nature methods

Resource

https://doi.org/10.1038/s41592-023-01769-3

Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations

Received: 3 June 2022

Accepted: 17 January 2023

Published online: 23 February 2023

Check for updates

Karthik Gangavarapu[1,2] ✉, Alaa Abdel Latif[2], Julia L. Mullen[3], Manar Alkuzweny[2,4], Emory Hufbauer[2], Ginger Tsueng[3], Emily Haag[3], Mark Zeller[2], Christine M. Aceves[2], Karina Zaiets[3], Marco Cano[3], Xinghua Zhou[3], Zhongchao Qian[3], Rachel Sattler[5], Nathaniel L. Matteson[2], Joshua I. Levy[2], Raphael T. C. Lee[6,7], Lucas Freitas[6,8], Sebastian Maurer-Stroh[6,7,9,10], GISAID Core and Curation Team*, Marc A. Suchard[1,11,12], Chunlei Wu[3,13,14], Andrew I. Su[3,13,14], Kristian G. Andersen[2,13] & Laura D. Hughes[3] ✉

In response to the emergence of SARS-CoV-2 variants of concern, the global scientific community, through unprecedented effort, has sequenced and shared over 11 million genomes through GISAID, as of May 2022. This extraordinarily high sampling rate provides a unique opportunity to track the evolution of the virus in near real-time. Here, we present outbreak.info, a platform that currently tracks over 40 million combinations of Pango lineages and individual mutations, across over 7,000 locations, to provide insights for researchers, public health officials and the general public. We describe the interpretable visualizations available in our web application, the pipelines that enable the scalable ingestion of heterogeneous sources of SARS-CoV-2 variant data and the server infrastructure that enables widespread data dissemination via a high-performance API that can be accessed using an R package. We show how outbreak.info can be used for genomic surveillance and as a hypothesis-generation tool to understand the ongoing pandemic at varying geographic and temporal scales.

In December 2019, a series of cases of pneumonia of unknown origin appeared in Wuhan, China and on 7 January 2020, the virus responsible for the diseases was identified as a novel coronavirus, SARS-CoV-2 (ref. [1]). The first SARS-CoV-2 genome was made publicly available on 10 January 2020 (refs. [2,3]). Since then, the global scientific community, through an unprecedented effort, has sequenced and shared over 11 million genomes through GISAID (https://gisaid.org/), as of May 2022 (ref. [4]). To keep track of the evolving genetic diversity of SARS-CoV-2, Rambaut

[1]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. [2]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA. [3]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. [4]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA. [5]Skaggs Graduate School of Biological and Chemical Sciences, The Scripps Research Institute, La Jolla, CA, USA. [6]GISAID Global Data Science Initiative, Munich, Germany. [7]Bioinformatics Institute & ID Labs, Agency for Science Technology and Research, Singapore, Singapore. [8]Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil. [9]National Centre for Infectious Diseases, Ministry of Health, Singapore, Singapore. [10]Department of Biological Sciences, National University of Singapore, Singapore, Singapore. [11]Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. [12]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, USA. [13]Scripps Research Translational Institute, La Jolla, CA, USA. [14]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA. *A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: gkarthik@ucla.edu; lhughes@scripps.edu

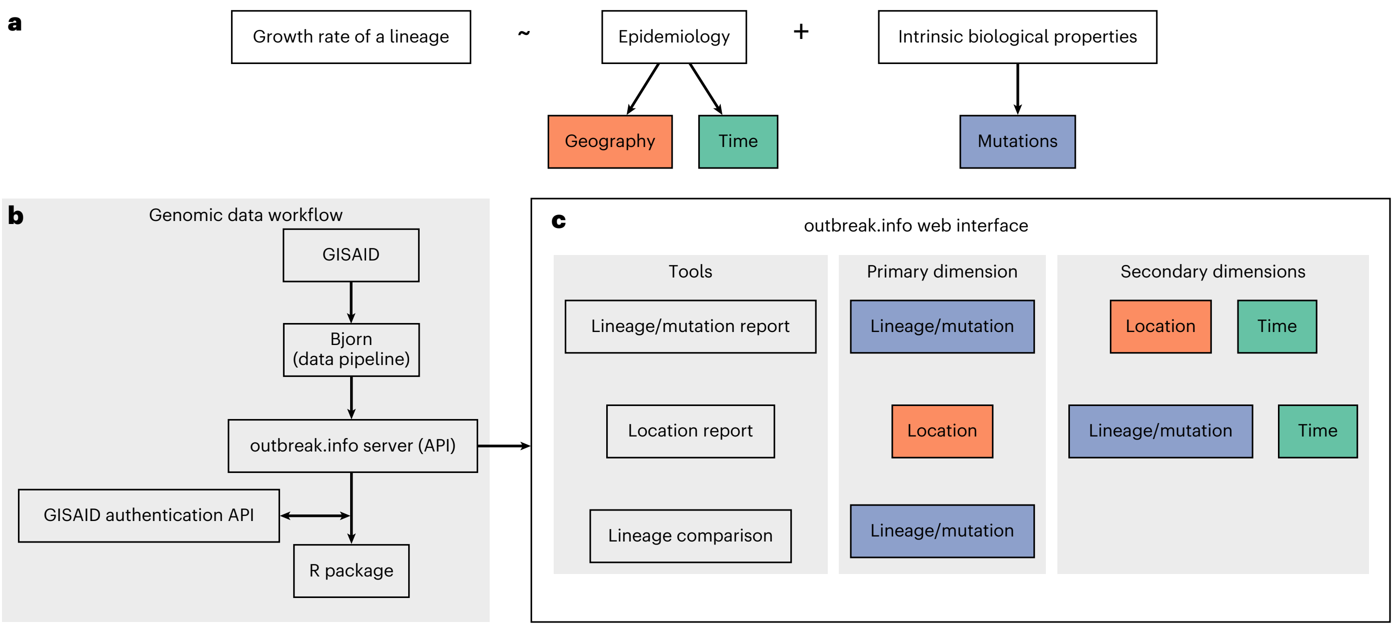


Fig. 1 | outbreak.info enables the exploration of genomic data across three dimensions. a, Growth rate of a lineage is a function of epidemiology and intrinsic biological properties of a lineage. Further, epidemiology varies over time and by geography, whereas intrinsic biological properties are determined by the mutations present in a given lineage. **b**, Genomic data are ingested from GISAID, processed using the custom-built data pipeline (`Bjorn`) and stored on a server that can be accessed via an API. The API is consumed by two clients: a JavaScript-based web client and an R package that provides programmatic access by authenticating against GISAID credentials. **c**, The web interface contains three tools that allow exploration of genomic data across three different dimensions: lineage/mutation, time and geography.

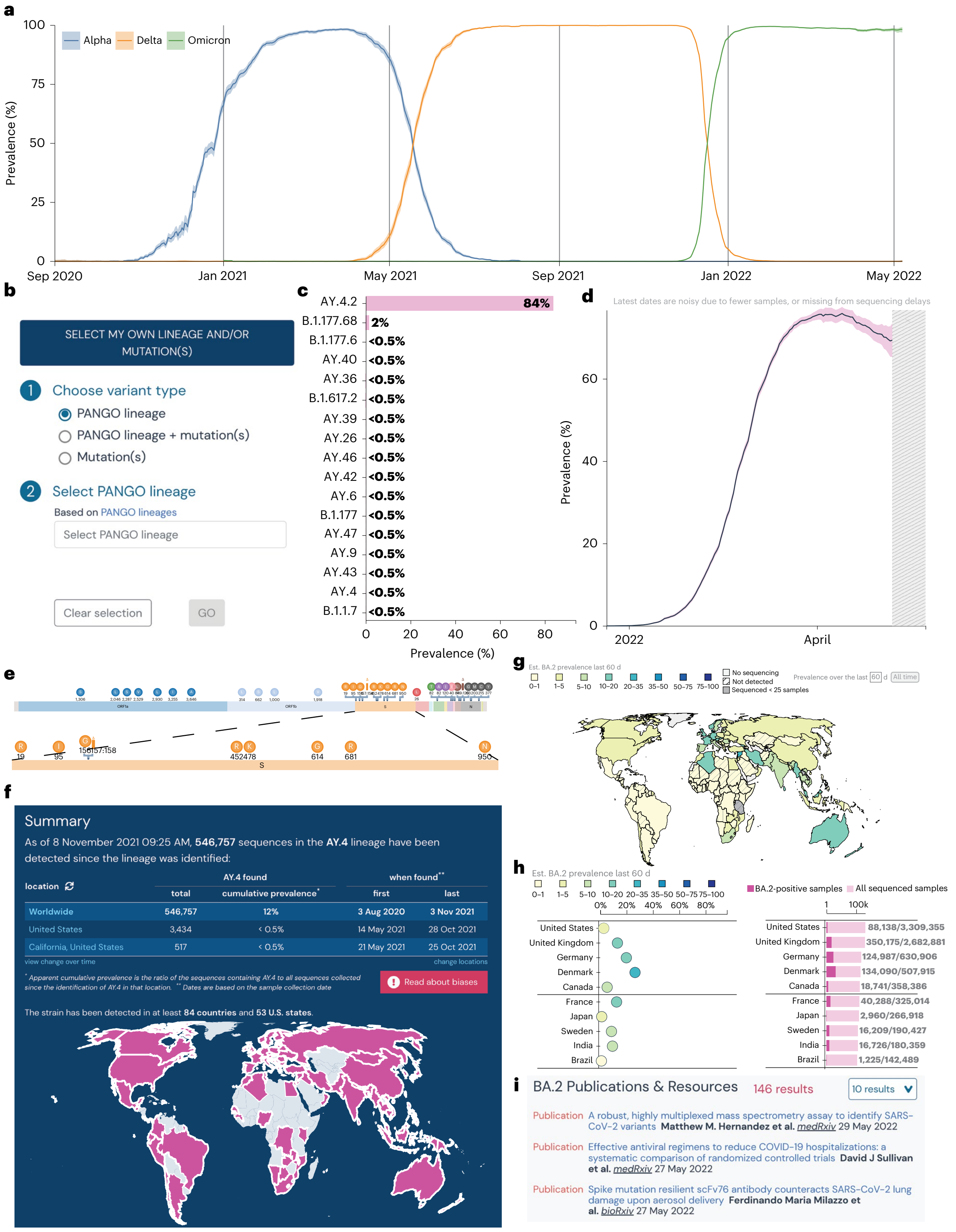
et al. developed a dynamic phylogeny-informed nomenclature (Pango) to classify SARS-CoV-2 lineages[5]. As of May 2022, over 2,000 lineages have been designated, which has enabled public health agencies such as Public Health England, the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) to identify variants of concern (VOCs), variants of interest/variants under investigation (VOIs/VUIs) and variants under monitoring/variants being monitored (VUMs/VBMs) based on the phenotypical characterization of these lineages[6]. Currently, the only designated VOC is the B.1.1.529* (Omicron) lineage, which exhibited very rapid growth and the ability to substantially avoid antibody neutralization[7,8]. There have been four VOCs previously designated over the course of the pandemic: the B.1.1.7* (Alpha; * denotes the lineage and any of its sublineages) lineage, resulting in increased transmissibility[9]; the B.1.351* (Beta) lineage, exhibiting immune evasion[10]; the P.1* (Gamma) lineage, exhibiting immune evasion[11]; and the B.1.617.2* (Delta) lineage, exhibiting increased transmissibility due to the P681R mutation in the spike gene[12].

The emergence of VOCs with fitness advantages has led to global ‘sweeps’, with newly emerged VOCs displacing previously circulating variants. More notably, the growth of each VOC has led to a renewed surge in infections worldwide. This has prompted the need for near real-time genomic surveillance to inform early public health interventions to control the rise of infections. In response to this need, thousands of academic, non-academic and public health laboratories have been depositing sequences predominantly on the sharing platform of the GISAID Initiative[4,13]. The extraordinarily high sampling rate of infecting viruses provides a unique opportunity to track the evolution of the virus in near real-time. For example, in December 2021 alone, over a million new genomes were submitted to GISAID[14]. Traditionally, phylodynamic approaches have been employed to retrospectively characterize lineage dynamics during outbreaks of viruses such as Zika[15–17], West Nile[18] and Ebola[19,20]. Existing tools such as NextStrain[21] and frameworks such as Microreact[22] primarily rely on a phylogeny to elucidate transmission chains and monitor the evolution of the virus. These tools, however, were not designed to track thousands of new genomes per day and given that building phylogenies for large sets of genomes is computationally intensive and time consuming, obtaining timely insights from the data is often problematic[23]; however, the high sampling rate of the virus has opened up the possibility of tracking the pandemic using available near real-time genomic data without the need for computationally intensive modeling.

Here, we present outbreak.info, a platform that currently tracks over 40 million combinations of Pango lineages and individual mutations, across over 7,000 locations, to provide insights for researchers, public health officials and the general public. In the following sections, we describe the data pipelines that enable the scalable ingestion and standardization of heterogeneous data on SARS-CoV-2 variants, the server infrastructure that enables the dissemination of the processed data and the client-side applications that provide intuitive visualizations of the underlying data.

Results

The growth rate of a given viral lineage is a function of epidemiology and its intrinsic biological properties (Fig. 1a). For example, the B.1.177 lineage, characterized by an A222V amino acid substitution in the spike gene, increased in prevalence in Europe during the summer of 2020 (ref. [24]). While initially thought to be more transmissible, it was eventually shown that the increase in prevalence was due to a resurgence in travel and not due to increased transmissibility. In contrast, a few months later, the B.1.1.7 lineage was shown to be 40–60% more transmissible than previously circulating lineages and this intrinsic biological property led to the rapid growth in its prevalence worldwide[25,26]. Epidemiological factors such as mobility[27,28], mask usage[29] and public health interventions[30] vary over time and across geographies worldwide, whereas biological properties are a function of the mutations found in a given lineage (Fig. 1a). Hence, to maximize the utility of genomic data for surveillance, we built outbreak.info to enable the exploration of genomic data across three dimensions: geography, time and



SELECT MY OWN LINEAGE AND/OR MUTATION(S)

1 Choose variant type

● PANGO lineage

○ PANGO lineage + mutation(s)

○ Mutation(s)

2 Select PANGO lineage

Based on PANGO lineages

Select PANGO lineage

Clear selection GO

Summary

As of 8 November 2021 09:25 AM, **546,757** sequences in the **AY.4** lineage have been detected since the lineage was identified:

location	AY.4 found		when found**	
	total	cumulative prevalence*	first	last
Worldwide	546,757	12%	3 Aug 2020	3 Nov 2021
United States	3,434	< 0.5%	14 May 2021	28 Oct 2021
California, United States	517	< 0.5%	21 May 2021	25 Oct 2021

view change over time change locations

* Apparent cumulative prevalence is the ratio of the sequences containing AY.4 to all sequences collected since the identification of AY.4 in that location. ** Dates are based on the sample collection date

Read about biases

The strain has been detected in at least **84 countries** and **53 U.S. states**.

	BA.2-positive samples/All sequenced samples
United States	88,138/3,309,355
United Kingdom	350,175/2,682,881
Germany	124,987/630,906
Denmark	134,090/507,915
Canada	18,741/358,386
France	40,288/325,014
Japan	2,960/266,918
Sweden	16,209/190,427
India	16,726/180,359
Brazil	1,225/142,489

BA.2 Publications & Resources 146 results 10 results

Publication A robust, highly multiplexed mass spectrometry assay to identify SARS-CoV-2 variants **Matthew M. Hernandez et al.** *medRxiv* 29 May 2022

Publication Effective antiviral regimens to reduce COVID-19 hospitalizations: a systematic comparison of randomized controlled trials **David J Sullivan et al.** *medRxiv* 27 May 2022

Publication Spike mutation resilient scFv76 antibody counteracts SARS-CoV-2 lung damage upon aerosol delivery **Ferdinando Maria Milazzo et al.** *bioRxiv* 27 May 2022

Fig. 2 | Lineage and/or Mutation Tracker. a, Prevalence of VOCs in the United Kingdom from Sep 2020 to May 2022. The error bands show the 95% binomial proportion confidence interval calculated using Jeffrey's interval. **b**, Search and filter options for Lineage/Variant of Concern tracker. **c**, Prevalence of S:Y145H+ S:A222V mutations across different lineages globally. **d**, Prevalence of BA.2 in the United Kingdom. The error bands show the 95% binomial proportion confidence interval calculated using Jeffrey's interval. **e**, Mutation map showing the characteristic mutations of AY.4. **f**, Summary statistics of BA.2 lineage. **g**, Geographic distribution of the cumulative prevalence of BA.2 lineage over the last 60 d globally. **h**, Cumulative prevalence of BA.2 in each country over the last 60 d globally. **i**, Research articles and datasets related to BA.2.

Table 1 | Questions addressed by the Lineage and/or Mutation Tracker

Question	Relevant visual elements
What is the prevalence of a set of mutations within different lineages?	Mutations such as S:N501Y, S:DEL69/70 and S:E484K have been shown to have functional impact on the phenotype exhibited by a lineage such as increased pathogenicity or immune evasion[33,60,61]. Furthermore, these mutations have been acquired independently by many lineages. Convergent evolution can be used as a metric to assess the importance of any advantage conferred on a lineage by a mutation. Hence, if a query contains a set of mutations (for example, S:E484K and S:N501Y), we estimate the prevalence of that set of mutations across all lineages globally (Fig. 2c).
What is the trend shown by the prevalence of a lineage and/or a set of mutations over time?	Tracking the growth rate of a lineage or a set of mutations over time is very important to inform public health interventions. We estimate the prevalence of a given query as a proportion of the total number of sequences collected on a given day at a given location. To convey the uncertainty in estimating the prevalence, we calculate binomial proportion confidence intervals using Jeffrey's interval (Fig. 2d).
What are the 'characteristic mutations' of a lineage?	The mutations that are characteristic of a lineage can be used to generate hypotheses about the phenotype exhibited by a lineage based on previous studies on the functional impact of mutations. This is especially important to assess any potential impact a lineage might have on therapeutics such as monoclonal antibody drugs. We define the 'characteristic mutations' of a lineage as those mutations found in at least 75% of the genomes classified as the lineage. We determined the 75% threshold empirically by examining the frequency of prevalent mutations within a variant (Extended Data Fig. 1a). We found that this threshold effectively excluded any mutations that occurred at low prevalence and identified the defining mutations of a lineage. We verified this using a manual check against known defining mutations of existing lineages (Extended Data Fig. 1b). Further, when a new lineage is designated, we perform manual checks to ensure that this 75% threshold is working as expected. The characteristic mutations identified in this manner are displayed in a 'mutation map' (Fig. 2e).
What is the total number of sequences that belong to a lineage and/or a set of mutations? In how many countries was a lineage and/or a set of mutations detected? When was this lineage and/or a set of mutations first detected?	To assess how quickly a variant spread and the extent of the geographic spread, we show summary of relevant statistics such as the total number of sequences that match the query, the cumulative prevalence of these mutations, the first and last date a sequence matching the query was detected worldwide for a customizable set of locations (Fig. 2f).
What is the geographic prevalence of a lineage and/or a set of mutations?	Many lineages, including VOCs Beta and Gamma, show variation in growth rates across different locations. Hence, it is essential to be able to access the geographic distribution of a given lineage. To facilitate this, we show the cumulative prevalence of lineages over a flexible time window across the sub-administrative levels of a given location for a lineage/mutation query (Fig. 2g). Choropleths are useful visual elements to map geographic variation in prevalence but to further highlight the uncertainty in these estimates and to account for cognitive biases in evaluating locations with different land areas, we use a dot chart to show the uncertainty in the point estimate of prevalence over the a flexible time window and a bar chart to show the number of sequences used to calculate it (Fig. 2h). The flexible time window for the cumulative prevalence estimates (Fig. 2g,h) is set to the last 60 d by default but can be changed as required. These two charts can be sorted by the prevalence of the query or the total number of sequences that match the query. This allows the user to account for the effects of sampling bias on prevalence estimates.
What is the latest research available on this lineage and/or set of mutations?	With the growth of new variants over the pandemic, we have seen many studies that focus on important aspects of a lineage such as the ability to evade immune response and the impact on vaccine efficacy. To aid in the discoverability of preprints, publications, datasets and other resources, we show the entries that match a given lineage or mutation query from our up-to-date research library[32] (Fig. 2i). The details of the Research Library have been described previously[32].

lineages/mutations. We use the Pango nomenclature to estimate the prevalence of SARS-CoV-2 lineages over time and at varying geographic scales. Using a phylogenetically-informed nomenclature allows us to determine genetic features such as the 'characteristic mutations' of a lineage without directly building a global phylogeny. By avoiding a global phylogeny, we can update our databases daily using the continuously growing number of SARS-CoV-2 genomes. In addition, we closely track reports from health agencies such as Public Health England, CDC and WHO that designate VOCs/VOIs/VUMs based on epidemiological analyses. In addition to genomic data, the server also ingests two other types of data: (1) epidemiological data curated by Johns Hopkins University[31] and (2) public literature, clinical trial, protocol and dataset metadata from sources such as bioRxiv, medRxiv and LitCovid[32]. Here, we describe how each of these data sources is used collectively to assist in genomic surveillance.

The overall workflow of genomic data is shown in Fig. 1b. Genomic data are ingested from GISAID, processed via a custom-built data pipeline, `Bjorn`, and stored on a server that can be accessed via an application programming interface (API). We built two client-side applications, a web interface and an R package that consume this API (Fig. 1b). The web interface consists of three main tools focusing on different facets of the underlying genomic data: (1) Lineage and/or Mutation Tracker, (2) Location Tracker and (3) Lineage Comparison Tool. We designed an interface for each tool that focuses on one primary dimension of the genomic data with additional customizability of one or more secondary dimensions (Fig. 1c). The Lineage and/or Mutation Tracker focuses on a specific lineage, mutation or a combination of these. The Location Tracker focuses on a given location and provides a snapshot of currently circulating lineages. Finally, the Lineage Comparison Tool can be used to explore the prevalence of mutations across different lineages. In addition to the web interface, we built an R package that authenticates against GISAID credentials and allows programmatic access to the processed data for downstream analyses.

Lineage and/or Mutation Tracker

The ongoing SARS-CoV-2 pandemic has been punctuated by the emergence of VOCs with fitness advantages over previously circulating variants, resulting in 'waves' of infections. Figure 2a shows the changing prevalence of the three most-dominant VOCs in the United Kingdom, but this phenomenon is observed globally with heterogeneity across geography. A fundamental part of genomic surveillance is to identify the emergence of such variants by closely tracking the growth of circulating lineages. Given the geographic variation in epidemiological, social and economic factors, it is important to estimate variant prevalence at varying geographic scales. The Lineage/Mutation Tracker can be used to dynamically query the temporal and geographic variation in the prevalence of a (1) VOC/VOI and its sublineages (for example, Delta and its sublineages); (2) a lineage (for example, B.1.1.7); (3) a lineage and one or more mutations (for example, B.1.1.7 with S:E484K); (4) a mutation (for example, S:E484K); or (5) a group of mutations (for example, S:E484K and S:N501Y) (Fig. 2b). When a mutation is part of a query, we also generate and display a link to Aquaria[33] so that users can visualize the amino acid site in the context of the three-dimensional structure of the relevant protein. In addition to customizing the lineages and mutations, users can specify various location scales, such as a country, state or county (or their local equivalents), to estimate the prevalence of a given lineage and/or mutations. To provide meaningful insights from these prevalence estimates, we designed an interface to address a specific set of questions listed in Table 1.

Location Tracker

Some variants of the virus have only been dominant in certain regions of the world. For example, the VOCs Beta and Gamma were dominant in South Africa[10] and Brazil[34], respectively. Similarly, B.1.621 (ref. [35]) was only dominant in Columbia, A.2.5 was only dominant in Panama and B.1.177 exhibited a high growth rate only in European countries due to a resurgence of travel in the summer of 2020 (refs. [24,36]). Factors such as the attack rate, population immunity due to previous infection or vaccination and social mobility vary by geographic region and have an impact on the growth rates exhibited by a given lineage. To account for such localized factors, it is important to have the ability to track the growth of lineages at different geographic scales. We built the Location Tracker on outbreak.info to facilitate the surveillance of SARS-CoV-2 lineages at a country, state/province or county/city level. The Location Tracker provides a snapshot of circulating lineages with a focus on the last 60 d and allows users to compare the prevalence of a customizable set of lineages/mutations over time in that location. Furthermore, the tracker also integrates reported cases over time to provide insights on the impact of growth of various lineages on caseloads in the region. As with the Lineage/Mutation Tracker, we designed the user interface to answer a set of specific questions as shown in Table 2.

Case study: outbreak.info as a hypothesis-generation tool

As the pandemic has continued to progress, we have seen the emergence of VOCs with fitness advantages that were able to outcompete previously circulating lineages. As of May 2022, there has been one currently designated VOC (Omicron, B.1.1.529+ sublineages, indicated by *) and four previously designated VOCs: Alpha (B.1.1.7*), Beta (B.1.351*), Gamma (P.1*) and Delta (B.1.617.2*). Of these, Alpha, Beta and Gamma were estimated to have emerged between September and December 2020 (refs. [10,11,37]) and were subsequently outcompeted globally by the Delta variant that was first detected in December 2020 (ref. [38]). The Omicron lineage, first detected in November 2021 (ref. [7]), was able to outcompete Delta and grew much more rapidly relative to previous VOCs during their emergence (Fig. 4a). Where Delta and Omicron variants exhibited high growth rates with little variation globally, Alpha continued to circulate in low prevalence in Brazil and South Africa, where Gamma and Beta variants were dominant, respectively (Fig. 4b,c). Additionally, the prevalence of sublineages within Delta and Omicron variants varies geographically. The Location Tracker on outbreak.info can be used to track the growth of VOCs within a given location, thus facilitating the comparison of lineage growth rates across locations. The Location Tracker can also be used to track the relative prevalence of sublineages within these VOCs, shedding light on any geographic variation in these dynamics. Here, we examine trends in the prevalence of the five VOCs globally and highlight the geographic variation in growth rates of Alpha, Beta, Gamma, Delta and Omicron variants.

The earliest samples of the Alpha variant were sequenced in southern England in late September 2020 (ref. [37]). There were multiple introductions of the lineage into the United States as early as late November[26]. The Alpha variant showed a transmission advantage of 40–50% in the United States[26], in line with observations in the United Kingdom and the Netherlands. In the United States, Alpha was able to outcompete previously circulating lineages and continued to increase in prevalence until the introduction of the Delta variant around April 2021 (Fig. 4d). In contrast to the United States, the Alpha variant circulated at very low prevalence in Brazil, whereas the Gamma variant remained dominant in the country[11] until the introduction of the Delta variant around April 2021 (Fig. 4b). Similarly, in South Africa, the Beta variant continued to spread until the emergence of the Delta variant and the Alpha variant never became dominant (Fig. 4c). While the Beta and Gamma variants were able to outcompete Alpha in South Africa and Brazil, respectively, Gamma only reached a maximum prevalence of 8% in the United States in May 2020 and Beta circulated at a prevalence of <1% (Fig. 4d). The growth of a lineage is determined by epidemiological factors such as number of introductions, travel between locations and by intrinsic biological properties such as transmission advantage or immune evasion. Both Beta and Gamma variants show varying degrees of immune evasion[39]. Regions of Brazil had attack rates as high as 75% in October 2020 (ref. [40]), indicating that immune evasion was the primary reason for the rapid growth of the P.1 lineage in Brazil. In contrast,

Table 2 | Questions addressed by the Location Tracker

Question	Relevant visual elements
What are the most prevalent lineages over the last 60 d?	To quickly provide a snapshot of the lineages currently circulating in a given location, we show a streamgraph of the prevalence of lineages over the last 60 d (Fig. 3a). To increase interpretability, we grouped lineages that are below 3% prevalence for at least 5 d over the last 60 d into a separate category, 'Other'. The prevalence over time can be skewed especially in recent days due to the lag between sample collection, sequencing and the deposition of sequence data. To convey this uncertainty, the total number of samples collected are shown in an inverted bar graph below the streamgraph. In addition, a stacked bar graph shows a snapshot of the cumulative prevalence of the lineages over the last 60 d (Fig. 3b). Additionally, the user can adjust this window to look at different time windows, for example 180 d.
What is the distribution of mutations across these lineages?	The Location Tracker shows a snapshot of currently circulating lineages which will help identify a newly emerging lineage that exhibits a high relative growth rate. Often in such cases, the mutations found in the lineage might provide preliminary evidence on phenotypes exhibited by the virus such as increased transmissibility or immune evasion. To facilitate this process, we show the prevalence of mutations that are present in the spike gene of at least 75% of the sequences of currently circulating lineages (Fig. 3c). A Lineage Comparison Tool is also available, which expands upon this functionality with customizable queries to add lineages based on the name, VOC/VOI classification, prevalence of mutations and prevalence within a location.
How does the prevalence of different lineages or mutations within this location change over time?	In addition to showing a snapshot of the lineages circulating over the last 60 d, we developed a component to show the temporal variation in the prevalence of a customizable set of lineages/mutations for a given location. This offers additional flexibility to dynamically select lineages or mutations of interest and compare their prevalence over time with a customizable time window (Fig. 3d).
How does the lineage prevalence over time correspond to the number of daily reported cases in this region?	The impact of lineage dynamics on the reported cases over time is of primary concern to public health. To accomplish this, we cross-linked the reported cases for each location using a standardized location identifier and this is shown in a line graph below the prevalence of a lineage (Fig. 3e). In addition, users can select a time range within the prevalence chart or the reported cases chart to compare trends over a shorter time span.

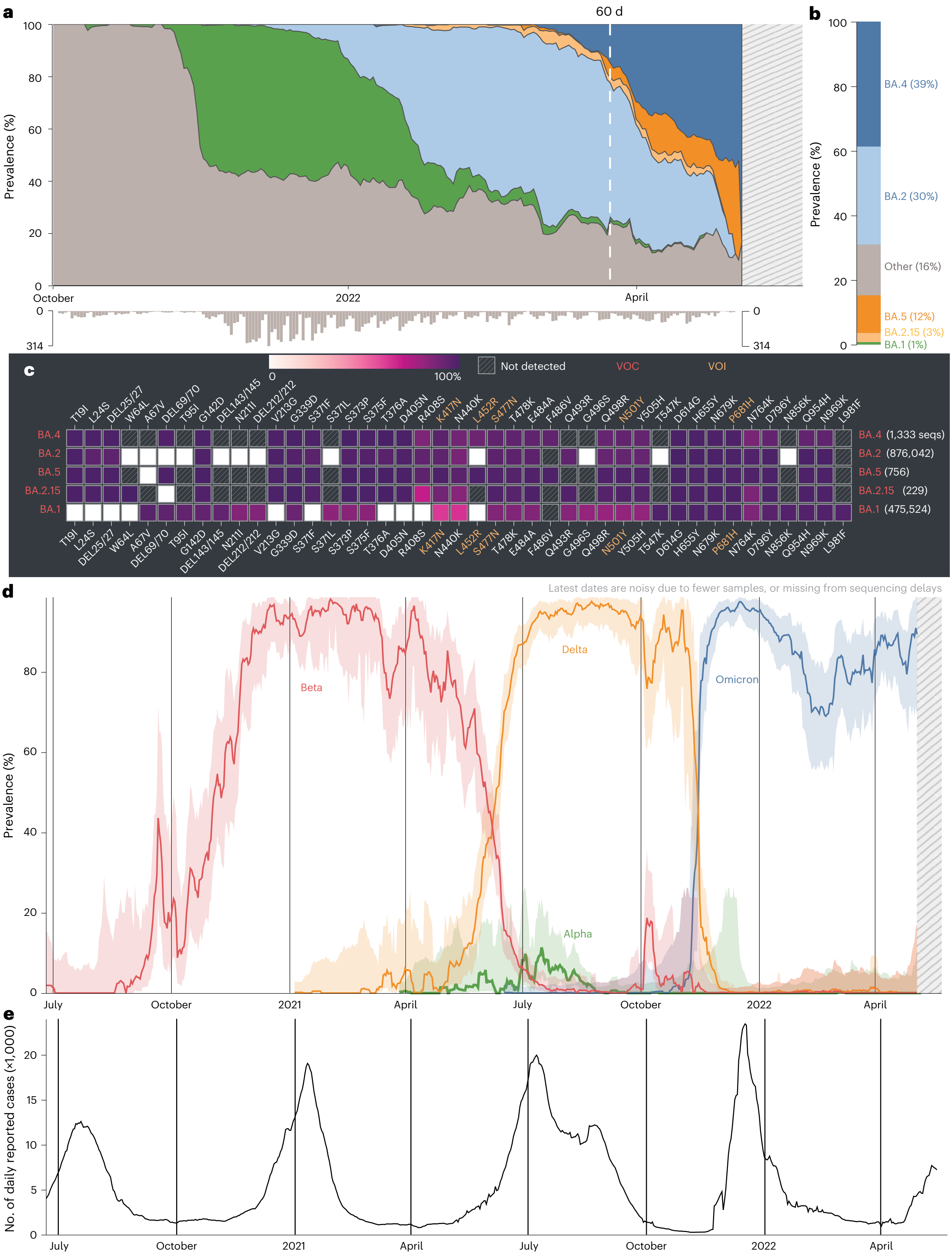


Fig. 3 | Location report. a, Relative prevalence of all lineages over time in South Africa. Total number of sequenced samples collected per day are shown in the bar chart below. **b**, Relative cumulative prevalence of all lineages over the last 60 d in South Africa. **c**, Mutation prevalence across the most prevalent lineages in South Africa over the last 60 d. **d**, Comparison of the prevalence of VOCs grouped by WHO classification: Alpha, Beta, Delta and Omicron over time in South Africa. The error bands show the 95% binomial proportion confidence interval calculated using Jeffrey's interval. **e**, Daily reported cases in South Africa are shown in the line chart.

states in the United States had an estimated attack rate between 0.1% and 16% in June 2020 (ref. [41]). Given this difference in attack rates, we can hypothesize that the intrinsic transmission advantage of the Alpha variant was able to outcompete the advantage conferred by immune evasion of Gamma in the United States, but the opposite was true in Brazil and South Africa. In all three countries, the introduction of the Delta lineage displaced previously circulating Alpha, Beta and/or Gamma lineages in the summer of 2021.

The Delta variant of SARS-CoV-2 was first detected in Maharashtra, India in December 2020 (ref. [38]), has been shown to be 40%–60% more transmissible than Alpha[42,43] and causes a reduction in vaccine efficacy relative to previously circulating lineages[44]. Vaccination campaigns against COVID-19 started in December 2020 and despite the progress of these campaigns[45], the Delta variant continued to cause a renewed surge in infections globally. The Delta variant report, which can be accessed directly on the landing page of the Lineage and/or Mutation Tracker, can be used to understand the dynamics of its sublineages. Figure 4a shows the global prevalence of the Delta variant over time. This growth reflects the transmission advantage that Delta has over previously circulating lineages including VOCs Alpha, Beta and Gamma. As the Delta variant continued to spread, its genetic diversity increased and as of May 2022, over 200 sublineages of Delta have been designated[46].

The Omicron variant was first detected in November 2021 by genomic surveillance teams in South Africa and Botswana. This variant was associated with a rapid resurgence of infections in Gauteng Province, South Africa and was designated a VOC by the WHO within 3 d of uploading the first genome[7]. The variant grew in prevalence very rapidly: within 3 weeks, the variant was detected in 87 countries and as of May 2022, Omicron has a prevalence of over 95% globally (Fig. 4a). While increased transmissibility confers a bigger fitness advantage compared to immune evasion when population immunity is low, the opposite is true as population immunity increases either due to vaccination or previous infection[47]. The Omicron variant was found to have a fivefold higher chance of reinfection compared to Delta[48] and Omicron infections presented with a higher viral load than wild type but still lower than Delta[49]. As viral load is one of the determinants of transmissibility, this indicates that Omicron is intrinsically not as transmissible as Delta, but it exhibits better immune evasion. This combination gave Omicron a large fitness advantage over Delta as evidenced by its rapid growth rate worldwide (Fig. 4a). The continued spread of the variant has resulted in the emergence of many sublineages and as of May 2022, over 100 sublineages of Omicron have been designated. Notably, there is considerable geographic variation in the relative prevalence of newly designated sublineages such as BA.2.12.1, BA.4 and BA.5. While BA.2 continues to be the dominant sublineage within Omicron in countries such as Denmark and the United Kingdom (Fig. 4e,f), we see the BA.2.12.1 sublineage slowly displacing BA.2 in the United States (Fig. 4g). In South Africa, sublineages BA.4 and BA.5 have completely displaced the previously dominant BA.2 (Fig. 4h) and have led to another surge in reported cases (Fig. 3e). The three variants, BA.2.12.1, BA.4 and BA.5 have been shown to evade antibodies elicited by previous BA.1 infection in in vitro neutralization studies[50,51]. This observed escape was higher than what was observed for BA.2 (ref. [52]), highlighting the possibility that these variants led to a renewed surge in infections as these variants continue to spread globally. While the growth of Alpha and Delta variants globally was driven primarily by higher intrinsic transmissibility, the growth of the new variants within Omicron seems to be driven primarily by enhanced immune evasion. The increasing prevalence of immunity due to vaccination or previous infection worldwide, further supports this hypothesis.

This case study illustrates how outbreak.info can be used to not only track and compare the prevalence of lineages across locations, but also to derive and support hypotheses regarding the complex interplay between epidemiology and the intrinsic phenotypic characteristics of emerging SARS-CoV-2 lineages as the virus continues to spread.

Discussion

The Omicron variant, first detected in late November 2021, has outcompeted Delta and as of May 2022, it is the dominant lineage globally; however, it is important to note that regardless of how prevalent previously circulating VOCs were, all five VOCs emerged independent of each other. While the current hypothesis for the emergence of VOCs is prolonged virus evolution in a chronically infected individual[53], we still lack a thorough understanding of this process. Given the underlying stochasticity of this process, predicting the emergence of a new VOC is not currently feasible. As a result, continued surveillance of all currently circulating lineages is of utmost importance to public health globally—particularly as SARS-CoV-2 continues to spread and evolve worldwide.

The global community has generated over 11 million genomes of SARS-CoV-2 as of May 2022, shared on platforms such as GISAID[14]. The availability of such a large volume of genomic data has led to the development of several applications that combine genomic data with associated datasets to track the evolution of variants and their properties such as immune evasion and transmission advantage. Databases such as the Stanford Coronavirus Antiviral & Resistance Database[54] rely on manual curation of publications to identify and extract relevant datasets. Applications such as CoVariants.org track the evolution of the virus by focusing on a specific set of VOCs/VOIs and key mutations. COG-UK Mutation Explorer[55] on the other hand tracks the virus within a specific location. Most of these applications rely on varying degrees of manual and automated workflows.

In terms of the degree of automation and the generalized approach to tracking SARS-CoV-2 variants globally, the applications that are most comparable to outbreak.info include COVID-19 CG[56], CoV-Spectrum[57] and BV-BRC SARS-CoV-2 Early Warning System[58]. While each of these applications has their own strengths which can complement each other, we developed outbreak.info to serve as a template for tracking the spread of any pathogen over varying geographic and temporal scales at scale, across the world, in near real-time. Our unique paradigm centralizes the computation of key statistics based on the analysis of disparate data streams.

We designed the server infrastructure of outbreak.info keeping two goals in mind: scalability of the API as existing data sources increase in size and new data sources are incorporated and reusability of the computed data by providing programmatic access through an R package (Extended Data Fig. 2). We separated our data ingestion pipelines from the server-side application so that new data sources could be easily incorporated by instantiating a new instance of the same server-side database and API. This approach differs from other existing applications and enables us to quickly incorporate and link metadata across existing data and new modes of surveillance such as the CDC's National Wastewater Surveillance System[59]. Furthermore, the easy dissemination of any computed data on outbreak.info via the R package enables registered GISAID users to not only recreate visualizations available on the web interface but to further interrogate and utilize the processed data for more sophisticated downstream analyses. Extensive documentation of the R package is available at https://outbreak-info.github.io/R-outbreak-info/. To the best of our knowledge, we are the only application that has incorporated an authentication workflow to enable users to directly access and analyze the processed data. To maximize accessibility of these data, the web interface of outbreak.info was designed to offer a high degree of customizability, allowing users to answer specific biological questions and use the platform as a hypothesis-generation tool.

The guiding principles for the web interface were interactivity via responsive user interface elements powered by a high-performance API and interpretability via intuitive visualization of data based on

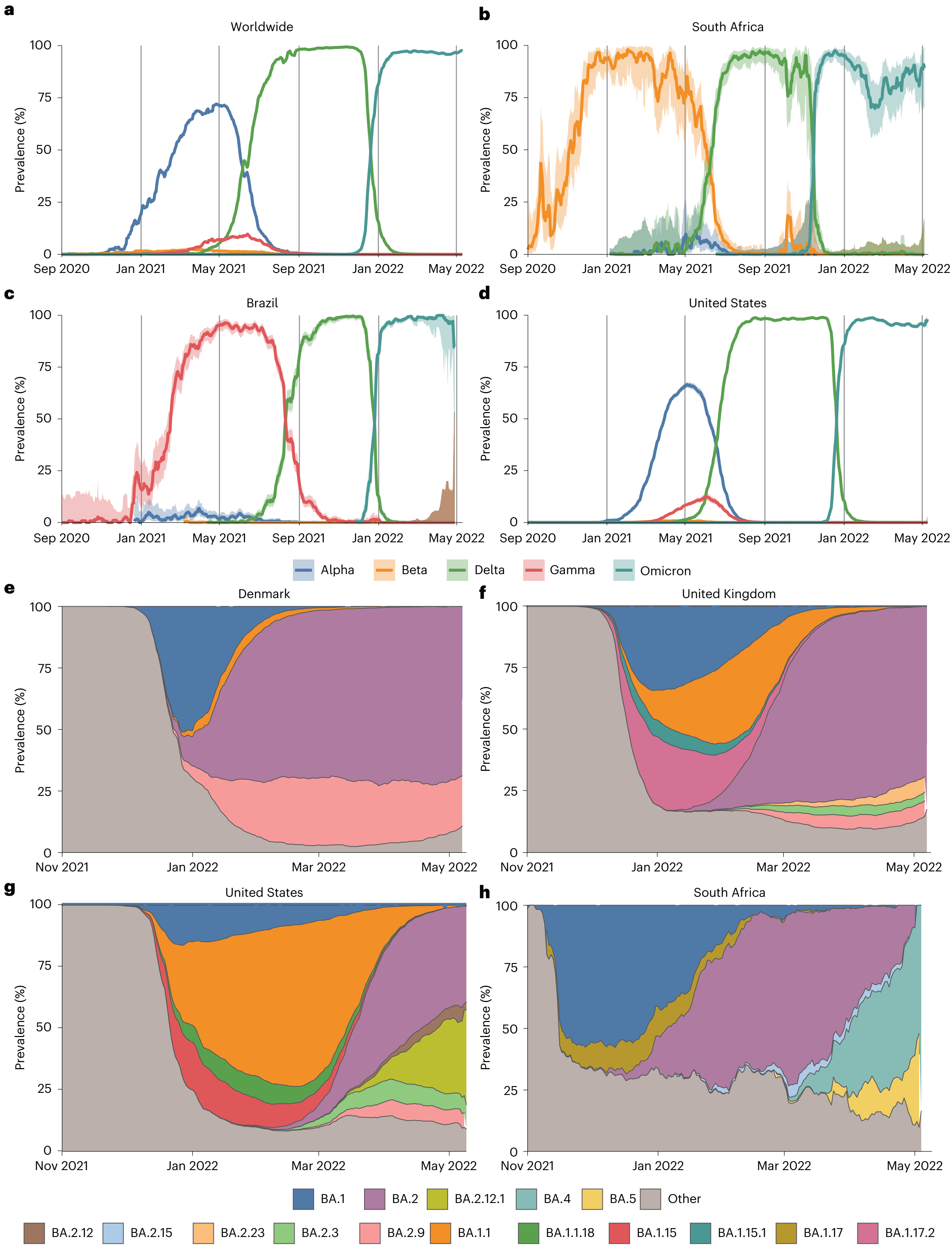


Fig. 4 | Prevalence of VOCs Alpha, Beta, Gamma, Delta and Omicron lineages over time. a–d, Prevalence worldwide (**a**), in South Africa (**b**), in Brazil (**c**) and in the United States (**d**). Error bands in **a**–**d** show 95% binomial proportion confidence intervals calculated using Jeffrey's interval. **e**–**h**, Lineages with a prevalence over 3% over the last 60 d in Denmark (**e**), United Kingdom (**f**), United States (**g**) and South Africa (**h**).

discussions with researchers, epidemiologists and public health officials. Besides informal discussions we also conducted two usability tests to evaluate user experience and redesigned parts of the user interface based on the feedback we received (see Usability studies in Methods). This rigorous approach to designing the interface and the overarching guiding principles have enabled us to create an interface that can be effectively used by the general public in addition to public health officials, academic and non-academic researchers.

outbreak.info has been enabled by unprecedented global genomic sequencing efforts and we developed every element of the application to fully leverage this capacity; however, genomic sampling varies globally with the vast majority of sequences coming from high income countries; even within well-sampled regions, there is geographic and temporal variation[13]. To communicate the increased uncertainty due to low sampling, we calculate confidence intervals of estimates wherever applicable, provide histograms of sampling density and mask data when there are very few data points available. Sampling strategies for sequencing could also bias the prevalence of a given variant. For example, samples that are prescreened for variants such as B.1.1.7 (Alpha) or BA.1 (Omicron) that show S gene dropout on a widely used qPCR assay could artificially inflate the prevalence of these variants. Efforts to investigate specific outbreaks within a location or preferentially sampling individuals based on their travel history could also bias prevalence estimates. The impact of such factors on prevalence estimates could be substantial especially when the variant has very few sequences and the prevalence of the variant is low; however, this bias decreases as more sequencing data are made available. We communicate these limitations in interpreting our reports through a dedicated 'caveats' page on the website and have warnings linking to this page in relevant sections throughout the interface (Extended Data Fig. 3a,b). Limited sequencing data can also impact the 'characteristic mutations' of a variant, identified using the 75% threshold. Oftentimes, when a new lineage is designated there are very few genomes assigned to it. In such cases, the 75% threshold might not be adequate to identify all the characteristic mutations due to many positions being masked using Ns or otherwise incomplete genomes. We highlight this limitation in the 'methods' page of our website and have a link to this page above the 'mutation map' in the interface (Extended Data Fig. 3c). Further, we also throw a warning regarding this limitation if a lineage has fewer than 1,000 sequences assigned to it (Extended Data Fig. 3d).

Since its launch, outbreak.info (https://outbreak.info/press) has been widely used not only by the scientific community but also as a source of information by a broader audience. On average, the website received over 270,000 page views per month, of which 75% were unique visitors. The average amount of time spent by users on a page was around 1 min. The API received an average of over 3 million unique hits per month, which included any requests from the website itself. The Lineage and/or Mutation Tracker, Location Tracker and the Lineage Comparison Tool accounted for 45%, 14% and 15% of overall traffic, respectively.

outbreak.info continues to provide a mechanism for researchers, epidemiologists and public health officials to easily track the growth of variants, across any number of locations. The platform, backed by robust infrastructure, allows users to quickly access key statistics for known VOCs, emerging variants and any combination of mutations without having to run any time-consuming analyses. While these statistics are useful to monitor the evolution of the virus, there exist several avenues to extend and improve outbreak.info as a platform to comprehensively monitor the evolution of the virus. First among these is to use existing statistics to calculate new quantities such as genetic divergence and the growth rate of emerging variants to aid in the early detection of VOCs and VOIs. The sparse availability of genomes especially for newly detected variants requires methods that can provide reliable confidence intervals in addition to point estimates of such quantities. Given that VOCs and VOIs of SARS-CoV-2 were first detected in many different locations globally including the United Kingdom, Brazil, South Africa and India, it is important to develop or incorporate new methods that can scale across thousands of locations and millions of combinations of lineages and mutations. The second avenue to improve outbreak.info is to leverage our ability to incorporate new types of data that could be analyzed in conjunction with sequencing data sampled in clinical settings. Such data sources include wastewater surveillance, which has emerged as an accurate and cost-efficient method for surveillance and human mobility data that could be used to track the geographic spread of the virus. Finally, we aim to improve our API by adding the ability to construct complex queries. For instance, a query such as 'CALCULATE prevalence FROM lineage:BA.1 NOT mutation:S:K417N AND location:USA' would calculate prevalence of BA.1 genomes without the S:K417N mutation sampled within the United States. This would allow researchers to build queries in an easy and intuitive manner so they can focus on data exploration, hypothesis generation and complex downstream analyses. Beyond the SARS-CoV-2 pandemic, outbreak.info serves as a model for providing scalable and reusable metrics to track the spread of any pathogen during an outbreak via interactive and interpretable visualizations.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01769-3.

References

1. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. Skowronski, D. M. & De Serres, G. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* **384**, 1576–1577 (2021).
3. Holmes, E. Novel 2019 coronavirus genome. *Virological* https://virological.org/t/novel-2019-coronavirus-genome/319 (2020).
4. Khare, S. et al. GISAID's role in pandemic response. *China CDC Wkly.* **3**, 1049–1051 (2021).
5. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
6. Konings, F. et al. SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823 (2021).
7. Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
8. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2021).
9. Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* https://doi.org/10.1126/science.abg3055 (2021).
10. Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
11. Faria, N. R. et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* https://doi.org/10.1126/science.abh2644 (2021).
12. Liu, Y. et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *Cell* **39**, 110829 (2022).
13. Brito, A. F. et al. Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
14. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* https://doi.org/10.1002/gch2.1018 (2017).

15. Faria, N. R. et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).
16. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
17. Grubaugh, N. D. et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401–405 (2017).
18. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
19. Park, D. J. et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).
20. Dudas, G. et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
21. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
22. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.* **2**, e000093 (2016).
23. Hodcroft, E. B. et al. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
24. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
25. Volz, E. et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
26. Washington, N. et al. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184**, 2587–2594 (2021).
27. Badr, H. S. et al. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
28. Nouvellet, P. et al. Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1–9 (2021).
29. Fischer, C. B. et al. Mask adherence and rate of COVID-19 across the United States. *PLoS ONE* **16**, e0249891 (2021).
30. Wade, A. et al. Variation in government responses to COVID-19. *Blavatnik School of Government* https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19 (2022).
31. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).
32. Tsueng, G. et al. Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. *Nat. Methods* https://doi.org/10.1038/s41592-023-01770-w (2023).
33. O'Donoghue, S. I. et al. Aquaria: simplifying discovery and insight from protein structures. *Nat. Methods* **12**, 98–99 (2015).
34. Faria, N. R. et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological* https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (2021).
35. Laiton-Donato, K. et al. Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2. *Infect. Genet. Evol.* **95**, 105038 (2021).
36. Lemey, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).
37. O'Toole, Á. et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Res.* **6**, 121 (2021).
38. Tracking of hCov-19 variants. *GISAID* https://www.gisaid.org/hcov19-variants/ (2023).
39. Hoffmann, M. et al. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell* **184**, 2384 (2021).
40. Buss, L. F. et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* **371**, 288–292 (2021).
41. Unwin, H. J. T. et al. State-level tracking of COVID-19 in the United States. *Nat. Commun.* **11**, 6189 (2020).
42. *SPI-M-O: Consensus Statement on COVID-19* (Assets Publishing Service UK, 2021); https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993321/S1267_SPI-M-O_Consensus_Statement.pdf
43. Dhar, M. S. et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* https://doi.org/10.1126/science.abj9932 (2021).
44. Mlcochova, P. et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).
45. Mathieu, E. et al. Coronavirus pandemic (COVID-19). *Our World in Data* https://ourworldindata.org/coronavirus (2020).
46. Lineage list. *Cov-Lineages* https://cov-lineages.org/lineage_list.html (2023).
47. Bushman, M., Kahn, R., Taylor, B. P., Lipsitch, M. & Hanage, W. P. Population impact of SARS-CoV-2 variants with enhanced transmissibility and/or partial immune escape. *Cell* **184**, 6229–6242 (2021).
48. Ferguson, N. et al. *Report 49: Growth, Population Distribution and Immune Escape of Omicron in England* (Imperial College COVID-19 Response Team, 2021); https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2021-12-16-COVID19-Report-49.pdf
49. UK Health Security Agency. *Technical Briefing 31: SARS-CoV-2 Variants of Concern and Variants Under Investigation in England* (Assets Publishing Service UK, 2021); https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1042367/technical_briefing-31-10-december-2021.pdf
50. Khan, K. et al. Omicron BA.4/BA.5 escape neutralizing immunity elicited by BA.1 infection. *Nat. Commun.* **13**, 4686 (2022).
51. Cao, Y. et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608**, 593–602 (2022).
52. Khan, K. et al. Omicron infection enhances Delta antibody immunity in vaccinated persons. *Nature* **607**, 356–359 (2022).
53. Hill, V. et al. The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol.* **8**, veac080 (2022).
54. Tzou, P. L., Tao, K., Kosakovsky Pond, S. L. & Shafer, R. W. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. *PLoS ONE* **17**, e0261045 (2022).
55. Wright, D. W. et al. Tracking SARS-CoV-2 mutations and variants through the COG-UK-mutation explorer. *Virus Evol.* **8**, veac023 (2022).
56. Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A. & Deverman, B. E. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife* **10**, e63409 (2021).
57. Chen, C. et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2021).
58. SARS-CoV-2 variants and lineages of concern. *BV-BRC* https://www.bv-brc.org/view/VariantLineage/#view_tab=overview (2023).
59. National Wastewater Surveillance System (NWSS). *Centers for Disease Control and Prevention* https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater-surveillance.html (2022).

60. Xie, X. et al. Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nat. Med.* **27**, 620–621 (2021).
61. Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, P1295-1310.E20 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Ingestion of genomic data

We built a data pipeline, `Bjorn`, to preprocess metadata and count mutations from a given set of genomes in a scalable manner on a daily basis (Extended Data Fig. 4). The pipeline consists of the following steps: (1) download SARS-CoV-2 genomes from the GISAID provision; (2) filter out records based on erroneous or incomplete genomes and associated metadata; (3) compare with cache to only process new and updated records; (4) divide sequences into chunks of 10,000 records (each chunk is subsequently processed in parallel using GNU parallel)[62]; (5) align these sequences using minimap2 v.2.24 (ref. [63]); (6) convert the alignment into a FASTA file using gofasta v.1.1.0 (ref. [64]); (7) count mutations and deletions from this alignment; (8) standardize and filter the metadata: country, division, location, Pango lineage, date of collection and date of submission; and (9) combine results from all chunks and convert to a JSON Lines object. We standardized geographic identifiers using shapefiles from GADM (https://gadm.org/). The final JSON Lines object was loaded into an Elasticsearch v.1.7 index within the BioThings framework v.0.11.1 (ref. [65]). The source code for `Bjorn` including a `Docker` container for easy deployment is available at https://github.com/andersen-lab/bjorn.

To exclude records with incomplete or erroneous genomes and associated metadata in step (2), we use the following conditions:

- Length of the genome is less than 20,000 nucleotides or greater than 40,000 nucleotides.
- Greater than one-eighth of nucleotides in the genome are masked using Ns.
- The genome differs from the reference genome at more than 50% of nucleotides.
- The genome has more than 500 insertions or deletions relative to the reference.
- The collection date is after the submission date.
- The submission date is in the future (beyond the current date).
- The collection date is before November 2019.
- The host is not human.
- The location string cannot be matched to any country in the GADM shapefiles.

Ingestion of epidemiological data

We built the EpiData pipeline to ingest reported global cases and deaths from Johns Hopkins University[31]. We used shapefiles from Natural Earth (https://www.naturalearthdata.com/) to standardize geographic identifiers and obtain populations for countries and states outside the United States. For the United States, we used the county-level shapefiles and population estimates from the 2019 population estimates by the Census Bureau to standardize geographic identifiers and get population estimates. We standardized reported date formats and geographic identifiers across the two resources. The code for the EpiData pipeline is available at https://github.com/outbreak-info/biothings_covid19.

Calculation of confidence intervals on prevalence

Most estimates of prevalence on outbreak.info are binomial proportions. We calculated 95% confidence intervals for these estimates using Jeffrey's interval, the 2.5 and 97.5 quantiles of the β distribution $\beta(x+0.5, n-x+0.5)$ where x is the number of successes and n is the number of trials.

Creation of outbreak.info API

To scale with the increasing size of existing data sources and the heterogeneity of newly emerging data sources, we used the BioThings framework[65]. The JSON outputs of our data pipelines are ingested by the BioThings framework and the processed data are stored in individual Elasticsearch indices. A Tornado server is used to create API endpoints that leverage the search capabilities of Elasticsearch to perform complex aggregations of the underlying data. These API endpoints allow the client-side applications to query the underlying data within reasonable query times while accounting for the scale of the ingested data. The BioThings Hub maintains historical data by default, allowing us to roll back to previous data backups if issues are discovered with new data after they are deployed. The code for the server-side application is available at https://github.com/outbreak-info/outbreak.api.

outbreak.info web application

The web application was built using Vue.js v.2.7.14 (https://vuejs.org/), a model–view–view model JavaScript framework that enables the two-way binding of user interface elements and the underlying data allowing the user interface to reflect any changes in underlying data and vice versa. The client-side application uses the high-performance API to interactively perform operations on the database. Customized data visualizations on the client were built using D3.js v.5.16.0 (ref. [66]), giving us the ability to develop novel and intuitive visual elements as part of the user interface. We designed these visualizations to answer specific questions of interest to epidemiologists, researchers and public health officials. We further added functionality to enable the one-click copy or download of every chart in the interface as a PNG or SVG. The code for the client-side application is available at https://github.com/outbreak-info/outbreak.info.

R package

We developed an R package for outbreak.info to allow researchers and other individuals to easily access the data via the API for downstream analyses and visualizations. The R package is composed of three parts: functions that allow the user to access genomic data, functions to access the epidemiological data and functions to access the Research Library metadata. They all consist of a base function that contains arguments for all possible parameters that can be used to query the API. While users can utilize this base function directly to access data, several wrapper functions are available that inherit the arguments from the base function in addition to prespecified parameters to simplify the process of querying the API. For example, while getGenomicData() can be used directly to access data regarding the daily global prevalence of a specified lineage, doing so would require a user to be familiar with the name of the end point as specified in the API URL (in this case, global prevalence). Instead, the user can access these data with the more intuitively named getPrevalence(). Therefore, these wrapper functions allow users to easily and quickly obtain the data they need. The R package also contains an authenticateUser() function that allows users to authenticate against their GISAID credentials and access computed statistics from the primary genomic data provided by GISAID.

In addition, as the API queries location by ISO3 code, rather than by location name, two functions have been created that allow users to forgo the step of searching for the ISO3 code themselves: getISO3Code() and getLocationIdGenomic(). The latter function uses the genomics API end point to obtain the ISO3 code for a given location. The ISO3 code can be obtained with either a full or incomplete location name; in the latter case, the user will be provided a list of matching locations and must specify the location they are interested in. This function is embedded in the parent getGenomicData() function and is therefore inherited in all wrapper functions. Therefore, searching for data by location in the R package replicates the experience on the client-side web application. Documentation is available at https://outbreak-info.github.io/R-outbreak-info with vignettes located at https://outbreak-info.github.io/R-outbreak-info/articles/index.html. The R package can be downloaded and installed using the remotes package function: install_github('outbreak-info/R-outbreak-info').

Usability studies

Two remote moderated usability tests of outbreak.info were conducted between 4 March 2021 and 2 April 2021 over Zoom. These tests explored the ease by which researchers could use outbreak.info's genomic tools to find information about SARS-CoV-2 lineages and mutations, including prevalence data, mutation characteristics and publications. In the first test, five biomedical researchers were recruited to complete a series of seven tasks using the Lineage and/or Mutation Tracker:

1. Please export a visualization of the average daily prevalence of the B.1.1.7 lineage in the United Kingdom.
2. Please use outbreak.info to find publications about the B.1.1.7 lineage.
3. Please tell me which lineages are VOCs and how to distinguish the difference between VOCs and VOIs using the website.
4. Please use outbreak.info to find which lineages contain the E484K mutation.
5. Please use outbreak.info to find the total number of E484K sequences that have been detected in New York state.
6. Please use outbreak.info to find the global prevalence of the P.2 lineage.
7. Please use outbreak.info to find the global prevalence of the S:D614G mutation.

In the second test, five biomedical researchers were recruited to complete a series of five tasks using the location reports:

1. Please find the prevalence of lineages in California. Subsequently, please find the prevalence of specific mutations (for example L452R) for the most common lineages in California.
2. Please compare the prevalence of these lineages/mutations in California over time: B.1.1.7, B.1.427, B.1.429 and S:L452R.
3. Please find cumulative B.1.429 prevalence by county in California.
4. Please find the global prevalence of the B.1.526 lineage with the S:E484K mutation.
5. Please find all publications related to B.1.1.7.

For both tests, five post-test questions were asked:

1. Overall, please rate how easy or difficult it is to use this app on a scale of 1–5, where 1 is very difficult and 5 is very easy.
2. How likely are you to continue using the Mutation Situation Reports to regularly access data or find information about the variants, on a scale of 1–5 where 1 is very unlikely and 5 is very likely?
3. What do you like most about the Mutation Situation Reports?
4. What would you improve about the app? Or what would you add to the app?
5. How would you compare this site to other sites you've used to find information about the variants?

The following metrics were recorded:

1. Ease and satisfaction about each task (five-point Likert scale).
2. Time on each task.
3. Number of successful task completions and errors.
4. Overall ease and satisfaction (five-point Likert scale).
5. Likelihood to use (five-point Likert scale).
6. Suggestions for improvement (likes, dislikes and recommendations).
7. Error-free rate.
8. Noted observations about the users' process.

The successful completion rate for both tests was 100% but users made non-critical errors and the error-free rate for the first and second tests was 10% and 20%, respectively. The average amount of time spent on tasks was a little over 1 min. Users rated the overall ease of using the interface with an average score of 80% and all users responded that they were very likely to continue using the site. Users specifically liked the site's interactivity, the visualizations and the wide range of features that could be found in one place.

Based on these usability tests, the following changes to the interface were made:

1. New descriptions of report categories were added, as well as the option to filter VOCs and VOIs.
2. The user interface to select mutations and lineage + mutation combinations was streamlined.
3. A second search bar for finding location reports was added to the home page.
4. New options were added to the navigation bar at the top of the interface.
5. An FAQs page was created.
6. A new site description and introductory video were added.
7. Several other design changes to the home page to highlight major features.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All SARS-CoV-2 virus sequence data were provided by the GISAID Global Data Science Initiative and are available at https://gisaid.org/. Additionally, the summarized variant prevalence data can be accessed through an R package (https://outbreak-info.github.io/R-outbreak-info/) using freely available GISAID credentials.

Code availability

All code used to generate the outbreak.info genomic reports is freely available on GitHub (https://github.com/outbreak-info) under open source licenses. This includes:

• outbreak.info web application: the code powering the outbreak.info front-end (https://github.com/outbreak-info/outbreak.info). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343497.

• outbreak.info R package: R package to access all the genomics and epidemiology data and Research Library metadata compiled and standardized on outbreak.info (https://github.com/outbreak-info/R-outbreak-info). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343501.

• Bjorn: pipeline for SARS-CoV-2 mutation counting, lineage classification and other processing (https://github.com/andersen-lab/bjorn). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343950.

• api.outbreak.info: the code to create the API to access Research Library metadata and cases and deaths data is available at api.outbreak.info (https://github.com/outbreak-info/outbreak.api). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343503.

References

62. Tange, O. GNU Parallel 20210622 ('Protasevich'). *Zenodo* https://doi.org/10.5281/zenodo.5013933 (2021).
63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
64. virus-evolution/gofasta. *GitHub* https://github.com/virus-evolution/gofasta (2023).
65. Lelong, S. et al. BioThings SDK: a toolkit for building high-performance data APIs in biomedical research. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac017 (2022).
66. Bostock, M. Data-driven documents. *D3.js* https://d3js.org/ (2021).

Acknowledgements

We thank the technical team, sequence curators and administrators of the GISAID database for helping us with this project and supporting rapid and transparent sharing of genomic data during the COVID-19 pandemic. We thank our colleagues for sharing genomic data on GISAID. This work was supported by the National Institute for Allergy and Infectious Diseases (5 U19 AI135995 (K.G., A.A.L., J.L.M., M.A., E. Hufbauer, G.T., E. Haag, M.Z., C.M.A., K.Z., M.C., R.S., N.L.M., J.I.L., M.A.S., C.W., A.I.S. and L.D.H.), 3 U19 AI135995-04S3 (K.G., J.L.M., E. Haag, G.T., E.H., M.C., C.W., A.I.S. and L.D.H.), 3 U19 AI135995-03S2 (K.G., J.L.M., E. Hufbauer, G.T., E. Haag, M.Z., C.M.A., K.Z., M.C., R.S., N.L.M., J.I.L., M.A.S., C.W., A.I.S. and L.D.H.), U01AI151812 (K.G., A.A.L., M.A., E. Hufbauer, M.Z., C.M.A., N.L.M., J.I.L., M.A.S., A.I.S., K.G.A. and L.D.H.), R01 AI162611 (K.G. and M.A.S.), R01 AI153044 (K.G. and M.A.S.)), National Center For Advancing Translational Sciences (5 U24 TR002306 (G.T., M.C., X.Z., Z.Q., C.W., A.I.S. and L.D.H.), UL1TR002550 (G.T., M.C., X.Z., Z.Q., C.W. and A.I.S.)), CDC (75D30120C09795 (K.G., A.A.L., M.A., E. Hufbauer, M.Z., C.M.A., N.L.M., J.I.L. and K.G.A.) and the National Institute of General Medical Sciences (R01GM083924 (G.T., M.C., X.Z., Z.Q., C.W. and A.I.S.)).

Author contributions

K.G., A.A.L., J.L.M., M.A., E.H., G.T., E.H., M.Z., C.M.A., K.Z., M.C., X.Z., Z.Q., R.S., N.L.M., J.I.L., M.A.S., C.W., A.I.S., K.G.A. and L.D.H. contributed to the design, construction and/or maintenance of the outbreak.info website and data pipelines. K.G., E.H., R.T.C.L., L.F. and S.M.S. designed the connection between the GISAID data stream and the Bjorn data pipeline. K.G., M.A. and L.D.H. designed and built the R outbreak.info package. K.G., A.A.L., J.L.M., M.A., E.H., G.T., E.H., M.Z., K.G.A. and L.D.H. contributed to writing and editing the manuscript.

Competing interests

M.A.S. receives grants from the US National Institutes of Health within the scope of this work and grants and contracts from the US Food and Drug Administration, the US Department of Veterans Affairs and Janssen Research & Development outside the scope of this work. M.A.S. and K.G.A. have received consulting fees and/or compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic. The other authors declare no competing interests.

Additional information

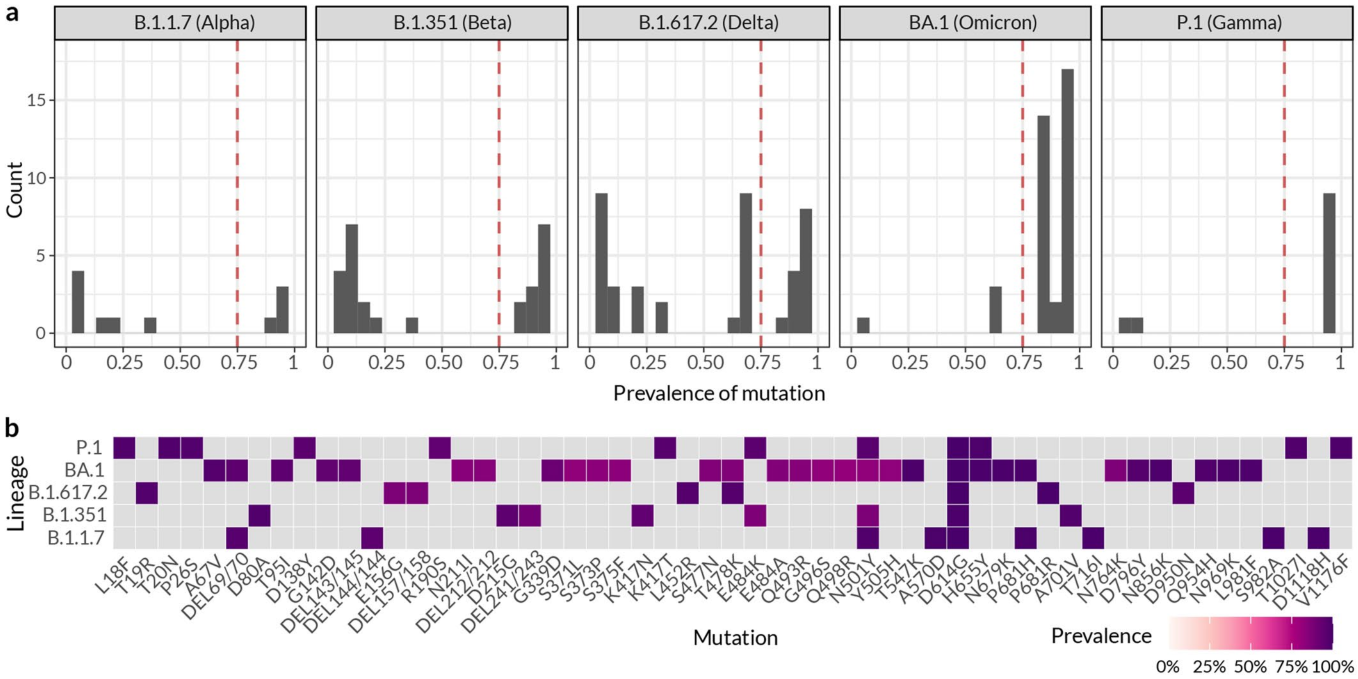
Extended data is available for this paper at https://doi.org/10.1038/s41592-023-01769-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01769-3.

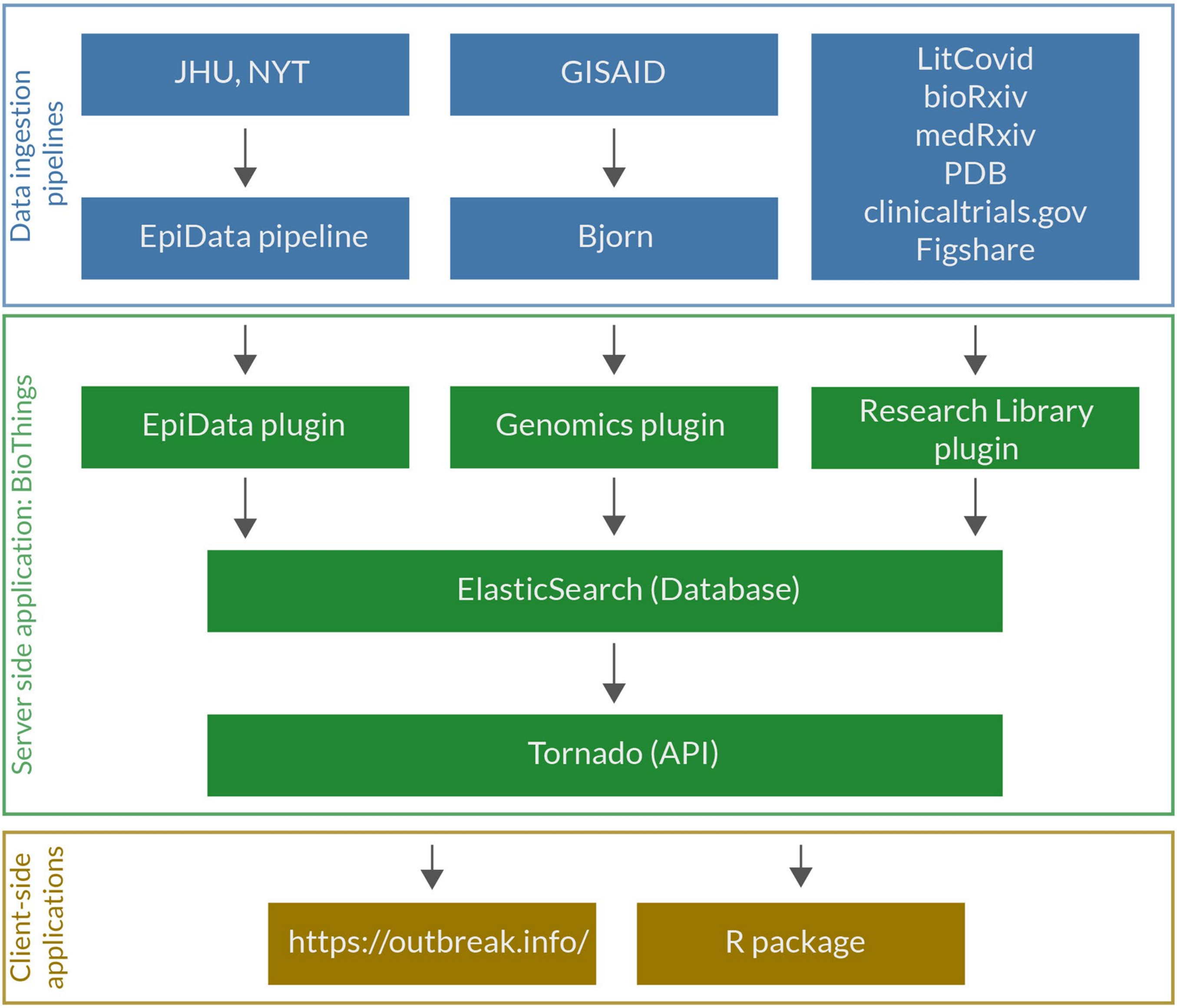
Correspondence and requests for materials should be addressed to Karthik Gangavarapu or Laura D. Hughes.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

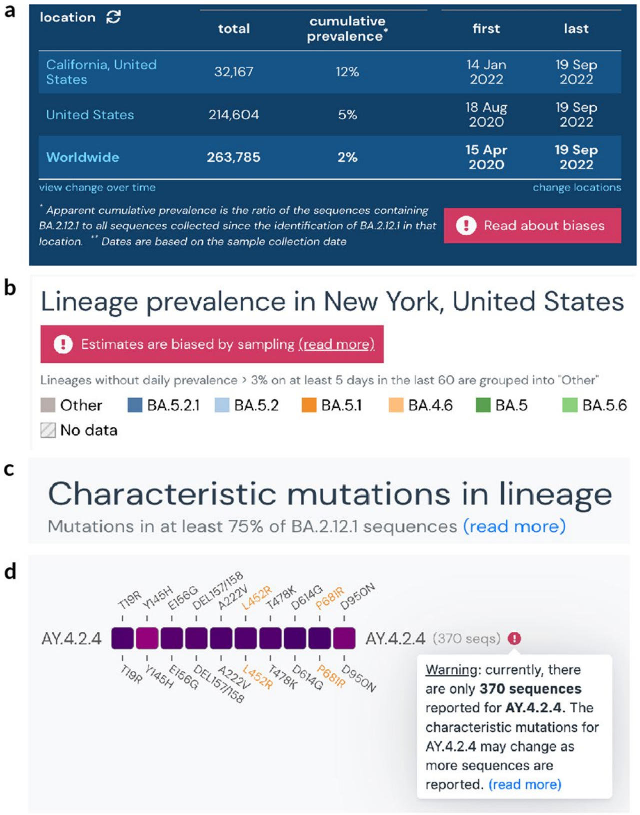
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Empirical basis for selecting 75% as a threshold to identify 'characteristic mutations' of a lineage. (a) The frequency of mutations above 5% prevalence in P.1 (Gamma), BA.1 (Omicron), B.1.617.2 (Delta), B.1.351 (Beta), and B.1.1.7 (Alpha) variants. **(b)** Mutations present in >= 75% of all sequences in P.1 (Gamma), BA.1 (Omicron), B.1.617.2 (Delta), B.1.351 (Beta), and B.1.1.7 (Alpha) variants.



Extended Data Fig. 2 | Software infrastructure of outbreak.info. The infrastructure can be broadly divided into (1) Data ingestion pipelines, (2) Server-side hosting the database and API server, and (3) Client-side applications that use the API from the server.



location	total	cumulative prevalence*	first	last
California, United States	32,167	12%	14 Jan 2022	19 Sep 2022
United States	214,604	5%	18 Aug 2020	19 Sep 2022
Worldwide	263,785	2%	15 Apr 2020	19 Sep 2022

view change over time

change locations

* Apparent cumulative prevalence is the ratio of the sequences containing BA.2.12.1 to all sequences collected since the identification of BA.2.12.1 in that location. ** Dates are based on the sample collection date

Read about biases

Lineage prevalence in New York, United States

Estimates are biased by sampling (read more)

Lineages without daily prevalence > 3% on at least 5 days in the last 60 are grouped into "Other"

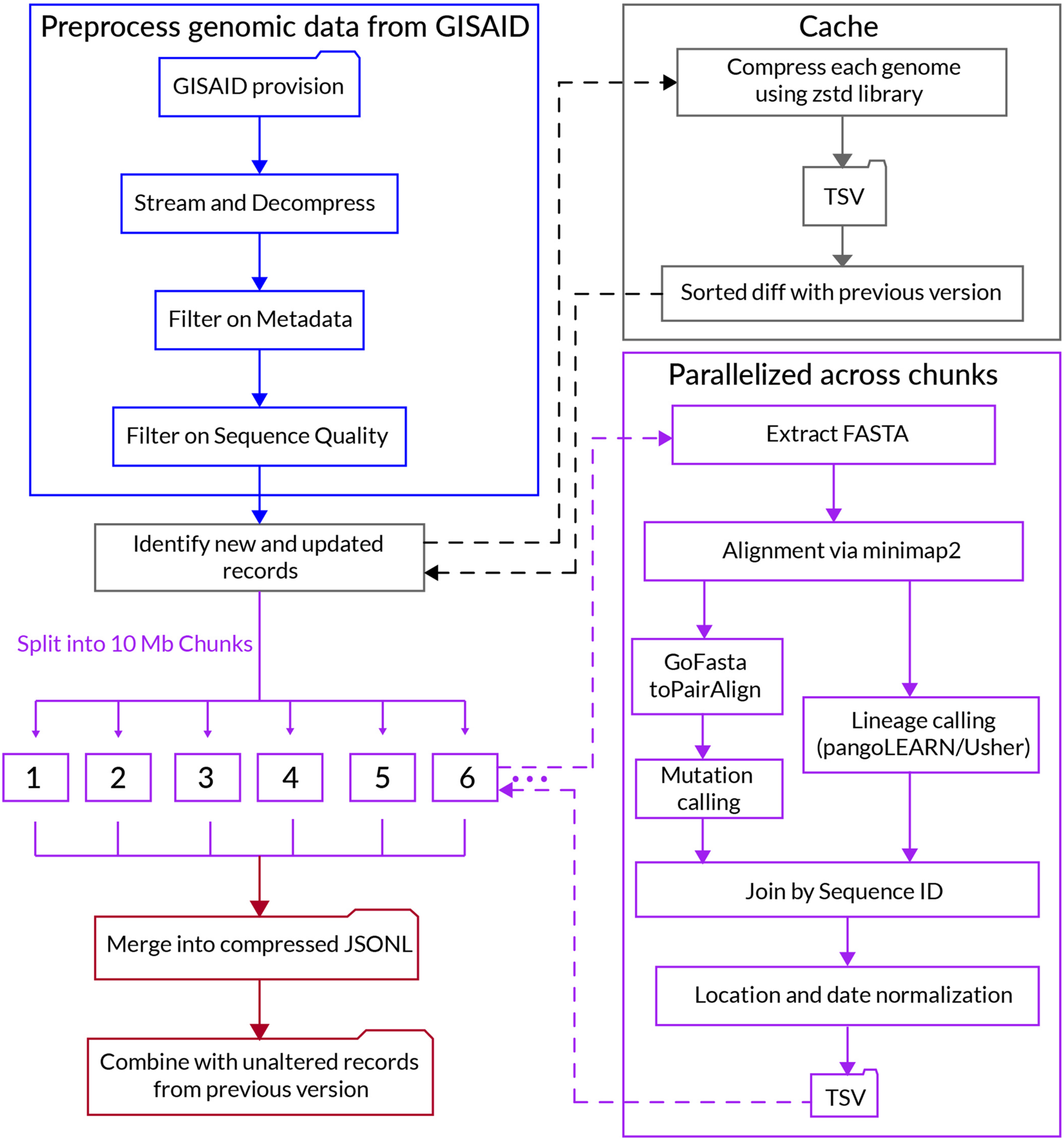
Other BA.5.2.1 BA.5.2 BA.5.1 BA.4.6 BA.5 BA.5.6 No data

Characteristic mutations in lineage

Mutations in at least 75% of BA.2.12.1 sequences (read more)

Warning: currently, there are only **370 sequences** reported for **AY.4.2.4.** The characteristic mutations for AY.4.2.4 may change as more sequences are reported. (read more)

Extended Data Fig. 3 | Examples of warnings to ensure users pay attention to possible biases while interpreting visualizations on the web interface. (**a**) Link ('Read about biases') to the caveats page on the web interface in the summary box section of the lineage/mutation tracker. (**b**) Link ('Estimates are biased by sampling (read more)') to the caveats page above the streamgraph on the web interface of the location tracker. (**c**) Link ('read more') to the methods page on the web interface about how characteristic mutations are identified and associated limitations. (**d**) Warning about the limitations of identifying characteristic mutations when less than 1000 sequences are assigned to a lineage.


Preprocess genomic data from GISAID
GISAID provision
Stream and Decompress
Filter on Metadata
Filter on Sequence Quality
Cache
Compress each genome using zstd library
TSV
Sorted diff with previous version
Identify new and updated records
Split into 10 Mb Chunks
1
2
3
4
5
6
Merge into compressed JSONL
Combine with unaltered records from previous version
Parallelized across chunks
Extract FASTA
Alignment via minimap2
GoFasta toPairAlign
Mutation calling
Lineage calling (pangoLEARN/Usher)
Join by Sequence ID
Location and date normalization
TSV

Extended Data Fig. 4 | Flowchart describing the steps in Bjorn. The genomic data and associated metadata from GISAID undergo preprocessing and filtering to exclude erroneous or incomplete records (depicted in blue). The preprocessed information is then compressed and compared to prior cached versions to determine new or updated records (depicted in black). These new genomes are aligned and mutations are counted, followed by lineage identification. The locations and dates in the new metadata are also normalized to enable standardized query access. These processing steps are executed in parallel by splitting the data into chunks of 10 Mb (depicted in purple). The processed data from the new records are combined with the processed data from the unaltered records (depicted in brown), following which they are stored in an Elasticsearch database.

nature portfolio

Corresponding author(s): Karthik Gangavarapu, Laura D. Hughes

Last updated by author(s): Nov 21, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All SARS-CoV-2 virus sequence data was provided by the GISAID Global Data Science Initiative and is available at https://gisaid.org/. Additionally, the summarized variant prevalence data can be accessed through an R package (https://outbreak-info.github.io/R-outbreak-info/) using freely available GISAID credentials.

Data analysis

All code used to generate the outbreak.info Genomic Reports is freely available on GitHub (https://github.com/outbreak-info) under open source licenses. This includes:

● outbreak.info web application: the code powering the outbreak.info front-end (https://github.com/outbreak-info/outbreak.info). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343497.

● outbreak.info R package: R package to access all the genomics and epidemiology data and Research Library metadata compiled and standardized on outbreak.info (https://github.com/outbreak-info/R-outbreak-info). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343501.

● Bjorn: pipeline for SARS-CoV-2 mutation counting, lineage classification, and other processing (https://github.com/andersen-lab/bjorn). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343950.

● api.outbreak.info: Code to create the application programming interface (API) to access Research Library metadata and cases & deaths data, available at api.outbreak.info (https://github.com/outbreak-info/outbreak.api). The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343503.

Bjorn further uses minimap2 v2.24 to align SARS-CoV-2 genomes and gofasta v1.1.0 to count mutations from the resulting alignment. Results from the pipeline are subsequently stored in an ElasticSearch v1.7 index within the BioThings framework v0.11.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All SARS-CoV-2 virus sequence data was provided by the GISAID Global Data Science Initiative and is available at https://gisaid.org/. Additionally, the summarized variant prevalence data can be accessed through an R package (https://outbreak-info.github.io/R-outbreak-info/) using freely available GISAID credentials.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender

Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We use all the SARS-CoV-2 genomes available on GISAID for the estimates shown on outbreak.info. We apply filters on the complete dataset to exclude records with incomplete or erroneous genomes or metadata. However, the genomes available on GISAID have some sampling issues, particularly when a variant has recently emerged, which we highlight within the manuscript. In such cases, we rely on our uncertainty estimates to clearly convey the bias that could be introduced due to sampling issues. Further, we highlight potential sampling issues throughout the interface to ensure that users take this account while interpreting the visualizations.

Data exclusions

No data were excluded

Replication

We have made all aspects of our code open source so users can deploy stand alone versions of our data pipelines, server, and the web interface. Further, we made all of the data accessible via an R package so users can perform their own analyses to either verify or replicate the results shown on the web interface.

Randomization

We use all the available data from GISAID and not perform any statistical comparisons between any groups in the data. Hence, we do not incorporate randomization into our estimates.

Blinding

Since these analyses were based on publicly available data provided by GISAID, the authors were blind to any properties about the sample, aside from its collection date and geographic origin. Samples were grouped into Pango lineages based on phylogenetic analysis of their sequences, as described in the manuscript.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging