MICROBIAL GENOMICS

BIORESOURCE
Rahman *et al.*, *Microbial Genomics* 2024;10:001188
DOI 10.1099/mgen.0.001188

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Mobilisation and analyses of publicly available SARS-CoV-2 data for pandemic responses

Nadim Rahman[1,*,†], Colman O'Cathail[1,*,†], Ahmad Zyoud[1], Alexey Sokolov[1], Bas Oude Munnink[2], Björn Grüning[3], Carla Cummins[1], Clara Amid[2], David F. Nieuwenhuijse[2], Dávid Visontai[4], David Yu Yuan[1], Dipayan Gupta[1], Divyae K. Prasad[2], Gábor Máté Gulyás[5], Gabriele Rinck[1], Jasmine McKinnon[1], Jeena Rajan[1], Jeff Knaggs[1], Jeffrey Edward Skiby[5], József Stéger[4], Judit Szarvas[5], Khadim Gueye[1], Krisztián Papp[4], Maarten Hoek[2], Manish Kumar[1], Marianna A. Ventouratou[1], Marie-Catherine Bouquieaux[2], Martin Koliba[5], Milena Mansurova[1], Muhammad Haseeb[1], Nathalie Worp[2], Peter W. Harrison[1], Rasko Leinonen[1], Ross Thorne[1], Sandeep Selvakumar[1], Sarah Hunt[1], Sundar Venkataraman[1], Suran Jayathilaka[1], Timothée Cezard[1], Wolfgang Maier[3], Zahra Waheed[1], Zamin Iqbal[1], Frank Møller Aarestrup[5], Istvan Csabai[4], Marion Koopmans[2], Tony Burdett[1] and Guy Cochrane[1,*]

Abstract

The COVID-19 pandemic has seen large-scale pathogen genomic sequencing efforts, becoming part of the toolbox for surveillance and epidemic research. This resulted in an unprecedented level of data sharing to open repositories, which has actively supported the identification of SARS-CoV-2 structure, molecular interactions, mutations and variants, and facilitated vaccine development and drug reuse studies and design. The European COVID-19 Data Platform was launched to support this data sharing, and has resulted in the deposition of several million SARS-CoV-2 raw reads. In this paper we describe (1) open data sharing, (2) tools for submission, analysis, visualisation and data claiming (e.g. ORCiD), (3) the systematic analysis of these datasets, at scale via the SARS-CoV-2 Data Hubs as well as (4) lessons learnt. This paper describes a component of the Platform, the SARS-CoV-2 Data Hubs, which enable the extension and set up of infrastructure that we intend to use more widely in the future for pathogen surveillance and pandemic preparedness.

DATA SUMMARY

The data described within this paper have all been shared and archived at the European Nucleotide Archive (ENA) at EMBL-EBI. Data can be found through the umbrella project accession: PRJEB45555 (https://www.ebi.ac.uk/ena/browser/view/PRJEB45555), or through the COVID-19 Data Portal: https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-analysis-covid19&size=15. More details on the structure of data can be found in Archival and Data Availability within the Results.

Received 04 May 2023; Accepted 14 January 2024; Published 15 February 2024

Author affiliations: [1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK; [2]Erasmus Medical Center, Wytemaweg 80, 3015 CN Rotterdam, Netherlands; [3]University of Freiburg, Friedrichstr. 39, 79098 Freiburg, Germany; [4]Eötvös Loránd University, H-1053 Budapest, Egyetem tér 1-3, Hungary; [5]Technical University of Denmark, Anker Engelunds Vej 101, 2800 Kongens Lyngby, Denmark.
***Correspondence:** Nadim Rahman, rahman@ebi.ac.uk; Colman O'Cathail, ocathail@ebi.ac.uk; Guy Cochrane, cochrane@ebi.ac.uk

Keywords: SARS-CoV-2; COVID-19; Open data; Data sharing; Genomics sequencing; Emerging variants.

Abbreviations: AF, allele frequency; API, application programming interface; DOI, digital object identifier; DTU, Technical University of Denmark; EBI, European Bioinformatics Institute; ELTE, Eötvös Loránd University; EMBL, European Molecular Biology Laboratory; EMC, Erasmus Medical Centre; ENA, European Nucleotide Archive; FAIR, findable, accesible, interoperable, reusable; FEGA, Federated European Genome-phenome Archive; GISAID, global initiative on sharing all influenza data; INSDC, International Nucleotide Sequence Database Collaboration; ORCiD, open researcher and contributor identifier; PAS, pathogen analysis system; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; VOC, variant of concern; VOI, variant of interest; WHO, World Health Organisation.

*These two authors contributed equally to this work

Data statement: One supplementary material are available with the online version of this article.

Significance as a BioResource to the community

The public SARS-CoV-2 Data Hub described in this paper, presents an example collection of tools that can be used for targeted use-cases by other users in the community in sharing data into the underlying resource(s) of the COVID-19 Data Portal, part of the European COVID-19 Data Platform. The dataset described in the paper, presents one of the single largest analysed public datasets, accessible through the portal. On the whole, including big data analysis, providing additional data structure and contextualisation for interpretation by the wider community.

INTRODUCTION

On 27 December 2019, public health authorities of Wuhan, China were notified of a small number of pneumonia-like cases, seemingly linked to visits to a market. The cluster was quickly identified as being caused by a novel coronavirus, soon thereafter named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) due to its genetic similarity to SARS-CoV, the virus causing the SARS outbreak in 2003 [1]. Despite initial optimism that the outbreak could be controlled by enforcing stringent public health measures, the virus spread to different cities in the region, and soon cases presenting with the disease (Coronavirus Disease 2019 or COVID-19) emerged on every continent. By 11 March 2020, the World Health Organisation (WHO) declared this COVID-19 outbreak a pandemic [2], triggering a massive global response. One of the hallmarks of the public health response during this pandemic was the extensive application of pathogen genomic sequencing, which brought in scientific institutes and consortia to mobilise sequencing efforts and participate in the sharing of COVID-19 biodata [3, 4].

The European COVID-19 Data Platform (Fig. 1) was launched in April 2020, as part of the European Molecular Biology Laboratory’s (EMBL) response to the COVID-19 pandemic. This data platform comprises three main components. First, the COVID-19 Data Portal provides an interface to access and browse a wealth of publicly available datasets, tools and resources [5]. Second, the Federated European Genome-phenome Archive (FEGA) supports the sharing of sensitive human genotypic and phenotypic data [6]. Third, the SARS-CoV-2 Data Hubs, the topic of this paper, offer a toolbox to those working with viral sequence data to support management, sharing, analysis and interpretation. A particular focus of our work with the SARS-CoV-2 Data Hubs was to make public data more easily available to open research around the world, and to ensure the sharing of raw sequencing data in addition to consensus sequences, along with a host of other data types (as presented within

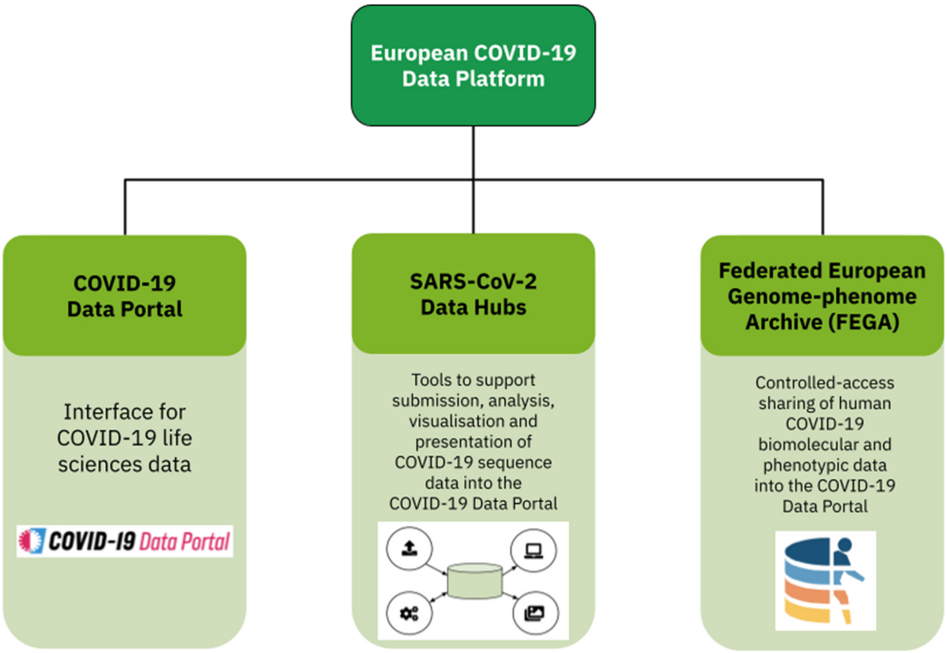

European COVID-19 Data Platform
COVID-19 Data Portal
Interface for COVID-19 life sciences data
COVID-19 Data Portal
SARS-CoV-2 Data Hubs
Tools to support submission, analysis, visualisation and presentation of COVID-19 sequence data into the COVID-19 Data Portal
Federated European Genome-phenome Archive (FEGA)
Controlled-access sharing of human COVID-19 biomolecular and phenotypic data into the COVID-19 Data Portal

Fig. 1. The European COVID-19 Data Platform and its three main components. 1) The COVID-19 Data Portal: the interface to a wealth of publicly available datasets relating to COVID-19. 2) The SARS-CoV-2 Data Hubs: the topic of this paper. 3) The FEGA: supporting controlled-access sharing of human related genotypic and phenotypic data.

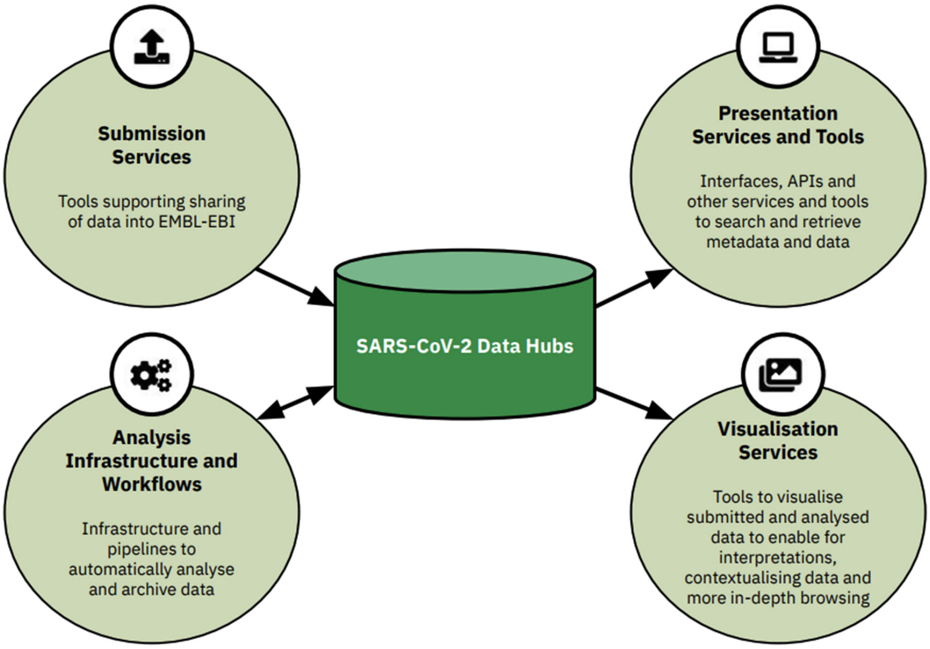

Submission Services
Tools supporting sharing of data into EMBL-EBI
Presentation Services and Tools
Interfaces, APIs and other services and tools to search and retrieve metadata and data
SARS-CoV-2 Data Hubs
Analysis Infrastructure and Workflows
Infrastructure and pipelines to automatically analyse and archive data
Visualisation Services
Tools to visualise submitted and analysed data to enable for interpretations, contextualising data and more in-depth browsing

Fig. 2. The SARS-CoV-2 Data Hubs and their components. This includes submission services, analysis infrastructure and workflows, presentation services and tools and visualisation services.

the COVID-19 Data Portal). By sharing raw sequencing data, re-analysis can be done allowing for evaluation of published consensus sequences, using other tools for assembly, minority variants determination, etc. This involved mobilisation of teams, infrastructure and data to support worldwide scientific research around SARS-CoV-2 sequence data.

The occurrence of a public health emergency at a time of progressively reducing sequencing costs has resulted in unprecedented levels of data sharing. Currently, nearly 25% (24.88%) of the raw sequence read records in the European Nucleotide Archive (ENA) are from SARS-CoV-2. Therefore, there was a need to facilitate submission and analysis of this explosively expanding dataset. Here we describe the development of a public SARS-CoV-2 Data Hub and its components, specifically designed to facilitate further analysis by the wider research community. We describe the community-owned/driven data hub model, the submission services and tools, analytical workflows, visualisation tools and search and retrieval via the COVID-19 Data Portal, in order to develop its potential usage for the general and more-targeted SARS-CoV-2 Data Hubs.

OVERVIEW

The SARS-CoV-2 Data Hubs (Fig. 2) build on a model conceived and developed through the COMPARE project [7]. This involved the provision of protected databases, in which groups of scientists could share pre-publication (private) data and the results of analyses with their collaborators prior to public data release. EMBL-EBI implemented a modular system centring around data storage modules, connecting to analysis compute, and including several components and tools, described further in the COMPARE Data Hubs [7]. To meet the need for rapid, large-scale data sharing at the height of the COVID-19 pandemic, a public SARS-CoV-2 Data Hub instance was established. This data hub builds on the components described in the COMPARE Data Hubs and includes a set of submission tools, analysis workflows and infrastructure, portals or interfaces for data and metadata search and retrieval, and visualisation tools. As data hubs are built upon the ENA [8], within EMBL-EBI [9] infrastructure, they enable integration with existing metadata and data models at the ENA and the connected databases of the International Nucleotide Sequence Database Collaboration (INSDC) [10]. The data hubs were built to allow easy re-use in the future, for other diseases and datasets, facilitating long-term sustainability and open science [7]. A specific focus has been on enabling the sharing of raw sequence reads to allow for standardised analysis of all public data, and future re-analyses using novel bioinformatic tools. The data hubs are configurable, enabling users to choose appropriate tools according to their needs. Datasets become the central focus, with tools being developed or mobilised. The development of key components of the data hubs was done in collaboration with partners in the VEO consortium [11]: Technical University of Denmark (DTU), Eötvös Loránd University (ELTE) of Hungary and Erasmus Medical Centre (EMC) of Netherlands.

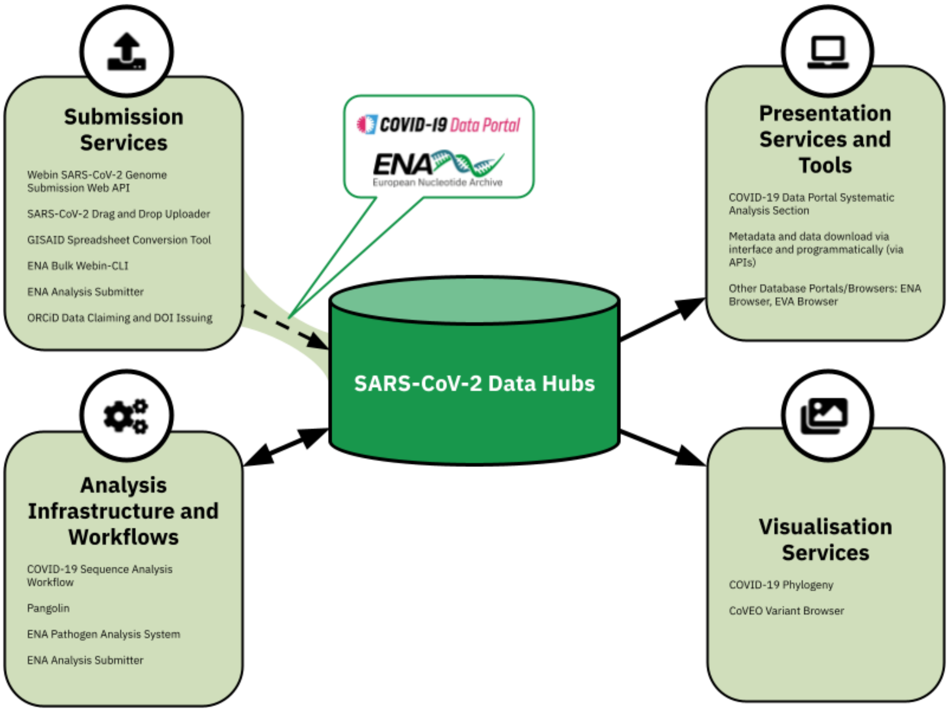

Submission Services
Webin SARS-CoV-2 Genome Submission Web API
SARS-CoV-2 Drag and Drop Uploader
GISAID Spreadsheet Conversion Tool
ENA Bulk Webin-CLI
ENA Analysis Submitter
ORCiD Data Claiming and DOI Issuing
COVID-19 Data Portal
ENA
European Nucleotide Archive
Presentation Services and Tools
COVID-19 Data Portal Systematic Analysis Section
Metadata and data download via interface and programmatically (via APIs)
Other Database Portals/Browsers: ENA Browser, EVA Browser
SARS-CoV-2 Data Hubs
Analysis Infrastructure and Workflows
COVID-19 Sequence Analysis Workflow
Pangolin
ENA Pathogen Analysis System
ENA Analysis Submitter
Visualisation Services
COVID-19 Phylogeny
CoVEO Variant Browser

Fig. 3. Diagram representing the developed tools mentioned in this paper and their contribution within the SARS-CoV-2 Data Hubs. As the data hubs sit on EMBL-EBI infrastructure, read data is submitted to and archived at the ENA, with a metadata view of this data provided by the COVID-19 Data Portal. Once shared to the ENA, this enables for usage as part of the data hubs.

METHODS

In April 2020, European Union president Ursula von der Leyen commissioned the development of the European COVID-19 Data Platform, an initiative aiming to leverage data sharing infrastructure for research during the pandemic. The VEO consortium assembled a taskforce to help address the software-based variability among shared genomic data. As part of the data mobilisation efforts for the European COVID-19 Data Platform, EMBL-EBI encouraged and supported mobilisation of raw read sequencing data from data providers, i.e. prior to any assembly and sharing of annotated genomes, as seen through (and in addition to) GISAID. Through meetings with users and data providers interested in sharing data, their challenges, feedback and requirements in handling the vast amount of data were identified, which helped drive tool development. Utilising this feedback, we developed, tested and reviewed tools to further support the mobilisation of data. Workflows, visualisation tools and components were developed for the analysis and interpretation of all mobilised data. These aspects supported the development of the public SARS-CoV-2 Data Hub, and included a focus on sustainability beyond the COVID-19 outbreak, to support re-use in any future outbreaks. This involved a large group of researchers and collaborators within the consortium, working on several different aspects, which have been described below.

Analysis workflow for processing raw read datasets

Systematic analysis

A flow for the systematic analysis of publicly submitted read data into the COVID-19 Data Portal was established, depicting an example of a public SARS-CoV-2 Data Hub (Figs 2 and 3). Various *submission services* at the ENA continue to be utilised for sharing and archiving multiple types of data publicly. This data became available in the public SARS-CoV-2 data hub as it is built upon the ENA infrastructure. A view of this public data and metadata is provided by the COVID-19 Data Portal. The *analysis* component of the data hub comprised two aspects, spanning first, the development, extension and integration of workflows and infrastructure for the analysis of incoming submitted raw read datasets. Second, an analysis management system, coupled with data flow management to effectively manage and track data analysis via the integrated workflows, whilst also sharing analysis results back to the ENA for dissemination via the presentation and visualisation tools. The COVID-19 Data Portal included a dedicated 'Systematic Analysis' page as a *presentation service*, with access to data and metadata download tools in the interface or programmatically. However as data was archived in the data hub (and therefore the ENA), the analysed data was also available via the ENA browser. Lastly, developed *visualisation tools* (along with their integration) enabled for support in browsing and interpreting the submitted and analysed data.

A key feature is that submitted raw sequence data were all uniformly processed through a sequencing platform-specific (Illumina or Oxford Nanopore) pipeline to generate consensus sequences and to identify mutations for variant calling. This resulted in approximately 80% of the total publicly submitted SARS-CoV-2 raw read data being available for analysis through the public SARS-CoV-2 data hub. The remainder of the public raw read datasets were generated using other sequencing platforms, such as PacBio. By using this approach, the data produced can be compared with confidence that variation introduced through implicit biases of different analysis approaches used in different laboratories is excluded. Therefore, when combined with quality thresholds and control for sequencing platforms used, variation observed between samples [12] most likely reflects true biological variation [13].

Workflows

COVID-19 sequence analysis workflow

Public read data submitted to the ENA were processed through the COVID-19 Sequence Analysis Workflow [14], if the metadata specified 'ILLUMINA' as the instrument platform. The workflow first trimmed reads using Trimmomatic [15] SLIDINGWINDOW:5:30 MINLEN:50. Reads were then mapped to the SARS-CoV-2 reference genome (NC_045512.2) using BWA-MEM [16]. Samtools [17] mpileup -Q 30 -d 8000 was used to generate a tab separated single base pair resolution coverage file. Variants were called from indexed and mapped reads using LoFreq [18]. Two different sets of variants were produced from this process, one that was unrestricted, and a second that only contained variants with an allele frequency (AF) greater than 0.25. Variants were annotated using SnpEff [19]. A consensus sequence was generated from these variant files using a custom python script [20].

Nanopore analysis workflow

Public raw read data submitted to the ENA were processed through the Nanopore Analysis Workflow [14] (incorporated within the COVID-19 Sequence Analysis Workflow), if the metadata specified 'OXFORD_NANOPORE' as the instrument platform. The workflow first trimmed the start and end of each read by 30 bps using cutadapt [21]. Trimmed reads were then mapped to the SARS-CoV-2 reference genome using Minimap2 [22]. Variants were called from the pileup results generated by the pysam python module, considering any variant occurring at more than a 10% variant frequency. Below this frequency the number of clearly erroneous variants (in particular single nucleotide deletions) increased dramatically. At the same time majority voting assigned variants as a 'MAJOR' variant by determining if the most abundant nucleotide at each position in the genome was the same or different from the reference. All called variants were stored in the unfiltered VCF, while only the 'MAJOR' variants were stored in the filtered VCF. The filtered VCF was in turn used to generate the consensus sequence by incorporating the variants in the reference genome and masking low coverage regions. All custom scripts to perform calling, filtering and consensus calling are available on GitHub [23].

Lineage assignment

To provide lineage annotations, the pangolin lineage assignment workflow [24] was integrated at EMBL-EBI to assign PANGO lineages using the pangolin [25] tool to both user-submitted and systematically analysed consensus sequences. In addition, World Health Organisation (WHO) variants of concern (VOCs) and variants of interest (VOIs) assignments were included, if applicable. The pangolin lineage assignment workflow integration was implemented in Snakemake [26].

Representative sequences

To support the wider scientific community in generating phylogenies or other comparative analyses, 'representative sequences' were identified [27]. This essentially provided a backbone of sequences, which users can retrieve, with a sequence per Pango lineage, allowing rapid assessment of placement of a newly submitted sequence. Here, sequences were selected based on a two step filtering process. First, the output of a Pangolin analysis of all consensus sequences archived at the ENA was parsed and sequences with a scorpio support value [25, 28] of less than 0.95 are removed. The scorpio support value was defined as the proportion of defining variants which have the alternative allele in the sequence. This step also removed low coverage sequences as these do not get assigned a scorpio support value. Second, the earliest appearance of a unique Pangolin lineage (e.g. A.1, B.1.1.7) in this list was selected based on the 'collection date' ENA metadata field, after first Sept 2020. This date was selected to filter out sequences with potentially erroneous collection dates that pre-date the earliest observation of a lineage. This was an automated curation process using a custom python script [29].

Visualisations

COVID-19 phylogeny

To visualise the evolution and lineages of SARS-CoV-2 from genomic data, a phylogenetic tree of submitted SARS-CoV-2 sequence data was set up using an upgraded version of the Evergreen phylogenetic analysis workflow [30]. The workflow [31] employed a reference based alignment strategy, where the input sequences were matched to references by genomic similarity calculated using KMA [32]. The workflow processed raw reads from single molecule sequencing technologies in addition to

Table 1. Requirements that were identified from discussions with stakeholders and users. This also includes tools to address trends and aspects of interest for the wider scientific community at the time, as noticed from literature and social media

Requirement	Description	Reason for requirement	Tool(s) developed
Support a more-bulk method to submit COVID-19 genomes.	To submit SARS-CoV-2 genomes, users were restricted to using Webin-CLI. This tool, although relatively fast and generally supportive towards larger scale submissions, was not efficient in support for high-throughput submissions to the levels seen during the COVID-19 pandemic.	To support the high volume of SARS-CoV-2 genomes from public health genome sequencing programmes.	Webin SARS-CoV-2 Genome Submission Web API ENA Bulk Webin-CLI Tool
Support for a simpler submissions interface.	Although interactive submissions to the ENA exist, this does require prior knowledge of the ENA submission process.	To provide a simple 'drag and drop' like tool to ease the submission process for prospective non-informatics based users.	SARS-CoV-2 Drag and Drop Uploader
Support large-scale analysis of SARS-CoV-2 raw read data.	To process the large amount of SARS-CoV-2 raw read data, sufficient management, scheduling, analysis, and finally submission of analysis products were required.	Appropriate scheduling, management, processing and data provenance/trail was required to adapt the process to the quantity of data.	ENA Pathogen Analysis System ENA Analysis Submitter
Support data claiming and attribution	When data is shared, affiliation and information detailing the data submitter or generator is held within and presented in data records.	Support researchers in data sharing by including additional methods in acknowledgement and attribution.	ORCiD Data Claiming Dataset DOIs
Annotate sequencing data in-line with WHO naming	Submitted SARS-CoV-2 genomes did not include lineage information as standard.	Support researchers in greater interpretation of data, and tailored searching for datasets of interest.	Integrated Pangolin pipeline setup
Tools to better support users in contextualising, visualising and interpreting genomic data	For SARS-CoV-2, there were no tools integrated into the platform to support the user in interpreting their data in greater detail.	Provide tools and visualisations to support researchers and other users in their interpretations	COVID-19 Phylogeny CoVEO

short read technologies and consensus sequences. Hamming distances were calculated between sequences in the alignment, excluding insertions and deletions, and a tree was inferred on the resulting distance matrix by a heuristic neighbour-joining algorithm implemented in CCPhylo [33]. Further details on the methods applied for the COVID-19 phylogeny can be found on the COVID-19 Data Portal.

PhyloDash [34] was used to present Evergreen output to users. PhyloDash is a reusable React component connecting interactive modules of Mapbox with OpenStreetMap [35], the phylogenetic tree with Phylocanvas.gl [36], a metadata table, and a timeline.

Variant browsing

The Kooplex collaboration platform [37] was used to develop CoVEO [38]. This application was developed and used for variant and sample browsing from systematically analysed raw reads. To obtain details on the number of new weekly COVID-19 cases, external data from the John Hopkins Coronavirus Resource Center [39] was required for an estimation of the percentage of new cases that have been sequenced in countries. Input samples were classified into variants of concern (VOC) or variants under investigation (VUI) groups based on key mutations in the database. Kooplex enables direct access to the database containing SARS-CoV-2 data and the system can be used for direct advanced data analysis and visualisation to answer scientific questions, for example identifying samples carrying mutations in SARS-CoV-2 PCR primer regions that may affect the testing efficiency [40].

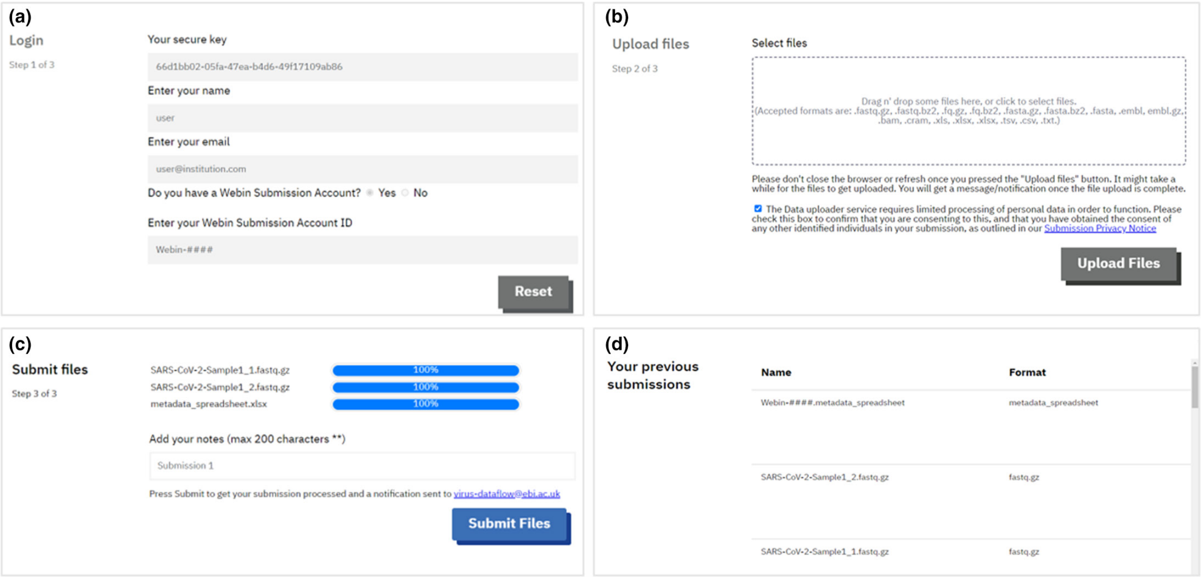
RESULTS

Submissions

Tools

Webin SARS-CoV-2 genome submission web API

Feedback from people submitting large-scale SARS-CoV-2 genomic data programmatically suggested existing routes for genome submission at the ENA, whereby metadata and sequence are separated, could be more intuitive and thereby reduce the effort needed to prepare and submit data (Table 1). The Webin SARS-CoV-2 Genome Submission Web Application Programming Interface (API) (launched in April 2021) is a dedicated tool for the programmatic submission of SARS-CoV-2 genome assemblies. During testing, the tool was seen to halve the amount of time taken per submission, compared to the existing method, Webin-CLI [41], both using a single thread. Assembled sequence data and metadata are provided within the



Login

Step 1 of 3

Your secure key: 66d1bb02-05fa-47ea-b4d6-49f17109ab86

Enter your name: user

Enter your email: user@institution.com

Do you have a Webin Submission Account? ● Yes ○ No

Enter your Webin Submission Account ID: Webin-####

Upload files

Step 2 of 3

Select files

Drag n' drop some files here, or click to select files.
(Accepted formats are: .fastq.gz, .fastq.bz2, .fq.gz, .fq.bz2, .fasta.gz, .fasta.bz2, .fasta, .embl, embl.gz, .bam, .cram, .xls, .xlsx, .xlsx, .tsv, .csv, .txt.)

Please don't close the browser or refresh once you pressed the "Upload files" button. It might take a while for the files to get uploaded. You will get a message/notification once the file upload is complete.

☑ The Data uploader service requires limited processing of personal data in order to function. Please check this box to confirm that you are consenting to this, and that you have obtained the consent of any other identified individuals in your submission, as outlined in our Submission Privacy Notice

Submit files

Step 3 of 3

SARS-CoV-2-Sample1_1.fastq.gz 100%
SARS-CoV-2-Sample1_2.fastq.gz 100%
metadata_spreadsheet.xlsx 100%

Add your notes (max 200 characters **): Submission 1

Press Submit to get your submission processed and a notification sent to virus-dataflow@ebi.ac.uk

Your previous submissions

Name	Format
Webin-####.metadata_spreadsheet	metadata_spreadsheet
SARS-CoV-2-Sample1_2.fastq.gz	fastq.gz
SARS-CoV-2-Sample1_1.fastq.gz	fastq.gz

Fig. 4. SARS-CoV-2 Drag and Drop Uploader Tool Interface. (a) Tool login page for users requiring a submission key (provided by the helpdesk team), brief contact details and their Webin Account ID. (b) Drag and drop area for users to deposit their data files and metadata spreadsheet, before uploading. (c) Upload progress area for users to monitor their file status. Submission occurs after successful upload for validation and processing at the ENA. (d) Submission history area where users can see their previous submissions through the tool.

same JSON payload, and are associated to a pre-registered ENA study and sample by specifying the respective accessions. The new tool includes a validation option to identify any errors prior to genome submission. Here, an API addresses a potential bottleneck, as it can be easily incorporated into scripts and tools written in different programming languages. It expands the existing assembly submission routes offered by the ENA, and is currently utilised by several national SARS-CoV-2 data brokers, and other high volume SARS-CoV-2 data submitters.

SARS-CoV-2 drag and drop uploader

Community feedback identified the need for a simpler interface and submission process to enable data sharing to keep up with the rapid expansion of the genomic sequencing effort (Table 1). The aim was to support sharing via another interactive route, without the need for extensive bioinformatics skills or prior knowledge of the ENA submission system. This resulted in the first new submission tool developed and launched in May 2020, the SARS-CoV-2 Drag and Drop Uploader, a browser-based ENA uploader [42].

This tool (Fig. 4) simplified the submission process at the ENA, by enabling users to drag and drop their sequence data files, accompanied by a metadata spreadsheet. Data is then stored on a cloud-based storage bucket prior to being submitted to the archive and data hub. For optimal performance and user-experience, individual uploads are limited to around 30 GB but there is no limit on the number of uploads a user can perform. The speed of upload will be dependent on the submitter's internet connectivity. To use the tool, submitters are first required to contact the ENA SARS-CoV-2 helpdesk via the form on the web page [42] and request the following: a secure login key unique to their ENA submission account, and an appropriate metadata spreadsheet for their data type.

Since its release in early 2020, there have been a total of 33 submissions with the Drag and Drop Uploader, with ~50% of users being new submitters to ENA. Of these data submissions 83% are raw reads (in FASTQ or BAM format) while the remaining 17% are assembled/consensus sequence submissions (mainly in FASTA format).

To support FAIR data principles, submissions to the ENA are structured via a metadata model. There was an even split between the number of end-to-end submissions made using the tool (from ENA top level project registration through to metadata and then data file submission), and part-submissions (e.g. only data file submission, whilst referencing pre-submitted project and metadata).

In a user survey with a total of nine respondents, six rated the tool 'excellent' and five said they would use the tool again. Consequently, five users have submitted data using the tool at least twice.

Additional submission tools

Gisaid Spreadsheet Conversion Tool. Data sharing via GISAID [43] has been a key aspect in the scientific community and in particular, the public health community. Therefore, to support users who have already submitted data to GISAID to also share it publicly with the INSDC, the GISAID spreadsheet converter was developed and released in November 2021. Utilising this tool, any spreadsheet that was used to submit sequence metadata to GISAID can be converted into the corresponding sample XML compatible with the ENA submission process. This was achieved by mapping metadata fields between the two repositories. This tool substantially reduces the burden on the submitter, can be accessed on GitHub [44].

In addition to this tool, the ENA enables submitters to link their sample submission to a corresponding GISAID record by supplying the GISAID Accession ID.

ENA bulk Webin-CLI tool. Webin-CLI, a Java based submission tool developed and maintained by the ENA, enables users to pre-validate their submission to identify any potential errors prior to submission. This tool often requires users to develop scripts to suit their submission, along with other preparatory steps for the submission process. This can create a bottleneck for bulk submissions. The ENA Bulk Webin-CLI tool [45], fully released in September 2021, is a wrapper tool on top of Webin-CLI which helps to simplify the submission process. The user submits a spreadsheet of data, which is used to streamline the submission process, including automated generation of manifest files and automated submission. Prior to usage, users are advised to familiarise themselves with the ENA submissions documentation. Although developed to support COVID-19 submitters and utilised in the backend for the SARS-CoV-2 drag and drop uploader tool, this tool was made available for all ENA submitters for any of the data types that can be submitted through Webin-CLI.

ENA analysis submitter. As part of the data hubs analysis component, raw read datasets were analysed, however could not be archived at the same time. This contributed to a lag in datasets being shared. To help reduce this backlog, automation of this step and running this step on-the-fly, was required (Table 1). The ENA Analysis Submitter [46] was developed and launched in March 2022. Although required as part of the data hubs, it was developed with external users in mind. Therefore, the tool takes several parameters appropriate for data submission and generates XMLs for submission via Webin REST API. The tool also uploads data files and carries out the submission process on behalf of the user. As mentioned, it is currently utilised within the ENA Pathogen Analysis System (PAS), described below, to automatically archive resulting analyses back to the data hubs.

Data ownership and attribution

Acknowledging ownership and provenance of publicly shared data was a key concern to collaborators (Table 1). Despite ownership already being stated within data and metadata records, we felt that more could be done to support researchers in open data sharing. To ease concerns and reduce the barrier to submissions, EMBL-EBI collaborates with ORCiD [47] to enable researchers/curators to add previously deposited datasets to their ORCiD profile, obtaining credit attribution for data records to which they have contributed. This tool allows you to retrospectively claim datasets to your ORCiD profile [48].

In addition to this, users can contact the helpdesk ena-path-collabs@ebi.ac.uk, where EMBL-EBI can issue Digital Object Identifiers (DOIs), traditionally used for publications [49], for their datasets. This is enabled via the BioStudies repository [50]; a BioStudy is registered and linked to the ENA project, then a DOI is assigned. Publications and/or datasets held in other repositories can also be linked to the BioStudy.

Claiming datasets through ORCiDs and DOI issuing are new ways for people to obtain credit for data submitted to open repositories and are anticipated to encourage greater open data sharing in the future. At the time of writing, the number of datasets shared at EMBL-EBI repositories that have been claimed through the ORCiD data claiming system is 2783.

Data mobilisation

Deposition of the initial Wuhan strain, collected in December 2019, within the INSDC - MN908947 [51], occurred in early January 2020. Subsequently, the health authorities of China and beyond, demanded further data sharing through GISAID, which resulted in rapid sharing of SARS-CoV-2 genomes. However, data shared via GISAID is not visible to all, and therefore not entirely open. To address this, an initial step in facilitating the public sharing of open data was the establishment of a specific support team for submitting centres to share COVID-19 data to the COVID-19 Data Portal. This support was largely provided in the form of detailed documentation, tracking of issues that data submitters faced, meetings with submitting institutes, groups and consortia.

Data sharing within the platform increased exponentially for over a year, as shown in Fig. 5. There was a large increase in the sharing of both genomes and raw reads around the summer of 2021, which continued into 2022, reflecting the Delta and Omicron variant waves [52]. By the end of summer 2022, the data portal had enabled the sharing of >6.3 million genomes and >5.5 million raw reads, and since then the rate of data sharing has slowed down. At the time of writing, there have been >6.6 million genomes and >6.2 million raw reads shared. The tools described here supported increased data sharing (Table 2), however, there were additional other reasons for the growth seen, such as increased genomic sequencing in certain

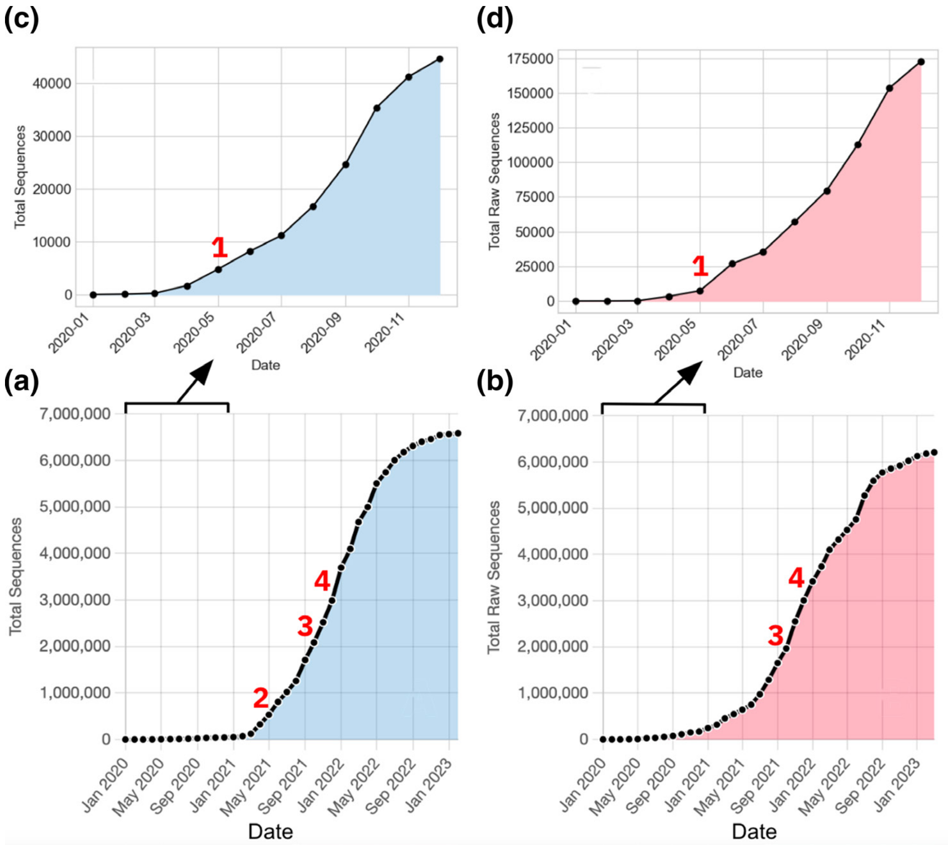


Fig. 5. Data mobilisation for (a) nucleotide sequences (genomes) and (b) raw reads from January 2020 up until February 2023. A zoom-in for 2020 has been shown for genome (c) and raw read (d) data mobilisation. Graphs have been annotated with release dates for submission tools. 1 – Drag and Drop Uploader; 2– Webin SARS-CoV-2 Genome API; 3 – ENA Bulk Webin-CLI tool; 4 – GISAID Spreadsheet Converter tool.

time periods for surveillance. Observed data sharing aligned relatively closely to sequencing efforts within countries sharing data. The median time between sample collection and dataset sharing was 24 days (Fig. 6).

To date, 120 and 103 countries have shared genomes and raw reads, respectively (Fig. 7). By 2020, genomes from 75 countries and raw reads from 54 countries had been shared, and by 2021, data had been shared by over half of the countries – 108 for genomes and 96 for raw reads. A summary of the number of datasets shared can be found on the COVID-19 Data Portal, under Data Statistics [53].

Systematic analysis – a public SARS-CoV-2 data hub

Broadly, the flow depicted in Fig. 8 was achieved, where public read data was processed via a reference-based mapping workflow. The pipelines call variants, generating an unfiltered and filtered VCF, and a fasta consensus sequence, per run processed. These analysis products are archived in the ENA and ingested by services and tools, including presentation services

Table 2. Number of SARS-CoV-2 raw read and genome submissions from January 2020 until March 2023, using the various submission tools. N/A reflects the fact that certain types of submissions cannot be carried out by specific tools. *Represents existing submission routes prior to COVID-19 pandemic

Submission tool	Raw read submissions	Genome submissions
Drag and Drop Uploader	5264	1658
Webin Submissions Portal (Interactive)*	110198	N/A
Webin-CLI*	182393	636850
Webin SARS-CoV-2 Genome Submission API	N/A	2858579
Webin-REST (Programmatic)*	2823226	N/A

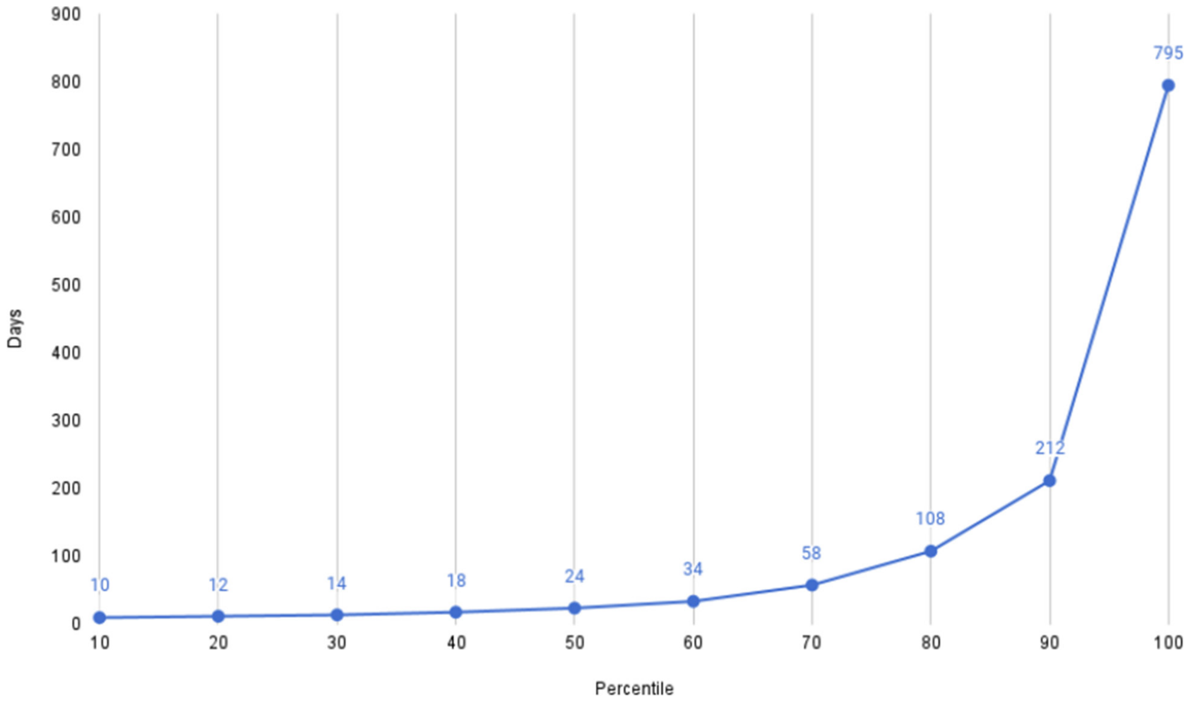


Fig. 6. Days from collection to first public release – A plot showing the percentile distribution of the length of time from collection of a sample as reported by the user, to the read data going public in ENA. A median time from collection to public of 24 days is observed.

such as the COVID-19 Data Portal, Pathogens Portal [54] and ENA browser. Additionally, data is ingested by visualisation tools, such as the COVID-19 Phylogeny and CoVEO variant explorer. To ensure appropriate tracking and staging of data processing and automated analysis, the ENA Pathogen Analysis System (PAS) ensures that pipelines are processing and archiving data and maintains a provenance trail from input to output, including technologies that utilise cloud compute. Overall, together this describes the public SARS-CoV-2 data hub, despite individual tools being available standalone and usable outside of the data hub.

As of writing this paper, 6239878 raw read datasets (sequenced using Illumina and Oxford Nanopore technologies) have been shared with the COVID-19 Data Portal, 3639538 raw read datasets (Fig. 9) have been successfully processed, equating to 58% of the entire raw dataset. The ENA Analysis Submitter tool supported the increase in processing after being incorporated into the ENA PAS to automatically submit data following analysis (Fig. 9).

Lineages

The COVID-19 Data Portal presents Pango lineage and (where necessary) WHO variant annotations for user-submitted [55] genomes and for systematically analysed (generated) [56] genomes (from raw reads). The integrated pangolin lineage assignment workflow runs three times per week on EMBL-EBI's high performance computing (HPC) cluster infrastructure. This running frequency was required due to the sheer volume of data submissions on a weekly basis. The resulting lineage assignments are fed back and presented within the COVID-19 Data Portal, which enables for querying and searchability. The workflow currently runs Pangolin in an incremental fashion, only running on sequences submitted in the previous 90 days, and keeping the previous assignments for all older sequences. To ensure scalability for future preparedness, the workflow chunks the sequences to be analysed into smaller multifasta datasets, so if there are large numbers of sequences to be analysed, e.g. greater than 500000 at once, memory overload can be avoided.

Figs 10 and 11 show that on the whole, submitted genome data correlated to the raw data being systematically analysed, depicted by similar percentages of WHO and Pango lineage annotations. Along with lineage annotations, the COVID-19 Data Portal also presents a table of submitted genomes that are representative of individual Pango lineages, the 'representative sequences' [27] (Supplementary Material).

The representative lineages data set that was created for the data hubs and portal is an approach that identifies representative sequences from real submitted sequences, as opposed to the alternative, which would be to construct representative consensus sequences, based on characteristic mutations of that lineage. The advantage of using 'real' submitted sequences is the attached

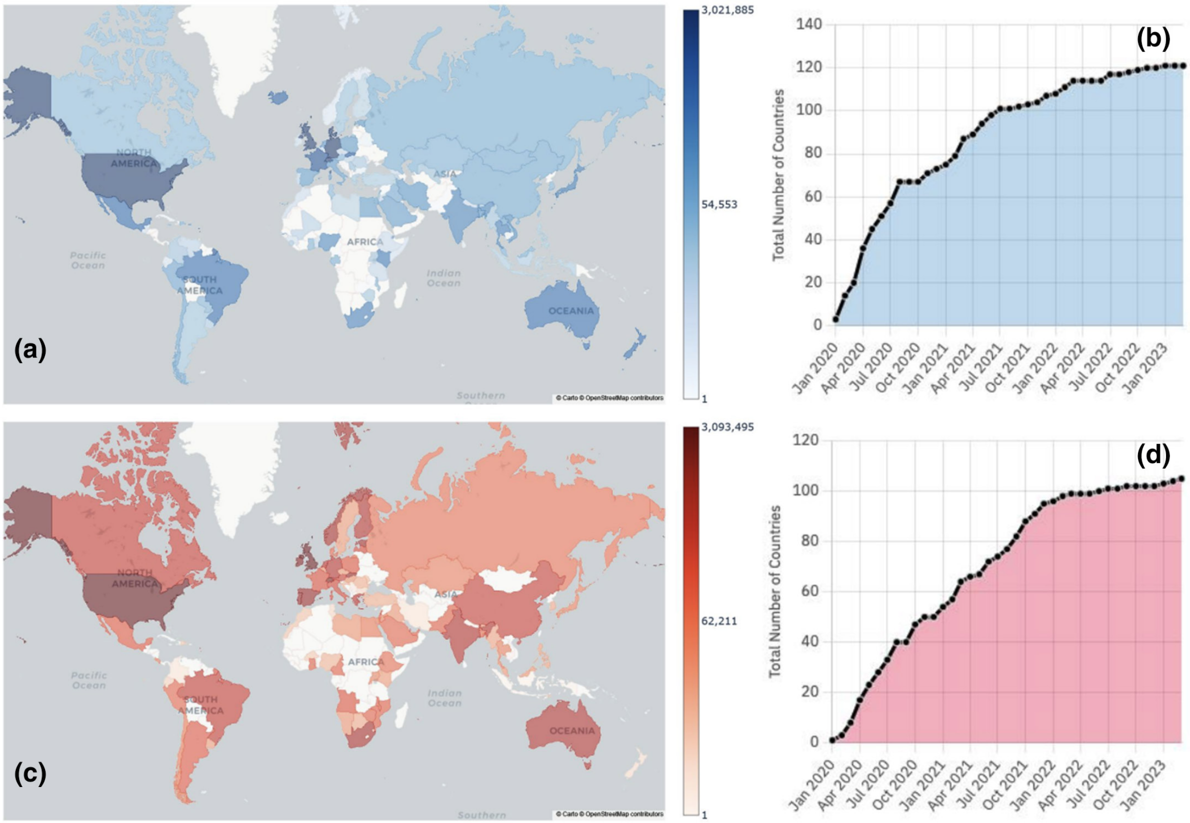


Fig. 7. Distributions of data sharing for SARS-CoV-2 genomes (a and b) and raw sequencing data (c and d) from January 2020, until February 2023. The maps present the geographic spread of data sharing, the deeper and darker the colour, the higher the amount of genomes (a) or raw reads (c) shared. The plots represent the number of countries worldwide that have shared genomes (b) and raw reads (d).

metadata, which means these sequences can be used in phylogenies with other sequences, giving context to sequences in a phylogeny without constructing a very large tree of millions of sequences.

ENA pathogen analysis system

The scale of the data to be analysed required a processing management system, to ensure data was not analysed more than once, whilst maintaining the integration of workflows. Therefore, the ENA PAS was developed to support management and scheduling of data processing (Fig. 12). It includes several tools and technologies, described below, that have been combined to retrieve information on datasets to analyse, provide infrastructure for analysis through integrated workflows (compute and storage), schedule dataset processing, run analysis, automatically submit resulting analyses, log processing and track overall processing.

Nextflow is a workflow management tool used here for scheduling and processing datasets in a parallel and compartmentalised manner [57]. It enables fine-grained control of compute requirements and provides logging capabilities. The workflows are containerised using Docker [58] for Google Cloud Compute and Singularity [59] for High-Performance Clusters. Input datasets are streamed just-in-time for processing and results are submitted on-the-fly using the ENA Analysis Submitter tool, described above. A Google BigQuery [60] database is used to track the number of analysed datasets and submitted results.

Archival and data availability

Analysis archival

The COVID-19 Sequence Analysis Workflow for Illumina and Nanopore data generates three products for each run processed: (1) an archive containing an unfiltered VCF file, a bam file, and a coverage file; (2) a filtered VCF file; and (3) a consensus sequence. These outputs are stored in individual projects under the umbrella project PRJEB45555 to make it easy for users and downstream services to find and retrieve them, while also preserving metadata and raw data links through sample and read references.

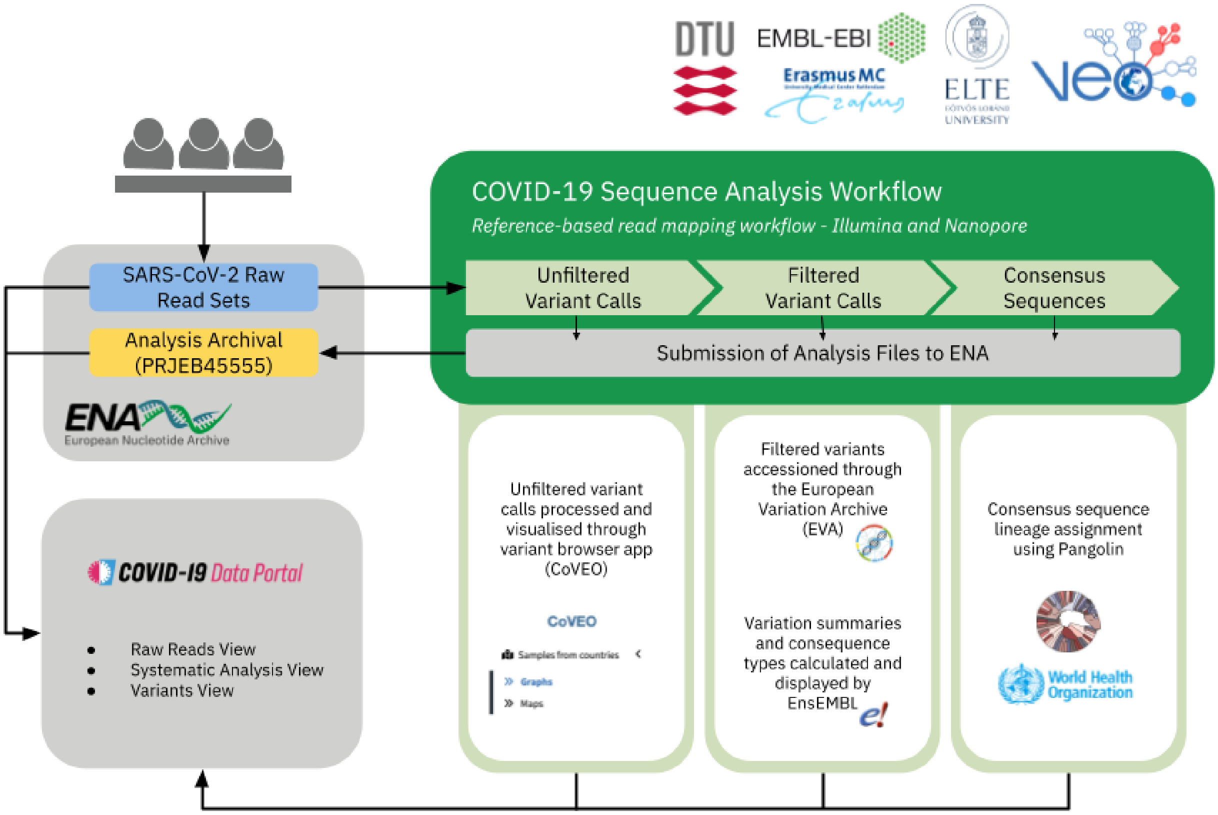

DTU
EMBL-EBI
Erasmus MC
ELTE
EÖTVÖS LORÁND UNIVERSITY
veo
SARS-CoV-2 Raw Read Sets
Analysis Archival (PRJEB45555)
ENA
European Nucleotide Archive
COVID-19 Sequence Analysis Workflow
Reference-based read mapping workflow - Illumina and Nanopore
Unfiltered Variant Calls
Filtered Variant Calls
Consensus Sequences
Submission of Analysis Files to ENA
COVID-19 Data Portal
Raw Reads View
Systematic Analysis View
Variants View
Unfiltered variant calls processed and visualised through variant browser app (CoVEO)
CoVEO
Samples from countries
Graphs
Maps
Filtered variants accessioned through the European Variation Archive (EVA)
Variation summaries and consequence types calculated and displayed by EnsEMBL
e!
Consensus sequence lineage assignment using Pangolin
World Health Organization

Fig. 8. Schematic representing an overview of the systematic analysis of public raw read data shared with INSDC and the COVID-19 Data Portal. Analysis outputs are ingested by various tools and services. Unfiltered variant calls, ingested directly by CoVEO, filtered variant calls by EVA and Ensembl and consensus sequences ingested by Pangolin.

To facilitate this archiving, two new analysis types were created at the ENA: COVID19_CONSENSUS, which is used for the computed consensus sequences (output 3); COVID19_FILTERED_VCF, which corresponds to the computed filtered VCFs (output 2); and re-use of PATHOGEN_ANALYSIS, which is used for the archive containing the unfiltered VCF file, bam file, and coverage file (output 1).

For additional discoverability, a view of these analysis products can be found on the COVID-19 Data Portal. This provides the user with more advanced search and download capabilities.

Variant archival and availability

The systematically generated filtered VCFs feed into a processing flow developed by the European Variation Archive (EVA) [61], which ingests and accessions the individual variants providing a globally unique identifier (rsIDs) for each one. These accessioned variants have been indexed and displayed on the COVID-19 Data Portal, providing a concise view of all variants, with advanced filtering options. Users can filter on variant types, or use the genome browser to filter on a specific gene or region of interest. The variants are also available to retrieve from the EVA portal via studies PRJEB45554, PRJEB55355 and PRJEB57993. Further downstream, variant calls are fed to Ensembl [62] and its Variant Effect Predictor (VEP) [63], to annotate aspects such as consequence types, presenting within the Ensembl COVID-19 browser [64].

Visualisation

COVID-19 phylogeny

Using Evergreen, a phylogenetic analysis workflow [30] was developed by the VEO consortium and integrated within the data hubs system at EMBL-EBI. PhyloDash enables users to view and search all fields of the metadata table, with the sample metadata optionally supplemented with genotyping data or other phenotypic information, enabling quick selection of samples of interest (Fig. 13). Immediate phylogenetic context of a sample can also be viewed by clicking the ‘Show sample context’ button, which provides a subtree containing the given sample.

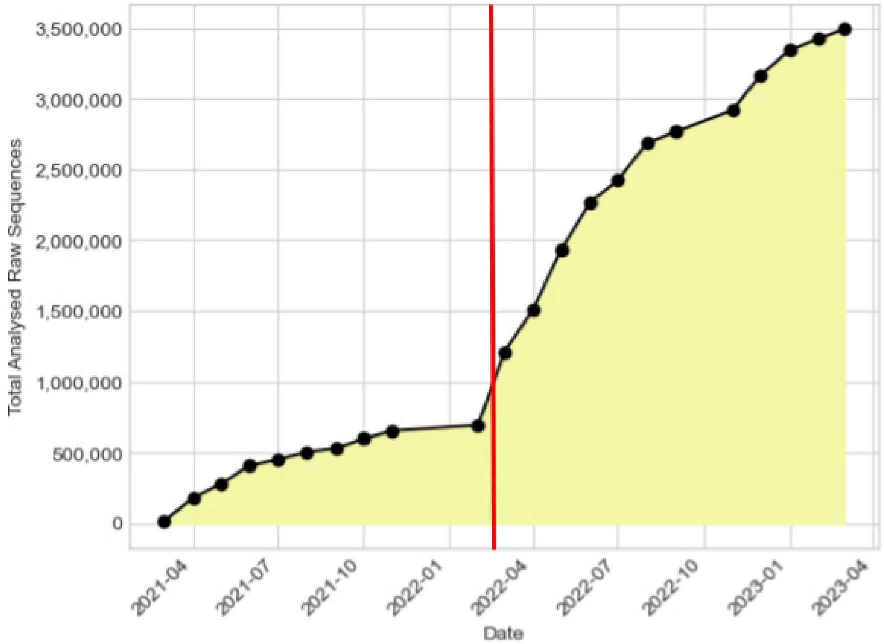


Fig. 9. Number of analyses archived following systematic analysis, since March 2021 to March 2023. Generated by tracking the contents of analysis projects detailed in Data Availability. The red vertical line represents the full release of the ENA Analysis Submitter and integration into the ENA PAS.

CoVEO

CoVEO [38] was developed by the VEO consortium and integrated within the data hubs system at EMBL-EBI. This includes an underlying PostgreSQL database which was updated with an unfiltered VCF and coverage file per sample processed in the systematically analysed dataset, aligning with the regular monthly VEO bulletin reports.

The CoVEO app presents an interactive visualisation on graphs and maps about the number of raw SARS-CoV-2 sequences submitted by countries across the world and the distribution of variants across time (Fig. 14). Users can input an inclusion/exclusion list of mutations of the S protein, with the app showing the distribution of samples containing the custom-selected mutations across time grouped by country. By enabling focus on samples with sufficient sequencing depth on predefined positions, a unique approach is provided to users. The underlying codebase has been shared on GitHub [65].

DISCUSSION

Open data sharing

Scale

The scale of open data sharing over the course of the SARS-CoV-2 pandemic is unprecedented [66]. To consider the numbers of raw read datasets in terms of records (~25% of all read datasets in ENA), does not even fully do justice to the scale of archived SARS-CoV–2 sequencing data. The publicly archived raw read data in the ENA for SARS-CoV-2 is 2 PetaBases worth of sequencing information. Contextually, that is enough sequencing data to sequence over 22000 human genomes from telomere to telomere at a depth of 30X. That is just the data that has been publicly shared and there are far more consensus sequences available in the biodata sphere, so it stands to reason that there must be even more raw read data not archived publicly. By comparison, the bacterial pathogen with the most raw read data shared to the ENA is *Salmonella enterica* at 313388 data sets. Other major viral pathogens have low volumes of data sharing, such as Ebola virus (3099 read data sets) or Zika virus (2858 read data sets). In fact, if you were to look for all publicly shared raw read data for the WHO priority pathogens [67] not including MERS, SARS or their descendants, you'd find 333087 read data sets. SARS-CoV-2 has 18 times more public read data sets than all of the other WHO priority pathogens combined. The volume of data being generated and submitted publicly to the ENA presented many challenges to overcome in the development of the public SARS-CoV-2 Data Hub. New challenges brought about new tools for submission, search and retrieval, scalable analysis workflows and visualisation tools.

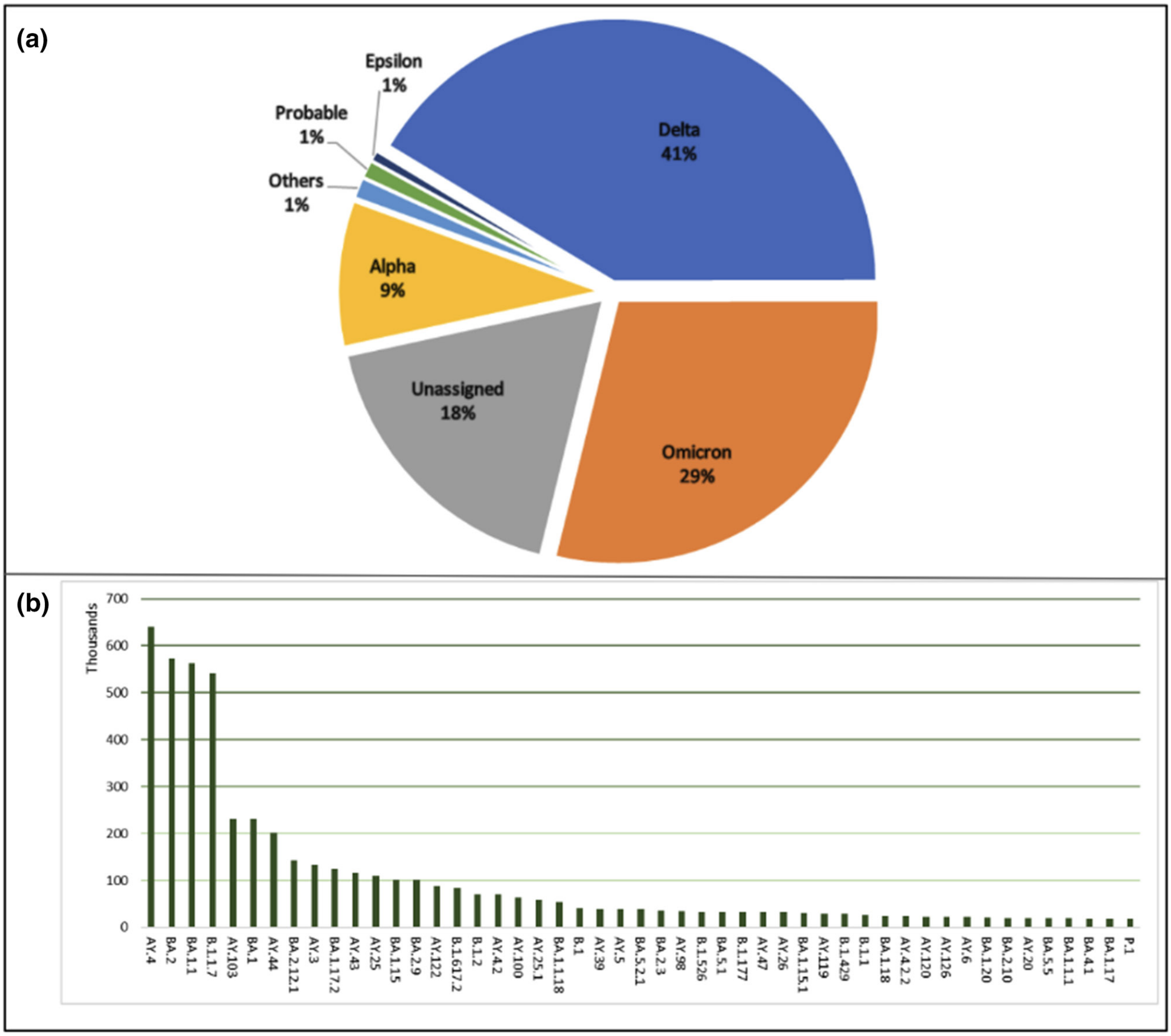


Fig. 10. Lineage distributions for SARS-CoV-2 submitted sequences, as of March 2023, a) WHO lineages, Others: consist of a collection of lineages with distribution under 1% which include A.23.1-like, A.23.1-like+E484K, AV.1-like, B.1.1.318-like, B.1.1.7-like+E484K, B.1.617.1-like, B.1.617.3-like, Beta, Eta, Gamma, Lambda, Mu, Theta, Zeta, Iota. b) Top 50 Pangolin lineages observed amongst the dataset. (N=6,051,081).

Data sharing

The unprecedented open data sharing mentioned above, does not include the wealth of data that has been shared to GISAID. Openly shared SARS-CoV-2 genomes number around 6.5 million to the INSDC, whereas GISAID currently contains approximately 15 million SARS-CoV-2 genomes [43]. There is likely overlap between the datasets shared, but the extent of this is not known fully. This also highlights that despite the immense and remarkable efforts undertaken and described here, more can be done to truly support open data sharing. Within INSDC, users can share their data, keeping it private, until public release at the time of publication, which appeals to more research focused sequencing efforts. GISAID offered the benefit of being incredibly rapid from submission to availability of genome sequences. Though as we move beyond a rapid data sharing ecosystem, the need for open sharing of both sequences and raw read data becomes the new priority. Through our GISAID conversion tool, we have made it easier for users who have already shared or have experience sharing data through GISAID to submit their data to open repositories like the ENA, thus facilitating more users to bring both their sequences and underlying read data into the public sphere. Furthermore, we have made it possible for users to get credit for the data they share openly through ORCiD data claiming and DOI issuing, helping scientists around the world get recognised for their contribution to open biodata. Each development is a small step forward to removing barriers to open data sharing, both practical and social.

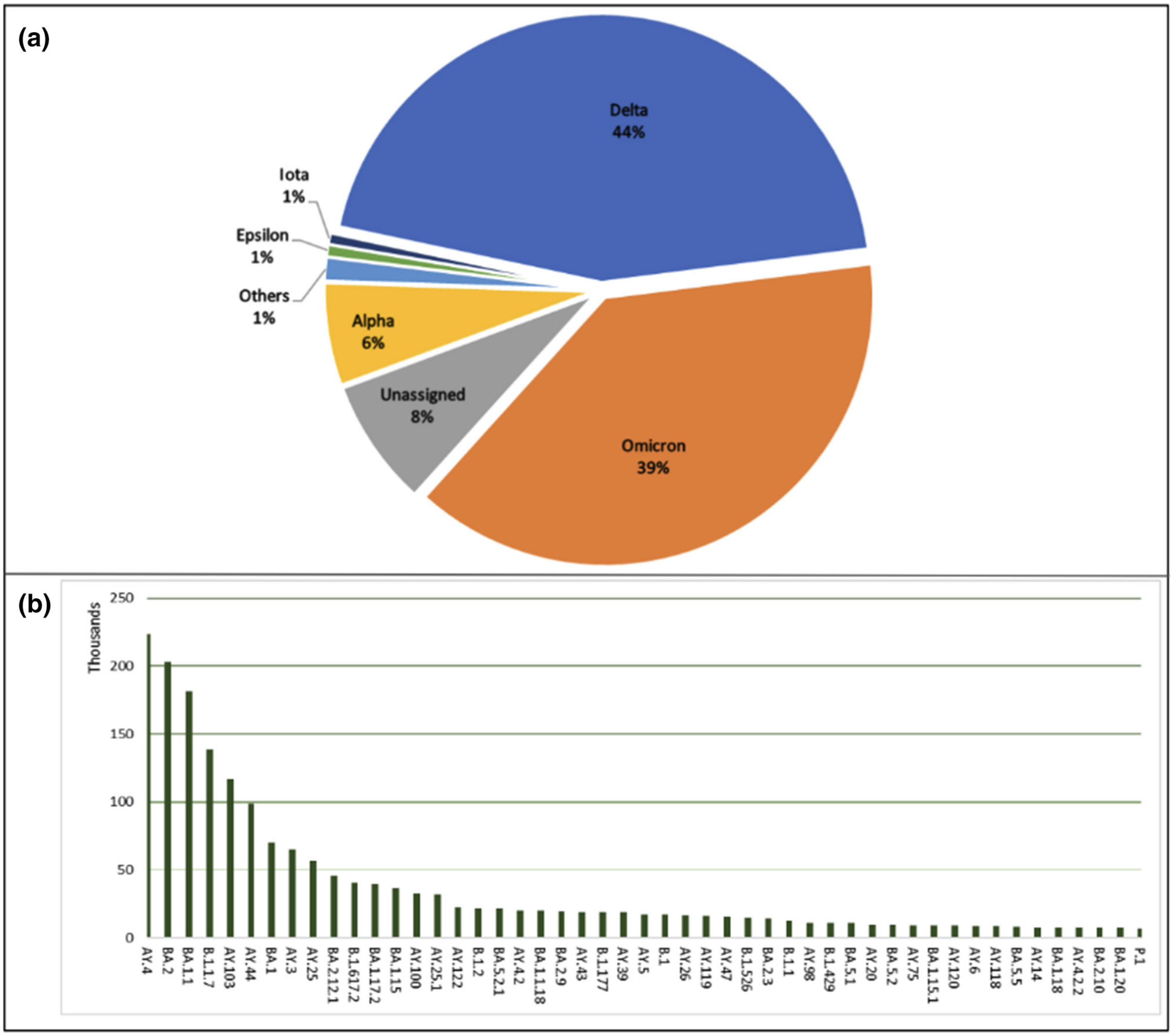


Fig. 11. Lineage distributions for consensus genomes generated from SARS-CoV-2 systematic analysis, as of March 2023, a) WHO lineages, Others: consist of a collection of lineages with distribution under 1% which include A.23.1-like, A.23.1-like+E484K, AV.1-like, B.1.1.318-like, B.1.1.7-like+E484K, B.1.617.1-like, B.1.617.3-like, Beta, Eta, Gamma, Lambda, Mu, Theta, Zeta and Probable Omicron. b) Top 50 Pangolin lineages observed amongst the dataset. (N=2,239,727).

The Data Hub has driven the development of a whole new set of submission routes for users looking to share data openly in a facile way, such as the Webin SARS-CoV-2 Genome API or the Drag and Drop uploader. Both of these tools are innovations covering two different demographics of users interested in sharing data. The former being targeted more at high volume, programmatically experienced users and the latter targeted at users with little to no knowledge of ENA submission systems.

Researchers now have a vast sea of data available to them, one that is made navigable by the search and retrieval tools powering the COVID-19 Data Portal and ENA. Thanks to the high quality metadata being collected by the majority of SARS-CoV-2 submitters, it is easier than ever for users to find exactly the right data set required to answer their question, either in the immediate present or long into the future. Users can rapidly filter through data based on numerous metadata fields such as country, collection date or in the case of read data, by sequencing platforms or sequencing approaches, e.g. WGS or amplicon.

Data dissemination

Dissemination of variants to EVA and Ensembl provides just a couple of examples of the interconnected web of life science resources at EMBL-EBI to which data from the ENA feeds into. As mentioned above, this provides greater context and annotations to sequencing-related datasets, and highlights the importance of the ENA (and other data repositories at EMBL-EBI) to the life sciences community. However, further than EMBL-EBI, datasets (both raw and systematically analysed) feed into downstream

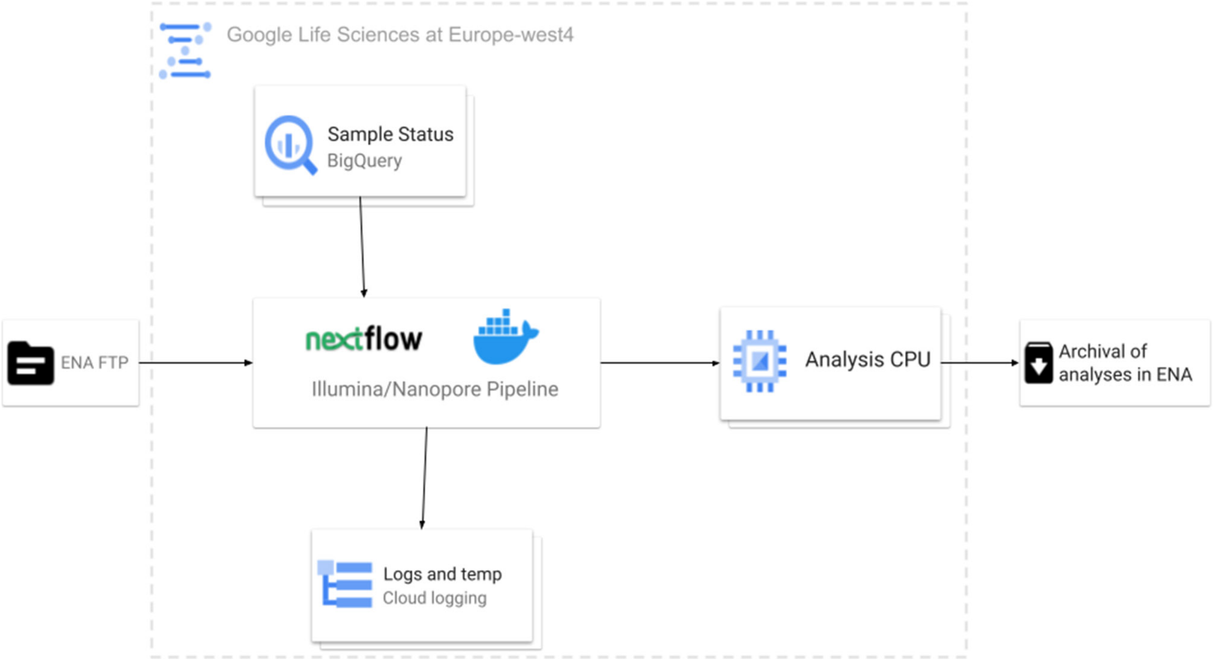

Google Life Sciences at Europe-west4
Sample Status
BigQuery
ENA FTP
nextflow
Illumina/Nanopore Pipeline
Analysis CPU
Archival of analyses in ENA
Logs and temp
Cloud logging

Fig. 12. Diagram presenting the flow of data within the ENA Pathogen Analysis System, including Google Cloud services – BigQuery and Cloud Storage. Datasets are pulled using File Transfer Protocol (FTP) and processed using Nextflow pipelines and Docker images, with resulting analyses submitted back to the ENA and the Data Hub.

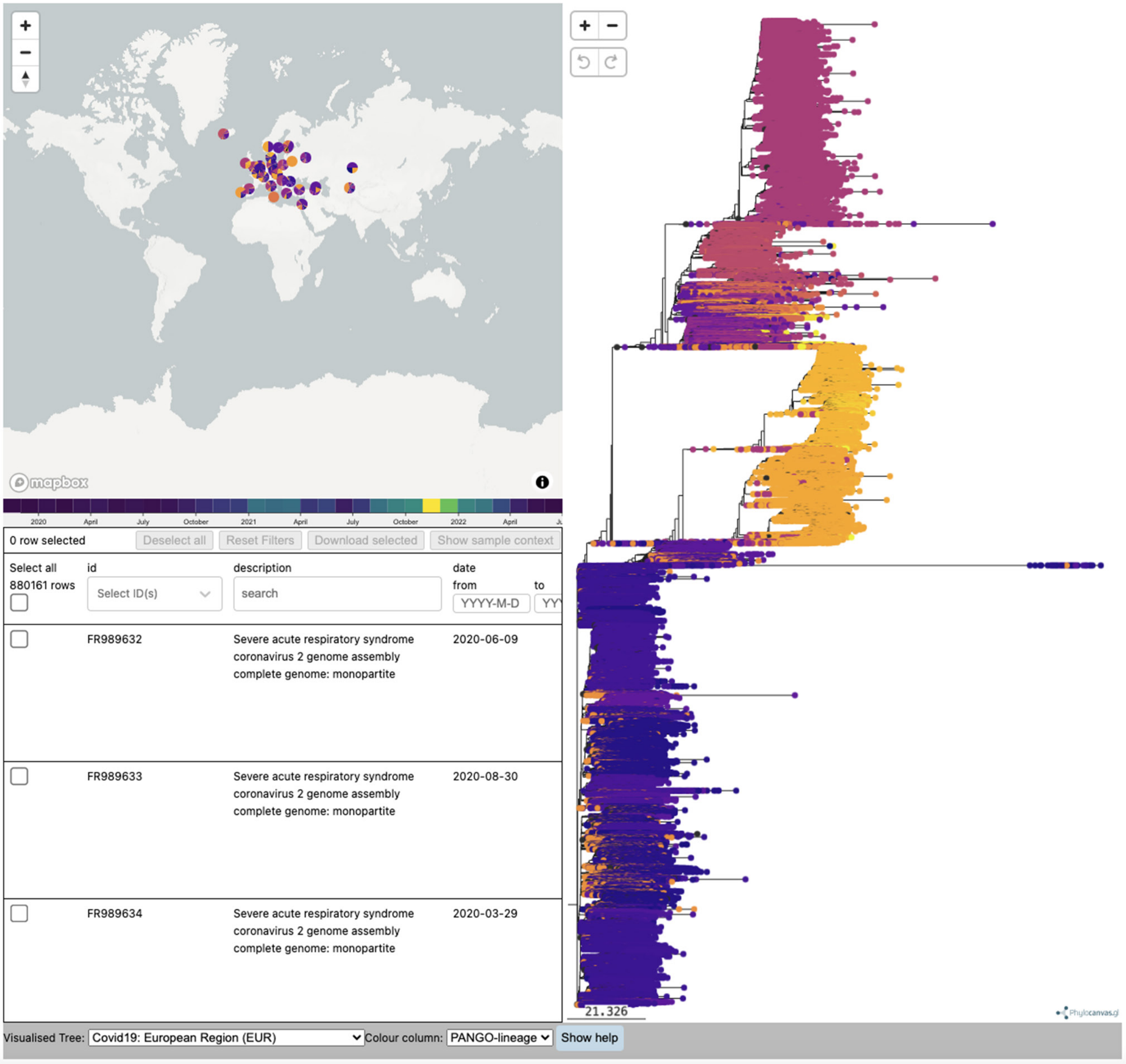
tools and services, including Galaxy workflows [68], CRG Viral Beacon [69] and CovSPECTRUM [70] to name a few. This contributes to a network of publicly accessible tools and visualisations that support the scientific community further in analysing and interpreting datasets.

It is worth highlighting that, as the sequencing data presenting within the COVID-19 Data Portal is sourced from the ENA, SARS-CoV-2 sequencing datasets become part of a larger database, that has existed for more than 40 years, with datasets spanning a plethora of different species and taxa. This provides a solid foundation supporting scientific research around targeted projects and resource development, e.g. European COVID-19 Data Platform.

Data analysis and processing

The volume of data processed through the analysis workflows is immense, with a growing stream of over six million raw read datasets publicly available in the ENA, and therefore the COVID-19 Data Portal. The public data hub has brought about two significant developments to facilitate a scalable analysis system. Firstly, a dedicated workflow to handle processing of Illumina and Oxford Nanopore datasets, which aim to provide a uniform approach to analysis, and written in Nextflow for cross compatibility. Having the resultant data products archived publicly in the ENA means that there is a systematically analysed, publicly accessible, set of consensus sequences and variants available to researchers around the globe to learn from the SARS-CoV-2 pandemic long into the future. Secondly, we demonstrated a successful deployment across local high-performance computing and distributed cloud platforms to maximise scalability and flexibility in processing. The utility of writing strong Nextflow based pipelines combined with distributed deployment means an elastic analysis management system has been developed, which, when combined with robust pipelines can be rapidly re-deployed to respond to future pandemic scenarios. Further, the utility would be available for more targeted SARS-CoV-2 Data Hubs (public or private), whereby specific groups and targeted efforts make use of tools, such as those described here.

The methods described in this paper covers analysis of a large and continuously growing dataset, however as discussed under the limitations below, this may come at a cost to accuracy of analysis. Other groups within the community have systematically analysed raw SARS-CoV-2 reads, including Galaxy Europe. Their approach [68] offers an equally valuable dataset, by assessing the incoming dataset, more-tailored analyses can be carried out. This results in fewer overall datasets analysed, but greater accuracy in the analysis. Finding an appropriate balance here is a major challenge, however through open sharing of both types of dataset, they can complement one another, as appropriate. What is common to both however, is they are powered by the open sharing of raw read data, and highlights the value of open FAIR data, as neither objective could be achieved without it.



0 row selected

Select all 880161 rows □	id Select ID(s)	description search	date from YYYY-M-D to YY
□	FR989632	Severe acute respiratory syndrome coronavirus 2 genome assembly complete genome: monopartite	2020-06-09
□	FR989633	Severe acute respiratory syndrome coronavirus 2 genome assembly complete genome: monopartite	2020-08-30
□	FR989634	Severe acute respiratory syndrome coronavirus 2 genome assembly complete genome: monopartite	2020-03-29

Visualised Tree: Covid19: European Region (EUR) Colour column: PANGO-lineage

Fig. 13. COVID-19 Phylogeny, as of February 2023, presenting in the COVID-19 Data Portal. This includes three main panels – a map on the top left, metadata table bottom left, and phylogenetic tree on the right.

The lineage workflow mentioned above, currently runs Pangolin for all sequences each time, but a move towards a more incremental approach is likely, only calling lineages for new sequences. However, a caveat of the Pangolin classification system is that it remains relatively dynamic, with new decision trees, lineages and sublineages released regularly. Therefore, when a new version of Pangolin is released, the workflow needs to run on the full sequence set again to ensure the latest decision tree is used. Pangolin itself requires quite involved development and curation, and will eventually lose support, meaning running this indefinitely is not sustainable.

Limitations

Big data processing

The unprecedented scale of data has posed a challenge when it comes to the amount, and in particular, cost of storage and analysis, for example when using commercial-cloud solutions. Increasingly, cloud providers are venturing into the 'Genomics' and 'Life Sciences' spaces, which provides an opportunity to better-handle big data [71], however navigating product costs is key, as these appear for all aspects – analysis itself, data storage and data ingress and egress. Hybrid compute solutions (as described in this paper), coupled with cost monitoring, offer a partial solution to this, especially when considering analysis of big datasets.

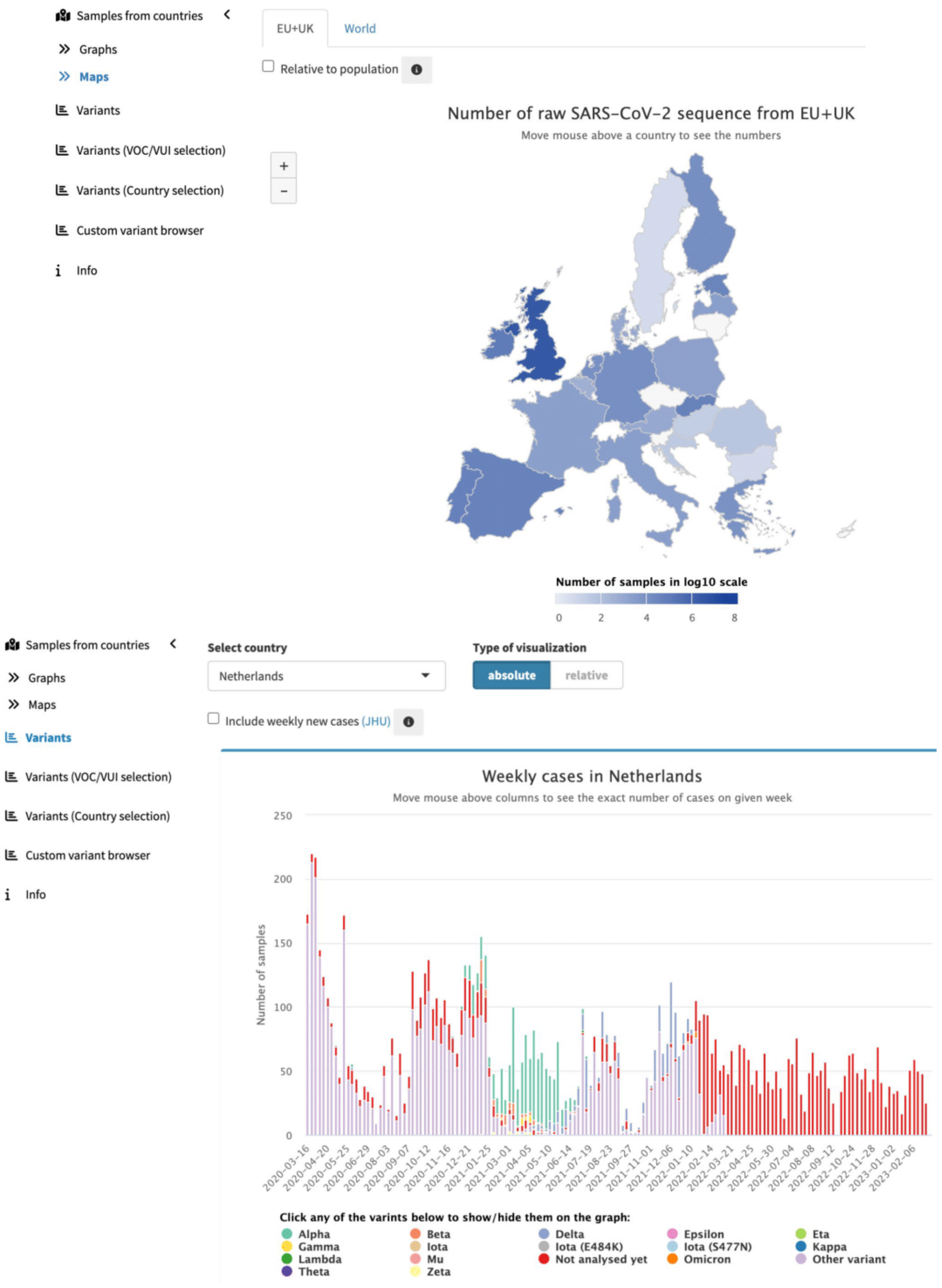


Fig. 14. CoVEO variant browser – presenting visualisations, as of February 2023, around the genomic data that is systematically analysed, in the context of VOCs and VOIs.

However, an alternative approach to alleviate processing issues and delays, may have been to consider working with more targeted sequencing efforts, combined with the usage of data hubs, as opposed to a single large public data hub, as described here.

Generalising data

By analysing in a systematic way, there are assumptions made on all of the raw read data ingested, for example when parameters are considered for pipelines. It is possible that in some cases, a generic parameter or threshold filters out major variants or fails to identify specific mutations amongst sequences – these are just a couple of examples. However, when analysing big data, it should also be taken into consideration that it is largely not possible to implement individual parameters and specifications in an efficient manner across a dataset. Furthermore, as raw data is openly available, this invites others in the scientific community to (re-)analyse datasets within the larger set to apply specific parameters or utilise other tools.

Phylogenetic trees at scale

The COVID-19 phylogeny aimed to present all appropriate data, filtering out some data according to criteria mentioned in the Methods. However, this became a challenge, as the number of shared datasets rose beyond hundreds of thousands to millions. Overall, this was due to the challenge in presenting a usable phylogeny report, with such a large number of points. As an example, the European tree (which is the largest regional tree), displays >880000 samples, which present big data challenges in efficiently processing, but also loading and navigating the tree itself.

Generally, this issue existed amongst the community, as how to best present so many samples on a functional phylogenetic tree report. Looking at other openly available tools such as Nextstrain [72], reports are restricted to around 4000 genomes, with subsampling to appropriately represent the global landscape of publicly shared SARS-CoV-2 genomes, balancing this with a smoother user experience. Looking to visualise all the public samples, Taxonium [73] can display phylogenetic trees with millions of leaves, but the user experience of navigating leaves much to be desired. From these solutions, and reflecting on our experience, another alternative model that our task force worked towards, and intend to use for future cases, included enabling the user to input a set of samples on top of a baseline tree of 'representative sequences'. This provides flexibility and a balance in enabling the user to view samples they are interested in, helping to target specific groups of samples (and therefore likely reduce the sample size), whilst maintaining an overall background view of samples across different lineages.

Looking forward

As many components of the public data hub started to mature and become available for more-targeted SARS-CoV-2 Data Hubs, there was an increase in training and outreach carried out. Here, it became evident that hands-on training in using tools and components of the public data hub is important and of great value to users, data providers and data consumers. Embedding this amongst regionally-relevant examples of the importance of open data sharing for life sciences and in particular, infectious diseases, provides wider context to the (further) development of platforms and components described here. Coupled with the data growth mentioned, this enabled us to focus on the user experience, for example what incentives are available for those who share data openly. Outreach also provides a forum to encourage direct feedback from users, which are taken on board and addressed, overall improving systems and components. We intend to continue training and outreach, advertised via platforms such as the News section of the COVID-19 Data Portal, X (formerly Twitter), EMBL-EBI training pages, amongst others.

Sustainability is a key aspect in setting up platforms and databases. One of the main take-away points from the efforts undertaken and detailed in this paper, was that the true value of the developments made, are in their re-use. The pandemic has provided an opportunity to utilise sustainable practices in developing tools, components and systems which can be re-used in the future. Since the COVID-19 outbreak, tools and components described in this paper have been repurposed for the Monkeypox (MPox) outbreak in May 2022 (a public MPox Data Hub). This provided a better indication of where efforts would have to be focused in repurposing tools for other outbreak scenarios. For example, spinning up an Outbreaks web page within a portal, providing submission instructions and repurposing of the drag and drop uploader were some aspects that were relatively quick. Other aspects, such as setting up systematic analysis of MPox runs took longer, which was expected, as pipeline testing was required to ensure that output was of sufficient quality, especially in identifying known mutations and variants. Therefore, going forward we are confident that tools and components can be repurposed for future potential outbreaks or generally for pathogen pandemic preparedness. The Pathogens Platform [54] provides an interface to facilitate and support these efforts.

CONCLUSION

The public SARS-CoV-2 Data Hub represents a fundamental change in how we approach open biodata in response to pandemics. Not only have more tools been developed to enable users to share data quickly, both privately and publicly, but added value can now be offered in exchange for sharing of data in the form of rapid systematic analysis and visualisations of data. This added value can have a significant impact on data submitters who may not have strong bioinformatic backgrounds and feel unable to do much with their data. They can now access a benefit in kind for their data sharing efforts, which will hopefully encourage more users to participate in the open biodata community. Having an adaptable pandemic response system is essential to our collective pandemic security, and the SARS-CoV-2 Data Hubs offer a strong platform on which to build.

Funding information

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874 735 (VEO), alongside others which have supported the European COVID-19 Data Platform: grant agreement No. 871 075 (ELIXIR-CONVERGE), grant agreement No. 824 087 (EOSC-Life) and grant agreement No. 825 746 (ReCoDID). Furthermore, this work has also received funding from the European Union Horizon project grant agreement No. 101 046 203 (BY-COVID). This work has been supported by the Member States of the European Molecular Biology Laboratory (EMBL).

Acknowledgements

The work described in this paper would not have been possible without the many data providers and data brokers who have shared their data openly. Each data provider has been mentioned within sample and run data records that our analyses link to directly (see Archival and Data Availability). This has been a highly collaborative effort, driven by the dedication of those within the VEO COVID-19 taskforce and the authors of this paper - thank you all.

Conflicts of interest

The author(s) declare that there are no conflicts of interest.

References

1. Lu R, Zhao X, Li J, Niu P, Yang B, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;395:565–574.
2. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. *Acta Biomed* 2020;91:157–160.
3. Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, *et al*. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat Med* 2021;27:1518–1524.
4. Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, *et al*. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol* 2021;22:196.
5. Harrison PW, Lopez R, Rahman N, Allen SG, Aslam R, *et al*. The COVID-19 data portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res* 2021;49:W619–W623.
6. Freeberg MA, Fromont LA, D'Altri T, Romero AF, Ciges JI, *et al*. The European genome-phenome archive in 2021. *Nucleic Acids Res* 2022;50:D980–D987.
7. Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O, *et al*. The COMPARE data hubs. *Database* 2019;2019:baz136.
8. Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, *et al*. The European Nucleotide Archive in 2022. *Nucleic Acids Res* 2023;51:D121–D125.
9. Cantelli G, Bateman A, Brooksbank C, Petrov AI, Malik-Sheriff RS, *et al*. The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Res* 2022;50:D11–D19.
10. International Nucleotide Sequence Database Collaboration [Internet]. (n.d.). https://www.insdc.org/ [accessed 18 April 2023].
11. About VEO - VEO Europe [Internet]; (n.d.). https://www.veo-europe.eu/about-veo [accessed 17 February 2023].
12. Lam C, Gray K, Gall M, Sadsad R, Arnott A, *et al*. SARS-CoV-2 genome sequencing methods differ in their abilities to detect variants from low-viral-load samples. *J Clin Microbiol* 2021;59:e0104621.
13. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, *et al*. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021;19:409–424.
14. covid-sequence-analysis-workflow [Internet]. European Nucleotide Archive; 2022. https://github.com/enasequence/covid-sequence-analysis-workflow [accessed 18 April 2023].
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
16. Li H. Aligning sequence reads, clone sequences and assembly Contigs with BWA-MEM. 2013.
17. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10:giab008.
18. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, *et al*. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–11201.
19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012;6:80–92.
20. vcf_to_consensus.py [Internet]. European Nucleotide Archive; 2022. https://github.com/enasequence/covid-sequence-analysis-workflow/blob/663fd128dc2af0c47e25a1c98adff9ca96bd4daf/illumina/bin/vcf_to_consensus.py [accessed 17 February 2023].
21. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;17:10.
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
23. covid-sequence-analysis-workflow/vcf2consensus.py at master · enasequence/covid-sequence-analysis-workflow [Internet]. GitHub; (n.d.). https://github.com/enasequence/covid-sequence-analysis-workflow [accessed 17 February 2023].
24. dca-analysis-tools/ena-pangolin-lineage at main · enasequence/dca-analysis-tools [Internet]. GitHub; (n.d.). https://github.com/enasequence/dca-analysis-tools [accessed 17 February 2023].
25. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, *et al*. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* 2021;7:veab064.
26. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH. Sustainable data analysis with Snakemake [Internet]. F1000Research; 2021. https://github.com/enasequence/dca-analysis-tools [accessed 17 February 2023].
27. COVID-19 Data Portal - Representative Sequences [Internet]; (n.d.). https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=representative-sequences&size=15 [accessed 20 February 2023].
28. scorpio [Internet]. CoV-lineages; 2022. https://github.com/cov-lineages/scorpio [accessed 18 April 2023].
29. ena-content-dataflow/get_repr_seqs.py at master · enasequence/ena-content-dataflow [Internet]; (n.d.). https://github.com/enasequence/ena-content-dataflow/blob/master/scripts/get_repr_seqs.py [accessed 17 February 2023].
30. Szarvas J, Ahrenfeldt J, Cisneros JLB, Thomsen MCF, Aarestrup FM, *et al*. Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform. *Commun Biol* 2020;3:137.
31. genomicepidemiology / ebi_viral_phylogeny — Bitbucket [Internet]; (n.d.). https://bitbucket.org/genomicepidemiology/ebi_viral_phylogeny/src/master/ [accessed 18 April 2023].
32. Clausen P, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* 2018;19:307.
33. Clausen PTLC. Scaling neighbor joining to one million taxa with dynamic and heuristic neighbor joining. *Bioinformatics* 2023;39:btac774.
34. genomicepidemiology / phylodash — Bitbucket [Internet]; (n.d.). https://bitbucket.org/genomicepidemiology/phylodash/src/main/ [accessed 18 April 2023].
35. OpenStreetMap contributors. Planet dump retrieved from; 2017. https://planet.osm.org
36. Phylocanvas.gl [Internet]. Phylocanvas.gl; (n.d.). https://www.phylocanvas.gl/
37. Kooplex [Internet]; (n.d.). https://k8plex-veo.vo.elte.hu/hub/
38. CoVEO: COVID-19 Data Portal. Internet; (n.d.). https://www.covid19dataportal.org/coveo [accessed 17 February 2023].

39. Johns Hopkins Coronavirus Resource Center; (n.d.). https://coronavirus.jhu.edu/map.html [accessed 17 February 2023].

40. Mentes A, Papp K, Visontai D, Stéger J, Csabai I, *et al*. Identification of mutations in SARS-CoV-2 PCR primer regions. *Sci Rep* 2022;12:18651.

41. Webin-CLI Submission — ENA Training Modules 1 documentation [Internet]; (n.d.). https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html [accessed 18 April 2023].

42. SARS-CoV-2 Drag and Drop Uploader [Internet]; (n.d.). https://ebi-ait.github.io/sars-cov2-data-upload/app-documentation [accessed 17 February 2023].

43. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017;22:30494.

44. ena-content-dataflow/scripts/gisaid_to_ena at master · enasequence/ena-content-dataflow [Internet]. GitHub; (n.d.). https://github.com/enasequence/ena-content-dataflow [accessed 17 February 2023].

45. ENA Webin-CLI Bulk Submission Tool [Internet]. European Nucleotide Archive; 2022. https://github.com/enasequence/ena-bulk-webincli [accessed 17 February 2023].

46. ena-analysis-submitter [Internet]. European Nucleotide Archive; 2022. https://github.com/enasequence/ena-analysis-submitter [accessed 17 February 2023].

47. Haak LL, Fenner M, Paglione L, Pentz E, Ratner H. ORCID: a system to uniquely identify researchers. *Learn Publish* 2012;25:259–264.

48. Institute EB. ORCID claiming | EBI Search | EMBL-EBI [Internet]; (n.d.). https://www.ebi.ac.uk/ebisearch/documentation/orcid-claiming [accessed 17 February 2023].

49. Liu J. Digital Object Identifier (DOI) and DOI services: an overview. *Libri* 2021;71:349–360.

50. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, *et al*. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res* 2018;46:D1266–D1270.

51. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;579:265–269.

52. CDC. Coronavirus Disease 2019 (COVID-19) [Internet]. Centers for Disease Control and Prevention; 2020. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html [accessed 18 April 2023].

53. COVID-19 Data Portal - accelerating scientific research through data [Internet]; (n.d.). https://www.covid19dataportal.org/statistics [accessed 18 April 2023].

54. The Pathogens Portal [Internet]; (n.d.). https://www.ebi.ac.uk/ena/pathogens/v2/ [accessed 17 February 2023].

55. COVID-19 Data Portal - Viral Seqeunces [Internet]; (n.d.). https://www.covid19dataportal.org/search/sequences [accessed 20 February 2023].

56. COVID-19 Data Portal - Systematic Analysis [Internet]; (n.d.). https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-analysis-covid19&size=15 [accessed 20 February 2023].

57. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, *et al*. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–319.

58. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014;2014:2.

59. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. *PLoS One* 2017;12:e0177459.

60. BigQuery API [Internet]. Google Cloud; (n.d.). https://cloud.google.com/bigquery/docs/reference/rest [accessed 17 February 2023].

61. Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, *et al*. The European variation archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res* 2022;50:D1216–D1220.

62. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, *et al*. Ensembl 2021. *Nucleic Acids Res* 2021;49:D884–D891.

63. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, *et al*. The ensembl variant effect predictor. *Genome Biol* 2016;17:122.

64. De Silva NH, Bhai J, Chakiachvili M, Contreras-Moreira B, Cummins C, *et al*. The ensembl COVID-19 resource: ongoing integration of public SARS-CoV-2 data. *Nucleic Acids Res* 2022;50:D765–D770.

65. pkrisz5. CoVEO [Internet]; (n.d.). https://github.com/pkrisz5/coveo [accessed 18 April 2023].

66. Chen Z, Azman AS, Chen X, Zou J, Tian Y, *et al*. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat Genet* 2022;54:499–507.

67. Asokan GV, Ramadhan T, Ahmed E, Sanad H. WHO global priority pathogens list: a bibliometric analysis of medline-PubMed for knowledge mobilization to infection prevention and control practices in Bahrain. *Oman Med J* 2019;34:184–193.

68. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 2022;50:W345–51.

69. CRG Viral Beacon - Info [Internet]; (n.d.). https://covid19beacon.crg.eu/info [accessed 17 February 2023].

70. Chen C, Nadeau S, Yared M, Voinov P, Xie N, *et al*. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* 2022;38:1735–1737.

71. Yang J. Cloud computing for storing and analyzing petabytes of genomic data. *J Ind Inf Integr* 2019;15:50–57.

72. Nextstrain / ncov / open / global / all-time [Internet]; (n.d.). https://nextstrain.org/ncov/open/global/all-time [accessed 18 April 2023].

73. Sanderson T. Taxonium, a web-based tool for exploring large phylogenetic trees. *Elife* 2022;11:e82392.