Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Epidemics 39 (2022) 100576

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Advancing genomic epidemiology by addressing the bioinformatics bottleneck: Challenges, design principles, and a Swiss example

Chaoran Chen[1], Sarah Nadeau[1], Ivan Topolsky, Niko Beerenwinkel, Tanja Stadler *

Department of Biosystems Science and Engineering, ETH Zürich, Basel, CH 4058, Switzerland
Swiss Institute of Bioinformatics, Lausanne, CH 1015, Switzerland

ARTICLE INFO

Keywords:
Genomic epidemiology
SARS-CoV-2
Data infrastructure
Relational database
Microservices

ABSTRACT

The SARS-CoV-2 pandemic led to a huge increase in global pathogen genome sequencing efforts, and the resulting data are becoming increasingly important to detect variants of concern, monitor outbreaks, and quantify transmission dynamics. However, this rapid up-scaling in data generation brought with it many IT infrastructure challenges. In this paper, we report about developing an improved system for genomic epidemiology. We (i) highlight key challenges that were exacerbated by the pandemic situation, (ii) provide data infrastructure design principles to address them, and (iii) give an implementation example developed by the Swiss SARS-CoV-2 Sequencing Consortium (S3C) in response to the COVID-19 pandemic. Finally, we discuss remaining challenges to data infrastructure for genomic epidemiology. Improving these infrastructures will help better detect, monitor, and respond to future public health threats.

0. Introduction

An increasingly important tool to help fight pathogenic diseases is genomic epidemiology. The analysis of pathogen genome sequences allows us to learn about pathogen evolution and epidemic or endemic transmission dynamics (Kraemer et al., 2019; Grenfell et al., 2004). However, the SARS-CoV-2 pandemic has highlighted a growing disparity between global sequencing data generation capacities and analysis capacities (Black et al., 2020). As Hodcroft et al. (2021) underscores, we seem to be drowning in data rather than swimming in information.

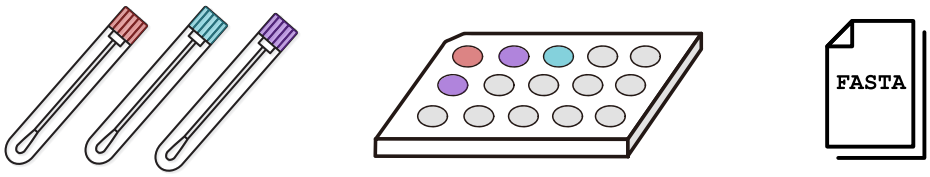
Genome sequence data are becoming increasingly important for epidemic response, as highlighted during the SARS-CoV-2 pandemic. In December 2019, when an unknown respiratory disease was identified in Wuhan, China, the first whole genome sequence from the causal virus helped classify the new human pathogen SARS-CoV-2 (Wu et al., 2020) and establish its likely origins (Andersen et al., 2020). Then, comparison of mutational differences in genomes collected from different regions helped distinguish imported cases from community transmission (Worobey et al., 2020). Next, genome surveillance efforts identified more transmissible variants of concern, e.g. the alpha variant (World Health Organization, 2021) in the UK in late 2020 (Volz et al., 2021). Finally, phylogenetic and phylodynamic methods use genome sequence data to quantify epidemic dynamics, including the reproductive number, transmission routes, effects of public health measures, and the role of super-spreading (Nadeau et al., 2021; Du Plessis et al., 2021; Miller et al., 2020). Thus, pathogen genome sequence data is instrumental for disease detection, outbreak tracking, and quantifying transmission dynamics.

The wealth and geographic distribution of available genomic data underlying these and other analyses indicates many groups around the world have developed their own infrastructures for genomic epidemiology. So far, several large national initiatives have published descriptions of their technical infrastructures. In particular, (Nicholls et al., 2021; Matthews et al., 2018; Egli et al., 2019) describe UK-, Canadian- and Swiss-specific infrastructures that enable linking of genome sequence data with associated metadata and integrate data from multiple regional contributors. Other examples are available as code bases, for instance that of the Spanish SARS-CoV-2 Sequencing Consortium (Spanish SARS-CoV-2 sequencing consortium, 2022).

Despite these successes, developing a data infrastructure for genome-based surveillance and genomic epidemiology remains a challenge (Black et al., 2020; Bernasconi et al., 2021). In the COVID-19 pandemic, bioinformatics capacity has proven to be a key bottleneck in pandemic response (Hodcroft et al., 2021). This is particularly true in countries without a well-supported national initiative, or in the period before such an initiative is established. As a US-focused report (Committee on Data Needs to Monitor Evolution of SARS-CoV-2 et al., 2020) highlights, a key priority for pandemic preparedness is to improve upon

* Corresponding author at: Department of Biosystems Science and Engineering, ETH Zürich, Basel, CH 4058, Switzerland.
E-mail address: tanja.stadler@bsse.ethz.ch (T. Stadler).
[1] These authors contributed equally.

https://doi.org/10.1016/j.epidem.2022.100576
Received 21 January 2022; Received in revised form 5 April 2022; Accepted 5 May 2022
Available online 14 May 2022



test

Sample No.	City
●	Basel
●	Basel
●	Zurich

plate

Plate No.	Seq. lab
1	GFB
2	GFB
3	FGCZ
4	H2030

sequence

Seq. No.	Sequence
1	TCTTTG
2	ACTTCG
3	TCTTCG
4	NNTTCG

test_plate_sequence_mapping

Sample No.	Plate No.	Well	Seq. No.
●	1	A1	1
●	1	A2	2
●	1	A3	3
●	1	B1	4

Fig. 1. An illustration of how three key entities – tests, plates, and sequences – are stored in database tables and the mapping table that links the information from each.

existing systems to integrate clinical and genomic data and better coordinate between different public health stakeholders. In this paper, we share lessons learned in the Swiss SARS-CoV-2 Sequencing Consortium (S3C) pertaining to three challenges that were particularly exacerbated by the COVID-19 pandemic: unstable data sources, rapid development of new tools, and the need for timely reporting. We outline design principles to address these challenges and describe our implementation of a relational database and containerized microservices as an example. Finally, we highlight remaining challenges in data management for genomic epidemiology.

The S3C began generating and analyzing SARS-CoV-2 genome sequences in March 2020. The Consortium started as a partnership between two academic groups, an associated academic sequencing facility, and a large Swiss medical diagnostics company (S3C, 2021). Since then, S3C has partnered with three core sequencing facilities in Switzerland to sequence over 44,000 samples from companies, hospitals, and research institutions. These data are made available on GISAID (Elbe and Buckland-Merrett, 2017) and the European Nucleotide Archive. To meet the demands of a growing genomic surveillance program in Switzerland, S3C benefited from early data infrastructure design choices that enabled rapid extension to new data sources, types, and users.

In the following sections we describe major implementation challenges for data infrastructure in light of the pandemic and outline design principles to address them. In particular, we discuss S3C's implementation of a relational database and microservices-based approach as an example fulfilling these design criteria using open source tools. Finally, we consider remaining challenges in data infrastructure for genomic epidemiology that must be met to improve future public health response to pathogenic diseases.

1. Unstable data sources

Emerging public health threats bring great uncertainties, including in data availability and formats. The basic data necessary for genomic surveillance are pathogen genome sequences and minimal patient metadata, e.g., sample collection date and location. Coupling these data and analyzing them in aggregate allows public health officials to track transmission and monitor key mutations. However, the format of these data may shift over the course of an outbreak, and new data may become available. For example, accommodating genomic restructuring by the pathogen itself (e.g., by insertion, deletion, recombination, or reassortment), annotating samples with the presence or absence of newly discovered key mutations, and newly available or re-formatted metadata all represent shifts in the basic data required for effective genomic surveillance. Furthermore, it might not be possible to define a fixed and sensible file format for data exchange in the early stages of outbreak response due to time pressure.

Recommendation: ensure clean data

Unreliable and shifting source data can quickly lead to messy data with, for example, missing values and different spellings of the same entity. Ideally, infrastructure developers will work with data submitters to develop a standardized data dictionary with clearly defined permitted values for each variable. However, it is also essential to strictly validate data upon import as a double-check. It should also be anticipated that changes and corrections to the data will be necessary over time. Therefore, data should be maintained in a non-redundant form so that changes to one attribute can be easily made without the danger of causing inconsistencies. Data relations should be tracked so that the effect of changes to one attribute on others are easy to identify. Data types should be strictly enforced so that changes to data formats are rapidly detected and mistakes are not incorporated. Finally, it should be easy to define custom data types and add attributes as new data is made available.

Example: relational database

Relational database management systems provide a good way to fulfill these design criteria. In a relational database management system, data are stored in a collection of tables, also known as the "relational format". Each table is independent from the others, but they may be linked (related) via shared keys, i.e. information common to two or

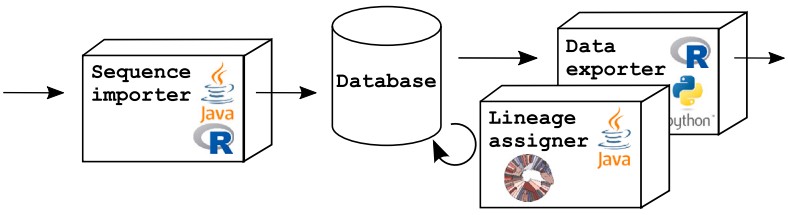


Fig. 2. Containerized microservices operate autonomously to add or extract data from the database.

more tables. This allows us to formulate complex queries by joining different tables together.

A relational database approach helps keep data clean in the face of unstable data sources. Each table's columns have fixed data types and it is possible to define custom types with a limited set of allowed values. Foreign keys, CHECK constraints and triggers allow definitions of arbitrarily complex validations. Invalid entries are rejected upon import so we know when corrections are necessary. This is especially important in the S3C, since we accept partially human-edited Excel files and non-documented output data from PCR machines as input. Non-redundancy between tables makes it easier to correct mistakes in these data when they arise. Finally, new and corrected data is simultaneously available to all database users.

Several relational database management systems are available. The S3C uses PostgreSQL,[2] which is freely available and open-source. In our implementation, we have three core database tables, one each for tests (samples), plates of RNA extracts, and SARS-CoV-2 genome sequences (Fig. 1). The test table contains sample metadata from the originating laboratory, the plate table tracks where each plate was sent for sequencing and when, and the sequence table stores the assembled SARS-CoV-2 whole-genome sequence and associated quality control statistics. Finally, a mapping table links the respective keys from each table. These tables represent the core of our database, though we have added other tables through time to accommodate new data. For example, we store the identifiers assigned by public databases and additional sample metadata provided by the Swiss Federal Office for Public Health (FOPH).

2. New tools

State-of-the art computational tools are also likely to change or are even being newly developed over the course of a public health response. This is exemplified in the COVID-19 pandemic by evolving nomenclature systems. Lineage assignment tools were frequently updated to keep up with nomenclature changes as new lineages arose. For example, the popular pangolin software for assigning SARS-CoV-2 genome sequences to global lineages has 75 releases since its development in April 2020 (O'Toole et al., 2021).

Recommendation: modular analysis workflows

Analysis workflows should be modular, rather than monolithic pipelines. It should be easy to update one component or swap it out for a different tool without having to re-run a full suite of analysis programs on the entire cohort. This modular structure allows individual components to be adapted or re-used for other pathogens or other projects. For use cases where software version tracking is especially important, workflow and software versions can be stored alongside the data in the database.

Example: containerized microservices

A microservices approach separates different tasks performed by different tools into loosely-coupled programs that operate autonomously, each performing a single, well-defined task. For the S3C, we implemented a growing set of microservices that import, export, and process data by adding or extracting data from the database (Fig. 2). The microservices each have their own code base, and, depending on the task, they are written in different languages.

We used a containerization technology to deploy these microservices. This packages software applications together with their dependencies into single units, called containers. For example, a Pango lineage assigner requires the pangolin tool (O'Toole et al., 2021), a Nextclade importer needs Nextclade (Aksamentov et al., 2021), and the metadata importer has to mount a network folder. The services can be written in different programming languages, perhaps even different versions of the same language to accommodate different dependencies.

Most services act only upon missing data. For example, we have a Nextclade importer service that runs the Nextclade program and imports resulting quality scores and mutations. This service queries the database every ten minutes and looks for entries in the sequence table where Nextclade quality scores were previously unpopulated. Other services avoid redundancy by maintaining a database table that stores a state, e.g. filenames which have already been processed and should not be re-imported. For example, our metadata importer service operates in this way.

The containerized microservices allow fast adoption of new or updated tools. Since they are packaged and deployed independently, they can be started or stopped without impacting other services. The containerization further serves to isolate each tool and remove dependency conflicts between tools. Finally, since services only act upon missing data or when a state is changed, we avoid redundant computation. Another complementary approach to achieving analysis modularity would be to use scientific workflow systems, such as Snakemake (Mölder et al., 2021) or Nextflow (Di Tommaso et al., 2017). These systems can be used together with containerization technologies and further simplify tracking of component software versions and workflow revisions used to generate output files.

3. Timely reporting

Timely reporting is crucial for an evidence-based public health response. Turn-around times for SARS-CoV-2 sequences to be made available on GISAID vary from a few days to a few weeks post-sampling, or more. Sample transport logistics, sequencing capacities, bioinformatics analysis, and report preparation all contribute to this turn-around time. Here, we focus on how to ensure rapid final reporting, as this is the aspect data managers have the most influence on.

Recommendation: Multiple levels of querying

A data management system needs to support rapid, ad-hoc querying in addition to generation of regular, stable reports. The prior is necessary for early outbreak detection and detection of new variants of concern, while the latter is essential for longer-term monitoring. Ideally, the system should be able to expose an application programming interface (API) for safe public data sharing.

[2] https://www.postgresql.org/.

```
select t.sample_no, t.city, p.plate, p.well
from
   test t
   join plate p on t.sample_no = p.sample_no
   join sequence s on p.plate = s.plate and p.well = s.well
   join mutation m on s.sequence_id = m.sequence_id
where
   m.mutation = 'S:N501Y';
```

Sample No.	City	Plate	Well
100	Basel	1	A1
101	Zurich	1	A2
101	Zurich	1	B1

Fig. 3. A SQL query that finds the samples with the S:N501Y mutation.

Example: Database queries

Relational database systems support querying in several ways, fulfilling the above design criteria. One way to interact with data in a relational database is by directly using structured query language (SQL), which is a high-level and declarative language specifically designed for efficient querying. In SQL, the user describes (declares) what data should be added or retrieved, but not exactly how. The language then works behind-the-scenes to optimize the necessary computations and return the desired information (Fig. 3). SQL is widely used by data analysts and does not require prior programming experience. Graphical user interfaces, for example DataGrip,[3] allow users to manually add or modify data and submit queries. For those who are programmers, popular languages like R and python have packages like dplyr and pandas that enable reading data from a database directly into data frames.

For recurring queries, for instance for regular reporting, the database enables easy aggregation and reporting using "views". These are derived tables that aggregate data from existing tables according to a query. For reporting purposes, we created a number of views, for instance a billing view that contains the number of sequenced and submitted samples per week and a surveillance view that aggregates per-sample lineage assignment and mutation information for the Swiss FOPH. These views are automatically updated with the correction or addition of data. We also have a microservice that exports the mutation information view on a daily basis to a drop-point for the Swiss FOPH.

Finally, for monitoring purposes, a relational database can also serve as the back-end to dashboards or websites. We offer two public-facing websites to interact with sequencing and case data stored in our database. One is a dashboard focused on Swiss case data[4] and the other enables monitoring of global SARS-CoV-2 variants [5] (Chen et al., 2021).

Discussion

The COVID-19 pandemic has underscored both the utility of genomic epidemiology for public health response and remaining challenges in supporting related data infrastructure. Here we highlighted three challenges that were exacerbated by the rapidly changing pandemic situation: unstable data sources, rapid development of new tools, and the need for timely reporting. Then, we outlined general design principles to address these challenges. As an example, we describe the S3C's implementation of a relational database and containerized microservices.

These design choices directly enabled genome-based outbreak detection, monitoring, and public health response in the Swiss SARS-CoV-2 epidemic. Even before a new variant could be reliably called by lineage classification tools, we could quickly query Swiss data for mutations characterizing variants of concern. This enabled us to detect the first instances of the Beta, Gamma, and Delta variants in Switzerland. Our database also enabled us to quickly develop two public-facing websites for epidemic monitoring. Finally, we collaborate with the Swiss FOPH as members of the Swiss National COVID-19 Science Task Force[6] to link genome sequences to patient metadata. Lineage assignment and mutation data are passed back to the FOPH to support the health authorities in their pandemic response.

Many labs around the world have developed a data infrastructure for genomic epidemiology over the course of the COVID-19 pandemic. In fact, there are over 4000 unique submitting labs in the GISAID EpiCoV database as of January 2022. Unfortunately, a paucity of published examples makes it difficult to compare the strengths and weaknesses of various implementations in light of the challenges outlined by Black et al. (2020), Bernasconi et al. (2021) and highlighted here. The largest pathogen genome sequencing consortium in the world is that of COG-UK. Like S3C, they use a relational database. On top of it, they developed an API and a web interface for the collaborators to submit and retrieve data (Nicholls et al., 2021). In comparison, we did not define a fixed metadata or sequence data format but adapted to the data provided by collaborators. Our aim was to reduce overhead for our collaborators. However, as data inputs stabilize, a future improvement would be to develop a more robust procedure for defining formats and updating data. An improved technical interface for data upload and correction by sequence submitters like that of COG-UK would also help.

There are also larger outstanding challenges to developing data infrastructures for genomic epidemiology. First, genome sequencing efforts are highly skewed towards high-income countries. In an interconnected world, local variants and fast epidemic spread are of global concern no matter where they arise. Expanding the technical and personnel resources for genome sequencing and data management in low and middle-income countries would enable a better, more coordinated public health response. Second, mistakes are common — from

[3] https://www.jetbrains.com/datagrip/.
[4] https://ibz-shiny.ethz.ch/covidDashboard/?_inputs_&tab=%22ts%22.
[5] https://cov-spectrum.org.
[6] https://sciencetaskforce.ch.

sequencing errors introducing spurious mutations, to sample contamination, to metadata errors. SARS-CoV-2 sequences and their metadata are regularly modified or deleted from public repositories. While some amount of mistakes are inevitable, better tools for tracking of changes to sequence data and their metadata would make correcting mistakes easier and promote reproducible science and transparency. Finally, we need robust infrastructures for safe linking of patient metadata with genome data. It can be a challenge to establish standardized, anonymized identifiers at the relevant scale for national sequencing projects, particularly in countries with decentralized health care services. Strong partnerships with government health ministries will help here, with metadata like vaccination and hospitalization status being provided to ensure actionable results for public health response.

In conclusion, generating pathogen genome sequence data and linking it to case-level metadata facilitates a rapid, evidence-based public health response to evolving infectious pathogens. Effective and timely generation of these data in rapidly changing situations relies on robust and agile data infrastructures, and improvements in the area should be a priority for pandemic preparedness.

Funding

TS, SN and CC are supported by the Swiss National Science Foundation (grant number 31CA30_196267). NB and IT are supported by the SIB Swiss Institute of Bioinformatics.

CRediT authorship contribution statement

Chaoran Chen: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Sarah Nadeau:** Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Ivan Topolsky:** Data curation, Resources, Software, Writing – review & editing. **Niko Beerenwinkel:** Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Tanja Stadler:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code availability

Our code is openly available under the LGPL-license on GitHub at https://github.com/cevo-public/harvester-database-and-automation.

References

Aksamentov, I., Roemer, C., Hodcroft, E.B., Neher, R.A., 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. 6 (67), 3773. http://dx.doi.org/10.21105/joss.03773.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 2020 26:4 (ISSN: 1546-170X) 26, 450–452, https://www.nature.com/articles/s41591-020-0820-9,

Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., Ceri, S., 2021. A review on viral data sources and search systems for perspective mitigation of COVID-19. Brief. Bioinform. (ISSN: 1477-4054) 22 (2), 664–675. http://dx.doi.org/10.1093/bib/bbaa359.

Black, A., MacCannell, D.R., Sibley, T.R., Bedford, T., 2020. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat. Med. (ISSN: 1546-170X) 26 (6), 832–841. http://dx.doi.org/10.1038/s41591-020-0935-z.

Chen, C., et al., 2021. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics (ISSN: 1367-4803) 38 (6), 1735–1737. http://dx.doi.org/10.1093/bioinformatics/btab856.

Committee on Data Needs to Monitor Evolution of SARS-CoV-2, et al., 2020. Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies. National Academies Press, Washington, D.C., ISBN: 978-0-309-68091-2, http://dx.doi.org/10.17226/25879, Pages: 1-110, https://www.nap.edu/catalog/25879.

Spanish SARS-CoV-2 sequencing consortium, 2022. FISABIO-NGS / SARS-CoV2-mapping. https://gitlab.com/fisabio-ngs/sars-cov2-mapping.

Di Tommaso, P., et al., 2017. Nextflow enables reproducible computational workflows. Nature Biotechnol. (ISSN: 1546-1696) 35 (4), 316–319. http://dx.doi.org/10.1038/nbt.3820.

Du Plessis, L., et al., 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 371 (6530), 708–712. http://dx.doi.org/10.1126/science.abf2946.

Egli, A., et al., 2019. Improving the quality and workflow of bacterial genome sequencing and analysis: paving the way for a Switzerland-wide molecular epidemiological surveillance platform. Swiss Med. Weekly (49), https://smw.ch/article/doi/smw.2018.14693.

Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Chall. 1 (1), 33–46, https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.1018.

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., Holmes, E.C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303 (5656), 327–332. http://dx.doi.org/10.1126/science.1090727.

Hodcroft, E.B., et al., 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature 2021 591:7848 (ISSN: 14764687) 591, 30–33. http://dx.doi.org/10.1038/d41586-021-00525-x.

Kraemer, M.U.G., et al., 2019. Reconstruction and prediction of viral disease epidemics. Epidemiol. Infect. 147, e34. http://dx.doi.org/10.1017/S0950268818002881.

Matthews, T.C., et al., 2018. The integrated rapid infectious disease analysis (IRIDA) platform. BioRxiv http://dx.doi.org/10.1101/381830.

Miller, D., et al., 2020. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nature Commun. (ISSN: 2041-1723) http://dx.doi.org/10.1038/s41467-020-19248-0.

Mölder, F., et al., 2021. Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. F1000Research 10 (33), http://dx.doi.org/10.12688/f1000research.29032.2.

Nadeau, S.A., Vaughan, T.G., Scire, J., Huisman, J.S., Stadler, T., 2021. The origin and early spread of SARS-CoV-2 in Europe. Proc. Natl. Acad. Sci. (ISSN: 0027-8424) 118, http://dx.doi.org/10.1073/PNAS.2012008118.

Nicholls, S.M., et al., 2021. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. Genome Biol. (ISSN: 1474-760X) 22 (1), 196. http://dx.doi.org/10.1186/s13059-021-02395-y.

O'Toole, A., et al., 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. (ISSN: 2057-1577) 7 (2), http://dx.doi.org/10.1093/ve/veab064, veab064.

S3C, 2021. Swiss SARS-CoV-2 sequencing consortium (S3C). https://bsse.ethz.ch/cevo/research/sars-cov-2/swiss-sars-cov-2-sequencing-consortium.html.

Volz, E., et al., 2021. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 2021 593:7858 (ISSN: 1476-4687) 593, 266–269. http://dx.doi.org/10.1038/s41586-021-03470-x.

World Health Organization, 2021. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.

Worobey, M., et al., 2020. The emergence of SARS-CoV-2 in Europe and North America. Science 370 (6516), 564–570, https://www.science.org/doi/abs/10.1126/science.abc8169.

Wu, F., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature (ISSN: 14764687) 579, 265–269. http://dx.doi.org/10.1038/s41586-020-2008-3.