

UC San Diego

UC San Diego Previously Published Works

Title

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

Permalink

<https://escholarship.org/uc/item/2x40z5h4>

Journal

Nature, 600(7889)

ISSN

0028-0836

Authors

Robson, Samuel C

Connor, Thomas R

Loman, Nicholas J

et al.

Publication Date

2021-12-16

DOI

10.1038/s41586-021-04069-y

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

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

<https://doi.org/10.1038/s41586-021-04069-y>

Received: 22 May 2021

Accepted: 29 September 2021

Published online: 14 October 2021

Open access

 Check for updates

Harald S. Vöhringer¹, Theo Sanderson^{2,3}, Matthew Sinnott², Nicola De Maio¹, Thuy Nguyen², Richard Goater², Frank Schwach^{2,4}, Ian Harrison⁴, Joel Hellewell⁵, Cristina V. Ariani², Sonia Gonçalves², David K. Jackson², Ian Johnston², Alexander W. Jung¹, Callum Saint², John Sillitoe², Maria Suci², Nick Goldman¹, Jasmina Panovska-Griffiths⁶, The Wellcome Sanger Institute COVID-19 Surveillance Team*, The COVID-19 Genomics UK (COG-UK) Consortium*, Ewan Birney¹, Erik Volz⁷, Sebastian Funk⁵, Dominic Kwiatkowski², Meera Chand^{4,8}, Inigo Martincorena², Jeffrey C. Barrett^{2,9} & Moritz Gerstung^{1,9}✉

The evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus leads to new variants that warrant timely epidemiological characterization. Here we use the dense genomic surveillance data 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 subepidemics that peaked in early autumn 2020, followed by a jump in transmissibility of the B.1.1.7/Alpha lineage. The Alpha variant 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 the Alpha variant and eliminated nearly all other lineages in early 2021. Yet a series of variants (most of which contained the spike E484K mutation) defied these trends and persisted at moderately increasing proportions. However, by accounting for sustained introductions, we found that the transmissibility of these variants is unlikely to have exceeded the transmissibility of the Alpha variant. Finally, B.1.617.2/Delta was repeatedly introduced in England and grew rapidly in early summer 2021, constituting approximately 98% of sampled SARS-CoV-2 genomes on 26 June 2021.

The SARS-CoV-2 virus accumulates approximately 24 point mutations per year, or 0.3 mutations per viral generation^{1–3}. Most of these mutations appear to be evolutionarily neutral but, as the SARS-CoV-2 epidemic spread around the world during spring 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 that was isolated in Wuhan, China⁴.

A broad range of lineages have been defined since that 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 2020 and spread across Europe through travel⁷. Subsequently, four variants of concern (VOCs) have been identified by the WHO and other public health authorities: the B.1.351/Beta lineage was discovered in South Africa⁸, where it spread rapidly in late 2020. The B.1.1.7/Alpha lineage was first observed in Kent in September 2020 (ref.⁹) from where it swept through the United Kingdom and large parts of the world due to a 50–60% increase^{10–13} 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 coronavirus disease 2019 (COVID-19) in India in April 2021 and subsequently around the world.

Epidemiology of SARS-CoV-2 in England

In the United Kingdom, by late June 2021 the COVID-19 Genomics UK Consortium (COG-UK) had sequenced close to 600,000 viral samples. These data have enabled a detailed reconstruction of the dynamics of the first wave of the epidemic in the United Kingdom between February and August 2020 (ref.¹⁶). Here we leverage a subset of those data—genomic surveillance data generated at the Wellcome Sanger Institute—to characterize the growth rates and geographical spread of different SARS-CoV-2 lineages and reconstruct how newly emerging variants changed the course of the epidemic.

Our data cover England between 1 September 2020 and 26 June 2021, encompassing three epidemic waves and two national lockdowns (Fig. 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 of the positive tests from PCR testing for the wider population,

¹European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton, UK. ²Wellcome Sanger Institute, Hinxton, UK. ³The Francis Crick Institute, London, UK. ⁴Public Health England, London, UK. ⁵London School of Hygiene & Tropical Medicine, London, UK. ⁶The Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ⁷MRC Centre for Global Infectious Disease Analysis, Jameel Institute for Disease and Emergency Analytics, Imperial College London, London, UK. ⁸Guy's and St Thomas' NHS Foundation Trust, London, UK. ⁹Division for AI in Oncology, German Cancer Research Centre DKFZ, Heidelberg, Germany. *Lists of authors and their affiliations appear online. ✉e-mail: jb26@sanger.ac.uk; moritz.gerstung@ebi.ac.uk

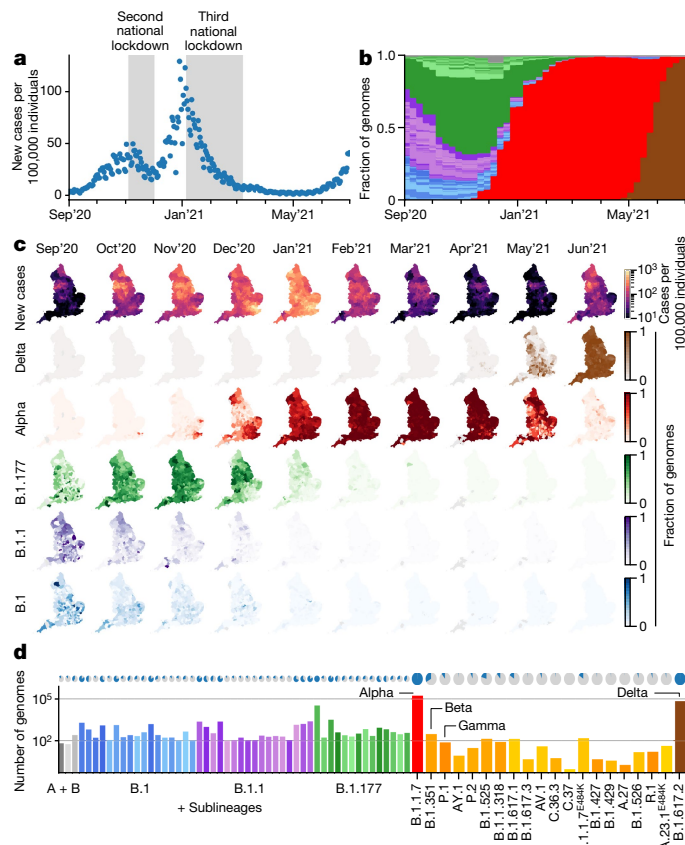


Fig. 1 | SARS-CoV-2 surveillance sequencing in England between September 2020 and June 2021. **a**, Positive Pillar 2 SARS-CoV-2 tests in England. **b**, The relative frequency of 328 different PANGO lineages, representing approximately 7.2% of the tests shown in **a**. **c**, Positive tests (row 1) and the frequency of 4 major lineages (rows 2–5) across 315 English lower tier local authorities. **d**, The absolute frequency of sequenced genomes mapped to 71 PANGO lineages. The blue areas in the pie charts are proportional to the fraction of LTLAs in which a given lineage was observed.

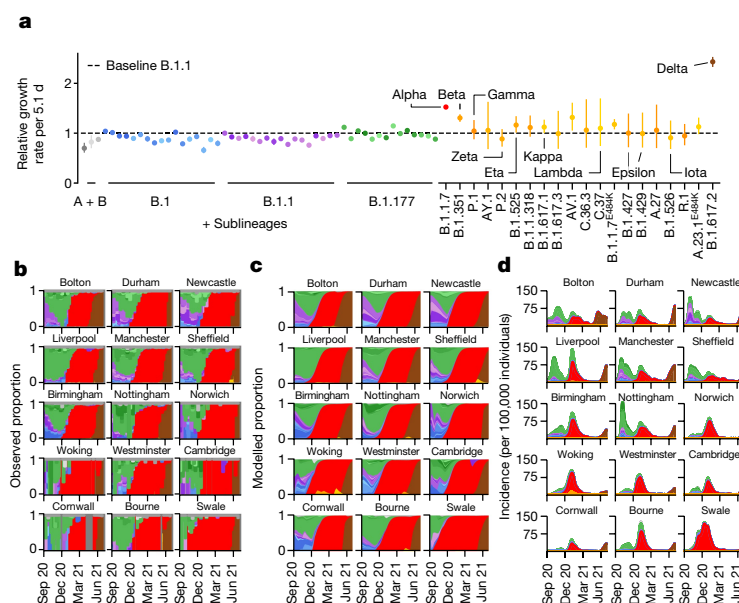


Fig. 2 | Spatiotemporal model of 71 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and June 2021. **a**, The average growth rates for 71 lineages. Data are median \pm 95% CI. **b**, Lineage-specific relative frequency for 15 selected LTLAs, arranged by longitude and latitude to geographically cover England. **c**, Fitted lineage-specific relative frequency for the same LTLAs as in **b**. **d**, Fitted lineage-specific incidence for the same LTLAs as in **b**.

ranging from 5% in winter 2020 to 38% in early summer 2021, and filtered to remove cases that were associated with international travel (Methods and Extended Data Fig. 1a, b). Overall, a total of 328 SARS-CoV-2 lineages were identified using the PANGO lineage definition⁵. As some of these lineages were only rarely and intermittently detected, we collapsed these on the basis of the underlying phylogenetic tree into a set of 71 lineages for modelling (Fig. 1b–d and Supplementary Tables 1 and 2).

These data reveal a diversity of lineages in the fall of 2020 followed by sweeps of the Alpha and Delta variants (Fig. 1b and Supplementary Tables 2 and 3). Figure 1c shows the geographical 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.

Modelling the dynamics of SARS-CoV-2

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 (Methods and Extended Data Fig. 2). The multivariate logistic regression model is conceptually similar to previous approaches in its estimation of relative growth rates^{10,11}. It accounts for differences in the epidemiological dynamics between LTLAs, and enables the introduction of new lineages (Fig. 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 (Fig. 2b, c and Supplementary Notes 1 and 2). The quality of fit is confirmed by different probabilistic model selection criteria (Extended Data Fig. 3) and also evident at the aggregated regional level (Extended Data Fig. 4).

Although the relative growth rate of each lineage is modelled as identical across LTLAs, the local viral proportions change dynamically due to the timing and rate of introduction of different lineages. The model also calculates total and lineage-specific local incidences and time-dependent growth rates and approximate reproduction numbers R_t by negative binomial spline fitting of the number of daily positive PCR tests (Methods, Fig. 2d and Extended Data Fig. 2c). Together, this enables a quantitative reconstruction of different periods of the epidemic, which we will discuss in chronological order.

Multiple subepidemics in autumn 2020

Autumn 2020 was characterized by a surge of cases—concentrated in the north of England—that peaked in November, triggering a second national lockdown (Fig. 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 (Fig. 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 of between 8% (growth per 5.1 d; 95% confidence interval (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 Fig. 5a) and involved a number of monophyletic sublineages that arose in the UK, and similar patterns were observed in Denmark¹⁷ (Extended Data Fig. 5b). Such behaviour cannot easily be explained by international travel, which was the major factor in the initial spread of B.1. throughout Europe in summer 2020 (ref. 7). However, the underlying biological mechanism is unclear as the characteristic A222V spike variant is not believed to confer a growth advantage⁷.

The spread of Alpha during restrictions

The subsequent third wave from December 2020 to February 2021 was almost exclusively driven by Alpha/B.1.1.7, as described previously^{10,11,18}. The rapid sweep of Alpha was due to an estimated transmissibility advantage of 1.52 compared with B.1.1 (growth per 5.1 d; 95% CI = 1.50–1.55; Fig. 2a), assuming an unchanged generation interval distribution¹⁹. The growth advantage is thought to stem, at least in part, from spike mutations that facilitate ACE2 receptor binding (N501Y)^{20,21} and furin cleavage (P681H)²². Alpha grew during a period of restrictions, which proved to be insufficient to contain its spread (Fig. 3a).

The second national lockdown from 5 November to 1 December 2020 successfully reduced the total number of cases, but this masked a lineage-specific increase ($R_t > 1$; defined as growth per 5.1 d) in Alpha and a simultaneous decrease in other hitherto dominant lineages ($R_t < 1$) in 78% (246/315) of LTLAs²³ (Fig. 3b, c). This pattern of Alpha-specific growth during lockdown is supported by a model-agnostic analysis of raw case numbers and proportions of Alpha genomes (Fig. 3e).

Three levels of regionally tiered restrictions were introduced in December 2020 (ref. 24) (Fig. 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 (Fig. 3a–c). The reopening caused a surge of cases across all tiers with $R_t > 1$, which is also evident in selected time series (Fig. 3d). As Alpha cases surged, more areas were placed under tier 3 restrictions, and stricter tier 4 restrictions were introduced. Nevertheless, Alpha continued to grow ($R_t > 1$) in most areas, presumably driven by increased social interaction over Christmas (Fig. 3c).

After the peak of 72,088 daily cases on 29 December 2020 (Fig. 1a), a third national lockdown was announced on 4 January 2021 (Fig. 3a). The lockdown and increasing immunity derived from infection and increasing vaccination²⁵ led to a sustained contraction of the epidemic to approximately 5,500 daily cases by 8 March, when restrictions began to be lifted by reopening schools (further steps of easing occurred on 12 April and 17 May). In contrast to the second national lockdown 93% (296/315) of LTLAs exhibited a contraction in both Alpha and other lineages (Fig. 3e).

Elimination of lineages in early 2021

The lineage-specific rates of decline during the third national lockdown and throughout March 2021 resulted in large differences in lineage-specific incidence. Cases of Alpha contracted nationally from a peak of around 50,000 daily new cases to approximately 2,750 on

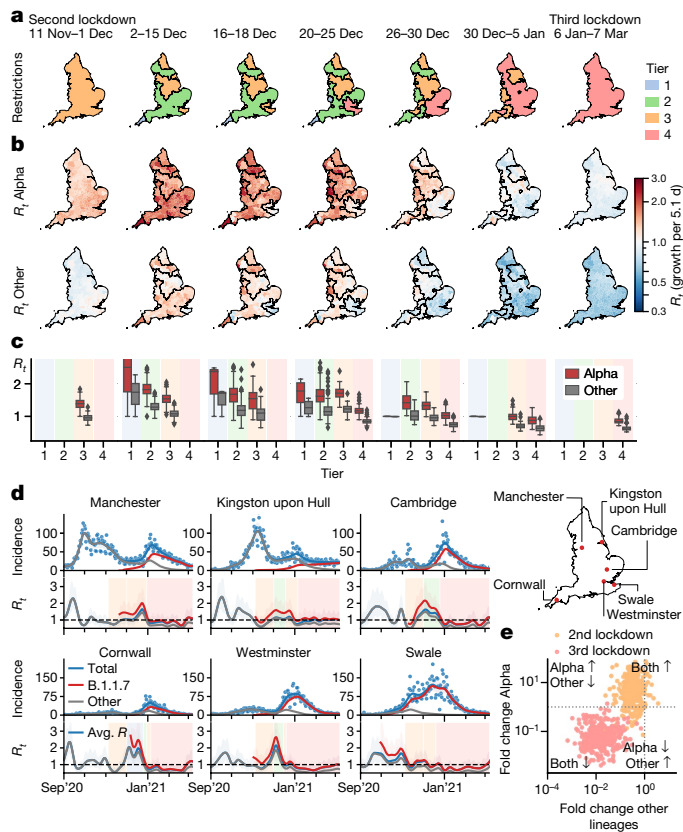


Fig. 3 | Growth of B.1.1.7/Alpha and other lineages in relation to lockdown restrictions between November 2020 and March 2021. **a**, Maps and dates of national and regional restrictions in England. Second national lockdown: closed hospitality businesses; contacts ≤ 2 , outdoors only; open schools; reasonable excuse needed for leaving home⁴⁵. Tier 1: private indoor gatherings of ≤ 6 persons. Tier 2: as tier 1 plus restricted hospitality services; gatherings of ≤ 6 in public outdoor places. Tier 3: as tier 2 plus most hospitality businesses closed. Tier 4: as tier 3 but 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 5.1 d) of all of the other lineages in the same periods. **c**, R_t values from $n = 315$ LTLA shown in **b**. The box centre horizontal line indicates the median, box limits show the quartiles, the whiskers extend to $1.5 \times$ the interquartile range. **d**, Total and lineage-specific incidence (top) and R_t values (bottom) for six selected LTLAs during the period of restrictions. **e**, Crude lineage-specific fold changes (odds ratios) in Alpha and other lineages across the second (orange) and third national lockdown (red).

1 April 2021 (Fig. 4a). At the same time, B.1.177—the most prevalent lineage in November 2020—fell to less than an estimated 10 cases per day. Moreover, the incidence of most other lineages present in autumn 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 (Fig. 4b). Although this may be attributed in part 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 Fig. 6).

Refractory variants with E484K mutations

Parallel to the elimination of many formerly dominant SARS-CoV-2 lineages, a number of new variants were imported or emerged (Fig. 4a). These include the VOCs B.1.351/Beta and P.1/Gamma, which carry the spike variant N501Y that is also found in B.1.1.7/Alpha and a similar pair of mutations (K417N/T and E484K) that were each shown to reduce the

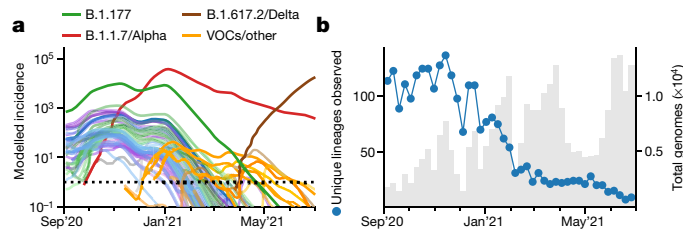


Fig. 4 | Elimination of SARS-CoV-2 lineages during spring 2021. **a**, Modelled lineage-specific incidence in England. The colours resemble major lineages as indicated and shades thereof indicate the respective sublineages. **b**, The observed number of PANGO lineages per week.

binding affinity of antibodies from vaccine-derived or convalescent sera^{20,26–29}. The ability to escape from previous immunity is consistent with the epidemiology of Beta in South Africa⁸ and especially the surge of Gamma in Manaus¹⁵. The variants B.1.525/Eta, B.1.526/Iota, 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 that acquired E484K were found in England (Fig. 5a, b).

The proportion of these E484K-containing variants 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 (Fig. 5a, b). Yet, the dynamics were largely stochastic and characterized by a series of individual and localized outbreaks, possibly curtailed by local surge testing efforts against Beta and Gamma variants (Fig. 5c). Consistent with the transient nature of these outbreaks, the estimated growth rates of these variants were typically lower than Alpha (Fig. 2a).

Sustained imports from international travel were a critical driving mechanism behind the observed number of non-Alpha cases. A phylogeographical analysis establishing the most parsimonious sets of monophyletic and exclusively domestic clades, which can be interpreted as individual introductions, confirmed that A.23.1 with E484K (1 clade) probably has a domestic origin as no genomes of the same clade were observed internationally (Methods, Fig. 5d and Extended Data

Fig. 7). The estimated number of introductions was lowest for B.1.1.318 (3 introductions, range = 1–6), and highest for Beta (49 introductions, range = 45–58) and Eta (30 introductions, range = 18–34). Although our data exclude genomes sampled directly from travellers, these repeated introductions show that the true rate of transmission is lower than the observed increase in the number of surveillance genomes.

The rise of Delta from April to June 2021

The B.1.617.1/Kappa and B.1.617.2/Delta lineages, which were first detected in India in 2020, first appeared in English surveillance samples in March 2021. In contrast to other VOCs, Delta/Kappa do not contain N501Y or E484K mutations, but their L452R mutation may reduce antibody recognition²⁷ and P681R enhances furin cleavage³⁰, similar to the P681H mutation of Alpha. The frequency of Delta, which harbours further spike mutations of unknown function, increased rapidly and reached levels of 98% (12,474/12,689) on 26 June 2021 (Fig. 5a, b). Although initially constrained to a small number of large local clusters, such as in Bolton, in May 2021 (Fig. 5c), Delta was detected in all LTLAs by 26 June 2021 (Fig. 1c). The sweep of Delta occurred at a rate of around 59% (growth per 5.1 d, CI = 53–66) higher than Alpha with minor regional variation (Fig. 2a, Extended Data Fig. 4e and Supplementary Table 4).

The rapid rise of Delta contrasts with Kappa, which grew more slowly despite being introduced at a similar time and into a similar demographic background (Figs. 2a and 5b). This is also evident in the phylogeographical analysis (based on data as of 1 May 2021). The 224 genomes of Delta derive from larger clades (23 introductions, range = 6–40; around 10 genomes for every introduction) compared with the 80 genomes of Kappa (17 introductions, range = 15–31; around 3–4 genomes per introduction) and also other variants (Fig. 5d and Extended Data Fig. 8). The AY.1 lineage, derived from Delta and containing an additional K417N mutation, appeared only transiently (Fig. 5b).

The sustained domestic growth of Delta and its international spread³¹ relative to the Alpha lineage are the first evidence of a biological growth advantage. The causes appear to be a combination of increased transmissibility and immune evasion. Evidence for higher transmissibility

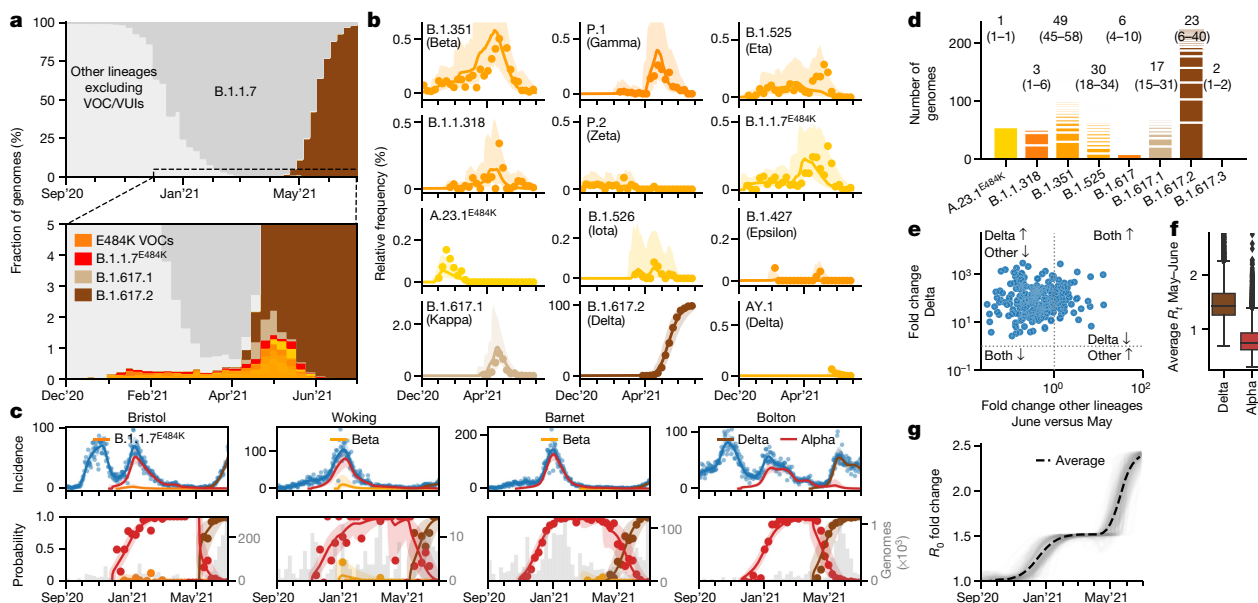


Fig. 5 | Dynamics of E484K variants and Delta between January and June 2021.

a, The observed relative frequency of other lineages (light grey), Alpha/B.1.1.7 (dark grey), E484K variants (orange) and Delta/B.1.617.2 (brown). **b**, The observed and modelled relative frequency of variants in England. **c**, The total and relative lineage-specific incidence in four selected LTLAs. For **b** and **c**, the shaded areas indicate the 95% CIs. **d**, Estimated UK 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**, Lineage-specific R_0 values of $n = 315$ LTLA in the same period, defined as in Fig. 3c. **g**, Changes in the average transmissibility across 315 LTLAs during the study period.

includes the fast growth in younger unvaccinated age groups, reports of elevated secondary attack rates³² and a higher viral load³³. Furthermore, vaccine efficacy against infection by Delta is diminished, depending on the type of vaccine^{34,35}, and reinfection is more frequent³⁶, both supported by experimental research demonstrating the reduced antibody neutralization of Delta by vaccine-derived and convalescent sera^{37,38}.

The higher growth rate of Delta—combined with gradual reopening and preceding vaccination—repeated the dichotomous pattern of lineage-specific decline and growth, although now with declining Alpha ($R_t < 1$) and growing Delta ($R_t > 1$; Fig. 5e, f). Overall, we estimate that the spread of more transmissible variants between August 2020 and early summer 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; Fig. 5g). Thus, previously effective interventions may prove to be insufficient to contain newly emerging and more transmissible variants.

Discussion

Our dense genomic surveillance analysis identified lineages that consistently grew faster than others in each local authority and, therefore, at the same time, under the same restrictions and in a comparable population. This pinpointed a series of variants with elevated transmissibility, in broad agreement with other reports^{10,11,13,15,31}. However, a number of limitations exist. The growth rates of rare new variants are stochastic due to introductions and superspreading. Local outbreaks of the Beta and Gamma variants triggered asymptomatic surge testing, which may have reduced their spread. Furthermore, 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 period³⁹. This will influence the growth rates of variants 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 limitations are discussed in the Limitations section of the Methods.

The third and fourth waves in England were each caused by more transmissible variants, which outgrew restrictions that were 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 declining. The fact that such growth was initially masked by the falling cases of dominant lineages highlights the need for dense genomic surveillance and rapid analysis 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 spring 2020 (ref. ⁴⁰) to the range of 6–7 for Delta. This is likely to spur new waves of the epidemic in countries that have to date been able to control the epidemic despite low vaccination rates, and it may exacerbate the situation elsewhere. Although 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 (ref. 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 therefore 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 humans⁴⁴. To date, variants with considerably higher transmissibility

have had strongest positive selection, and swept through England during the 10 months of this investigation. However, 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-021-04069-y>.

- Rambaut, A. *Phylogenetic Analysis of nCoV-2019 Genomes* (Virological, 2020); <https://virological.org/t/phylogenetic-analysis-176-genomes-6-mar-2020/356>
- Nextstrain Team *Genomic Epidemiology of Novel Coronavirus—Global Subsampling* (Nextstrain, 2020); <https://nextstrain.org/ncov/global?l=clock>
- Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
- Volz, E. et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75 (2021).
- Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
- O’Toole, Á. et al. *Global Report Investigating Novel Coronavirus Haplotypes* https://cov-lineages.org/global_report.html (2021).
- Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
- Tegally, H., Wilkinson, E., Lessells, R.J. et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021).
- Rambaut, A. et al. *Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations* (Virological, 2020); <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>
- Volz, E., Mishra, S., Chand, M. et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
- Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
- O’Toole, Á. et al. *Tracking the International Spread of SARS-CoV-2 Lineages B.1.1.7 and B.1.351/501Y-V2* (Virological, 2021); <https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592>
- Washington, N. L. et al. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184**, 2587–2594 (2021).
- Faria, N. R. et al. *Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in Manaus: Preliminary Findings* (Virological, 2021); <https://www.icpccovid.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>
- Faria, N. R. et al. Genomics and Epidemiology of the P.1 SARS-CoV-2 Lineage in Manaus, Brazil. *Science* **372**, 815–21 (2021).
- du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
- Danish Covid-19 Genome Consortium *Genomic Overview of SARS-CoV-2 in Denmark* (2021); <https://www.covid19genomics.dk/statistics>
- Kraemer, M. U. G. et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
- Park, S. W. et al. Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. Preprint at *medRxiv* <https://doi.org/10.1101/2021.05.03.21256545> (2021).
- Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 (2020).
- Zahradnik, J., Marciano, S., Shemesh, M. et al. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat. Microbiol.* **6**, 1188–1198 (2021).
- Brown, J. C. et al. 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. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.02.24.432576> (2021).
- Vöhringer, H. et al. *Lineage-specific Growth of SARS-CoV-2 B.1.1.7 During the English National Lockdown* (Virological, 2020); <https://virological.org/t/lineage-specific-growth-of-sars-cov-2-b-1-1-7-during-the-english-national-lockdown/575/2>
- The Health Protection (Coronavirus, Restrictions) (All Tiers) (England) Regulations 2020. *Wikipedia* [https://en.wikipedia.org/w/index.php?title=The_Health_Protection_\(Coronavirus,_Restrictions\)__\(All_Tiers\)__\(England\)_Regulations_2020&oldid=1014831173](https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)__(All_Tiers)__(England)_Regulations_2020&oldid=1014831173) (2021).
- Steel, K. & Davies, B. *Coronavirus (COVID-19) Infection Survey, Antibody and Vaccination Data for the UK* (ONS, 2021); <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveyantibodydatafortheuk/28april2021>
- Greaney, A. J. et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host & Microbe* **29**, 463–476 (2021).
- Greaney, A. J. et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44–57 (2021).

28. Zhou, D. et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348–2361 (2021).
29. Planas, D. et al. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. *Nat. Med.* **27**, 917–924 (2021).
30. Peacock, T. P. et al. The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.28.446163> (2021).
31. Campbell, F. et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26**, 2100509 (2021).
32. *Investigation of Novel SARS-CoV-2 Variants of Concern* Technical briefing 10 (Public Health England, 2021); <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>
33. Li, B. et al. Viral infection and transmission in a large well-traced outbreak caused by the Delta SARS-CoV-2 variant. Preprint at *medRxiv* <https://doi.org/10.1101/2021.07.07.21260122> (2021).
34. Nasreen, S. et al. Effectiveness of COVID-19 vaccines against variants of concern in Ontario, Canada. Preprint at *medRxiv* <https://doi.org/10.1101/2021.06.28.21259420> (2021).
35. Lopez Bernal, J. et al. Effectiveness of Covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
36. *Investigation of Novel SARS-CoV-2 Variants of Concern* Technical briefing 19 (Public Health England, 2021); https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1005517/Technical_Briefing_19.pdf
37. Ferreira, I. et al. SARS-CoV-2 B.1.617 emergence and sensitivity to vaccine-elicited antibodies. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.08.443253> (2021).
38. Wall, E. C. et al. 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. Steel, K. & Haughton, P. *Coronavirus (COVID-19) Infection Survey, Antibody and Vaccination Data, UK* (ONS, 2021); <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveyantibodyandvaccinationdatafortheuk/21july2021>
40. Anderson, R. et al. *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).
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. et al. 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. *Euro Surveill.* **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.* **50**, 40–48 (2021).
45. The Health Protection (Coronavirus, Restrictions) (England) (No. 4) Regulations 2020. *Wikipedia* [https://en.wikipedia.org/w/index.php?title=The_Health_Protection_\(Coronavirus,_Restrictions\)_\(_England\)_\(_No._4\)_Regulations_2020&oldid=1014701607](https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)_(_England)_(_No._4)_Regulations_2020&oldid=1014701607) (2021).

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021, corrected publication 2022

Article

The Wellcome Sanger Institute COVID-19 Surveillance Team

Irina Abnizova², Louise Aigrain², Alex Alderton², Mozam Ali², Laura Allen², Roberto Amato², Ralph Anderson², Cristina Ariani², Siobhan Austin-Guest², Sendu Bala², Jeffrey Barrett², Andrew Bassett², Kristina Battleday², James Beal², Mathew Beale², Charlotte Beaver², Sam Bellamy², Tristram Bellerby², Katie Bellis², Duncan Berger², Matt Berriman², Emma Betteridge², Paul Bevan², Simon Binley², Jason Bishop², Kirsty Blackburn², James Bonfield², Nick Boughton², Sam Bowker², Timothy Brendler-Spaeth², Iraad Bronner², Tanya Brooklyn², Sarah Kay Buddenborg², Robert Bush², Catarina Caetano², Alex Cagan², Nicola Carter², Joanna Cartwright², Tiago Carvalho Monteiro², Liz Chapman², Tracey-Jane Chillingworth², Peter Clapham², Richard Clark², Adrian Clarke², Catriona Clarke², Daryl Cole², Elizabeth Cook², Maria Coppola², Linda Cornell², Clare Cornwell², Craig Corton², Abby Crackett², Alison Cranage², Harriet Craven², Sarah Craw², Mark Crawford², Tim Cutts², Monika Dabrowska², Matt Davies², Robert Davies², Joseph Dawson², Callum Day², Aiden Densem², Thomas Dibling², Cat Dockree², David Dodd², Sunil Dogga², Matthew Dorman², Gordon Dougan², Martin Dougherty², Alexander Dove², Lucy Drummond², Eleanor Drury², Monika Dudek², Jillian Durham², Laura Durrant², Elizabeth Easthope², Sabine Eckert², Pete Ellis², Ben Farr², Michael Fenton², Marcella Ferrero², Neil Flack², Howard Fordham², Grace Forsythe², Luke Foulser², Matt Francis², Audrey Fraser², Adam Freeman², Anastasia Galvin², Maria Garcia-Casado², Alex Gedny², Sophia Girgis², James Glover², Sonia Goncalves², Scott Goodwin², Oliver Gould², Marina Gourtovaia², Andy Gray², Emma Gray², Coline Griffiths², Yong Gu², Florence Guerin², Will Hamilton², Hannah Hanks², Ewan Harrison², Alexandria Harrott², Edward Harry², Julia Harvison², Paul Heath², Anastasia Hernandez-Koutoucheva², Rhiannon Hobbs², Dave Holland², Sarah Holmes², Gary Hornett², Nicholas Hough², Liz Huckle², Lena Hughes-Hallett², Adam Hunter², Stephen Inglis², Sameena Iqbal², Adam Jackson², David Jackson², Keith James², Dorota Jamrozny², Carlos Jimenez Verdejo², Ian Johnston², Matthew Jones², Kalyan Kallepally², Leanne Kane², Keely Kay², Sally Kay², Jon Keatley², Alan Keith², Alison King², Lucy Kitchin², Matt Kleanthous², Martina Klimekova², Petra Korlevic², Ksenia Krasheninnikova², Dominic Kwiatkowski², Greg Lane², Cordelia Langford², Adam Laverack², Katharine Law², Mara Lawniczak², Stefanie Lensing², Steven Leonard², Laura Letchford², Kevin Lewis², Amanah Lewis-Wade², Jennifer Liddle², Quan Lin², Sarah Lindsay², Sally Linsdell², Rich Livett², Stephanie Lo², Rhona Long², Jamie Lovell², Jon Lovell², Catherine Ludden², James Mack², Mark Maddison², Aleksei Makunin², Irfan Mamun², Jenny Mansfield², Neil Marriott², Matt Martin², Inigo Martincorena², Matthew Mayo², Shane McCarthy², Jo McClintock², Samantha McGuigan², Sandra McHugh², Liz McMinn², Carl Meadows², Emily Mobley², Robin Moll², Maria Morra², Leanne Morrow², Kathryn Murie², Sian Nash², Claire Nathwani², Plamena Naydenova², Alexandra Neaverson², Rachel Nelson², Ed Nerou², Jon Nicholson², Tabea Nimz², Guillaume G. Noell², Sarah O'Meara², Valeriu Ohan², Karen Oliver², Charles Olney², Doug Ormond², Agnes Oszlanczi², Steve Palmer², Yoke Fei Pang², Barbora Pardubska², Naomi Park², Aaron Parmar², Gaurang Patel², Minal Patel², Maggie Payne², Sharon Peacock¹⁰, Arabella Petersen², Deborah Plowman², Tom Preston², Liam Prestwood², Christoph Puethel², Michael Quail², Diana Rajan², Shavanthi Rajatileka², Richard Rance², Suzannah Rawlings², Nicholas Redshaw², Joe Reynolds², Mark Reynolds², Simon Rice², Matt Richardson², Connor Roberts², Katrina Robinson², Melanie Robinson², David Robinson², Hazel Rogers², Eduardo Martin Rojo², Daljit Roopra², Mark Rose², Luke Rudd², Ramin Sadri¹⁰, Nicholas Salmon², David Saul², Frank Schwach², Carol Scott², Phil Seekings², Lesley Shirley², John Sillitoe², Alison Simms², Matthew Sinnott², Shanthi Sivadasan², Bart Siwek², Dale Sizer², Kenneth Skeldon², Jason Skelton², Joanna Slater-Tunstall², Lisa Sloper², Nathalie Smerdon², Chris Smith², Christen Smith², James Smith², Katie Smith², Michelle Smith², Sean Smith², Tina Smith², Leighton Sneade², Carmen Diaz Soria², Catarina Sousa¹⁰, Emily Souster², Andrew Sparkes², Michael Spencer-Chapman², Janet Squares², Robert Stanley², Claire Steed², Tim Stickland², Ian Still², Michael R. Stratton², Michelle Strickland², Allen Swann², Agnieszka Swiatkowska², Neil Sycamore², Emma Swift², Edward Symons², Suzanne Szluha², Emma Taluy², Nunu Tao², Katy Taylor², Sam Taylor², Stacey Thompson², Mark Thompson², Mark Thomson², Nicholas Thomson², Scott Thurston², Gerry Tonkin-Hill², Dee Toombs², Benjamin Topping², Jaime Tovar-Corona², Daniel Ungureanu², James Uphill², Jana Urbanova², Philip Jansen Van Vuuren², Valerie Vancollie², Paul Voak², Danielle Walker², Matthew Walker², Matt Waller², Gary Ward², Charlie Weatherhogg², Niki Webb², Danni Weldon², Alan Wells², Eloise Wells², Luke Westwood², Theo Whipp², Thomas Whiteley², Georgia Whitton², Andrew Whitwham², Sara Widaa², Mia Williams², Mark Wilson² & Sean Wright²

¹⁰Department of Medicine, University of Cambridge, Cambridge, UK.

The COVID-19 Genomics UK (COG-UK) Consortium

Funding acquisition, leadership and supervision, metadata curation, project administration, samples and logistics, sequencing and analysis, software and analysis tools, and visualization
Samuel C. Robson^{11,12}

Funding acquisition, leadership and supervision, metadata curation, project administration, samples and logistics, sequencing and analysis, and software and analysis tools
Thomas R. Connor^{13,14} & Nicholas J. Loman¹⁵

Leadership and supervision, metadata curation, project administration, samples and logistics, sequencing and analysis, software and analysis tools, and visualization
Tanya Golubchik⁶

Funding acquisition, leadership and supervision, metadata curation, samples and logistics, sequencing and analysis, and visualization
Rocio T. Martinez Nunez¹⁶

Funding acquisition, leadership and supervision, project administration, samples and logistics, sequencing and analysis, and software and analysis tools
David Bonsall⁶

Funding acquisition, leadership and supervision, project administration, sequencing and analysis, software and analysis tools, and visualization
Andrew Rambaut¹⁷

Funding acquisition, metadata curation, project administration, samples and logistics, sequencing and analysis, and software and analysis tools
Luke B. Snell¹⁸

Leadership and supervision, metadata curation, project administration, samples and logistics, software and analysis tools, and visualization
Rich Livett²

Funding acquisition, leadership and supervision, metadata curation, project administration, and samples and logistics
Catherine Ludden^{4,10}

Funding acquisition, leadership and supervision, metadata curation, samples and logistics, and sequencing and analysis
Sally Corden¹⁴ & Eleni Nastouli^{19,20,21}

Funding acquisition, leadership and supervision, metadata curation, sequencing and analysis, and software and analysis tools
Gaia Nebbia¹⁸

Funding acquisition, leadership and supervision, project administration, samples and logistics, and sequencing and analysis
Ian Johnston²

Leadership and supervision, metadata curation, project administration, samples and logistics, and sequencing and analysis
Katrina Lythgoe⁶, M. Estee Torok^{10,22} & Ian G. Goodfellow²³

Leadership and supervision, metadata curation, project administration, samples and logistics, and visualization
Jacqui A. Prieto^{24,25} & Kordo Saeed^{24,26}

Leadership and supervision, metadata curation, project administration, sequencing and analysis, and software and analysis tools
David K. Jackson²

Leadership and supervision, metadata curation, samples and logistics, sequencing and analysis, and visualization
Catherine Houlihan^{19,27}

Leadership and supervision, metadata curation, sequencing and analysis, software and analysis tools, and visualization
Dan Frampton^{20,27}

Metadata curation, project administration, samples and logistics, sequencing and analysis, and software and analysis tools
William L. Hamilton²² & Adam A. Witney²⁹

Funding acquisition, samples and logistics, sequencing and analysis, and visualization
Giselda Bucca²⁸

Funding acquisition, leadership and supervision, metadata curation and project administration
Cassie F. Pope^{29,30}

Funding acquisition, leadership and supervision, metadata curation, and samples and logistics
Catherine Moore¹⁴

Funding acquisition, leadership and supervision, metadata curation, and sequencing and analysis
Emma C. Thomson³¹

Funding acquisition, leadership and supervision, project administration, and samples and logistics
Ewan M. Harrison^{2,32}

Funding acquisition, leadership and supervision, sequencing and analysis, and visualization
Colin P. Smith²⁸

Leadership and supervision, metadata curation, project administration, and sequencing and analysis
Fiona Rogan³³

Leadership and supervision, metadata curation, project administration, and samples and logistics

Shaun M. Beckwith³⁴, Abigail Murray³⁴, Dawn Singleton³⁴, Kirstine Eastick³⁵, Liz A. Sheridan³⁶, Paul Randell³⁷, Leigh M. Jackson³⁸, Cristina V. Ariani² & Sónia Gonçalves²

Leadership and supervision, metadata curation, samples and logistics, and sequencing and analysis

Derek J. Fairley^{33,39}, Matthew W. Loose⁴⁰ & Joanne Watkins¹⁴

Leadership and supervision, metadata curation, samples and logistics, and visualization

Samuel Moses^{41,42}

Leadership and supervision, metadata curation, sequencing and analysis, and software and analysis tools

Sam Nicholls¹⁵, Matthew Bull¹⁴ & Roberto Amato²

Leadership and supervision, project administration, samples and logistics, and sequencing and analysis

Darren L. Smith^{43,44,45}

Leadership and supervision, sequencing and analysis, software and analysis tools, and visualization

David M. Aanensen^{2,46} & Jeffrey C. Barrett²

Metadata curation, project administration, samples and logistics, and sequencing and analysis

Dinesh Aggarwal^{2,4,10}, James G. Shepherd³¹, Martin D. Curran¹⁷ & Surendra Parmar⁴⁷

Metadata curation, project administration, sequencing and analysis, and software and analysis tools

Matthew D. Parker⁴⁸

Metadata curation, samples and logistics, sequencing and analysis, and software and analysis tools

Catryn Williams¹⁴

Metadata curation, samples and logistics, sequencing and analysis, and visualization

Sharon Glaysher⁴⁹

Metadata curation, sequencing and analysis, software and analysis tools, and visualization

Anthony P. Underwood^{2,46}, Matthew Bashton^{43,44}, Nicole Pacchiarini¹⁴, Katie F. Loveson¹² & Matthew Byott^{19,20}

Project administration, sequencing and analysis, software and analysis tools, and visualization

Alessandro M. Carabelli¹⁰

Funding acquisition, leadership and supervision, and metadata curation

Kate E. Templeton^{17,50}

Funding acquisition, leadership and supervision, and project administration

Thushan I. de Silva⁴⁸, Dennis Wang⁴⁸, Cordelia F. Langford² & John Sillitoe²

Funding acquisition, leadership and supervision, and samples and logistics

Rory N. Gunson⁵¹

Funding acquisition, leadership and supervision, and sequencing and analysis

Simon Cottrell¹⁴, Justin O'Grady^{52,53} & Dominic Kwiatkowski^{2,54}

Leadership and supervision, metadata curation and project administration

Patrick J. Lillie³⁵

Leadership and supervision, metadata curation, and samples and logistics

Nicholas Cortes⁵⁵, Nathan Moore⁵⁶, Claire Thomas⁵⁵, Phillipa J. Burns³⁵, Tabitha W. Mahungu⁵⁶ & Steven Liggett⁵⁷

Leadership and supervision, metadata curation, and sequencing and analysis

Angela H. Beckett^{11,58} & Matthew T. G. Holden⁵⁹

Leadership and supervision, project administration, and samples and logistics

Lisa J. Levett⁶⁰, Husam Osman^{4,61} & Mohammed O. Hassan-Ibrahim³⁷

Leadership and supervision, project administration, and sequencing and analysis

David A. Simpson³³

Leadership and supervision, samples and logistics, and sequencing and analysis

Meera Chand⁴, Ravi K. Gupta³², Alistair C. Darby⁶² & Steve Paterson⁶²

Leadership and supervision, sequencing and analysis, and software and analysis tools

Oliver G. Pybus⁶³, Erik Volz⁷, Daniela de Angelis⁶⁴, David L. Robertson³¹, Andrew J. Page⁵² & Inigo Martincorena²

Leadership and supervision, sequencing and analysis, and visualization

Louise Aigrain² & Andrew R. Bassett²

Metadata curation, project administration, and samples and logistics

Nick Wong⁶⁵, Yusri Taha⁶⁶, Michelle J. Erkiert³⁷ & Michael H. Spencer Chapman^{2,32}

Metadata curation, project administration, and sequencing and analysis

Rebecca Dewar⁵⁰ & Martin P. McHugh^{50,67}

Metadata curation, project administration, and software and analysis tools

Siddharth Mookerjee^{68,69}

Metadata curation, project administration and visualization

Stephen Aplin²⁴, Matthew Harvey²⁴, Thea Sass²⁴, Helen Umpleby²⁴ & Helen Wheeler²⁴

Metadata curation, samples and logistics, and sequencing and analysis

James P. McKenna³⁹, Ben Warne⁷⁰, Joshua F. Taylor⁷¹, Yasmin Chaudhry²³, Rhys Izuagbe²³, Aminu S. Jahun²³, Gregory R. Young^{43,44}, Claire McMurray¹⁵, Clare M. McCann^{44,45}, Andrew Nelson^{44,45} & Scott Elliott⁴⁹

Metadata curation, samples and logistics, and visualization

Hannah Lowe⁴¹

Metadata curation, sequencing and analysis, and software and analysis tools

Anna Price¹³, Matthew R. Crown¹⁴, Sara Rey¹⁴, Sunando Roy¹⁹ & Ben Temperton³⁸

Metadata curation, sequencing and analysis, and visualization

Sharif Shaaban⁵⁹ & Andrew R. Hesketh²⁸

Project administration, samples and logistics, and sequencing and analysis

Kenneth G. Laing²⁹, Irene M. Monahan²⁹ & Judith Heaney^{19,20,60}

Project administration, samples and logistics, and visualization

Emanuela Pelosi²⁴, Siona Silveira²⁴ & Eleri Wilson-Davies²⁴

Samples and logistics, software and analysis tools, and visualization

Helen Fryer⁶

Sequencing and analysis, software and analysis tools, and visualization

Helen Adams⁷², Louis du Plessis⁵³, Rob Johnson⁷, William T. Harvey^{31,73}, Joseph Hughes³¹, Richard J. Orton³¹, Lewis G. Spurgin⁷⁴, Yann Bourgeois⁵⁸, Chris Ruis³², Aine O'Toole¹⁷, Marina Gourtova² & Theo Sanderson²

Funding acquisition, and leadership and supervision

Christophe Fraser⁶, Jonathan Edgeworth¹⁸, Judith Breuer^{19,75}, Stephen L. Michell³⁸ & John A. Todd⁷⁶

Funding acquisition and project administration

Michaela John⁷⁷ & David Buck⁷⁶

Leadership and supervision, and metadata curation

Kavitha Gajee³⁵ & Gemma L. Kay⁵²

Leadership and supervision, and project administration

Sharon J. Peacock⁴¹⁰ & David Heyburn¹⁴

Leadership and supervision, and samples and logistics

Katie Kitchman³⁵, Alan McNally^{15,78}, David T. Pritchard⁶⁵, Samir Dervisevic⁷⁹, Peter Muir⁴, Esther Robinson^{4,61}, Barry B. Vipond⁴, Newara A. Ramadan⁸⁰, Christopher Jeanes⁸¹, Danni Weldon², Jana Catalan⁸² & Neil Jones⁸²

Leadership and supervision, and sequencing and analysis

Ana da Silva Filipe³¹, Chris Williams¹⁴, Marc Fuchs³³, Julia Miskelly³³, Aaron R. Jeffries³⁸, Karen Oliver² & Naomi R. Park²

Metadata curation, and samples and logistics

Amy Ash⁸³, Cherian Koshy⁸³, Magdalena Barrow⁸⁴, Sarah L. Buchan⁸⁴, Anna Mantzouratou⁸⁴, Gemma Clark⁸⁵, Christopher W. Holmes⁸⁶, Sharon Campbell⁸⁷, Thomas Davis⁸⁸, Ngee Keong Tan⁷¹, Julianne R. Brown⁷⁵, Kathryn A. Harris^{75,89}, Stephen P. Kidd⁵⁵, Paul R. Grant⁶⁰, Li Xu-McCrae⁹¹, Alison Cox^{68,90}, Pinglawathee Madona^{68,90}, Marcus Pond^{68,90}, Paul A. Randell^{68,90}, Karen T. Withell⁹¹, Cheryl Williams⁹², Clive Graham⁹³, Rebecca Denton-Smith³⁴, Emma Swindells³⁴, Robyn Turnbull³⁴, Tim J. Sloan⁹⁵, Andrew Bosworth^{4,61}, Stephanie Hutchings⁴, Hannah M. Pymont⁴, Anna Casey⁹⁶, Liz Ratcliffe⁹⁶, Christopher R. Jones^{38,97}, Bridget A. Knight^{48,97}, Tanzina Haque⁵⁶, Jennifer Hart⁵⁶, Dianne Irish-Tavares⁵⁶, Eric Witele⁵⁶, Craig Mower⁵⁷, Louisa K. Watson⁵⁷, Jennifer Collins⁶⁶, Gary Eltringham⁵⁶, Dorian Crudgington³⁶, Ben Macklin³⁶, Miren Iturriza-Gomara⁶², Anita O. Lucaci⁶² & Patrick C. McClure⁹⁸

Metadata curation, and sequencing and analysis

Matthew Carlile⁴⁰, Nadine Holmes⁴⁰, Christopher Moore⁴⁰, Nathaniel Storey⁷⁵, Stefan Rooke⁵⁹

Article

Gonzalo Yebra⁵⁹, Noel Craine¹⁴, Malorie Perry¹⁴, Nabil-Fareed Alikhan⁵⁹, Stephen Bridgett³³, Kate F. Cook¹², Christopher Fearn¹², Salman Goudarzi¹², Ronan A. Lyons⁹⁹, Thomas Williams¹⁷, Sam T. Haldenby⁶², Jillian Durham² & Steven Leonard²

Metadata curation, and software and analysis tools

Robert M. Davies²

Project administration, and samples and logistics

Rahul Batra¹⁸, Beth Blane¹⁰, Moira J. Spyer^{19,20,21}, Perinder Smith^{8,100}, Mehmet Yavuz^{48,101}, Rachel J. Williams¹⁹, Adhyana I. K. Mahanama²⁴, Buddhini Samaraweera²⁴, Sophia T. Giris³², Samantha E. Hansford⁴⁸, Angie Green⁷⁶, Charlotte Beaver², Katherine L. Bellis^{2,32}, Matthew J. Dorman², Sally Kay², Liam Prestwood² & Shavanthi Rajatileka²

Project administration, and sequencing and analysis

Joshua Quick¹⁵

Project administration, and software and analysis tools

Radoslaw Poplawski¹⁵

Samples and logistics, and sequencing and analysis

Nicola Reynolds¹⁰², Andrew Mack¹³, Arthur Morriss¹³, Thomas Whalley¹³, Bindi Patel¹⁸, Iliana Georgana²³, Myra Hosmilto²³, Malte L. Pinckert²³, Joanne Stockton¹⁵, John H. Henderson⁴⁴, Amy Hollis⁴⁴, William Stanley⁴⁴, Wen C. Yew⁴⁴, Richard Myers⁴, Alicia Thornton⁴, Alexander Adams¹⁴, Tara Annett¹⁴, Hibo Asad¹⁴, Alec Birchley¹⁴, Jason Coombes¹⁴, Johnathan M. Evans¹⁴, Laia Fina¹⁴, Bree Gatica-Wilcox¹⁴, Lauren Gilbert¹⁴, Lee Graham¹⁴, Jessica Hey¹⁴, Ember Hilvers¹⁴, Sophie Jones¹⁴, Hannah Jones¹⁴, Sara Kumziene-S ummerhayes¹⁴, Caoimhe McKerr¹⁴, Jessica Powell¹⁴, Georgia Pugh¹⁴, Sarah Taylor¹⁴, Alexander J. Trotter⁶², Charlotte A. Williams¹⁹, Leanne M. Kermack³², Benjamin H. Foulkes⁴⁸, Marta Gallis⁴⁸, Hailey R. Hornsby⁴⁸, Stavroula F. Louka⁴⁸, Manoj Pohare⁴⁸, Paige Wolverson⁴⁸, Peijun Zhang⁴⁸, George MacIntyre-Cockett⁷⁶, Amy Trebes⁷⁶, Robin J. Mol¹², Lynne Ferguson¹⁰³, Emily J. Goldstein¹⁰³, Alasdair Maclean¹⁰³ & Rachael Tomb¹⁰³

Samples and logistics, and software and analysis tools

Igor Starinski³¹

Sequencing and analysis, and software and analysis tools

Laura Thomson⁶, Joel Southgate^{3,14}, Moritz U. G. Kraemer⁶³, Jayna Raghwan⁶³, Alex E. Zarebski⁶³, Olivia Boyd⁷, Lily Geidelberg⁷, Chris J. Illingworth⁶⁴, Chris Jackson⁶⁴, David Pascall⁶⁴, Sreenu Vattipally³¹, Timothy M. Freeman⁴⁸, Sharon N. Hsu⁴⁸, Benjamin B. Lindsey⁴⁸, Keith James², Kevin Lewis², Gerry Tonkin-Hill² & Jaime M. Tovar-Corona²

Sequencing and analysis, and visualization

MacGregor Cox¹⁰

Software and analysis tools, and visualization

Khalil Abudahab^{2,46}, Mirko Menegazzo⁴⁶, Ben E. W. Taylor^{2,46}, Corin A. Yeats⁴⁶, Afrida Mukaddas³¹, Derek W. Wright³¹, Leonardo de Oliveira Martins⁵², Rachel Colquhoun¹⁷, Verity Hill¹⁷, Ben Jackson¹⁷, J. T. McCrone¹⁷, Nathan Medd¹⁷, Emily Scher¹⁷ & Jon-Paul Keatley²

Leadership and supervision

Tanya Curran³⁹, Sian Morgan⁷⁷, Patrick Maxwell¹⁰, Ken Smith¹⁰, Sahar Eldirdir⁸⁸, Anita Kenyon⁸⁸, Alison H. Holmes^{68,69}, James R. Price^{68,69}, Tim Wyatt¹⁰⁴, Alison E. Mather⁵², Timofey Skvortsov³³ & John A. Hartley¹⁹

Metadata curation

Martyn Guest¹³, Christine Kitchen¹³, Ian Merrick¹³, Robert Munn¹³, Beatrice Bertolusso⁵⁵, Jessica Lynch⁵⁵, Gabrielle Vernet⁵⁵, Stuart Kirk⁶⁰, Elizabeth Wastnedge⁵⁰, Rachael Stanley⁷⁹, Giles Idle¹⁰⁵, Declan T. Bradley^{33,104}, Jennifer Poyner⁴⁹⁷ & Matilde Mori¹⁰⁶

Project administration

Owen Jones¹³, Victoria Wright⁴⁰, Ellena Brooks¹⁰, Carol M. Churcher¹⁰, Mireille Fragakis¹⁰, Katerina Galai^{4,10}, Andrew Jermy¹⁰, Sarah Judges¹⁰, Georgina M. McManus¹⁰, Kim S. Smith¹⁰, Elaine Westwick¹⁰, Stephen W. Attwood⁶³, Frances Bolt^{68,69}, Alisha Davies¹⁴, Elen De Lacy¹⁴, Fatima Downing¹⁴, Sue Edwards¹⁴, Lizzie Meadows⁵², Sarah Jeremiah²⁴, Nikki Smith⁴⁸ & Luke Foulser²

Samples and logistics

Themoula Charalampous^{16,18}, Amita Patel¹⁸, Louise Berry⁸⁵, Tim Boswell⁸⁵, Vicki M. Fleming⁸⁵, Hannah C. Howson-Wells⁸⁵, Amelia Joseph⁸⁵, Manjinder Khakh⁸⁵, Michelle M. Lister⁸⁵, Paul W. Bird⁸⁶, Karlie Fallon⁸⁶, Thomas Helmer⁸⁶, Claire L. McMurray⁸⁶, Mina Oedra⁸⁶, Jessica Shaw⁸⁶, Julian W. Tang⁸⁶, Nicholas J. Willford⁸⁶, Victoria Blakey⁸⁷, Veena Raviprakash⁸⁷, Nicola Sheriff⁶⁷, Lesley-Anne Williams⁶⁷, Theresa Feltwell¹⁰, Luke Bedford¹⁰⁷, James S. Cargill¹⁰⁸, Warwick Hughes¹⁰⁸, Jonathan Moore¹⁰⁹, Susanne Stonehouse¹⁰⁹, Laura Atkinson⁷⁵, Jack C. D. Lee⁷⁵, Divya Shah⁷⁵, Adela Alcolea^{8,100}, Natasha Ohemeng-Kumi^{8,100}, John Ramble^{8,100}, Jasveen Sehmi^{8,100}, Rebecca Williams⁵⁵, Wendy Chatterton⁶⁰, Monika Pusok⁶⁰, William Everson³⁵, Anibolina Castigador¹¹⁰, Emily Macnaughton¹¹⁰, Kate El Bouzidi¹¹¹, Temi Lampejo¹¹¹, Malur Sudhanva¹¹¹, Cassie Breen¹¹², Graciela Sluga⁹³, Shazaad S. Y. Ahmad^{4,113}, Ryan P. George¹¹³, Nicholas W. Machin^{4,113}, Debbie Binns⁵⁵, Victoria James⁵⁵, Rachel Blacow³¹, Lindsay Coupland⁷⁹, Louise Smith⁷⁴, Edward Barton⁹³, Debra Padgett⁹³, Garren Scott⁹³, Aidan Cross¹¹⁴, Mariyam Mirfenderesky¹¹⁴, Jane Greenaway⁸⁴, Kevin Cole¹⁰⁵, Phillip Clarke⁸⁵, Nichola Duckworth⁸⁵, Sarah Walsh⁸⁵, Kelly Bicknell⁴⁹, Robert Impney⁴⁹, Sarah Wyllie⁴⁹, Richard Hopes⁴, Chloe Bishop⁴

Vicki Chalker⁴, Ian Harrison⁴, Laura Gifford¹⁴, Zoltan Molnar³³, Cressida Auckland⁹⁷, Cariad Evans^{48,101}, Kate Johnson^{48,101}, David G. Partridge^{48,101}, Mohammad Raza^{48,101}, Paul Baker³⁷, Stephen Bonner³⁷, Sarah Essex³⁷, Leanne J. Murray³⁷, Andrew I. Lawton¹¹⁵, Shirelle Burton-Fanning⁶⁶, Brendan A. I. Payne⁶⁶, Sheila Waugh⁶⁶, Andrea N. Gomes¹¹⁶, Maimuna Kimuli¹¹⁶, Darren R. Murray¹¹⁶, Paula Ashfield¹¹⁷, Donald Dobie¹¹⁷, Fiona Ashford⁷⁸, Angus Best¹⁶, Liam Crawford⁷⁸, Nicola Cumley⁷⁸, Megan Mayhew⁷⁸, Oliver Megram⁷⁸, Jeremy Mirza⁷⁸, Emma Moles-Garcia⁷⁸, Benita Percival⁷⁸, Leah Ensell¹⁹, Helen L. Lowe¹⁹, Laurentiu Maftel¹⁹, Matteo Mondani¹⁹, Nicola J. Chaloner³⁷, Benjamin J. Cogger³⁷, Lisa J. Easton³⁷, Hannah Huckson³⁷, Jonathan Lewis³⁷, Sarah Lowdon³⁷, Cassandra S. Malone³⁷, Florence Munemo³⁷, Manasa Mutingwende³⁷, Roberto Nicodemi³⁷, Olga Podplomyk³⁷, Thomas Somassa³⁷, Andrew Beggs¹¹⁸, Alex Richter¹¹⁸, Claire Cormie³², Joana Dias³², Sally Forrest³², Ellen E. Higginson³², Mailis Maes³², Jamie Young³², Rose K. Davidson⁵³, Kathryn A. Jackson⁶², Lance Turtle⁶², Alexander J. Keeley⁴⁸, Jonathan Ball⁹⁸, Timothy Byaruhanga⁹⁸, Joseph G. Chappell⁹⁸, Jayasree Dey⁹⁸, Jack D. Hill⁹⁸, Emily J. Park⁹⁸, Arezou Fanaie¹¹⁹, Rachel A. Hilson¹¹⁹, Geraldine Yaze¹¹⁹ & Stephanie Lo²

Sequencing and analysis

Safiah Affifi⁷⁷, Robert Beer⁷⁷, Joshua Maksimovic⁷⁷, Kathryn McCluggage Masters⁷⁷, Karla Spellman⁷⁷, Catherine Bresner¹³, William Fuller¹³, Angela Marchbank¹³, Trudy Workman¹³, Ekaterina Shelest^{11,58}, Johnny Debebe⁴⁰, Fei Sang⁴⁰, Marina Escalera Zamudio⁶³, Sarah Francisco⁶³, Bernardo Gutierrez⁶³, Tetyana I. Vasylyeva⁶³, Flavia Flaviani¹²⁰, Manon Ragonnet-Cronin⁷, Katherine L. Smollett⁷³, Alice Broos³¹, Daniel Mai³¹, Jenna Nichols³¹, Kyriaki Nomikou³¹, Lily Tong³¹, Ioulia Tsatsani³¹, Sarah O'Brien²¹, Steven Rushton¹²¹, Roy Sanderson¹²¹, Jon Perkins⁵¹, Seb Cotton⁵⁰, Abbie Gallagher⁵⁰, Elias Allara^{4,32}, Clare Pearson^{4,32}, David Bibby⁴, Gavin Dabrera⁴, Nicholas Ellaby⁴, Eileen Gallagher⁴, Jonathan Hubb⁴, Angie Lackenby⁴, David Lee⁴, Nikos Manesis⁴, Tamyo Mbisa⁴, Steven Platt⁴, Katherine A. Twohig⁴, Mari Morgan¹⁴, Alp Aydin⁵², David J. Baker⁵², Ebenezer Foster-Nyrako⁵², Sophie J. Prosolek⁵², Steven Rudder⁵², Chris Baxter³³, Silvia F. Carvalho³³, Deborah Lavin³³, Arun Mariappan³³, Clara Radulescu³³, Aditi Singh³³, Miao Tang³³, Helen Morcrette³⁷, Nadua Bayzid¹⁹, Marius Cotic¹⁹, Carlos E. Balcazar¹⁷, Michael D. Gallagher¹⁷, Daniel Maloney¹⁷, Thomas D. Stanton¹⁷, Kathleen A. Williamson¹⁷, Robin Manley³⁸, Michelle L. Michelsen³⁸, Christine M. Sambles³⁸, David J. Studholme³⁸, Joanna Warwick-Dugdale³⁸, Richard Eccles⁵², Matthew Gemmill⁶², Richard Gregory⁶², Margaret Hughes⁶², Charlotte Nelson⁶², Lucille Rainbow⁶², Edith E. Vamos⁶², Hermione J. Webster⁶², Mark Whitehead⁶², Claudia Wierzbicki⁶², Adrienn Angyal⁴⁸, Luke R. Green⁴⁸, Max Whiteley⁴⁸, Emma Betteridge², Iraad F. Bronner², Ben W. Farr², Scott Goodwin², Stefanie V. Lensing², Shane A. McCarthy^{2,32}, Michael A. Quail², Diana Rajan², Nicholas M. Redshaw², Carol Scott², Lesley Shirley² & Scott A. J. Thurston²

Software and analysis tools

Will Rowe¹⁵, Amy Gaskin¹⁴, Thanh Le-Viet⁵², James Bonfield², Jennifer Liddle² & Andrew Whitwham²

¹¹Centre for Enzyme Innovation, University of Portsmouth, Portsmouth, UK. ¹²School of Pharmacy & Biomedical Sciences, University of Portsmouth, Portsmouth, UK. ¹³Cardiff University, Cardiff, UK. ¹⁴Public Health Wales, Cardiff, UK. ¹⁵Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. ¹⁶King's College London, London, UK. ¹⁷University of Edinburgh, Edinburgh, UK. ¹⁸Centre for Clinical Infection and Diagnostics Research, Department of Infectious Diseases, Guy's and St Thomas' NHS Foundation Trust, London, UK. ¹⁹University College London Hospitals NHS Foundation Trust, London, UK. ²⁰Advanced Pathogen Diagnostics Unit, University College London Hospital, London, UK. ²¹Great Ormond Street Institute of Child Health, University College London, London, UK. ²²Department of Infectious Diseases and Microbiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. ²³Division of Virology, Department of Pathology, University of Cambridge, Cambridge, UK. ²⁴University Hospital Southampton NHS Foundation Trust, Southampton, UK. ²⁵School of Health Sciences, University of Southampton, Southampton, UK. ²⁶School of Medicine, University of Southampton, Southampton, UK. ²⁷Division of Infection and Immunity, University College London, London, UK. ²⁸University of Brighton, Brighton, UK. ²⁹Institute for Infection and Immunity, St George's University of London, London, UK. ³⁰Infection Care Group, St George's University Hospitals NHS Foundation Trust, London, UK. ³¹MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. ³²University of Cambridge, Cambridge, UK. ³³Queen's University Belfast, Belfast, UK. ³⁴Blackpool Teaching Hospitals NHS Foundation Trust, Blackpool, UK. ³⁵Hull University Teaching Hospitals NHS Trust, Hull, UK. ³⁶University Hospitals Dorset NHS Foundation Trust, Poole, UK. ³⁷University Hospitals Sussex NHS Foundation Trust, Worthing, UK. ³⁸University of Exeter, Exeter, UK. ³⁹Belfast Health & Social Care Trust, Belfast, UK. ⁴⁰Deep Seq, School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK. ⁴¹East Kent Hospitals University NHS Foundation Trust, Canterbury, UK. ⁴²University of Kent, Canterbury, UK. ⁴³Hub for Biotechnology in the Built Environment, Northumbria University, Newcastle upon Tyne, UK. ⁴⁴Northumbria University, Newcastle upon Tyne, UK. ⁴⁵NU-OMICS, Northumbria University, Newcastle upon Tyne, UK. ⁴⁶Centre for Genomic Pathogen Surveillance, University of Oxford, Oxford, UK. ⁴⁷Public Health England, Cambridge, Cambridge, UK. ⁴⁸University of Sheffield, Sheffield, UK. ⁴⁹Portsmouth Hospitals University NHS Trust, Portsmouth, UK. ⁵⁰NHS Lothian, Edinburgh, UK. ⁵¹NHS Greater Glasgow and Clyde, Glasgow, UK. ⁵²Quadram Institute Bioscience, Norwich, UK. ⁵³University of East Anglia, Norwich, UK. ⁵⁴University of Oxford, Oxford, UK. ⁵⁵Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK. ⁵⁶Royal Free London NHS Foundation Trust, London, UK. ⁵⁷South Tees Hospitals NHS Foundation Trust, Middlesbrough, UK. ⁵⁸School of Biological Sciences, University of Portsmouth, Portsmouth, UK. ⁵⁹Public Health Scotland, Edinburgh, UK. ⁶⁰Health Services Laboratories, London, UK. ⁶¹Heartlands Hospital, Birmingham, UK. ⁶²University of Liverpool, Liverpool, UK. ⁶³Department of Zoology, University of Oxford, Oxford, UK. ⁶⁴MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. ⁶⁵Microbiology Department, Buckinghamshire Healthcare NHS Trust, Aylesbury, UK. ⁶⁶The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. ⁶⁷University of St Andrews, St Andrews, UK. ⁶⁸Imperial College Healthcare NHS Trust, London, UK. ⁶⁹NIHR Health Protection Research Unit in HCAI and AMR, Imperial College London, London, UK. ⁷⁰Cambridge University Hospitals NHS Foundation Trust,

Cambridge, UK. ⁷¹Department of Microbiology, South West London Pathology, London, UK. ⁷²Betsi Cadwaladr University Health Board, Wrexham, UK. ⁷³Institute of Biodiversity, Animal Health & Comparative Medicine, Glasgow, UK. ⁷⁴Norfolk County Council, Norfolk, UK. ⁷⁵Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. ⁷⁶Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ⁷⁷Cardiff and Vale University Health Board, Cardiff, UK. ⁷⁸Turnkey Laboratory, University of Birmingham, Birmingham, UK. ⁷⁹Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, UK. ⁸⁰Royal Brompton and Harefield Hospitals, London, UK. ⁸¹The Queen Elizabeth Hospital King's Lynn NHS Foundation Trust, King's Lynn, UK. ⁸²Whittington Health NHS Trust, London, UK. ⁸³Barking, Havering and Redbridge University Hospitals NHS Trust, London, UK. ⁸⁴Bournemouth University, Bournemouth, UK. ⁸⁵Clinical Microbiology Department, Queens Medical Centre, Nottingham University Hospitals NHS Trust, Nottingham, UK. ⁸⁶Clinical Microbiology, University Hospitals of Leicester NHS Trust, Leicester, UK. ⁸⁷County Durham and Darlington NHS Foundation Trust, Durham, UK. ⁸⁸Department of Microbiology, Kettering General Hospital, Kettering, UK. ⁸⁹Barts Health NHS Trust, London, UK. ⁹⁰North West London Pathology, London, UK. ⁹¹Maidstone and Tunbridge Wells NHS Trust, Maidstone, UK. ⁹²Microbiology, Royal Oldham Hospital, Oldham, UK. ⁹³North Cumbria Integrated Care NHS Foundation Trust, Carlisle, UK. ⁹⁴North Tees and Hartlepool NHS Foundation Trust, Stockton on Tees, UK. ⁹⁵Path Links, Northern Lincolnshire

and Goole NHS Foundation Trust, Grimsby, UK. ⁹⁶Queen Elizabeth Hospital, Birmingham, UK. ⁹⁷Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. ⁹⁸Virology, School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK. ⁹⁹Swansea University, Swansea, UK. ¹⁰⁰Viapath, Guy's and St Thomas' NHS Foundation Trust, and King's College Hospital NHS Foundation Trust, London, UK. ¹⁰¹Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. ¹⁰²Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. ¹⁰³West of Scotland Specialist Virology Centre, NHS Greater Glasgow and Clyde, Glasgow, UK. ¹⁰⁴Public Health Agency, Belfast, UK. ¹⁰⁵Northumbria Healthcare NHS Foundation Trust, Newcastle upon Tyne, UK. ¹⁰⁶University of Southampton, Southampton, UK. ¹⁰⁷East Suffolk and North Essex NHS Foundation Trust, Colchester, UK. ¹⁰⁸East Sussex Healthcare NHS Trust, St Leonards-on-Sea, UK. ¹⁰⁹Gateshead Health NHS Foundation Trust, Gateshead, UK. ¹¹⁰Isle of Wight NHS Trust, Newport, UK. ¹¹¹King's College Hospital NHS Foundation Trust, London, UK. ¹¹²Liverpool Clinical Laboratories, Liverpool, UK. ¹¹³Manchester University NHS Foundation Trust, Manchester, UK. ¹¹⁴North Middlesex University Hospital NHS Trust, London, UK. ¹¹⁵Southwest Pathology Services, Taunton, UK. ¹¹⁶The Royal Marsden NHS Foundation Trust, London, UK. ¹¹⁷The Royal Wolverhampton NHS Trust, Wolverhampton, UK. ¹¹⁸University of Birmingham, Birmingham, UK. ¹¹⁹Watford General Hospital, Watford, UK. ¹²⁰Guy's and St Thomas' Biomedical Research Centre, London, UK. ¹²¹Newcastle University, Newcastle, UK.

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) were downloaded from <https://coronavirus.data.gov.uk/> spanning the date range from 1 September 2020 to 30 June 2021 for 315 English LTLAs (downloaded on 20 July 2021). 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 the 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 were downloaded from the Office of National Statistics (ONS; <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>).

SARS-CoV-2 surveillance sequencing

In total, 281,178 tests (September 2020 to June 2021) were collected as part of random surveillance of positive tests of residents of England from four Pillar 2 Lighthouse laboratories. The samples were collected between 1 September 2020 and 26 June 2021. A random selection of samples was taken, after excluding those that were 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 Pangolin⁵, according to the latest lineage definitions at the time, except for B.1.617, which we reanalysed after the designation of sublineages B.1.617.1, B.1.617.2 and B.1.617.3. Lineage prevalence was computed from 281,178 genome sequences. The genomes were mapped to the same 315 English LTLAs as for the 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.

Finally, 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, unless the lineage has been designated a VOC, variant under investigation (VUI) or variant in monitoring by Public Health England³².

Spatiotemporal 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 historical prevalence and transmission parameters. In the following, t denotes time and is measured in days. We use the convention that bold lowercase symbols, such as \mathbf{b} , indicate vectors.

Motivation

Suppose that $\mathbf{x}'(t) = (\mathbf{b} + r_0(t)) \cdot \mathbf{x}(t)$ describes the ordinary differential equation (ODE) for the viral dynamics for a set of l different lineages. Here $r_0(t)$ is a scalar time-dependent logarithmic growth rate that is thought to reflect lineage-independent transmission determinants, which changes over time in response to behaviour, non-pharmaceutical interventions (NPIs) and immunity. This reflects a scenario in which the lineages differ only in terms of the intensity of transmission, but

not the intergeneration time distribution. The ODE is solved by $\mathbf{x}(t) = e^{c+\mathbf{b}t + \int_{t_0}^t r_0(t)dt} = e^{c+\mathbf{b}t} v(t)$. The term $v(t)$ contributes the same factor to each lineage and therefore drops from the relative proportions of lineages $\mathbf{p}(t) = \frac{\mathbf{x}(t)}{\sum \mathbf{x}(t)} \propto e^{c+\mathbf{b}t}$.

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

Incidence

In the following text, 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 $\boldsymbol{\lambda}(t) = \log \boldsymbol{\mu}(t) - \log(s)$ the logarithmic daily incidence per capita at time t in each of the 315 LTLAs.

Suppose $f(t)$ is the daily number of new infections caused by the number of people infected at time t . As new cases are noticed and tested only after a delay u with distribution g , the observed number of cases $f^*(t)$ will be given by the convolution

$$f^*(t) = \int_0^{\infty} g(u)f(t-u)du = (g * f)(t).$$

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 5 d of symptom onset. To account for these factors, the log normal incubation time distribution from ref. ⁴⁶ is scaled by the equivalent of changing the mean by 2 d. The convolution shifts cases approximately 6 d into the future and also spreads them out according to the width of g (Extended Data Fig. 2a).

To parameterize 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 $\mathbf{f}(t) = (f_1^*(t), \dots, f_k^*(t))$. 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 , $\mathbf{f}^*(t) = (f_1^*(t), \dots, f_k^*(t))$ as outlined above and used to fit the logarithmic incidence. The derivatives of the original basis $\mathbf{f}'(t)$ are used to calculate the underlying growth rates and R_t values, as shown further below. The convolved spline basis $\mathbf{f}^*(t)$ is used to fit the per capita incidence in each LTLA as (Extended Data Fig. 2b):

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

This implies that fitting the incidence function for each of the m local

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

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

The s.d. of the prior regularizes 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 regularized monthly splines reflect trends on the scale of several weeks and are therefore subject to less noise.

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

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

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

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

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

The equation above assumes that all elements of $\mathbf{X}(t)$ 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 logarithmic growth rate $\mathbf{r}(t) = \boldsymbol{\lambda}'(t)$ of the local epidemic simplifies to:

$$\mathbf{r}(t) = \mathbf{B} \times \mathbf{f}'(t),$$

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

To express the daily growth rate as an approximate reproductive number R_t , one needs to consider the distribution of the intergeneration time, which is assumed to be gamma distributed with mean 6.3 d ($\alpha = 2.29, \beta = 0.36$)⁴⁶. The R_t value can be expressed as a Laplace transform of the intergeneration time distribution⁴⁷. Effectively, this shortens the relative time period because the exponential dynamics put disproportionately 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 $\bar{\tau}_e = 5.1$ d, 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),

$$\log(\boldsymbol{\rho}(t)) \approx \frac{d \log(\boldsymbol{\mu}(t))}{dt} \bar{\tau}_e = \mathbf{r}(t) \bar{\tau}_e$$

Thus, the overall growth rate scaled to an effective inter generation time of 5.1 d 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 ref. ⁴⁸, but shifted according to the typical delay from infection to test (Extended Data Fig. 2b).

Genomic prevalence

The dynamics of the relative frequency $\mathbf{P}(t)$ of each lineage was modelled using a logistic-linear model in each LTLA, as described above. The logistic prevalence of each lineage in each LTLA is defined as $\mathbf{L}(t) = \text{logit}(\mathbf{P}(t))$. This is modelled using the piecewise linear expression

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

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

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

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 a higher growth rate compared with 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 $\mathbf{t}_0 \sim \text{Unif}(-14, 0) + \mathbf{t}_0^{\text{obs}}$ prior to the first observation $\mathbf{t}_0^{\text{obs}}$ was chosen.

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

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

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

$$\mathbf{b} \sim N(0, 0.2)$$

$$\mathbf{c} \sim N(-10, 5)$$

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

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

The scalar parameter $\alpha_0 = 0.01$ can be interpreted as a weak prior with expectation $1/n$, making the model less sensitive to the introduction of single new lineages, which can otherwise exert a very strong effect. Furthermore, the array $\boldsymbol{\alpha}_1 = \frac{\text{cases}}{2}$ increases the variance to account for the fact that, especially at high sequencing coverage (genomes \approx cases), cases and therefore genomes are likely to be correlated and overdispersed as they may derive from a single transmission event. Other choices such as $\alpha_1 = 1,000$, which make the model converge to a standard multinomial, leave the conclusions qualitatively unchanged. This model aspect is illustrated in Extended Data Fig. 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 $\mathbf{P}(t)$ for lineage j at each time point

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

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

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

By adding the log-transformed growth rate fold changes \mathbf{b} and subtracting the average log-transformed growth rate change $\mathbf{P}(t) \times \mathbf{b}$, it follows that $\mathbf{R}_{i,\cdot}(t) = \mathbf{R}_{i,0}(t) e^{\bar{\tau}_e \mathbf{b}}$, where $\mathbf{R}_{i,0}(t)$ is the R_t value of the reference lineage $j = 0$ (for which $\mathbf{b}_0 = 0$) 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 $e^{\bar{\tau}_e \mathbf{b}}$.

Inference

The model was implemented in `numpyro`^{49,50} and fitted using stochastic variational inference⁵¹. Guide functions were multivariate normal distributions for each row (corresponding to an LTLA) of **B**, **C** to preserve the correlations across lineages and time as well as for (**b**, **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 the release dates 17 April 2021 (for the analysis of the lineages A.23.1, B.1.1.318, B.1.351 and B.1.525) and 5 May 2021 (for the analysis of the B.1.617 sublineages). We next extracted a subalignment from each lineage (according to the 1 April 2021 version of PANGOLin for the 17 April 2021 alignment and the 23 April 2021 version of PANGOLin for the 5 May 2021 alignment) and, for each subalignment, we inferred a phylogeny through maximum likelihood using `FastTree2` (v.2.1.11)⁵² with the default options and GTR substitution model⁵³.

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 measured clade sizes for one specific example parsimonious migration history. We counted only introduction events into the UK that resulted in at least one descendant from the set of UK samples that we considered in this work for our hierarchical Bayesian model; similarly, we measured clade sizes by the number of UK samples considered here included in such clades. Multiple occurrences of identical sequences were counted as separate cases, as this helped us to identify rapid SARS-CoV-2 spread.

When using parsimony, we considered only 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 of the non-UK locations into a single one). Furthermore, as SARS-CoV-2 phylogenies present substantial numbers of polytomies, that is, phylogenetic nodes where the tree topology cannot be reconstructed due to a 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 inference⁵⁴ in that it enables 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 ref.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^{55,56}. The main advantage of our approach is that, owing 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 enables 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 enables 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 world⁵⁵ 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: (1) 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; (2) 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; (3) the size so far of the UK clade the current node is part of, conditional on it being UK; and (4) 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 or 1 migration events 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 also used the new Bayesian phylogenetic approach Thorney BEAST¹⁶ (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; we also used `treetime`⁵⁷ v.0.8.2 to estimate a timed tree and branch divergences for use in Thorney BEAST. We used a two-state (UK and non-UK) migration model⁵⁴ of migration to infer introductions into the UK but again counted, from the posterior sample trees, only UK clades with at least one UK surveillance sample. We used a `Skygrid`⁵⁸ tree coalescent prior with six time intervals. The comparison of parsimony and Bayesian estimates is shown in Extended Data Fig. 8d.

ONS infection survey analysis

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

Comparison of ONS incidence estimates with hospitalization, case and death rates was conducted by estimating infection trajectories separately from observed cases, hospitalizations and deaths^{59,60}, involving them with estimated PCR detection curves⁶¹, and dividing the resulting PCR prevalence estimates by the estimated prevalence from the ONS Community Infection Survey at the midpoints of the two-week intervals over which prevalence was reported in the survey.

Maps

Maps were plotted using LTLA shapefiles (<https://geoportal.statistics.gov.uk/datasets/69dc11c7386943b4ad8893c45648b1e1>), sourced from the ONS, which is licensed under the Open Government Licence v.3.0.

Limitations

A main limitation of the analysis is that the transmission model is deterministic, whereas the spread of variants is a stochastic process. Although the logistic growth assumption is a consistent estimator of the average transmission dynamics, individual outbreaks may deviate from these averages and therefore produce unreliable estimates.

Stochastic growth effects are accounted for only in terms of (uncorrelated) overdispersion and the offset at the time of the introduction. For these reasons, the estimated growth rates may not accurately reflect the viral transmissibility, especially at a low prevalence. It is therefore important to assess whether consistent growth patterns in multiple independent areas are observed. We note that the posterior distribution

of the growth rates of rare variants tends to be biased to the baseline due to the centred prior.

In its current form, the model accounts for only a single introduction event per LTLA. Although 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 through 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 of altered transmission. Biological mechanisms include a higher viral load, longer infectivity or greater susceptibility. Lineages could potentially differ by their intergeneration time, which would lead to nonlinear scaling. Here we did not find convincing evidence in incidence data for such effects, in contrast to previous reports²³. However, contact-tracing data indicate that the intergeneration time may be shortening for more transmissible lineages such as Delta^{33,62}. Cases of the Beta and Gamma VOCs may have been more intensely contact traced and triggered asymptomatic surge testing in some postcode areas. This may have reduced the observed growth rates relative to other lineages.

Lineages, such as Beta, Gamma or Delta also differ in their ability to evade previous 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 definition⁵ assigns lineages to geographical clusters, which have by definition expanded, and this 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 on the basis of the total number of positive PCR tests, it may be influenced by testing capacity; the total number of tests approximately tripled between September 2020 and March 2021. This can potentially lead to a time trend in recorded cases and therefore baseline R_t values if the access to testing changed, for example, by too few tests being available during periods of high incidence, or changes to the eligibility to intermittently test with fewer symptoms. Generally, the observed incidence was in good agreement with representative cross-sectional estimates from the ONS^{63,64}, except for a period of peak incidence from late December 2020 to January 2021 (Extended Data Fig. 1d). Values after 8 March 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. Positive cases were usually confirmed by PCR and counted only once.

The modelled curves are smoothed over intervals of approximately 7 d using cubic splines, creating the 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 (that is, constant R_t values per week) leaves the overall conclusions and extracted parameters broadly unchanged.

Ethical approval

This study was performed 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 COG-UK study protocol was approved by the Public Health England Research Ethics Governance Group.

Reporting summary

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

Data availability

PCR test data are publicly available online (<https://coronavirus.data.gov.uk/>). A filtered, privacy conserving version of the lineage–LTLA–week dataset is publicly available online (<https://covid19.sanger.ac.uk/downloads>) and enables strong reproduction of our results, despite a small number of cells having been suppressed to avoid disclosure. Full SARS-CoV-2 genome data and geolocations can be obtained under controlled access from <https://www.cogconsortium.uk/data/>. Application for full data access requires a description of the planned analysis and can be initiated at coguk_DataAccess@medschl.cam.ac.uk. The data and a version of the analysis with fewer lineages can be interactively explored at <https://covid19.sanger.ac.uk>. Source data are provided with this paper.

Code availability

The genomic surveillance model is implemented in Python and available at GitHub (<https://github.com/gerstung-lab/genomicsurveillance>) and as a PyPI package (genomicsurveillance). Specific code for the analyses of this study can be found as individual Google colab notebooks in the same repository. These were run using Python v.3.7.1 (packages: matplotlib (v.3.4.1), numpy (v.1.20.2), pandas (v.1.2.3), scikit-learn (v.0.19.1), scipy (v.1.6.2), seaborn (v.0.11.1), jax (v.0.2.8), genomicsurveillance (v.0.4.0), numpyro (v.0.4.0)). The phylogeographic analyses were performed using Thorney Beast (v.0.1.1) and <https://github.com/NicolaDM/phylogeographySARS-CoV-2>. Code for the ONS infection survey analysis is available at GitHub (https://github.com/jhellewell14/ons_severity_estimates).

46. Bi, Q. et al. 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).
47. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. Biol. Sci.* **274**, 599–604 (2007).
48. 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).
49. Bingham, E. et al. Pyro: deep universal probabilistic programming. Preprint at <http://arxiv.org/abs/1810.09538> (2018).
50. Phan, D., Pradhan, N. & Jankowiak, M. Composable effects for flexible and accelerated probabilistic programming in NumPyro. Preprint at <http://arxiv.org/abs/1912.11554> (2019).
51. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
52. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
53. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
54. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
55. 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).
56. Turakhia, Y. et al. Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
57. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: maximum-likelihood phylogenetic analysis. *Virus Evol.* **4**, vex042 (2018).
58. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
59. Sherratt, K. et al. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England. *Philos. Trans. Royal Soc. B* **376** <https://doi.org/10.1098/RSTB.2020.0283> (2021).
60. Abbott, S. et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* **5**, 112 (2020).
61. Hellewell, J. et al. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Med.* **19**, 106 (2021).
62. Hart, W. S. et al. Inference of SARS-CoV-2 generation times using UK household data. Preprint at [medRxiv https://doi.org/10.1101/2021.05.27.21257936](https://doi.org/10.1101/2021.05.27.21257936) (2021).
63. Pouwels, K. B. et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Publ. Health* **6**, e30–e38 (2021).

Article

64. Donnarumma, K. S. *Coronavirus (COVID-19) Infection Survey, UK* (ONS, 2021); <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveysurvey/23april2021>

Acknowledgements We thank E. Allara (Cambridge) and G. Whitton (Sanger) for providing outer postcodes to LTLA mappings; R. Beale for comments and J. McCrone for setting up Thorney Beast analysis; all of the contributors who submitted genome sequences to GISAID (acknowledgement tables for individual sequences are provided at GitHub; <https://github.com/NicolaDM/phylogeographySARS-CoV-2>); and our colleagues at EMBL-EBI, the Wellcome Sanger Institute and COG-UK for discussions and comments on this manuscript. 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. Additional sequence generation was funded by the Department of Health and Social Care. H.S.V., J.P.G. and M.G. are supported by a grant from the Department of Health and Social Care. A.W.J., E.B. and M.G. are beneficiaries from grant NNF17OC0027594 from the Novo Nordisk Foundation. E.V. is supported by Wellcome Trust grant 220885/Z/20/Z. T.S. is supported by grant 210918/Z/18/Z, and J.H. and S.F. by grant 210758/Z/18/Z from the Wellcome Trust. H.S.V., N.D.M., A.W.J., N.G., E.B. and M.G. are supported by EMBL.

Author contributions H.S.V. and M.G. developed the analysis code, which H.S.V. implemented with input from A.W.J.; H.S.V. created most of the figures. M.S. analysed, annotated and

aggregated viral genome data. N.D.M. conducted phylogeographic analyses supervised by N.G.; T.S., R.G., M.S. and H.S.V. developed the interactive spatiotemporal viewer. T.N., F.S., I.H., R.A., C.A., S.G., D.J., I.J., C.S., J.S., T.S. and M.S. analysed genomic surveillance data under the supervision of D.K., M.C., I.M. and J.C.B.; J.H. and S.F. analysed ONS data and helped with epidemiological modelling and data interpretation. E.V. analysed growth rates and helped with data interpretation. E.B. and J.P.G. supervised H.S.V. and helped with data interpretation. J.C.B. and M.G. supervised the analysis with advice from I.M.; M.G., H.S.V., M.S., N.D.M., T.S., I.M. and J.C.B. wrote the manuscript with input from all of the co-authors.

Funding Open access funding provided by Deutsches Krebsforschungszentrum (DKFZ).

Competing interests E.B. is a paid consultant of Oxford Nanopore.

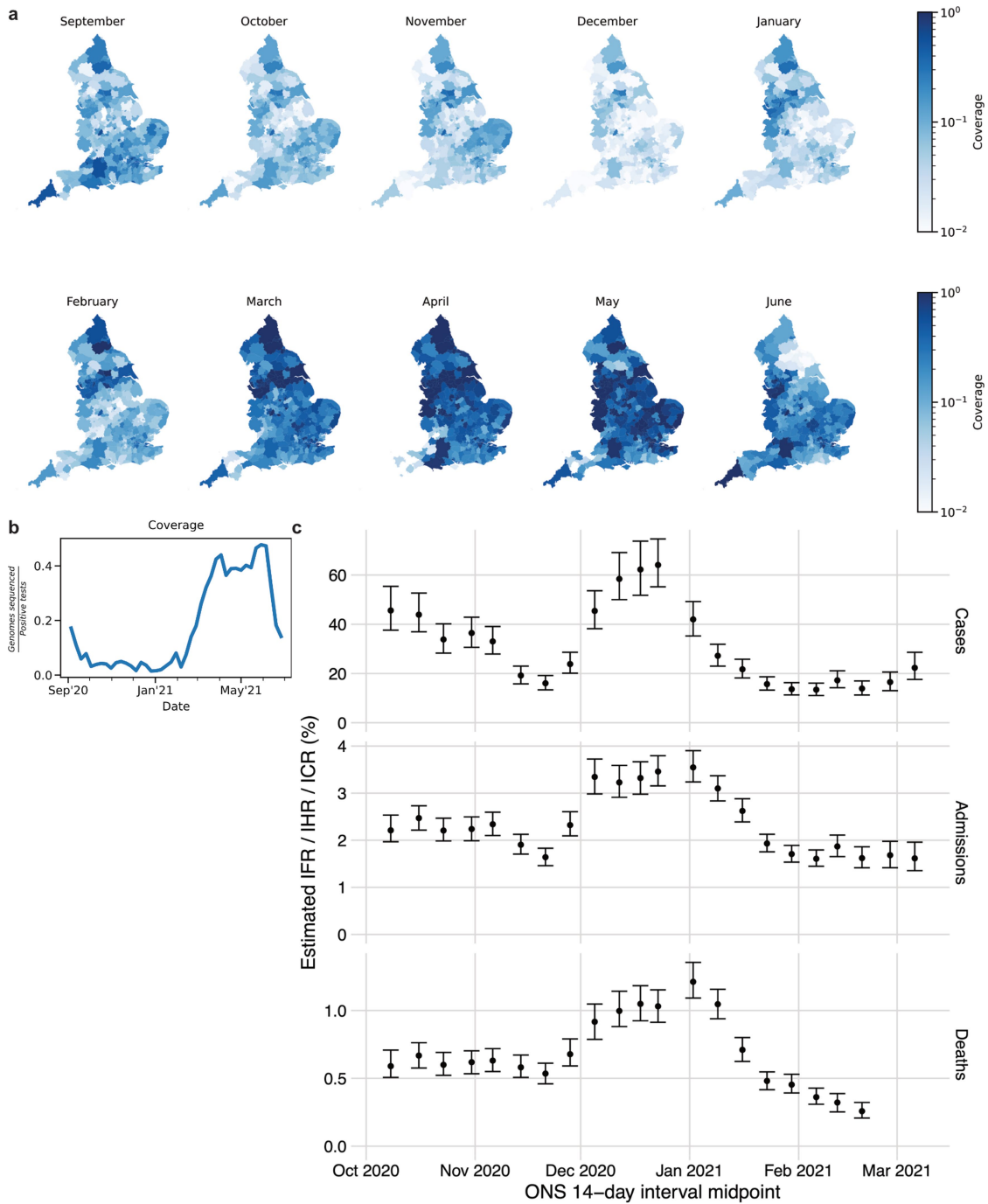
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-021-04069-y>.

Correspondence and requests for materials should be addressed to Jeffrey C. Barrett or Moritz Gerstung.

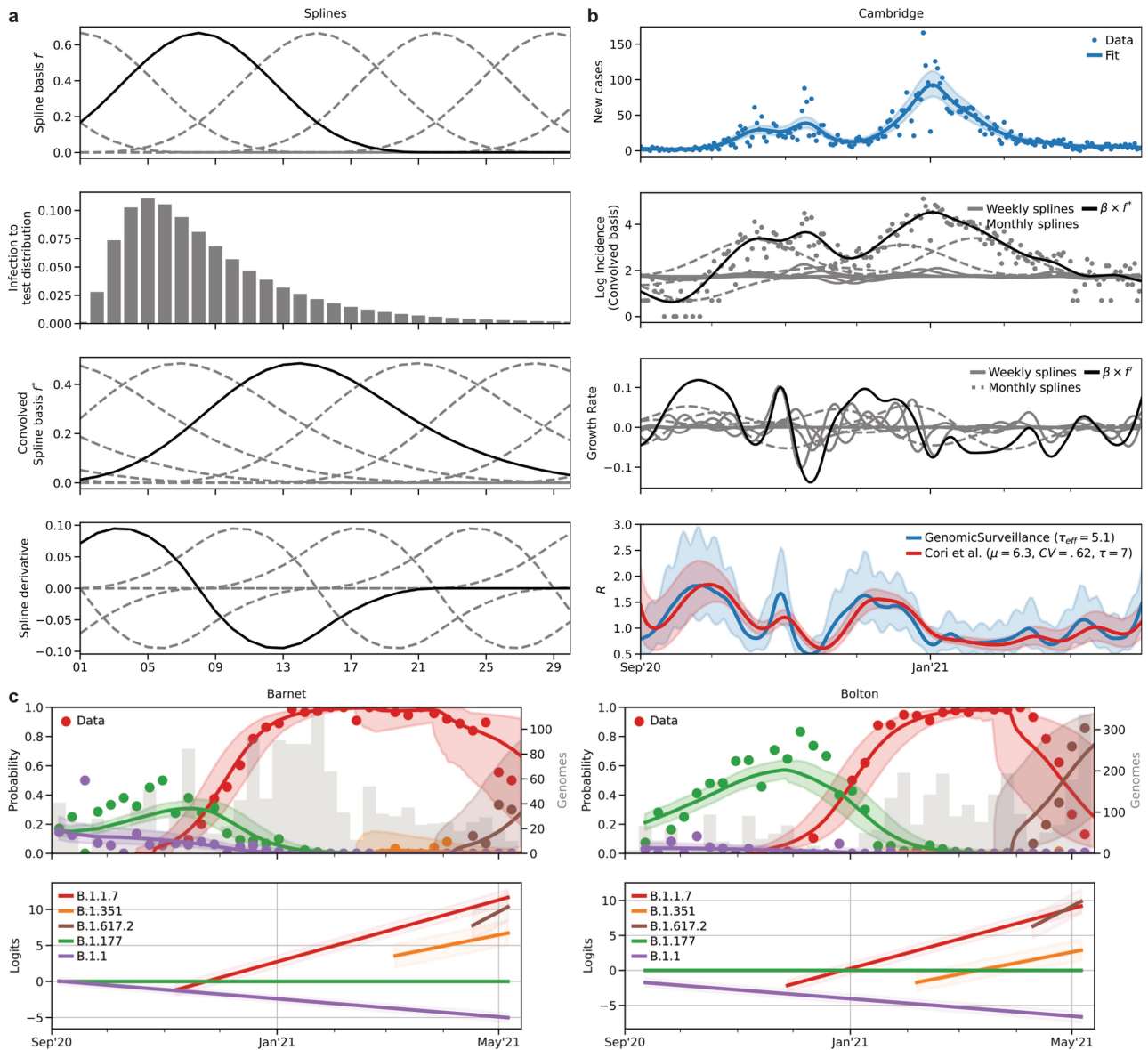
Peer review information *Nature* thanks Tulio De Oliveira, Philippe Lemey and Matthew Scotch for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



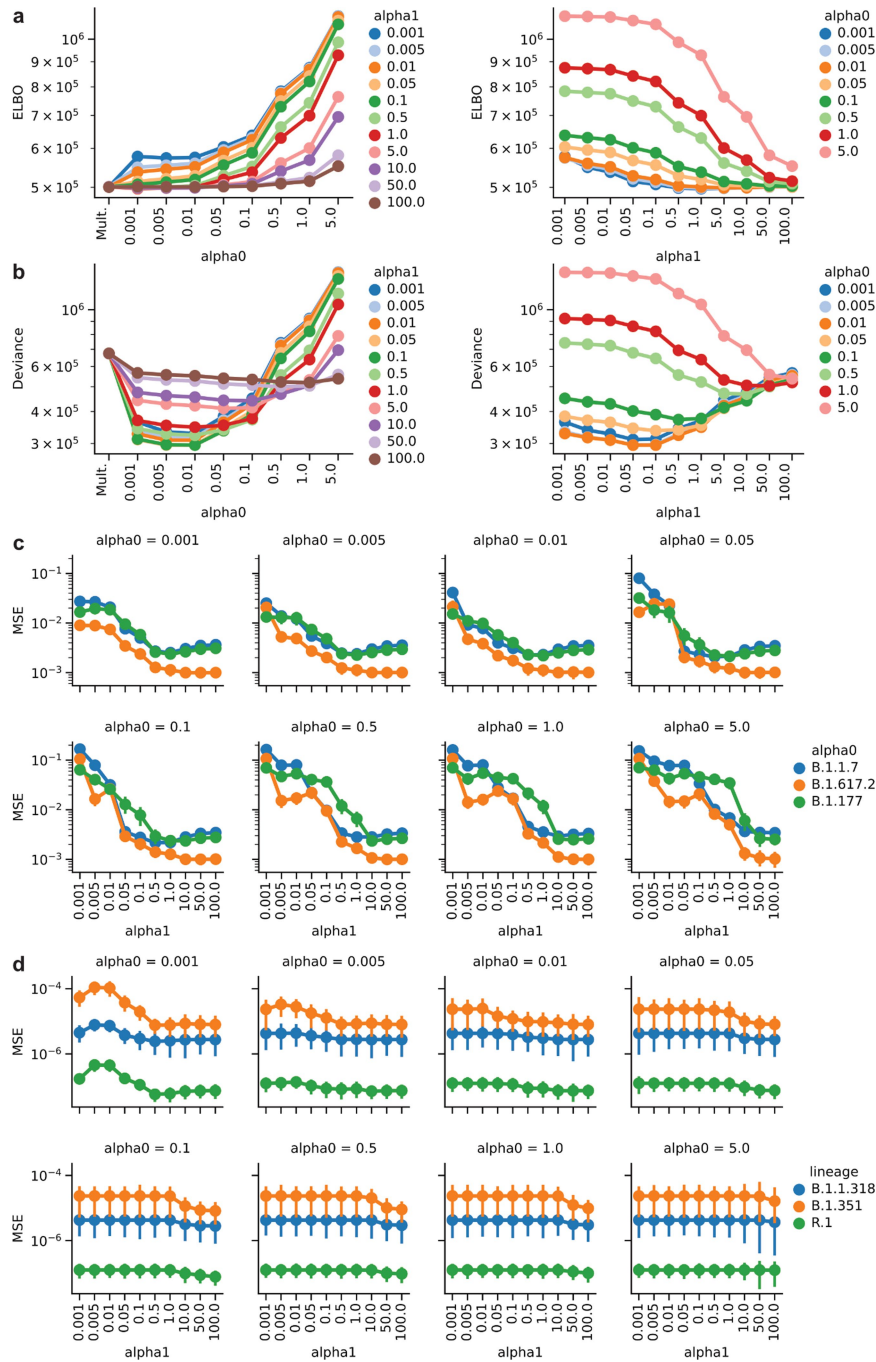
Extended Data Fig. 1 | SARS-CoV-2 surveillance sequencing in England between September 2020 and June 2021. a. Local monthly coverage across 315 LTLAs. **b.** Weekly coverage of genomic surveillance sequencing.

c. Hospitalization, case and infection fatality rates relative to ONS prevalence. Dots denote mean estimates and error bars 95% CIs.



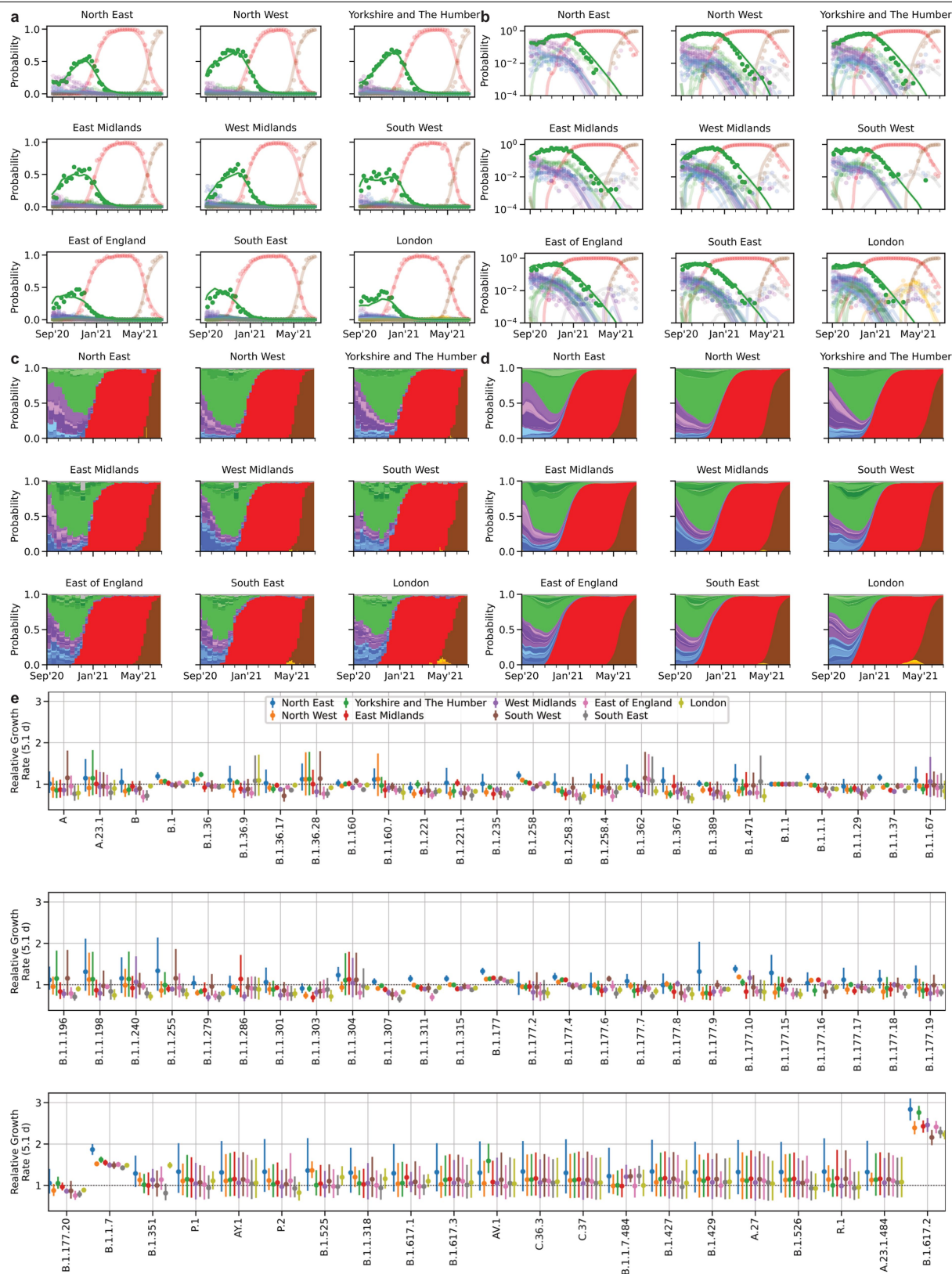
Extended Data Fig. 2 | Genomic surveillance model of total incidence and lineage-specific frequencies. a. Cubic basis splines (top row) are convolved with the infection to test distribution (row 2 and 3) and used to fit the log incidence in a LTLA and its corresponding derivatives (growth rates; bottom row). **b.** Example incidence (top row), logarithmic incidence with individual convolved basis functions (dashed lines, row 2), and resulting (case) reproduction

numbers (growth rate per 5.1d) from our approach (GenomicSurveillance) and estimates by EpiEstim⁴⁸, 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.



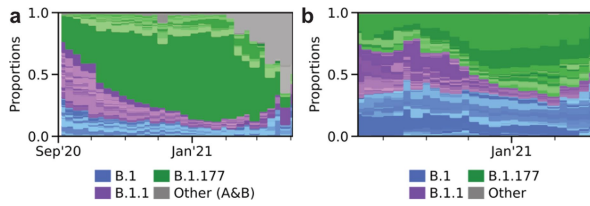
Extended Data Fig. 3 | Genomic surveillance model selection. a. Model loss in terms of the ELBO objective function and the model hyperparameters α_0 and α_1 (see Methods). **b.** Model deviance (calculated as $-2 \times \log$ pointwise predictive density) with respect to the model hyperparameters

α_0 and α_1 (see Methods). **c.** Mean squared error (MSE) of modelled weekly proportions of highly prevalent lineages with respect to the model parameters α_0 and α_1 (see Methods). **d.** Same as in **c**, but for lineages exhibiting low frequencies (VOCs).

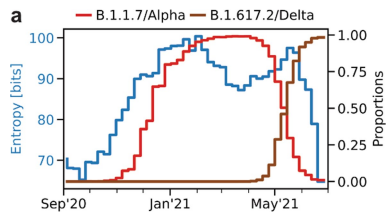


Extended Data Fig. 4 | Spatiotemporal model of 71 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and June 2021. a. Regional lineage specific relative frequency of lineages contributing more than 50 genomes during the time period shown. Dots denote observed data, lines the fits aggregated to each region. **b.** Same as **a**, but on a log scale. **c.** Same data as

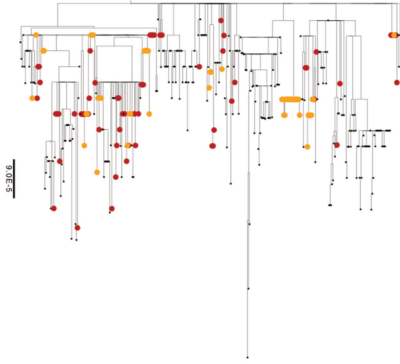
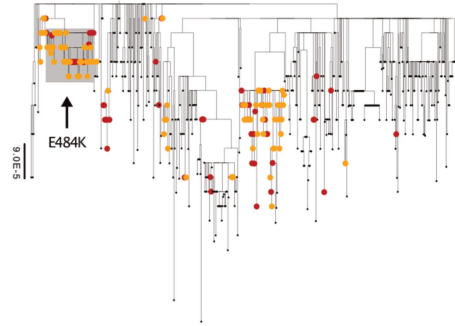
in **a**, shown as stacked bar charts. Colours resemble major lineages as indicated and shadings thereof indicate sublineages. **d.** Same fits as in **a**, shown as stacked segments. **e.** Average growth rates for 71 SARS-CoV-2 lineages estimated in different regions in England. Dots denote median estimates and error bars 95% CIs.



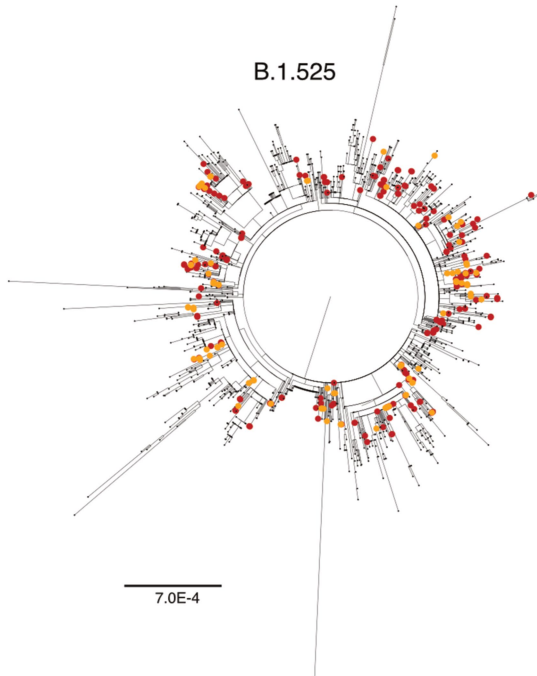
Extended Data Fig. 5 | Relative growth of B.1.177. a. Lineage-specific relative frequency data in England, excluding B.1.1.7 and other VOCs/VUIs (Category Other includes: A, A.18, A.20, A.23, A.25, A.27, A.28, B, B.29, B.40, None). Colours resemble major lineages as indicated and shadings thereof indicate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



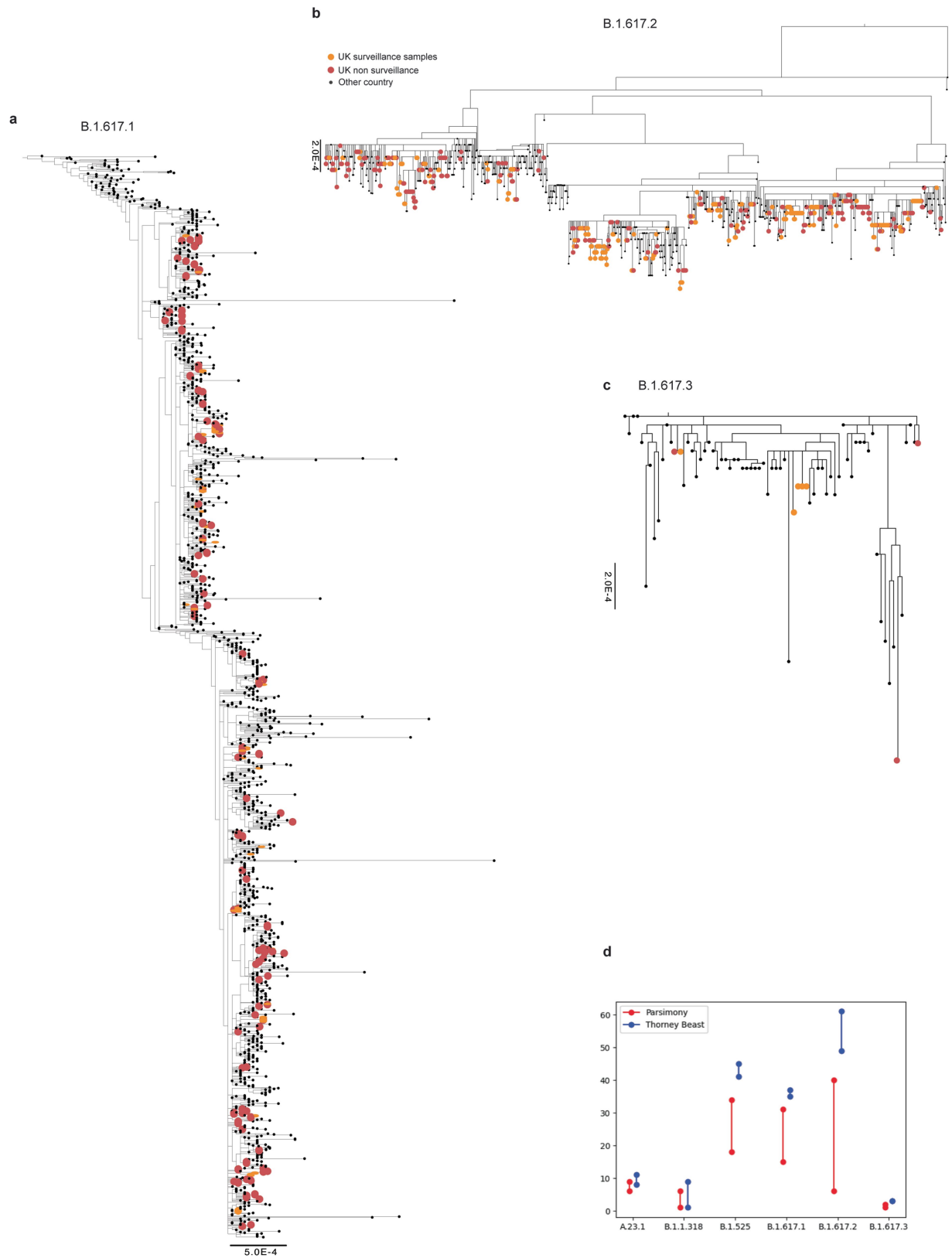
Extended Data Fig. 6 | Genomic diversity of the SARS-CoV-2 epidemic. Shown is the entropy (blue), total number of observed Pango lineages (grey, divided by 4), as well as the proportion of B.1.1.7 (orange, right axis). The sweep of B.1.1.7 causes an intermittent decline of genomic diversity as measured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

a**B.1.1.318****b****A.23.1**

- UK surveillance
- UK non surveillance
- other country

c**B.1.525****d****B.1.351**

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



Extended Data Fig. 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.

Reporting Summary

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

Statistics

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

n/a Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Consensus Fasta sequences were created using the ARTIC nextflow processing pipeline and SARS-CoV-2 lineage assignments using the Pangolin software (01-04-2021 and 23-04-2021 version) and FastTree2 (2.1.11).
Data analysis	Code for spatio-temporal modeling of viral lineages is available at https://github.com/gerstung-lab/genomicsurveillance and as a PyPI package (genomicsurveillance). Analyses were performed in Python 3.7.1 (Packages: matplotlib (3.4.1), numpy (1.20.2), pandas (1.2.3), scikit-learn (0.19.1), scipy (1.6.2), seaborn (0.11.1), jax (0.2.8), genomicsurveillance (0.4.0), numpyro (0.4.0)). The phylogeographic analyses were performed using Thorney Beast (0.1.1) and https://github.com/NicolaDM/phylogeographySARS-CoV-2 . Code for ONS infection survey analysis is available at https://github.com/jhellewell14/ons_severity_estimates .

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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>.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	The study is based on data from 281,178 viral genomes and 3,894,234 positive PCR tests collected in England during the time period from September 1, 2020 to June 30, 2021. This is an observational study based on an existing data set compiled by COG-UK, therefore no sample size calculation is applicable.
Data exclusions	No data was excluded in the analysis.
Replication	This is an observational study based on an existing data set compiled by COG-UK, therefore no replication is applicable.
Randomization	This is an observational study based on an existing data set compiled by COG-UK, therefore no randomization is applicable.
Blinding	This is an observational study based on an existing data set compiled by COG-UK, therefore no blinding is applicable.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging