of interest (VOI) lineage as of June 15th, 2021 (''SARS-CoV−2 Variant Classifications and Definitions'', 2021). At that time, the list included 10 PANGO lineages: six VOCs (B.1.1.7, B.1.351, B.1.427, B.1.429, B.1.617.2, P.1) and four VOIs (B.1.525, B.1.526, P.2, B.1.617.1). Using CDC internally curated consensus sequences (Table S1) for each lineage as the reference, we developed an automatic workflow to select representative Illumina paired-end reads generated with ARTIC V3 primers from NCBI SRA that satisfied our QC metric cutoffs. See the automated data mining workflow section below for more details. Additionally, we verified that a corresponding consensus sequence record was also present in GISAID EpiCov. This dataset is intended for benchmarking PANGO lineage-assignment pipelines, particularly for those classified as VOI/VOC lineages.

Dataset 5—more lineages

This dataset contains a complementary selection of 39 non-VOI/non-VOC lineages, chosen from a collection of CDC internally curated consensus sequences. In addition to raw sequencing reads, accessions to the consensus genome sequences available in both GISAID and NCBI Reference Sequence Database (RefSeq) for each sample in this dataset are provided. The intended use for this dataset is to benchmark lineage-assignment bioinformatics pipelines, nonspecific to VOI/VOCs. Lineages were assigned to all sequences using the Pangolin v3.1.3 classification software (*O'Toole et al., 2021*). All sequences were aligned to a SARS-CoV-2 reference genome (NCBI accession number: MT019531) using the SSW library, an extension of Farrar's Striped Smith-Waterman algorithm (*Zhao et al., 2013*). Representative sequences for each lineage were obtained to avoid artifacts introduced from the aligned and classified sequences by taking the earliest sample from the most abundant genome alignment profile (sorted by genome length) per lineage.

Dataset 6—QC failures

Sequence data that failed at least one QC metric described below were manually selected for this dataset from a large collection of sequence runs performed by the CDC and Utah Public Health Laboratory for national surveillance. These QC failures include low Phred score, low coverage breadth, low mean coverage depth, human sequence contamination, long stretches of ambiguous nucleotides in the consensus (>100 Ns in a row), and amplicon dropout. All human contamination was anonymized by replacing patient read data with the corresponding sequence from a published reference. Briefly, reads containing human sequences were first separated using NCBI's Human Read Removal Tool (*Katz et al., 2021*) and mapped to the T2T reference genome (*Nurk et al., 2022*) using bowtie2 (*Langmead*

& Salzberg, 2012). The sequences of mapped read pairs were then replaced with aligned reference sequence and the anonymized reads subsequently recombined with the remaining non-human reads. In this way, Personal Identifiable Information (PII) has been removed from the dataset while retaining observed frequency and location of human reads. The intended use of this dataset is to help calibrate bioinformatics pipelines against common quality control thresholds for analyses or submission to public data repositories.

AUTOMATED DATA MINING WORKFLOW

To find genomes that met eligibility criteria for the VOC/VOI dataset, a custom workflow which included in-house python scripts was developed, noted with an asterisk in Fig. 1. To begin, 383,000 consensus genomes from the GISAID database assigned to each designated VOC/VOI were downloaded on April 23rd, 2021. Initially, Pangolin (v2.4.2) was run to confirm the lineage was unchanged from the time of submission using PangoLEARN (container dated 2021-05-19).

For each VOC/VOI lineage, a multi-FASTA file was created by filtering the original file containing 383,000 sequences using seqkit with the subcommand grep (v1.0) (*Shen et al., 2016*). Next, Snippy (v4.3.8) was run on the multi-FASTA file against the internal CDC reference sequence (Table S1). These data mining bioinformatics workflows were incorporated on our GitHub repository. For each VOI/VOC lineage, the workflow output four metrics: number of single nucleotide polymorphisms (SNPs) and LowCov for each individual sequence, a SNP range and LowCov range for the multi-FASTA (*Seemann, 2019*). LowCov indicated the number of low coverage sites with a depth cutoff of 10. We selected the 5 samples for each VOI/VOC lineage whose GISAID consensuses have the fewest number of SNPs and LowCov from the corresponding reference. We linked the selected GISAID consensus to the NCBI Sequence Read Archive (SRA) accession. Next, the raw reads from each sample's SRA accession were run through TheiaCov v1.4.4, using the Terra.bio interface, to ensure that they met our minimum QC thresholds (Table S2). As part of TheiaCov's workflow, Pangolin (v3.1.3; container dated 2021-06-15) was rerun on samples with UShER (v0.3.1) being used as the inference engine for lineage calls. This version of Pangolin notably utilizes Scorpio (v0.3.1) and Constellations (v0.0.5). If the read sequences failed any of our QC thresholds, we removed them from consideration until we found the most representative genome for each VOI/VOC lineage. In some cases, we had to analyze all publicly available sequences for a specific lineage to meet our goal.

We also confirmed that the selected SARS-CoV-2 sequences were paired-end, sequenced by the Illumina platform, and used ARTIC V3 primers for amplification through querying their SRA run accessions in the NCBI SRA page. Instead of manually checking this information, we generated another automatic process using the browser automation package Selenium (v3.141.0) for Python3 to harvest the pertinent information (*Muthukadan, 2018*). This script, named NCBI_Scraping.py, accounted for differences in the location and description of the construction protocol (*i.e.*, Artic protocol V3, ARTIC V3 PCR-tiling of viral cDNA, ARTIC v3 amplicons etc.). From this process a CSV file containing a list of SRR accessions that met our criteria was generated. The methods for

our entire automated data mining workflow are encapsulated in scripts that are available in the project repository including a readme.md file describing how to use them.

RESULTS AND DISCUSSION

Datasets

We provide the community with six benchmark datasets of SARS-CoV-2 genomic sequencing data that can be used for a variety of applications (Table 1). These curated datasets consist of a defined outbreak (dataset 1), different sequencing approaches and platforms (dataset 2 and 3), different lineages (datasets 4 and 5), and commonly encountered sequencing failures (dataset 6). We also provide a script that downloads the data behind these datasets. Our benchmark datasets can be applied to many different use-cases, such as bioinformatics pipeline evaluation, QC verification, cross sequencing platform validation, personnel training, or competency testing. We hope this effort will facilitate public health laboratories at different development stages to build robust sequencing and bioinformatics infrastructure for accurate real-time sequence-based virus outbreak investigation and surveillance.

These benchmark datasets can be accessed at https://github.com/CDCgov/datasets-sars-cov-2. To ensure consistency, the benchmark data have been downloaded independently onto different computers at multiple institutions and have been confirmed to be identical. Continuous integration with GitHub Actions is maintained, such that whenever the repository changes, a remote computer at GitHub downloads all the datasets and checks them against the hashsums.

Future datasets

As the SARS-CoV-2 pandemic continues, new important lineages are certain to emerge, and additional lineages can be added to these benchmark datasets as needed with the use of version control on GitHub. This platform is also able to accept new benchmark datasets that address scenarios not covered by our current collection, such as replicate sequencing following amplification with varied primer sets. Users can prepare datasets for consideration by copying our spreadsheet formula, adding their own data to it, and submitting it to the GitHub repository via a pull request. More detailed instructions are available in our GitHub repository. Spreadsheets of new benchmark datasets will be evaluated to confirm that all required fields have been provided in the correct format for the GenFSGopher.pl script, and for data integrity via a hashsums for the expected sequence files downloaded from the SRA. If the dataset passes these criteria, then we can use the pull request feature in GitHub to include these new entries in the repository. In the future, some of these checks may be automated using continuous integration GitHub Actions. In this way, we hope that this GitHub repository becomes a central location for SARS-CoV-2 benchmark datasets.

CONCLUSION

This work describes six benchmark datasets of importance to the global effort in tracking ongoing SARS-CoV-2 evolution and in public health surveillance. These datasets are useful

for testing bioinformatics pipelines, both those already established and in production as well as those under development. In addition, these data can be used as a resource for training laboratory personnel and building regional sequencing capacity. Under each of these circumstances, individual datasets can assist users to: test phylogenetic signals, corroborate cross-platform chemistries, confirm SARS-CoV-2 lineage designations, and verify QC thresholds.

By adopting an open access format this effort helps set the stage for future genomic datasets for SARS-CoV-2 to be included in this publicly available benchmarking repository available via GitHub. Additionally, we have also developed a mechanism through GitHub to accept new benchmark datasets through standard repository pull requests, making the repository a central location for benchmark datasets to support SARS-CoV-2 bioinformatic analyses.

ACKNOWLEDGEMENTS

We acknowledge Jessica C. Chen at CDC for constructive suggestions on manuscript writing. We also thank all data contributors. The authors of the specimens retrieved from GISAID are named in the Supplemental File. Thank you to all who have made contributions on GitHub through pull requests. Thank you to StaPH-B for creating and maintaining a Docker container and to Peter van Heusden for creating and maintaining a Conda environment.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Andrew J. Page and Nabil-Fareed Alikhan were supported by the Biotechnology and Biological Sciences Research Council (BBSRC); their research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10352, also Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). Daniel J. Park was supported by the National Institute of Allergy and Infectious Diseases (U19AI110818) and the Bill and Melinda Gates Foundation (INV-002717). Lingzi Xiaoli, Jill V. Hagey, Chris Gulvik, Adrian Lawsin, Norman Hassell, Kristen Knipe, Adam C. Retchless, Benjamin J. Metcalf, Michelle Su, Clinton R. Paden, Andrew D. Huang, Dawn Roeillig, Ketan Patel, Kathryn Winglee, Michael R. Weigand, and Lee S. Katz were funded by Federal Appropriations to the Centers for Disease Control and Prevention. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Biotechnology and Biological Sciences Research Council (BBSRC).
BBSRC Institute Strategic Programme Microbes in the Food Chain.

Quadram Institute Bioscience BBSRC: BB/CCG1860/1.
National Institute of Allergy and Infectious Diseases: U19AI110818.
Bill and Melinda Gates Foundation: INV-002717.
Federal Appropriations to the Centers for Disease Control and Prevention.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Lingzi Xiaoli performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jill V. Hagey performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Daniel J. Park performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Christopher A. Gulvik performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Erin L. Young performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Nabil-Fareed Alikhan performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Adrian Lawsin performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Norman Hassell conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Kristen Knipe conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Kelly F. Oakeson conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Adam C. Retchless conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Migun Shakya conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Chien-Chi Lo conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Patrick Chain conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Andrew J. Page conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

- Benjamin J. Metcalf conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Michelle Su conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Jessica Rowell conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Eshaw Vidyaprakash conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Clinton R. Paden conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Andrew D. Huang conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Dawn Roellig conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Ketan Patel conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Kathryn Winglee conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Michael R. Weigand conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Lee S. Katz conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The genome sequence accessions are available at GitHub: https://github.com/CDCgov/datasets-sars-cov-2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13821#supplemental-information.

REFERENCES

Abdool Karim SS, De Oliveira T. 2021. New SARS-CoV-2 variants—clinical, public health, and vaccine implications. *The New England Journal of Medicine* **384**(**19**):1866–1868 DOI 10.1056/NEJMc2100362.

Andrews S. 2010. Babraham bioinformatics—FastQC a quality control tool for high throughput sequence data. *Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (accessed on 03 November 2021).

ARTIC. 2020. Home—artic pipeline. *Available at https://artic.readthedocs.io/en/latest/?badgelatest* (accessed on 30 November 2021).

Baker DJ, Aydin A, Le-Viet T, Kay GL, Rudder S, De Oliveira Martins L, Tedim AP, Kolyva A, Diaz M, Alikhan N-F, Meadows L, Bell A, Gutierrez AV, Trotter AJ, Thomson NM, Gilroy R, Griffith L, Adriaenssens EM, Stanley R, Charles IG, Elumogo N, Wain J, Prakash R, Meader E, Mather AE, Webber MA, Dervisevic S, Page AJ, O'Grady J. 2021. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. *Genome Medicine* **13**:21 DOI 10.1186/s13073-021-00839-5.

BBMap. 2021. *Available at https://sourceforge.net/projects/bbmap/* (accessed on 03 November 2021).

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120 DOI 10.1093/bioinformatics/btu170.

Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, Ho J, Tegally H, Githinji G, Agoti CN, Matkin LE, Whittaker C, Howden BP, Sintchenko V, Zuckerman NS, Mor O, Blankenship HM, Oliveira TD, Lin RTP, Siqueira MM, Resende PC, Vasconcelos ATR, Spilki FR, Aguiar RS, Alexiev I, Ivanov IN, Philipova I, Carrington CVF, Sahadeo NSD, Gurry C, Maurer-Stroh S, Naidoo D, Von Eije KJ, Perkins MD, Van Kerkhove M, Hill SC, Sabino EC, Pybus OG, Dye C, Bhatt S, Flaxman S, Suchard MA, Grubaugh ND, Baele G, Faria NR, Danish Covid-19 Genome Consortium, COVID-19 Impact Project, Network for Genomic Surveillance in South Africa (NGS-SA), GISAID core curation team. 2021. Global disparities in SARS-CoV-2 genomic surveillance. *Epidemiology* DOI 10.1101/2021.08.21.21262393.

CanCOGeN. 2020. CanCOGeN | Genome Canada. *Available at https://www.genomecanada.ca/en/cancogen* (accessed on 30 November 2021).

CDC. 2020a. Cases, Data, and Surveillance. *Available at https://www.cdc.gov/coronavirus/2019-ncov/variants/spheres.html* (accessed on 30 November 2021).

CDC. 2020b. Coronavirus Disease 2019 (COVID-19). *Available at https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html*.

CDC. 2020c. COVID Data Tracker. *Available at https://covid.cdc.gov/covid-data-tracker* (accessed on 01 November 2021).

Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, Xu X, Wu Y, Lu W, Ge S, Zhao Z, Yang J, Leung DT, Domman DB, Yu H. 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics* **54**:499–507 DOI 10.1038/s41588-022-01033-y.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**:1767–1771 DOI 10.1093/nar/gkp1137.

COG-UK. 2020. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe* **1**:e99–e100 DOI 10.1016/S2666-5247(20)30054-9.

Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, Pearson CAB, Russell TW, Tully DC, Washburne AD, Wenseleers T, Gimma A, Waites W, Wong KLM, Van Zandvoort K, Silverman JD, Diaz-Ordaz K, Keogh R, Eggo RM, Funk S, Jit M, Atkins KE, Edmunds WJ, CMMID, COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium. 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**:eabg3055 DOI 10.1126/science.abg3055.

Elliott P, Haw D, Wang H, Eales O, Walters CE, Ainslie KEC, Atchison C, Fronterre C, Diggle PJ, Page AJ, Trotter AJ, Prosolek SJ, Ashby D, Donnelly CA, Barclay W, Taylor G, Cooke G, Ward H, Darzi A, Riley S, The COVID-19 Genomics UK (COG-UK) Consortium. 2021. Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* **374(6574)**:eabl9551 DOI 10.1126/science.abl9551.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**:4121–4123 DOI 10.1093/bioinformatics/bty407.

Helmy M, Awad M, Mosa KA. 2016. Limited resources of genome sequencing in developing countries: challenges and solutions. *Applied & Translational Genomics* **9**:15–19 DOI 10.1016/j.atg.2016.03.003.

Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. 2021. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. *Briefings in Bioinformatics* **22**:631–641 DOI 10.1093/bib/bbaa386.

CSIR-Institute of Genomics and Integrative Biology. 2020. COVID-19 Genomic Surveillance. *Available at https://clingen.igib.res.in/covid19genomes/* (accessed on 30 November 2021).

Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2021. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biology* **22**:270 DOI 10.1186/s13059-021-02490-0.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, Adams G, Fink T, Tomkins-Tinch CH, Krasilnikova LA, De Ruff KC, Rudy M, Bauer MR, Lagerborg KA, Normandin E, Chapman SB, Reilly SK, Anahtar MN, Lin AE, Carter A, Myhrvold C, Kemball ME, Chaluvadi S, Cusick C, Flowers K, Neumann A, Cerrato F, Farhat M, Slater D, Harris JB, Branda JA, Hooper D, Gaeta JM, Baggett TP, O'Connell J, Gnirke A, Lieberman TD, Philippakis A, Burns M, Brown CM, Luban J, Ryan ET, Turbett SE, LaRocque RC, Hanage WP, Gallagher GR, Madoff LC, Smole S, Pierce VM, Rosenberg E, Sabeti PC, Park DJ, MacInnis BL. 2021. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**:eabe3261 DOI 10.1126/science.abe3261.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The

Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Libuit K, Petit IIIRA, Ambrosio F, Kapsak C, Smith E, Sevinsky J. 2022. Public health viral genomics: bioinformatics workflows for genomic characterization, submission preparation, and genomic epidemiology of viral pathogens, especially the SARS-CoV-2 virus.

Muthukadan B. 2018. Selenium with Python. *Available at https://selenium-python.readthedocs.io/*.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Cheng H, Chin C-S, Chow W, De Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM. 2022. The complete sequence of a human genome. *Science* **376**:44–53 DOI 10.1101/2021.05.26.445798.

O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, Yeats C, Du Plessis L, Maloney D, Medd N, Attwood SW, Aanensen DM, Holmes EC, Pybus OG, Rambaut A. 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution* **7**(**2**) DOI 10.1093/ve/veab064.

Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, Mechai S, Domselaar GV, Wu J, Earn DJD, Ogden NH. 2021. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Current Biology* **31**:R918–R929 DOI 10.1016/j.cub.2021.06.049.

Pascall DJ, Mollett G, Blacow R, Bulteel N, Campbell R, Campbell A, Clifford S, Davis C, da SFilipeA, Fjodorova L, Forrest R, Goldstein E, Gunson R, Haughney J, Holden MTG, Honour P, Hughes J, James E, Lewis T, Lycett S, McHugh M, Onishi Y, Parcell B, Robertson DL, Sakka NE, Shabaan S, Shepherd JG, Smollett K, Templeton K, Vink E, Wastnedge E, Williams T, Thomson EC, Consortium TC-19 GU COG-U. 2021. The SARS-CoV-2 Alpha variant causes increased clinical severity of disease.

PulseNet. 2016. Announcement: 20th Anniversary of PulseNet: the National Molecular Subtyping Network for Foodborne Disease Surveillance—United States. *Morbidity and Mortality Weekly Report* **65**(**24**):636 DOI 10.15585/mmwr.mm6524a5.

Quick J. 2020. nCoV-2019 sequencing protocol v3 (LoCost). *Available at https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye* (accessed on 19 October 2021).

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Á Camino-Sánchez, Carter LL, Doerrbecker J, Enkirch T, IG Dorival, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal PJ, Turner D, Pollakis G, Hiscox JA, Matthews DA, Shea MKO, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Wölfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keïta S, Rambaut A, Formenty P, Günther S, Carroll MW. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**:228–232 DOI 10.1038/nature16996.

Schäffer AA, Hatcher EL, Yankie L, Shonkwiler L, Brister JR, Karsch-Mizrachi I, Nawrocki EP. 2020. VADR: validation and annotation of virus sequence submissions to GenBank. *BMC Bioinformatics* **21**:211 DOI 10.1186/s12859-020-3537-3.

Seemann T. 2019. Snippy. *Available at https://github.com/tseemann/snippy*.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE* **11**:e0163962 DOI 10.1371/journal.pone.0163962.

Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* **22**:30494 DOI 10.2807/1560-7917.ES.2017.22.13.30494.

Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, Davis S, Tillman GE, Defibaugh-Chavez S, Carleton HA, Klimke WA, Katz LS. 2017. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **5**:e3893 DOI 10.7717/peerj.3893.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics* **53**:809–816 DOI 10.1038/s41588-021-00862-7.

Vega E, Barclay L, Gregoricus N, Williams K, Lee D, Vinjé J. 2011. Novel surveillance network for norovirus gastroenteritis outbreaks, United States. *Emerging Infectious Diseases* **17**:1389–1395 DOI 10.3201/eid1708.101837.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**:R46 DOI 10.1186/gb-2014-15-3-r46.

Zhao M, Lee W-P, Garrison EP, Marth GT. 2013. SSW Library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLOS ONE* **8**:e82138 DOI 10.1371/journal.pone.0082138.

Zhbannikov IY, Hunter SS, Foster JA, Settles ML. 2017. SeqyClean: a pipeline for high-throughput sequence data preprocessing. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM-BCB '17. Boston, Massachusetts, USA: Association for Computing Machinery*. 407–416 DOI 10.1145/3107411.3107446.