# Genomic reconstruction of the SARS-CoV-2 epidemic in England

Harald S. Vöhringer[1,2], Theo Sanderson[3,4], Matthew Sinnott[3], Nicola De Maio[1], Thuy Nguyen[3], Richard Goater[3], Frank Schwach[3,5], Ian Harrison[5], Joel Hellewell[6], Cristina Ariani[3], Sonia Gonçalves[3], David Jackson[3], Ian Johnston[3], Alexander W. Jung[1], Callum Saint[3], John Sillitoe[3], Maria Suciu[3], Nick Goldman[1], Jasmina Panovska-Griffiths[2,7], The Wellcome Sanger Institute Covid-19 Surveillance Team[8], The COVID-19 Genomics UK (COG-UK) Consortium[9], Ewan Birney[1], Erik Volz[10], Sebastian Funk[6], Dominic Kwiatkowski[3], Meera Chand[5], Inigo Martincorena[3], Jeffrey C. Barrett[3,*], Moritz Gerstung[1,11,*]

1. European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton, UK
2. Joint Biosecurity Center JBC
3. Wellcome Sanger Institute, Hinxton, UK
4. The Francis Crick Institute, London, UK
5. Public Health England PHE
6. London School of Hygiene & Tropical Medicine, London, UK
7. Big Data Institute, University of Oxford, UK
8. https://www.sanger.ac.uk/project/wellcome-sanger-institute-covid-19-surveillance-team/
9. Full list of consortium names and affiliations are in the Appendix
10. Imperial College, London, UK
11. German Cancer Research Centre dkfz, Heidelberg, Germany

* Correspondence to Jeffrey C. Barrett jb26@sanger.ac.uk or Moritz Gerstung moritz.gerstung@ebi.ac.uk

Moritz Gerstung
European Bioinformatic Institute EMBL-EBI
Wellcome Genome Campus
Hinxton
CB10 2LP
United Kingdom

Jeffrey C. Barrett
Wellcome Sanger Institute
Wellcome Genome Campus
Hinxton
CB10 1SA
United Kingdom

# Abstract

The evolution of the SARS-CoV-2 pandemic continuously produces new variants, which warrant timely epidemiological characterisation. Here we use the dense genomic surveillance generated by the COVID-19 Genomics UK Consortium to reconstruct the dynamics of 71 different lineages in each of 315 English local authorities between September 2020 and June 2021. This analysis reveals a series of sub-epidemics that peaked in the early autumn of 2020, followed by a jump in transmissibility of the B.1.1.7/Alpha lineage. Alpha grew when other lineages declined during the second national lockdown and regionally tiered restrictions between November and December 2020. A third more stringent national lockdown suppressed Alpha and eliminated nearly all other lineages in early 2021. However, a series of variants (mostly containing the spike E484K mutation) defied these trends and persisted at moderately increasing proportions. Accounting for sustained introductions, however, indicates that their transmissibility is unlikely to have exceeded that of Alpha. Finally, B.1.617.2/Delta was repeatedly introduced to England and grew rapidly in the early summer of 2021, constituting approximately 98% of sampled SARS-CoV-2 genomes on June 26.

# Main

The SARS-CoV-2 virus accumulates approximately 24 point mutations per year, or 0.3 per viral generation [1–3]. Tracking mutations across successive virus generations enables researchers to follow transmission clusters, define distinct viral lineages and model their behaviour. Most of these mutations appear to be evolutionarily neutral, but as the SARS-CoV-2 epidemic swept around the world in the spring of 2020, it became apparent that the virus is continuing to adapt to its human host. An initial sign was the emergence and global spread of the spike protein variant D614G in the second quarter of 2020. Epidemiological analyses estimated that this mutation, which defines the B.1 lineage, confers a 20% transmissibility advantage over the original A lineage isolated in Wuhan, China [4].

A broad range of lineages have been defined since, which can be used to track SARS-CoV-2 transmission across the globe [5,6]. For example, B.1.177/EU-1, emerged in Spain in early summer of 2020 and spread across Europe through travel [7]. Subsequently, four variants of concern (VOC) have been identified by the WHO and other public health authorities: The B.1.351/Beta lineage was discovered in South Africa[8], where it spread rapidly in late 2020. The B.1.1.7/Alpha lineage was first observed in Kent in September 2020 [9] from where it swept through the United Kingdom and large parts of the world due to a 50-60% [10–13] increase in transmissibility. P.1/Gamma originated in Brazil [14,15] and has spread throughout South America. Most recently, B.1.617.2/Delta was associated with a large surge of COVID-19 in India in April 2021 and subsequently around the world.

## Spatiotemporal genomic epidemiology of SARS-CoV-2 lineages in England

In the United Kingdom, by late June 2021 the COVID-19 Genomics UK Consortium (COG-UK) has sequenced close to 600,000 viral samples. These data have enabled

detailed reconstruction of the dynamics of the first wave of the epidemic in the UK between February and August 2020 [16]. Here, we leverage a subset of those data: genomic surveillance generated by the Wellcome Sanger Institute Covid-19 Surveillance Team as part of COG-UK, to characterise the growth rates and geographic spread of different SARS-CoV-2 lineages and reconstruct how newly emerging variants changed the course of the epidemic. We will discuss the key events of the reconstructed epidemic in chronological order.

Our data covers England between September 1, 2020 and June 26, 2021 encompassing three epidemic waves and two national lockdowns (**Figure 1a**). In this time period, we sequenced 281,178 viral genomes, corresponding to an average of 7.2% (281,178/3,894,234) of all positive tests from PCR testing for the wider population outside the National Health Service (Pillar 2), ranging from 5% in the winter of 2020 to 38% in the early summer of 2021, and filtered to remove cases associated with international travel (**Methods**; **Extended Data Figure 1a,b**). Overall a total of 328 SARS-CoV-2 lineages were identified using the PANGO lineage definition [5]. As some of these lineages were only rarely and intermittently detected, we collapsed these based on the underlying phylogenetic tree into a set of 71 lineages such that each resulting lineage constituted at least 100 genomes, unless the lineage has been designated a VOC, variant under investigation (VUI) or variant in monitoring by Public Health England[17] (**Figure 1b-d, Supplementary Table 1, 2**).

These data reveal a diversity of lineages in the fall of 2020 followed by sweeps of the Alpha and Delta variants (**Figure 1b, Supplementary Table 2, 3**). **Figure 1c** shows the geographic distribution of cases and of different lineages, studied at the level of 315 English Lower Tier Local Authorities (LTLAs), administrative regions with approximately 100,000-200,000 inhabitants.

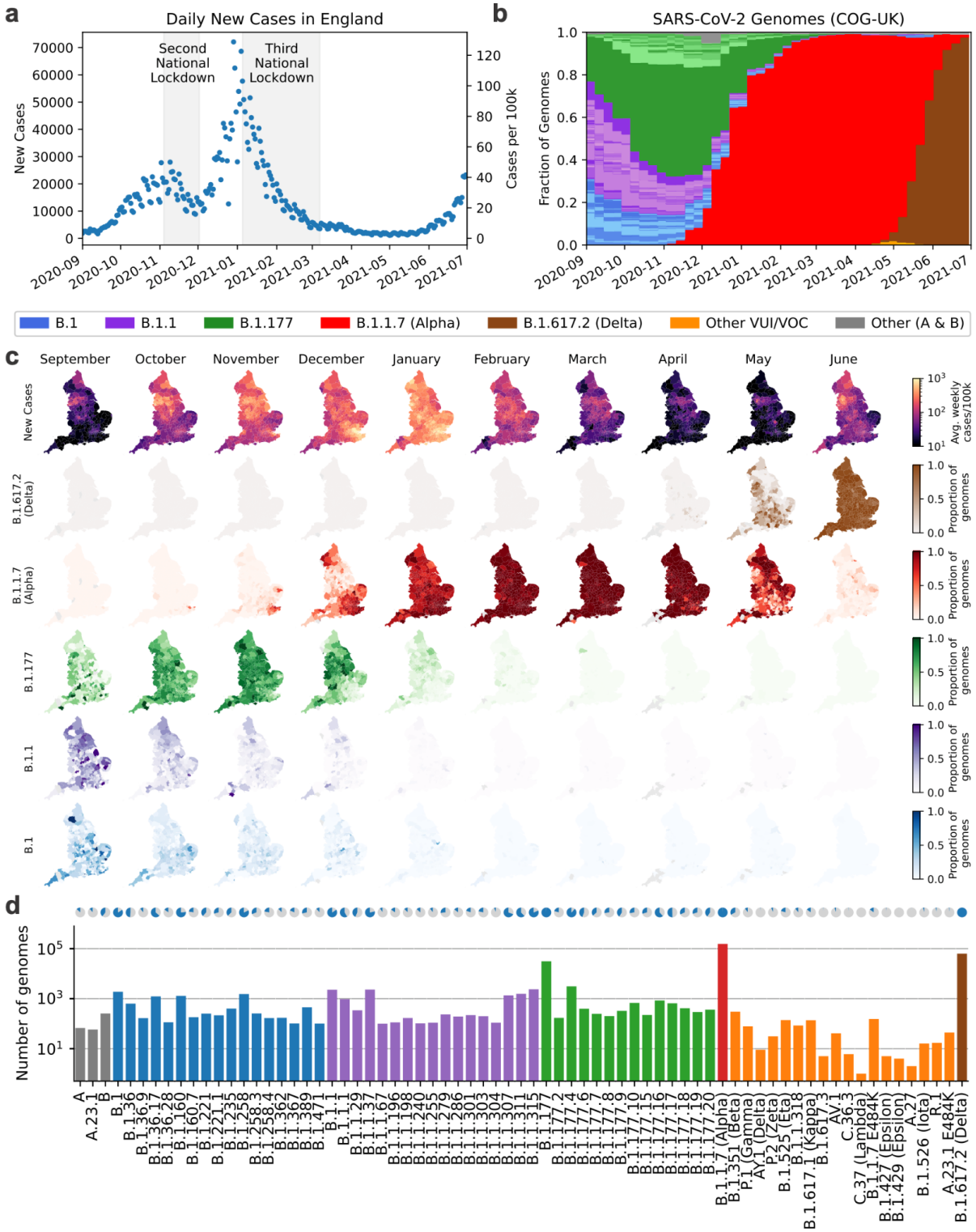



**Figure 1. SARS-CoV-2 surveillance sequencing in England between September 2020 and June 2021. a.** Positive Pillar 2 SARS-CoV-2 tests in England. **b**. Relative frequency of 328 different PANGO lineages, representing approximately 7.2% of tests shown in **a**. **c**. Positive tests (top row) and frequency of 4 major lineages across 315 English lower tier local authorities. **d**. Absolute frequency of sequenced genomes mapped to 71 PANGO lineages. Blue areas in the pie charts are proportional to the fraction of LTLAs where a given lineage was observed.

## Modeling the dynamics of SARS-CoV-2 lineages

We developed a Bayesian statistical model that tracks the fraction of genomes from different lineages in each LTLA in each week and fits the daily total number of positive Pillar 2 tests (**Extended Data Figure 2**; **Methods**). The multivariate logistic regression model is conceptually similar to previous approaches in its estimation of relative growth rates[10,11] and accounts for differences in the epidemiological dynamics between LTLAs, and allows for the introduction of new lineages (**Figure 2a-c**). Despite the sampling noise in a given week, the fitted proportions recapitulate the observed proportions of genomes as revealed by 35 example LTLAs covering the geography of England (**Figure 2b,c, Supplementary Note 1,2**). The quality of fit is confirmed by different probabilistic model selection criteria (**Extended Data Figure 3**) and also evident at the aggregated regional level (**Extended Data Figure 4**).

While the relative growth rate of each lineage is identical, the fitted patterns of viral proportions in each LTLA are dynamic due to the timing and rate of introduction of each lineage. The model also calculates total and lineage-specific local incidences and time-dependent $R_t$ values by negative binomial spline fitting of the number of daily positive PCR tests (**Methods; Figure 2d**; **Extended Data Figure 2c**). Together, this enables a quantitative reconstruction of different periods of the epidemic.

## Multiple sub-epidemics and expansion of B.1.177 in the autumn of 2020

The autumn of 2020 was characterised by a surge of cases, concentrated in the north of England, which peaked in November 2020 triggering a second national lockdown (**Figure 1a,c**). This second wave initially featured B.1 and B.1.1 sublineages, which were slightly more prevalent in the south and north of England, respectively (**Figure 2b,c**). Yet the proportion of B.1.177 and its geographically diverse sublineages steadily increased across LTLAs from around 25% at the beginning of September to 65% at the end of October. This corresponds to a growth rate between 8% (growth per 5.1d; 95% CI: 7-9) and 12% (95% CI: 11-13) greater than that of B.1 or B.1.1. The trend of B.1.177 expansion relative to B.1 persisted throughout January (**Extended Data Figure 5a**) and involved a number of monophyletic sublineages that arose in the UK and similar patterns in Denmark[18] (**Extended Data Figure 5b**). Such behavior cannot easily be explained by international travel, which was the major factor in B.1.177's initial spread throughout Europe in the summer of 2020 [7]. However, the biological explanation for this growth advantage is unclear as the characteristic A222V spike variant is not believed to confer a growth advantage [7].

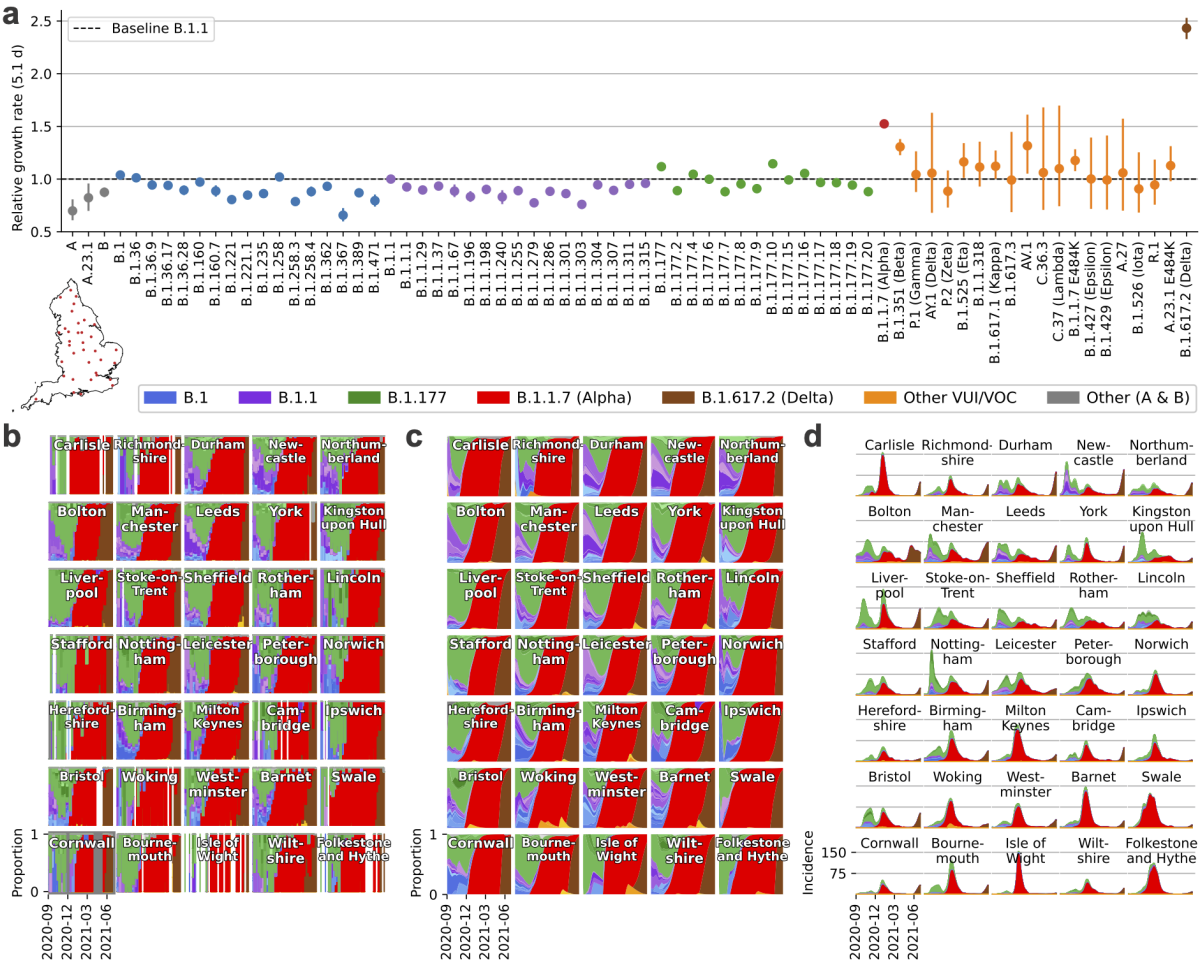



**Figure 2. Spatiotemporal model of 71 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and June 2021. a.** Average growth rates for 71 lineages. **b.** Lineage specific relative frequency for 35 selected LTLAs, arranged by longitude and latitude to geographically cover England. **c**. Fitted lineage-specific relative frequency for the same LTLAs as **b**. **d**. Fitted lineage-specific incidence for the same LTLAs as in **b**.

## Alpha-specific growth during restrictions from November 2020 to March 2021

The subsequent third wave from December 2020 to February 2021 was almost exclusively driven by Alpha/B.1.1.7 as described previously [10,11,19]. The rapid sweep of Alpha was due to an estimated transmissibility advantage of 1.52 compared to B.1.1 (growth per 5.1d; 95% CI 1.50-1.55; **Figure 2a**), assuming an unchanged generation interval distribution [20]. The growth advantage is thought to stem from spike mutations facilitating ACE2 receptor binding (N501Y)[21,22] and furin cleavage (P681H)[23]. Alpha grew during a period of restrictions, which were insufficient to contain a much more transmissible variant (**Figure 3a**).

The second national lockdown from November 5 to December 1, 2021 successfully reduced total cases, but this masked a lineage-specific rise ($R_t > 1$; defined as growth per 5.1d) of Alpha and simultaneous decline of other hitherto dominant lineages ($R_t < 1$) in 78% (246/315) of LTLAs (**Figure 3b,c**)[24]. This pattern of Alpha-specific growth during lockdown is

supported by a model-agnostic analysis of raw case numbers and proportions of Alpha genomes (**Figure 3e**).

Three levels of regionally-tiered restrictions were introduced in December 2020 [25] (**Figure 3a**). The areas under different tiers of restrictions visibly and quantitatively coincide with the resulting local $R_t$ values, with greater $R_t$ values in areas with lower restrictions (**Figure 3a-c**). The reopening caused a surge of cases across all tiers with $R_t > 1$, which is also evident in selected time series (**Figure 3d**). As Alpha cases surged, more areas were placed under tier 3 and a higher tier 4 was introduced. Nevertheless, Alpha continued to grow ($R_t > 1$) in most areas, presumably driven by increased social interaction over Christmas (**Figure 3c**).

Following the peak of 72,088 daily cases on December 29 (**Figure 1a**), a third national lockdown was announced on January 4 (**Figure 3a**). The lockdown and increasing immunity derived from infection and increasing vaccination[26] led to a sustained contraction of the epidemic to approximately 5,500 daily cases by March 8, when restrictions began to be lifted by reopening schools (further steps of easing occurred on April 12 and May 17). In contrast to the second national lockdown 93% (296/315) of LTLAs exhibited a contraction of both Alpha and other lineages (**Figure 3e)**.

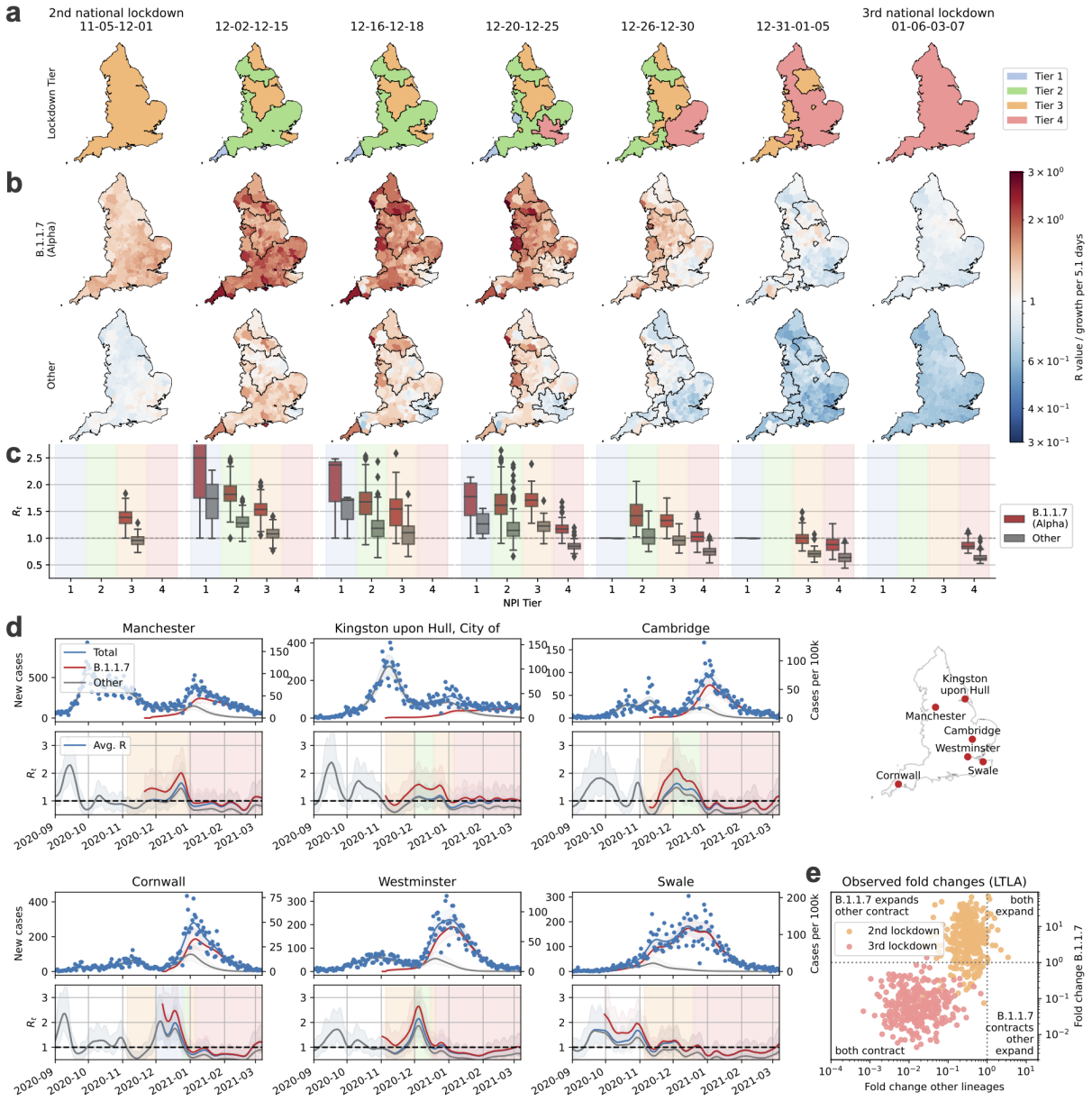



**Figure 3. Growth of B.1.1.7/Alpha and other lineages in relation to lockdown restrictions between November 2020 and March 2021. a.** Maps and dates of national and regional restrictions in England. Second national lockdown: closed hospitality businesses, contacts ≤ 2 outdoors only, open schools, reasonable excuse needed for leaving home[64]. Tier 1: private indoor gatherings ≤ 6 persons. Tier 2: as tier 1, restricted hospitality services, gatherings ≤ 6 in public outdoor places. Tier 3: as tier 2, most hospitality businesses closed. Tier 4: as tier 3, single outdoor contact. Third national lockdown: Closed schools with the exception of key workers. **b**. Local lineage-specific $R_t$ values for Alpha and the average $R_t$ value (growth per 5.1d) of all other lineages in the same periods. **c**. Boxplots of $R_t$ values shown in **b**, boxes show quartiles, whiskers extend to 1.5x the inter quartile range. **d**. Total and lineage-specific incidence (top) and $R_t$ values (bottom) for 6 selected LTLAs during the period of restrictions. **e**. Crude lineage-specific fold changes (odds ratios) for Alpha and other lineages across the second (orange) and third national lockdown (red).

## Elimination of SARS-CoV-2 lineages from January to April 2021

The lineage-specific rates of decline during the third national lockdown and throughout March 2021 resulted in large differences in lineage-specific incidence (**Figure 4a**). Cases of Alpha contracted nationally from a peak of around 50,000 daily new cases to around 2,739 (CI: 2,666-2,806) on April 1. At the same time B.1.177, the most prevalent lineage in November 2020 fell to only about 6 (95% CI 4-10) detected cases per day. Moreover, the incidence of most other lineages present in the autumn of 2020 was well below 1 after April 2021, implying that the majority of them have been eliminated. The number of observed distinct PANGO lineages declined from a peak of 137 to only 22 in the first week of April 2021 (**Figure 4b**). While this may in part be attributed to how PANGO lineages were defined, we note that the period of contraction did not replenish the genetic diversity lost due to the selective sweep by Alpha (**Extended Data Figure 6**).

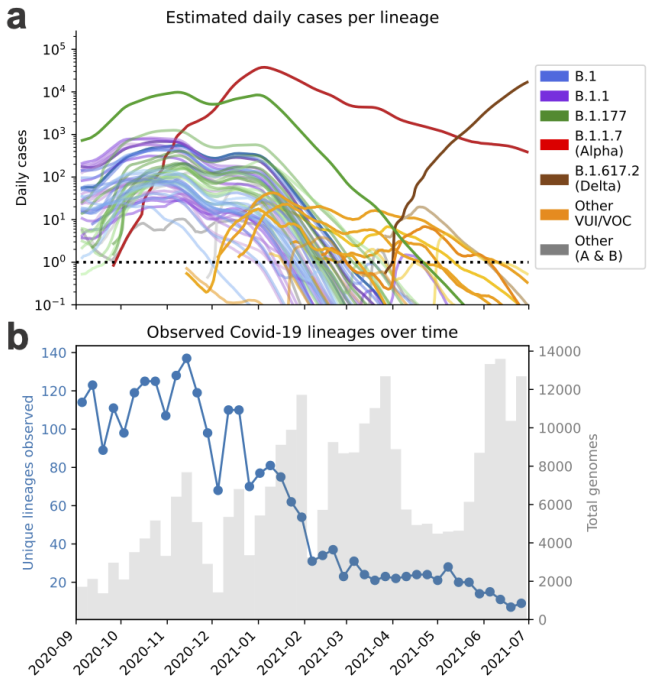



**Figure 4. Elimination of SARS-CoV-2 lineages during spring 2021. a** modelled lineage-specific incidence in England. Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **b**. Observed number of PANGO lineages per week.

## Emergence of refractory variants with E484K mutations between December 2020 and May 2021

At the same time during which many formerly dominant SARS-CoV-2 lineages were eliminated, a number of new variants were imported or emerged (**Figure 4a**). These include the VOCs B.1.351/Beta, P.1/Gamma, which carry the spike variant N501Y also found in B.1.1.7/Alpha and a similar pair of mutations (K417N/T, E484K) each shown to reduce the binding affinity of antibodies from vaccine derived or convalescent sera [21,27–30] . The ability to escape from prior immunity is consistent with the epidemiology of Beta in South Africa [8] and especially the surge of Gamma in Manaus [15]. The VUIs B.1.525/Eta, B.1.1.318 and P.2/Zeta also harbour E484K spike mutations as per their lineage definition, and sublineages of Alpha and A.23.1 acquired E484K were found in England (**Figure 5a,b**).

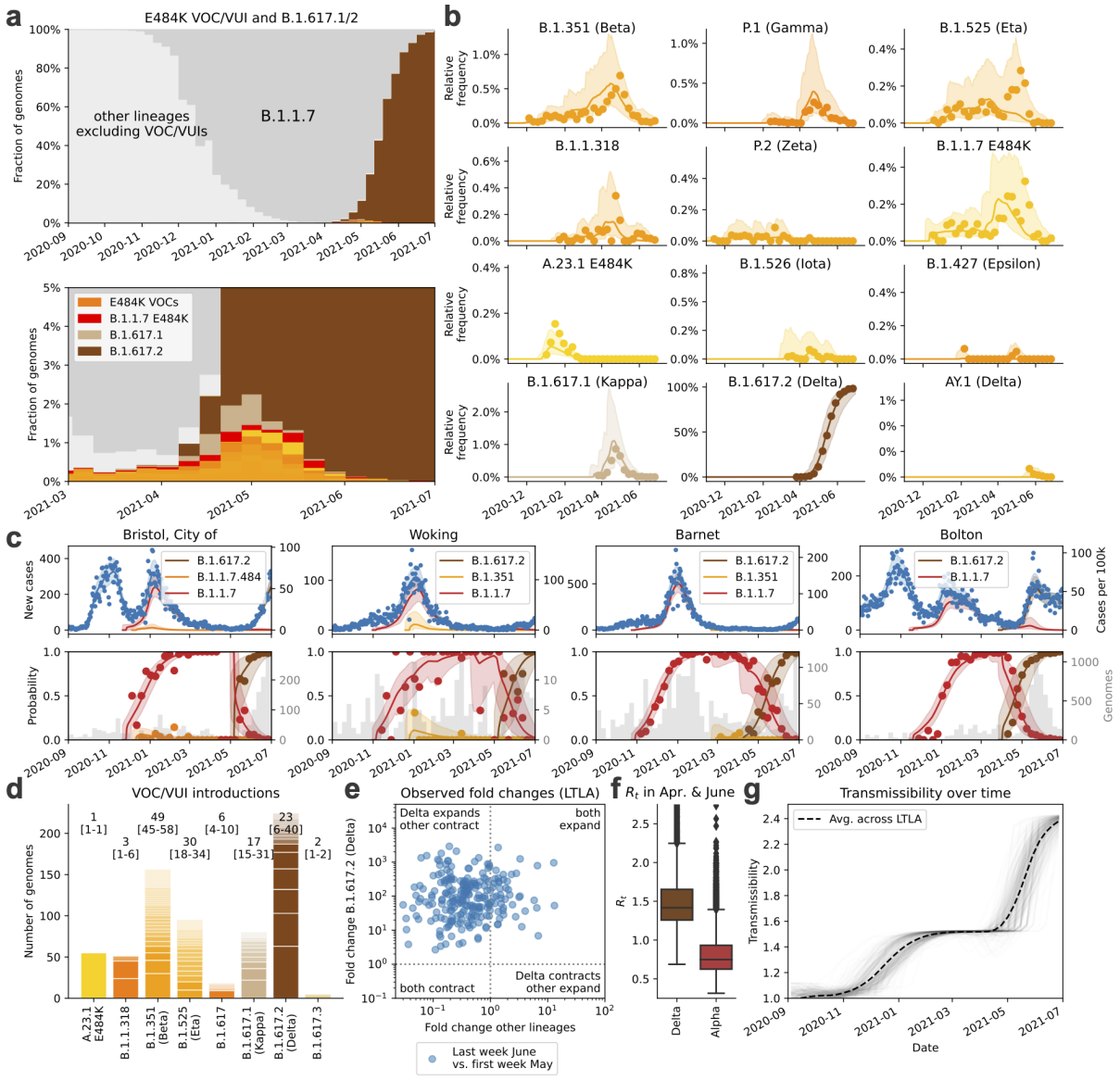



**Figure 5. Dynamics of VOC and VUIs between January and June 2021. a**. Observed relative frequency of other lineages (light grey), Alpha/B.1.1.7 (dark grey), VOC/VUIs (orange), and Delta/B.1.617.2 (brown). **b**. Observed and modelled relative frequency of VOC/VUIs in England. **c**. Total and relative lineage-specific incidence in four selected LTLAs. **d**. Estimated UK VOC/VUI clade numbers (numbers in square parentheses represent minimum and maximum numbers) and sizes. **e.** Crude growth rates (odds ratios) of Delta and Alpha between April and June 2021, as in **Fig. 3e**. **f.** , Boxplots of lineage-specific $R_t$ values in the same period, as in **Fig. 3d**. **g**. Changes of the average transmissibility across 315 LTLAs during the study period.

The proportion of these E484K containing VOCs and VUIs was consistently 0.3-0.4% from January to early April 2021. A transient rise especially of the Beta and Gamma variants was observed in May 2021 (**Figure 5a,b**). Yet the dynamics were largely stochastic and characterised by a series of individual and localised outbreaks, possibly curtailed by local surge testing efforts against Beta and Gamma variants (**Figure 5c**). Consistent with this

transient nature of the outbreaks, the estimated growth rates of these variants were typically lower than Alpha (**Figure 2a**).

Sustained imports from international travel were a critical driving mechanism behind the observed number of non-Alpha cases. A detailed phylogeographic analysis establishing the most parsimonious sets of monophyletic and exclusively domestic clades, which can be interpreted as individual introductions, confirms that A.23.1 with E484K (1 clade) is likely to have been of domestic origin as no genomes of the same clade were observed internationally (**Figure 5d**; **Extended Data Figure 7; Methods**). The estimated number of introductions was lowest for B.1.1.318 (3 introductions, range 1-6), and highest for Beta (49; range 45-58) and Eta (30; range 18-34). While our data explicitly exclude genomes sampled directly from travellers, these repeated introductions make it clear that the true growth rate due to transmission is lower than the observed increase in the number of surveillance genomes.

## Rapid rise of Delta from April to June 2021

The B.1.617.1/Kappa and B.1.617.2/Delta lineages, first detected in India in 2020, began to appear in English surveillance samples in March 2021. Unlike other VOCs, Delta/Kappa do not contain N501Y or E484K mutations, but their L452R mutation may reduce antibody recognition[28] and P681R enhances furin cleavage[31], similar to Alpha's P681H. The frequency of Delta, which harbours further spike mutations of unknown significance, increased rapidly and reached levels of 98% (12,474/12,689) on June 26, 2021 (**Figure 5a,b**). While initially constrained to a small number of large local clusters, such as in Bolton, in May 2021 (**Figure 5c**), Delta has been detected in all LTLAs by June 26 (**Figure 1c**). The sweep of Delta occurred at a rate around 59% (growth per 5.1d, CI 53-66) higher than Alpha with minor regional variation (**Figure 2a, Extended Data Figure 4e**).

The rapid rise of Delta contrasts to Kappa, which grew more slowly despite being introduced at a similar time and into a similar demographic background (**Figure 2a**, **Figure 5b**). This is also evident in the phylogeographic analysis (based on data as of May 1): Delta's 224 genomes derive from larger clades (23 introductions, range 6-40; ~10 genomes for every introduction) compared to the 80 genomes of Kappa (17 introductions, range 15-31, ~3-4 genomes per introduction) and also other VOCs/VUIs (**Figure 5d; Extended Data Figure 8**). The AY.1 lineage, derived from Delta and containing an additional K417N mutation, appeared only transiently (**Figure 5b**).

Delta's sustained domestic growth and international spread[32] relative to the Alpha lineage are first evidence of a biological growth advantage. Causes appear to be a combination of increased transmissibility and immune evasion. Evidence for higher transmissibility are the high rates of spread in younger, unvaccinated age groups, reports of elevated secondary attack rates[17] and a higher viral load[33]. Further, vaccine efficacy against infection by Delta is diminished, depending on the type of vaccine [34,35] and reinfection is more frequent[36], both supported by experimental work demonstrating reduced antibody neutralisation of Delta by vaccine derived and convalescent sera [37,38].

The higher growth rate of Delta, combined with gradual reopening and proceeding vaccination, repeated the dichotomous pattern of lineage-specific decline and growth,

however now with declining Alpha ($R_t < 1$) and growing Delta ($R_t > 1$; **Figure 5e,f**). Overall, we estimate that the spread of more transmissible variants between August 2020 and the early summer of 2021, increased the average growth rate of circulating SARS-CoV-2 in England by a factor of 2.39 (95% CI 2.25-2.42) (**Figure 5g**). Thus previously effective interventions may prove insufficient to contain newly emerging and more transmissible variants.

# Discussion

Here we reconstructed the SARS-CoV-2 epidemic in England from September 2020 to June 2021 in unprecedented genomic, spatial and temporal detail thanks to dense genomic surveillance data. Identifying lineages which consistently grew faster than others in each local authority – and thus at the same time, under the same restrictions and in a comparable population – pinpointed a series of variants with elevated transmissibility, in broad agreement with other reports [10,11,13,15,32]. We note our precise growth rate estimates have a number of limitations. The growth rates of novel and thus rare variants is stochastic due to introductions and local outbreaks. Transmission depends both on the viral variant and the immunity of the host population, which changed from less than 20% to over 90% in the study period[39]. This will influence the growth rates of VOCs/VUIs with immune evasion capabilities over time. The effect of immunity is currently not modelled, but may become more important in the future as SARS-CoV-2 becomes endemic. Further technical considerations are discussed at the end of the **Methods** section.

The third and fourth waves in England were each caused by more transmissible variants, which outgrew restrictions sufficient to suppress previous variants. During the second national lockdown, Alpha grew despite falling numbers for other lineages and, similarly, Delta took hold in April and May when cases of Alpha were falling (**Figure 4a**). The fact that such growth was initially masked by the falling cases of dominant lineages highlights the need of dense genomic surveillance and rapid analysis in order to devise optimal and timely control strategies. Such surveillance should ideally be global, as even though Delta was associated with a large wave of cases in India, its transmissibility remained unclear at the time due to a lack of systematic genomic surveillance data.

The 2.4-fold increase in growth rate during the study period as a result of new variants is also likely to have consequences for the future course of the pandemic. If this increase in growth rate was explained solely by higher transmissibility it would raise the basic reproduction number $R_0$ from a value of around 2.5-3 in the spring of 2020[40] the range of 6-7 for Delta. This is likely to spur new waves of the epidemic in countries which have so far been able to control the epidemic despite low vaccination rates and may exacerbate the situation elsewhere. Even though the exact herd immunity threshold depends on contact patterns and the distribution of immunity across age groups [41,42], it is worth considering that Delta may increase the threshold to values around 0.85. Given current estimates of vaccine efficacy [34,35,43] this would require nearly 100% vaccination coverage. Even though more than 90% of adults had antibodies against SARS-CoV-2[39] and close to 70% had received two doses of vaccination, England saw rising Delta variant cases in the first weeks of July 2021. It can thus be expected that other countries with high vaccination coverage are also likely to experience rising cases when restrictions are lifted.

SARS-CoV-2 is likely to continue its evolutionary adaptation process to humans[44]. Thus far variants with considerably higher transmissibility have had strongest positive selection, and swept through England during the 10 months of this investigation. But the possibility that an increasingly immune population may now select for variants with better immune escape highlights the need for continued systematic, and ideally global, genomic surveillance of the virus.

# Methods

## Pillar 2 SARS-CoV-2 testing data

Publicly available daily SARS-CoV-2 test result data from testing for the wider population outside the National Health Service (Pillar 2 newCasesBySpecimenDate) was downloaded from https://coronavirus.data.gov.uk/ spanning the date range from 2020-09-01 to 2021-06-30 for 315 English lower tier local authorities (downloaded on 2021-07-20). These data are mostly positive PCR tests, with about 4% of results from lateral flow tests without PCR confirmation. In this dataset, the City of London is merged with Hackney, and Isles of Scilly are merged with Cornwall due to their small number of inhabitants, thereby reducing the number of English LTLAs from 317 to 315. Population data for each LTLA was downloaded from the Office of National Statistics, https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

## SARS-CoV-2 surveillance sequencing

281,178 (Sep-June) were collected as part of random surveillance of positive tests of residents of England from four Pillar 2 Lighthouse Labs. The samples were collected between 2020-09-01 and 2021-06-26. A random selection of samples was taken, after excluding those known to be taken during quarantine of recent travellers, and samples from targeted and local surge testing efforts. The available metadata made this selection imperfect, but these samples should be an approximately random selection of infections in England during this time period, and the large sample size makes our subsequent inferences robust.

We amplified RNA extracts from these tests with $C_t < 30$ using the ARTIC amplicon protocol, https://www.protocols.io/workspaces/coguk/publications. We sequenced 384-sample pools on Illumina NovaSeq, and produced consensus fasta sequences according to the ARTIC nextflow processing pipeline https://github.com/connor-lab/ncov2019-artic-nf. Lineage assignments were made using Pangolin [5], according to the latest lineage definitions at the time, except for B.1.617, which we re-analysed after the designation of sub-lineages B.1.617.1, .2 and .3. Lineage prevalence was computed from 281,178 genome sequences. The genomes were mapped to the same 315 English LTLAs for testing data described above. Mapping was performed from outer postcodes to LTLA, which can introduce some misassignment to neighbouring LTLAs. Furthermore, lineages in each LTLA were aggregated to counts per week for a total of 43 weeks, defined beginning on Sunday and ending on Saturday.

Lastly, the complete set of 328 SARS-CoV-2 PANGO lineages was collapsed into $l = 71$ lineages using the underlying phylogenetic tree, such that each resulting lineage constituted at least 100 genomes each during the study period with the exception of VOCs or selected VUIs, which were included regardless.

## Spatio-temporal genomic surveillance model

A hierarchical Bayesian model was used to fit local incidence data in a given day in each local authority and jointly estimate the relative historic prevalence and transmission parameters. In the following $t$ denotes time and is measured in days. We use the convention that bold uppercase variables such as $\boldsymbol{B}$ are matrix-variate, usually a combination of region and lineage. Bold lowercase variables $\boldsymbol{b}$ are vector-variate.

### Motivation

Suppose that $\boldsymbol{x}'(t) = (\boldsymbol{b} + r(t)) \cdot \boldsymbol{x}(t)$ describes the ODE for the viral dynamics for a set of $l$ different lineages. Here $r(t)$ is a scalar time-dependent logarithmic growth rate thought to reflect lineage-independent transmission determinants, which changes over time in response to behavior, NPIs and immunity. This reflects a scenario where the lineages only differ in terms of the intensity of transmission, but not the inter generation time distribution. The ODE is solved by $\boldsymbol{x}(t) = e^{\boldsymbol{c}+\boldsymbol{b}t+\int_{t_0}^{t} r(t)dt} = e^{\boldsymbol{c}+\boldsymbol{b}t}\nu(t)$. The term $\nu(t)$ contributes the same factor to each lineage and therefore drops from the relative proportions of lineages $\boldsymbol{p}(t) = \boldsymbol{x}(t)/\sum \boldsymbol{x}(t) \propto e^{\boldsymbol{c}+\boldsymbol{b}t}$.

In the given model the lineage prevalence $\boldsymbol{p}(t)$ follows a multinomial logistic-linear trajectory. Moreover the total incidence factorises into $\mu(t) = \nu(t) \sum e^{\boldsymbol{c}+\boldsymbol{b}t}$, which provides a basis to separately estimate the total incidence $\mu(t)$ from Pillar 2 test data and lineage-specific prevalence $\boldsymbol{p}(t)$ from genomic surveillance data (which is taken from a varying proportion of positive tests). Exploiting the equations above one can subsequently calculate lineage-specific estimates by multiplying $\mu(t)$ with the respective genomic proportions $\boldsymbol{p}(t)$.

### Incidence

In the following we describe a flexible semi-parametric model of the incidence. Let $\boldsymbol{\mu}(t)$ be the expected daily number of positive Pillar 2 tests and $s$ the population size in each of 315 LTLAs. Denote $\boldsymbol{\lambda}(t) = \log\boldsymbol{\mu}(t) - \log\boldsymbol{s}$ the logarithmic daily incidence per capita at time $t$ in each of the 315 LTLAs.

Suppose $f'(t)$ is the daily growth rate of the epidemic, i.e., the number of new infections caused by the number of people infected at time $t$. As new cases are only noticed and tested after a delay $u$ with distribution $g$, the resulting number of cases $f^*(t)$ will be given by the convolution

$$f^*(t) = \int_0^\infty g(s)f'(t-u)du = (g * f)(t).$$

The time from infection to test is given by the incubation time plus the largely unknown distribution of the time from symptoms to test, which in England was required to take place within 5d of symptom onset. To account for these factors the log normal incubation time distribution from [45] is scaled by the equivalent of changing the mean by 2d. The convolution shifts cases approximately 6d into the future and also spreads them out according to the width of $g$ (**Extended Data Figure 2a**).

In order to parametrise the short and longer term changes of the logarithmic incidence $\boldsymbol{\lambda}(t)$, we use a combination of $h$ weekly and $k-h$ monthly cubic basis splines $\boldsymbol{f}(t) = (\boldsymbol{f}_1(t), \ldots, \boldsymbol{f}_k(t))$. The knots of the $h$ weekly splines uniformly tile the observation period except for the last 6 weeks.

Each spline basis function is convolved with the time to test distribution $g$, $\boldsymbol{f}^*(t) = (\boldsymbol{f}_1^*(t), \ldots, \boldsymbol{f}_k^*(t))$ as outlined above and used to fit the logarithmic incidence. The derivatives of the original basis $\boldsymbol{f}'(t)$ are used to calculate the underlying growth rates and $R_t$ values, as shown further below. The convolved spline basis $\boldsymbol{f}^*(t)$ is used to fit the per capita incidence in each LTLA as (**Extended Data Figure 2b**):

$$\boldsymbol{\lambda}(t) = \boldsymbol{B} \times \boldsymbol{f}^*(t).$$

This implies that fitting the incidence function for each of the $m$ local authorities is achieved by a suitable choice of coefficients $\boldsymbol{B} \in \mathbb{R}^{m \times k}$, that is one coefficient for each spline function for each of the LTLAs. The parameters $\boldsymbol{B}$ have a univariate Normal prior distribution each, which reads for LTLA $i$ and spline $j$:

$$\boldsymbol{B}_{i,j} \sim N(0, \sigma_j).$$

The standard deviation of the prior regularises the amplitude of the splines and is chosen as $\sigma_j = 0.2$ for weekly splines and $\sigma_j = 1$ for monthly splines. This choice was found to reduce the overall variance resulting from the high number of weekly splines, meant to capture rapid changes in growth rates, but which can lead to instabilities particularly at the end of the time series, when not all effects of changes in growth rates are observed yet. The less regularised monthly splines reflect trends on the scale of several weeks and are therefore subject to less noise.

Lastly, we introduce a term accounting for periodic differences in weekly testing patterns (there are typically 30% lower specimens taken on weekends; **Figure 1A**).

$$\tilde{\boldsymbol{\mu}}(t) = \boldsymbol{\mu}(t) \cdot \delta(t),$$

Where the scalar $\delta(t) = \delta(t - i \times 7) \forall i \in \mathbb{N}$ and prior distribution $\delta(t) \sim \mathrm{LogNormal}(0, 1)$ for $t = 1, \ldots, 6$ and $\delta(0) = 1$.

The total incidence was fitted to the observed number of positive daily tests $\boldsymbol{X}$ by a negative binomial with a dispersion $\omega = 10$. The overdispersion buffers against non-Poissonian uncorrelated fluctuations in the number of daily tests.

$$\boldsymbol{X}(t) \sim \mathrm{NB}(\tilde{\boldsymbol{\mu}}(t), \omega).$$

The equation above assumes that all elements of $\boldsymbol{X}(t)$ are independent, conditional on $\tilde{\boldsymbol{\mu}}(t)$ .

### Growth rates and $R_t$ values

A convenient consequence of the spline basis of $\log \boldsymbol{\mu} = \boldsymbol{\lambda}$, is that the delay-adjusted daily growth rate of the local epidemic, simplifies to:

$$\boldsymbol{\lambda}'(t) = (\boldsymbol{B} \times \boldsymbol{f}'(t))$$

where $\boldsymbol{f}'_j(t)$ represents the first derivative of the $j$th cubic spline basis function.

In order to express the daily growth rate as an approximate reproductive number $R_t$, one needs to consider the distribution of the inter generation time, which is assumed to be Gamma distributed with mean 6.3 days ($\alpha$=2.29, $\beta$=0.36) [45]. The $R_t$ value can be expressed as a Laplace transform of the inter generation time distribution [46]. Effectively, this shortens the relative time period because the exponential dynamics put disproportionally more weight on stochastically early transmissions over late ones. For reasons of simplicity and being mindful also of the uncertainties of the intergeneration time distribution, we approximate $R_t$ values by multiplying the logarithmic growth rates with a value of $\bar{\tau}_e$ = 5.1d, which was found to be a reasonable approximation to the convolution required to calculate $R_t$ values (denoted here by the lower case symbol $\boldsymbol{\rho}(t)$ in line with our convention for vector-variate symbols and to avoid confusion with the epidemiological growth rate $r_t$),

$$\log \boldsymbol{\rho}(t) \approx \frac{d \log \boldsymbol{\mu}(t)}{dt} \bar{\tau}_e = \boldsymbol{\lambda}'(t) \bar{\tau}_e$$

Hence the overall growth rate scaled to an effective inter generation time of 5.1d can be readily derived from the derivatives of the spline basis and the corresponding coefficients. The values derived from the approach are in very close agreement with those of the method of [47], but shifted according to the typical delay from infection to test (**Extended Data Figure 2b**).

### Genomic prevalence

The dynamics of the relative frequency $\boldsymbol{P}(t)$of each lineage was modelled using a logistic-linear model in each LTLA, as motivated earlier. Define the logistic prevalence of each lineage in each LTLA as $\boldsymbol{L}(t) = \operatorname{logit} \boldsymbol{P}(t)$. This is modelled using the piecewise linear expression

$$\boldsymbol{L}(t) = \boldsymbol{C} + \boldsymbol{b} \cdot \boldsymbol{t}_+$$

where $b$ may be interpreted as a lineage specific growth advantage and $\boldsymbol{C}$ as an offset term of dimension (LTLA x lineages). Time $\boldsymbol{t}_+$ is measured since introduction $\boldsymbol{t}_0$ and defined as

$$\boldsymbol{t}_+ = t - \boldsymbol{t}_0 \text{ if } t > \boldsymbol{t}_0 \text{ else } -\infty$$

and accounts for the fact that lineages can be entirely absent prior to a stochastically distributed time period preceding their first observation. This is because in the absence of such a term the absence of a lineage prior to the point of observation can only be explained by higher growth rate compared to the preceding lineages, which may not necessarily be the case. As the exact time of introduction is generally unknown a stochastic three week period of $\boldsymbol{t}_0 \sim \mathrm{Unif}(-14, 0) + \boldsymbol{t}_0^{\mathrm{obs}}$ prior to the first observation $\boldsymbol{t}_0^{\mathrm{obs}}$ was chosen.

As the inverse logit transformation projects onto the $l-1$ dimensional simplex $S_{l-1}$ and thus loses one degree of freedom, B.1.177 was set as a baseline with

$$\boldsymbol{L}_{\cdot,0}(t) = 0.$$

The offset parameters $\boldsymbol{C}$ are modelled across LTLAs as independently distributed multivariate Normal random variables with a lineage specific mean $\boldsymbol{c}$ and covariance $\Sigma = 10 \cdot I_{l-1}$, where $I_{l-1}$ denotes a $(l-1) \times (l-1)$ identity matrix. The lineage specific parameters growth rate $\boldsymbol{b}$ and average offset $\boldsymbol{c}$ are modelled using IID Normal prior distributions

$$\begin{aligned} \boldsymbol{b} &\sim \mathrm{N}(0, 0.2) \\ \boldsymbol{c} &\sim \mathrm{N}(-10, 5) \end{aligned}$$

The time-dependent relative prevalence $\boldsymbol{P}(t)$ of SARS-CoV2 lineages was fitted to the number of weekly genomes $\boldsymbol{Y}(t)$ in each LTLA by a Dirichlet-Multinomial distribution with expectation $\mathbb{E}[\boldsymbol{Y}(\boldsymbol{t})] \approx \boldsymbol{P}(\boldsymbol{t}) \cdot \boldsymbol{G}(\boldsymbol{t})$ where $\boldsymbol{G}(t)$ are the total number of genomes sequenced from each LTLA in each week. For LTLA $i$ this is defined as:

$$\boldsymbol{Y}_{i,\cdot}(t) \sim \mathrm{DirMult}(\alpha_0 + \boldsymbol{\alpha}_1 \boldsymbol{P}_{i,\cdot}(t), \boldsymbol{G}_{i,\cdot}(t)).$$

The scalar parameter $\alpha_0 = 0.01$ can be interpreted as a weak prior with expectation $1/n$, which makes the model less sensitive to the introduction of single new lineages, which can otherwise exert a very strong effect. Further, the array $\boldsymbol{\alpha}_1 = \mathrm{cases}/2$ increases the variance to account for the fact that, especially at high sequencing coverage (genomes ≈ cases), cases and thus genomes are likely to be correlated and overdispersed as they may derive from a single transmission event. Other choices such as $\alpha_1 = 1000$, which make the model converge to a standard Multinomial, leave the conclusions qualitatively unchanged. This model aspect is illustrated in **Extended Data Figure 2c.**

### Lineage-specific incidence and growth rates

From the two definitions above it follows that the lineage specific incidence is given by multiplying the total incidence in each LTLA $\boldsymbol{\mu}(t)$ with the corresponding lineage frequency estimate $\boldsymbol{P}(t)$ for lineage *j* at each time point

$$\boldsymbol{M}_{\cdot,j}(t) = \boldsymbol{\mu}(t) \cdot \boldsymbol{P}_{\cdot,j}(t) \quad \text{for } j = 0, \ldots, l-1$$

Further corresponding lineage-specific $R_t$ values $\boldsymbol{R}(t)$ in each LTLA can be calculated from the lineage agnostic average $R_t$ value $\boldsymbol{\rho}(t)$ and the lineage proportions $\boldsymbol{P}(t)$ as

$$\log \boldsymbol{R}(t) = \log \boldsymbol{\rho}(t) + \bar{\tau}_e(\boldsymbol{b} - \boldsymbol{P}(t) \times \boldsymbol{b})$$

By adding the log growth rate fold changes $\boldsymbol{b}$ and subtracting the average log growth rate change $\boldsymbol{P}(t) \times \boldsymbol{b}$ . It follows that $\boldsymbol{R}_{i,\cdot}(t) = \boldsymbol{R}_{i,0}(t)e^{\bar{\tau}_e \boldsymbol{b}}$, where $\boldsymbol{R}_{i,0}(t)$ is the $R_t$ value of the reference lineage $j$ = 0 (for which $\boldsymbol{b}_0$ = 0) in LTLA $i$. It follows that all other lineage-specific the $R_t$ values are proportional to this baseline at any given point in time with factor $e^{\bar{\tau}_e \boldsymbol{b}}$.

### Inference

The model was implemented in numpyro [48,49] and fitted using stochastic variational inference [50]. Guide functions were multivariate normal distributions for each row (corresponding to an LTLA) of $\boldsymbol{B}$, $\boldsymbol{C}$ to preserve the correlations across lineages and time as well as for $(\boldsymbol{b}, \boldsymbol{c})$ to also model correlations between growth rates and typical introduction.

## Phylogeographic analyses

To infer VOC introduction events into the UK and corresponding clade sizes, we investigated VOC genome sequences from GISAID https://www.gisaid.org/ available from any country. We downloaded multiple sequence alignments of genome sequences with release dates 17-04-2021 (for the analysis of lineages A.23.1, B.1.1.318, B.1.351, B.1.525) and 05-05-2021 (for the analysis of B.1.617 sublineages). We then extracted a sub-alignment from each lineage (following the 01-04-2021 version of PANGOlin for the 17-04-2021 alignment and the 23-04-2021 version of PANGOlin for the 05-05-2021 alignment), and, for each sub-alignment, we inferred a phylogeny via maximum likelihood using FastTree2 version 2.1.11 [51] with default options and GTR substitution model [52].

On each VOC/VUI phylogeny we inferred the minimum and maximum number of introductions of the considered SARS-CoV-2 lineage into the UK compatible with a parsimonious migration history of the ancestors of the considered samples; we also measure clade sizes for one specific example parsimonious migration history. We only count introduction events into the UK that result in at least one descendant from the set of UK samples that we consider in this work for our hierarchical Bayesian model; similarly, we measure clade sizes by the number of UK samples considered here included in such clades. Multiple occurrences of identical sequences were counted as separate cases, since this helps us identify rapid SARS-CoV-2 spread.

When using parsimony, we only consider migration histories along a phylogenetic tree that are parsimonious in terms of the number of migration events from and to the UK (in practice we collapse all the non-UK locations into a single one). Also, since SARS-CoV-2 phylogenies present substantial numbers of polytomies, that is, phylogenetic nodes where the tree topology cannot be reconstructed due to lack of mutation events on certain branches, we developed a tailored dynamic programming approach to efficiently integrate over all possible splits of polytomies and over all possible parsimonious migration histories. The idea of this method is somewhat similar to typical Bayesian phylogeographic inference (e.g. [53]) in that it allows us to at least in part integrate over phylogenetic uncertainty and

uncertainty in migration history; however, it also represents a very simplified version of these analyses, more so than [16], as it considers most of the phylogenetic tree as fixed, ignores sampling times, and uses parsimony instead of a likelihood-based approach. Parsimony is expected to represent a good approximation in the context of SARS-CoV-2, due to the shortness (both in time and substitutions) of the phylogenetic branches considered [54,55]. The main advantage of our approach is that, thanks to the dynamic programming implementation, it is more computationally efficient than Bayesian alternatives, as the most computationally demanding step is the inference of the maximum likelihood phylogenetic tree. This allows us to infer plausible ranges for numbers of introduction events for large datasets and to quickly update our analyses as new sequences become available. The other advantage of this approach is that it allows us to easily customize the analysis and to focus on inferred UK introductions that result in at least one UK surveillance sample, while still making use of non-surveillance UK samples to inform the inferred phylogenetic tree and migration history. Note that possible biases due to uneven sequencing rates across the world (e.g. [54]) apply to our approach as well as other popular phylogeographic methods.

Our approach works by traversing the maximum likelihood tree starting from the terminal nodes and ending at the root (postorder traversal). Here, we define a “UK clade” as a maximal subtree of the total phylogeny for which all terminal nodes are from the UK, all internal nodes are inferred to be from the UK, and at least one terminal node is a UK surveillance sample; the size of a UK clade is defined as the number of UK surveillance samples in it. At each node, using values already calculated for all children nodes (possibly more than two children in the case of a multifurcation), we calculate the following quantities: i) the maximum and minimum number of possible descendant UK clades of the current node, over the space of possible parsimonious migration histories, and conditional on the current node being UK or non-UK; ii) the number of migration events compatible with a parsimonious migration history in the subtree below the current node, and conditional on the current node being UK or non-UK; iii) the size so far of the UK clade the current node is part of, conditional on it being UK; iv) A sample of UK clade sizes for the subtree below the node. To calculate these quantities, for each internal node, and conditional on each possible node state (UK or non-UK), we consider the possible scenarios of having 0 of 1 migration event between the internal node and its children nodes (migration histories with more than 1 migration event between the node and its children are surely not parsimonious in our analysis and can be ignored).

To confirm the results of our analyses based on parsimony, we have also used the new Bayesian phylogenetic approach Thorney BEAST[16] (https://beast.community/thorney_beast) for VOCs for which it was computationally feasible, that is, excluding B.1.351. For each VOC, we used in Thorney BEAST the same topology inferred with FastTree2 as for our parsimony analysis; in addition, we used treetime[56] 0.8.2 to estimate a timed tree and branch divergences for use in Thorney BEAST. We used a 2-state (“UK” and “non-UK”) mugration model[53] of migration to infer introductions into the UK, but again, only counted, from the posterior sample trees, UK clades with at least one UK surveillance sample. We used a Skygrid[57] tree coalescent prior with 6 time intervals. The comparison of parsimony and Bayesian estimates is shown in **Extended Data Figure 8d**.

## ONS infection survey analysis

Data from the cross sectional infection survey was downloaded from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/30april2021.

Comparison of ONS incidence estimates with hospitalisation, case and death rates was conducted by estimating infection trajectories separately from observed cases, hospitalisations and deaths [58,59], convolving them with estimated PCR detection curves [60], and dividing the resulting PCR prevalence estimates by the estimated prevalence from the ONS Community Infection Survey at the midpoints of the 2-week intervals over which prevalence was reported in the survey.

## Limitations

A main limitation of the model is that the underlying transmission dynamics are deterministic and stochastic growth dynamics are only accounted for in terms of (uncorrelated) overdispersion. For that reason the estimated growth rates may not accurately reflect the viral transmissibility, especially a low prevalence. While the logistic growth assumption is a consistent estimator of the average transmission dynamics, individual outbreaks may deviate from these dynamics and therefore provide unreliable estimates. It is therefore important to assess whether consistent growth patterns in multiple independent areas are observed.

In its current form the model only accounts for a single introduction event per LTLA. While this problem is in part alleviated by the high spatial resolution, which spreads introductions across 315 LTLAs, it is important to investigate whether sustained introductions inflate the observed growth rates, as in the case of the Delta variant or other VOCs and VUIs. This can be achieved by a more detailed phylogeographic assessment and the assessment of monophyletic sublineages.

Furthermore there is no explicit transmission modelled from one LTLA to another. As each introduction is therefore modelled separately, this makes the model conservative in ascertaining elevated transmission as single observed cases across different LTLAs can be explained by their introduction.

The inferred growth rates also cannot identify a particular mechanism which may be caused by higher viral load, longer infectivity or greater susceptibility. Lineages could potentially differ by their inter-generation time, which would lead to a non linear scaling. Here we did not find convincing evidence in incidence data for such effects. in contrast to previous reports [24]. However, contact tracing data indicates that the inter-generation time may be shortening for more transmissible lineages such as Delta[33,61].

Also lineages, such as Beta, Gamma or Delta differ in their ability to evade prior immunity. As immunity changes over time, this might lead to a differential growth advantage over time. It is therefore advisable to assess whether a growth advantage is constant over periods in which immunity changes considerably.

A further limitation underlies the nature of lineage definition and assignment. The PANGO lineage definition[5] assigns lineages to geographic clusters, which have by definition expanded, which can induce a certain survivor bias, often followed by winner's curse. Another issue results from the fact that very recent variants may not be classified as a lineage despite having grown, which can inflate the growth rate of ancestral lineages over sublineages.

As the total incidence is modelled based on the total number of positive PCR tests it may be influenced by testing capacity with the total number of tests having approximately tripled between September 2020 and March 2021. This can potentially lead to a time trend in recorded cases and thus baseline $R_t$ values if the access to testing changed, e.g. by too few available tests during high incidence, or changes to the eligibility to test with fewer symptoms intermittently. Generally, the observed incidence was in good agreement with representative cross-sectional estimates from the Office of National Statistics [62,63], except for a period of peak incidence from late December 2020 to January 2021 (**Extended Data Figure 1d**). Values after March 8, 2021 need to be interpreted with caution as pillar 2 PCR testing was supplemented by lateral flow devices, which increased the number of daily tests to more than 1.5 million.

The modelled curves are smoothed over intervals of approximately 7 days using cubic splines, creating a possibility that later time points influence the period of investigation and cause a certain waviness of the $R_t$ value pattern. An alternative parameterization using piecewise linear basis functions per week (i.e., constant $R_t$ values per week) leaves the overall conclusions and extracted parameters broadly unchanged.

## Code availability

Code for spatio-temporal modeling of different viral lineage is available at https://github.com/gerstung-lab/genomicsurveillance and as a PyPI package (genomicsurveillance). This phylogeographic model has been implemented in python scripts, and the code is available from https://github.com/NicolaDM/phylogeographySARS-CoV-2. Code for ONS infection survey analysis is available at https://github.com/jhellewell14/ons_severity_estimates.

## Data availability

PCR test data are publicly available at https://coronavirus.data.gov.uk/. SARS-CoV-2 genome data and geolocations can be obtained under controlled access from https://www.cogconsortium.uk/data/. A filtered, privacy conserving version of the data set is publicly available at https://covid19.sanger.ac.uk/downloads. The data and a version of the analysis with fewer lineages can be interactively explored at https://covid19.sanger.ac.uk.

## Acknowledgements

COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. We would like to

thank our colleagues at EMBL-EBI, the Wellcome Sanger Institute and from COG-UK for stimulating discussions and helpful comments on this manuscript. HSV and MG are supported by a grant from the Department of Health and Social Care. AWJ, EB and MG are beneficiaries from grant NNF17OC0027594 from the Novo Nordisk Foundation. TS is supported by grant 210918/Z/18/Z, and JH and SF by grant 210758/Z/18/Z from the Wellcome Trust. HSV, NDM, AWJ, NG, EB and MG are supported by EMBL. We would like to thank Elias Allara (Cambridge) and Georgia Whitton (Sanger) for providing outer postcode to LTLA mappings, Rupert Beale for comments and John McCrone for setting up Thorney Beast analysis. We thank all the contributors who submitted genome sequences to GISAID. Acknowledgement tables for individual sequences are deposited at https://github.com/NicolaDM/phylogeographySARS-CoV-2.

## Conflicts of Interest

None declared.

## Ethical approval

This study was done as part of surveillance for COVID-19 under the auspices of Section 251 of the National Health Service Act 2006. It therefore did not require individual patient consent or ethical approval. The COVID-19 Genomics UK (COG-UK) study protocol was approved by the Public Health England Research Ethics Governance Group.

## Author contributions

HSV and MG developed the analysis code, which HSV implemented with input from AWJ. HSV created most Figures. MS analysed, annotated and aggregated viral genome data. NDM conducted phylogeographic analyses supervised by NG. TS, RG, MS, and HSV developed the interactive spatiotemporal viewer. TN, FS, IH, RA, CA, SG, DJ, IJ, CS, JS, TS, MS analysed genomic surveillance data under supervision of DK, MC, IM and JCB. JH and SF analysed ONS data and helped with epidemiological modeling and data interpretation. EV analysed growth rates and helped with data interpretation. EB and JPG supervised HSV and helped with data interpretation. JCB and MG supervised the analysis with advice from IM. MG, HSV, MS, NDM, TS, IM and JCB wrote the manuscript with input from all co-authors.

## References

1. Rambaut, A. Phylogenetic analysis of nCoV-2019 genomes. (2020). at <https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356>
2. Nextstrain team. Genomic epidemiology of novel coronavirus - Global subsampling. (2020). at <https://nextstrain.org/ncov/global?l=clock>

3. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. & Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34,** 4121–4123 (2018).
4. Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., Rey, S. M., Nicholls, S. M., Colquhoun, R. M., da Silva Filipe, A., Shepherd, J., Pascall, D. J., Shah, R., Jesudason, N., Li, K., Jarrett, R., Pacchiarini, N., Bull, M., Geidelberg, L., Siveroni, I., COG-UK Consortium, Goodfellow, I., Loman, N. J., Pybus, O. G., Robertson, D. L., Thomson, E. C., Rambaut, A. & Connor, T. R. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184,** 64–75.e11 (2021).
5. Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L. & Pybus, O. G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5,** 1403–1407 (2020).
6. O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., du Plessis, L., Aanensen, D., Holmes, E., Pybus, O. & Rambaut, A. Global Report Investigating Novel Coronavirus Haplotypes. (2021). at <https://cov-lineages.org/global_report.html>
7. Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H. D., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., Bloom, J. D., Veesler, D., Mateo, D., Hernando, A., Comas, I., González-Candelas, F., Stadler, T. & Neher, R. A. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595,** 707–712 (2021).
8. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N. & Others. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* (2020). at <https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1.full>
9. Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T.,

Peacock, T., Robertson, D. L., Volz, E. & on behalf of COVID-19 Genomics Consortium UK. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. (2020). at <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>

10. Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D. K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D. P., COVID-19 Genomics UK (COG-UK) consortium, Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A. & Ferguson, N. M. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 1–17 (2021).
11. Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., Group1‡, C. C.-19 W., COVID-19 Genomics UK (COG-UK) Consortium‡, Diaz-Ordaz, K., Keogh, R., Eggo, R. M., Funk, S., Jit, M., Atkins, K. E. & John Edmunds, W. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372,** (2021).
12. O'Toole, Á., Hill, V., Pybus, O. G., Watts, A., Bogoch, I. I., Khan, K., Messina, J. P., The COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, Tegally, H., Lessells, R. R., Giandhari, J., Pillay, S., Tumedi, K. A., Nyepetsi, G. & Others. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. (2021). at <https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592>
13. Washington, N. L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E. T., Schiabor Barrett,

K. M., Larsen, B. B., Anderson, C., White, S., Cassens, T., Jacobs, S., Levan, G., Nguyen, J., Ramirez, J. M., 3rd, Rivera-Garcia, C., Sandoval, E., Wang, X., Wong, D., Spencer, E., Robles-Sikisaka, R., Kurzban, E., Hughes, L. D., Deng, X., Wang, C., Servellita, V., Valentine, H., De Hoff, P., Seaver, P., Sathe, S., Gietzen, K., Sickler, B., Antico, J., Hoon, K., Liu, J., Harding, A., Bakhtar, O., Basler, T., Austin, B., MacCannell, D., Isaksson, M., Febbo, P. G., Becker, D., Laurent, M., McDonald, E., Yeo, G. W., Knight, R., Laurent, L. C., de Feo, E., Worobey, M., Chiu, C. Y., Suchard, M. A., Lu, J. T., Lee, W. & Andersen, K. G. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184,** 2587–2594.e7 (2021).

14. Faria, N. R., Claro, I. M., Candido, D., Moyses Franco, L. A., Andrade, P. S., Coletti, T. M., Silva, C. A. M., Sales, F. C., Manuli, E. R., Aguiar, R. S. & Others. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological* (2021). at <https://www.icpcovid.com/sites/default/files/2021-01/Ep%20102-1%20Genomic%20characterisation%20of%20an%20emergent%20SARS-CoV-2%20lineage%20in%20Manaus%20Genomic%20Epidemiology%20-%20Virological.pdf>

15. Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., Pereira, R. H. M., Peixoto, P. S., Kraemer, M. U., Gaburo, N., Camilo, C. da C., Hoeltgebaum, H., Souza, W. M., Rocha, E. C., de Souza, L. M., de Pinho, M. C., Araujo, L. J. T., Malta, F. S. V., de Lima, A. B., Silva, J. do P., Zauli, D. A. G., Ferreira, A. C. de S., Schnekenberg, R. P., Laydon, D. J., Walker, P. G. T., Schlüter, H. M., Dos Santos, A. L. P., Vidal, M. S., Del Caro, V. S., Filho, R. M. F., Dos Santos, H. M., Aguiar, R. S., Modena, J. L. P., Nelson, B., Hay, J. A., Monod, M., Miscouridou, X., Coupland, H., Sonabend, R., Vollmer, M., Gandy, A., Suchard, M. A., Bowden, T. A., Pond, S. L. K., Wu, C.-H., Ratmann, O., Ferguson, N. M., Dye, C., Loman, N. J., Lemey, P., Rambaut, A., Fraiji, N. A., Carvalho, M. do P. S. S., Pybus, O. G., Flaxman, S., Bhatt,

S. & Sabino, E. C. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. *medRxiv* (2021). doi:10.1101/2021.02.26.21252554

16. du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., Faria, N. R., Jackson, B., Loman, N. J., O'Toole, Á., Nicholls, S. M., Parag, K. V., Scher, E., Vasylyeva, T. I., Volz, E. M., Watts, A., Bogoch, I. I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium†, Aanensen, D. M., Kraemer, M. U. G., Rambaut, A. & Pybus, O. G. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371,** 708–712 (2021).
17. Public Health England. Investigation of novel SARS-CoV-2 variants of concern. Technical briefing 10. (2021). at <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>
18. Danish Covid-19 Genome Consortium. Genomic overview of SARS-CoV-2 in Denmark. (2021). at <https://www.covid19genomics.dk/statistics>
19. Kraemer, M. U. G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J. T., Baele, G., Parag, K. V., Battle, A. L., Gutierrez, B., Jackson, B., Colquhoun, R., O'Toole, Á., Klein, B., Vespignani, A., The COVID-19 Genomics UK (CoG-UK) consortium‡, Volz, E., Faria, N. R., Aanensen, D., Loman, N. J., du Plessis, L., Cauchemez, S., Rambaut, A., Scarpino, S. V. & Pybus, O. G. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* (2021). doi:10.1126/science.abj0113
20. Park, S. W., Bolker, B. M., Funk, S., Metcalf, C. J. E., Weitz, J. S., Grenfell, B. T. & Dushoff, J. Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. *medRxiv* (2021). at <https://www.medrxiv.org/content/10.1101/2021.05.03.21256545v1.abstract>
21. Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D. & Bloom, J. D. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182,** 1295–1310.e20 (2020).

22. Zahradník, J., Marciano, S., Shemesh, M., Zoler, E., Chiaravalli, J., Meyer, B., Rudich, Y., Dym, O., Elad, N. & Schreiber, G. SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv* 2021.01.06.425392 (2021). doi:10.1101/2021.01.06.425392
23. Brown, J. C., Goldhill, D. H., Zhou, J., Peacock, T. P., Frise, R., Goonawardane, N., Baillon, L., Kugathasan, R., Pinto, A. L., McKay, P. F., Hassard, J., Moshe, M., Singanayagam, A., Burgoyne, T., the ATACCC Investigators, PHE Virology Consortium & Barclay, W. S. Increased transmission of SARS-CoV-2 lineage B.1.1.7 (VOC 2020212/01) is not accounted for by a replicative advantage in primary airway cells or antibody escape. *bioRxiv* 2021.02.24.432576 (2021). doi:10.1101/2021.02.24.432576
24. Vöhringer, H., Sinnott, M., Amato, R., Martincorena, I., Kwiatkowski, D., Barrett, J. C., Gerstung, M. & on behalf of The COVID-19 Genomics UK (COG-UK) consortium. Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown. *virological.org* (2020). at <https://virological.org/t/lineage-specific-growth-of-sars-cov-2-b-1-1-7-during-the-english-national-lockdown/575/2>
25. Wikipedia contributors. The Health Protection (Coronavirus, Restrictions) (All Tiers) (England) Regulations 2020. *Wikipedia, The Free Encyclopedia* (2021). at <https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)_(All_Tiers)_(England)_Regulations_2020&oldid=1014831173>
26. Davies, K. S. A. Coronavirus (COVID-19) Infection Survey, antibody and vaccination data for the UK - Office for National Statistics. (2021). at <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveyantibodydatafortheuk/28april2021>
27. Greaney, A. J., Loes, A. N., Crawford, K. H. D., Starr, T. N., Malone, K. D., Chu, H. Y. & Bloom, J. D. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *bioRxiv*

2020.12.31.425021 (2021). doi:10.1101/2020.12.31.425021

28. Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., Hilton, S. K., Huddleston, J., Eguia, R., Crawford, K. H. D., Dingens, A. S., Nargi, R. S., Sutton, R. E., Suryadevara, N., Rothlauf, P. W., Liu, Z., Whelan, S. P. J., Carnahan, R. H., Crowe, J. E. & Bloom, J. D. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29,** 44–57.e9 (2021).
29. Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., Zhao, Y., Duyvesteyn, H. M. E., Tuekprakhon, A., Nutalai, R., Wang, B., Paesen, G. C., Lopez-Camacho, C., Slon-Campos, J., Hallis, B., Coombes, N., Bewley, K., Charlton, S., Walter, T. S., Skelly, D., Lumley, S. F., Dold, C., Levin, R., Dong, T., Pollard, A. J., Knight, J. C., Crook, D., Lambe, T., Clutterbuck, E., Bibi, S., Flaxman, A., Bittaye, M., Belij-Rammerstorfer, S., Gilbert, S., James, W., Carroll, M. W., Klenerman, P., Barnes, E., Dunachie, S. J., Fry, E. E., Mongkolsapaya, J., Ren, J., Stuart, D. I. & Screaton, G. R. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* (2021). doi:10.1016/j.cell.2021.02.037
30. Planas, D., Bruel, T., Grzelak, L., Guivel-Benhassine, F., Staropoli, I., Porrot, F., Planchais, C., Buchrieser, J., Rajah, M. M., Bishop, E. & Others. Sensitivity of infectious SARS-CoV-2 B. 1.1. 7 and B. 1.351 variants to neutralizing antibodies. *Nat. Med.* 1–8 (2021).
31. Peacock, T. P., Sheppard, C. M., Brown, J. C., Goonawardane, N., Zhou, J., Whiteley, M., PHE Virology Consortium, de Silva, T. I. & Barclay, W. S. The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin. *bioRxiv* 2021.05.28.446163 (2021). doi:10.1101/2021.05.28.446163
32. Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., Pavlin, B., Vandemaele, K., Van Kerkhove, M. D., Jombart, T., Morgan, O. & le Polain de Waroux, O. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26,** (2021).

33. Li, B., Deng, A., Li, K., Hu, Y., Li, Z., Xiong, Q., Liu, Z., Guo, Q., Zou, L., Zhang, H., Zhang, M., Ouyang, F., Su, J., Su, W., Xu, J., Lin, H., Sun, J., Peng, J., Jiang, H., Zhou, P., Hu, T., Luo, M., Zhang, Y., Zheng, H., Xiao, J., Liu, T., Che, R., Zeng, H., Zheng, Z., Huang, Y., Yu, J., Yi, L., Wu, J., Chen, J., Zhong, H., Deng, X., Kang, M., Pybus, O. G., Hall, M., Lythgoe, K. A., Li, Y., Yuan, J., He, J. & Lu, J. Viral infection and Transmission in a large well-traced outbreak caused by the Delta SARS-CoV-2 variant. *bioRxiv* (2021). doi:10.1101/2021.07.07.21260122

34. Nasreen, S., Chung, H., He, S., Brown, K. A., Gubbay, J. B., Buchan, S. A., Fell, D. B., Austin, P. C., Schwartz, K. L., Sundaram, M. E., Calzavara, A., Chen, B., Tadrous, M., Wilson, K., Wilson, S. E. & Kwong, J. C. Effectiveness of COVID-19 vaccines against variants of concern in Ontario, Canada. *bioRxiv* (2021). doi:10.1101/2021.06.28.21259420

35. Lopez Bernal, J., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., Stowe, J., Tessier, E., Groves, N., Dabrera, G., Myers, R., Campbell, C. N. J., Amirthalingam, G., Edmunds, M., Zambon, M., Brown, K. E., Hopkins, S., Chand, M. & Ramsay, M. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* (2021). doi:10.1056/NEJMoa2108891

36. Public Health England. Investigation of novel SARS-CoV-2 variants of concern. Technical briefing 19. (2021). at <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1005517/Technical_Briefing_19.pdf>

37. Ferreira, I., Datir, R., Papa, G., Kemp, S., Meng, B., Rakshit, P., Singh, S., Pandey, R., Ponnusamy, K., Radhakrishnan, V. S., INSACOG CONSORTIUM, COG-UK CONSORTIUM, Sato, K., James, L., Aggarwal, A. & Gupta, R. K. SARS-CoV-2 B.1.617 emergence and sensitivity to vaccine-elicited antibodies. *bioRxiv* 2021.05.08.443253 (2021). doi:10.1101/2021.05.08.443253

38. Wall, E. C., Wu, M., Harvey, R., Kelly, G., Warchal, S., Sawyer, C., Daniels, R., Hobson, P., Hatipoglu, E., Ngai, Y., Hussain, S., Nicod, J., Goldstone, R., Ambrose, K.,

Hindmarsh, S., Beale, R., Riddell, A., Gamblin, S., Howell, M., Kassiotis, G., Libri, V., Williams, B., Swanton, C., Gandhi, S. & Bauer, D. L. Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination. *Lancet* **397,** 2331–2333 (2021).

39. Haughton, K. S. A. Coronavirus (COVID-19) Infection Survey, antibody and vaccination data, UK - Office for National Statistics. (2021). at <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveyantibodyandvaccinationdatafortheuk/21july2021>
40. Anderson, R., Donnelly, C., Hollingsworth, D., Keeling, M., Vegvari, C., Baggaley, R. & Maddren, R. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. *The Royal Society* **2020,** (2020).
41. Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369,** 846–849 (2020).
42. Funk, S., Knapp, J. K., Lebo, E., Reef, S. E., Dabbagh, A. J., Kretsinger, K., Jit, M., Edmunds, W. J. & Strebel, P. M. Combining serological and contact data to derive target immunity levels for achieving and maintaining measles elimination. *BMC Med.* **17,** 180 (2019).
43. Hodgson, D., Flasche, S., Jit, M., Kucharski, A. J. & CMMID COVID-19 Working Group. The potential for vaccination-induced herd immunity against the SARS-CoV-2 B.1.1.7 variant. *Eurosurveillance* **26,** 2100428 (2021).
44. van Dorp, L., Houldcroft, C. J., Richard, D. & Balloux, F. COVID-19, the first pandemic in the post-genomic era. *Curr. Opin. Virol.* (2021). doi:10.1016/j.coviro.2021.07.002
45. Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., Gao, W., Cheng, C., Tang, X., Wu, X., Wu, Y., Sun, B., Huang, S., Sun, Y., Zhang, J., Ma, T., Lessler, J. & Feng, T. Epidemiology and transmission of COVID-19 in

391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20,** 911–919 (2020).

46. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. Biol. Sci.* **274,** 599–604 (2007).
47. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178,** 1505–1512 (2013).
48. Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. & Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *arXiv [cs.LG]* (2018). at <http://arxiv.org/abs/1810.09538>
49. Phan, D., Pradhan, N. & Jankowiak, M. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv [stat.ML]* (2019). at <http://arxiv.org/abs/1912.11554>
50. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic Variational Inference. *J. Mach. Learn. Res.* **14,** 1303–1347 (2013).
51. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).
52. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17,** 57–86 (1986).
53. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian Phylogeography Finds Its Roots. *PLoS Comput. Biol.* **5,** e1000520 (2009).
54. De Maio, N., Wu, C.-H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet.* **11,** e1005421 (2015).
55. Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D. & Corbett-Detig, R. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53,** 809–816 (2021).
56. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic

analysis. *Virus Evol* **4,** vex042 (2018).

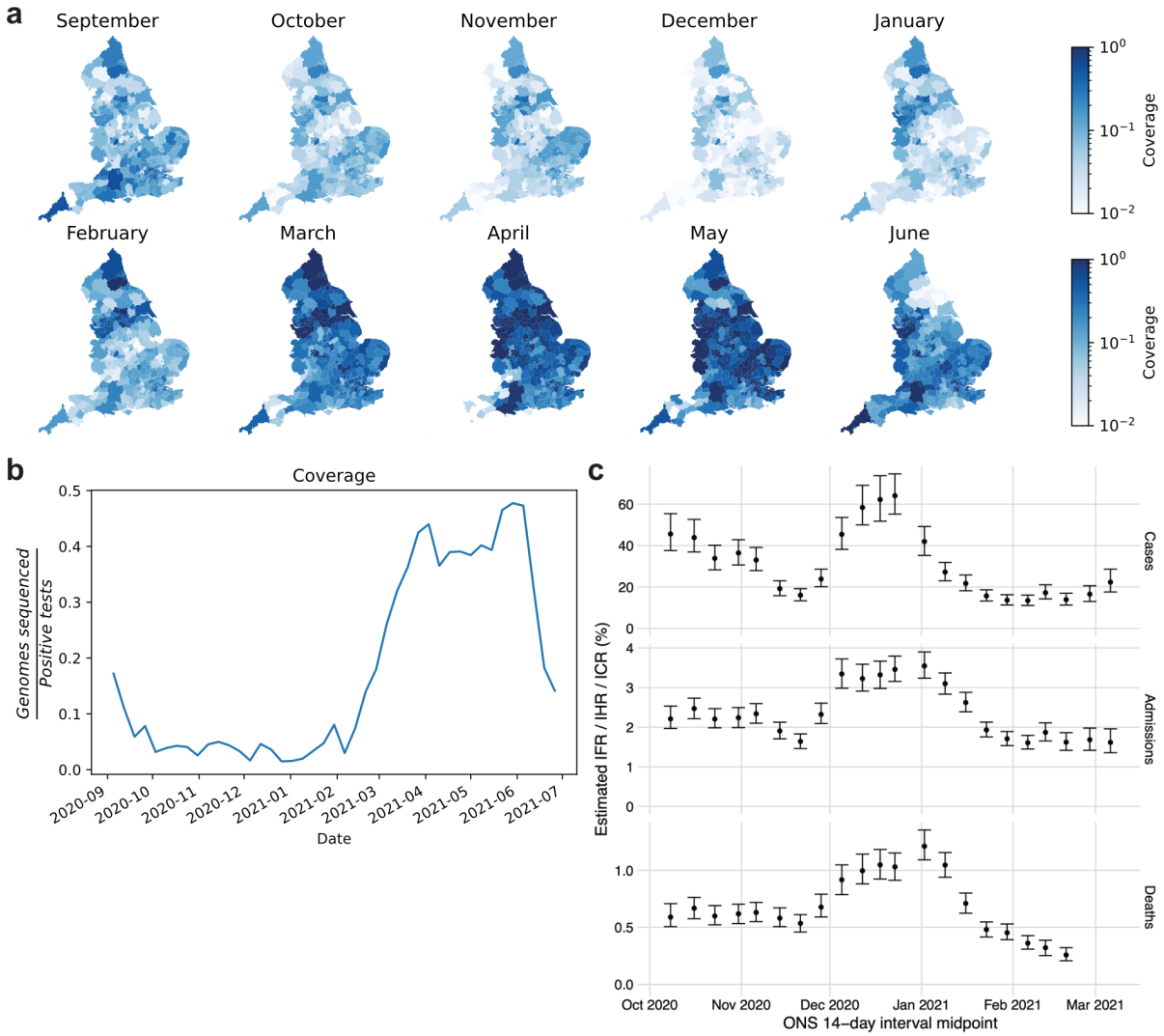
57. Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B. & Suchard, M. A. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2013).
58. Sherratt, K., Abbott, S., Meakin, S. R., Hellewell, J., Munday, J. D., Bosse, N., Jit, M., Funk, S. & CMMID Covid-19 working group. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of Covid-19 in England. *bioRxiv* (2020). doi:10.1101/2020.10.18.20214585
59. Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., Chan, Y.-W. D., Finger, F., Campbell, P., Endo, A., Pearson, C. A. B., Gimma, A., Russell, T., Flasche, S., Kucharski, A. J., Eggo, R. M., Funk, S. & CMMID COVID modelling group. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* **5,** 112 (2020).
60. Hellewell, J., Russell, T. W., SAFER Investigators and Field Study Team, Crick COVID-19 Consortium, CMMID COVID-19 working group, Beale, R., Kelly, G., Houlihan, C., Nastouli, E. & Kucharski, A. J. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Med.* **19,** 106 (2021).
61. Hart, W. S., Abbott, S., Endo, A., Hellewell, J., Miller, E., Andrews, N., Maini, P. K., Funk, S. & Thompson, R. N. Inference of SARS-CoV-2 generation times using UK household data. *medRxiv* (2021). at <https://www.medrxiv.org/content/10.1101/2021.05.27.21257936v1.abstract>
62. Pouwels, K. B., House, T., Pritchard, E., Robotham, J. V., Birrell, P. J., Gelman, A., Vihta, K.-D., Bowers, N., Boreham, I., Thomas, H., Lewis, J., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Benton, P., Walker, A. S. & COVID-19 Infection Survey Team. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Public Health* **6,** e30–e38

(2021).

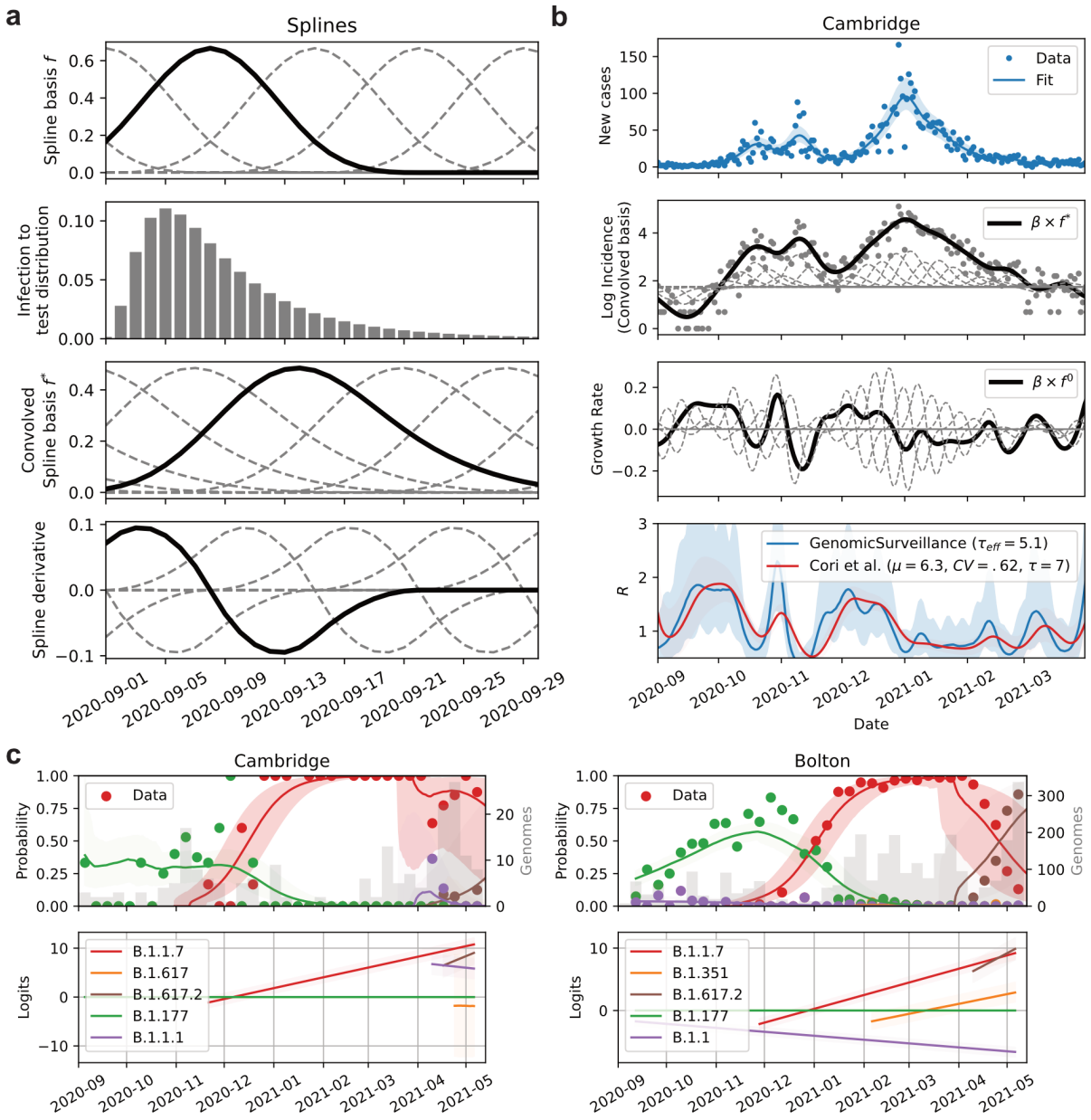
63. Donnarumma, K. S. A. Coronavirus (COVID-19) infection survey, UK - office for national statistics. (2021). at <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/23april2021>

64. Wikipedia contributors. The Health Protection (Coronavirus, Restrictions) (England) (No. 4) Regulations 2020. *Wikipedia, The Free Encyclopedia* (2021). at <https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)_(England)_(No._4)_Regulations_2020&oldid=1014701607>

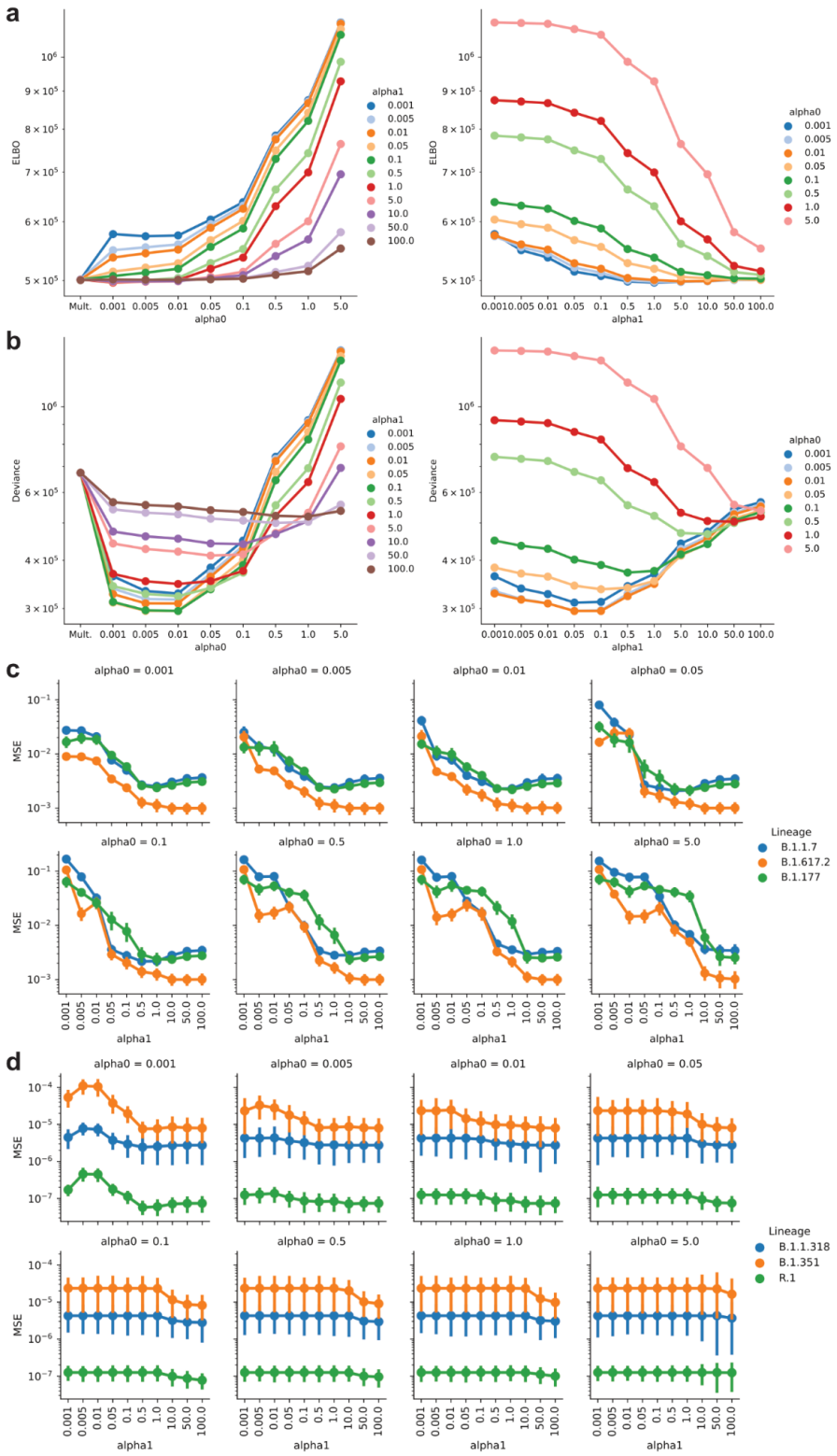
# Extended Data Figures



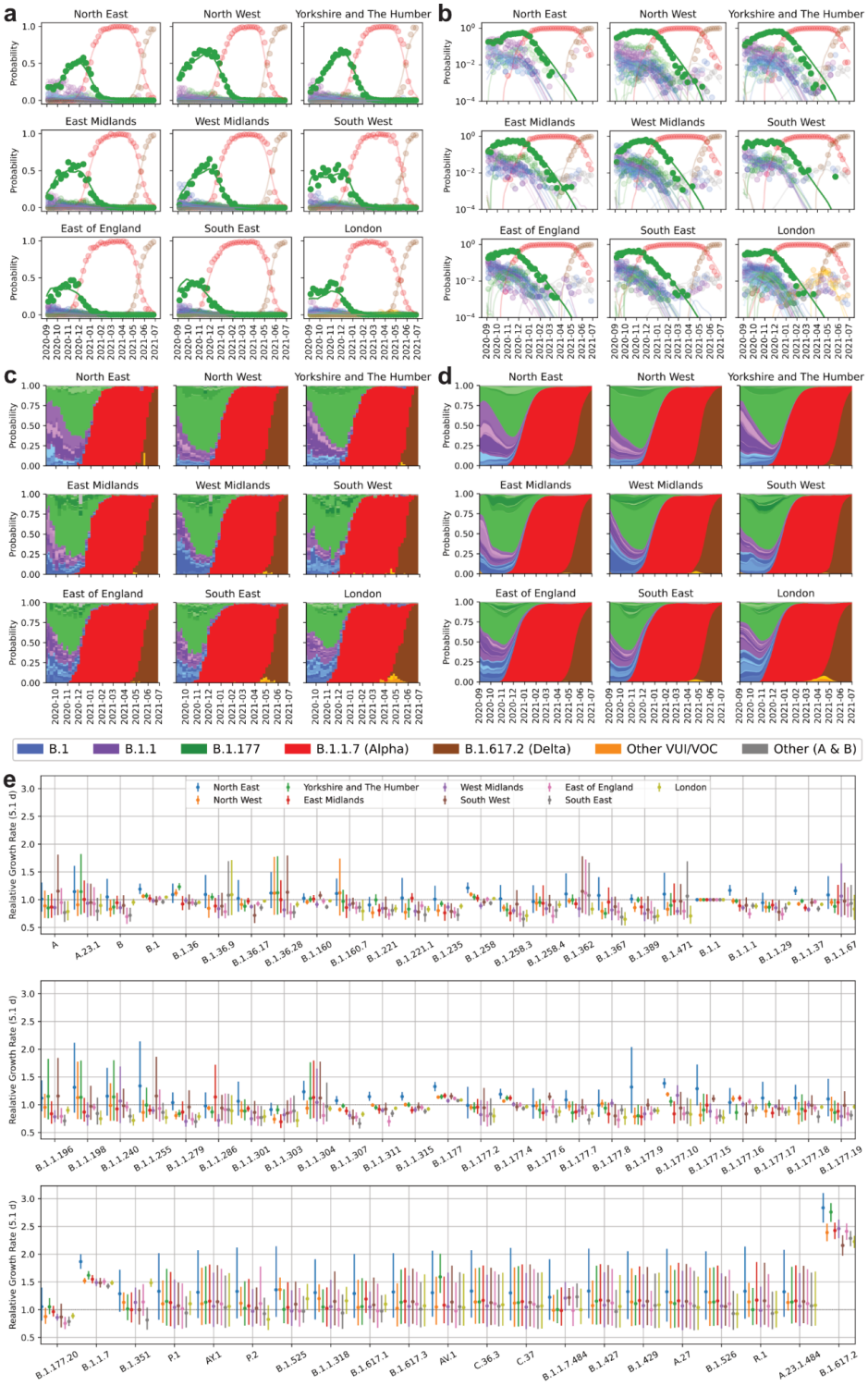

**Extended Data Figure 1, related to Figure 1. SARS-CoV-2 surveillance sequencing in England between September 2020 and June 2021. a**. Local monthly coverage across 315 LTLAs. **b**. Weekly coverage of genomic surveillance sequencing. **c**. Hospitalisation, case and infection fatality rates relative to ONS prevalence.



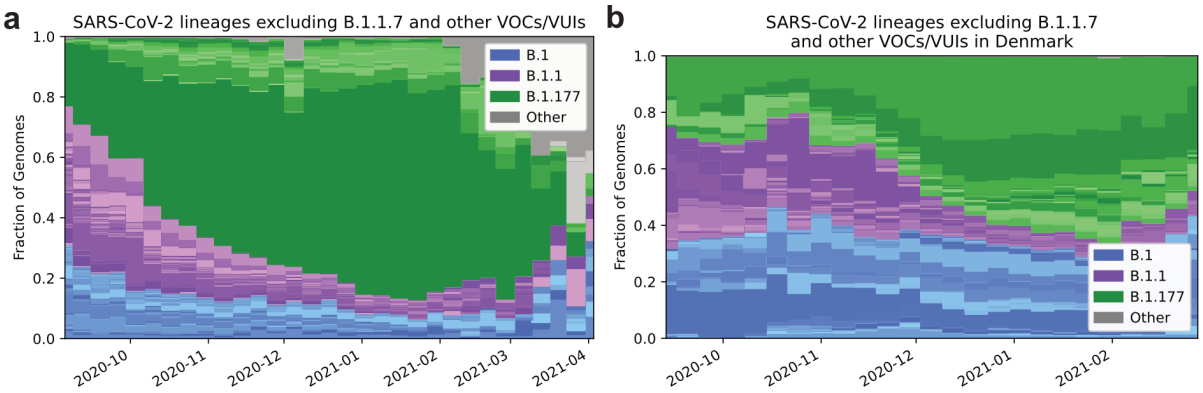

**Extended Data Figure 2: Genomic surveillance model of total incidence and lineage-specific frequencies. a.** Cubic basis splines (top row) are convolved with the infection to test distribution (row 2 and 3) and used to fit the log incidence in a LTLA and its corresponding derivatives (growth rates; bottom row). **b**. Example incidence (top row), logarithmic incidence with individual convolved basis functions (dashed lines, row 2), growth rate with individual spline basis derivatives (dashed lines, row 3) and resulting (case) reproduction numbers (growth rate per 5.1d) from our approach (GenomicSurveillance) and estimates by EpiEstim [47], shifted by 10d to approximate a case reproduction number. **c.** The relative frequencies of 62 different lineages are modelled using piecewise multinomial logistic regression. The linear logits are modelled to jump stochastically within 21d prior to first observation to account for the effects of new introductions. Shown are the logits of 5 selected lineages in two different LTLAs.



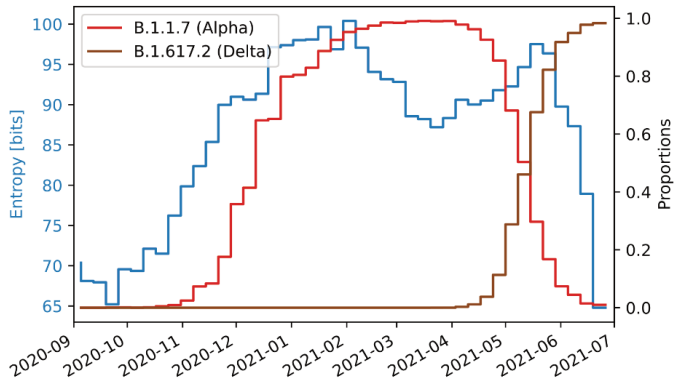

**Extended Data Figure 3: Genomic surveillance model selection. a.** Model loss in terms of the ELBO objective function and the model hyperparameters alpha0 and alpha1 (see **Methods**). **b.** Model deviance (calculated as -2 x log pointwise predictive density) with respect to the model hyperparameters alpha0 and alpha1 (see **Methods**). **c.** Mean squared error (MSE) of modelled weekly proportions of highly prevalent lineages with respect to the model parameters alpha0 and alpha1 (see **Methods**). **d.** Same as in **c**, but for lineages exhibiting low frequencies (VOCs).



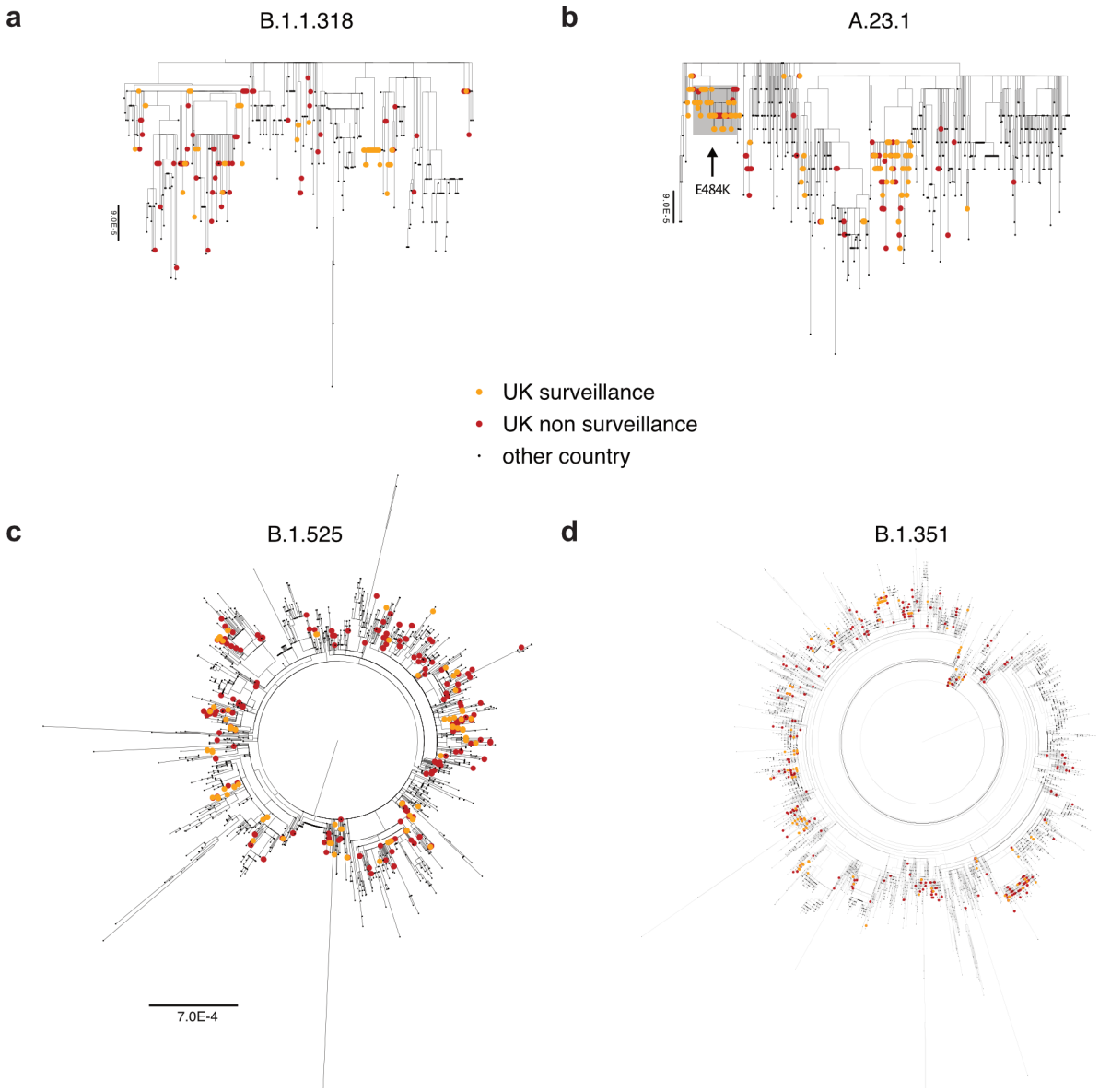

**Extended Data Figure 4. Spatiotemporal model of 71 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and June 2021. a.** Regional lineage specific relative frequency of lineages contributing more than 50 genomes during the time period shown. Dots denote observed data, lines the fits aggregated to each region. **b.** Same as **a**, but on a log scale. **c.** Same data as in a, shown as stacked bar charts. Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **d**. Same fits as in **a**, shown as stacked segments. **e.** Average growth rates for 71 SARS-Cov2 lineages estimated in different regions in England.



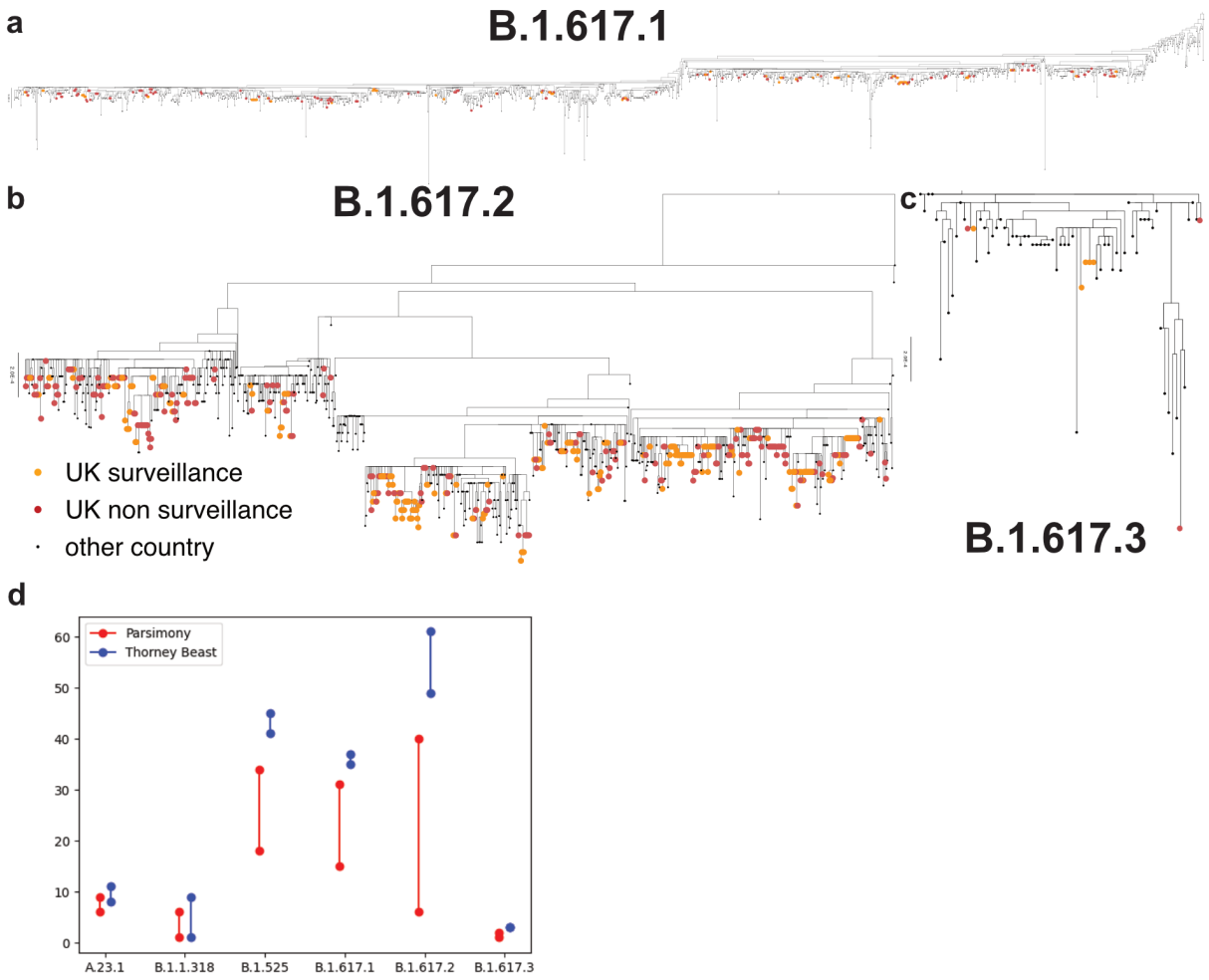

**Extended Data Figure 5. Relative growth of B.1.177. a.** Lineage-specific relative frequency data in England, excluding B.1.1.7 and other VOCs/VUIs (Category Other includes: A, A.18, A.20, A.23, A.25, A.27, A.28, B, B.29, B.40, None). Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **b**. Lineage-specific relative frequency data in Denmark, excluding B.1.1.7 and other VOCs/VUIs. Colors resemble major lineages as indicated and shadings thereof indicate sublineages.




**Extended Data Figure 6. Genomic diversity of the SARS-CoV-2 epidemic.** Shown is the entropy (blue), total number of observed Pango lineages (grey, divided by 4), as well as the proportion of B.1.1.7 (orange, right axis). The sweep of B.1.1.7 causes an intermittent decline of genomic diversity as measured by the entropy.




**Extended Data Figure 7. Global phylogenetic trees of selected VOCs/VUIs**. English surveillance and other (targeted and quarantine) samples are highlighted respectively orange and red.




**Extended Data Figure 8. Global phylogenetic trees of B.1.617 sublineages**. **a, b and c.** English surveillance and other (targeted and quarantine) samples are highlighted respectively orange and red. The trees of B.1.617.1 and B.1.617.2 are rooted. **d.** Number of UK introductions inferred by parsimony (minimum and maximum numbers) and by Thorney BEAST (95% posterior CI) for each VOC.