# Genomic reconstruction of the SARS-CoV-2 epidemic across England from September 2020 to May 2021

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

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

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

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

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

## Abstract

Despite regional successes in controlling the SARS-CoV-2 pandemic, global cases have reached an all time high in April 2021 in part due to the evolution of more transmissible variants. Here we use the dense genomic surveillance generated by the COVID-19 Genomics UK Consortium to reconstruct the dynamics of 62 different lineages in each of 315 English local authorities between September 2020 and April 2021. This analysis reveals a series of sub-epidemics that peaked in the early autumn of 2020, followed by a singular jump in transmissibility of the B.1.1.7 lineage. B.1.1.7 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 eventually suppressed B.1.1.7 and eliminated nearly all other lineages in early 2021. However, a series of variants (mostly containing the spike E484K mutation) defied these trends and persisted at moderately increasing proportions. Accounting for sustained introductions, however, indicates that their transmissibility is unlikely to exceed that of B.1.1.7. Finally, B.1.617.2 was repeatedly introduced to England and grew rapidly in April 2021, constituting approximately 40% of sampled COVID-19 genomes on May 15.

## Introduction

The spread and evolution of viruses reflects the continuous process of mutation and selection. For SARS-CoV-2 the rate of evolution is approximately 24 point mutations per year, or 0.3 per viral generation (Hadfield et al., 2018; Nextstrain team, 2020; Rambaut, 2020). As mutations are passed on to successive virus generations, this enables researchers to follow transmission clusters, define distinct viral lineages and model their behaviour.

After the first wave of the SARS-CoV-2 epidemic swept around the world in the spring of 2020 it soon became apparent that the virus would continue its evolutionary adaptation process to its human host. The first sign of this was the emergence and spread of the spike protein variant D614G, which steadily gained ground across different parts of the world in the second quarter of 2020 and led to defining the B.1 lineage of SARS-CoV-2. Detailed epidemiological analyses estimated that the mutation confers a 20% transmissibility advantage over the original A lineage as isolated in Wuhan, China (Volz, Hill, et al., 2021).

A broad range of lineages have been defined since, which allow the characterization of the dynamics of SARS-CoV-2 transmission across the globe. Noteworthy lineages include B.1.177 (EU-1), which emerged in Spain in early summer of 2020 and spread across Europe through travel (Hodcroft et al., 2020). More recently, the B.1.351 lineage was discovered and characterized in South Africa, where it has rapidly spread due to a combination of mutations thought to increase its transmissibility (N501Y) and ability to evade prior immunity (E484K) (Tegally et al., 2020). Further, a new lineage termed B.1.1.7 was first observed in Kent in September 2020 (Rambaut, Loman, et al., 2020), which has since swept through the United Kingdom and large parts of the world due to a transmissibility elevated by approximately 70% (N. G. Davies et al., 2021; O'Toole, Hill, et al., 2021; Volz, Mishra, et al., 2021). Finally, multiple other variants have been identified that gave rise to distinct epidemiological clusters and harboured a range of mutations suspected to change viral biology. These include P.1, thought to have originated in Brazil (Faria, Claro, et al., 2021; Faria, Mellan, et al., 2021),

A.23.1, with clusters in Uganda and Rwanda, which acquired the E484K mutation after introduction in the UK, B.1.525 with many cases found in Nigeria, the UK and North America, and B.1.1.318 which is prevalent in the UK, the US and also Germany (O'Toole, Scher, et al., 2021). Another recent addition to this list is B.1.617, which has been associated with a large surge of COVID-19 in India in April 2021.

In the United Kingdom, by early May 2021 the COVID-19 Genomics UK Consortium (COG-UK) has sequenced more than 450,000 viral samples. These data have enabled the reconstruction of the dynamics of the first 6 months of the epidemic in the UK in great detail (du Plessis et al., 2021). Here, we leverage a subset of those data: genomic surveillance generated by the Wellcome Sanger Institute Covid-19 Surveillance Team as part of COG-UK, to perform a dense reconstruction and characterisation of the spread and diversity of different SARS-CoV-2 lineages in space and time. We analyse the impact of two national lockdowns and tiered restrictions in the winter of 2020-21, track the elimination of previously dominant SARS-CoV-2 lineages in the early spring of 2021, and assess relative growth rates of variants introduced in the first months of 2021.

# Results

## Spatio-temporal genomic epidemiology of SARS-CoV-2 lineages in England

In this study, we focus on England between September 1, 2020 and April 3, 2021 with an outlook covering the period until May 15 2021 at the end of the manuscript. There were two epidemic waves peaking in early November 2020 and early January 2021, each followed by national lockdown (**Figure 1a**). In this time period, we sequenced 186,585 viral genomes, corresponding to an average of 5.3% (186,585/3,486,417) of all positive tests from PCR testing for the wider population outside the National Health Service (Pillar 2), ranging from 5% in the winter of 2020 to 50% in April 2021, and filtered to remove cases associated with international travel (**Methods**; **Supplementary Figure 1a, b**). Overall a total of SARS-CoV-2 323 lineages were identified using the PANGO lineage definition (Rambaut, Holmes, et al., 2020). As some of these lineages were only rarely and intermittently detected, we collapsed these based on the underlying phylogenetic tree into a set of 62 lineages such that each resulting lineage constituted at least 100 genome, unless the lineage has been designated a variant of concern (VOC) or variant under investigation (VUI) by Public Health England (Public Health England, 2020) (**Figure 1b-d, Supplementary Table 1, 2**).

In the first half of the study period, the most prevalent major lineages were B.1, B.1.1, and B.1.177, each consisting of a series of further sub lineages, usually providing between 100-1000 genomes each to the data set. These data reveal a dynamic pattern of SARS-CoV-2 lineage frequencies in this time period with the evident sweep of B.1.1.7 (**Figure 1b, Supplementary Table 2, 3**). A further characteristic is also the geographic distribution of cases and of different lineages, which provide further insight into the dynamics of the epidemic (**Figure 1c**). Here we aggregated weekly lineage counts in each of 315 English Lower Tier Local Authorities (LTLAs), administrative regions with approximately 100,000-200,000 inhabitants.

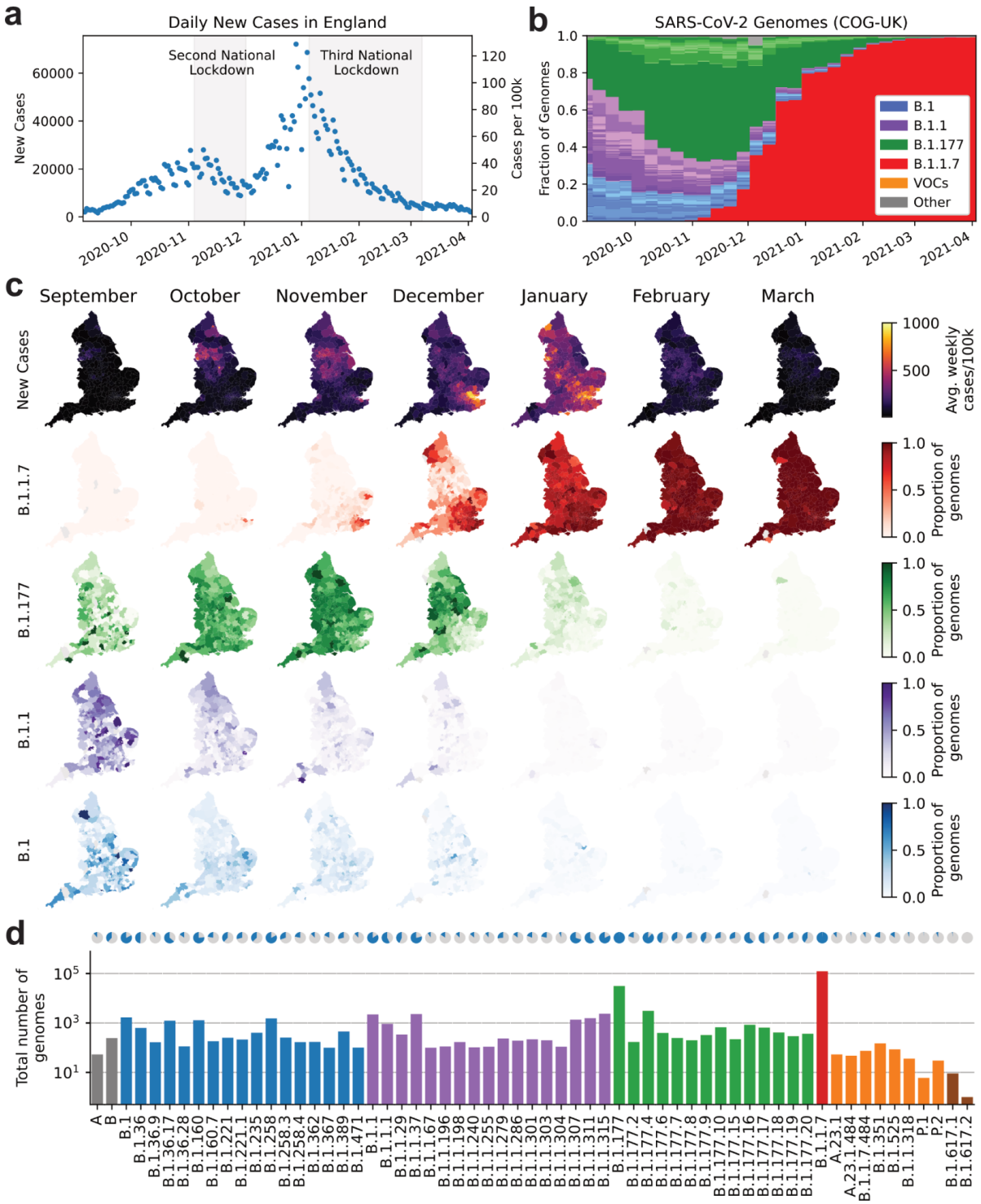



**Figure 1. SARS-CoV-2 surveillance sequencing in England between September 2020 and April 2021. a.** Positive Pillar 2 SARS-CoV-2 tests in England. **b**. Relative frequency of 62 different PANGO lineages, representing approximately 5.3% of tests shown in **a**. **c**. Positive tests (top row) and frequency of 4 major lineages across 315 English lower tier local authorities. **d**. Absolute frequency of sequenced genomes mapped to 62 PANGO lineages. Blue areas in the pie charts are proportional to the fraction of LTLAs where a given lineage was observed.

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

We developed a 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 (**Supplementary Figure 2**; **Methods**). The multivariate logistic regression model is conceptually similar to previous approaches in its estimation of relative growth rates (N. G. Davies et al., 2021; Volz, Mishra, et al., 2021). It accounts for differences in the overall epidemiological dynamics within an LTLA, and allows for the introduction of new lineages (**Figure 2a-c**). Modeling of multiple lineages across space and time can reveal consistent growth patterns. While these are indicative of a transmissibility advantage, we note that these can still be subject to stochastic growth and possible founder effects, and may be influenced by a survivor bias from the designation of PANGO lineages as observed clusters and therefore need to be cautiously interpreted (see limitations at the end of **Methods**).

Despite the sampling noise in a given week, the fitted proportions recapitulate the observed proportions of genomes as revealed by 50 example LTLAs chosen to cover the geography of England (**Figure 2b, c**). The quality of the fit can be better appreciated when the data are aggregated to the level of larger regions, each reflecting areas with approximately 5-8 million inhabitants (**Supplementary Figure 3**). While the relative growth rate of each lineage is not assumed to change, the fitted patterns of viral proportions in each LTLA differ due to the timing and rate of introduction of each lineage, as shown in **Figure 2c**.

We estimate the overall and lineage-specific local incidence by coupling the model to regression of the number of daily positive PCR tests using cubic basis splines, adjusted for testing delays, and an overdispersed negative binomial distribution (**Figure 2d**). These estimates agree with cross-sectional survey data up to a scale factor (Donnarumma, 2021; Pouwels et al., 2021). This method also allows us to calculate lineage-specific growth rates with a temporal accuracy of around 3-4 days; the derived R values, approximated by an effective inter generation time of 5 days, are in good agreement with previous approaches (Cori et al., 2013) (**Supplementary Figure 2c**).

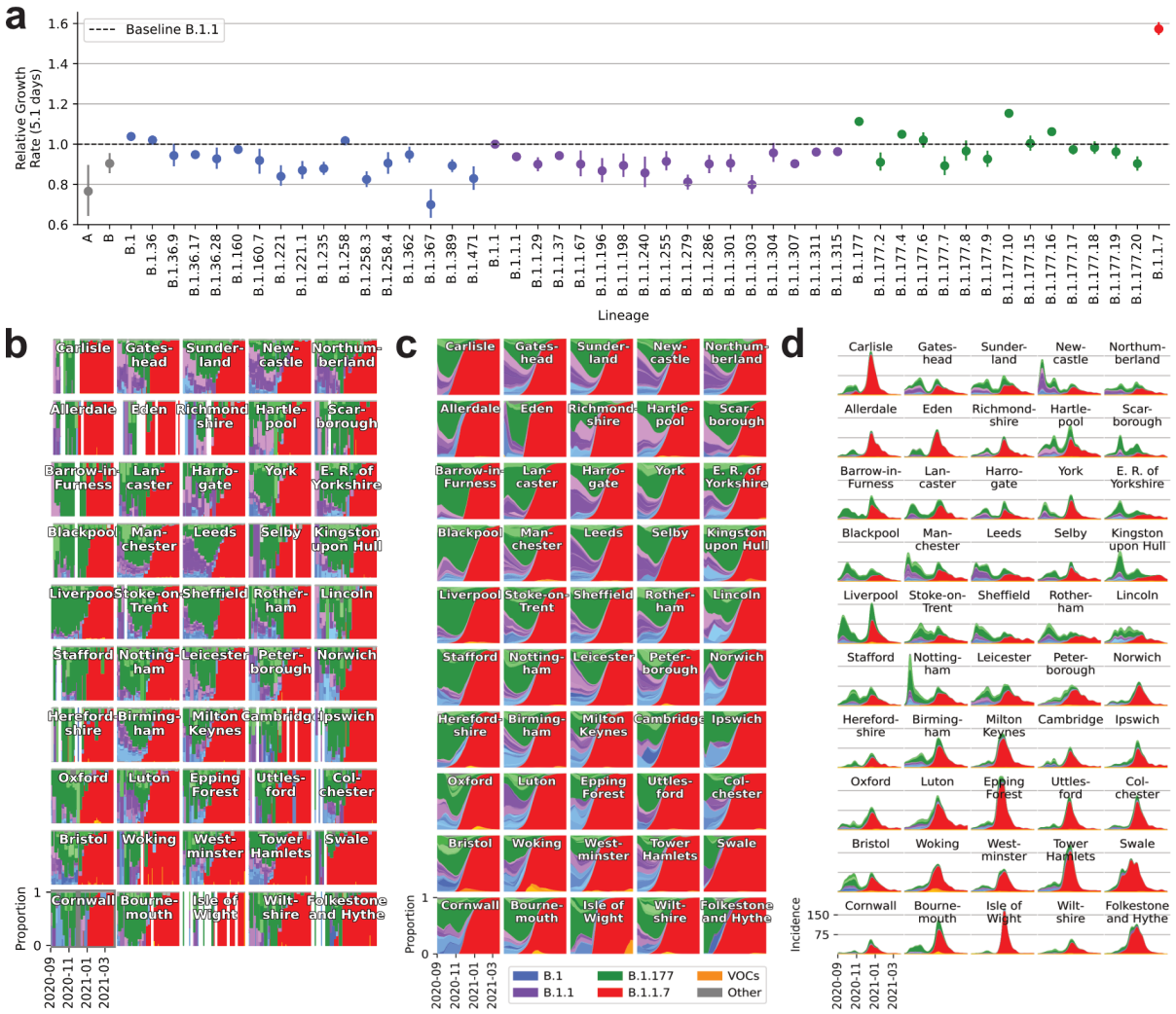



**Figure 2. Spatiotemporal model of 62 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and April 2021. a.** Average growth rates for 50 lineages, excluding VOCs/VUIs other than B.1.1.7 (shown in Figure **5e**). **b.** Lineage specific relative frequency for 50 selected LTLAs, arranged by longitude and latitude. **c**. Fitted lineage-specific relative frequency for the same LTLAs as **b**. **d**. Fitted lineage-specific incidence for the same LTLAs as in **b**.

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

The autumn of 2020 was characterised by a surge of cases, which was concentrated in the north of England and peaked in November 2020, triggering a second national lockdown (**Figure 1a, c**). This wave was jointly caused by multiple lineages with noticeable regional differences in the frequency of different B.1 and B.1.1 sublineages, which were slightly more prevalent in the south and north of England, respectively (**Figure 2b, c**). A similar pattern was observed for sublineages of B.1.177, which were geographically very diverse. Yet a common feature was a steady increase in the proportion of the B.1.177 lineage and its sublineages across LTLAs.

Most sublineages of B.1 and B.1.1 displayed an average growth rate comparable to the reference value of 1 (B.1.1, chosen as a relatively stable lineage circulating during the time period analysed here) with only minor deviations (**Figure 2a**), which can most likely be

attributed to the stochastic growth of these sub lineages. The increasing B.1.177 proportions from around 25% at the beginning of September to 65% at the end of October, corresponds to the growth rate being between 6% (95% CI: 5-7) and 11% (95% CI: 10-12) greater than that of B.1 or B.1.1, depending on the sublineage studied. Of note, the trend of B.1.177 expansion relative to B.1 persisted throughout the end of January (**Supplementary Figure 4a**) and also involved a number of monophyletic sublineages that arose in the UK. The latter cannot easily be explained by international travel, which was the major factor in B.1.177's initial spread throughout Europe in the summer of 2020 (Hodcroft et al., 2020). Distinct B.1.177 lineages also expanded in Denmark in between November 2020 and January 2021 (Danish Covid-19 Genome Consortium, 2021) (**Supplementary Figure 4b**). The biological explanation for this growth advantage is unclear as the characteristic A222V spike variant is not believed to confer a growth advantage (Hodcroft et al., 2020).

## B.1.1.7-specific proliferation during restrictions from November 2020 to March 2021

The pattern of multiple sub-epidemics changed in the subsequent second wave from December 2020 to February 2021 that was almost exclusively driven by B.1.1.7. Its epicenter was in the Southeast of England, but a number of local hotspots in the north such as Carlisle also emerged (**Figure 2d**). This rapid expansion is due to B.1.1.7's transmissibility advantage of 1.62 compared to B.1.1 (95% CI 1.59-1.65; **Figure 2a**), corresponding to an advantage of approximately 1.45 over B.1.177 (95% CI: 1.43-1.48), assuming an unchanged generation interval distribution (Park et al., 2021). This dramatically increased transmissibility, corresponding to an approximately 20x higher monthly increase in cases compared to B.1.1, and 10x greater than B.1.177, led to the rapid dominance of B.1.1.7, a trend which has since been observed in more than 100 other countries (O'Toole, Hill, et al., 2021; O'Toole, Scher, et al., 2021). It is worth noting that the growth of B.1.1.7 in England occurred even though a number of restrictions being in place, which we will discuss in the following.

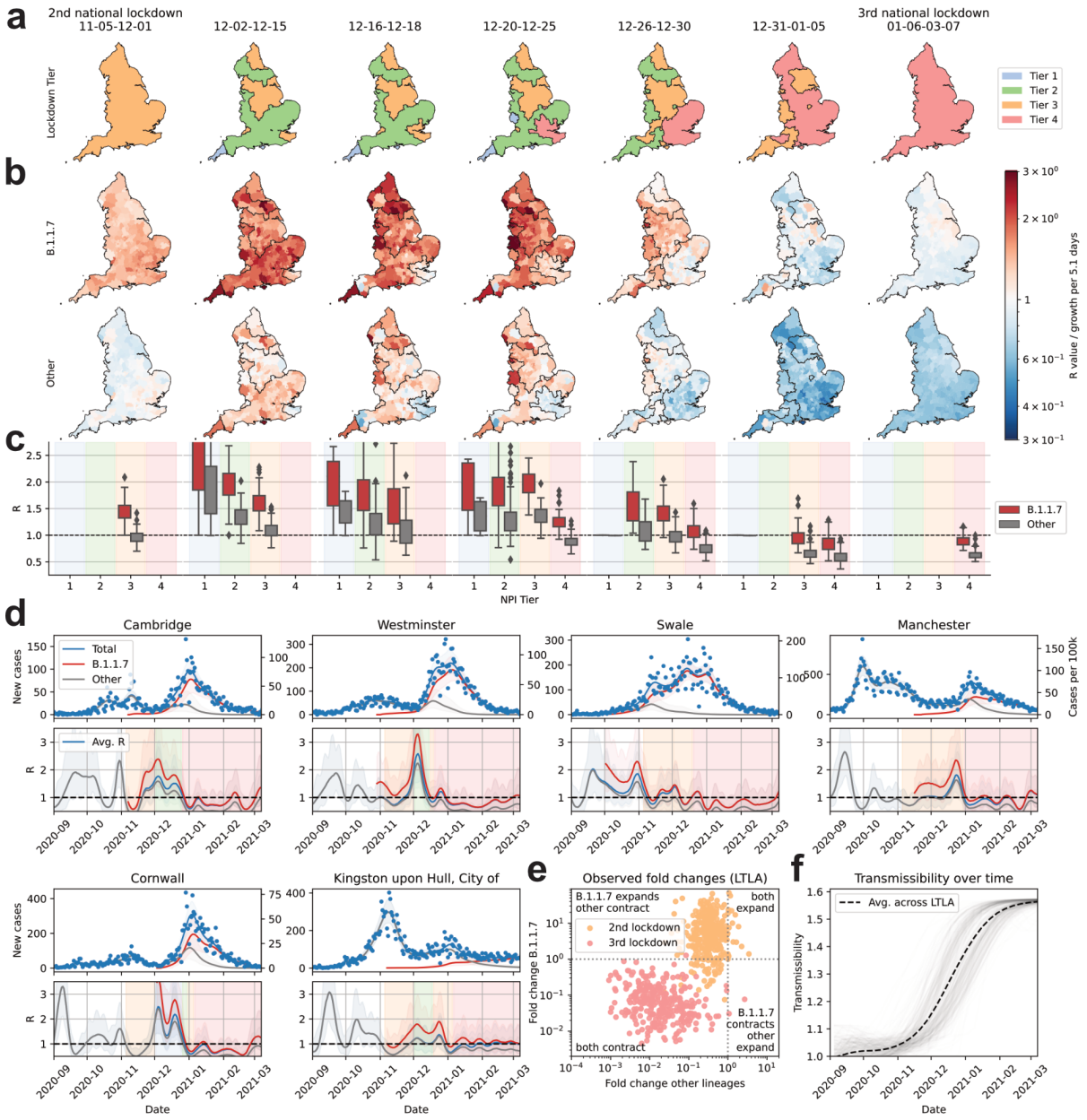



**Figure 3. Growth of B.1.1.7 and other lineages in relation to restrictions. a.** National and regional restrictions in England between November 2020 and March 2021. **b**. Local lineage-specific R values for B.1.1.7 and the average R value (growth per 5.1d) of all other lineages in the same periods. **c**. Boxplots of R values shown in **b**. **d**. Total and lineage-specific incidence (top) and R values (bottom) for 6 selected LTLAs. **e**. Crude lineage-specific fold changes (odds ratios) for B.1.1.7 and other lineages across the second (orange) and third national lockdown (red). **f**. Modelled temporal evolution of the average transmissibility across 315 LTLAs.

As cases rose in October 2020 a second national lockdown was instituted from November 5 to December 1, 2020. It closed hospitality businesses, limited social contacts to no more than 2 outdoors, but kept schools open and allowed leaving one's home with reasonable excuse (Wikipedia contributors, 2021a). This caused a suppression of total cases nationally and a corresponding R value (defined as the growth rate over a period of 5d) of hitherto

dominant lineages below 1 (**Figure 3a-c**). Interestingly, the dense temporal analysis reveals that this pattern of contraction was flanked by an initial peak of proliferation during the week leading to the lockdown, possibly due to the lockdown being announced on October 30 by leaked newspaper reports resulting in behavioural anticipation of consequently banned social activities (**Figure 3d**) (Hunter et al., 2021).

Despite falling total case numbers during the second national lockdown, a rise (R>1) of B.1.1.7 and simultaneous decline of other lineages (R<1) was observed in 68% (214/315) of LTLAs, as described previously (Vöhringer et al., 2020). This demonstrates that failure to contain B.1.1.7 was generally not caused by an overall lack of SARS-CoV-2 control but rather by the higher transmissibility of the B.1.1.7 lineage (**Figure 3d**). This pattern of B.1.1.7-specific growth during lockdown is supported by a model-agnostic analysis of the change in raw case numbers, calculated simply by multiplying the number of positive tests in a given LTLA by the proportion of B.1.1.7 genomes in the weeks prior to and after the lockdown (**Figure 3e**).

The end of the second lockdown was followed by regionally-tiered restrictions, which were defined according to the local incidence, as summarised in (Wikipedia contributors, 2021b). Tier 3 closed most hospitality businesses and allowed gatherings of up to 6 people in public outdoor places, while in tier 2 hospitality venues were allowed to open with restricted service. Tier 1 was the lowest level of restrictions and also allowed private indoor gathering of up to 6 persons. Initially most of England was put under tier 2 restrictions, with the exception of the region around Kent and in the north of England due to refractory case levels. The areas under different tiers of restrictions visibly coincide with the resulting local R values (**Figure 3a-c**). The reopening led to a surge of cases across all tiers with R values above 1. The highest values of R ~1.8 were recorded in tier 1 areas and lowest values in tier 3, which is also evident in the time series shown in **Figure 3d** (i.e. Cornwall and Swale).

As the scale of resurgence due to the nature of B.1.1.7 became apparent, the regional measures were successively increased, with more areas placed under tier 3. This was again reflected in the local R values, which nevertheless were mostly insufficient to contain B.1.1.7 (**Figure 3c**). Consequently, a higher tier 4 was introduced on December 20, which limited interactions to a single outdoor contact. There was generally 20-50% lower proliferation in tier 4 areas compared to tier 3, depending on the time period and LTLA, consistent with previous estimates (Abbott & Funk, 2021), however estimation of R values from case data during this period may be complicated as reflected by a temporal divergence from cross-sectional surveys (**Supplementary Figure 1c, d**). Nevertheless B.1.1.7 nevertheless grew (R>1) in most areas, presumably also driven by increased social interaction over Christmas (**Figure 3c**).

Following the peak of 72,090 daily cases on December 29 (**Figure 1a**), a third national lockdown was announced on January 4. This lockdown extended the restrictions of tier 4 to the entire country and closed educational settings with the exemption of children of key workers. This lockdown, with a high level of compliance – and also increasing immunity, which rose from an estimated 12.8% of adults at the beginning of December 2020 to 28.3% by the end of January 2021 and finally to 47.6% in the first week of March 2021 due to a combination of natural immunity from the large December wave and increasing vaccination (K. S. A. Davies, 2021) – led to a sustained contraction of the epidemic to approximately

5500 daily cases by March 8. The pattern of lineage-specific decline during the third national lockdown was starkly different to the second national lockdown in that approximately 87% (277/315) of LTLAs exhibited a contraction of all lineages at an average R value of 0.9 for B.1.1.7 and 0.6 for other lineages (**Figure 3e)**.

Overall, we estimate that the average transmissibility (R value) of SARS-CoV-2 in England rose from a reference value of 1 in August 2020 to approximately 1.6 in the spring of 2021, as more transmissible variants took over (**Figure 3f**). This highlights not only the general issue of designing appropriate interventions relative to a dynamic epidemic in which data about the change of infection rates lags by around 10 days, but also the problem that previously effective interventions may prove insufficient to contain newly emerging and more transmissible variants.

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

The lineage-specific effects of the third national lockdown had dramatic consequences on the genomic diversity of the epidemic. In the week to April 4, 2021, 99.1% (8,789/8,872) of genomes sampled were B.1.1.7, leaving only 0.9% (83/8,872) non B.1.1.7 genomes. This change in proportion coincided with a dramatic decline in cases from nearly 60,000 in late December 2020 to around 2,500 at the beginning of April 2021.

These changes brought about large differences in lineage-specific incidence (**Figure 4a**). Cases of B.1.1.7 contracted nationally from a peak of around 50,000 cases to around 2,500. At the same time B.1.177, the most prevalent lineage in November 2020, giving rise to approximately 10,000 cases per day in December 2020 is estimated to have fallen to only about 5 detected cases per day. Moreover, the incidence of most other lineages present in the autumn of 2020 was well below 1 in April 2021, implying that the majority of them have been eliminated between January and April 2021. In aggregate, there were an estimated 15 non-B.1.1.7 cases per day at the beginning of April 2021, which constituted the lowest rate since the start of the epidemic in England.

The number of observed PANGO lineages declined steadily from a peak of 137 to only 22 different lineages in the first week of April 2021 (**Figure 4b**). While this figure may in part be attributed to the definition of lineages as emerging clusters – which has a tendency to trail behind the evolving epidemic – it's worth noting that the period of contraction also did not replenish the genetic diversity lost due to the selective sweep by B.1.1.7 (**Supplementary Figure 5**). The conclusion that the majority of lineages present in the autumn of 2020 has been eliminated in the spring of 2021, and are not simply unobserved, is further supported by the fact that the fraction of positive cases that were sequenced has increased from around 5% in the fall of 2020 to 50% in the week ending April 3, 2021 (**Supplementary Figure 1a, b**). Though only around 50% of new infections in the community are ascertained by Pillar 2 testing (**Supplementary Figure 1d**; (Colman et al., 2021)), consistent data from consecutive weeks leaves relatively little room for many lineages to go undetected over sustained periods.

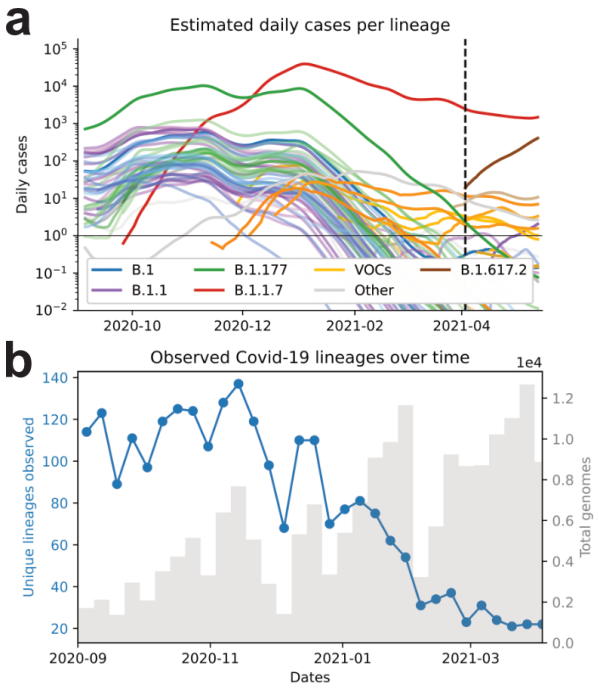



**Figure 4. Elimination of SARS-CoV-2 lineages during spring 2021. a** modelled lineage-specific incidence in England. Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **b**. Observed number of PANGO lineages per week.

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

Parallel to the elimination of SARS-CoV-2 lineages that arose in 2020, a number of variants with potentially higher growth rates surfaced (**Figure 4a**). These include the variants of concern B.1.351, P.1, and also the variants under investigation B.1.525, B.1.1.318 and P.2, which all harbour E484K spike mutations as per their lineage definition, as well as B.1.1.7 and A.23.1 with acquired E484K (**Figure 5a**). The E484K mutation has been found to reduce the binding affinity of antibodies (Greaney, Loes, et al., 2021; Greaney, Starr, et al., 2021; Planas et al., 2021; Starr et al., 2020; Zhou et al., 2021) and is therefore believed to be better at infecting individuals with immunity from previous infection or vaccination. This is consistent with the spread of B.1.351 in South Africa (Tegally et al., 2020) and especially the surge of P.1 in Manaus (Faria, Mellan, et al., 2021), a metropolitan area in Brazil where two thirds of the population were estimated to have antibodies to SARS-CoV-2 after a large epidemic in May 2020 (Buss et al., 2021).

To gain further insights into the most recent VOC/VUI dynamics, we included a provisional data set covering an extra 6 weeks until May 15, 2021, consisting of a further 23,897 viral genomic sequences. We note that local coverage was variable, especially in the last four weeks with some areas in the South and North East underrepresented. A cutoff date of May 8 was used for modeling results presented here (**Supplementary Figure 9**).

In England, the proportion of VOCs and VUIs as a whole was relatively constant from January to early April 2021, comprising approximately 0.3-0.4% of genomes from surveillance samples (**Figure 5b**). However this situation changed in April 2021, when the frequency of B.1.617.2 started to rise rapidly as revealed by provisional data extending to

May 15 (see next section for detailed analysis of this variant). A comparably moderate increase was also seen for B.1.351 and P.1 in this period.

The relatively low numbers of each of these variants in the early spring implies that their dynamics were largely stochastic and characterised by a series of individual and localised outbreaks (**Figure 5c**). The A.23.1 lineage with E484K mutation peaked in January 2021. Similarly, one out of multiple clusters of B.1.1.7 with acquired E484K emerged in the Bristol area in January, but has subsequently subsided. Cases of B.1.525, B.1.351 and B.1.1.318 have been observed in multiple LTLAs, where they usually transiently peaked (**Figure 5c**).

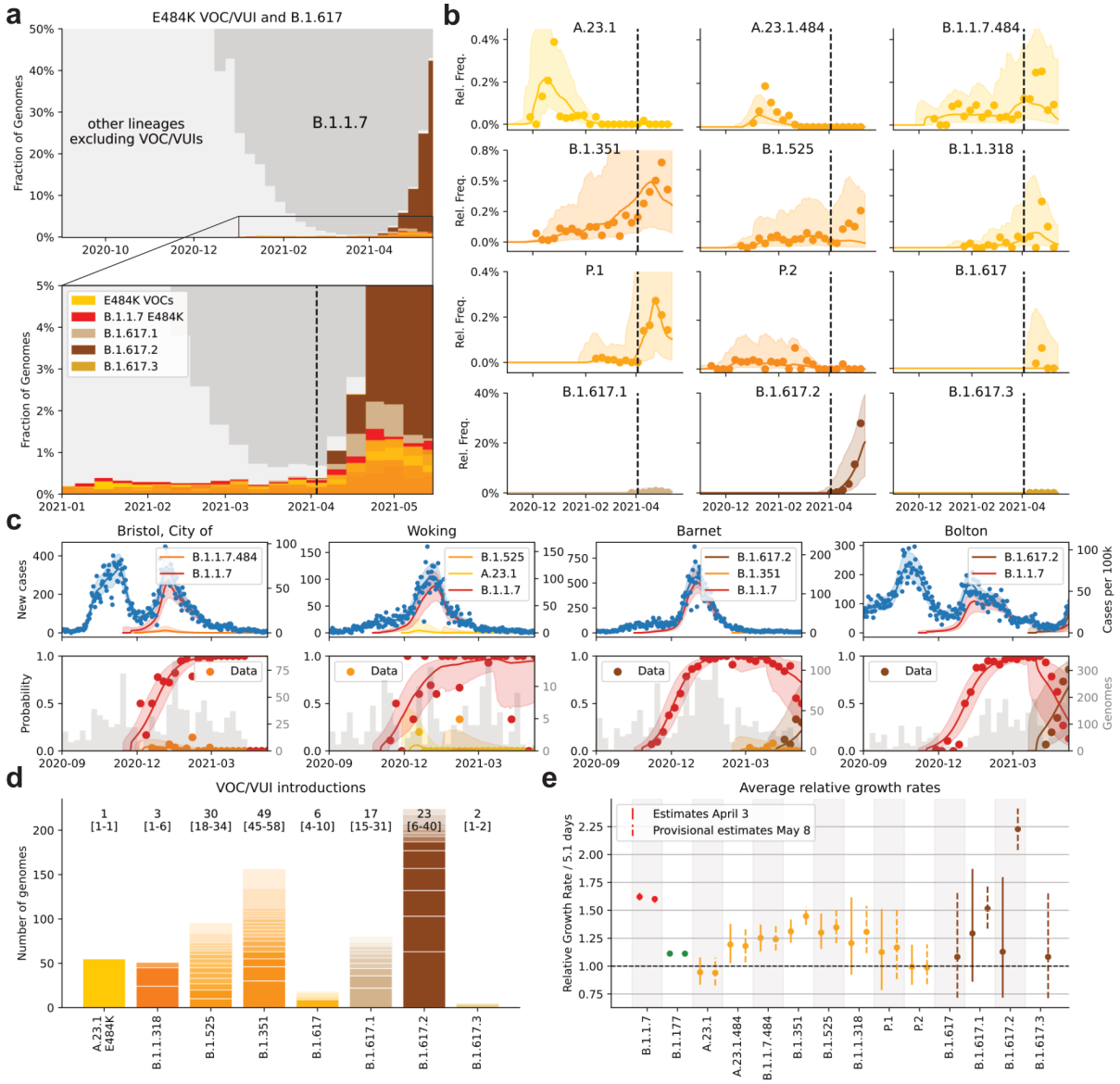



**Figure 5. Dynamics of VOC and VUIs between January and April 2021. a**. Observed relative frequency of B.1.1.7 (dark grey), VOC/VUIs (orange and brown), and other lineages (light grey). **b**. Observed and modelled relative frequency of VOC/VUIs in England. **c**. Total and relative lineage-specific incidence in two selected LTLAs. **d**. Estimated UK VOC/VUI clade numbers (numbers in square parentheses represent minimum and maximum numbers) and sizes. **e**. Estimated average logistic growth rates based on data until 3.4 and 1.5.2021.

Sustained imports from international travel are a critical driving mechanism behind the observed number of non-B.1.1.7 cases. A detailed phylogeographic analysis establishing the most parsimonious sets of monophyletic and exclusively domestic clades, which can be interpreted as individual introductions, confirms that A.23.1 with E484K (1 clade) is likely to have been of domestic origin as no genomes of the same clade were observed internationally (**Figure 5d**; **Supplementary Figure 6; Methods**). The estimated number of introductions was lowest for B.1.1.318 (3 introductions, range 1-6), and highest for B.1.351 (49; range 45-58) and B.1.525 (30; range 18-34). While our data explicitly exclude genomes sampled directly from travellers, these repeated introductions make it clear that the true growth rate due to transmission is lower than the observed increase even in the number of surveillance genomes.

Calculating the growth rates of novel variants is challenging as the stochastic event of introduction needs to be accounted for as an explicit jump in the logistic growth model. Furthermore, at low frequency growth can be stochastic and inflated by repeated introductions, which is partly addressed by overdispersion and by studying the lineage dynamics in small regions, making multiple concomitant introductions unlikely. Still, the estimated growth rates of newly emerging variants need to be interpreted with caution. The model generally infers transmissibilities between those of B.1.177 and B.1.1.7 (**Figure 5e**). This is plausible for the following two reasons. First, it is worth noting that the initially relatively constant proportion displayed in **Figure 5a** corresponds to a reduction of absolute cases from an estimated 240 daily cases in early January to approximately 10 by the end of March, which is a reduction similar to the fold changes seen for B.1.1.7 (**Figure 4a**). As these variants have so far been successfully controlled, their transmissibility is unlikely to exceed that of B.1.1.7. Furthermore, variants emerged and largely persisted while B.1.177 and other lineages were largely eliminated suggesting that their transmissibility is elevated over B.1 or B.1.177. This would be consistent also with independent international observations (Faria, Mellan, et al., 2021; Tegally et al., 2020).

## Rapid rise of B.1.617.2 in April and early May 2021

The B.1.617 lineages, first detected in India in 2020, began to appear in English surveillance samples in March 2021. The frequency of the sublineage B.1.617.2 has rapidly increased since, reaching levels of 41% (761/1851) of surveillance sequences on May 15, 2021 (**Figure 5a, b**). This growth derived from a large influx of cases due to travel from India to England and onward transmission from those cases despite mandatory quarantine (at individuals' homes prior to April 23, in managed quarantine facilities thereafter) and testing (Public Health England, 2020). B.1.617.2 was subsequently observed in a number of large local clusters, such as in Bolton (**Figure 5b**) and has been detected in 128/264 informative LTLAs across most of England by May 8 (**Supplementary Figure 7a, b**).

The most recent data indicate a relative growth rate of B.1.617.2 around 37% (growth per 5.1d, CI 1.26-1.49) in excess of that of B.1.1.7 and unlike all other VOCs/VUIs (**Figure 5e, Supplementary Figure 7c**). Estimates vary between regions and in hotspots (range 1.07-1.41 for informative regions; **Supplementary Figure 7d**), yet the cause and duration of this increased growth are unclear. It is likely that at least three factors have contributed: intrinsic differences in the biology of this variant (transmissibility and/or immune escape), the very high rate of introductions in a short time, and epidemiological factors specific to the

communities where it spread most quickly. Thus the long-term growth advantage may change as it circulates more broadly in the population.

The rapid rise of B.1.617.2 contrasts to B.1.617.1, which was introduced at a similar time and into a similar demographic background, but which grew more slowly than B.1.1.7 (**Figure 5b, e**). This is also evident in the phylogeographic analysis (based on data as of May 1): The 80 genomes of B.1.617.1 are estimated to derive from 17 introductions (range: 15-31), which is similar to the patterns observed for B.1.525 and B.1.351 with around 3-4 genomes for every introduction (**Figure 5d; Supplementary Figure 8**). In contrast, B.1.617.2's 224 genomes are from fewer, but larger, clades (23 introductions, range 6-40), equating to around 10 genomes for every introduction, indicative of elevated domestic transmission throughout April 2021.

The phylogeny of B.1.617.2, in contrast to other VOCs, consists of several clades, which are likely to have originated in the late summer of 2020 and coexisted since as they are split by long branches (**Supplementary Figure 8)** (Nextstrain team, 2021). This suggests that the growth rate of B.1.617.2 was lower in the past. Even though B.1.617.2 increased to 40% frequency in English surveillance samples throughout April and May 2021, this has been largely offset by the falling number of B.1.1.7 cases. Lastly, it is worth noting that the B.1.617.2 lineage is E484 wildtype, unlike B.1.617.1 and B.1.617.3 which harbour a E484Q mutation which might be functionally similar to E484K. Further surveillance and research is thus needed to fully understand the mechanisms that drive transmission of B.1.617.2.

While the incidence of most VOCs and VUIs in England has been successfully controlled in the early months of 2021 by a combination of national lockdown, genomic surveillance, quarantine of international travelers, vaccination, surge testing in affected areas and preferential test, trace and isolation of cases, the example of B.1.617.2 demonstrates how rapidly the situation can change. Variants currently circulating at low levels may also grow in the future if a propensity to evade prior immunity increases their fitness relative to other variants. Constant genomic surveillance is essential to identify and respond to rapid changes caused by pre-existing and new variants.

## Discussion

The period of the SARS-CoV-2 epidemic in England from September 2020 to April 2021, reconstructed here at unprecedented genomic, spatial and temporal detail, can be described in four chapters. First, the introduction of B.1.177 over the summer months led to a slow rise to dominance owing to a small transmissibility advantage of approximately 10% over hitherto existing A, B.1 and B.1.1 lineages. The peak of the epidemic in October 2020, however, was due to many sub-epidemics of different lineages.

Second, B.1.1.7 rapidly outgrew all extant lineages and, despite a series of restrictions in November and December which had reasonable transmission control against those other lineages, led to a large surge of cases in most parts of England as described elsewhere (N. G. Davies et al., 2021). Third, the wave of COVID-19 caused by the rapid rise of B.1.1.7 forced the UK to adopt progressively stricter restrictions in December 2020 and January 2021 to avoid a catastrophic saturation of the health system. The resulting reduction of

B.1.1.7 was also accompanied by an even faster contraction of other lineages, most of which were eliminated in the spring of 2021.

Fourth, from December 2020 to April 2021 a series of variants with novel biological characteristics were repeatedly introduced in the UK, persisting at low levels and slowly rising in frequency. A common feature of these variants was the E484K spike mutation which reduces the rate of neutralisation by antibodies developed by previous infections. However, it is worth noting that the E484K variant was usually seen in conjunction with other mutations of biological importance and also that B.1.1.7 with acquired E484K did not appear to exhibit a growth advantage. In fact, the apparently slower growth of B.1.1.7 with E484K raises the possibility of a transmissibility disadvantage of the E484K mutation in the background of B.1.1.7. It will be crucial to monitor and characterise these and other biologically similar variants in the future. Lastly, the introduction of B.1.617.2, which does not have E484K, and its rapid growth throughout April and early May 2021 are a stark reminder that viral evolution can take unexpected paths and that rapid genomic surveillance remains as important as ever.

Overall, our analysis exemplifies how a genomically complex epidemic can be reconstructed comprehensively using systematic genomic surveillance and spatio-temporal statistical models. These analyses highlight the interplay between selection of more transmissible variants and human interventions, which determine the spread, scale and course of the epidemic. Over the period of investigation, we estimate that the average transmissibility of SARS-CoV-2 in the UK has increased by around 70% by the end of March 2021. While this increase can be compensated by stronger social distancing, more effective test and trace and vaccination, the evolution of more transmissible variants and potentially immunity-evading variants is likely to remain a challenge in the absence of near-complete suppression.

The global presence of SARS-CoV-2 in human populations, the existence of human-to-animal and animal-to-animal transmission means it is unlikely that global eradication is possible at least in the short to mid term. Therefore a global perspective on controlling and surveilling the virus is essential.

# Methods

## Pillar 2 SARS-CoV-2 testing data

Publicly available daily SARS-CoV-2 test result data from testing for the wider population outside the National Health Service (Pillar 2 newCasesBySpecimenDate) was downloaded from https://coronavirus.data.gov.uk/ spanning the date range from 2020-09-01 to 2020-04-03 for 315 English lower tier local authorities (downloaded on 2021-05-19). These data are mostly positive PCR tests, with about 4% of results from lateral flow tests without PCR confirmation. In this dataset, the City of London is merged with Hackney, and Isles of Scilly are merged with Cornwall due to their small number of inhabitants, thereby reducing the number of English LTLAs from 317 to 315. Population data for each LTLA was downloaded from the Office of National Statistics,

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

## SARS-CoV-2 surveillance sequencing

185,847 (Sept-Apr) and 18,650 (Apr-May) samples were collected as part of random surveillance of positive tests of residents of England from four Pillar 2 Lighthouse Labs. The samples were collected between 2020-09-01 and 2021-04-03 and later extended to 2021-05-01. A random selection of samples was taken, after excluding those known to be taken during quarantine of recent travellers, and samples from targeted and local surge testing efforts. The available metadata made this selection imperfect, but these samples should be an approximately random selection of infections in England during this time period, and the large sample size makes our subsequent inferences robust.

We amplified RNA extracts from these tests with Ct < 30 using the ARTIC amplicon protocol, https://www.protocols.io/workspaces/coguk/publications. We sequenced 384-sample pools on Illumina NovaSeq, and produced consensus fasta sequences according to the ARTIC nextflow processing pipeline https://github.com/connor-lab/ncov2019-artic-nf. Lineage assignments were made using Pangolin (Rambaut, Holmes, et al., 2020), according to the latest lineage definitions at the time, except for B.1.617, which we re-analysed after the designation of sub-lineages B.1.617.1, .2 and .3. Lineage prevalence was computed from 186,585 genome sequences. The genomes were mapped to the same 315 English LTLAs for testing data described above. Mapping was performed from outer postcodes to LTLA, which can introduce some misassignment to neighbouring LTLAs. Furthermore, lineages in each LTLA were aggregated to counts per week for a total of 35 weeks, defined beginning on Sunday and ending on Saturday.

Lastly, the complete set of 323 SARS-CoV-2 PANGO lineages was collapsed into $l = 62$ lineages using the underlying phylogenetic tree, such that each resulting lineage constituted at least 100 genomes each during the study period with the exception of VOCs or selected VUIs, which were included regardless.

## Spatio-temporal genomic surveillance model

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

### Motivation

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

each lineage and therefore drops from the relative proportions of lineages $\boldsymbol{p}(t) = \boldsymbol{x}(t)/\sum \boldsymbol{x}(t) \propto e^{\boldsymbol{c}+\boldsymbol{b}t}$.

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

## Incidence

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

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

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

The time from infection to test is given by the incubation time plus the largely unknown distribution of the time from symptoms to test, which in England was required to take place within 5d of symptom onset. To account for these factors the log normal incubation time distribution from (Bi et al., 2020) is scaled by the equivalent of changing the mean by 2d. The convolution shifts cases approximately 7d into the future and also spreads them out according to the width of $g$ (**Supplementary Figure 2a**).

In order to parametrise the logarithmic incidence $\boldsymbol{\lambda}(t)$ we use a vector of $k = 38$ ($k = 43$ for the analysis of the provisional dataset until May 8) cubic basis functions $\boldsymbol{f}(t)$ equidistantly tiled across the study interval every 7 days, that is $\boldsymbol{f}_{i+1}(t) = \boldsymbol{f}_i(t-7) \quad \forall i$. Each spline basis function is convolved with the time to test distribution $g$, $\boldsymbol{f}^*(t) = (\boldsymbol{f}_1^*(t), \ldots, \boldsymbol{f}_k^*(t))$ as outlined above and used to fit the logarithmic incidence. The derivatives of the original basis $\boldsymbol{f}'(t)$ are used to calculate the underlying growth rates and R values, as shown further below. The convolved spline basis $\boldsymbol{f}^*(t)$ is used to fit the per capita incidence in each LTLA as (**Supplementary Figure 2b**):

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

This implies that fitting the incidence function for each of the $m$ local authorities is achieved by a suitable choice of coefficients $\boldsymbol{B} \in \mathbb{R}^{m\times k}$, that is one coefficient for each spline function

for each of the LTLAs. The parameter $\boldsymbol{B}$ is modelled in each LTLAs by a multivariate Normal prior distribution, which read for LTLA $i$:

$$\boldsymbol{B}_{i\cdot} \sim \mathrm{MVN}(-10, \boldsymbol{L}\boldsymbol{L}^T).$$

The prior expectation of -10 was chosen as it approximately reflects the logarithmic daily incidence of 10/100,000 inhabitants observed during the study period. No prior correlation is modelled across the dimension of LTLAs, but the temporal sequence of $k$ splines are assumed to be correlated, which can also be used to control the variance of the estimated time series. This is achieved by an appropriate choice of the Cholesky factor of the covariance $\boldsymbol{L}$, which has the following ARMA (auto regressive moving average) structure

$$L = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ ((-1)\rho)^1 & \ddots & 0 & & & \\ ((-1)\rho)^2 & \ddots & 1 & 0 & & \vdots \\ ((-1)\rho)^3 & ((-1)\rho)^2 & ((-1)\rho)^1 & 1 & 0 & \\ \vdots & ((-1)\rho)^3 & 1 & -2 & 1 & 0 \\ & & ((-1)\rho)^3 & 1 & -2 & 1 & 0 \\ ((-1)\rho)^{k-1} & & \dots & ((-1)\rho)^3 & 1 & -2 & 1 \end{pmatrix}^{-1}$$

First, there is a weak prior correlation (MA) of $\rho = 0.5$ between consecutive spline functions, which makes the spline somewhat stiffer. Second, there is a linear autoregressive (AR) dependency of the last three spline functions, which are only partially observed owing to the shifted convolved basis $\boldsymbol{f}^*$and otherwise have a tendency to diverge.

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

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

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

## Growth rates and R values

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

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

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

In order to express the daily growth rate as an approximate reproductive number R, one needs to consider the distribution of the inter generation time, which is assumed to be Gamma distributed with mean 6.3 days ($\alpha$=2.29, $\beta$=0.36) (Bi et al., 2020). The R value can be expressed as a Laplace transform of the inter generation time distribution (Wallinga & Lipsitch, 2007). Effectively, this shortens the relative time period because the exponential

dynamics put disproportionally more weight on stochastically early transmissions over late ones. For reasons of simplicity and being mindful also of the uncertainties of the intergeneration time distribution, we approximate R values by multiplying the logarithmic growth rates with a value of $\bar{\tau}_e$=5.1d, which was found to be a reasonable approximation to the convolution required to calculate R values,

$$\log \boldsymbol{r}_0(t) \approx \frac{d \log \boldsymbol{\mu}(t)}{dt} \bar{\tau}_e = \boldsymbol{\lambda}'(t) \bar{\tau}_e$$

Hence the overall growth rate scaled to an effective inter generation time of 5.1d can be readily derived from the derivatives of the spline basis and the corresponding coefficients. The values derived from the approach are in very close agreement with those of the method of (Cori et al., 2013), but shifted according to the typical delay from infection to test (**Supplementary Figure 2b**).

### Genomic prevalence

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

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

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

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

and accounts for the fact that lineages can be entirely absent prior to a stochastically distributed time period preceding their first observation. This is because in the absence of such a term the absence of a lineage prior to the point of observation can only be explained by higher growth rate compared to the preceding lineages, which may not necessarily be the case. As the exact time of introduction is generally unknown a stochastic three week period of $\boldsymbol{t}_0 \sim \operatorname{Unif}(-21, 0) + \boldsymbol{t}_0^{\text{obs}}$ prior to the first observation $\boldsymbol{t}_0^{\text{obs}}$ was chosen.

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

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

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

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

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

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

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

### Lineage-specific incidence and growth rates

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

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

Further corresponding lineage-specific R value can be calculated as

$$\log \boldsymbol{R}(t) = \log \boldsymbol{r}_0(t) - \bar{\tau}_e(\boldsymbol{P}(t) \times \boldsymbol{b} + \boldsymbol{b})$$

where the term $\boldsymbol{P}(t) \times \boldsymbol{b}$ is the average log growth rate fold change derived from the varying genomic composition, which is subtracted from the observed R value and then the log transmissibility fold changes $\boldsymbol{b}$ are added for each lineage. This implies that $\boldsymbol{R}_i(t) = \boldsymbol{r}_0(t) \frac{e^{\boldsymbol{b}_i}}{e^{\boldsymbol{P}_i(t) \times \boldsymbol{b}_i}}$, ie the R values of lineages are proportional to one another at any given point in time with factor $e^b$ .

### Inference

The model was implemented in numpyro (Bingham et al., 2018; Phan et al., 2019) and fitted using stochastic variational inference (Hoffman et al., 2013). Guide functions were multivariate normal distributions for each row (corresponding to an LTLA) of $\boldsymbol{B}$, $\boldsymbol{C}$ to preserve the correlations across lineages and time as well as for $(\boldsymbol{b}, \boldsymbol{c})$ to also model correlations between growth rates and typical introduction.

## Phylogeographic analyses

To infer VOC introduction events into the UK and corresponding clade sizes, we investigated VOC genome sequences from multiple countries as available from GISAID https://www.gisaid.org/ . We downloaded multiple sequence alignments of genome sequences with release dates 17-04-2021 (for the analysis of lineages A.23.1, B.1.1.318, B.1.351, B.1.525) and 05-05-2021 (for the analysis of B.1.617 sublineages). We then extracted a sub-alignment from each lineage (following the 01-04-2021 version of PANGOlin for the 17-04-2021 alignment and the 23-04-2021 version of PANGOlin for the 05-05-2021 alignment), and, for each sub-alignment, we inferred a phylogeny via maximum likelihood using FastTree2 version 2.1.11 (Price et al., 2010) with default options and GTR substitution model (Tavaré, 1986).

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

When using parsimony, we only consider migration histories along a phylogenetic tree that are parsimonious in terms of migration events from and to the UK (in practice we collapse all the non-UK locations into a single one). Also, since SARS-CoV-2 phylogenies present substantial numbers of polytomies, that is, phylogenetic nodes where the tree topology cannot be reconstructed due to lack of mutation events on certain branches, we developed a tailored dynamic programming approach to efficiently integrate over all possible splits of polytomies and over all possible parsimonious migration histories. The idea of this method is somewhat similar to typical Bayesian phylogeographic inference (e.g. (Lemey et al., 2009)) in that it allows us to at least in part integrate over phylogenetic uncertainty and uncertainty in migration history; however, it also represents a very simplified version of these analyses, more so than (du Plessis et al., 2021), as it considers most of the phylogenetic tree as fixed, ignores sampling times, and uses parsimony instead of a likelihood-based approach. Other caveats of phylogeographic approaches also apply here, in particular possible biases due to uneven sequencing rates across the world (e.g. (De Maio et al., 2015)). The benefit of our approach is however that it allows us to quickly investigate large phylogenetic trees.

## ONS infection survey analysis

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

Comparison of ONS incidence estimates with hospitalisation, case and death rates was conducted by estimating infection trajectories separately from observed cases, hospitalisations and deaths (Abbott et al., 2020; Sherratt et al., 2020), convolving them with estimated PCR detection curves (Hellewell et al., 2021), and dividing the resulting PCR prevalence estimates by the estimated prevalence from the ONS Community Infection

Survey at the midpoints of the 2-week intervals over which prevalence was reported in the survey.

## Limitations

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

In its current form the model only accounts for a single introduction event per LTLA, which seems to be a reasonable assumption based on the estimated number of introductions of newly emergent lineages by international travel. Nevertheless it is important to investigate whether sustained introductions inflate the observed growth rates. This can be achieved by a more detailed phylogeographic assessment and the assessment of monophyletic sublineages.

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

The inferred growth rates also cannot identify a particular mechanism which may be caused by higher viral load, longer infectivity or greater susceptibility. Lineages could potentially differ by their inter-generation time, which would lead to a non linear scaling. Here we did not find convincing evidence for such effects in contrast to previous reports (Vöhringer et al., 2020). Also lineages may differ in their ability to evade prior immunity, which might provide a differential growth advantage over time; it is therefore advisable to assess whether a growth advantage is constant over periods in which immunity changes considerably.

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

As the total incidence is modelled based on the total number of positive PCR tests it may be influenced by testing capacity with the total number of tests having approximately tripled between September 2020 and March 2021. This can potentially lead to a time trend in recorded cases and thus baseline R values if the access to testing changed, e.g. by too few available tests during high incidence, or changes to the eligibility to test with fewer symptoms intermittently. Generally, the observed incidence was in good agreement with representative cross-sectional estimates from the Office of National Statistics (Donnarumma, 2021; Pouwels et al., 2021), except for a period of peak incidence from late December 2020 to

January 2021 (**Supplementary Figure 1d**). Values after March 8, 2021 need to be interpreted with caution as pillar 2 PCR testing was supplemented by lateral flow devices, which increased the number of daily tests to more than 1.5 million.

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

## Code availability

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

## Data availability

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

## Acknowledgements

COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. We would like to thank our colleagues at EMBL-EBI, the Wellcome Sanger Institute and from COG-UK for stimulating discussions and helpful comments on this manuscript. HSV and MG are supported by a grant from the Department of Health and Social Care. AWJ, EB and MG are beneficiaries from grant NNF17OC0027594 from the Novo Nordisk Foundation. TS is supported by grant 210918/Z/18/Z, and JH and SF by grant 210758/Z/18/Z from the Wellcome Trust. HSV, NDM, AWJ, NG, EB and MG are supported by EMBL. We would like to thank Elias Allara (Cambridge) and Georgia Whitton (Sanger) for providing outer postcode to LTLA mappings. We thank all the contributors who submitted genome sequences to GISAID. Acknowledgement tables for individual sequences are deposited at https://github.com/NicolaDM/phylogeographySARS-CoV-2.

## Conflicts of Interest

None declared.

## Ethical approval

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

## Author contributions

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

## References

Abbott, S., & Funk, S. (2021, January 28). *Local area reproduction numbers and S-gene target failure*. https://github.com/epiforecasts/covid19.sgene.utla.rt

Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., Chan, Y.-W. D., Finger, F., Campbell, P., Endo, A., Pearson, C. A. B., Gimma, A., Russell, T., Flasche, S., Kucharski, A. J., … CMMID COVID modelling group. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, *5*, 112.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1810.09538

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., Gao, W., Cheng, C., Tang, X., Wu, X., Wu, Y., Sun, B., Huang, S., Sun, Y., Zhang, J., … Feng, T. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of

their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases*, *20*(8), 911–919.

Buss, L. F., Prete, C. A., Jr, Abrahim, C. M. M., Mendrone, A., Jr, Salomon, T., de Almeida-Neto, C., França, R. F. O., Belotti, M. C., Carvalho, M. P. S. S., Costa, A. G., Crispim, M. A. E., Ferreira, S. C., Fraiji, N. A., Gurzenda, S., Whittaker, C., Kamaura, L. T., Takecian, P. L., da Silva Peixoto, P., Oikawa, M. K., … Sabino, E. C. (2021). Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science*, *371*(6526), 288–292.

Colman, E., Enright, J., Puspitarani, G. A., & Kao, R. R. (2021). Estimating the proportion of SARS-CoV-2 infections reported through diagnostic testing. *medRxiv*. https://www.medrxiv.org/content/10.1101/2021.02.09.21251411v1.abstract

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, *178*(9), 1505–1512.

Danish Covid-19 Genome Consortium. (2021). *Genomic overview of SARS-CoV-2 in Denmark*. https://www.covid19genomics.dk/statistics

Davies, K. S. A. (2021, April 27). *Coronavirus (COVID-19) Infection Survey, antibody and vaccination data for the UK - Office for National Statistics*. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveyantibodydatafortheuk/28april2021

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., Group1‡, C. C.-19 W., COVID-19 Genomics UK (COG-UK) Consortium‡, Diaz-Ordaz, K., … John Edmunds, W. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, *372*(6538). https://doi.org/10.1126/science.abg3055

De Maio, N., Wu, C.-H., O'Reilly, K. M., & Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*,

*11*(8), e1005421.

Donnarumma, K. S. A. (2021, April 22). *Coronavirus (COVID-19) infection survey, UK - office for national statistics*. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/23april2021

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., Faria, N. R., Jackson, B., Loman, N. J., O'Toole, Á., Nicholls, S. M., Parag, K. V., Scher, E., Vasylyeva, T. I., Volz, E. M., … Pybus, O. G. (2021). Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, *371*(6530), 708–712.

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

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., … Sabino, E. C. (2021). Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. *medRxiv : The Preprint Server for Health Sciences*. https://doi.org/10.1101/2021.02.26.21252554

Greaney, A. J., Loes, A. N., Crawford, K. H. D., Starr, T. N., Malone, K. D., Chu, H. Y., & Bloom, J. D. (2021). Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. In *bioRxiv* (p. 2020.12.31.425021). https://doi.org/10.1101/2020.12.31.425021

Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., Hilton, S. K.,

Huddleston, J., Eguia, R., Crawford, K. H. D., Dingens, A. S., Nargi, R. S., Sutton, R. E., Suryadevara, N., Rothlauf, P. W., Liu, Z., Whelan, S. P. J., Carnahan, R. H., Crowe, J. E., & Bloom, J. D. (2021). Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host & Microbe*, *29*(1), 44–57.e9.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* , *34*(23), 4121–4123.

Hellewell, J., Russell, T. W., SAFER Investigators and Field Study Team, Crick COVID-19 Consortium, CMMID COVID-19 working group, Beale, R., Kelly, G., Houlihan, C., Nastouli, E., & Kucharski, A. J. (2021). Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Medicine*, *19*(1), 106.

Hodcroft, E. B., Zuber, M., Nadeau, S., Crawford, K. H. D., Bloom, J. D., Veesler, D., Vaughan, T. G., Comas, I., Candelas, F. G., Stadler, T., & Neher, R. A. (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv : The Preprint Server for Health Sciences*. https://doi.org/10.1101/2020.10.25.20219063

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research: JMLR*, *14*, 1303–1347.

Hunter, P. R., Brainard, J. S., & Grant, A. R. (2021). The Impact of the November 2020 English National Lockdown on COVID-19 case counts. *medRxiv*. https://www.medrxiv.org/content/10.1101/2021.01.03.21249169v1.abstract

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian Phylogeography Finds Its Roots. *PLoS Computational Biology*, *5*(9), e1000520.

Nextstrain team. (2020). *Genomic epidemiology of novel coronavirus - Global subsampling*. https://nextstrain.org/ncov/global?l=clock

Nextstrain team. (2021). *Genomic epidemiology of novel coronavirus - Asia-focused*

*subsampling*. https://nextstrain.org/ncov/asia

O'Toole, Á., Hill, V., Pybus, O. G., Watts, A., Bogoch, I. I., Khan, K., Messina, J. P., The COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, Tegally, H., Lessells, R. R., Giandhari, J., Pillay, S., Tumedi, K. A., Nyepetsi, G., & Others. (2021, February 4). *Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2*. https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., du Plessis, L., Aanensen, D., Holmes, E., Pybus, O., & Rambaut, A. (2021). *PANGO lineages*. https://cov-lineages.org/pango_lineages.html

Park, S. W., Bolker, B. M., Funk, S., Metcalf, C. J. E., Weitz, J. S., Grenfell, B. T., & Dushoff, J. (2021). Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. *medRxiv*. https://www.medrxiv.org/content/10.1101/2021.05.03.21256545v1.abstract

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1912.11554

Planas, D., Bruel, T., Grzelak, L., Guivel-Benhassine, F., Staropoli, I., Porrot, F., Planchais, C., Buchrieser, J., Rajah, M. M., Bishop, E., & Others. (2021). Sensitivity of infectious SARS-CoV-2 B. 1.1. 7 and B. 1.351 variants to neutralizing antibodies. *Nature Medicine*, 1–8.

Pouwels, K. B., House, T., Pritchard, E., Robotham, J. V., Birrell, P. J., Gelman, A., Vihta, K.-D., Bowers, N., Boreham, I., Thomas, H., Lewis, J., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Benton, P., Walker, A. S., & COVID-19 Infection Survey Team. (2021). Community prevalence of SARS-CoV-2 in England from April to November,

2020: results from the ONS Coronavirus Infection Survey. *The Lancet. Public Health*, *6*(1), e30–e38.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*, *5*(3), e9490.

Public Health England. (2020, December 21). *Investigation of novel SARS-CoV-2 variants of concern. Technical briefing 10*. GOV.UK. https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201

Rambaut, A. (2020, January 29). *Phylogenetic analysis of nCoV-2019 genomes*. https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D. L., Volz, E., & on behalf of COVID-19 Genomics Consortium UK. (2020, December 18). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*. 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

Sherratt, K., Abbott, S., Meakin, S. R., Hellewell, J., Munday, J. D., Bosse, N., Jit, M., Funk, S., & CMMID Covid-19 working group. (2020). Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of Covid-19 in England. In *bioRxiv*. medRxiv. https://doi.org/10.1101/2020.10.18.20214585

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, *182*(5), 1295–1310.e20.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*(2), 57–86.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., & Others. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1.full

Vöhringer, H., Sinnott, M., Amato, R., Martincorena, I., Kwiatkowski, D., Barrett, J. C., Gerstung, M., & on behalf of The COVID-19 Genomics UK (COG-UK) consortium. (2020, December 30). *Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown*. Virological.org. https://virological.org/t/lineage-specific-growth-of-sars-cov-2-b-1-1-7-during-the-english-national-lockdown/575/2

Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., Rey, S. M., Nicholls, S. M., Colquhoun, R. M., da Silva Filipe, A., Shepherd, J., Pascall, D. J., Shah, R., Jesudason, N., Li, K., … Connor, T. R. (2021). Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*, *184*(1), 64–75.e11.

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., … Ferguson, N. M. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, 1–17.

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings. Biological Sciences / The Royal Society*, *274*(1609), 599–604.

Wikipedia contributors. (2021a, March 28). *The Health Protection (Coronavirus, Restrictions) (England) (No. 4) Regulations 2020*. Wikipedia, The Free Encyclopedia.

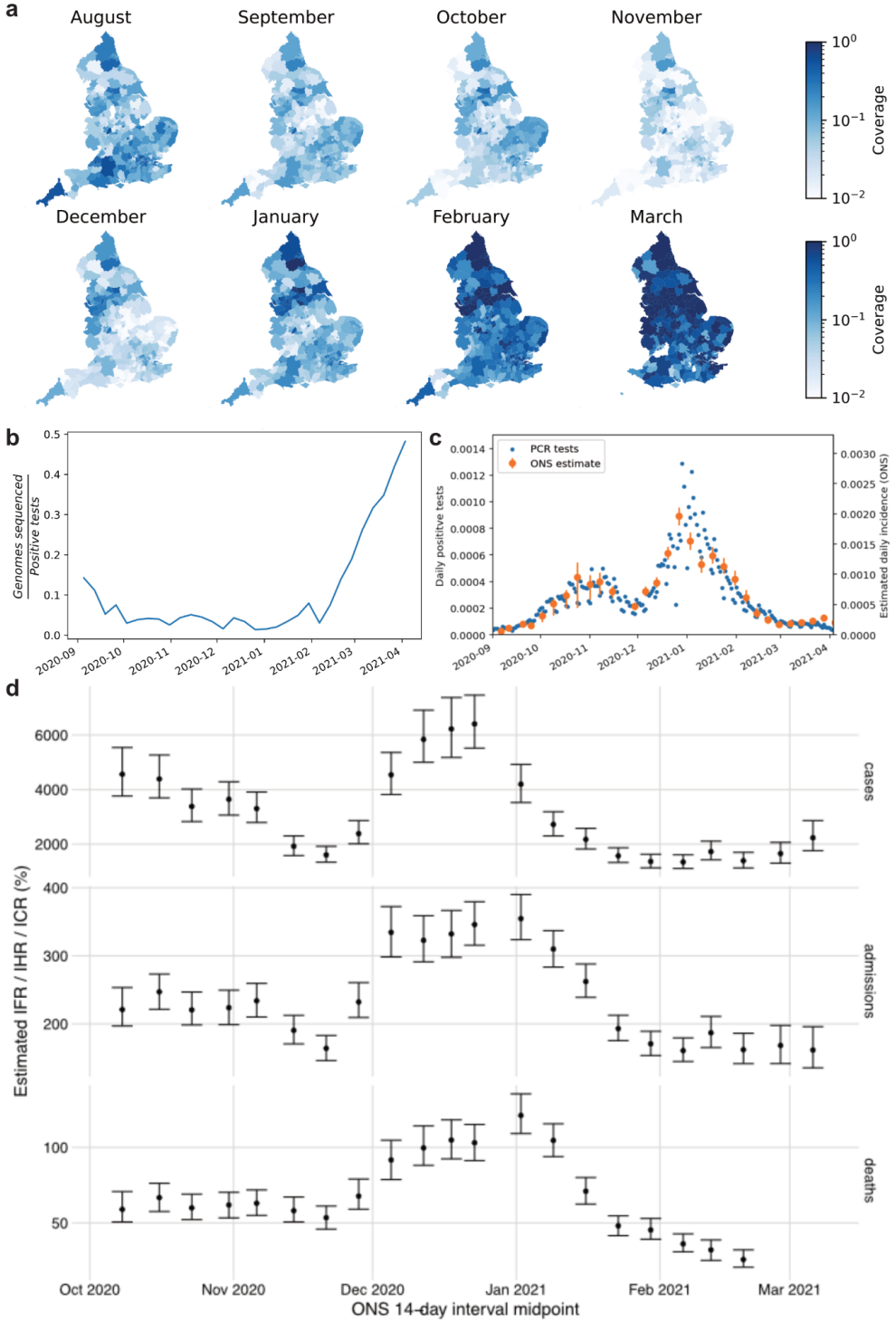
https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)_(England)_(No._4)_Regulations_2020&oldid=1014701607

Wikipedia contributors. (2021b, March 29). *The Health Protection (Coronavirus, Restrictions) (All Tiers) (England) Regulations 2020*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=The_Health_Protection_(Coronavirus,_Restrictions)_(All_Tiers)_(England)_Regulations_2020&oldid=1014831173

Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., Zhao, Y., Duyvesteyn, H. M. E., Tuekprakhon, A., Nutalai, R., Wang, B., Paesen, G. C., Lopez-Camacho, C., Slon-Campos, J., Hallis, B., Coombes, N., Bewley, K., Charlton, S., Walter, T. S., … Screaton, G. R. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell*. https://doi.org/10.1016/j.cell.2021.02.037

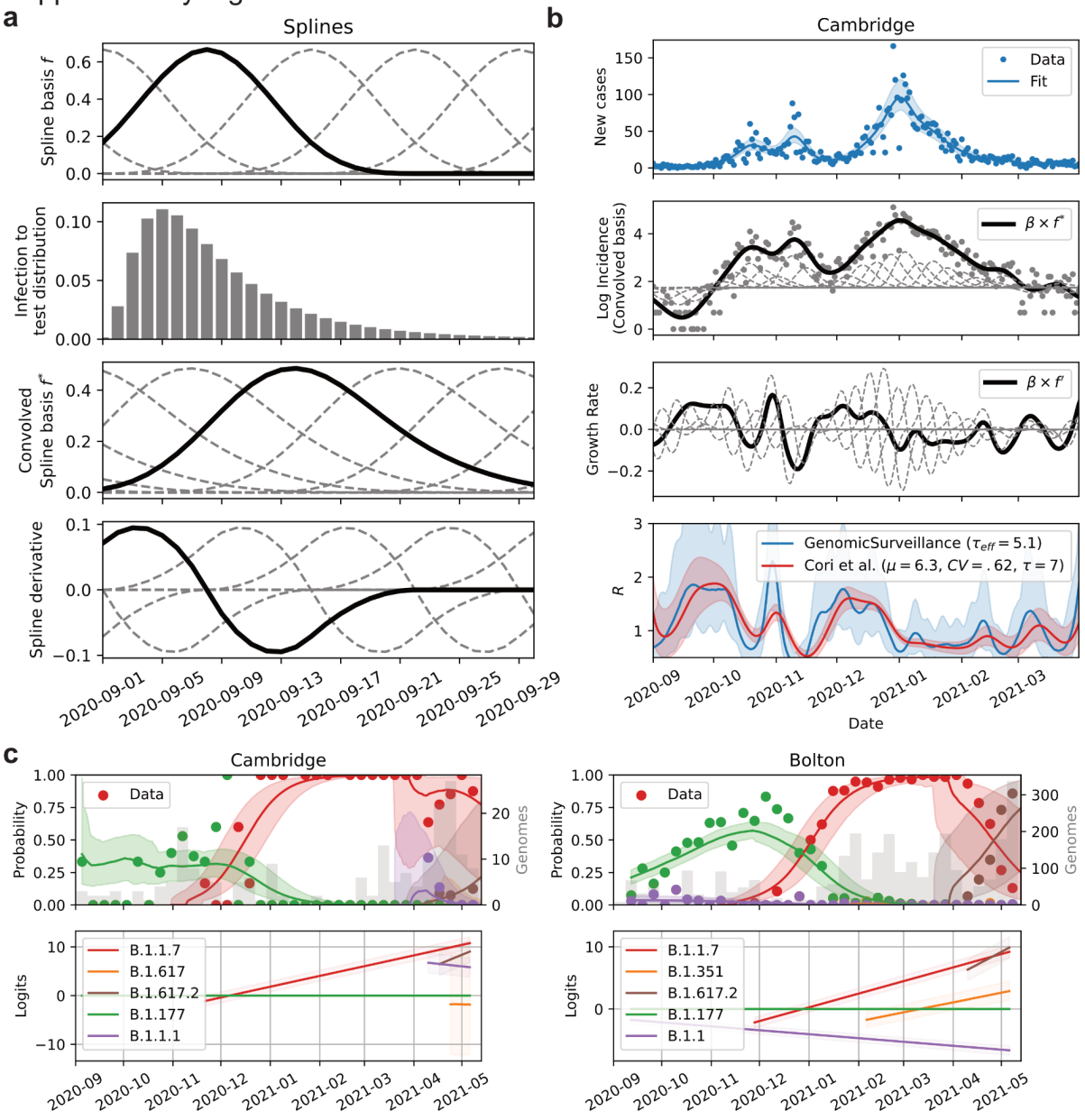
# Supplementary Figures

Supplementary Figure 1



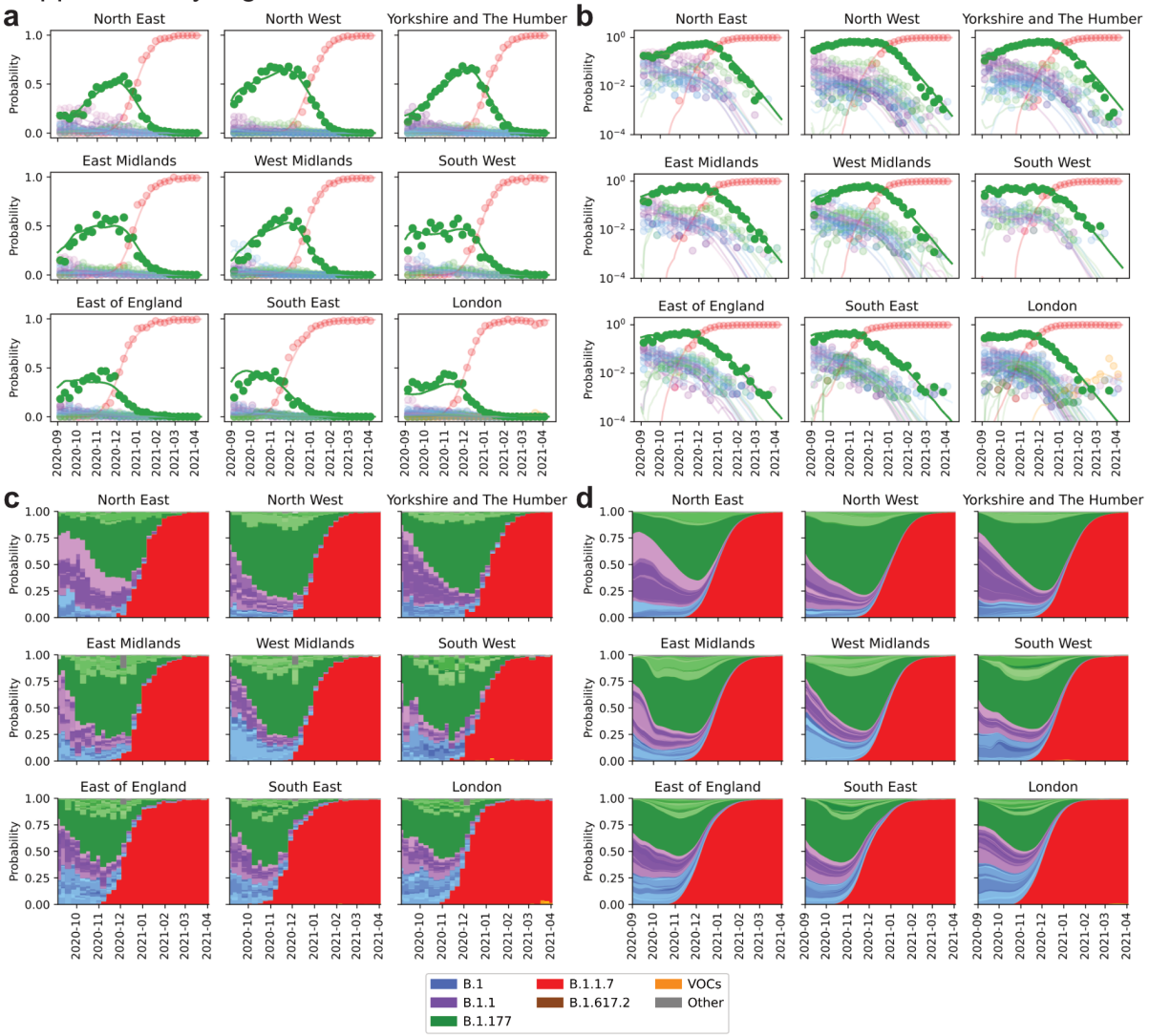

**Supplementary Figure 1, related to Figure 1. SARS-CoV-2 surveillance sequencing in England between September 2020 and April 2021. a**. Local monthly coverage across 315 LTLAs. **b**. Weekly coverage of genomic surveillance sequencing. **c**. Comparison of Pillar 2 SARS-CoV-2 incidence and cross-sectional incidences inferred from household surveys by the Office of National Statistics (ONS). **d**. Hospitalisation, case and infection fatality rates relative to ONS prevalence.

Supplementary Figure 2



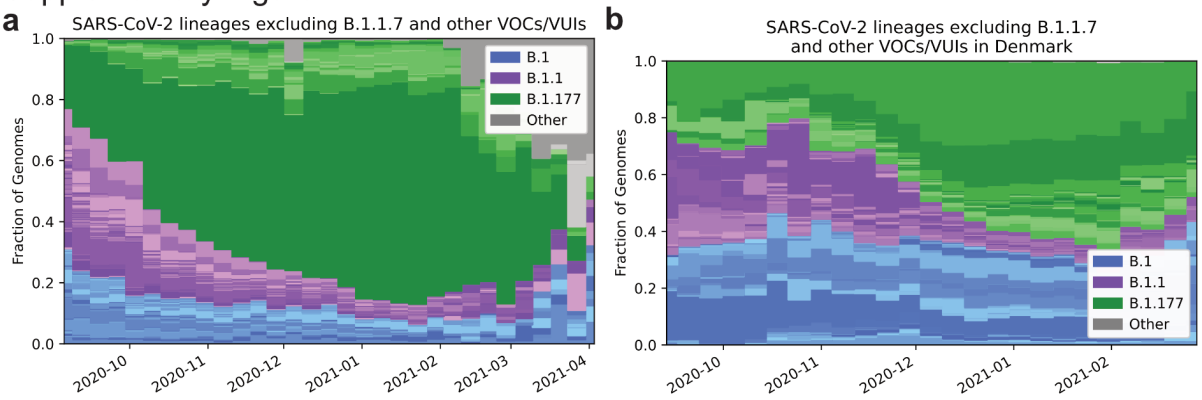

**Supplementary Figure 2: Genomic surveillance model of total incidence and lineage-specific frequencies. a.** Cubic basis splines (top row) are convolved with the infection to test distribution (row 2 and 3) and used to fit the log incidence in a LTLA and its corresponding derivatives (growth rates; bottom row). **b**. Example incidence (top row), logarithmic incidence with individual convolved basis functions (dashed lines, row 2), growth rate with individual spline basis derivatives (dashed lines, row 3) and resulting (case) reproduction numbers (growth rate per 5.1d) from our approach (GenomicSurveillance) and estimates by EpiEstim (Cori et al., 2013), 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.

Supplementary Figure 3



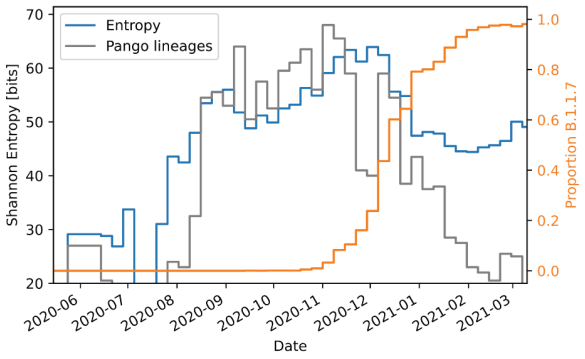

**Supplementary Figure 3, related to Figure 2. Spatiotemporal model of 62 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and April 2021. a.** Regional lineage specific relative frequency of lineages contributing more than 50 genomes during the time period shown. Dots denote observed data, lines the fits aggregated to each region. **b.** Same as **a**, but on a log scale. **c.** Same data as in a, shown as stacked bar charts. Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **d**. Same fits as in **a**, shown as stacked segments.

Supplementary Figure 4



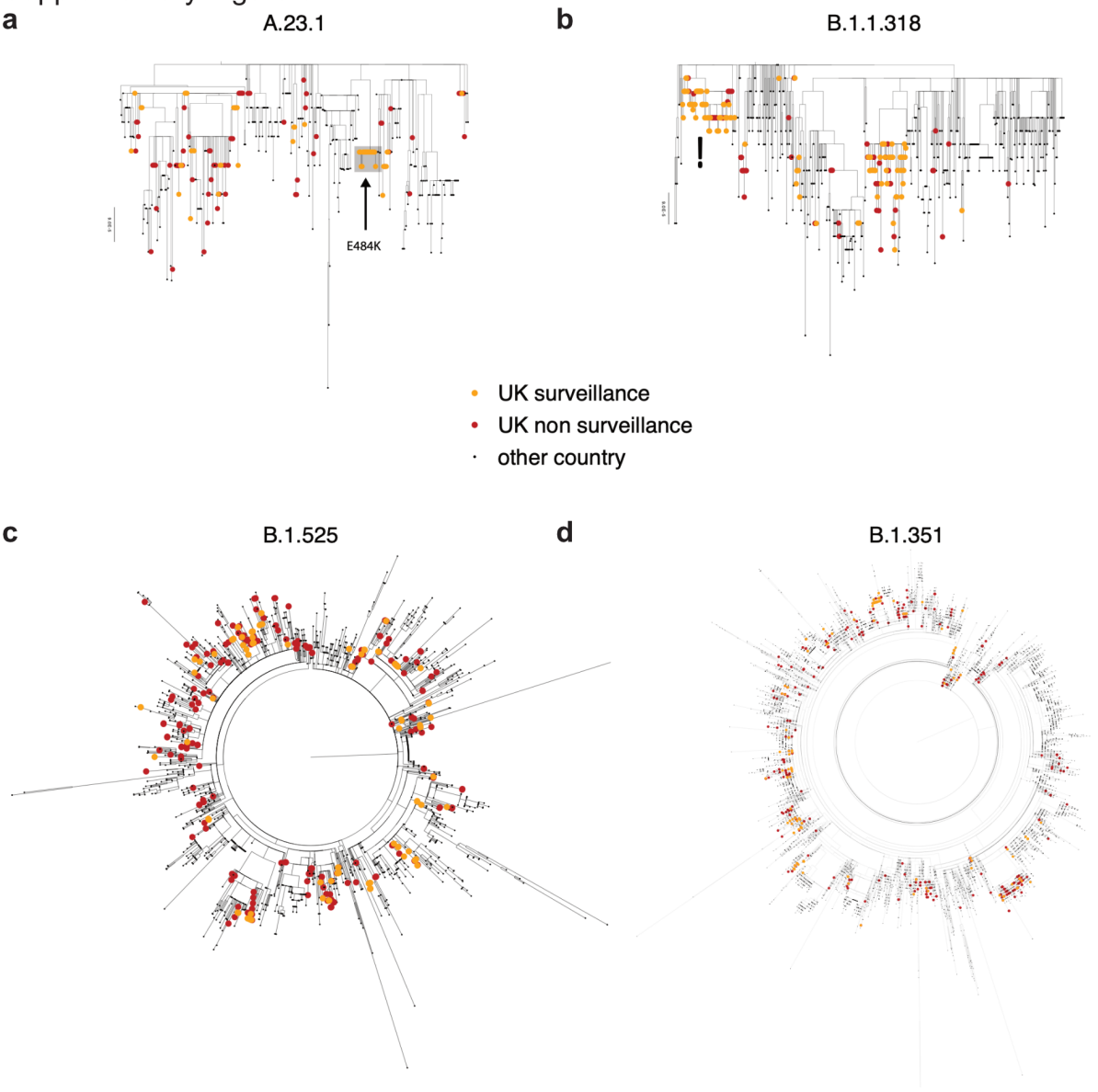

**Supplementary Figure 4, related to Figure 2. Relative growth of B.1.177. a.** Lineage-specific relative frequency data in England, excluding B.1.1.7 and other VOCs/VUIs (Category Other includes: A, A.18, A.20, A.23, A.25, A.27, A.28, B, B.29, B.40, None). Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **b.** Lineage-specific relative frequency data in Denmark, excluding B.1.1.7 and other VOCs/VUIs. Colors resemble major lineages as indicated and shadings thereof indicate sublineages.

Supplementary Figure 5



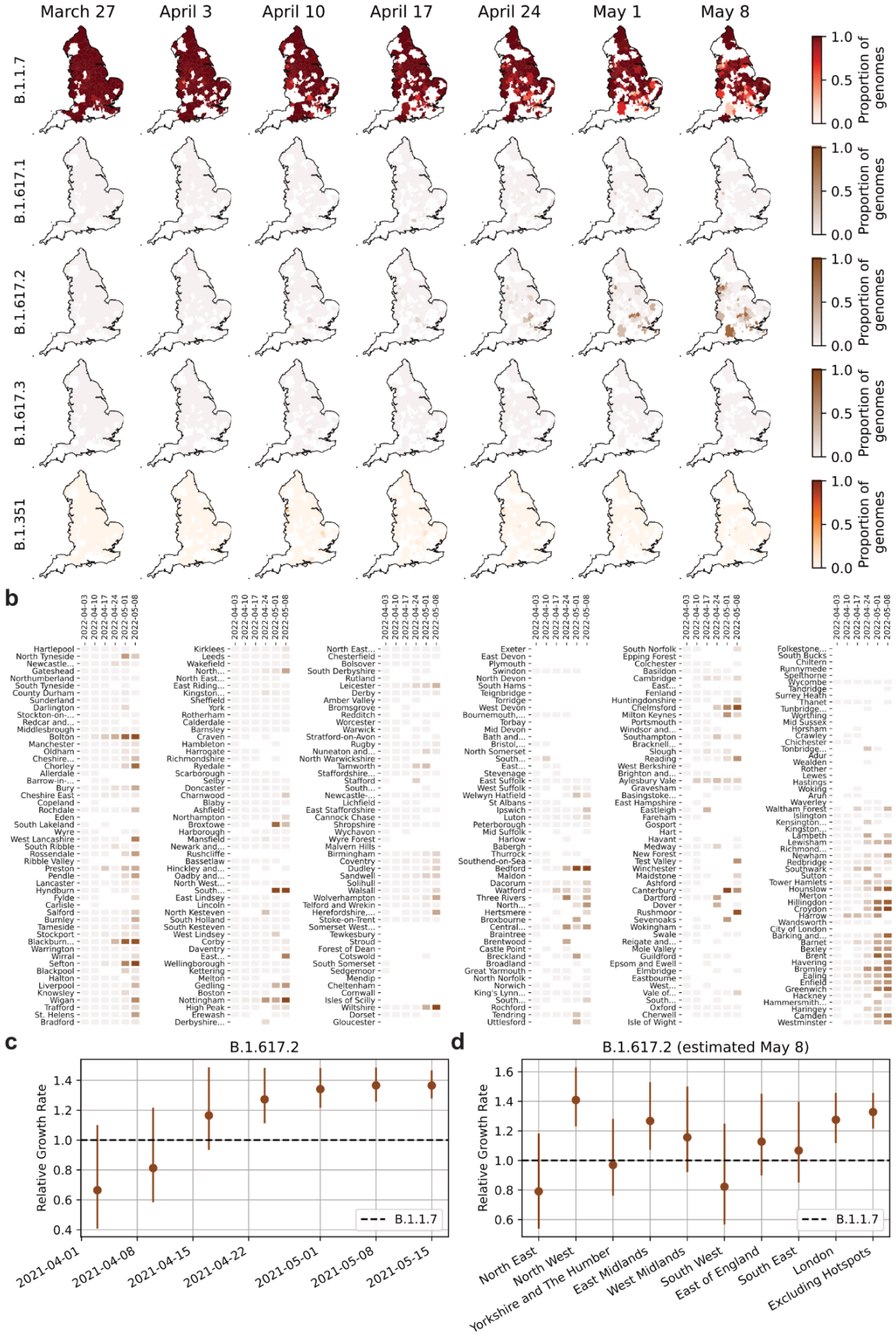

**Supplementary Figure 5, related to Figure 4.** Shown is the entropy (blue), total number of observed Pango lineages (grey, divided by 4), as well as the proportion of B.1.1.7 (orange, right axis). The sweep of B.1.1.7 causes an intermittent decline of genomic diversity as measured by the entropy.

Supplementary Figure 6

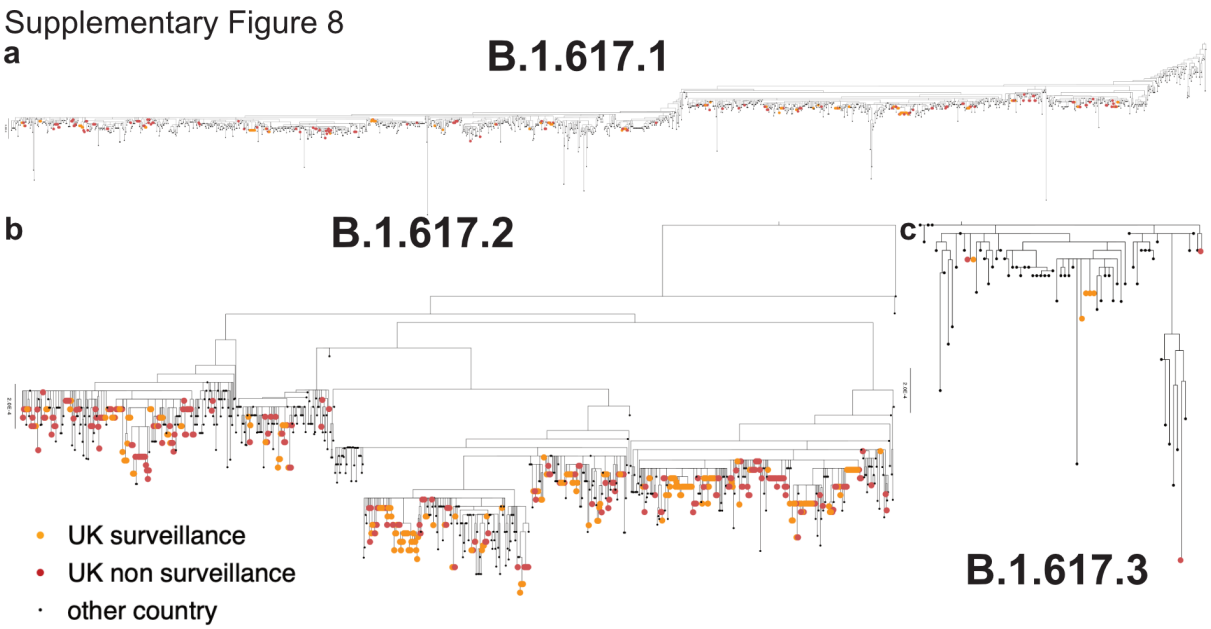



**Supplementary Figure 6, related to Figure 5. Global phylogenetic trees of selected VOCs/VUIs**. English surveillance and other (targeted and quarantine) samples are highlighted respectively orange and red.

Supplementary Figure 7

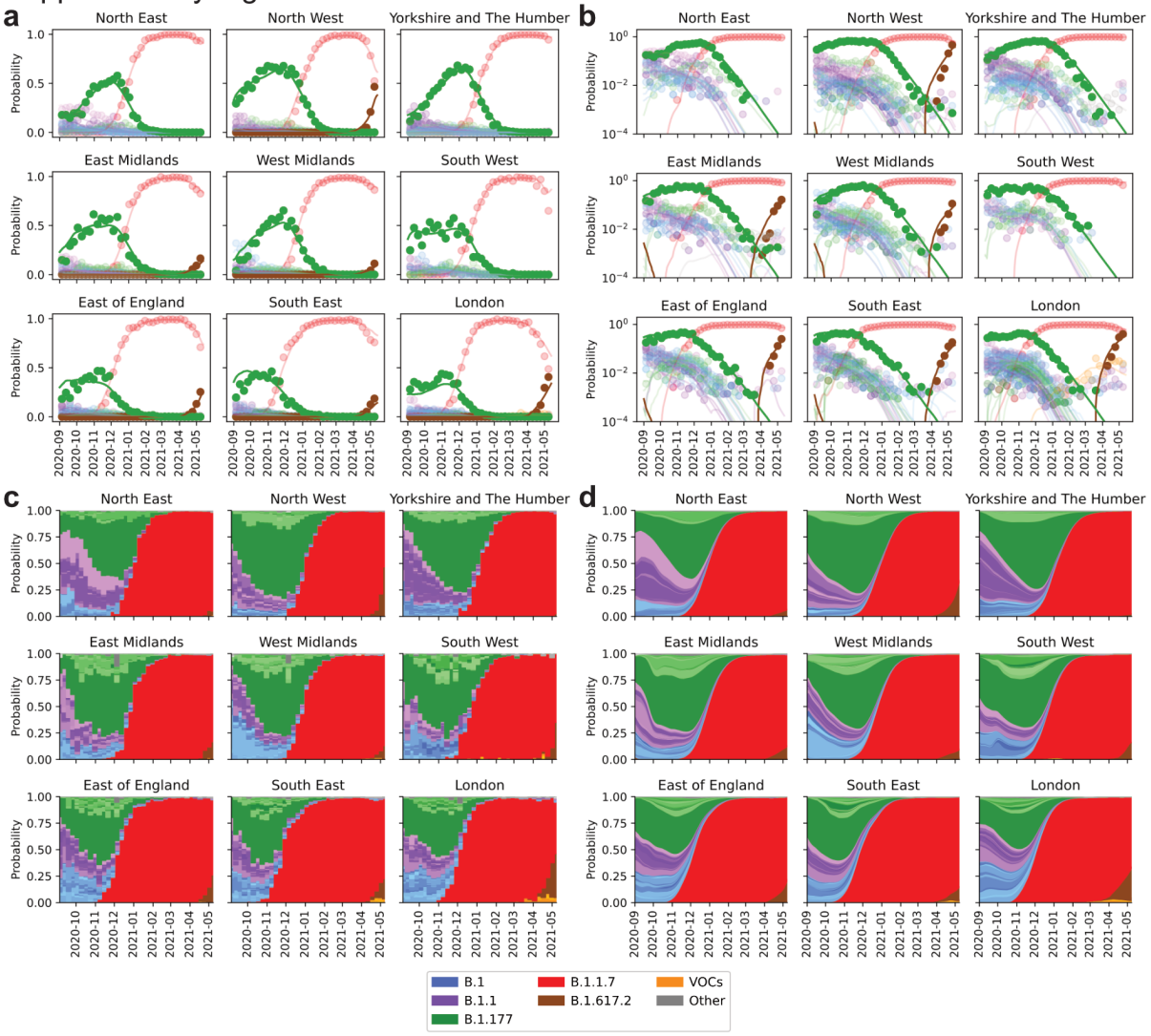


**Supplementary Figure 7, related to Figure 5: The rise of B.1.617 in April 2021. a.** Relative frequency of the lineages B.1.1.7, B.1.617.1/2/3 and B.1.351 during April and May 2021. **b.** Relative frequency of B.1.617.2 genomes in 317 LTLA (sorted by regions). LTLAs with a total of less than 5 genomes in a given week are shown in white in **a** and **b**. **c.** Relative growth rates of B.1.617.2 relative to B.1.1.7, estimated based on data ending in different weeks. **d.** Relative growth rates of B.1.617.2 relative to B.1.1.7, estimated in different regions of England and excluding B.1.617.2 hotspots (South Northamptonshire, Bedford, Blackburn with Darwen, Bromley, Bolton, Colchester). The regions North East, Yorkshire and the Humber as well as South West have low sequencing coverage, as shown in **a** and **b**, and therefore the estimated growth rates are lower and have greater uncertainty.




**Supplementary Figure 8, related to Figure 5. Global phylogenetic trees of B.1.617 sublineages**. 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.

Supplementary Figure 9




**Supplementary Figure 9, related to Figure 5. Spatiotemporal model of 64 SARS-CoV-2 lineages in 315 English LTLAs between September 2020 and May 2021 using provisional data. a.** Regional lineage specific relative frequency of lineages contributing more than 50 genomes during the time period shown. Dots denote observed data, lines the fits aggregated to each region. **b.** Same as **a**, but on a log scale. **c.** Same data as in **a**, shown as stacked bar charts. Colors resemble major lineages as indicated and shadings thereof indicate sublineages. **d**. Same fits as in **a**, shown as stacked segments.