Bioinformatics, 38(10), 2022, 2700–2704
https://doi.org/10.1093/bioinformatics/btac176
Advance Access Publication Date: 24 March 2022
Original Paper

OXFORD

Sequence analysis

EDGE COVID-19: a web platform to generate submission-ready genomes from SARS-CoV-2 sequencing efforts

Chien-Chi Lo [1,*,†], Migun Shakya [1,†], Ryan Connor[2], Karen Davenport[1], Mark Flynn[1], Adán Myers y Gutiérrez [1], Bin Hu [1], Po-E Li[1], Elais Player Jackson[1], Yan Xu[1] and Patrick S. G. Chain[1,*]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and [2]National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Can Alkan

Received on June 22, 2021; revised on January 14, 2022; editorial decision on March 21, 2022; accepted on March 23, 2022

Abstract

Summary: Genomics has become an essential technology for surveilling emerging infectious disease outbreaks. A range of technologies and strategies for pathogen genome enrichment and sequencing are being used by laboratories worldwide, together with different and sometimes *ad hoc*, analytical procedures for generating genome sequences. A fully integrated analytical process for raw sequence to consensus genome determination, suited to outbreaks such as the ongoing COVID-19 pandemic, is critical to provide a solid genomic basis for epidemiological analyses and well-informed decision making. We have developed a web-based platform and integrated bioinformatic workflows that help to provide consistent high-quality analysis of SARS-CoV-2 sequencing data generated with either the Illumina or Oxford Nanopore Technologies (ONT). Using an intuitive web-based interface, this workflow automates data quality control, SARS-CoV-2 reference-based genome variant and consensus calling, lineage determination and provides the ability to submit the consensus sequence and necessary metadata to GenBank, GISAID and INSDC raw data repositories. We tested workflow usability using real world data and validated the accuracy of variant and lineage analysis using several test datasets, and further performed detailed comparisons with results from the COVID-19 Galaxy Project workflow. Our analyses indicate that EC-19 workflows generate high-quality SARS-CoV-2 genomes. Finally, we share a perspective on patterns and impact observed with Illumina versus ONT technologies on workflow congruence and differences.

Availability and implementation: https://edge-covid19.edgebioinformatics.org, and https://github.com/LANL-Bioinformatics/EDGE/tree/SARS-CoV2.

Contact: chienchi@lanl.gov or pchain@lanl.gov

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

1 Introduction

Public health laboratories and scientists around the world have been sequencing the genome of SARS-CoV-2, the etiological agent of COVID-19. In just few years, more than 5 million genomes have been sequenced and made publicly available via genome repositories, such as GISAID (Shu and McCauley, 2017) and GenBank (Clark *et al.*, 2016). Multiple sequencing platforms (Illumina, Oxford Nanopore and PacBio) and experimental approaches are being used to obtain these genomes, including enriching for the virus before sequencing, amplifying and sequencing overlapping genomic regions, or performing deep, random, 'shotgun' sequencing. While multiple strategies and platforms provide different trade-offs that can be tailored to address specific scientific questions, they complicate the development of a bioinformatic workflow that can process diverse input data to produce high-quality genome sequences.

There are many available software options for each individual bioinformatic step in genome consensus generation, and often for each sequencing technology and/or approach. However, most of these can only be executed from the command-line in a Linux

environment and are not readily available as an automated workflow. This accessibility barrier and lack of standardization (e.g. on consensus base calling and thresholds for minimum read coverage and support), presents a major challenge to most bench scientists without extensive training in bioinformatics, and can result in *ad hoc* procedures for deciding the final genome. Finally, the lack of automated genome sequence validation and submission to public repositories results in paucity or delays in accurate publicly available data.

There is a need for an easy-to-use, GUI-based, bioinformatics application that helps automate the routine, though complex, bioinformatics steps required to generate high-quality genomes for submission to public repositories. Such an application would position labs with limited computational resources and bioinformatics expertise to perform their own analyses, while encouraging standardization of SARS-CoV-2 genome processing.

To address this need, we developed EDGE COVID-19 (EC-19), a user-friendly, web-based platform that enables rapid and automated processing of FASTQ read datasets from either Illumina or Oxford Nanopore Technology (ONT) platforms generated using either amplicon or random shotgun-based methods. The output of this workflow is a consensus genome, an inventory of all single nucleotide variants (SNVs) and insertion/deletion events (indels), the lineage of the consensus genome and metadata information sufficient for submitting genomes to GenBank, GISAID and Sequence Read Archive (SRA). EC-19 was validated on a number of datasets and further compared with another popular analytic workflow provided by the Galaxy Project (Maier *et al.*, 2021). Detailed examination and comparison of the results demonstrate EC-19 provides highly robust SNV and indel calls, resulting in high-quality genomes. This platform is freely available as a Docker container for local installation and as a public web-service, where users can register for accounts, upload data, run analyses and download results.

2 Design, implementation and features

EC-19 is a tailored version of the more generic open-source, Empowering the Development of Genomics Expertise (EDGE) bioinformatics platform (Li *et al.*, 2017) that was developed to democratize genomic analysis and make complex bioinformatic tools available to non-experts. Although it includes a similar user interface and other portions of the original platform, EC-19 has been customized for SARS-CoV-2 genomes.

EC-19 allows users to input their own Illumina or ONT FASTQ files, or automatically obtain them from the International Nucleotide Sequence Database Collaboration (INSDC) raw read repositories. The workflow (Fig. 1) uses many commonly used tools for analysis of Illumina or ONT data, and this includes quality control using FaQCs (Lo and Chain, 2014), mapping reads to a SARS-CoV-2 reference genome using Minimap2 (Li, 2018) (default for ONT) or BWA mem (default for Illumina), removal of primer sequences if an amplicon method is specified or primer sets are known, generation of consensus genomes, variant calling and the ability to take sample metadata and deposit genomes to GISAID and GenBank, and raw reads to INSDC. While users can specify any SARS-CoV-2 genomes from GenBank as a reference, or upload their own reference for variant and consensus calling, a slightly modified RefSeq genome (NC_045512.2 with the 33 nt poly-A tail from the 3′ end of the genome removed) is used as the default reference. Upon selection of the reference genome, each position in the consensus genome is reported as either a gap (*n*) when no reads are mapped to the position, an ambiguous base (*N*) when there are five or fewer reads to support a nucleotide. EC-19 reports indels in the consensus genome if the indels are supported by an Allele Frequency (AF) >0.5 for Illumina reads, by and AF > 0.6 for ONT data to account for the higher error rates with this technology, and >0.8 within homopolymer sequences (Cretu Stancu *et al.*, 2017; Rang *et al.*, 2018). A separate consensus genome is also generated with IUPAC codes where multiple alleles are observed above 0.2 AF, and a single allelic nucleotide (SNV or reference nucleotide) is reported for AFs of at least 0.8.

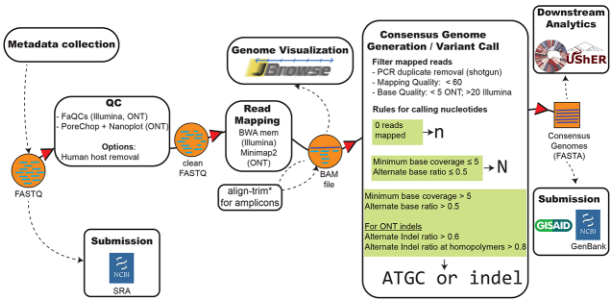

Fig. 1. Overview of the EDGE COVID-19 workflow. EC-19 includes Quality Control, mapping reads to a SARS-CoV-2 reference genome sequence, removing primer sequences when needed, variant allele analysis and generation of consensus genomes, lineage determination and phylogenetic placement. Dotted lines indicate optional steps. Greater detail, including underlying tools and versioning can be found in Supplementary Table S2 and in Supplementary Material

EC-19 automatically updates which workflow is used based on the sequencing technology, with default parameters taking into account user-provided (or SRA metadata-derived) information on methods used (shotgun versus amplicon) and primer details if they are used and available. Detailed documentation for each step in the workflow can be found in the Supplementary Material, and most of the parameters can be modified readily directly from the web GUI. Lineage assignment is conducted with Pangolin (O'Toole *et al.*, 2021) using the *–usher* mode (Turakhia *et al.*, 2021). To assure that the latest release of Pangolin is used, EC-19 automatically checks and downloads the latest available version for each sample run. Outputs of the workflow are presented within the web browser and include statistics and figures for quality control, read mapping and consensus genome generation, as well access to all original output files from the tools used during the analysis. The integrated genome browser JBrowse (Buels *et al.*, 2016) further allows for a more detailed, visual inspection of SNVs and indels and their underlying data.

While several workflows exist to process either Illumina or ONT data, from quality filtering to consensus calling and even lineage assignment, few provide a web-based environment that is easy to use with many integrated features that lend themselves to deeper investigation and rapid hassle-free data submission to public repositories. A recently published and popular SARS-CoV-2 analysis workflow from the Galaxy project (Maier *et al.*, 2021) does provide a number of similar features, but differs with respect to the workflow itself (i.e. the tools used for quality control, for calling nucleotide variant alleles and for generating the consensus genome) and does not provide a project-based view of all the results, interactive genome browsing or the ability to automatically submit the data to public repositories. A detailed comparison of different features offered by the two platforms is summarized in Supplementary Table S1.

3 Testing and evaluation

We evaluated the EC-19 workflow for its usability by generating consensus genomes from a number of samples sequenced at LANL from positive cases in New Mexico (USA), and then using the EC-19 platform to submit them to GenBank and GISAID. The accuracy of the EC-19 workflow has also been validated using benchmark datasets put together by the U.S. Centers for Disease Control and Prevention (CDC) (https://github.com/CDCgov/datasets-sars-cov-2), and detailed comparisons between the EC-19 and the COVID-19 Galaxy Project workflows were performed, with respect to detected mutations (both SNV and indel calling). For these comparisons, we used a set of data and results that have been collated as part of an ongoing collaborative effort across a variety of partner institutions (the ACTIV TRACE Deep Sequencing Subgroup). A total of 239 SRA datasets allowed us to compare the workflow results across several amplicon-sequencing technology combinations. A total of 151 ONT and 88 Illumina SRA runs sequenced using one of the

Table 1. SNVs, insertions and deletions that are shared, or specific to the EC-19 and Galaxy workflows

	SNVs			Insertions			Deletions		
	EC-19	Shared	Galaxy	EC-19	Shared	Galaxy	EC-19	Shared	Galaxy
Illumina	16	1149	49	0	4	0	3	50	0
ONT	34	1919	14	0	3	44	10	69	4

ARTIC protocols from six Bioprojects and 98 Biosamples were processed using the EC-19 and the results compared with the same samples run using the Galaxy workflow (https://github.com/veg/SARS-CoV-2/blob/compact/ACTIV-TRACE/platform-comparison/galaxy.csv; last accessed on October 1, 2021). Because the consensus genomes generated by EC-19 only report SNVs that have AF > 0.5, as well as indels by AF >0.5 for Illumina (and AF > 0.6 for ONT) data, the core comparisons performed only focused on variant and indel calls that passed these thresholds for both EC-19 and Galaxy workflows. All of the processed SRA data can also be accessed from the public project list on the EC-19 website (https://edge-covid19.edgebioinformatics.org/) and the compiled results can be accessed from https://github.com/LANL-Bioinformatics/Lo-et-al-Bioinformatics-EC-19-data.

4 Results and discussion

4.1 Real-world data processing and submission to GISAID and GenBank

The EC-19 platform was made available early on in the pandemic and was initially tested on dozens of SARS-CoV-2 genomes sequenced (using either ONT or Illumina technologies with either the ARTIC or SWIFT PCR enrichment protocols) as part of a surveillance program in New Mexico, USA. Single nucleotide differences, as well as indels and lineage assignments were manually inspected, and the consensus genomes were successfully processed and submitted to both GISAID and GenBank genome repositories using the EC-19 user interface (Supplementary Table S3).

4.2 Genomes generated using EC-19 correctly assigned to expected lineages

We further validated the lineage assignments made by the EC-19 workflow due to the fact that this type of analysis is one of the most pertinent in terms of public health impact. Using two sets of benchmark data with known corresponding lineages (55 SRR datasets in total) provided by CDC (https://github.com/CDCgov/datasets-sars-cov-2), EC-19 correctly predicted all 55 lineages based on the EC-19-derived consensus genome. (Supplementary Table S4).

4.3 Reprocessing SRA datasets for cross-workflow comparisons

To ascertain that the EC-19 workflow converges on the same mutation results provided by other established workflows, we compared all the SNVs and indels reported in EC-19 consensus genomes to those determined by the COVID-19 Galaxy Project workflow (Maier *et al.*, 2021). A total of 239 samples sequenced by Illumina or ONT were processed using EC-19 with default parameters specific to the sequencing platform and strategy (Supplementary Table S5). The test datasets were generally of high quality, with an average fold coverage of 1834× and average linear coverage of 88% of the reference genome per sample, however a subset of samples (36) we refer to as low quality with <75% linear genome coverage (and covered as little as 1.4% of the genome). A total of 3115 SNVs and 139 indels were reported by EC-19 for all projects, with an average of 13 SNVs and 0.6 indels detected per sample.

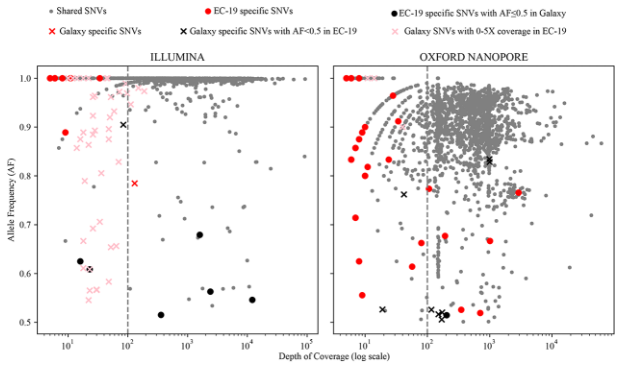

Fig. 2. Distribution of all SNV mutations detected by the EC-19 and Galaxy workflows, displaying depth of coverage (DP) (*X* axis) and allele frequency (AF) (*Y* axis). SNVs detected by both workflows for Illumina (left panel) and ONT (right panel) data are shown in gray (detected and called by both workflows) and black (detected by both but above 0.5 AF cutoff in only one workflow). SNV mutations detected by only one of the workflows yet not detected in the other are colored in red. A dotted vertical line is drawn at 100× Depth of coverage to separate SNVs with high and low depth of coverage. All the AFs and SNVs are based on EC-19 mapping, except for Galaxy specific SNVs

4.4 Strong concordance among workflows with Illumina data

For each sequencing sample, we compared variant calls reported by EC-19 to the ones reported by the Galaxy workflows. For Illumina datasets, both SNVs and indels showed a high (95%) percentage of agreement (Table 1). Moreover, the majority (59/68) of all differences (49 SNVs specific to Galaxy and 16 SNVs and 3 deletions specific to EC-19) between the workflows were in regions of low coverage (Depth coverage (DP) <100×) (Fig. 2 and Supplementary Table S6).

Among the 16 EC-19 specific SNVs, 6 were in fact reported by the Galaxy workflow, but had AFs of lower than ≤ 0.5 with an average of 0.46 (Fig. 2). For example, position 11 123 in ERR4969200 is reported as the majority SNV (G to C with AF = 0.55) in EC-19, but the COVID-19 Galaxy Project reported an AF of 0.46 (Supplementary Fig. S1). The remaining SNVs were majority variants (AF > 0.68) but were in regions with low-fold coverage (an average of only 9.6×), where individual reads could have an impact on the AF. We attribute these EC-19 specific SNVs to differences in the raw data quality control step which results in mapping and AF differences compared with the Galaxy workflow.

Deeper investigation of the 49 Galaxy-specific SNVs revealed that all 49 are either within ARTIC primer binding sites or within 5 nt of these locations. Thirty-eight of these sites originate from only 12 poor-quality samples (<22% linear coverage on average), exhibited low (32× average)-fold coverage based on the Galaxy output, and in our analyses, did not recruit any reads (i.e. not supported by any data). Ten of the remaining 11 Galaxy-specific SNV sites had extremely low DP (<10) in EC-19. For the remaining SNV, Galaxy reported an A to G transition at position 19 865 in sample ERR4969215, but all the mapped reads in this region in EC-19 matched the reference (Supplementary Fig. S2).Taken together, these observations with Illumina data suggest that the overall per sample data quality and the local fold coverage are important determinants for the discrepancies observed between these two workflows, and that extra care should be taken when determining

mutations in these regions and in regions adjacent to or overlapping primer locations.

4.5 Increased differences observed among workflows with ONT data

For ONT data, EC-19 and Galaxy workflow output comparisons were more discordant. While the workflows generally agreed (95%) when detecting SNVs, agreement among indels was much lower, with only 55% of all detected indels in agreement (Fig. 2; Supplementary Table S7). The source of the majority of the indel discrepancies observed are due to 44 insertions uniquely called by the Galaxy workflow. Many (21) of these Galaxy-specific insertions were found to be present in the EC-19 results, but fell below the default threshold cutoff ($\leq$0.6) when examining the read mapping data in detail. Another 21 Galaxy-specific insertions were found in only five samples and had no underlying read support when examining the EC-19 read mapping data. These insertion calls by Galaxy were at genomic locations (position 23, 30 and 31) that either overlap or precede the first ARTIC primer, and theoretically should not have any coverage as the first primer and preceding sequence are generally trimmed from the analyses. The two final Galaxy-specific T insertions were both at position 3034 of the genome, in samples ERR4969052 and ERR5501241 (see Supplementary Fig. S3). This position is at the beginning of a short TTT homopolymer (in the reference genome). We note that in these samples, this has become a TTTT homopolymer due to a C to T transition at position 3037. An alternate encoding of this mutation event could be an insertion of a T in the homopolymer (the Galaxy call of a T insertion at 3034 is consistent with this) followed by a C deletion at 3037. However, in the Galaxy workflow results of both samples, the C to T transition was also called, which indicates this Galaxy-specific insertion is not simply the result of an encoding difference of an indel but is an insertion call that results in a frameshift in the *orf1ab* gene.

With respect to ONT deletions found by the Galaxy and EC-19 workflows, 82% (69) were in agreement while four were specifically found with the Galaxy workflow and ten were specific to EC-19. Three of the four Galaxy-specific deletions were from one sample (SRR11593353) that exhibited low coverage (~11×), and were at positions upstream of the first ARTIC primer binding site (Supplementary Table S7). The fourth Galaxy-specific deletion was found at position 12 790 in ERR4969052, which was also detected with EC-19 but was not higher than the cutoff of 0.6. The ten EC-19 specific deletions were either from low coverage areas or low-quality genomes, with eight of the deletions covered at an average DP of 20× (see Supplementary Fig. S4 for an example) and the remaining two were from samples with <20% linear coverage of the genome (see Supplementary Fig. S5).

While the large majority (Fig. 2) of SNVs were found by both workflows, there were 14 Galaxy-specific SNVs and 34 EC-19 specific SNVs. Among the 14 Galaxy-specific SNVs, eight were also detected by EC-19 but at an AF lower than the default threshold cutoff of 0.5 (Table 1). The final six were reported within or preceding the first ARTIC primer, similar to other anomalous observations discussed above. Of the 34 EC-19-specific SNVs, only one SNV was also found by the Galaxy workflow just below the AF cutoff value of 0.5 at AF 0.49. Fifteen of the remaining EC-19-specific SNVs were in low quality projects with <75% linear coverage and another 15 were from regions with low coverage (average DP of 7.6). The final three were in locations with reasonable fold coverage (average DP of 387) and AF values between 0.51 and 0.77, yet none of these SNVs was reported by the Galaxy workflow (an example in Supplementary Fig. S6). While these three SNV differences are difficult to explain, the inconsistencies in ONT workflows, like their Illumina counterparts, are concentrated in low-quality samples and regions of low-fold coverage.

4.6 Differences in workflow outputs and cautionary tales in comparing results

The majority of all differences examined between the EC-19 and Galaxy Illumina and ONT workflows reside in either samples that are poorly sequenced (low linear coverage), in regions that display lower-fold coverage and/or in genomic locations overlapping amplification primers (Fig. 2). The linear and depth of coverage varies greatly among samples and this significantly impacts the consistency of these two established platforms. The examined differences are likely the result of how strict the parameters are for quality control, read mapping and post-mapping variant calling.

Given the constantly evolving protocols (e.g. amplification primers, library preparation and sequencing technologies), some dictated by a constantly evolving virus, bioinformatics workflows will continue to evolve and adapt as well. Having an open and transparent set of tools, parameters and variant calling rules, and reporting these rules in a consistent fashion as metadata when depositing genomes, or otherwise submitting the raw data alongside genome depositions will allow the community to validate the veracity of the data. Without these supporting data, efforts to document 'problematic' sites (http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473) in the genome (ascertained post-alignment and phylogenetic reconstruction) will undoubtedly continue to expand, and much time and effort will continue to be unnecessarily wasted by scientists to infer the possible methodological basis underlying questionable sequences.

In addition, we suggest that comparing workflow outputs be approached with caution and with rigorous examination of not only the output files summarizing (i) the detected mutations, (ii) their genomic positions within the reference genome and (iii) the allelic frequencies at those positions, but also examination of the underlying raw data that support the variant calls. For comparison of workflow results, the formatting describing the allelic variants observed in each workflow can differ, which can then lead to interpreted differences in variant calls, where none exist. For example, at position 27 396 in ERR4969365, the encoding of a deletion TG to A can be misinterpreted as TG → A- instead of TG → -A (see Supplementary Fig. S7). This encoding was not consistent with other reported deletions, in the output file (where the deletion is reported in the second nucleotide position) and led to two detected differences between EC-19 and Galaxy instead of none. Only detailed investigations can identify these types of reporting anomalies; therefore, we strongly recommend a fully standardized output format that can be directly compared instead of manipulating outputs from different workflows and tools for comparisons.

5 Conclusion

While the SARS-CoV-2 pandemic has both highlighted the utility of genome sequencing and our collective ability to generate enormous numbers of genomes during an outbreak, it is often unclear how individual genome sequences have been generated. This is rendered enormously complex by the number of different sequencing protocols available, the number of different sequencing technologies used, the number of different bioinformatics strategies being applied, and the number of different labs (with diverse expertise and experience with genomics) generating and submitting the data. Efforts such as the EC-19 platform are meant to lower the barrier for submitting genomes for public use, while helping to generate and interactively investigate high quality genomes and their underlying data. EC-19 provides a potential solution toward reproducible genomic surveillance capabilities for the global community using a robust suite of bioinformatics workflows in a user-friendly web browser.

Acknowledgements

The authors thank ACTIV TRACE deep sequencing analysis subgroup for providing test datasets, and the Galaxy team for processing the samples and making these data available for analysis.

Funding

This work was supported by the Los Alamos National Laboratory LDRD [20200732ER], DTRA [CB10152 and CB10623] and DOE [KP160101 and 4000150817]. Hosting of edge-covid19.edgebioinformatics.orgis provided by the Extreme Science and Engineering Discovery Environment (XSEDE),

which is supported by National Science Foundation [ACI-1548562, allocation id is MCB180107]. R.C was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

Conflict of Interest: none declared.

Data availability

All new data generated as part of this manuscript has been deposited to NCBI SRA database (Bioproject: PRJNA714680), GenBank, and GISAID. Individual accession IDs for GenBank and GISAID can be found in Supplementary Table S3.

References

Buels,R. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.

Clark,K. *et al.* (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.

Cretu Stancu,M. *et al.* (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.*, **8**, 1326.

Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li,P.E. *et al.* (2017) Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.*, **45**, 67–80.

Lo,C.C., and Chain,P.S. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, **15**, 366.

Maier,W. *et al.* (2021) Ready-to-use public infrastructure for global SARS-CoV-2 monitoring. *Nat. Biotechnol.*, **39**, 1178–1179.

O'Toole,A. *et al.* (2021) Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.*, **7**, veab064.

Rang,F.J. *et al.* (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.*, **19**, 90.

Shu,Y., and McCauley,J. (2017) GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill*, **22**, 30494.

Turakhia,Y. *et al.* (2021) Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, **53**, 809–816.