

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

3 Karthik Gangavarapu<sup>1,\*</sup>, Alaa Abdel Latif<sup>2</sup>, Julia L. Mullen<sup>3</sup>, Manar Alkuzweny<sup>4</sup>, Emory Hufbauer<sup>2</sup>,  
4 Ginger Tsueng<sup>3</sup>, Emily Haag<sup>3</sup>, Mark Zeller<sup>2</sup>, Christine M. Aceves<sup>2</sup>, Karina Zaiets<sup>3</sup>, Marco Cano<sup>3</sup>, Jerry  
5 Zhou<sup>3</sup>, Zhongchao Qian<sup>3</sup>, Rachel Sattler<sup>5</sup>, Nathaniel L Matteson<sup>2</sup>, Joshua I. Levy<sup>2</sup>, Raphael TC Lee<sup>6,7</sup>,  
6 Lucas Freitas<sup>6,8</sup>, Sebastian Maurer-Stroh<sup>6,7,9,10</sup>, GISAID core and curation team<sup>#</sup>, Marc A.  
7 Suchard<sup>1,11,12</sup>, Chunlei Wu<sup>3,13,14</sup>, Andrew I. Su<sup>3,13,14</sup>, Kristian G. Andersen<sup>2,13</sup>, Laura D. Hughes<sup>3,\*</sup>

8  
9 <sup>1</sup>Department of Human Genetics, David Geffen School of Medicine, University of California Los  
10 Angeles, Los Angeles, CA 90095, USA

11 <sup>2</sup>Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037,  
12 USA

13 <sup>3</sup>Department of Integrative, Structural and Computational Biology, The Scripps Research Institute,  
14 La Jolla, CA 92037, USA

15 <sup>4</sup>Department of Biological Sciences, University of Notre Dame, IN 46556, USA

16 <sup>5</sup>Skaggs Graduate School of Biological and Chemical Sciences, The Scripps Research Institute, La  
17 Jolla, CA 92037, USA

18 <sup>6</sup>GISAID Global Data Science Initiative (GISAID), Munich, Germany

19 <sup>7</sup>Bioinformatics Institute & ID Labs, Agency for Science Technology and Research, Singapore

20 <sup>8</sup>Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

21 <sup>9</sup>National Centre for Infectious Diseases, Ministry of Health, Singapore

22 <sup>10</sup>Department of Biological Sciences, National University of Singapore, Singapore

23 <sup>11</sup>Department of Biomathematics, David Geffen School of Medicine, University of California  
24 Los Angeles, Los Angeles, CA 90095, USA

25 <sup>12</sup>Department of Biostatistics, Fielding School of Public Health, University of California Los  
26 Angeles, Los Angeles, CA 90095, USA

27 <sup>13</sup>Scripps Research Translational Institute, La Jolla, CA 92037, USA

28 <sup>14</sup> Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA

29 <sup>#</sup>Author list in **Supplementary File 1**

30 <sup>\*</sup>Corresponding authors: [gkarthik@ucla.edu](mailto:gkarthik@ucla.edu), [lhughes@scripps.edu](mailto:lhughes@scripps.edu)

31

## 32 **Abstract**

33 The emergence of SARS-CoV-2 variants of concern has prompted the need for near real-time  
34 genomic surveillance to inform public health interventions. In response to this need, the global  
35 scientific community, through unprecedented effort, has sequenced and shared over 10 million  
36 genomes through GISAID, as of May 2022. This extraordinarily high sampling rate provides a unique  
37 opportunity to track the evolution of the virus in near real-time. Here, we present *outbreak.info*,  
38 a platform that currently tracks over 40 million combinations of PANGO lineages and individual  
39 mutations, across over 7,000 locations, to provide insights for researchers, public health officials,  
40 and the general public. We describe the interpretable and opinionated visualizations in the variant  
41 and location focussed reports available in our web application, the pipelines that enable the scalable  
42 ingestion of heterogeneous sources of SARS-CoV-2 variant data, and the server infrastructure that  
43 enables widespread data dissemination via a high performance API that can be accessed using an  
44 R package. We present a case study that illustrates how *outbreak.info* can be used for genomic  
45 surveillance and as a hypothesis generation tool to understand the ongoing pandemic at varying

46 geographic and temporal scales. With an emphasis on scalability, interactivity, interpretability, and  
47 reusability, `outbreak.info` provides a template to enable genomic surveillance at a global and  
48 localized scale.

49

## 50 **Introduction**

51 In December 2019, a series of cases of pneumonia of unknown origin appeared in Wuhan, China,  
52 and on 7 January 2020, the virus responsible for the diseases was identified as a novel coronavirus,  
53 SARS-CoV-2<sup>1</sup>. The first SARS-CoV-2 genome was made publicly available on 10 January 2020<sup>2</sup>. Since  
54 then, the global scientific community, through an unprecedented effort, has sequenced and shared  
55 over 10 million genomes through GISAID, as of May 2022<sup>3,4</sup>. To keep track of the evolving genetic  
56 diversity of SARS-CoV-2, Rambaut *et al.* developed a dynamic phylogeny-informed nomenclature  
57 (PANGO) to classify SARS-CoV-2 lineages<sup>5</sup>. As of May 2022, over 2,000 lineages have been  
58 designated, which has enabled public health agencies such as Public Health England (PHE), the  
59 Centers for Disease Control (CDC), and the World Health Organization (WHO) to identify Variants of  
60 Concern (VOC), Variants of Interest (VOI/VUI), and Variants Under Monitoring (VUM/VBM) based on  
61 the phenotypical characterization of these lineages<sup>6</sup>. Currently, there are five designated VOCs:  
62 B.1.1.7\* (Alpha; \* denotes the lineage and any of its sub lineages) lineage resulting in increased  
63 transmissibility<sup>7</sup>, B.1.351\* (Beta) lineage exhibiting immune evasion<sup>8</sup>, the P.1\* (Gamma) lineage  
64 exhibiting immune evasion<sup>9</sup>, the B.1.617.2\* lineage exhibiting increased transmissibility due to the  
65 P681R mutation in the Spike gene<sup>10</sup>, and more recently, the B.1.1.529\* (Omicron) lineage exhibiting  
66 very rapid growth and the ability to substantially avoid antibody neutralization<sup>11,12</sup>.

67

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

85

86 Here, we present `outbreak.info`, a platform that currently tracks over 40 million combinations of  
87 PANGO lineages and individual mutations, across over 7,000 locations, to provide insights for  
88 researchers, public health officials, and the general public. In the following sections, we describe the  
89 data pipelines that enable the scalable ingestion and standardization of heterogeneous data on

90 SARS-CoV-2 variants, the server infrastructure that enables the dissemination of the processed data,  
91 and the client-side applications that provide intuitive visualizations of the underlying data.

92

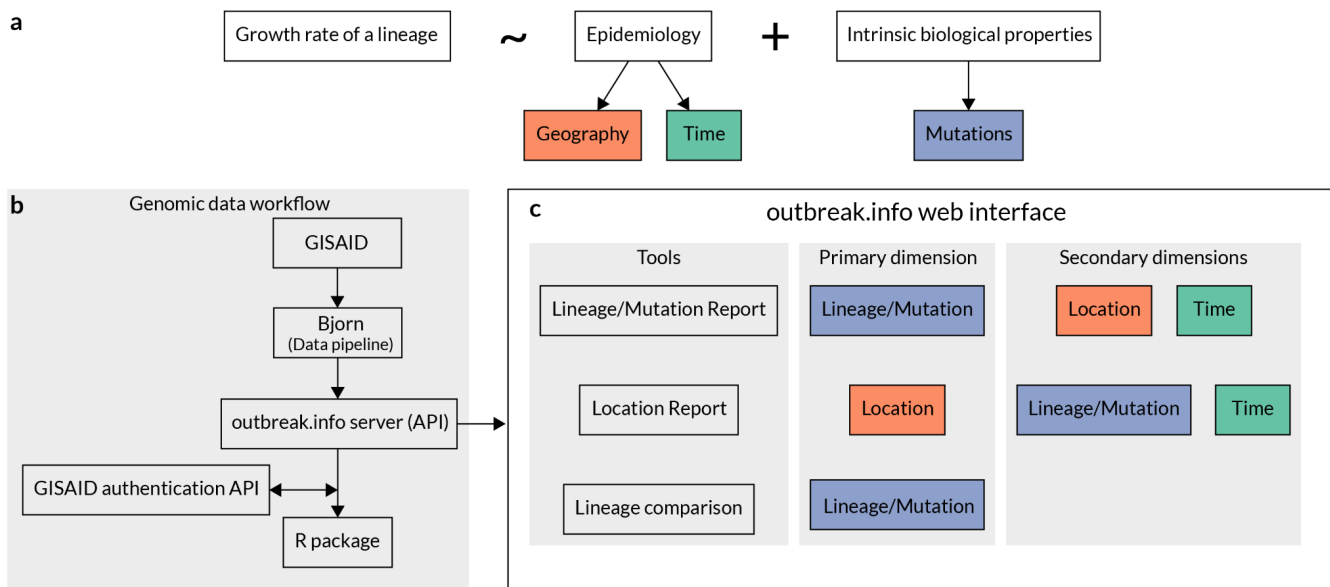
## 93 **Results**

94 The growth rate of a given viral lineage is a function of epidemiology and its intrinsic biological  
95 properties (**Fig 1a**). For example, the B.1.177 lineage, characterized by an A222V amino acid  
96 substitution in the spike gene, increased in prevalence in Europe during the summer of 2020<sup>24</sup>.  
97 While initially thought to be more transmissible, it was eventually shown that the increase in  
98 prevalence was due to a resurgence in travel and not due to increased transmissibility. In contrast,  
99 a few months later, the B.1.1.7 lineage was shown to be 40-60% more transmissible than previously  
100 circulating lineages and this intrinsic biological property led to the rapid growth in its prevalence  
101 worldwide<sup>25,26</sup>. Epidemiological factors such as mobility<sup>27,28</sup>, mask usage<sup>29</sup>, and public health  
102 interventions<sup>30</sup> vary over time and across geographies worldwide, while biological properties are a  
103 function of the mutations found in a given lineage (**Fig 1a**). Hence, to maximize the utility of genomic  
104 data for surveillance, we built `outbreak.info` to enable the exploration of genomic data across  
105 three dimensions: geography, time, and lineages/mutations. We use the PANGO nomenclature to  
106 estimate the prevalence of SARS-CoV-2 lineages over time and at varying geographic scales. Using  
107 a phylogenetically-informed nomenclature allows us to determine genetic features such as the  
108 “characteristic mutations” of a lineage without directly building a global phylogeny. By avoiding a  
109 global phylogeny, we can update our databases daily using the continuously growing number of  
110 SARS-CoV-2 genomes. In addition, we closely track reports from health agencies such as the PHE,  
111 the CDC and the WHO that designate VOC/VOI/VUMs based on epidemiological analyses. In addition  
112 to genomic data, the server also ingests two other types of data: (1) epidemiological data curated  
113 by Johns Hopkins University<sup>31</sup>, and (2) public literature, clinical trial, protocol, and dataset metadata  
114 from sources such as bioRxiv, medRxiv, and LitCovid<sup>32</sup>. Here, we describe how each of these data  
115 sources is used in cohesion to assist in genomic surveillance.

116

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

130



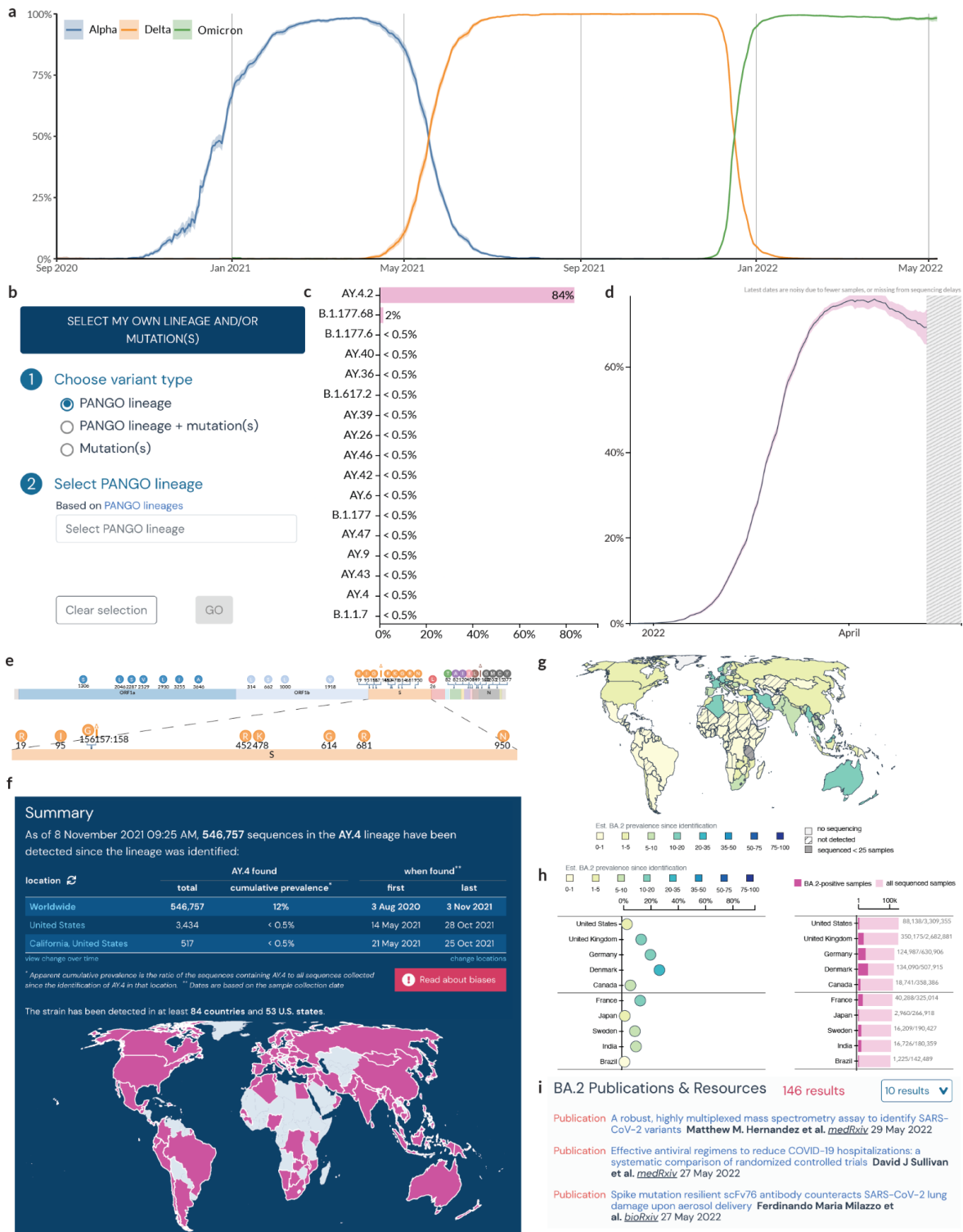
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141

**Figure 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 while intrinsic biological properties are determined by the mutations present in a given lineage. **b**, Genomic data is ingested from GISAID, processed using the custom-built data pipeline, Bjorn, and stored on a server which can be accessed via an Application Programming Interface (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.

### 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. **Fig 1a** 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 (i) VOC/VOI and its sublineages (e.g., Delta and its sublineages), (ii) a lineage (e.g., B.1.1.7), (iii) a lineage and one or more mutations (e.g., B.1.1.7 with S:E484K), (iv) a mutation (e.g., S:E484K), or (v) a group of mutations (e.g., S:E484K and S:N501Y) (**Fig 1b**). In addition, 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 opinionated interface to address a specific set of questions listed in Table 1.

157



158  
159  
160  
161  
162  
163  
164  
165  
166

**Figure 2. Lineage and/or Mutation Tracker.** **a**, Prevalence of VOCs in the United Kingdom from Sep 2020 to May 2022. **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. **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 globally. **h**, Cumulative prevalence of BA.2 in each country 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
<p>What is the prevalence of a set of mutations within different lineages?</p>	<p>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<sup>33,34</sup>. 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 (e.g., S:E484K and S:N501Y), we estimate the prevalence of that set of mutations across all lineages globally. (<b>Fig 2c</b>).</p>
<p>What is the trend shown by the prevalence of a lineage and/or a set of mutations over time?</p>	<p>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 (<b>Fig 2d</b>).</p>
<p>What are the "characteristic mutations" of a lineage?</p>	<p>The mutations that are characteristic of a lineage can be used to generate hypotheses about the phenotype exhibited by a lineage based on prior 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 (<b>Fig 2e</b>). These mutations are displayed in a "mutation map".</p>
<p>What is the total number of sequences that belong to a lineage and/or a set of mutations?</p> <p>In how many countries was a lineage and/or a set of mutations detected?</p> <p>When was this lineage and/or a set of mutations first detected?</p>	<p>In order 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 (<b>Fig 2f</b>).</p>
<p>What is the geographic prevalence of a lineage and/or a set of mutations?</p>	<p>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 since they were first detected across the sub-admin levels of a given location for a lineage/mutation query (<b>Fig 2g</b>). 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 areas, we use a dot chart to show the uncertainty in the point estimate of prevalence over the last 60 days and a bar chart to show the number of sequences used</p>

	to calculate it ( <b>Fig 2h</b> ). 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. In order 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 <sup>32</sup> ( <b>Fig 2i</b> ).

168

169

### Location Tracker

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

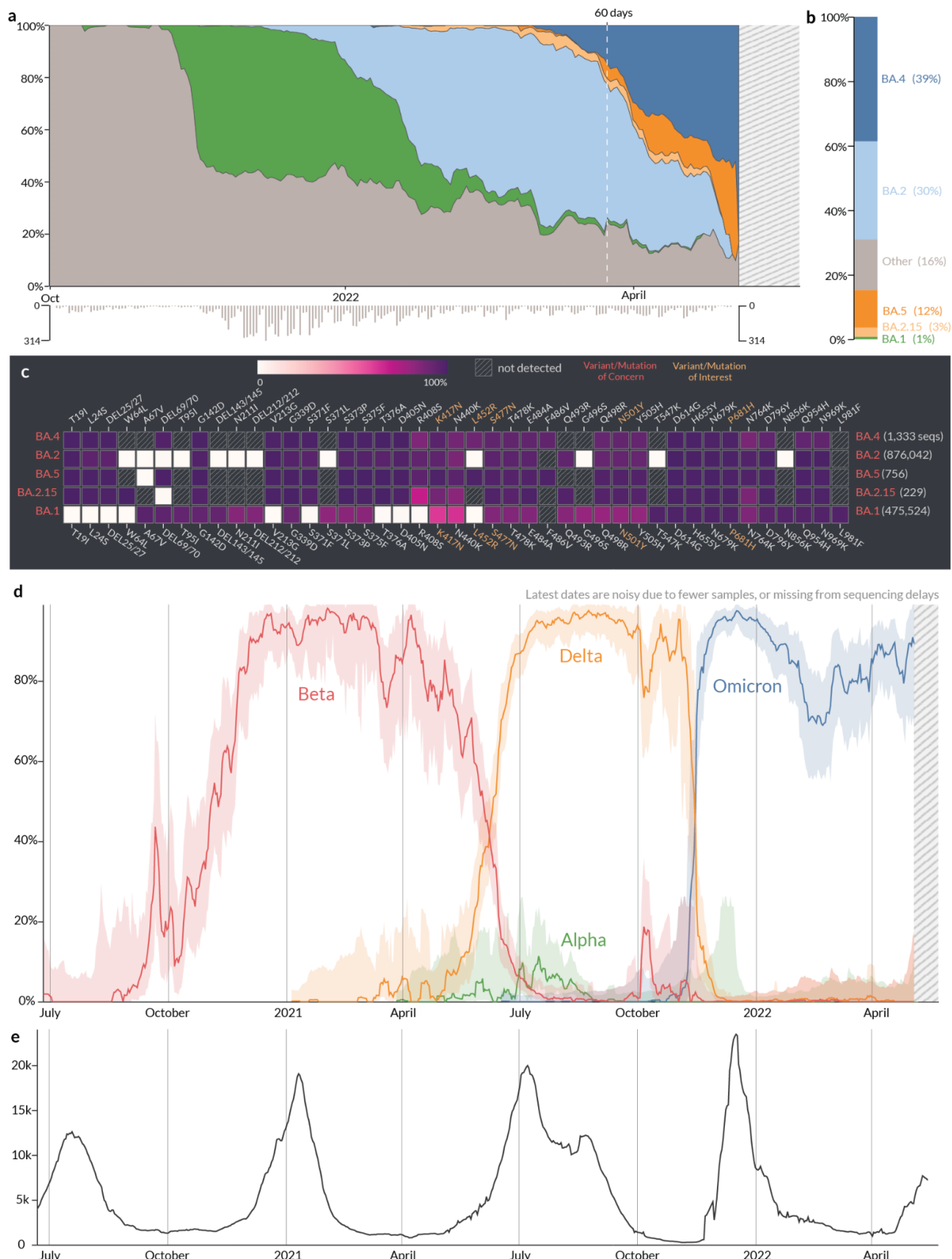
Some VOCs have only been regionally dominant. For example, Beta and Gamma were dominant in South Africa<sup>8</sup> and Brazil<sup>35</sup> respectively. Similarly, B.1.621<sup>36</sup> was only dominant in Columbia, A.2.5 was only dominant in Panama, and B.1.177 exhibited a high growth rate only in European countries due to a resurgence of travel in the summer of 2020<sup>24,37</sup>. Factors such as the attack rate, population immunity due to previous infection or vaccination, and social mobility vary from one region to the next and have a significant 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 days, 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.

**Table 2. Questions addressed by the Location Tracker**

Question	Relevant visual elements
What are the most prevalent lineages over the last 60 days?	In order to quickly provide a snapshot of the lineages currently circulating in a given location, we show a stream graph of the prevalence of lineages over the last 60 days ( <b>Fig 3a</b> ). In order to increase interpretability, we grouped lineages that are below 3% prevalence for at least five days over the last 60 days 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 stream graph. In addition, a stacked bar graph shows a snapshot of the cumulative prevalence of the lineages over the last 60 days ( <b>Fig 3b</b> ). Additionally, the user can adjust this window to look at different time windows, e.g. 180 days.
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

	<p>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 (<b>Fig 3c</b>). 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.</p>
<p>How does the prevalence of different lineages or mutations within this location change over time?</p>	<p>In addition to showing a snapshot of the lineages circulating over the last 60 days, 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 (<b>Fig 3d</b>).</p>
<p>How does the lineage prevalence over time correspond to the number of daily reported cases in this region?</p>	<p>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 (<b>Fig 3e</b>). 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.</p>





187  
 188 **Figure 3. Location report.** **a**, Relative prevalence of all lineages over time in South Africa. Total number of  
 189 sequenced samples collected per day are shown in the bar chart below. **b**, Relative cumulative prevalence of  
 190 all lineages over the last 60 days in South Africa. **c**, Mutation prevalence across the most prevalent lineages in  
 191 South Africa over the last 60 days. **d**, Comparison of the prevalence of VOCs grouped by WHO classification:  
 192 Alpha, Beta, Delta, and Omicron over time in South Africa. **e**, Daily reported cases in South Africa are shown  
 193 in the line chart below.

194 **Case Study: outbreak.info as a hypothesis generation tool to investigate geographic**  
195 **variation in lineage dynamics of VOCs**

196 As the pandemic has continued to progress, we have seen the emergence of VOCs with significant  
197 fitness advantages that were able to outcompete previously circulating lineages. As of May 2022,  
198 there have been five designated VOCs: Alpha (B.1.1.7 + sublineages, indicated by \*), Beta (B.1.351\*),  
199 Gamma (P.1\*), Delta (B.1.617.2\*), and Omicron (B.1.1.529\*). Of these, Alpha, Beta and Gamma were  
200 estimated to have emerged between September and December 2020<sup>9,38,39</sup> and were subsequently  
201 outcompeted globally by the Delta variant that was first detected in December 2020<sup>40</sup>. The Omicron  
202 lineage, first detected in November 2021<sup>11</sup>, was able to outcompete Delta and grew much more  
203 rapidly relative to previous VOCs during their emergence (**Fig 4a**). Whereas Delta and Omicron  
204 variants exhibited high growth rates with little variation globally, Alpha continued to circulate in low  
205 prevalence in Brazil and South Africa, where Gamma and Beta variants were dominant respectively  
206 (**Fig 4b, 5c**). Additionally, the prevalence of sublineages within Delta and Omicron variants varies  
207 geographically. The Location Tracker on `outbreak.info` can be used to track the growth of VOCs  
208 within a given location, thus facilitating the comparison of lineage growth rates across locations.  
209 The Location Tracker can also be used to track the relative prevalence of sublineages within these  
210 VOCs, shedding light on any geographic variation in these dynamics. Here, we examine trends in  
211 the prevalence of the five VOCs globally and highlight the geographic variation in growth rates of  
212 Alpha, Beta, Gamma, Delta, and Omicron variants.

213  
214 The earliest samples of the Alpha variant were sequenced in Southern England in late September  
215 2020<sup>38</sup>. There were multiple introductions of the lineage into the United States (U.S.) as early as late  
216 November<sup>41</sup>. The Alpha variant showed a transmission advantage of 40-50% in the U.S.<sup>26</sup>, in line  
217 with observations in the United Kingdom and the Netherlands. In the U.S., Alpha was able to  
218 outcompete previously circulating lineages and continued to increase in prevalence until the  
219 introduction of the Delta variant around April 2021 (**Fig 4d**). In contrast to the U.S., the Alpha variant  
220 circulated at very low prevalence in Brazil, while the Gamma variant remained dominant in the  
221 country<sup>9</sup> until the introduction of the Delta variant around April 2021 (**Fig 4b**). Similarly, in South  
222 Africa, the Beta variant continued to spread until the emergence of the Delta variant and the Alpha  
223 variant never became dominant (**Fig 4c**). Whereas the Beta and Gamma variants were able to  
224 outcompete Alpha in South Africa and Brazil respectively, Gamma only reached a maximum  
225 prevalence of 8% in the U.S. in May 2020, and Beta circulated at a prevalence of <1% (**Fig 4d**). The  
226 growth of a lineage is determined by epidemiological factors such as number of introductions, travel  
227 between locations, and by intrinsic biological properties such as transmission advantage or immune  
228 evasion. Both Beta and Gamma variants show varying degrees of immune evasion<sup>42</sup>. Regions of  
229 Brazil had attack rates as high as 75% in October 2020<sup>43</sup>, indicating that immune evasion was the  
230 primary reason for the rapid growth of the P.1 lineage in Brazil. In contrast, states in the U.S. had  
231 an estimated attack rate between 0.1% and 16% in June 2020<sup>44</sup>. Given this difference in attack rates,  
232 we can hypothesize that the intrinsic transmission advantage of the Alpha variant was able to  
233 outcompete the advantage conferred by immune evasion of Gamma in the U.S., but the opposite  
234 was true in Brazil and South Africa. In all three countries, the introduction of the Delta lineage  
235 displaced previously circulating Alpha, Beta, and/or Gamma lineages in the summer of 2021.

236  
237 The Delta variant of SARS-CoV-2 was first detected in Maharashtra, India in December 2020<sup>40</sup>, has  
238 been shown to be 40%-60% more transmissible than Alpha<sup>45,46</sup>, and causes a reduction in vaccine

239 efficacy relative to previously circulating lineages<sup>47</sup>. Vaccination campaigns against COVID-19 started  
240 in December 2020 and despite the progress of these campaigns<sup>48</sup>, the Delta variant continued to  
241 cause a renewed surge in infections globally. The Delta variant report, which can be accessed  
242 directly on the landing page of the lineage reports view, can be used to understand the dynamics  
243 of its sublineages. **Fig 4a** shows the global prevalence of the Delta variant over time. This growth  
244 reflects the transmission advantage that Delta has over previously circulating lineages including  
245 VOCs Alpha, Beta, and Gamma. As the Delta variant continued to spread, its genetic diversity  
246 increased and as of May 2022, over 200 sublineages of Delta have been designated<sup>49</sup>.

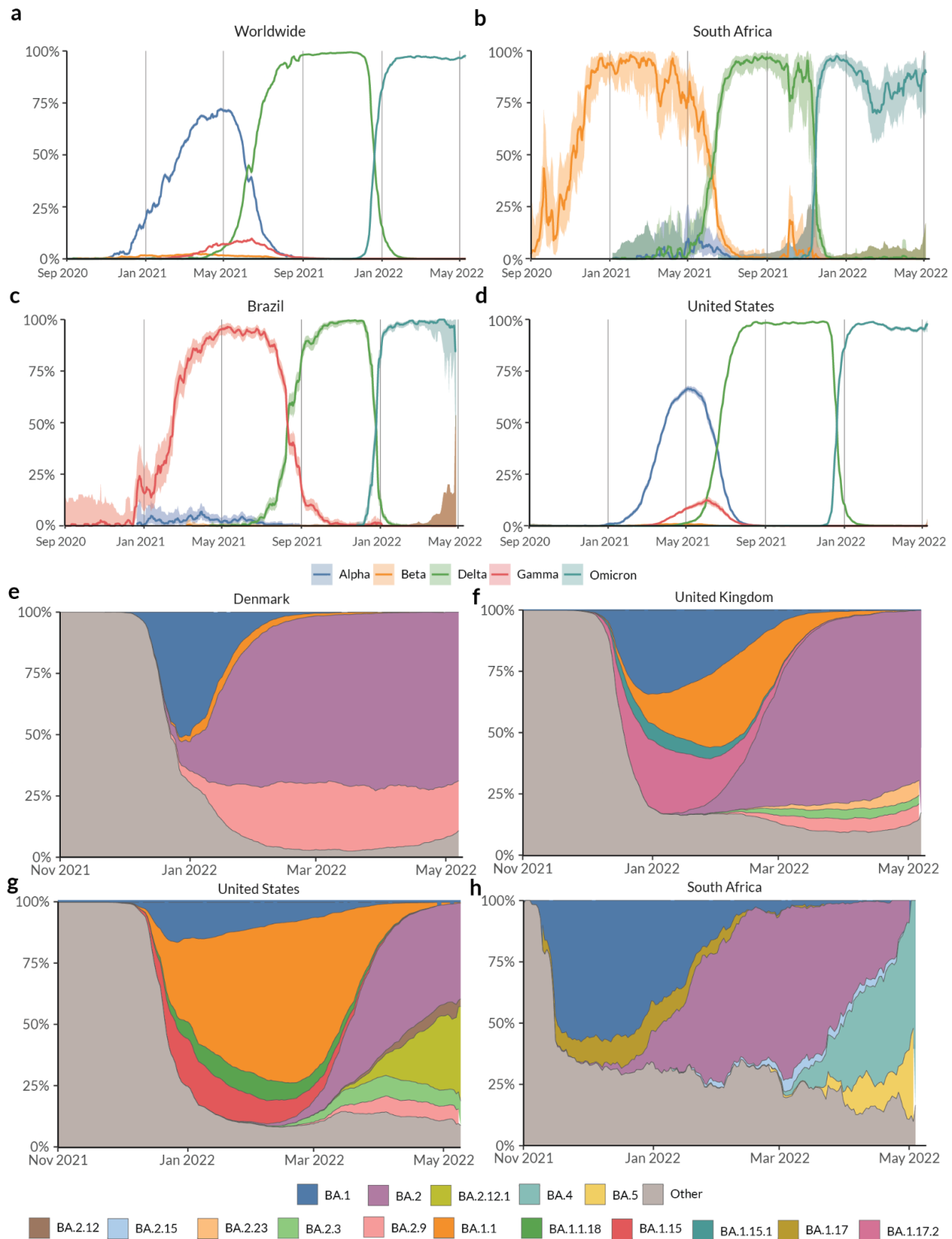
247

248 The Omicron variant was first detected in November 2021 by genomic surveillance teams in South  
249 Africa and Botswana. This variant was associated with a rapid resurgence of infections in Gauteng  
250 Province, South Africa and was designated a VOC by the WHO within 3 days of uploading the first  
251 genome<sup>11</sup>. The variant grew in prevalence very rapidly and within three weeks, the variant was  
252 detected in 87 countries and as of May 2022, Omicron has a prevalence of over 95% globally (**Fig**  
253 **4a**). While increased transmissibility confers a bigger fitness advantage compared to immune  
254 evasion when population immunity is low, the opposite is true as population immunity increases  
255 either due to vaccination or previous infection<sup>50</sup>. The Omicron variant was found to have a five fold  
256 higher chance of reinfection compared to Delta<sup>51</sup>, and Omicron infections presented with a higher  
257 viral load than wild type but still lower than Delta<sup>52</sup>. As viral load is one of the determinants of  
258 transmissibility, this indicates that Omicron is intrinsically not as transmissible as Delta, but it  
259 exhibits better immune evasion. This combination gave Omicron a large fitness advantage over  
260 Delta as evidenced by its rapid growth rate worldwide (**Fig 4a**). The continued spread of the variant  
261 has resulted in the emergence of many sublineages and as of May 2022, over 100 sublineages of  
262 Omicron have been designated. Importantly, there is significant geographic variation in the relative  
263 prevalence of newly designated sublineages such as BA.2.12.1, BA.4, and BA.5. While BA.2 continues  
264 to be the dominant sublineage within Omicron in countries such as Denmark and the United  
265 Kingdom (**Fig 4e, 4f**), we see the BA.2.12.1 sublineage slowly displacing BA.2 in the United States (**Fig**  
266 **4g**). In South Africa, sublineages BA.4 and BA.5 have completely displaced the previously dominant  
267 BA.2 (**Fig 4h**) and have led to another surge in reported cases (**Fig 3e**). The three variants, BA.2.12.1,  
268 BA.4, and BA.5 have been shown to evade antibodies elicited by prior BA.1 infection in *in vitro*  
269 neutralization studies<sup>53,54</sup>. This observed escape was higher than what was observed for BA.2<sup>55</sup>,  
270 highlighting the possibility that these variants led to a renewed surge in infections as these variants  
271 continue to spread globally. While the growth of Alpha and Delta variants globally was driven  
272 primarily by higher intrinsic transmissibility, the growth of the new variants within Omicron seems  
273 to be driven primarily by enhanced immune evasion. The increasing prevalence of immunity due to  
274 vaccination or prior infection worldwide, further supports this hypothesis.

275

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

280



281  
282  
283  
284  
285  
286

**Figure 4.** Prevalence of Variants of Concern: Alpha, Beta, Gamma, Delta, and Omicron lineages over time in the (a) Worldwide, (b) South Africa, (c) Brazil, and (d) United States. Lineages with a prevalence over 3% over the last 60 days in (e) Denmark, (f) United Kingdom, (g) United States, and (h) South Africa.

## 287 Discussion

288 The Omicron variant, first detected in late November 2021, has outcompeted Delta and is currently  
289 the most dominant lineage globally. However, it is important to note that regardless of how  
290 prevalent previously circulating VOCs were, all five VOCs emerged independent of each other. While  
291 the current hypothesis for the emergence of VOCs is prolonged virus evolution in a chronically  
292 infected individual<sup>56</sup>, we still lack a thorough understanding of this process. Given the underlying  
293 stochasticity of this process, *predicting* the emergence of a new VOC is not currently feasible. As a  
294 result, continued surveillance of all currently circulating lineages is of utmost importance to public  
295 health globally — particularly as SARS-CoV-2 continues to spread and evolve worldwide.

296  
297 The global community has generated over 10 million genomes of SARS-CoV-2 as of May 2022,  
298 shared on platforms such as GISAID<sup>14</sup>. The wealth of primary genomic data can enable downstream  
299 applications such as tracking the prevalence of different virus lineages in near real-time. However,  
300 the sheer volume of genomic data that continues to increase daily presents challenges to running  
301 analyses ad hoc. We developed `outbreak.info` to serve as a template for tracking the spread of  
302 the pandemic over varying geographic and temporal scales at scale, across the world, in near-real  
303 time. This new paradigm centralizes the computation of key statistics based on the analysis of  
304 disparate data streams. We designed the server infrastructure of `outbreak.info` keeping two  
305 goals in mind: **scalability** of the API as existing data sources increase in size and new data sources  
306 are incorporated and **reusability** of the computed data by providing programmatic access through  
307 an R package (**Fig 5**). This approach enables us to quickly adapt to and incorporate new modes of  
308 surveillance such as the CDC's National Wastewater Surveillance System<sup>57</sup>. Furthermore, the easy  
309 dissemination of any computed data on `outbreak.info` via the R package enables users to further  
310 interrogate and utilize this data for more sophisticated downstream analyses. To maximize  
311 accessibility of these data, the web interface of `outbreak.info` has been designed to offer a high  
312 degree of customizability, allowing users to answer specific biological questions and use the  
313 platform as a hypothesis generation tool. The guiding principles for the web interface have been  
314 **interactivity** via responsive user interface (UI) elements powered by a high performance API, and  
315 **interpretability** via intuitive visualization of data based on discussions with researchers,  
316 epidemiologists, and public health officials.

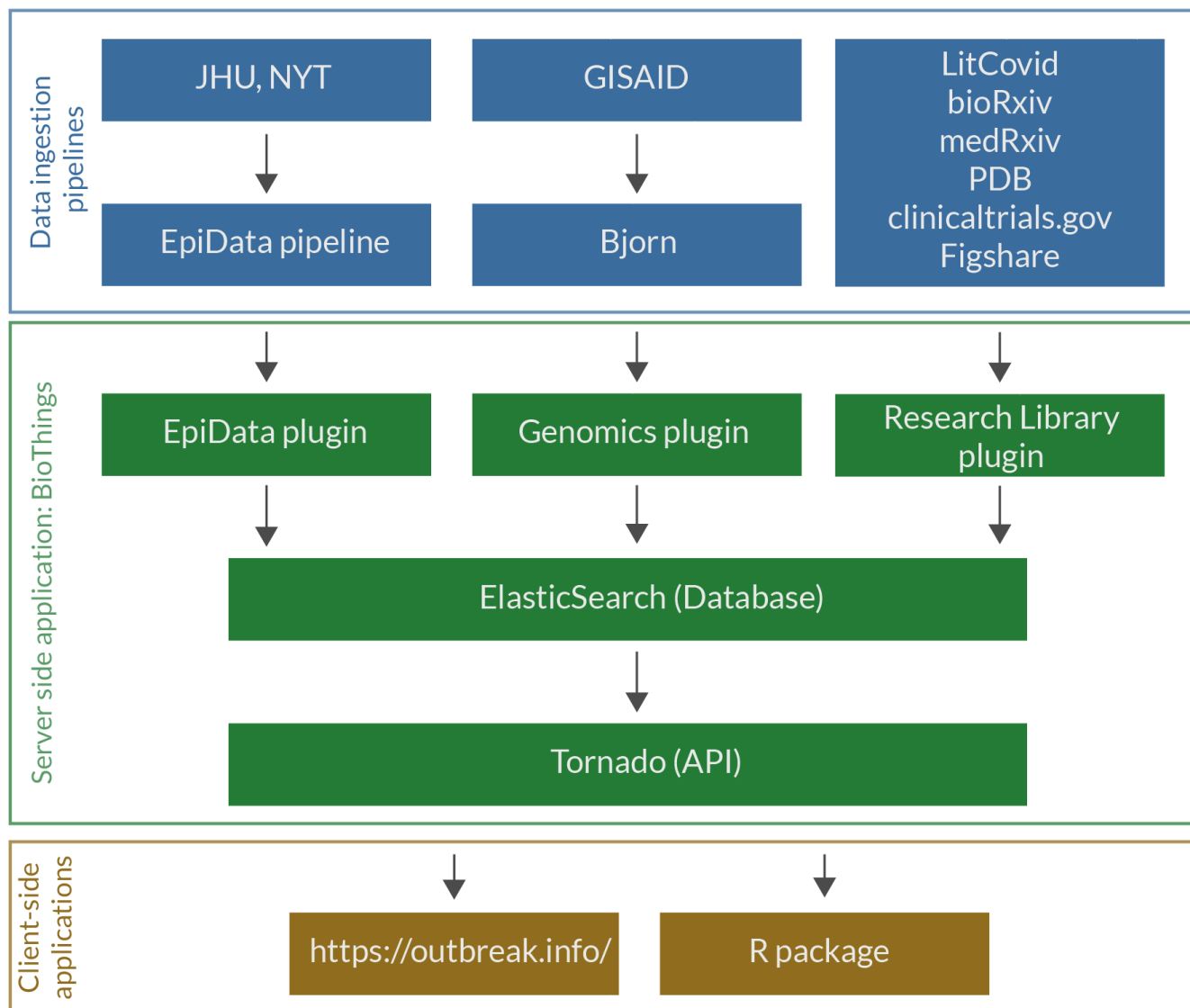
317  
318 `outbreak.info` has been enabled by unprecedented global genomic sequencing efforts, and we  
319 developed every element of the application to fully leverage this capacity. However, genomic  
320 sampling varies globally with the vast majority of sequences coming from high income countries;  
321 even within well-sampled regions, there is geographic and temporal variation<sup>13</sup>. To communicate  
322 the increased uncertainty due to low sampling, we calculate confidence intervals of estimates  
323 wherever applicable, provide histograms of sampling density, and mask data when there are very  
324 few data points available. Furthermore, sampling can be selective as samples of the Alpha variant  
325 and BA.1 lineage (sublineage of Omicron) show S gene target failure on a widely used qPCR assay.  
326 Such sampling biases impact the insights that can be derived from quantities such as the prevalence  
327 of a lineage/mutation. We communicate these limitations through a dedicated “caveats” page with  
328 warnings regarding the interpretation of data interspersed throughout the interface.

329 `outbreak.info` continues to provide a mechanism for researchers, epidemiologists, and public  
330 health officials to easily track the growth of variants, across any number of locations. The platform,  
331 backed by robust infrastructure, allows users to quickly access key statistics for known VOCs,



332 emerging variants, and any combination of mutations without having to run any time-consuming  
333 analyses. This allows researchers to focus on data exploration, hypothesis generation and more  
334 complex downstream analyses. Beyond the SARS-CoV-2 pandemic, `outbreak.info` serves as a  
335 model for providing **scalable** and **reusable** metrics to track the spread of any pathogen during an  
336 outbreak via interactive and interpretable visualizations.

337  
338



339  
340 **Figure 5.** Software infrastructure of `outbreak.info`. The infrastructure can be broadly divided into (1) Data  
341 ingestion pipelines, (2) Server-side hosting the database and API server, and (3) Client-side applications that  
342 use the API from the server.

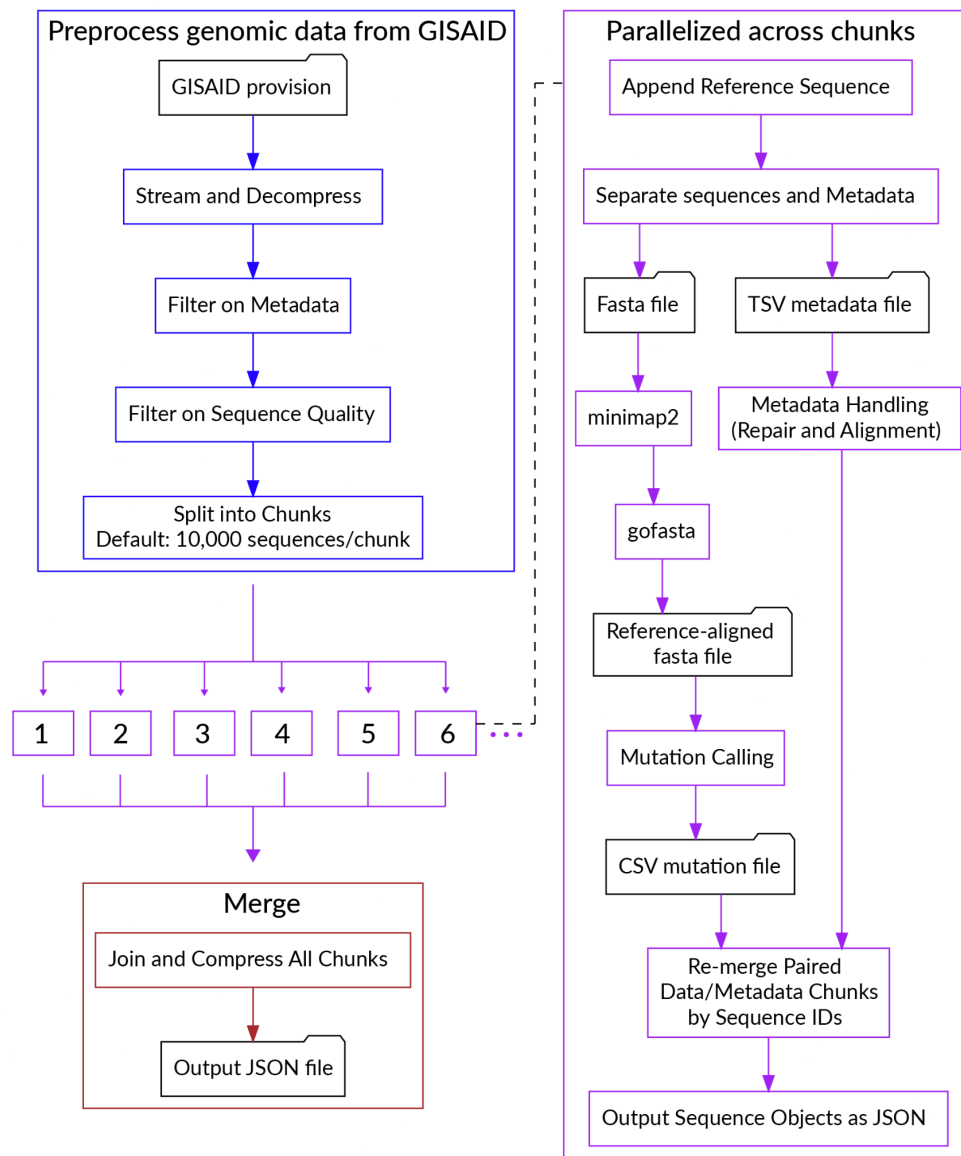
## 343 344 **Methods**

### 345 346 **Ingestion of genomic data**

347 We built a data pipeline, `Bjorn`, to count mutations from a given set of genomes in a scalable  
348 manner daily (**Fig 6**). The pipeline consists of the following steps: (1) Download SARS-CoV-2 genomes  
349 from the GISAID provision; (2) Divide sequences into chunks of 10,000; (3) Align these sequences



350 using minimap2<sup>58</sup>; (4) Convert the alignment into a FASTA file using gofasta  
351 (<https://github.com/virus-evolution/gofasta>); (5) count mutations and deletions from this alignment;  
352 (6) standardize and filter the metadata: country, division, location, pangolin lineage, date of  
353 collection, and date of submission and (7) combine results from all chunks and convert to a JSON  
354 object. We standardized geographic identifiers using shapefiles from GADM<sup>59</sup>. The final JSON object  
355 is loaded into an Elasticsearch index within the BioThings framework<sup>60</sup>. The code for Bjorn  
356 is available at <https://github.com/andersen-lab/bjorn>.



357  
358 **Figure 6.** Flowchart describing the steps in Bjorn.  
359

### 360 Ingestion of epidemiological data

361 We built the EpiData pipeline to ingest reported global cases, and deaths from Johns Hopkins  
362 University<sup>31</sup>. We used shapefiles from Natural Earth<sup>61</sup> to standardize geographic identifiers, and  
363 obtain populations for countries and states outside the U.S. For the U.S., we used the county level  
364 shapefiles and population estimates from the 2019 population estimates by the Census Bureau to  
365 standardize geographic identifiers and get population estimates. We standardized reported date

366 formats, and geographic identifiers across the two resources. The code for the EpiData pipeline is  
367 available at [https://github.com/outbreak-info/biothings\\_covid19](https://github.com/outbreak-info/biothings_covid19).

368

### 369 **Calculation of confidence intervals on prevalence**

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

374

### 375 **Creation of `outbreak.info` API**

376 In order to scale with the increasing size of existing data sources and the heterogeneity of newly  
377 emerging data sources, we used the BioThings framework<sup>60</sup>. The JSON outputs of our data pipelines  
378 are ingested by the BioThings framework and the processed data is stored in individual  
379 Elasticsearch indices. A Tornado server is used to create API endpoints that leverage the search  
380 capabilities of Elasticsearch to perform complex aggregations of the underlying data. These API  
381 endpoints allow the client-side applications to query the underlying data within reasonable query  
382 times while accounting for the scale of the ingested data. The BioThings Hub maintains historical  
383 data by default, allowing us to roll back to previous data backups if issues are discovered with new  
384 data after they are deployed. The code for the server-side application is available at  
385 <https://github.com/outbreak-info/outbreak.api>.

386

### 387 **`outbreak.info` web application**

388 The web application was built using Vue.js<sup>62</sup>, a model-view-viewmodel JavaScript framework which  
389 enables the two-way binding of user interface elements and the underlying data allowing the user  
390 interface to reflect any changes in underlying data and vice versa. The client-side application uses  
391 the high performance API to interactively perform operations on the database. Customized data  
392 visualizations on the client were built using D3.js<sup>63</sup>, giving us the ability to develop novel, and  
393 intuitive visual elements as part of the user interface. We designed these visualizations to answer  
394 specific questions of interest to epidemiologists, researchers, and public health officials. We further  
395 added functionality to enable the 1-click copy or download of every chart in the interface as a png  
396 or svg. The code for the client-side application is available at: [https://github.com/outbreak-](https://github.com/outbreak-info/outbreak.info)  
397 [info/outbreak.info](https://github.com/outbreak-info/outbreak.info)

398

### 399 **R package**

400 We developed an R package for `outbreak.info` to allow researchers and other individuals to easily  
401 access the data via the API for downstream analyses and visualizations. The R package is composed  
402 of three parts: functions that allow the user to access genomic data, functions to access the  
403 epidemiological data, and functions to access the Research Library metadata. They all consist of a  
404 base function that contains arguments for all possible parameters that can be used to query the  
405 API. While users can utilize this base function directly to access data, several wrapper functions are  
406 available that inherit the arguments from the base function in addition to pre-specified parameters  
407 to simplify the process of querying the API. For example, while `getGenomicData()` can be used  
408 directly to access data regarding the daily global prevalence of a specified lineage, doing so would  
409 require a user to be familiar with the name of the endpoint as specified in the API URL (in this case,  
410 global-prevalence). Instead, the user can access this data with the more intuitively named

411 `getPrevalence()`. Therefore, these wrapper functions allow users to easily and quickly obtain the  
412 data they need. The R package also contains an `authenticateUser()` function that allows users to  
413 authenticate against their GISAID credentials and access computed statistics from the primary  
414 genomic data provided by GISAID.

415  
416 In addition, as the API queries location by ISO3 code, rather than by location name, two functions  
417 have been created that allow users to forgo the step of searching for the ISO3 code themselves:  
418 `getISO3Code()` and `getLocationIdGenomic()`. The latter function uses the genomics API  
419 endpoint to obtain the ISO3 code for a given location. The ISO3 code can be obtained with either a  
420 full or incomplete location name; in the latter case, the user will be provided a list of matching  
421 locations and must specify the location they are interested in. This function is embedded in the  
422 parent `getGenomicData()` function, and is therefore inherited in all wrapper functions. Therefore,  
423 searching for data by location in the R package replicates the experience on the client-side web  
424 application. Documentation is available at: <https://outbreak-info.github.io/R-outbreak-info> with  
425 vignettes located at <https://outbreak-info.github.io/R-outbreak-info/articles/index.html>. The R  
426 package can be downloaded and installed using the `remotes` package function:  
427 `install_github("outbreak-info/R-outbreak-info")`.

428

### 429 **Acknowledgements**

430 We thank the technical team, sequence curators, and administrators of the GISAID database for  
431 helping us with this project and supporting rapid and transparent sharing of genomic data during  
432 the COVID-19 pandemic. We thank our colleagues for sharing genomic data on GISAID. This work  
433 was supported by the National Institute for Allergy and Infectious Diseases (5 U19 AI135995, 3  
434 U19 AI135995-04S3, 3 U19 AI135995-03S2, U01AI151812, R01 AI162611, R01 AI153044), National  
435 Center For Advancing Translational Sciences (5 U24 TR002306, UL1TR002550), Centers for Disease  
436 Control and Prevention (75D30120C09795), and the National Institute of General Medical Sciences  
437 (R01GM083924).

438

### 439 **Conflicts of Interest**

440 MAS receives grants from the US National Institutes of Health within the scope of this work, and  
441 grants and contracts from the US Food & Drug Administration, the US Department of Veterans  
442 Affairs and Janssen Research & Development outside the scope of this work. MAS and KGA have  
443 received consulting fees and/or compensated expert testimony on SARS-CoV-2 and the COVID-19  
444 pandemic.

445

446

447

448

449

450

451

452

453

454

455

## 456 References

- 457 1. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.*  
458 **382**, 727–733 (2020).
- 459 2. Edward Holmes *et al.* Novel 2019 coronavirus genome. [https://virological.org/t/novel-2019-](https://virological.org/t/novel-2019-coronavirus-genome/319)  
460 [coronavirus-genome/319](https://virological.org/t/novel-2019-coronavirus-genome/319) (2020).
- 461 3. GISAID - Initiative. <https://gisaid.org>.
- 462 4. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC weekly* **3**, (2021).
- 463 5. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist  
464 genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).
- 465 6. Konings, F. *et al.* SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for  
466 global discourse. *Nature microbiology* **6**, (2021).
- 467 7. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in  
468 England. *Science* (2021) doi:10.1126/science.abg3055.
- 469 8. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-  
470 related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa.  
471 *bioRxiv* (2020) doi:10.1101/2020.12.21.20248640.
- 472 9. Faria, N. R. *et al.* Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil.  
473 *medRxiv* (2021) doi:10.1101/2021.02.26.21252554.
- 474 10. Liu, Y. *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant.  
475 *bioRxiv* 2021.08.12.456173 (2021) doi:10.1101/2021.08.12.456173.
- 476 11. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern  
477 Africa. *Nature* (2022) doi:10.1038/d41586-021-03832-5.
- 478 12. Liu, L. *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2.  
479 *Nature* (2021) doi:10.1038/d41586-021-03826-3.
- 480 13. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *medRxiv*  
481 2021.08.21.21262393 (2021).
- 482 14. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution  
483 to global health. *Global Challenges* vol. 1 33–46 (2017).
- 484 15. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the  
485 Americas. *Nature* **546**, 406–410 (2017).
- 486 16. Metsky, H. C. *et al.* Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415  
487 (2017).
- 488 17. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into  
489 the United States. *Nature* **546**, 401–405 (2017).
- 490 18. Volz, E. M., Koelle, K. & Bedford, T. Viral Phylodynamics. *PLoS Comput. Biol.* **9**, (2013).
- 491 19. Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in  
492 Sierra Leone. *Cell* **161**, 1516–1526 (2015).
- 493 20. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic.  
494 *Nature* **544**, 309–315 (2017).
- 495 21. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–  
496 4123 (2018).
- 497 22. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and  
498 phylogeography. *Microbial Genomics* **2**, e000093 (2016).
- 499 23. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck.

- 500 *Nature* **591**, 30–33 (2021).
- 501 24. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020.  
502 *Nature* **595**, 707–712 (2021).
- 503 25. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**,  
504 266–269 (2021).
- 505 26. Website. <https://doi.org/10.1016/j.cell.2021.03.052> doi:10.1016/j.cell.2021.03.052.
- 506 27. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA:  
507 a mathematical modelling study. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
- 508 28. Nouvellet, P. *et al.* Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1–9  
509 (2021).
- 510 29. Fischer, C. B. *et al.* Mask adherence and rate of COVID-19 across the United States. *PLoS One*  
511 **16**, e0249891 (2021).
- 512 30. Variation in government responses to COVID-19.  
513 <https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19>.
- 514 31. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real  
515 time. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.
- 516 32. Tsung, G. *et al.* Outbreak.info: A standardized, searchable platform to discover and explore  
517 COVID-19 resources and data. *bioRxiv* 2022.01.20.477133 (2022)  
518 doi:10.1101/2022.01.20.477133.
- 519 33. Xie, X. *et al.* Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by  
520 BNT162b2 vaccine-elicited sera. *Nat. Med.* **27**, 620–621 (2021).
- 521 34. Website. <https://doi.org/10.1016/j.cell.2020.08.012> doi:10.1016/j.cell.2020.08.012.
- 522 35. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus:  
523 preliminary findings. *virological.org* (2021).
- 524 36. Laiton-Donato, K. *et al.* Characterization of the emerging B.1.621 variant of interest of SARS-  
525 CoV-2. *Infect. Genet. Evol.* **95**, 105038 (2021).
- 526 37. Lemey, P. *et al.* Untangling introductions and persistence in COVID-19 resurgence in Europe.  
527 *Nature* **595**, 713–717 (2021).
- 528 38. O’Toole, Á. *et al.* Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and  
529 B.1.351/501Y-V2 with grinch. *Wellcome Open Research* **6**, (2021).
- 530 39. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* (2021)  
531 doi:10.1038/s41586-021-03402-9.
- 532 40. GISAID - hCov19 Variants. <https://www.gisaid.org/hcov19-variants/>.
- 533 41. Washington, N. L. *et al.* Genomic epidemiology identifies emergence and rapid transmission of  
534 SARS-CoV-2 B.1.1.7 in the United States. *medRxiv* (2021) doi:10.1101/2021.02.06.21251159.
- 535 42. Hoffmann, M. *et al.* SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies.  
536 *Cell* **184**, 2384 (2021).
- 537 43. Buss, L. F. *et al.* Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a  
538 largely unmitigated epidemic. *Science* **371**, (2021).
- 539 44. Unwin, H. J. T. *et al.* State-level tracking of COVID-19 in the United States. *Nat. Commun.* **11**, 1–9  
540 (2020).
- 541 45. [No title].  
542 [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_da](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993321/S1267_SPI-M-O_Consensus_Statement.pdf)  
543 [ta/file/993321/S1267\\_SPI-M-O\\_Consensus\\_Statement.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993321/S1267_SPI-M-O_Consensus_Statement.pdf).
- 544 46. Dhar, M. S. *et al.* Genomic characterization and Epidemiology of an emerging SARS-CoV-2



- 545 variant in Delhi, India. *medRxiv* 2021.06.02.21258076 (2021).
- 546 47. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion.  
547 *Nature* 1–8 (2021).
- 548 48. Ritchie, H. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
- 549 49. Cov-Lineages. [https://cov-lineages.org/lineage\\_list.html](https://cov-lineages.org/lineage_list.html).
- 550 50. Bushman, M., Kahn, R., Taylor, B. P., Lipsitch, M. & Hanage, W. P. Population impact of SARS-  
551 CoV-2 variants with enhanced transmissibility and/or partial immune escape. *Cell* **184**, 6229–  
552 6242.e18 (2021).
- 553 51. [No title]. [https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2021-12-16-](https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2021-12-16-COVID19-Report-49.pdf)  
554 [COVID19-Report-49.pdf](https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2021-12-16-COVID19-Report-49.pdf).
- 555 52. [No title].  
556 [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_da](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1042367/technical_briefing-31-10-december-2021.pdf)  
557 [ta/file/1042367/technical\\_briefing-31-10-december-2021.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1042367/technical_briefing-31-10-december-2021.pdf).
- 558 53. Khan, K. *et al.* Omicron sub-lineages BA.4/BA.5 escape BA.1 infection elicited neutralizing  
559 immunity. doi:10.1101/2022.04.29.22274477.
- 560 54. Xie, X. *et al.* BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron BA.1 infection.  
561 doi:10.21203/rs.3.rs-1611421/v1.
- 562 55. Khan, K. *et al.* Omicron infection enhances Delta antibody immunity in vaccinated persons.  
563 *Nature* 1–3 (2022).
- 564 56. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK.  
565 doi:10.1101/2022.03.08.481609.
- 566 57. CDC. National Wastewater Surveillance System (NWSS). *Centers for Disease Control and*  
567 *Prevention* [https://www.cdc.gov/healthywater/surveillance/wastewater-](https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater-surveillance.html)  
568 [surveillance/wastewater-surveillance.html](https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance.html) (2022).
- 569 58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100  
570 (2018).
- 571 59. GADM. <https://gadm.org/>.
- 572 60. Lelong, S. *et al.* BioThings SDK: a toolkit for building high-performance data APIs in biomedical  
573 research. *Bioinformatics* (2022) doi:10.1093/bioinformatics/btac017.
- 574 61. Natural Earth - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales.  
575 <https://www.naturalearthdata.com/>.
- 576 62. Vue.js. <https://vuejs.org/>.
- 577 63. Bostock, M. Data-Driven Documents. <https://d3js.org/>.