A phylogenetics and variant calling pipeline to support SARS-CoV-2 genomic epidemiology in the UK

Rachel Colquhoun[1*], Áine O'Toole[1], Verity Hill[1,2], JT McCrone[1,6], Xiaoyu Yu[1], Samuel M. Nicholls[3], Radoslaw Poplawski[3], Thomas Whalley[4], Natalie Groves[5], Nicholas Ellaby[5], Nick Loman[3], Tom Connor[4,7], Andrew Rambaut[1]

1 Institute of Ecology and Evolution, University of Edinburgh, Edinburgh EH9 3FL, UK

2 Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510, USA

3 Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK

4 School of Biosciences, Cardiff University, Cardiff CF10 3AX, Wales, UK

5 UK Health Security Agency, 10 South Colonnade, Canary Wharf, London, E14 4PU, UK

6 Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

7 Public Health Wales, Number 2 Capital Quarter, Tyndall Street, Cardiff, CF10 4BZ

Abstract

In response to the escalating SARS-CoV-2 pandemic, in March 2020 the COVID-19 Genomics UK (COG-UK) consortium was established to enable national-scale genomic surveillance in the United Kingdom. By the end of 2020, 49% of all SARS-CoV-2 genome sequences globally had been generated as part of the COG-UK programme and to date this system has generated more than 3 million SARS-CoV-2 genomes. Rapidly and reliably analysing this unprecedented number of genomes was an enormous challenge. To fulfil this need and to inform public health decision making, we developed a centralised pipeline that performs quality control, alignment and variant calling, and provides the global phylogenetic context of sequences. We present this pipeline and describe how we tailored it as the pandemic progressed to scale with the increasing amounts of data and to provide the most relevant analyses on a daily basis.

Keywords

SARS-CoV-2, genomic surveillance, genomic epidemiology, phylogenetics, software.

Introduction

In the decade prior to 2020, viral genomic epidemiology emerged as a dynamic and rapidly evolving field. Phylogenetic analysis was used to infer the origins and diversity of HIV [1] and influenza A virus, including during the 2009 swine flu epidemic [2]. The decreasing cost of sequencing allowed it to be applied further to "large-scale" datasets to infer transmission dynamics and influence public health decisions, first during the 2013-2016 West African Ebola epidemic [3,4], and for each major epidemic since (Zika [5,6], MERS[7], Ebola in DRC [8]). Concurrently, visualisation tools like Nextstrain [9] had been developed to enable interactive tracking of viral evolution. When the SARS-CoV-2 pandemic started, a global sequencing effort provided an unprecedented opportunity to use genomic surveillance to inform the public health response.

In March 2020, the COVID-19 Genomics UK (COG-UK) consortium was set up to provide a framework for national-scale, rapid whole-genome sequencing of SARS-CoV-2 within the UK in order to understand viral transmission and evolution and inform public health responses in real-time. This national partnership included the four UK Public Health Agencies, NHS organisations, regional sequencing centres and academic partners [10]. The data generation arm of the consortium operated as a decentralised network of labs, in both healthcare and academic settings, collecting and genome sequencing SARS-CoV-2 samples. The genome sequences and associated metadata were then submitted to a central platform, CLIMB-COVID [11], where it was collected into a single canonical dataset.

A genome sequence without appropriate associated metadata is of limited use, so we quickly established a minimal metadata standard that could contextualise genomes in time and space. This minimum standard facilitated informative analysis of the data whilst limiting the burden of participation. It required a collection date, or the date a sample was received by a lab, country and county level geographic information (administration levels 1 and 2 in the UK) and a record of whether the sample was collected as part of a random surveillance strategy or a targeted outbreak analysis. Additional metadata fields could be supplied and, in practice, this has resulted in a rich and detailed dataset with a consistency of useful information that has been invaluable to consortium members across the UK. This level of private metadata sharing was only possible within a controlled UK-based shared computing environment.

To interpret any new genome sequence it needs to be compared to and contextualised within the recent local and global diversity, most commonly in the form of a phylogenetic tree. Most phylogenetic methods were not developed with this unprecedented amount of data in mind and require large computational resources that scale poorly with increasing numbers of sequences. To overcome this, we needed to develop an analysis pipeline that processed this dataset centrally on a daily basis, performing alignment and variant calling, and amalgamated

the COG-UK data with publicly available global sequences to provide the phylogenetic context. Outputs of this pipeline were made available within the consortium and provided interpretable results for both local NHS health protection teams and the UK public health agencies. They were consumed by public data explorers including COG-UK-ME [12], Microreact [13] and the UCSC Genome Browser phylogenetic tree [14] and provided the basis of individual local outbreak investigations using CIVET [15].

Results

The analysis pipeline that supported the UK efforts is divided into two workflows (*Datapipe* and *Phylopipe*, Figure 1) written in the nextflow [18] workflow language. Datapipe (https://github.com/COG-UK/datapipe) performs alignment and variant calling, and Phylopipe (https://github.com/virus-evolution/phylopipe) constructs the phylogenetic tree. These replace the original single snakemake [19] pipeline, grapevine (https://github.com/COG-UK/grapevine), which performed both workflows until February 2021. A high level overview of each pipeline is provided here along with the design decisions, with further details provided in the Supplementary Materials.

Datapipe

This consumes the FASTA and metadata TSV file(s) generated by the ELAN pipeline (https://github.com/SamStudio8/elan-nextflow/) and runs variant calling and alignment. First we clean non-nucleotide symbols from sequences and reformat metadata.New samples are assigned a Pango lineage with pangolin [20], with all samples reassigned if the underlying model has been updated.. Sequences are deduplicated if they correspond to the same biological isolate and background/global sequences with the same sample_name are deduplicated by date.

A trimmed FASTA alignment is generated using minimap v2.17 [21] to pairwise align each sequence to Wuhan-Hu-1 (GenBank: MN908947.3; https://www.ncbi.nlm.nih.gov/nuccore/MN908947) and gofasta [22] to combine and mask 5' and 3' ends. Insertions relative to the reference are discarded. All nucleotide mutations, insertions and deletions with respect to the reference are noted in the metadata. Sequences that consist of more than 5% unknown sites ('N') per genome after mapping are discarded. Geographical metadata is cleaned (https://github.com/COG-UK/geography_cleaning) and the UK dataset is combined with all non-UK sequences from GISAID [23], which we process similarly on a weekly basis.

Finally, subsets of the metadata and alignment files are published using a configuration JSON. This includes outputs with sensitive data removed that can be made available to the public via the consortium website and s3 buckets (https://www.cogconsortium.uk/), as well

as specific subsets which are consumed by data explorers including the COG-UK Mutation Explorer [12], and CIVET [15].

Design decisions

Pango data releases were regular for most of the pandemic and often resulted in the most recent sequences being re-classified as new lineages were defined. For this reason we added a step to the pipeline to check if a new version had been released, in which case we would re-classify all sequences.

The very first implementations of the pipeline used Multiple Sequence Alignment, however this approach scales quadratically with the number of sequences and a pairwise approach using gofasta was necessary from very early in 2020. The alignment was trimmed with the 5' and 3' ends masked with N's as these regions were typically more error prone.

Whilst outputs were initially hardcoded, we later started generating outputs using a JSON "recipe" file. This allowed different subsets of the sequences and metadata to be defined easily as the downstream requirements for analysis changed over time.

Phylopipe

This consumes the FASTA and metadata CSV file(s) generated by the datapipe pipeline and either constructs a phylogenetic tree using FastTree [16], or adds to an existing tree with UShER [17].

First, globally problematic sites (flagged homoplasic sites, sites with mutations arising multiple times across the canonical global phylogeny, and nanopore adaptor sites) are masked in the combined (UK and global) alignment, and sequences with too many ambiguous bases are excluded.

To construct a new phylogenetic tree, non-unique sequences are hashed to a single representative. Optionally sequences are further reduced by heavy downsampling by date and lineage diversity. The reduced alignment is split based on Pango lineage assignment into 6 large distinct sub-lineages and sub-trees are built independently for each using the Jukes-Cantor model [24] in FastTree [16] v2.1.10 (double precision). The resulting sub-trees are rooted and grafted together by attaching the root of incoming trees to the same taxon's tip in the parent tree, and non-unique sequences are inserted alongside their representative. Branch lengths less than 5E-6, which represent distances smaller than one SNP and result from ambiguities between sequences, are collapsed to 0.

UShER [17] is used to update this tree with additional sequences using maximum parsimony. Branches with more than 30 private mutations are pruned from the tree as artifactual, branch lengths rescaled and the tree rerooted on Wuhan/WH04/2020.

The tips of the full tree are annotated with binary UK/non-UK trait information, and fine scale uk_linages representing independent UK introductions from other countries.

The resulting annotated tree and metadata are disseminated to the consortium. Again specific outputs are published using a configuration JSON, including those for Microreact [13].

Design decisions

As full tree construction typically scales worse than quadratically with the number of sequences, we implemented several steps to try and reduce the number of sequences considered by a tree-building algorithm. In the first case we hashed non-unique sequences, including only a single representative sequence type during the tree building step. These sequences were then inserted alongside their representative in the resulting tree. This approach is mirrored internally by some tree building methods such as FastTree but not all. Secondly we partitioned the sequences into groups based on their Pango lineage assignment, which we expected to represent well defined subtrees with a clear outgroup. When these groups are relatively evenly sized, this approach effectively divides the expected total time to construct the tree by the number of groups, with a much greater time saving if the subtrees are then constructed in parallel instead of consecutively. Finally, updating the tree with UShER allows only new sequences to be added in approximately linear time.

In addition to defining UK introductions, the subtree for each UK lineage was annotated with phylotypes by codifying the internal nodes of the tree, effectively representing parent/child/sibling phylogenetic relationships in metadata. This proved hugely beneficial to public health agencies as it enabled interpretation of phylogenetic relationships in a format which could be represented on reports and without requiring tree visualisation software or interpretation.

Discussion

Scaling with the pandemic

The pipelines described above have had to evolve considerably as the pandemic has progressed both in order to stay relevant to the questions being investigated by public health bodies, and in order to continue to scale with exponentially growing levels of data.

A relevant resource

In the early phase of the pandemic, key questions on a national and local health level focused on quantifying the number of introductions into an area and assessing subsequent spread (e.g. [25–27]). At this time, the pipeline automatically generated weekly reports that summarised the latest data at the national (UK-wide) level, within each of the four constituent countries of the UK (Wales, Scotland, Northern Ireland and England), and at a further regional level corresponding to several of the COG-UK sequencing partners. These reports included case counts of individual lineages as well as estimates of the numbers of new introductions and subsequent spread based on uk_lineage information. As cases rose the outputs fed into the COG-UK coverage maps used by government and as more bespoke investigative reports were in demand, we added specific pipeline outputs to support local report generation using CIVET [15], in addition to those that already existed to support Microreact [13].

Following the first lockdown and as the initial variants of concern (VOCs) emerged, the focus of investigations shifted to mutations, lineages and constellations for VOC and variants under investigation (VUI), rather than the previous focus on introductions. Relevant steps were added into the pipeline to type them and these classifications were fed into the COG-UK Mutation Explorer [12] and GRINCH [28].

Timely results

The initial pipeline grapevine (https://github.com/COG-UK/grapevine) was written in snakemake [19]. By January 2021, it became clear that the phylogenetics steps of the pipeline were becoming prohibitively slow. To enable the continued rapid dissemination of sequence data to the consortium, the pipeline was separated into a data processing pipeline https://github.com/COG-UK/datapipe which rapidly performed the initial alignment, variant calling and lineage assignment steps, and a phylogenetics pipeline https://github.com/cov-ert/phylopipe which consumed the output of the data processing pipeline and performed the tree building and post-processing steps. This allowed the data processing pipeline to run reliably every day, while the phylogenetics pipeline was allowed to run less frequently.

During this rewrite, we moved from using Snakemake workflow language to Nextflow. This was motivated by the observation that the Nextflow workflow manager seemed better able to handle issues arising on the SLURM [29] computing cluster resource manager due to the large resource requirements, for example when a node became unresponsive.

Scaling phylogenetic methods

The phylogenetic tree construction steps have also had to adapt considerably with the growth of global data. Initially we used IQTREE to estimate the global phylogeny [30,31]. Then, for speed we introduced a process of assigning sequences to three distinct lineages A, B

or B.1, estimating these trees independently with IQTREE, and subsequently grafting together these subtrees to form our global tree. By June 2020, after a series of benchmarking experiments, we adopted FastTree [16] as the inference engine for parsimony-based tree reconstruction and this method was sufficient for our needs for the rest of 2020, with new subtrees (representing emerging sub-lineages such as B.1.1) added when appropriate.

By January 2021, with the advent of variants of concern leading to a global surge in SARS-CoV-2 genome sequence generation, the pipeline was again struggling to complete regularly and it was not possible to parallelize further with new subtree splits. As a result, in February we began downsampling the data before tree-building. We initially subset to the previous 6 months, but as scaling continued to be a problem this was further restricted to 5 months, then 100 days, and finally to 30 days plus background data. Even with this heavy downsampling, construction of the split-grafted tree using FastTree at this time took more than 2 days, meaning an interim solution would be needed for "real-time" analysis. As such, we introduced daily tree updates with UShER [17], with less frequent tree rebuilds as and when the daily tree became unwieldy or gained errors.

Surprising bottlenecks

Because of the sheer scale of numbers of genome sequences, every aspect of the pipeline has been evaluated for both time and memory efficiencies. Simple processes such as reading and manipulating FASTA and metadata CSV files became significant bottlenecks because of how large these files had become. We found that the highly optimised pandas [32] dataframe library required too much memory and had to replace it with a custom metadata reader module based on the DictReader class from the python csv module. This custom suite of utility functions (https://github.com/cov-ert/fastafunk) instead streamed the metadata file twice and used set manipulation in order to hold minimal data in memory.

During the rewrite from grapevine to datapipe and phylopipe, we moved from a system where rows/sequences filtered at each stage were removed from the relevant files to one where the metadata table remained complete, with a column tracing why any given sequence had been eliminated from output FASTA. While there would have been performance gains from reducing metadata size, in practice, this method can easily log information for members of the consortium to know exactly why their sequences were missing from final metadata tables without becoming a time consuming tracing exercise for the pipeline maintainers.

Recommendations for next time

There are a number of design choices we would recommend for a centralised analysis pipeline in a future outbreak scenario. Firstly, genomic epidemiological analysis depends heavily on being able to index between genome sequence data and metadata: filtering to a

subset of sequences, applying some analysis and updating the metadata table with the results. We recommend working from the start either with a custom database, or with the most lightweight, optimised and well tested software libraries available. Secondly, we recommend using restricted metadata fields wherever possible (rather than free text) in order to remove an ongoing burden of maintenance. We also recommend that the first step in the pipeline is to check and clean sequences, sequence headers and metadata, including for non-unicode characters. Thirdly, we recommend building in sample traceability from the start, so that it is easy to identify why a given sample may not have been retained to the final analysis steps. Finally, the success of this pipeline was in our ability to adapt it to the most pressing analysis questions. We recommend designing code to be as clean and modular as possible, so that less relevant analysis steps can be removed and more relevant ones added over time, whilst retaining a consistency of output for downstream tools.

The impact of centralised phylogenetic analysis

The UK was one of the earliest countries to adopt a national genomic surveillance program for SARS-CoV-2 [10]. After the emergence of the Alpha variant, many more countries began to use genomics for surveillance. Many surveillance approaches make use of Nextstrain [9] builds, however these are heavily downsampled and only include a small subset of the data. Early in 2020, the sarscov2phylo public tree [33] or later the daily updated UCSC tree [14] provided a full global phylogeny of public data. However the early and extensive genomic sequencing within the UK, the detailed private metadata collected with governance to be shared within the distributed network of data users, and a requirement for numerous custom downstream analyses based on the full global phylogeny, all contributed to the requirement for a local pipeline.

The real power of performing these analyses centrally was that all members of the consortium were able to access detailed analysis and relevant information about their submitted data without extensive bioinformatics knowledge or the requirement for large computational resources to build a phylogenetic tree.

At the government level, simple representations of the data including the number of lineages over time fed national dashboards. COG-UK reports also fed into SAGE meetings. At the Public Health Agency level, unpublished summaries formed the basis of national surveillance projects and informed response (e.g. [34]). At the local or regional hospital level, the outputs enabled investigations of hospital onset SARS-CoV-2 infections [35] such as with the CIVET tool [15] informing outbreak management, and wider infection prevention and control measures. This tool directly accessed the output of initially grapevine, and later datapipe and phylopipe. Some of the resolution within these reports came from the uk_lineage and phylotype metadata fields. These provided a fine-scale text representation of the

phylogenetic relationships between samples in the uk and global tree which could be interpreted without specialist tree viewing software or bioinformatics expertise.

More generally, outputs from this centralised analysis pipeline were used in analyses to reveal the multiple introductions of SARS-CoV-2 from mainland Europe into Scotland in 2020 [26], and to show that Alpha variant was associated with increased clinical severity [36]. They were used to identify and verify early recombinant genomes [37] and to test lineage frequencies from wastewater surveillance sequencing by comparison with conventional surveillance sequencing in the same geographic location [38]. The phylogenetic trees were used to provide routine early tracking of emerging variants [39], investigate how genetic drift changes over time [40], and investigate the impact of viral mutations on recognition by T cells [41]. They also provided the means to select a targeted downsample for more in depth analyses, such as into the emergence and growth of the SARS-CoV-2 Delta variant in the UK [42].

Finally at a public level, data explorers such as COG-UK-ME [12], Microreact [13], GRINCH [28] and the UCSC Genome Browser phylogenetic tree [14] were able to ingest pipeline outputs and allow exploration of the vast data resource more widely. As a result, the outputs of this pipeline have fed into many applications and impacted analyses both within the UK and globally.

Data availability

All code is open-source and available under a GNU GPL-3.0 licence in the GitHub repositories https://github.com/COG-UK/datapipe and https://github.com/virus-evolution/phylopipe.

Funding

COG-UK was supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. CLIMB is funded by the Medical Research Council (MRC) through grant MR/L015080/1, with CLIMB-COVID also receiving funding from the UK Department of Health and Social Care. AR, RC, JTM acknowledge support from the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network). AOT was supported by the Wellcome Trust Hosts, Pathogens & Global Health Programme (grant number: grant.203783/Z/16/Z) and Fast Grants (award number: 2236). TRC acknowledges funding from, and TW was supported by, the Wellcome Trust (Innovator Award: Digital Technologies, grant number 215800/Z/19/Z). VH was supported

by the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/M010996/1).

Acknowledgements

The authors would like to thank all those who generated sequences for their rapid data sharing as part of COG-UK. We also thank the research labs and public health bodies who made their genome data accessible on GISAID. We would like to thank Ben C. Jackson for the huge amount of work he put into development and maintenance of the pipeline in its first year. We thank Angie Hinrichs for her eagle eye and Derek Wright, Anthony Underwood, and Matt Bull for helpful suggestions along the way.

References

1. Sharp PM, Hahn BH. Origins of HIV and the AIDS Pandemic. Cold Spring Harb Perspect Med. 2011;1:a006841.
2. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459:1122–5.
3. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369–72.
4. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular Evidence of Sexual Transmission of Ebola Virus. N Engl J Med. 2015;373:2448–54.
5. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.
6. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546:401–5.
7. Sabir JSM, Lam TT-Y, Ahmed MMM, Li L, Shen Y, E. M. Abo-Aba S, et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science. 2016;351:81–4.
8. Kinganda-Lusamaki E, Black A, Mukadi DB, Hadfield J, Mbala-Kingebeni P, Pratt CB, et al. Integration of genomic sequencing into the response to the Ebola virus outbreak in Nord Kivu, Democratic Republic of the Congo. Nat Med. 2021;27:710–6.
9. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3.
10. An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe. 2020;1:e99–100.
11. Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, et al. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. Genome Biology. 2021;22:196.
12. Wright DW, Harvey WT, Hughes J, Cox M, Peacock TP, Colquhoun R, et al. Tracking SARS-CoV-2 mutations and variants through the COG-UK-Mutation Explorer. Virus Evolution. 2022;8:veac023.
13. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microbial

Genomics [Internet]. 2016 [cited 2023 Jun 5];2. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000093
14. McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, et al. A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. Mol Biol Evol. 2021;38:5819–24.
15. O'Toole Á, Hill V, Jackson B, Dewar R, Sahadeo N, Colquhoun R, et al. Genomics-informed outbreak investigations of SARS-CoV-2 using civet. PLOS Global Public Health. 2022;2:e0000704.
16. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5:e9490.
17. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 2021;53:809–16.
18. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
19. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake [Internet]. F1000Research; 2021 [cited 2023 Jun 5]. Available from: https://f1000research.com/articles/10-33
20. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution. 2021;7:veab064.
21. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
22. Jackson B. gofasta: command-line utilities for genomic epidemiology research. Bioinformatics. 2022;38:4033–5.
23. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017;1:33–46.
24. Jukes TH, Cantor CR. CHAPTER 24 - Evolution of Protein Molecules. In: Munro HN, editor. Mammalian Protein Metabolism [Internet]. Academic Press; 1969 [cited 2024 Aug 13]. p. 21–132. Available from: https://www.sciencedirect.com/science/article/pii/B9781483232119500097
25. Plessis L du, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371:708–12.
26. da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, et al. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. Nat Microbiol. 2021;6:112–22.
27. Lycett SJ, Hughes J, McHugh MP, Filipe A da S, Dewar R, Lu L, et al. Epidemic waves of COVID-19 in Scotland: a genomic perspective on the impact of the introduction and relaxation of lockdown on SARS-CoV-2. medRxiv. 2021;2021.01.08.20248677.
28. O'Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. Wellcome Open Res. 2021;6:121.
29. Jette MA, Wickberg T. Architecture of the Slurm Workload Manager. In: Klusáček D, Corbalán J, Rodrigo GP, editors. Job Scheduling Strategies for Parallel Processing. Cham: Springer Nature Switzerland; 2023. p. 3–23.
30. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32:268–74.
31. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020;37:1530–4.
32. McKinney W. Data Structures for Statistical Computing in Python. Austin, Texas; 2010 [cited

2024 Jan 22]. p. 56–61. Available from: https://conference.scipy.org/proceedings/scipy2010/mckinney.html
33. roblanf, Mansfield R. roblanf/sarscov2phylo: 13-11-20 [Internet]. Zenodo; 2020 [cited 2024 Aug 14]. Available from: https://zenodo.org/records/4289383
34. COG-UK SAGE Reports [ACHIVED} [Internet]. [cited 2024 May 31]. Available from: https://webarchive.nationalarchives.gov.uk/ukgwa/20220701104759/https://www.cogconsortium.uk/about/archive/sage-reports/
35. Stirrup O, Hughes J, Parker M, Partridge DG, Shepherd JG, Blackstone J, et al. Rapid feedback on hospital onset SARS-CoV-2 infections combining epidemiological and sequencing data. Osterhaus A, van der Meer JW, Osterhaus A, editors. eLife. 2021;10:e65828.
36. Pascall DJ, Vink E, Blacow R, Bulteel N, Campbell A, Campbell R, et al. The SARS-CoV-2 Alpha variant was associated with increased clinical severity of COVID-19 in Scotland: A genomics-based retrospective cohort analysis. PLOS ONE. 2023;18:e0284187.
37. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 2021;184:5179-5188.e8.
38. Brunner FS, Payne A, Cairns E, Airey G, Gregory R, Pickwell ND, et al. Utility of wastewater genomic surveillance compared to clinical surveillance to track the spread of the SARS-CoV-2 Omicron variant across England. Water Research. 2023;247:120804.
39. Drake KO, Boyd O, Franceschi VB, Colquhoun RM, Ellaby NAF, Volz EM. Phylogenomic early warning signals for SARS-CoV-2 epidemic waves. eBioMedicine [Internet]. 2024 [cited 2024 May 2];100. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00505-4/fulltext
40. Yu Q, Ascensao JA, Okada T, Consortium TC-19 GU (COG-U, Boyd O, Volz E, et al. Lineage frequency time series reveal elevated levels of genetic drift in SARS-CoV-2 transmission in England. PLOS Pathogens. 2024;20:e1012090.
41. De Silva TI, Liu G, Lindsey BB, Dong D, Moore SC, Hsu NS, et al. The impact of viral mutations on recognition by SARS-CoV-2 specific T cells. iScience. 2021;24:103353.
42. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. Nature. 2022;610:154–60.

Figure 1: