Establishment & lineage dynamics of the SARS-CoV-2 epidemic in the UK

Louis du Plessis[†1], John T. McCrone[†2], Alexander E. Zarebski[†1], Verity Hill[†2], Christopher Ruis[†3,4], Bernardo Gutierrez[1,5], Jayna Raghwani[1], Jordan Ashworth[2], Rachel Colquhoun[2], Thomas R. Connor[6,7], Nuno R. Faria[1,8], Ben Jackson[2], Nicholas J. Loman[9], Áine O'Toole[2], Samuel M. Nicholls[9], Kris V. Parag[8], Emily Scher[2], Tetyana I. Vasylyeva[1], Erik M. Volz[8], Alexander Watts[12,13], Isaac I. Bogoch[10,11], Kamran Khan[10,12,13], the COVID-19 Genomics UK (COG-UK) Consortium[+,14], David M. Aanensen[15,16], Moritz U. G. Kraemer[‡1], Andrew Rambaut[*‡2], Oliver G. Pybus[*‡1,17]

[1] Department of Zoology, University of Oxford, Oxford, UK.

[2] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[3] Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, UK.

[4] Department of Veterinary Medicine, University of Cambridge, Cambridge, UK.

[5] School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador.

[6] School of Biosciences, Cardiff University, UK.

[7] Pathogen Genomics Unit, Public Health Wales NHS Trust, UK.

[8] MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, UK.

[9] Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK.

[10] Department of Medicine, University of Toronto, Toronto, Canada.

[11] Divisions of General Internal Medicine and Infectious Diseases, University Health Network, Toronto, Canada.

[12] Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Canada.

[13] BlueDot, Toronto, Canada.

[14] https://www.cogconsortium.uk

[15] Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK.

[16] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK.

[17] Department of Pathobiology & Population Sciences, Royal Veterinary College London, UK.

[*] Correspondence to: a.rambaut@ed.ac.uk; oliver.pybus@zoo.ox.ac.uk

[†] These authors contributed equally; [‡] These authors contributed equally.

[+] Full list of consortium names and affiliations are in Supplementary Material.

Abstract

The UK's COVID-19 epidemic during early 2020 was one of world's largest and unusually well represented by virus genomic sampling. Here we reveal the fine-scale genetic lineage structure of this epidemic through analysis of 50,887 SARS-CoV-2 genomes, including 26,181 from the UK sampled throughout the country's first wave of infection. Using large-scale phylogenetic analyses, combined with epidemiological and travel data, we quantify the size, spatio-temporal origins and persistence of genetically-distinct UK transmission lineages. Rapid fluctuations in virus importation rates resulted in >1000 lineages; those introduced prior to national lockdown were larger and more dispersed. Lineage importation and regional lineage diversity declined after lockdown, whilst lineage elimination was size-dependent. We discuss the implications of our genetic perspective on transmission dynamics for COVID-19 epidemiology and control.

Introduction

Infectious disease epidemics are composed of chains of transmission, yet surprisingly little is known about how co-circulating transmission lineages vary in size, spatial distribution, and persistence, and how key properties such as epidemic size and duration arise from their combined action. Whilst individual-level contact tracing investigations can reconstruct the structure of small-scale transmission clusters (e.g. *1-3*) they cannot be extended practically to large national epidemics. However, recent studies of Ebola, Zika, influenza and other viruses have demonstrated that virus emergence and spread can be instead tracked using large-scale pathogen genome sequencing (e.g. *4-7*). Such studies show that regional epidemics can be highly dynamic at the genetic level, with recurrent importation and extinction of transmission chains within a given location. In addition to measuring genetic diversity, understanding pathogen lineage dynamics can help target interventions effectively (e.g. *8, 9*), track variants with potentially different phenotypes (e.g. *10, 11*), and improve the interpretation of incidence data (e.g. *12, 13*).

The rate and scale of virus genome sequencing worldwide during the COVID-19 pandemic has been unprecedented, with >100,000 SARS-CoV-2 genomes shared online by 1 October 2020 (*14*). Notably, approximately half of these represent UK infections and were generated by the national COVID-19 Genomics UK (COG-UK) consortium (*15*). The UK experienced one of the largest epidemics worldwide during the first half of 2020. Numbers of positive SARS-CoV-2 tests rose in March and peaked in April; by 26 June there had been 40,453 nationally-notified COVID-19 deaths in the UK (deaths occurring ≤28 days of first positive test; *16*). Here, we combine this large genomic data set with epidemiological and travel data to provide a full characterisation of the genetic structure and lineage dynamics of the UK epidemic.

Our study encompasses the initial epidemic wave of COVID-19 in the UK and comprises all SARS-CoV-2 genomes available before 26 June 2020 (50,887 genomes, of which 26,181 were from the UK; Fig 1A). The data represents genomes from 9.29% of confirmed UK COVID-19 cases by 26 June (*16*). Further, using an estimate of the actual size of the UK epidemic (*17*) we infer virus genomes were generated for 0.66% (95% CI=0.46-0.95%) of all UK infections by 5th May.

Genetic structure and lineage dynamics of the UK epidemic

We first sought to identify and enumerate all independently introduced, genetically-distinct chains of infection within the UK. We developed a large-scale molecular clock phylogenetic pipeline to identify "UK transmission lineages" that (i) contain two or more UK genomes and (ii) descend from an ancestral lineage inferred to exist outside of the UK (Fig. S1, S2). Sources of statistical uncertainty in lineage assignation were taken into account. We identified a total of 1179 (95% HPD=1143-1286) UK transmission lineages. Although each is intended to capture a chain of local transmission arising from a single importation event, some UK transmission lineages will be unobserved or aggregated due to limited SARS-CoV-2 genetic diversity (*18*) or incomplete or uneven genome sampling (*19, 20*). Therefore we expect this number to be an underestimate (see Methods). In our phylogenetic analysis 1650 (95% HPD=1611-1783) UK genomes could not be allocated to a UK transmission lineage (singletons).

Most transmission lineages are small and 72.4% (95% HPD=69.3-72.9%) contain <10 genomes (Fig. 1B). However the lineage size distribution is strongly skewed and follows a power-law distribution (Fig. 1B inset), such that the 8 largest UK transmission lineages contain >25% of all sampled UK genomes (Fig. 1C, Figs. S4-S7 show further visualisations). Although the two largest transmission lineages are estimated to comprise >1500 UK genomes each, there is phylogenetic uncertainty in their sizes (95% HPDs=1280-2133 and 1342-2011 genomes). All 8 largest lineages were first detected before the UK national lockdown on 23 March and, as expected, larger lineages were observed for longer (Pearson's r=0.82; 95% CI=0.8-0.83; Fig. S9). The sampling frequency of lineages of varying sizes differed over time (Fig. 1D); whilst UK transmission lineages containing >100 genomes consistently accounted for >40% of weekly sampled genomes, the proportion of small transmission lineages (≤10 genomes) and singletons decreased over the course of the epidemic (Fig. 1D).

The detection of UK transmission lineages in our data changed markedly through time. In early March the epidemic was characterised by lineages first observed within the previous week (Fig. 1E). The per-genome rate of appearance of new lineages was initially high, then declined throughout March and April (Fig. 1F), such that by 1st May 96.2% of sampled genomes belonged to transmission lineages that were first observed >7 days previously. By 1st June, a growing number of lineages (>73%) had not been detected by genomic sampling for >4 weeks, suggesting that they were rare or had gone extinct, a result that is robust to the sampling rate (Fig. 1F, 1A). Together, these results indicate that the UK's first epidemic wave resulted from the concurrent growth of many hundreds of independently-introduced transmission lineages, and that the introduction of non-pharmaceutical interventions (NPIs) was followed by the apparent extinction of lineages in a size-dependent manner.

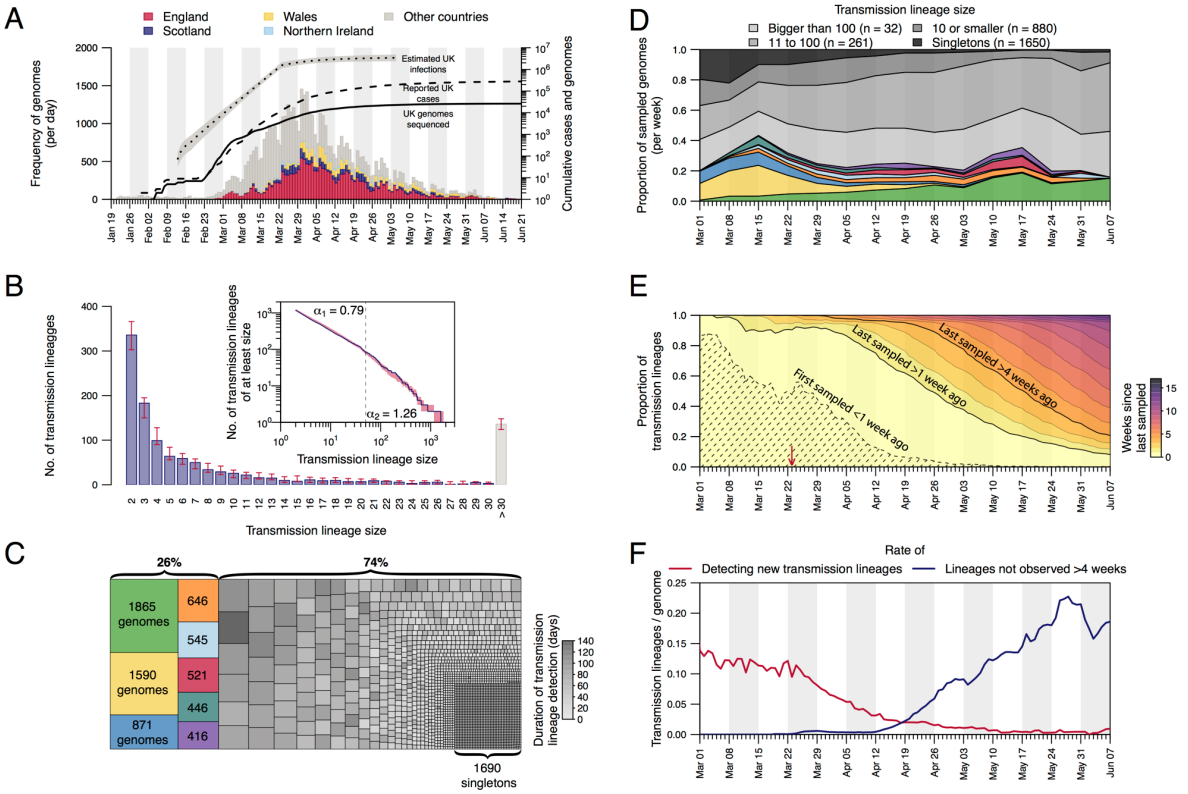

Fig. 1. Structure and dynamics of UK transmission lineages. (A) Collection dates of the 50,887 genomes analysed here (left-hand axis). Genomes are coloured by sampling location (England=red, Scotland=dark blue, Wales=yellow, Northern Ireland=light blue, elsewhere=grey). The solid line shows the cumulative number of UK virus genomes (right-hand axis). The dashed and dotted lines show, respectively, the cumulative number of laboratory-confirmed UK cases (by specimen date) and the estimated number of UK infections (17; grey shading=95% CI; right-hand axis). Due to retrospective screening, the cumulative number of genomes early in the epidemic exceeds that of confirmed cases. (B) Distribution of UK transmission lineage sizes. Blue bars show the number of transmission lineages of each size (red bars=95% HPD of these sizes across the posterior tree distribution). Inset: the corresponding cumulative frequency distribution of lineage size (blue line), on double logarithmic axes (red shading=95% HPD of this distribution across the posterior tree distribution). Values either side of vertical dashed line show coefficients of power-law distributions ($P[X \geq x] \sim x^{\alpha}$) fitted to lineages containing ≤50 (α_1) and >50 (α_2) virus genomes, respectively. (C) Partition of 26,181 UK genomes into UK transmission lineages and singletons, coloured by (i) lineage, for the 8 largest lineages, or (ii) duration of lineage detection (time between the lineage's oldest and most recent genomes) for the remainder. (D) Lineage size breakdown of UK genomes collected each week. Colours of the 8 largest lineages are as depicted in (C). (E) Trends through time in the detection of UK transmission lineages. For each day, all lineages detected up to that day are coloured by the time since the transmission lineage was last sampled. Isoclines correspond to weeks. Shaded area=transmission lineages that were first sampled <1 week ago. The red arrow indicates the start of the UK lockdown. (F) Red line=daily rate of detecting new transmission lineages. Blue line=rate at which lineages have not been observed for >4 weeks.

Transmission lineage diversity and geographic range

We also characterised the spatial distribution of UK transmission lineages using available data on 107 virus genome sampling locations, which correspond broadly to UK counties or metropolitan regions. Although genomes were not collected randomly (some lineages and regions will be over-represented due to targeted investigation of local outbreaks; e.g. *21*) the number of UK lineages detected in each region correlates with the number of genomes sequenced (Fig. 2A, Pearson's r=0.96, 95% CI=0.95-0.98) and the number of reported cases (Fig. S10, Pearson's r=0.6, 95% CI=0.44-0.72) in each region. Further, larger lineages were observed in more locations; every 100 additional genomes in a lineage increases its observed range by 6-7 regions (Fig. 2B; Pearson's r=0.8, 95% CI=0.78-0.82). Thus, bigger regional epidemics comprised a greater diversity of transmission lineages, and larger lineages were more geographically widespread. These observations indicate substantial dissemination of a subset of lineages across the UK and suggest many regions experienced a series of introductions of new lineages from elsewhere, potentially hindering the impact of local interventions.

We quantified the substantial variation among regions in the diversity of transmission lineages present using Shannon's index (SI; this value increases as both the number of lineages and the evenness of their frequencies increase; Fig. 2C). We observed the highest SIs in Hertfordshire (4.77), Greater London (4.62) and Essex (4.49); these locations are characterised by frequent commuter travel to/within London and proximity to major international airports (*22*). Locations with the three lowest non-zero SIs were in Scotland (Stirling=0.96, Aberdeenshire=1.04, Inverclyde=1.32; Fig. 2C).

To illustrate temporal trends in transmission lineage diversity, we plot SI through time for each of the UK's national capital cities (Fig. 2D). Lineage diversities in each peaked in late March and declined after the UK national lockdown, congruent with Figure 1E, F. Greater London's epidemic was the most diverse and characterised by an early, rapid rise in SI (Fig. 2C), consistent with epidemiological trends there (*16, 23*). Belfast's lineage diversity was notably lower.

We observe variation in the spatial range of individual UK transmission lineages. Although some lineages are widespread, most are more localised and the range size distribution is right-skewed (Fig. S11), congruent with an observed abundance of small lineages (Fig. 1B, 2B) and biogeographic theory (e.g. *24*). For example, lineage DTA_13 is geographically dispersed (>50% of sequence pairs sampled >234km apart) whereas DTA_290 is strongly local (95% of sequence pairs sampled <100km apart) and DTA_62 has multiple foci of sampled genomes (Fig. 2E, S12). The national distribution of cases therefore arose from the aggregation of multiple heterogeneous lineage-specific patterns.

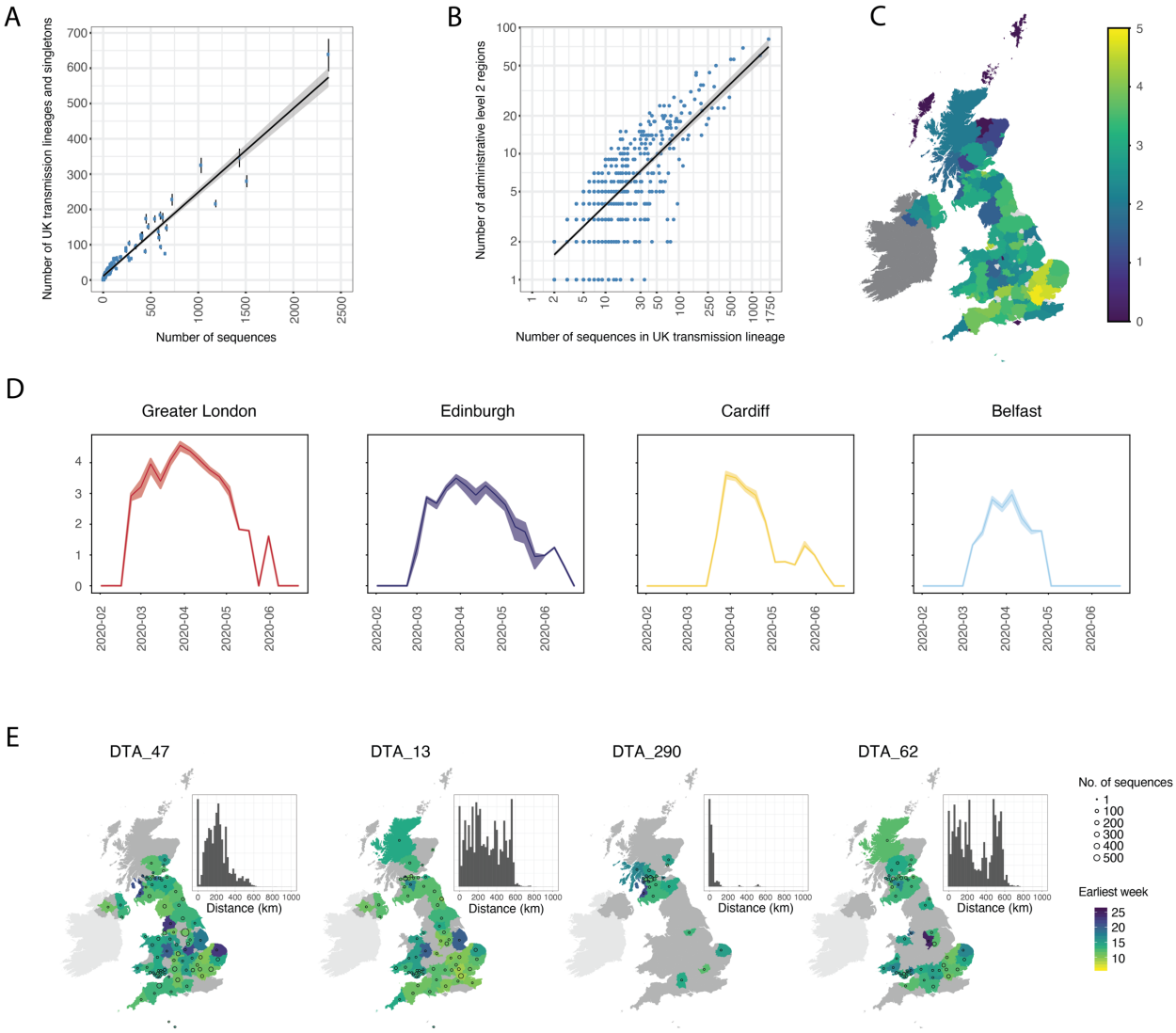

Fig. 2. Spatial distribution of UK transmission lineages. (A) Correlation between the number of transmission lineages detected in each region (points=median values, bars=95% HPD intervals) and the number of UK virus genomes from each region (Pearson's r=0.96, 95% CI=0.95-0.98, p<0.001). (B) Correlation between the spatial range of each transmission lineage and the number of virus genomes it contains (Pearson's *r=0.8, 95% CI=0.78-0.82, p<0.001*) *(C) Map showing Shannon's index (SI) for each region, calculated across the study period (2nd Feb-26th Jun). Yellow colours indicate higher SI values and darker colours lower values. (D) SI through time for the UK national capital cities. (E) Illustration of the diverse spatial range distributions of UK transmission lineages. Colours represent the week of the first detected genome in the transmission lineage in each location. Circles show the number of sampled genomes per location. Insets show the distribution of geographic distances for all sequence pairs within the lineage (see Fig S12 for further details).*

Dynamics of international introduction of transmission lineages

The process by which transmission lineages are introduced to an area is an important aspect of early epidemic growth (e.g. *25*). To investigate this at a national scale we estimated the rate and source of SARS-CoV-2 importations into the UK. Since standard phylogeographic approaches were precluded by strong biases in genome sampling among countries (*19*), we developed a new approach that combines virus phylogenetics with epidemiological and travel data. First, we estimated the TMRCA (time to the most recent common ancestor) of each UK transmission lineage. The TMRCAs of most UK lineages are dated to March and early April (median=21st March; IQR=14th-29th March). UK lineages with earlier TMRCAs are larger and longer-lived than those whose TMRCAs postdate the national lockdown (Fig. 3A, S15).

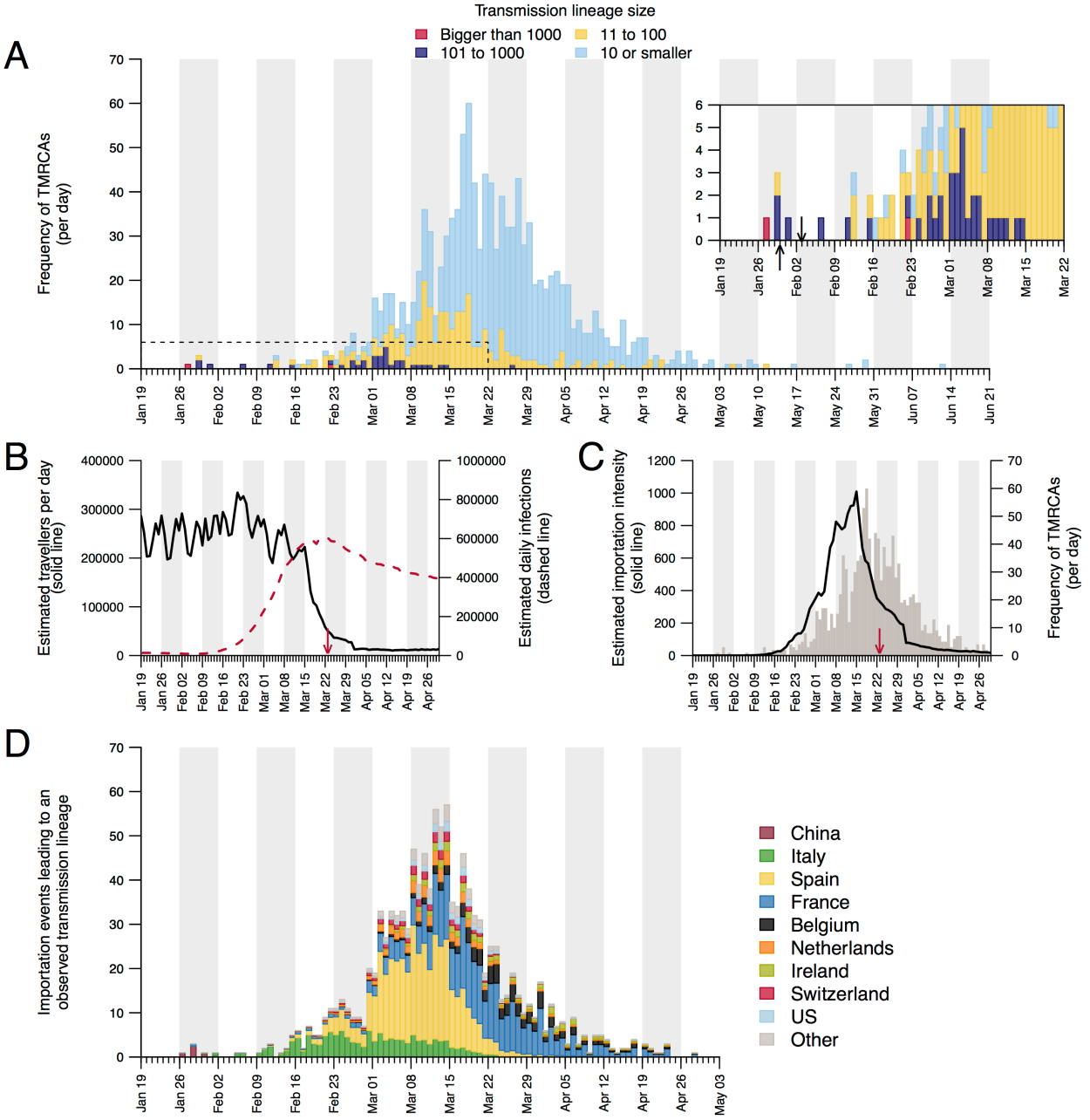

Fig. 3. Dynamics of transmission lineage importation (A) Histogram of lineage TMRCAs, coloured by lineage size. Inset: expanded view of the days prior to UK lockdown. Left-hand arrow=collection date of the UK's first laboratory-confirmed case; right-hand arrow=collection date of the earliest UK virus genome in our dataset. (B) Estimated number of inbound travellers to the UK per day (black) and estimated number of infectious cases worldwide (dashed red). Arrow here shows the start of the UK lockdown. (C) Estimated importation intensity (EII) curve (black) and the histogram of lineage TMRCAs (grey). (D) Estimated histogram of virus lineage importation events per day, obtained from our lag model. Colours show the proportion attributable each day to inbound travel from various countries (see Figs. S19, S20, Table S4). This assignment is statistical, i.e. we cannot ascribe a specific source location to any given lineage.

Due to incomplete sampling, TMRCAs best represent the date of the first inferred transmission event in a lineage, not its importation date (Fig. S2). To infer the latter, and quantify the delay between importation and onward within-UK transmission, we generated daily estimates of the number of travellers arriving in the UK and of global SARS-CoV-2 infections (see Methods) worldwide. Before March, the UK received ~1.75m inbound travellers per week (school holidays explain the end-February ~10% increase; Fig. 3B). International arrivals fell by ~95% during March and this reduction was maintained through April. Elsewhere, estimated numbers of infectious cases peaked in late March (Fig. 3B). We combined these two trends to generate an estimated importation intensity (EII) - a daily empirical measure of the intensity of SARS-CoV-2 importation into the UK. The EII peaks in mid-March, when high UK inbound travel volumes coincided with growing numbers of infectious cases elsewhere (Fig. 3B, C).

Crucially, the EII's temporal profile closely matches, but precedes, that of the TMRCAs of UK transmission lineages (Fig. 3A, C). The difference between the two represents the "importation lag", the time elapsed between lineage importation and the first detected local transmission event (Fig. S2). Using a statistical model, we estimate importation lag to be on average 8.22 ± 5.21 days (IQR=3.35-15.18) across all transmission lineages. Further, importation lag is strongly size-dependent; average lag is ~10 days for lineages comprising ≤10 genomes and <1 day for lineages of >100 genomes (Table S2). This size-dependency likely arises because the earliest transmission event in a lineage is more likely to be captured if it contains many genomes (Fig. S2; see Methods). We use this model to impute an importation date for each UK transmission lineage (Fig. 3D). Importation was unexpectedly dynamic, rising and falling substantially over only 4 weeks, hence 80% of importations (that gave rise to detectable UK transmission lineages) occurred between 27 February and 30 March. The delay between the inferred date of importation and the first genomic detection of each lineage was 14.13 ± 5.61 days on average (IQR=10-18) and declined through time (Table S2, S3).

To investigate country-specific contributions to virus importation we generated separate importation intensity (EII) curves for each country (Fig. S17). Using these values, we estimated the numbers of inferred importations each day attributable to inbound travel from each source location. As with the rate of importation (Fig. 3A), the relative contributions of arrivals from different countries were dynamic (Fig. 3D). Dominant source locations shifted rapidly in February and March and the diversity of source locations increased in mid-March (Fig. S17). Earliest importations were most likely from China or elsewhere in Asia but were rare compared to those from Europe. Over our study period we infer ~33% of UK transmission lineages

stemmed from arrivals from Spain, 29% from France, 12% from Italy and 26% from elsewhere (Fig. S20; Table S4). These large-scale trends were not apparent from individual-level travel histories; routine collection of such data ceased on 12 March (*26*).

Conclusions

The exceptional size of our genomic survey provides insight into the micro-epidemiological patterns that underlie the features of a large, national COVID-19 epidemic, allowing us to quantify the abundance, size distribution, and spatial range of transmission lineages. Pre-lockdown, high travel volumes and few restrictions on international arrivals (Table S5; Fig. 3B) led to the establishment and co-circulation of >1000 identifiable UK transmission lineages (Fig. 3A), jointly contributing to accelerated epidemic growth that quickly exceeded national contact tracing capacity (*26*). The relative contributions of importation and local transmission to initial epidemic dynamics under such circumstances warrants further investigation. We expect similar trends occurred in other countries with comparably large epidemics and high international travel volumes; virus genomic studies from regions with smaller or controlled COVID-19 epidemics have reported high importation rates followed by more transient lineage persistence (e.g. *27-29*).

Earlier lineages were larger, more dispersed, and harder to eliminate, highlighting the importance of rapid or pre-emptive interventions in reducing transmission (e.g. *30-32*). The high heterogeneity in SARS-CoV-2 transmission at the individual level (*33-35*) appears to extend to whole transmission lineages, such that >75% of sampled viruses belong to the top 20% of lineages ranked by size. Whilst the national lockdown coincided with limited importation and reduced regional lineage diversity, its impact on lineage extinction was size-dependent (Fig. 1E, F). The over-dispersed nature of SARS-CoV-2 transmission likely exacerbated this effect (*36*), thereby favouring, as R_t declined, greater survival of larger widespread lineages and faster local elimination of lineages in low prevalence regions. The degree to which the surviving lineages contributed to the UK's ongoing second epidemic is currently under investigation. The transmission structure and dynamics measured here provide a new context in which future public health actions at regional, national, and international scales should be planned and evaluated.

References

1. Centers for Disease Control and Prevention, Severe acute respiratory syndrome - Singapore, 2003. *MMWR* **52**, 405-411 (2003).

2. O. Faye, P.Y. Boëlle, E. Heleze, O. Faye, C. Loucoubar, N. Magassouba, B. Soropogui, S. Keita, T. Gakou, E.H.I. Bah, L. Koivogui, A.A. Sall, S. Cauchemez, Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *Lancet Infect. Dis.* **15**, 320-326 (2015).

3. K.H. Kim, T.E. Tandi, J.W. Choi, J.M. Moon, M.S. Kim, Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak in South Korea, 2015: epidemiology, characteristics and public health implications. *J. Hosp. Infect.* **95**, 207-213 (2017).

4. J. Bahl, M.I. Nelson, K.H. Chan, R. Chen, D. Vijaykrishna, R.A. Halpin, T.B. Stockwell, X. Lin, D.E. Wentworth, E. Ghedin, Y. Guan, J.S.M. Peiris, S. Riley, A. Rambaut, E.C. Holmes,

G.J.D. Smith, Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19359-19364 (2011).

5. G.J. Baille, M. Galiano, P.M. Agapow, R. Myers, R. Chiam, A. Gall, A.L. Palser, S.J. Watson, J. Hedge, A. Underwood, S. Platt, E. McLean, R.G. Pebody, A. Rambaut, J. Green, R. Daniels, O.G. Pybus, P. Kellam, M. Zambon, Evolutionary dynamics of local pandemic H1N1/09 influenza lineages revealed by whole genome analysis. *J. Virol.* **86**, 11–18 (2012).

6. G. Dudas, L.M. Carvalho, T. Bedford, A.J. Tatem, G. Beale, N.R. Faria, D.J. Park, J.T. Ladner, A. Arias, D. Asogun, F. Biejelec, S.L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J.W. Diclaro, S. Duraffour, M.J. Elmore, L.S. Fakoli, O. Faye, M.L. Gilbert, S.M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D.S. Grant, B.L. Haagmans, J.A. Hiscox, U. Jah, J.R. Kugelman, D. Liu, J. Lu, C.M. Malbeouf, S. Mate, D.A. Matthews, C.B. Matranga, L.W. Meredith, J. Qu, J. Quick, S.D. Pas, M.V.T. Phan, G. Pollakis, C.B. Reusken, M. Sanchez-Lockhart, S.F. Schaffner, J.S. Schieffelin, R.S. Sealfon, E. Simon-Loriere, S.L. Smits, K. Stoecker, L. Thorne, E.A. Tobin, M.A. Vandi, S.J. Watson, K. West, S. Whitmer, M.R. Wiley, S.M. Winnicki, S. Wohl, R. Wölfel, N.L. Yozwiak, K.G. Andersen, S.O. Blyden, F. Bolay, M.W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G.F. Gao, R.F. Garry, I. Goodfellow, S. Günther, C.T. Happi, E.C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M.P.G. Koopmans, J.H. Kuhn, N.J. Loman, N. Magassouba, D. Naidoo, S.T. Nichol, T. Nyenswah, G. Palacios, O.G. Pybus, P.C. Sabeti, A. Sall, U. Ströher, I. Wurie, M.A. Suchard, P. Lemey, A. Rambaut, Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309-315 (2017).

7. N.D. Grubaugh, J.T. Ladner, M.U.G. Kraemer, G. Dudas, A.L. Tan, K. Gangavarapu, M.R. Wiley, S. White, J. Thézé, D.M. Magnani, K. Prieto, D. Reyes, A.M. Bingham, L.M. Paul, R. Robles-Sikisaka, G. Oliveira, D. Pronty, C.M. Barcellona, H.C. Metsky, M.L. Baniecki, K.G. Barnes, B. Chak, C.A. Freije, A. Gladden-Young, A. Gnirke, C. Luo, B. MacInnis, C.B. Matranga, D.J. Park, J. Qu, S.F. Schaffner, C. Tomkins-Tinch, K.L. West, S.M. Winnicki, S. Wohl, N.L. Yozwiak, J. Quick, J.R. Fauver, K. Khan, S.E. Brent, R.C. Reiner Jr, P.N. Lichtenberger, M.J. Ricciardi, V.K. Bailey, D.I. Watkins, M.R. Cone, E.W. Kopp IV, K.N. Hogan, A.C. Cannons, R. Jean, A.J. Monoghan, R.F. Garry, N.J. Loman, N.R. Faria, M.C.Porcelli, C. Vasquez, E.R. Nagle, D.A.T. Cummings, D. Stanek, A. Rambaut, M. Sanchez-Lockhart, P.C. Sabeti, L.D. Gillis, S.F. Michael, T. Bedford, O.G. Pybus, S. Isern, G. Palacios, K.G. Andersen, Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401-405 (2017).

8. A.F.Y. Poon, R. Gustafson, P. Daly, L. Serr, S.E. Demlow, J. Wong, C.K. Woods, R.S. Hogg, M. Krajden, D. Moore, P. Kendall, J.S.G. Montaner, P.R. Harrigan, Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV* **3**, e231-e238 (2016).

9. J. Thomas, N. Govender, K.M. McCarthy, L.K. Erasmus, T.J. Doyle et al. Outbreak of Listeriosis in South Africa Associated with Processed Meat. *N. Engl. J. Med.* **382**, 632-643 (2020).

10. M.A. Beale, M. Marks, S.K. Sahi, L.C. Tantalo, A.V. Nori, P. French, S.A. Lukehart, C.M. Marra, N.R. Thomson, Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat. Commun.* **10**, 3255 (2019).

11. E.M. Volz, V. Hill, J.T. McCrone, A. Price, D. Jorgensen, A. O'Toole, J.A. Southgate, R. Johnson, B. Jackson, F.F. Nascimento, S.M. Rey, S.M. Nicholls, R.M. Colquhoun, A. da Silva Filipe, J.G. Shepherd, D.J. Pascall, R. Shah, N. Jesudason, K. Li, R. Jarrett, N. Pacchiarini, M. Bull, L. Geidelberg, I. Siveroni, I.G. Goodfellow, N.J. Loman, O. Pybus, D.L. Robertson, E.C. Thomson, A. Rambaut, T.R. Connor, The COVID-19 Genomics UK Consortium, Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *medRxiv* https://doi.org/10.1101/2020.07.31.20166082 (2020).

12. L.M. Li, N.C. Grassly, C. Fraser, Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evol.* **34**, 2982-2995 (2017).

13. N.D. Grubaugh, J.T. Ladner, P. Lemey, O.G. Pybus, A. Rambaut, E. Holmes, K.G. Andersen, Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10-19 (2019).

14. Y. Shi, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

15. COVID-19 Genomics UK (COG-UK) Consortium, An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99-e100 (2020).

16. GOV.UK. Coronavirus (COVID-19) in the UK. https://coronavirus.data.gov.uk/cases (2020).

17. S. Flaxman, S. Mishra, A. Gandy, H.J.T. Unwin, T.A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J.W. Eaton, M. Monod, Imperial College COVID-19 Response Team, A.C. Ghani, C.A. Donnelly, S. Riley, M.A.C. Vollmer, N.M. Ferguson, L.C. Okell, S. Bhatt. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257-261 (2020).

18. C.J. Villabona-Arenas, W.P. Hanage, D.C. Tully, Phylogenetic interpretation during outbreaks requires caution. *Nat. Microbiol.* **5**, 876-877 (2020).

19. M. Worobey, J. Pekar, B.B. Larsen, M.I. Nelson, V. Hill, J.B. Joy, A. Rambaut, M.A. Suchard, J.O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America. *Science*, eabc8169 (2020).

20. S.A. Nadeau, T.G. Vaughan, J. Sciré, J.S. Huisman, T. Stadler, The origin and early spread of SARS-CoV-2 in Europe. *medRxiv* https://doi.org/10.1101/2020.06.10.20127738 (2020).

21. GOV.WALES. Genomic analysis of Covid-19 lineages in Wales. https://gov.wales/genomic-analysis-covid-19-lineages-wales (2020).

22. Greater London Authority Intelligence and Analysis Unit, "Census Information Scheme: Commuting in London" (CIS 2014-11, GLA, 2014; https://londondatastore-upload.s3.amazonaws.com/Zho%3Dttw-flows.pdf).

23. C. Angus, CoVid Plots and Analysis. The University of Sheffield https://doi.org/10.15131/shef.data.12328226 (2020).

24. K.J. Gaston, F. He, The distribution of species range size: a stochastic process. *Proc. R. Soc. Lond. B* **269**, 1079-1086 (2002).

25. G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, Mathematical models to characterize early epidemic growth: A Review. *Phys. Life Rev.* **18**, 66-97 (2016).

26. C. Baraniuk, Covid-19 contact tracing: a briefing. *BMJ* **369**, m1859 (2020).

27. J.L. Geoghegan, X. Ren, M. Storey, J. Hadfield, L. Jelley, S. Jefferies, J. Sherwood, S. Paine, S. Huang, J. Douglas, F.K. Mendes, A. Sporle, M.G. Baker, D.R. Murdoch, N. French, C.R. Simpson, D. Welch, A.J. Drummond, E.C. Holmes, S. Duchene, J. de Ligt, Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *medRxiv* https://www.medrxiv.org/content/10.1101/2020.08.05.20168930v3 (2020).

28. J. Lu, L. du Plessis, Z. Liu, V. Hill, M. Kang, H. Lin, J. Sun, S. François, M.U.G. Kraemer, N.R. Faria, J.T. McCrone, J. Peng, Q. Xiong, R. Yuan, L. Zeng, P. Zhou, C. Liang, L. Yi, J. Liu, J. Xiao, J. Hu, T. Liu, W. Ma, W. Li, J. Su, H. Zheng, B. Peng, S. Fang, W. Su, K. Li, R. Sun, R. Bai, X. Tang, M. Liang, J. Quick, T. Song, A. Rambaut, N. Loman, J. Raghwani, O.G. Pybus, C. Ke, Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003 (2020).

29. T. Seemann, C.R. Lane, N.L. Sherry, S. Duchene, A.G. da Silva, L. Caly, M. Sait, S.A. Ballard, K. Horan, M.B. Schultz, T. Hoang, M. Easton, S. Dougall, T.P. Stinear, J. Druce, M. Catton, B. Sutton, A. van Diemen, C. Alpren, D.A. Williamson, B.P. Howden, Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* **11**, 4376 (2020).

30. C. Dye, R.C.H. Cheng, J.S. Dagpunar, B.G. Williams, The scale and dynamics of COVID-19 epidemics across Europe. *medRxiv* https://doi.org/10.1101/2020.06.26.20131144 (2020).

31. H. Tian, Y. Liu, Y. Li, C.H. Wu, B. Chen, M.U.G. Kraemer, B. Li, J.. Cai, B. Xu, Q. Yang, B. Wang, P. Yang, Y. Cui, Y. Song, P. Zheng, Q. Wang, O.N. Bjornstad, R. Yang, B.T. Grenfell, O.G. Pybus, C. Dye, An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**(6491), 638-642 (2020).

32. K. Leung, J.T. Wu, D. Liu, G.M. Leung, First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a

modelling impact assessment. *Lancet* **395**, 1382-1393 (2020).

33. D.C. Adam, P. Wu, J.Y. Wong, E.H.Y. Lau, T.K. Tsang, S. Cauchemez, G.M. Leung, B.J. Cowling, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* (2020).

34. A. Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, S. Abbott et al. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]. *Wellcome Open Res.* **5**, 67 (2020).

35. L. Wang, X. Didelot, J. Yang, G. Wong, Y. Shi, W. Liu, F. Gao, Y. Bi, Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat. Commun.* **11**, 5006 (2020).

36. J.O. Lloyd-Smith, S.J. Schreiber, P.E. Kopp, W.M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355-359 (2005).

37. S.M. Nicholls, R. Poplawski, M.J. Bull, A. Underwood, M. Chapman, K. Abu-Dahab, B. Taylor, B. Jackson et al. MAJORA: Continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.10.06.328328v1 (2020).

38. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

39. A. Rambaut, E.C. Holmes, Á. O'Toole, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* https://doi.org/10.1038/s41564-020-0770-5 (2020).

40. Á. O'Toole, E. Scher, J.T. McCrone, B. Jackson, V. Hill, A. Underwood, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. du Plessis, R. Lanfear, D. Aanensen, E. Holmes, O. Pybus, A. Rambaut, Pangolin: phylogenetic assignment of named global outbreak lineages. https://cov-lineages.org/pangolin (2020).

41. M.N. Price, P.S. Dehal, A.P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

42. M.A. Suchard, P. Lemey, G. Beale, D.L. Ayres, A.J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).

43. M.A.R. Ferreira, M.A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36**, 355-369 (2008).

44. M.S. Gill, P. Lemey, N.R. Faria, A. Rambaut, B. Shapiro, M.A. Suchard, Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Mol. Biol. Evol.* **30**, 713-724 (2013).

45. A. Rambaut, A.J. Drummond, D. Xie, G. Beale, M.A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**(5), 901-904 (2018).

46. J.L. Thorne, H. Kishino, I.S. Painter, Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647-1657 (1998).

47. E.M. Volz, S.D.W. Frost, Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3**, vex025 (2017).

48. X. Didelot, N.J. Croucher, S.D. Bentley, S.R. Harris, D.J. Wilson, Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

49. P. Sagulenko, V. Puller, R.A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

50. M. Plummer, N. Best, K. Cowles, K. Vines, CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).

51. P. Lemey, A. Rambaut, A.J. Drummond, M.A. Suchard, Bayesian Phylogeography Finds Its Roots. *PLOS Comp. Biol.* **5**, 1-16 (2009).

52. V.N. Minin, M.A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391-412 (2008).

53. P.J. Lillie, A. Samson, A. Li, K. Adams, R. Capstick, G.D. Barlow, N. Easom, E. Hamilton, P.J. Moss, A. Evans, M. Ivan, PHE Incident Team, Y. Taha, C.J.A. Duncan, M.L. Schimd, Airborne HCID Network, Novel coronavirus disease (Covid-19): The first two patients in the UK with person to person transmission. *J. Infection* **80**, 578-606 (2020).

54. E. Dong, H. Du, L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533-534 (2020).

55. United Nations, Department of Economic and Social Affairs, Population Division, "World Population Prospects 2019: Methodology of the United Nations population estimates and projections" (ST/ESA/SER.A/425, UN, 2019; https://population.un.org/wpp/Publications/Files/WPP2019_Methodology.pdf).

56. X. He, E.H.Y. Lau, P. Wu, J. Wang, X. Hao, Y.C. Lau, J.Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B.J. Cowling, F. Li, G.M. Leu, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672-675 (2020).

57. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S.M Leung, E.H.Y. Lau, J.Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T.T.Y Lam, J.T. Wu, G.F. Gao, B.J. Cowling, B. Yang, G.M Leung, Z. Feng, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *New Engl. J. Med.* **382**, 1199-1207 (2020).

58. R. Verity, L.C. Okell, I. Dorigati, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P.G.T. Walker, H. Fu, A. Dighe, J.T. Griffin, M. Baguelin, S. Bhatia, A. Boonsyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C.A. Donelly, A.C. Ghani, N.M. Ferguson, Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 785-794 (2020).

59. H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.M. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A.R. Akhmetzhanov, N.M. Linton, Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int. J. Infect. Dis.* **94**, 154-155 (2020).

60. L. Roques, E.K. Klein, J. Papaïx, A. Sar, S. Souberyand, Using Early Data to Estimate the Actual Infection Fatality Ratio from COVID-19 in France. *Biology* **9**, 97 (2020).

61. T.W. Russell, J. Hellewell, C.I. Jarvis, K. van Zandvoort, S. Abbott, R. Ratnayake, CMMID COVID-19 working group, S. Flasche, R.M. Eggo, W.J. Edmunds, A.J. Kucharski, Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill.* **25**, pii=2000256 (2020).

62. K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* **25**, pii=2000180 (2020).

63. M. Day, Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* **369**, m1375 (2020).

64. M.J. Sanderson, How Many Taxa Must Be Sampled to Identify the Root Node of a Large Clade? *Syst. Biol.* **45**, 168-173 (1996).

Acknowledgments

We are grateful to everyone worldwide involved in generating the virus genome data shared on GISAID. We thank Samir Bhatt, Philippe Lemey, and Christopher Dye for insightful discussion. Funding: COG-UK is funded by 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. VH was supported by BBSRC grant BB/M010996/1; CR by a Fondation Botnar Research Award (Programme grant 6063) and UK Cystic Fibrosis Trust (Innovation Hub Award 001); JR by the UKRI GCRF One Health Poultry

Hub (BB/S011269/1); MUGK by a Branco Weiss Fellowship and EU H2020 project MOOD; NRF by WT fellowship 204311/Z/16/Z and MRC-FAPESP awards MR/S0195/1 and 18/14389-0; TIV by a Branco Weiss Fellowship; IIB is funded by the Canadian Institutes for Health Research (02179 – 000); JTM, RMC, NJL and AR by WT Collaborators Award 206298/Z/17/Z; AR and ES by ERC grant 725422; DMA by NIHR Global Health Research Unit (16/136/111); OGP, MUGK, LDP and AEZ by the Oxford Martin School. Author contributions: Study design: LdP, JTM, MUGK, AR, OGP. Methods development/programming: LdP, JTM, MUGK, AR, OGP, AEZ, VH, CR, JA, RC, TC, BJ, NJL, AO, SN, DMA, ES. Data analysis: LdP, MUGK, AR, OGP, AEZ, VH, CR, DMA, JTM, BG, KVP, ES, TIV. Data collection/experiments: AW, IIB, KK. Wrote paper: LdP, JR, MUGK, OGP. Edited paper/figure creation: LdP, MUGK, JR, CR, BG, TIV, NRF, EMV. Competing interests: KK is the founder of BlueDot, a social enterprise that develops digital technologies for public health. AW and IIB received employment or consulting income from BlueDot during this research. Data and materials availability: UK SARS-CoV-2 genomes and public metadata are available from www.cogconsortium.uk/data/ and deposited at gisaid.org and ENA (bioproject PRJEB37886). Non-UK genomes were obtained from gisaid.org. Raw data and analysis files for this work are in supplementary materials or available from GitHub at https://github.com/COG-UK/UK-lineage-dynamics-analysis, which also contains a list of sequence accession numbers.

Methods

Genomic data

All SARS-CoV-2 genomes available on GISAID (*14*) on 23 June 2020 were downloaded and combined with all SARS-CoV-2 genomes sequenced by the COG-UK consortium (*15*) by 26 June 2020 (available at https://www.cogconsortium.uk/data/). The pipeline used to collect and process raw SARS-CoV-2 sequence data and sample-associated metadata across the national COG-UK network is described in (*37*). We removed sequences that were from duplicate or environmental samples, those without exact collection dates, and those with large clusters of substitutions or large indels. Each genome sequence was aligned to the reference (Wuhan-Hu-1, GenBank: MN908947.3) using *minimap v2.17* (*38*) and the resulting SAM alignment was converted to a FASTA alignment, with the 5' and 3' UTRs of each genome masked by Ns. Insertions relative to the reference were discarded and site 11,083 (site position relative to MN908947), which is globally homoplasic, was also masked. Genomes that contained >5% Ns after mapping and those with a genetic distance to WH04 (GISAID: EPI_ISL_406801) more than 4 standard deviations from the epi-week mean genetic distance to WH04 were discarded. The final dataset consisted of 50,887 genomes sampled between 24 December 2019 and 22 June 2020, of which 26,181 (~51%) were from the UK (see Fig. 1A).

Geographical metadata

Administrative level 2 (admin2) metadata for the sampling location of UK virus genome sequences in the dataset (roughly equivalent to counties in the UK) required cleaning in order to be mapped to official admin2 regions, as found in the Global Administrative Database (GADM, https://gadm.org).

Some sampling locations in the metadata could not be unambiguously mapped to a known location (e.g. "City Centre"), while others were for locations in overseas territories (e.g. Falklands and Gibraltar). Yet other genome sequences had uninformative spatial records (e.g. Yorkshire or Wales), or no admin2 level data at all. For these (3431 of 26,181) the admin2 region was not mapped. We carried out a simple one-to-one mapping where possible, which included correcting spelling mistakes and alternative entries for the same county (e.g. Durham versus County Durham). Locations recorded at a higher spatial resolution were mapped to the corresponding admin2 region (e.g. Solihull was mapped to Birmingham). Where the recorded locations were larger than the admin2 regions (e.g. "West Midlands"), and most of the sequences in the area were from this larger conglomeration as opposed to its higher-resolution components, these admin2 regions were combined. When creating the map figures, we also merged some

city authorities with no reported sequences with their surrounding county, on the assumption that the larger county was used to represent the location of city samples (e.g. for Leicester and Leicestershire). Finally, genome sequences from Northern Ireland reported locations as historical counties, rather than the official admin2 designations, and so these historical counties were used instead. The cleaning code is provided on the GitHub repository (https://github.com/COG-UK/UK-lineage-dynamics-analysis).

Phylogenetic analysis and molecular clock dating

We developed a new Bayesian molecular clock phylogenetic analysis pipeline in order to reconstruct a posterior set of time-scaled phylogenetic trees for our exceptionally large virus genome dataset. Using the standard Bayesian approach it is currently impractical to estimate time-scaled trees directly from genome sequence data for more than a few thousand sequences. Therefore, we employed a number of extensions to make the analysis tractable.

First, we divided the full genome sequence dataset (n = 50,887) into five smaller datasets. Genomes were assigned SARS-CoV-2 lineages according to the nomenclature defined in *(39)* using *Pangolin* (*40;* github.com/cov-lineages/pangolin). Each lineage (and its sublineages) represents a monophyletic clade in the global SARS-CoV-2 phylogeny and can thus be analysed independently. For each lineage in A (n = 3591), B (n = 8821), B.1 (n = 22,861), B.1.1 (n = 15,616), we estimated an approximately maximum-likelihood tree using the Jukes-Cantor model in *FastTree v2.1.10* (*41*), then collapsed branch lengths shorter than 5×10^{-6} substitutions per site, which corresponded to distances smaller than one substitution across the whole virus genome, and likely result from nucleotide ambiguity codes in the genome sequences. By pruning out a large monophyletic clade the maximum-likelihood tree for B.1 was further divided into two trees, B.1.pruned (n = 12,275) and B.1.X (n = 10,586).

Prior to analysing the full dataset, an initial analysis was performed on a subset of genomes to obtain estimates of the molecular clock rate and of the TMRCA of each large-scale phylogenetic tree defined above. The full dataset was subsampled as evenly as possible across epi-weeks and countries with a slight enrichment for samples immediately descended from five large polytomies in the global phylogeny. For each of these nodes, we always included the five oldest genomes, the most recent genome sequence and five other immediate descendants that were randomly chosen. The remaining genomes were sampled by allocating an even number of sequences per epi-week while maintaining a dataset size of <1,000 genomes. For each epi-week, genomes were sampled evenly by country until either its allocation was exhausted or there were no remaining genomes available. This subsampled dataset was analysed in *BEAST 1.10* (*42*) using a GTR+G+F substitution model, with a strict molecular clock model using a non-informative continuous-time Markov chain (CTMC) prior (*43*) and a Skygrid coalescent tree prior (*44*) with 40 grid points, roughly corresponding to weeks between 1 October 2019 and 2 July 2020. In the analysis, monophyly constraints were used to ensure that the clades corresponding to the large-scale phylogenetic trees identified in the previous step were monophyletic. We combined four independent Markov Chain Monte Carlo (MCMC) chains that were each run for 40 million steps, discarding the first 4 million steps of each chain as burn-in and resampling states every 4000 steps. Convergence was assessed using *Tracer* (*45*).

Next, we applied a commonly used approach, recently implemented in *BEAST 1.10*, to convert branches of the large-scale phylogenetic trees from units of substitutions per site to time. This model takes the place of the nucleotide substitution model in a traditional Bayesian molecular clock dating analysis. Briefly, each branch of a maximum-likelihood tree is first scaled to represent the number of substitutions that occurred along that branch. Polytomies are resolved by inserting branches of length 0 substitutions. The likelihood of a branch b_i of length s_i substitutions is defined by a Poisson distribution with mean $t_i m$ where t_i is the length of the branch in years and m is the clock rate. The log-likelihood of the whole tree is then the sum of the log-likelihoods of each branch, which represents a fixed, strict-clock model and follows a commonly implemented approach for scaling phylogenies into time-calibrated trees (e.g. *46-48*).

Each large-scale phylogenetic tree was analysed under a strict clock model, with the clock rate fixed to the median estimate from the preliminary analysis (7.5×10^{-4} substitutions/site/year) and a Laplace root-height prior with mean equal to the median TMRCA estimate of the corresponding subtree in the preliminary analysis and scale equal to the average distance from the median. Trees were sampled using MCMC under the model described above with a Skygrid coalescent tree prior (*44*) using the same grid-points as in the preliminary analysis. A randomly resolved time-calibrated tree estimated in *TreeTime* (*49*) was used as the starting tree. To maintain a mapping between the topology in the estimated time-calibrated tree and the input genetic distance tree, we constrained the topologies such that any tree-move that broke a clade present in the input tree was rejected. The resulting MCMC chain, therefore,

only samples different polytomy resolutions and branch durations. This approach allowed us to incorporate uncertainty in the polytomy resolutions and branch durations into our molecular clock analysis.

We ran between 8 and 24 chains for 60 to 100 million MCMC steps for each large-scale phylogeny. Upon completion, we discarded 15 million states as burn-in from each chain. Chains that did not converge or pass the burn-in in less than 15 million states were re-run. Chains were combined and resampled every 100,000 states using custom R-scripts, leaving between 6808 and 17,020 posterior samples of each large-scale phylogenetic tree. Convergence was assessed using *Tracer* (*45*) and the R-package *coda* (*50*).

Identifying transmission lineages

We define a "UK transmission lineage" as two or more UK infection cases that (i) descend from a shared, single importation of the virus into the UK from elsewhere, (ii) are the result of subsequent local transmission within the UK, and (iii) were present in our virus genome sequence dataset. This concept is illustrated in Figure S1 and is distinct from a transmission cluster, an epidemiological term commonly referring to a group of cases that occur close to each other in space and time (e.g. in a hospital or care home). Therefore, a large UK transmission lineage may comprise many different individual transmission clusters.

[It is important to note that the "UK transmission lineage" definition employed here is distinct from the lineage/phylotype designations used by other parts of the COG-UK consortium and that are displayed at https://microreact.org/project/cogconsortium. Those latter designations (which have the format "UK…") are defined on the basis of shared sets of mutations, rather than shared descent from an inferred single introduction event.]

We can identify UK transmission lineages in the time-calibrated trees estimated in the previous step as clades of two or more genomes sampled in the UK. The TMRCA of all genome sequences in a UK transmission lineage represents the earliest transmission event in the lineage revealed by the data; however, it does not necessarily represent the first transmission event in the lineage as a whole, nor does it represent the importation date (i.e. the arrival date of the index patient in the UK). The relationship between the TMRCA of a UK transmission lineage in our dataset and the importation date is illustrated in Figure S2. Specifically, if the transmission lineage is well-sampled, then the TMRCA represents the date of the first transmission event in the lineage (TMRCA A in Fig. S2, UK transmission lineage 2 in Fig. S1). However, if the transmission lineage is sparsely sampled then the TMRCA may represent a later transmission event (TMRCA B in Fig. S2, UK transmission lineage 1 in Fig. S1). The "importation date" of each UK transmission lineage is the date that an infected inbound traveller entered the UK.

We used a two-state asymmetric discrete trait analysis (DTA) model (*51*) implemented in *BEAST 1.10* (*42*) to infer ancestral node locations (UK, non-UK) on empirical distributions of 500 time-calibrated trees sampled from each of the posterior tree distributions estimated above. Additionally, we used a robust counting approach (*52*) to estimate the expected number of location state transitions into and out of the UK. For each large-scale subtree, we combined 2 independent chains, each run for 5 million MCMC steps and sampled every 4500 states. The first 10% of each run was discarded as burn-in, resulting in 2000 trees with estimates of the ancestral location for each internal node. Finally, *TreeAnnotator 1.10* was used to generate maximum clade credibility (MCC) trees for each subtree, where each internal node is assigned a posterior probability of representing a transmission event in the UK.

Transmission lineages were identified by first labelling each node in the MCC trees as UK or non-UK and then initiating a depth-first search from each UK genome in the MCC trees. All nodes with a median age after 23 January 2020 and posterior probability >0.5 of the ancestral location being located in the UK were labelled as UK nodes. The depth-first search is continued until a non-UK node is encountered or there are no nodes left to explore. At the end of the depth-first search, all nodes visited by the search are added to the same (arbitrarily named) UK transmission lineage. If only one tip is visited, the UK genome at the tip is marked as a singleton. This procedure is repeated iteratively until every UK genome in the tree has been assigned a transmission lineage or marked as a singleton. The same procedure was repeated on each of the 2000 posterior trees, for each subtree, from the DTA analyses described above to examine statistical uncertainty in the number, size and duration of UK transmission lineages and their TMRCA distribution.

Our methodology is likely to underestimate the true number of transmission lineages and singletons. Since only a small fraction of UK infections have been sequenced (Fig. 1A), many lineages will have gone undetected. Furthermore, the power to detect a transmission lineage in our sparsely-sampled dataset is dependent on its size (i.e.

the frequency of a lineage being sampled from a small random sample of infections), making it more likely for larger lineages to be detected. The low sampling fraction means that some singletons detected in our dataset likely belong to observed and unobserved UK transmission lineages. Nonetheless, the true number of singletons (importations not resulting in onward transmission) is likely to be significantly more than our estimate, because their small size makes them difficult to detect with a low sampling fraction. Finally, under-sampling of genomes from other countries could result in mistaken aggregation of separate importations, reducing the number of detected lineages. This mistaken aggregation will result in larger, older lineages being estimated. This was the motivation for placing an age limit on UK nodes in the tree. We chose 23 January 2020 as the oldest possible date for a transmission event in the UK as this represents the date that the first patient who tested positive for SARS-CoV-2 in the UK entered the country (*53*) (tested positive on 30 January 2020). Although older importations into the UK could in theory be possible, if they had resulted in large autochthonous outbreaks, we would have observed this in both epidemiological and genomic data.

We estimate a median of 2968 (95% HPD 2829-3103) non-UK to UK state transitions and an additional 1468 (95% HPD 1362-1566) UK to non-UK state transitions (Fig. S3, Table S1) using the robust counting approach (*52*). The former slightly exceeds the sum of transmission lineages and singletons as identified on the MCC tree (=2918) and across the 2000 posterior trees (median=2829, 95% HPD=2773-3048; Table S1). This result is expected, since multiple location state changes along long branches contribute to the total number of state transitions, but do not add to the total number of UK transmission lineages or singletons. The largest number of location state transitions occur on the B.1.1 phylogeny, with the fewest occurring on lineage A, which are the largest and smallest of the subtrees, respectively. Proportional to the number of tips, fewer state changes are inferred on the two B.1 phylogenies than other subtrees, while the number of UK to non-UK transitions on the B phylogeny exceeds that inferred on other lineages. We caution that UK to non-UK transitions are likely to be underestimated because of under-sampling in other countries and differences in the proportion of infections sequenced between countries.

The transmission lineage size distribution from the MCC trees falls within the HPD interval taken across the 2000 posterior trees (Fig. 1B). Although the sizes of the largest transmission lineages vary substantially across posterior trees, the cumulative size distributions are similar across all trees (Fig. 1B, inset). Similarly, the transmission lineage duration distribution on the MCC trees falls within the variation of the HPD interval taken across the 2000 posterior trees (Fig. S8).

We used the Jaccard index to compare the classification of UK genome sequences into transmission lineages and singletons between posterior trees and the MCC trees. Figure S13A shows the mean, median and 95% HPD interval of the Jaccard index for each posterior tree compared to the 1999 other posterior trees, across all subtrees. While most Jaccard indices are between 0.7 and 0.8, there is a noteworthy minority of trees with mean Jaccard indices <0.6 (n=100). Comparing the 2000 posterior trees to the classification on the MCC tree (Fig. S13B), results in a similar distribution of Jaccard indices, with most indices between 0.7 and 0.8 and minorities below 0.6 and above 0.8 (n=68, n=170 respectively).

We undertook a similar analysis of the sensitivity to phylogenetic uncertainty of the distribution of UK transmission lineage TMRCAs. We computed the median and 95% HPD interval of the number of transmission lineage TMRCAs on each date across the 2000 sampled posterior trees. Figure S14 shows that the TMRCA distribution computed from the MCC trees falls within the comparatively narrow HPD limits, and oscillates around the median estimate for each date.

UK epidemiological data

The number of reported COVID-19 cases in the UK, by specimen date, were downloaded from https://coronavirus.data.gov.uk/cases (date accessed: 1 September 2020). The number of reported COVID-19 cases for each Upper Tier Local Authority (UTLA) in England, Local Health Board (LHB) in Wales and regional NHS Board in Scotland, by specimen date, were downloaded from https://coronavirus.data.gov.uk/downloads/csv/coronavirus-cases_latest.csv, http://www2.nphs.wales.nhs.uk:8080/CommunitySurveillanceDocs.nsf (file: *"Rapid COVID-19 surveillance data.xlsx"*) and https://github.com/DataScienceScotland/COVID-19-Management-Information (file: *"COVID19 - Daily Management Information - Scottish Health Boards - Cumulative cases.csv"*), respectively (date accessed: 15 October 2020).

To enable comparison of case and sequence data, locations used to report case data were combined to correspond to those used for sequence data and vice-versa (see the *Geographical metadata* section). Northern Ireland was not included due to inconsistencies between the locations used for case and sequence data reporting that could not be easily resolved.

Global deaths due to COVID-19

The cumulative number of daily COVID-19 deaths for each country were downloaded from the JHU CSSE COVID-19 Database (date accessed: 19 August 2020) (*54*). We removed data pertaining to cruise ships, and aggregated data to the country level where data were reported for subnational divisions (e.g. Australia). For countries with overseas territories included in the dataset (e.g. United Kingdom), we excluded the cumulative death counts in those overseas territories. For each country we computed a time series of the daily number of deaths by taking the difference in the cumulative number on consecutive days. When this difference was negative, for example when corrections in the cumulative number were not propagated backwards, we set the value to zero. A relevant outlier in these time series is the addition of 1290 deaths in China on 17 April 2020, while on the days before and after no deaths were recorded. To account for these deaths, we uniformly distributed these deaths over the previous 85 days described by the epidemiological data.

Population data

Country population size estimates were downloaded from the UN Department of Economic and Social Affairs website (https://population.un.org/wpp/Download/Standard/Population/), using the *Medium* fertility projection for 2020 (*55*).

Travel and mobility data

To investigate temporal trends in SARS-CoV-2 importation intensity we sought information on the number of travellers entering the UK from each other country for the period from 1 January to 30 April 2020. Incoming travellers comprised both British nationals and resident and visiting citizens of other countries. Estimates were obtained by combining multiple data sources. First, the UK Home Office has provided statistics that describe the number of inbound travellers arriving in the UK by air on each day during this period (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/887655/statistics-relating-to-covid-19-and-the-immigration-system-tables-may-2020-arrivals.ods). This data set provides the daily number of incoming air passengers but not their source country. Second, we obtained the number of tickets sold for inbound flight journeys to the UK along with their origin location from the IATA (for passengers that transfer, the source location is the country from where the whole journey started). We used these numbers to calculate the percentage of arrivals from each country on a monthly basis from January to April 2020. We multiplied the monthly distribution of source destination by the total number of air passenger arrivals in the UK each day to estimate the number of arrivals from each country. Third, we augmented the above air passenger numbers with estimated numbers of incoming travellers arriving per day by short-sea ferry and through the Channel Tunnel (French: *Le tunnel sous la Manche*). Numbers of short-sea ferry passengers from France, Netherlands and the Republic of Ireland were estimated from monthly statistics obtained from the UK Department of Transport (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908445/spas0107.ods). Within that data set, values are provided for the Republic of Ireland and for "Other EU countries". The latter total was broken down by country using data from 2019 showing that 72.7% of UK short-sea journeys are with France, 13.6% with the Republic of Ireland, 10.4% with the Netherlands, and 3.3% with other countries (https://www.gov.uk/government/statistics/sea-passenger-statistics-2019-short-sea-routes). Eurotunnel Shuttle vehicle movements from France were obtained from publicly available monthly records (https://www.eurotunnelfreight.com/uk/2020/02/shuttle-traffic-for-january-2020). In the absence of other information we assumed (i) inbound and outbound vehicle movements via the Eurotunnel Shuttle services were equally frequent and (ii) one passenger per truck and 1.5 passengers per passenger vehicle. Inbound Eurostar rail passenger numbers from France and Belgium were estimated from available data and adjusted as far as possible for post-pandemic reduction in travel. Specifically, ~2m passengers travelled by Eurostar in the first quarter of 2020 (https://www.breakingtravelnews.com/news/article/eurostar-passenger-count-slips-by-a-fifth-in-early-2020). Monthly Eurostar passenger numbers were then calculated by assuming (i) inbound and outbound journeys were equally frequent, (ii) two thirds of inbound Eurostar journeys originated in France and one third in Belgium, in approximate proportion to the ratio of services, and (iii) the proportional decrease in Eurostar travel volumes during March and April 2020 was equal to that observed for vehicle movements via the Eurotunnel Shuttle. Our estimates do not incorporate estimates of movements across the land border between the UK and the Republic of Ireland. This

is unlikely to be problematic as the numbers of infections in the Republic of Ireland was relatively low compared to other potential source countries during the time period of interest.

Epidemiological model

We sought an estimate of the number of individuals in each source country who are (i) infected with SARS-CoV-2 and (ii) able to travel to the UK and initiate a transmission chain. In what follows we refer to these individuals as the "potential initiators of a transmission lineage" (PITL). We conservatively assumed that symptomatic individuals cannot initiate a transmission chain in the UK, either through being prevented from travelling or perfect isolation on arrival. Thus, our estimates of daily SARS-CoV-2 prevalence includes only pre-symptomatic and asymptomatic individuals. Asymptomatic individuals are counted among the PITL as those capable of initiating a transmission lineage at any time while they are still infectious. Figure S16 illustrates the ways in which individuals are counted towards the daily PITL and their potential disease outcomes.

We estimated the daily number of PITL by back-extrapolating the time series of daily numbers of deaths due to COVID-19 in each source country. COVID-19 deaths were used instead of confirmed cases, as we are primarily interested in temporal dynamics rather than absolute values, and death counts are believed to be less sensitive to changes in case definition, reporting delays and differences in the level of surveillance among countries and regions. Estimates of the latent period (infection to becoming infectious), incubation period (infection to onset of symptoms), the infectious duration, and the time between symptom onset and death (in fatal cases) were used to estimate the number of infected individuals who would go on to die from COVID-19, in each stage of the disease, on each day (Fig. S16). We then estimated the total number of infected individuals on each day by multiplying with the reciprocal of the infection fatality rate (IFR).

Estimates of the periods defined above were taken from peer-reviewed sources. Specifically, we assumed that the time from acquiring an infection to becoming infectious is 3 days (*56*) and the time to symptom onset 5 days (2 days after becoming infectious) (*57*). The infectious period for patients who recover from the disease was assumed to end 5 days after symptom onset (*56*) while those who die from the disease are assumed to do so 18 days after the onset of symptoms (*58*). Given the large numbers of deaths we expect that variation in these timings among individuals will be averaged out and is not considered. We further assumed an asymptomatic proportion of 31% (*59*) and an IFR of 1%, which is broadly consistent with those found in the literature for China, France, and passengers aboard the Diamond Princess (*58, 60*, *61*). These values correspond to our study period, the spring epidemic of COVID-19; more recent estimates of IFR may vary due to changing treatment regimes and other factors. To examine the sensitivity of our results to the asymptomatic proportion we re-ran our analysis with proportions of 0.18 and 0.78 (the range of published estimates; *62, 63*), and found that our results were robust over this range (data not shown). As our main results are fully determined by temporal trends in EII and not absolute numbers, they are invariant to the value of the IFR and we did not perform a sensitivity analysis on it. We did not account for changing levels of infectivity among individuals over the course of their infection.

Using the time series of deaths extracted from the JHU CSSE COVID-19 Database (*54*), as described above, we obtained estimates of the daily number of PITL in 183 countries from 31 December 2019 to 26 July 2020.

Estimated importation intensity

The daily "estimated importation intensity" (EII) of a country is defined as the product of the proportion of individuals in that country who make up the PITL (as described above) on each day, and the number of individuals who travelled from that country to the UK on that day. The former is estimated by dividing our estimate of the total number of individuals who could potentially initiate a lineage (for each day) by the total population of the country (see the *Epidemiological model* section). The latter corresponds to the total number of arrivals by air, ferry, and rail on that day (see the *Travel and mobility data* section). To assist in the subsequent use of the EII, we aggregated all countries with low PITL estimates into a single "other" category. The aggregated countries are those that comprised less than 1% of the cumulative total number of cases as of 1 May 2020 (excluding the UK). This left 53 primary source locations. Maximum EII (**Fig. S17**) was highest for Spain, (which experienced a large, early epidemic that peaked before inbound passenger numbers declined), followed by France (whose later epidemic peak coincided with high but declining international travel).

Importation lag model

We modelled the TMRCA of an observed transmission lineage (the data observation) as the arrival date of the index patient (of that transmission cluster) in the UK, G, plus a lag time, L, until the first transmission event in the lineage revealed by the data. Given the probability that an importation occurs on day g, $f_G(g)$, and the probability of a lag time of j days, $f_L(j)$, the probability of a TMRCA occurring on day k is v_k, is defined by

$$\hat{v}_k = \sum_g f_G(g) f_l(k-g)$$

with $v = \hat{v} / |\hat{v}|$. TMRCAs and importation dates are assumed to be independent, so the likelihood for all transmission lineages is the product of the corresponding v_k for each lineage.

This model does not account for incomplete sampling of patients from UK transmission lineages. It is likely that the TMRCA of a small transmission lineage is more recent than the first transmission event after the importation and this issue is potentially further exacerbated by non-random sampling of genome sequences from patients in the lineage (*64*). We therefore expect shorter lag times for bigger transmission lineages. To account for this size-dependence, we model the average importation lag as a function of lineage size. The functional form of this is given by the equation $\alpha + \beta / n$, where α corresponds to the minimal average lag time expected under complete sampling of the lineage and β accounts for the increase in lag time as a smaller proportion of sequences are included in the lineage.

We applied this model to the TMRCA estimates of individual transmission lineages and their sizes as obtained from the MCC trees (see the *Identifying transmission lineages* section). Values for α and β were found by numerically optimising the likelihood function using random draws from an exponential distribution as initial parameter values. The optimisation procedure was repeated several times to ensure that the algorithm did not become stuck in a local optimum. We further tested whether lineage size affects the importation lag through a likelihood ratio test (LRT) comparing the above model to a nested model without size dependence ($\beta = 0$) and found that the size-dependent model is preferred ($\chi^2_1 = 137.22$, $p < 0.001$). The maximum likelihood estimates for α and β are 0.72 and 28.91 (Fig. S18), respectively.

Travel advice in the UK

The travel advice issued by the Foreign and Commonwealth Office (FCO) of the United Kingdom pertaining to countries and regions affected by COVID-19 was primarily made available through their website (*FCO Travel advice: coronavirus (COVID-19)* at https://www.gov.uk/guidance/travel-advice-novel-coronavirus). The number of COVID-19 cases in the UK was available via the government website (*Coronavirus cases in the UK* at https://www.gov.uk/guidance/coronavirus-covid-19-information-for-the-public). Travel advice was also echoed by various news outlets and other information platforms, such as the Public Health Scotland/NHS Scotland *Fit for Travel* website (https://www.fitfortravel.nhs.uk/). We collected this information by mining archived FCO sites, manually retrieving HTML files corresponding to updates to the URLs provided above and available at the Internet Archive (https://archive.org/). Files were obtained and examined for all dates when changes to the URL were published (18 updates were published in total between 4 February and 23 May 2020). Furthermore, we compared this advice with the *Fit for Travel* online resource, collected through a similar approach. Where information was insufficient or unclear, we complemented it with data from news outlets to clarify travel advice, which was the case before February 4, when there was no official travel advice (only notifications for novel coronavirus). We collated all the travel advice information into a single standardised table containing types of advice, dates of implementation and countries or geographic regions covered by the advice. The types of advice included both suggestions against specific types of travel versus all but non-essential travel and the recommended period of self-isolation upon return from specific destinations. All of the changes in travel advice were between February 6 and March 23, when specific self-isolation recommendations applied to the general population and not just returning travellers. A summary of the main changes in the UK travel advice across time (in particular, dates when advice for new countries were issued) is presented in Table S5.

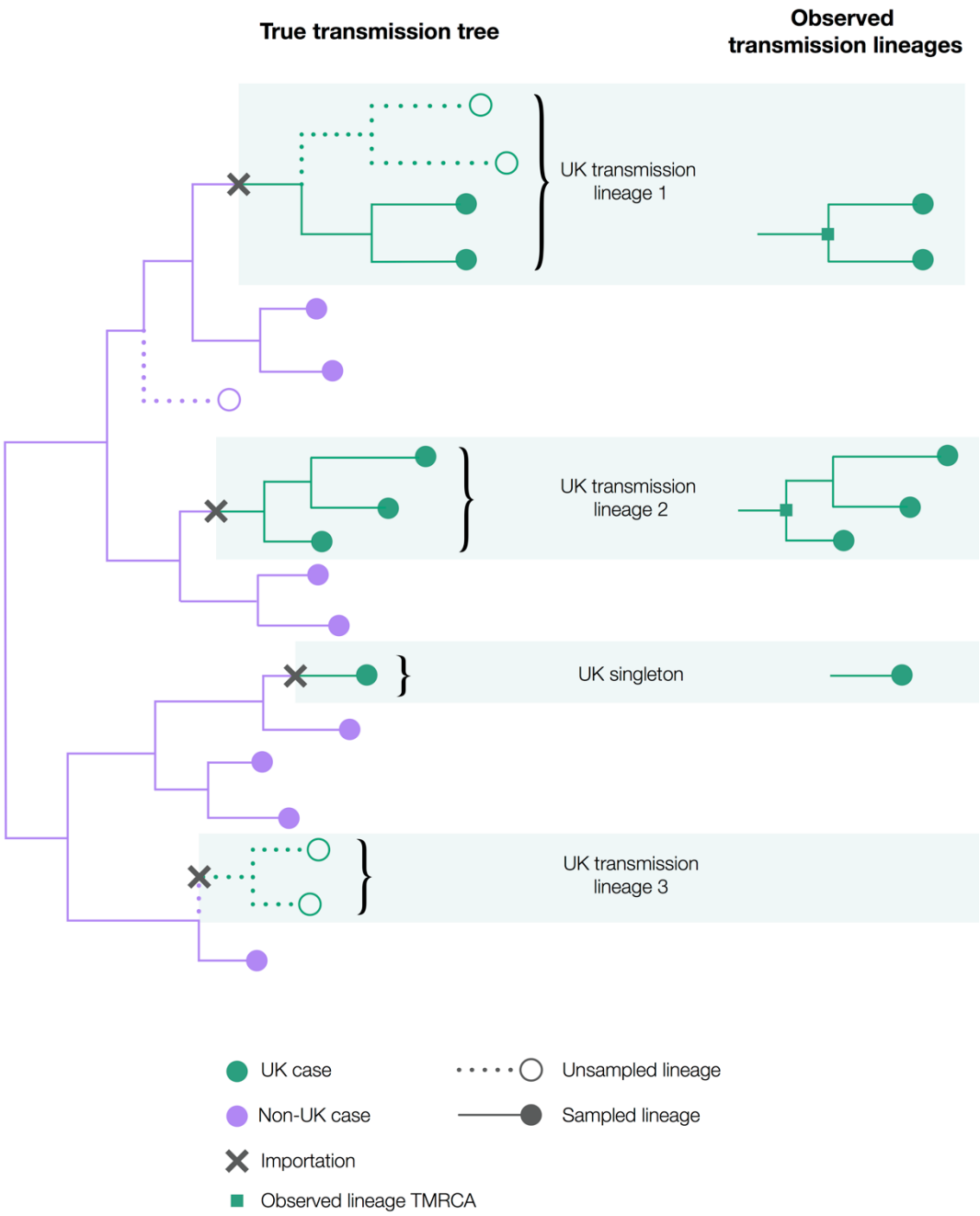

Fig. S1. *Hypothetical scenario illustrating the definition and international context of UK transmission lineages. Note that only half of cases in UK transmission lineage 1 are observed and that UK transmission lineage 3 is not observed at all. Singletons do not contribute to onward transmission within the UK and are not classified here as UK transmission lineages.*

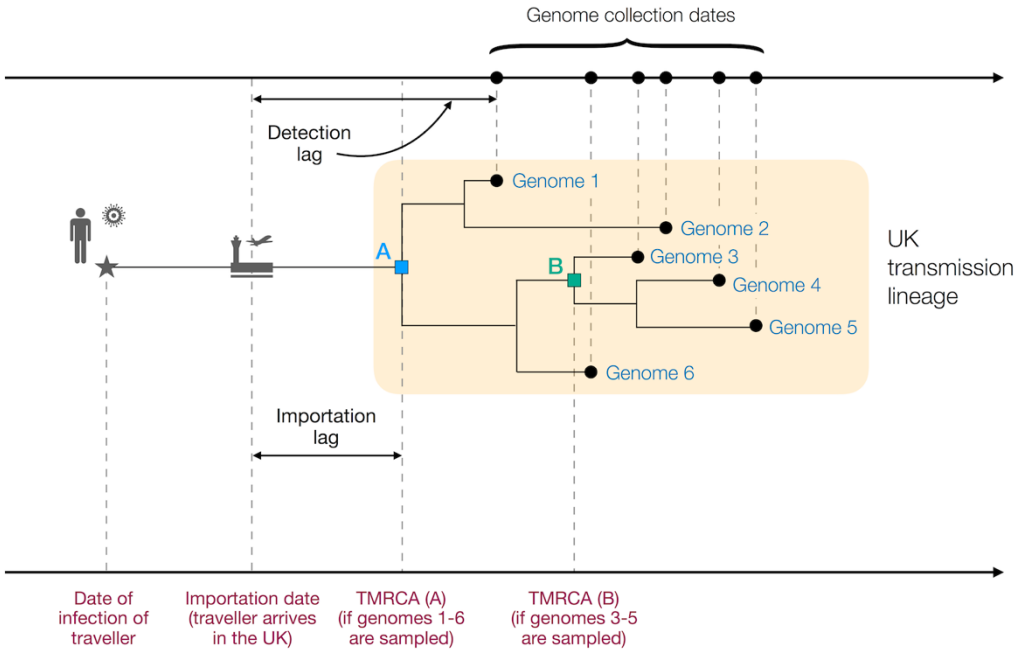

Fig. S2. *Figurative illustration of a UK transmission lineage detected through genome sampling. To be detected, a UK transmission lineage must contain two or more sampled genomes (see* ***Figure S1****). The terms TMRCA, detection lag, and importation lag can be understood with reference to this figure. The lineage TMRCA is sample dependent, for example, TMRCA A is observed if genomes 1–6 are sampled and TMRCA B is observed if only genomes 3–5 are sampled.*

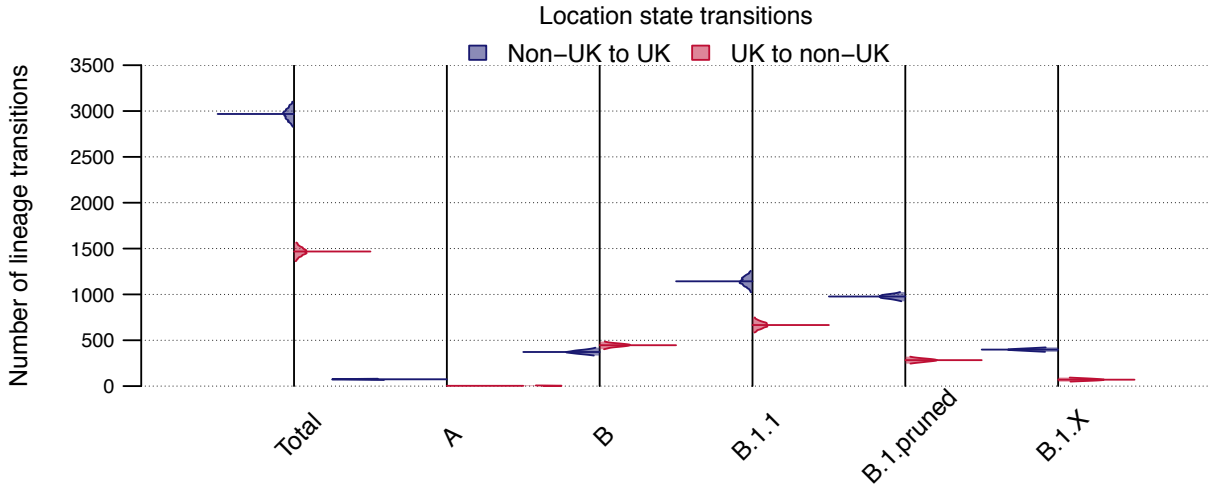

Fig. S3. *Number of location state transitions between the binary phylogenetic traits UK/non-UK detected by the robust counting approach implemented in BEAST 1.10. Non-UK to UK=blue, UK to non-UK=red. Posterior distributions are truncated at their 95% HPD interval limits and the horizontal lines indicate median estimates.*

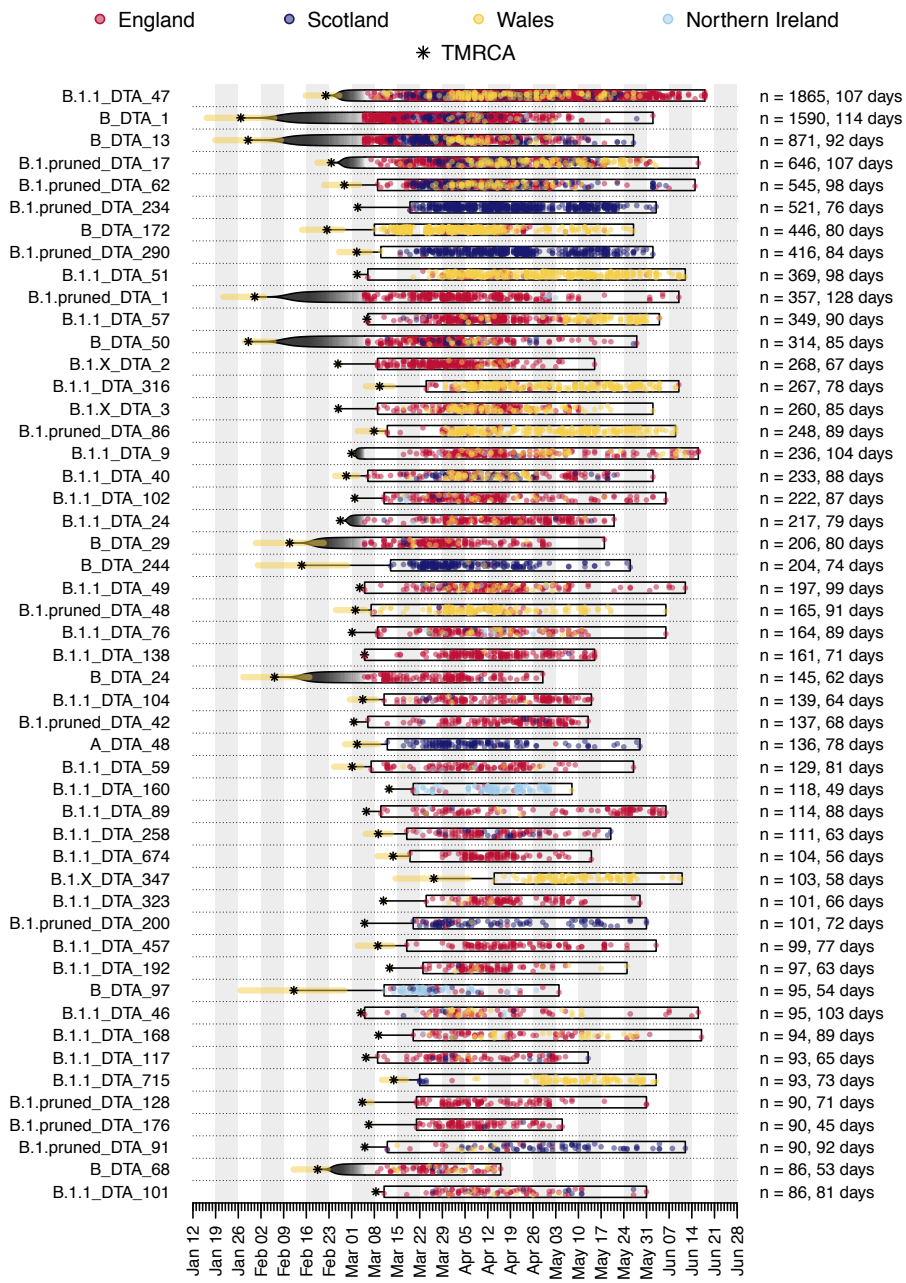

Fig. S4. *Illustration of the time course of the 50 largest UK transmission lineages in our dataset. Each row is a transmission lineage. Dots are genome sampling times (coloured by sampling location) and boxes show the range of sampling times for each transmission lineage (sampling duration). Asterisks show the median TMRCA of each lineage and the yellow bars show the 95% HPD of each TMRCA. On the right, n indicates the number of UK genomes in the lineage and the duration of lineage detection (time between the lineage's oldest and most recent genomes). Sampling times of the first 500 SARS-CoV-2 genomes collected in the UK have been obscured.*

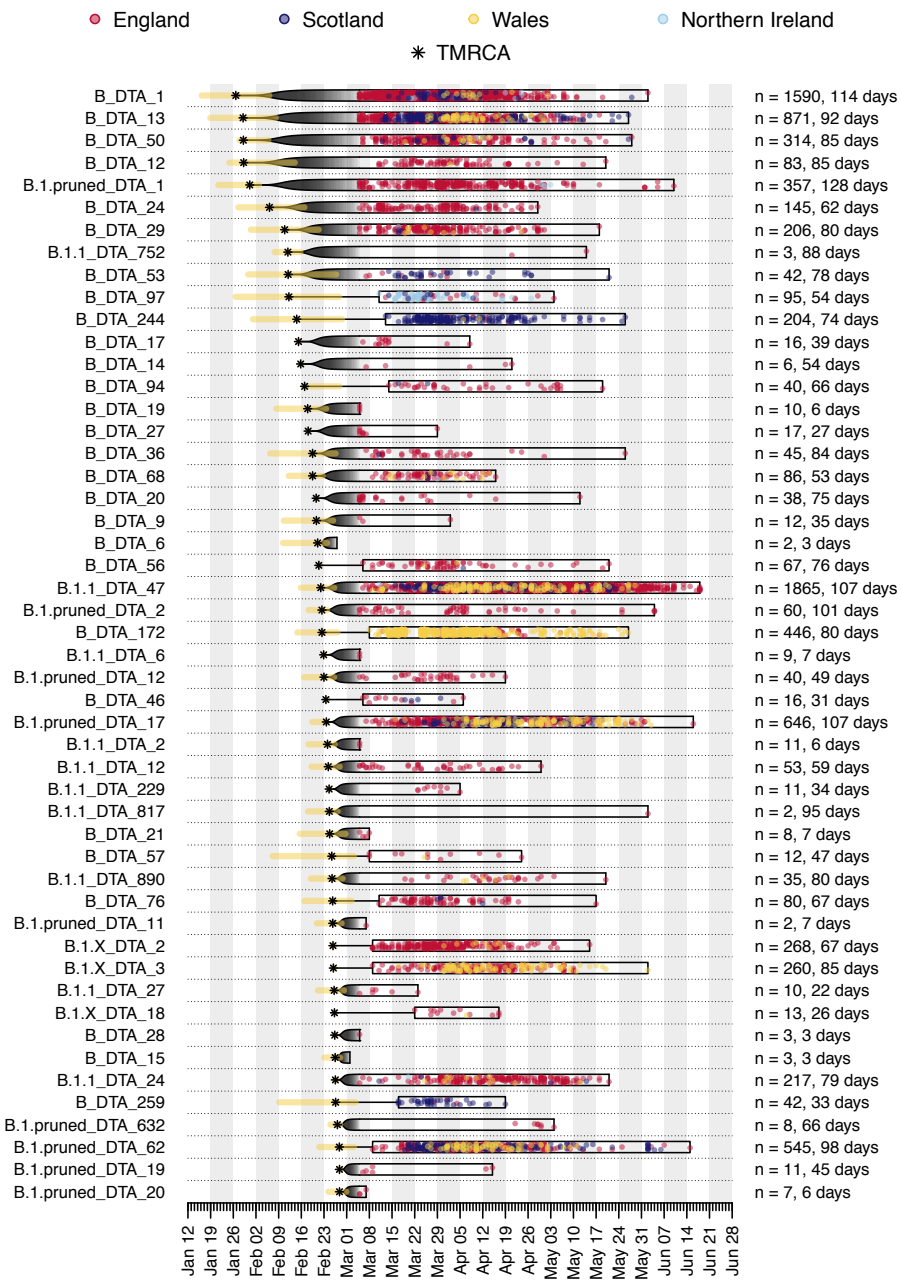

Fig. S5. *Illustration of the time course of the 50 earliest UK transmission lineages in our dataset. See Figure S4 caption for details.*

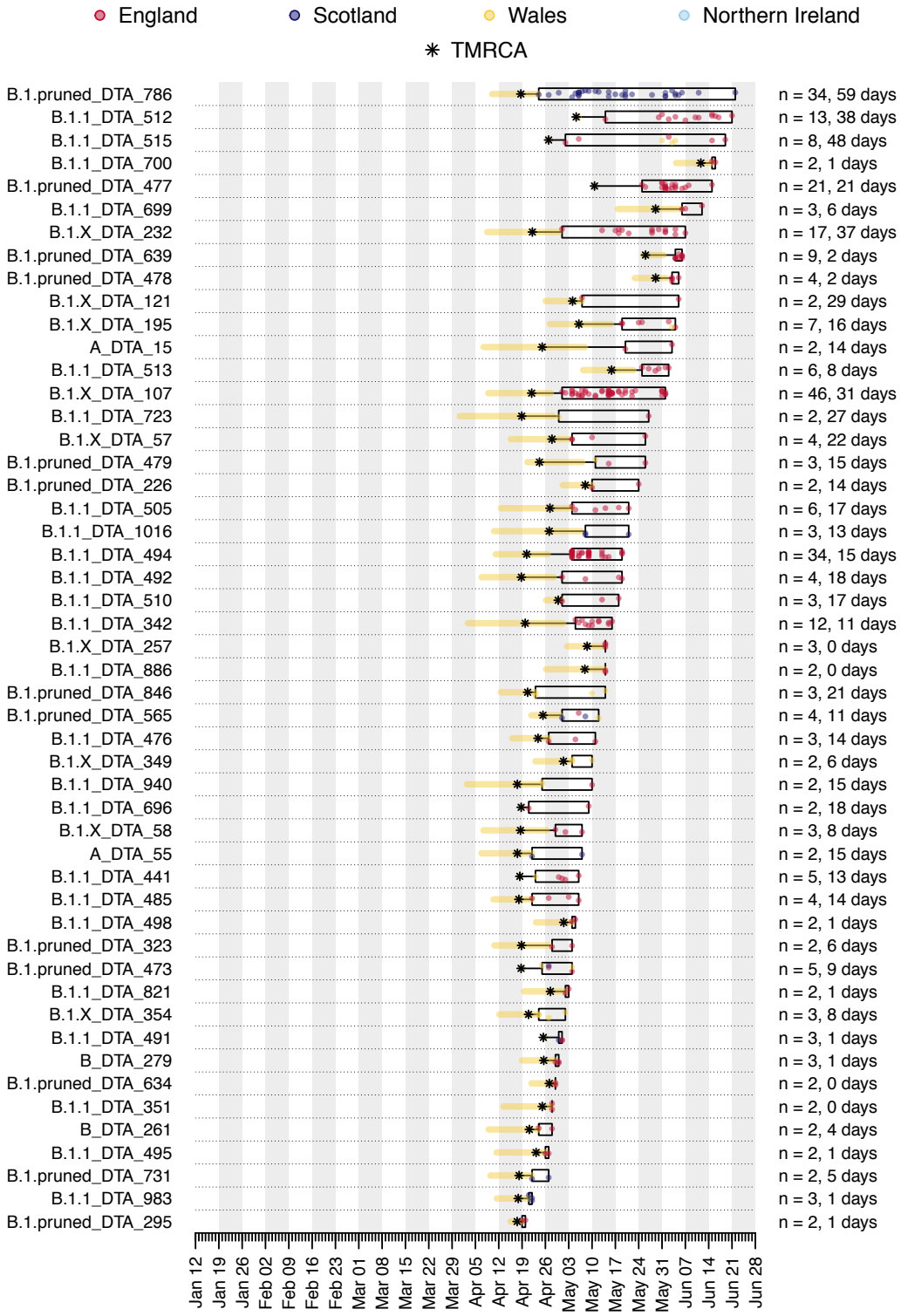

Fig. S6. *Illustration of the time course of the 50 most recent (by TMRCA) UK transmission lineages in our dataset. See Figure S4 caption for details.*

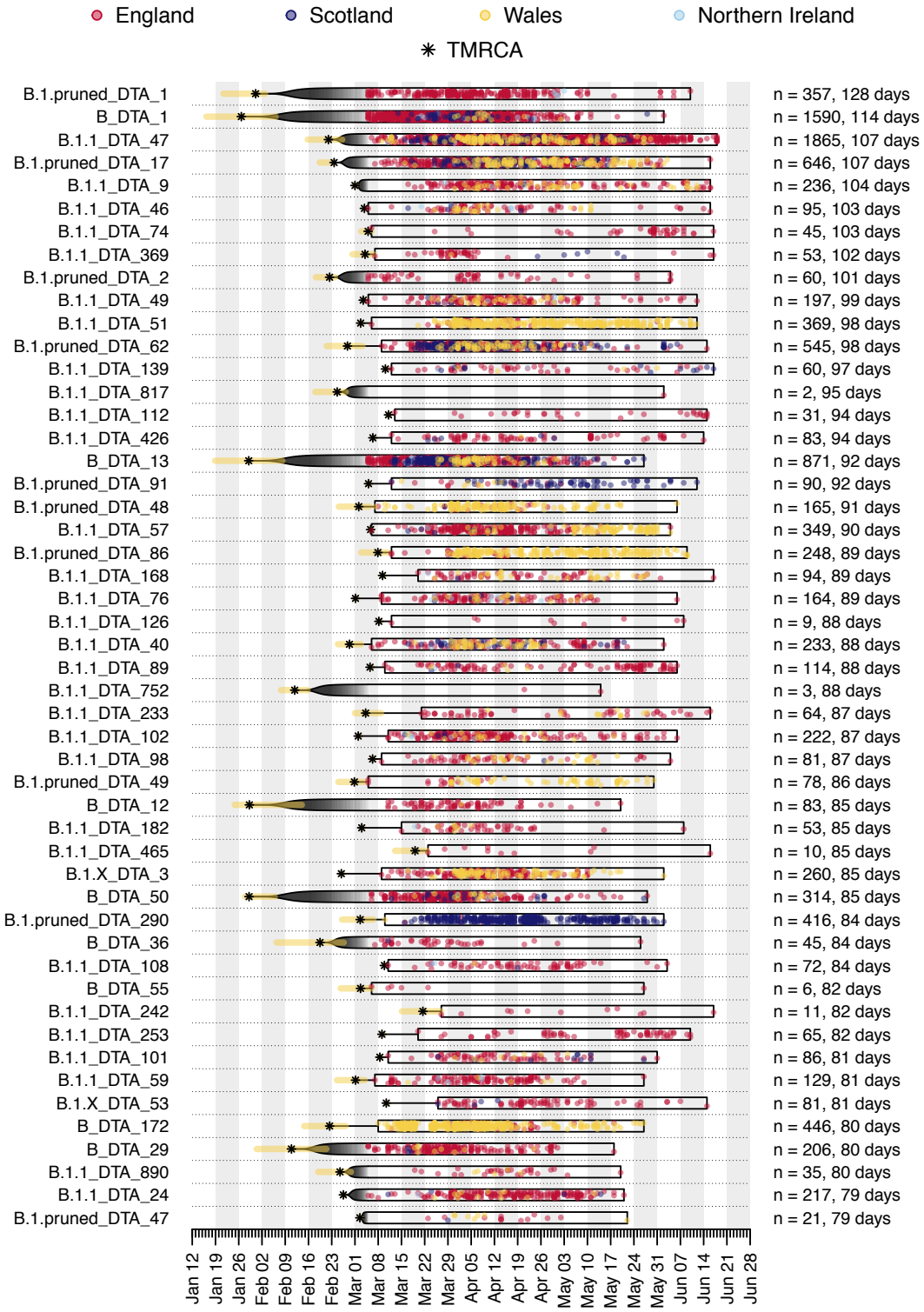

Fig. S7. *Illustration of the time course of the 50 UK transmission lineages with the longest sampling duration in our dataset. See Figure S4 caption for details.*

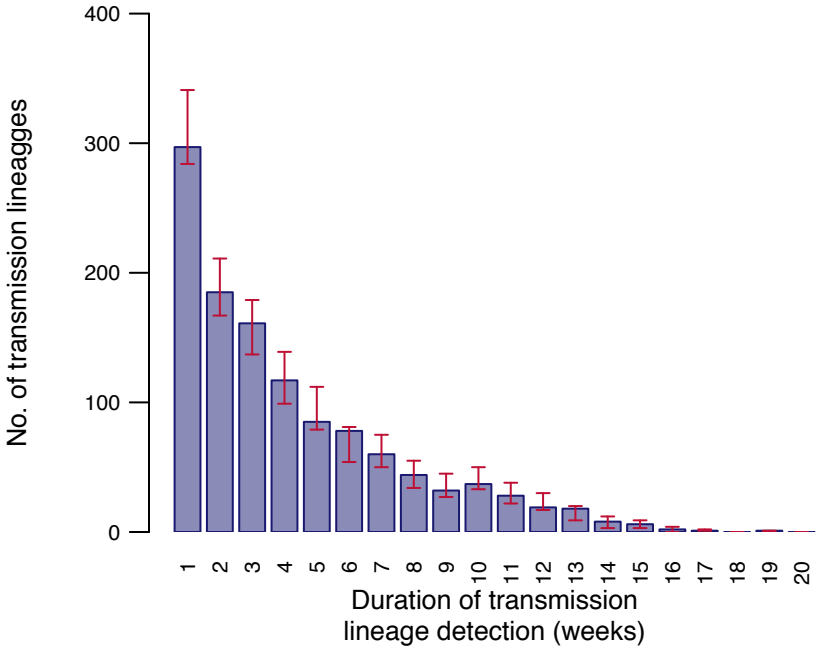

Fig. S8. *Distribution of UK transmission lineage sampling durations, aggregated by week. Blue bars show the number of transmission lineages that were observed over different durations in the MCC tree. Red bars show 95% HPD intervals for these numbers across the posterior tree distribution.*

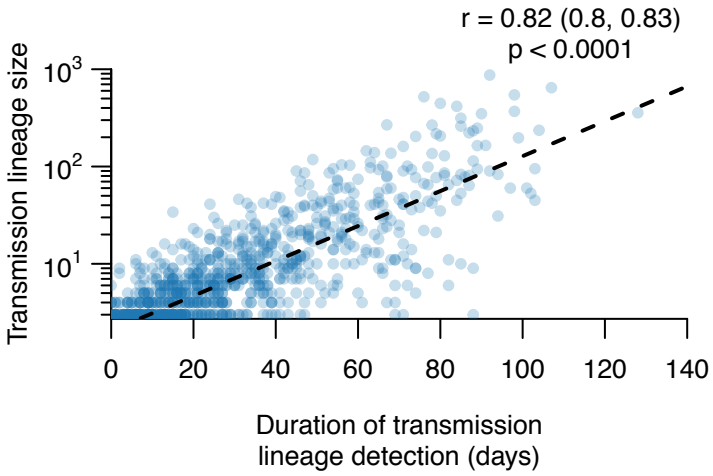

Fig. S9. *Scatterplot showing the strong relationship between UK transmission lineage size and sampling duration. The Pearson correlation coefficient, 95% CI and p-value are shown.*

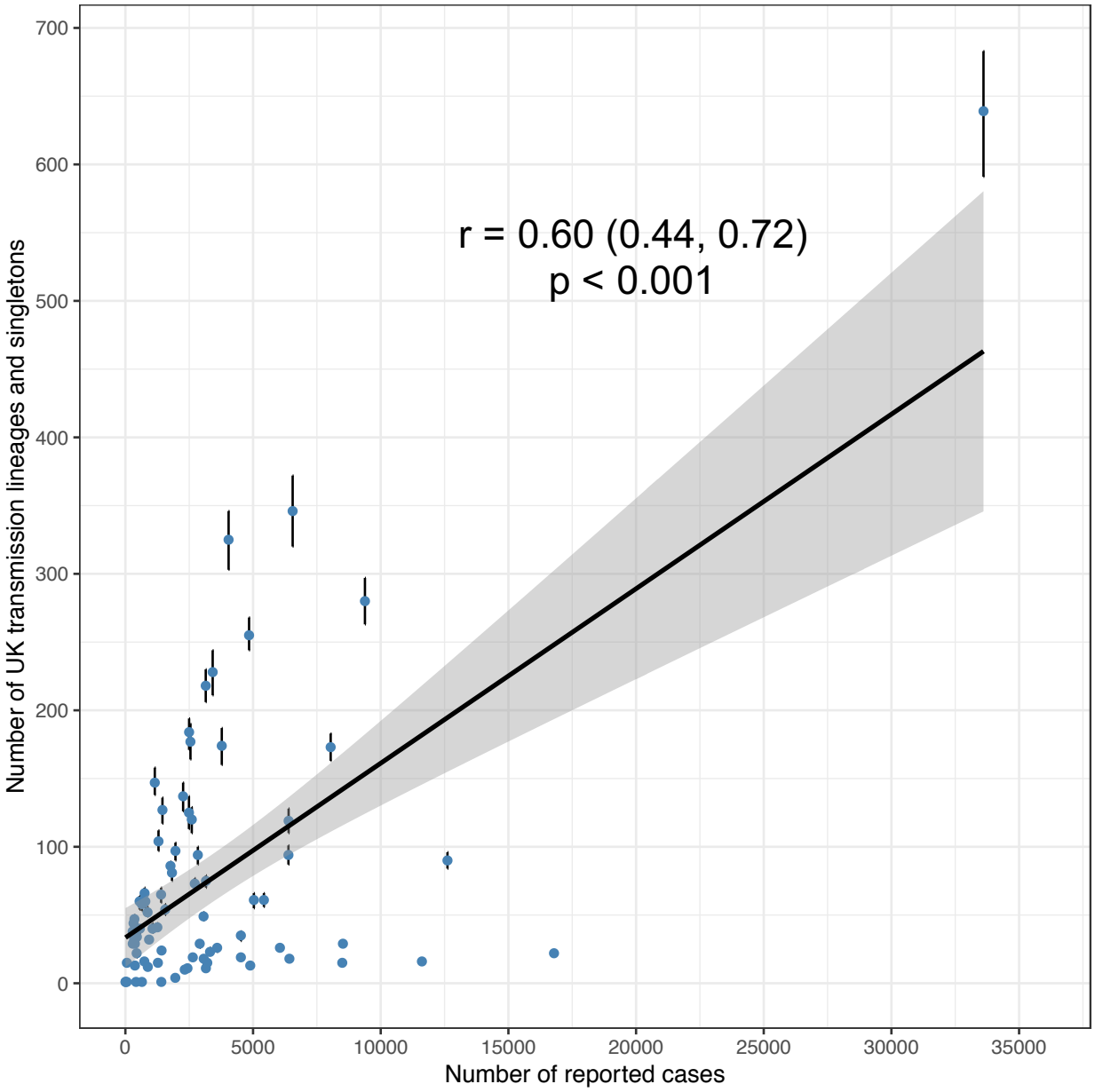

Fig. S10. *Scatterplot showing, for each geographic region, the relationship between the number of reported cases up to 26th June 2020 in that region and number of distinct UK transmission lineages and singletons detected in the region. Points show median estimates and error bars 95% HPDs from the posterior distribution of trees.*

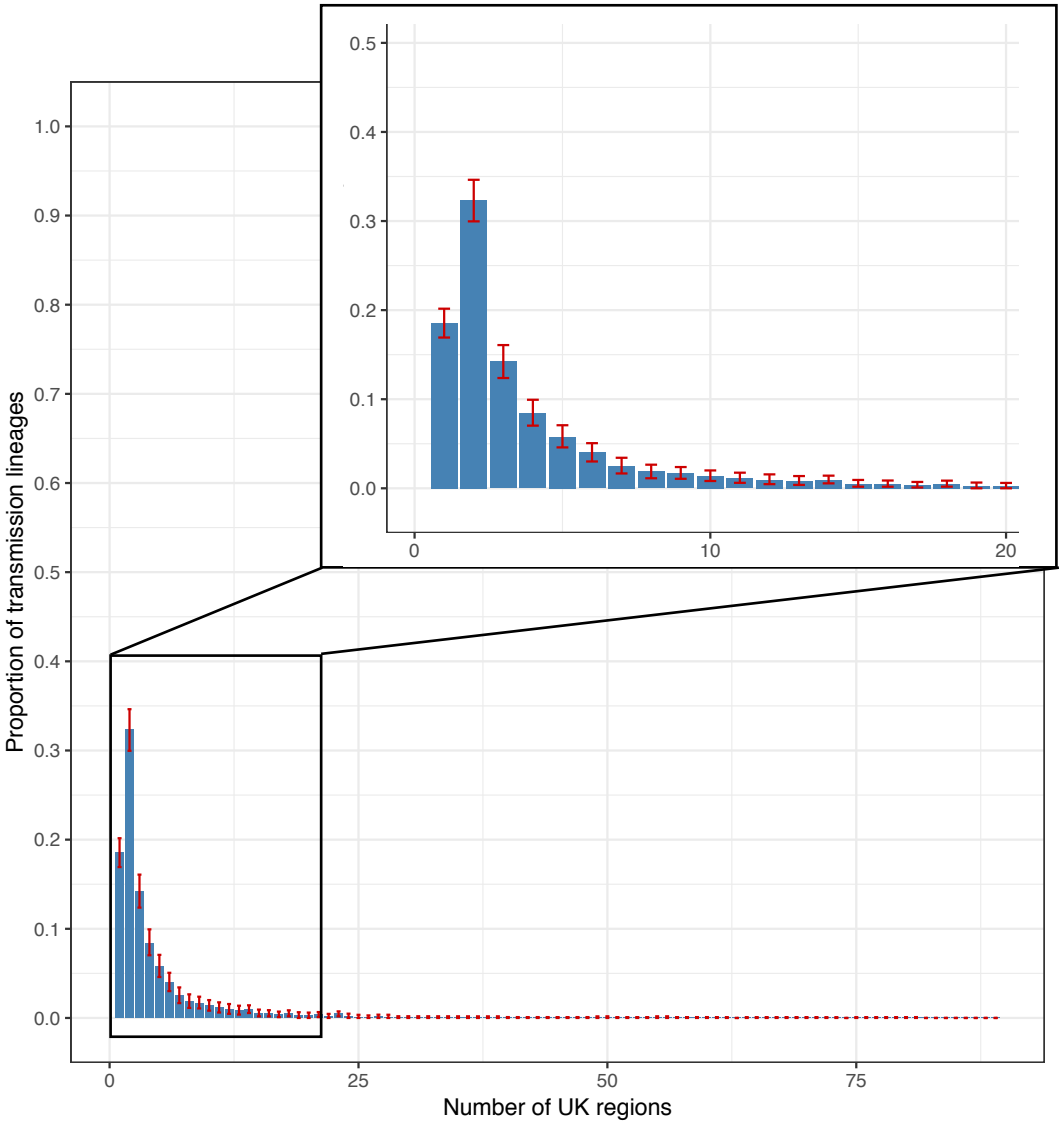

Fig. S11. *Geographic range size distribution of UK transmission lineages. Plot shows the distribution of the number of geographic regions in which each UK transmission lineage was sampled. Bars represent median proportions across the posterior distribution of trees and red bars show the 95% HPD intervals.*

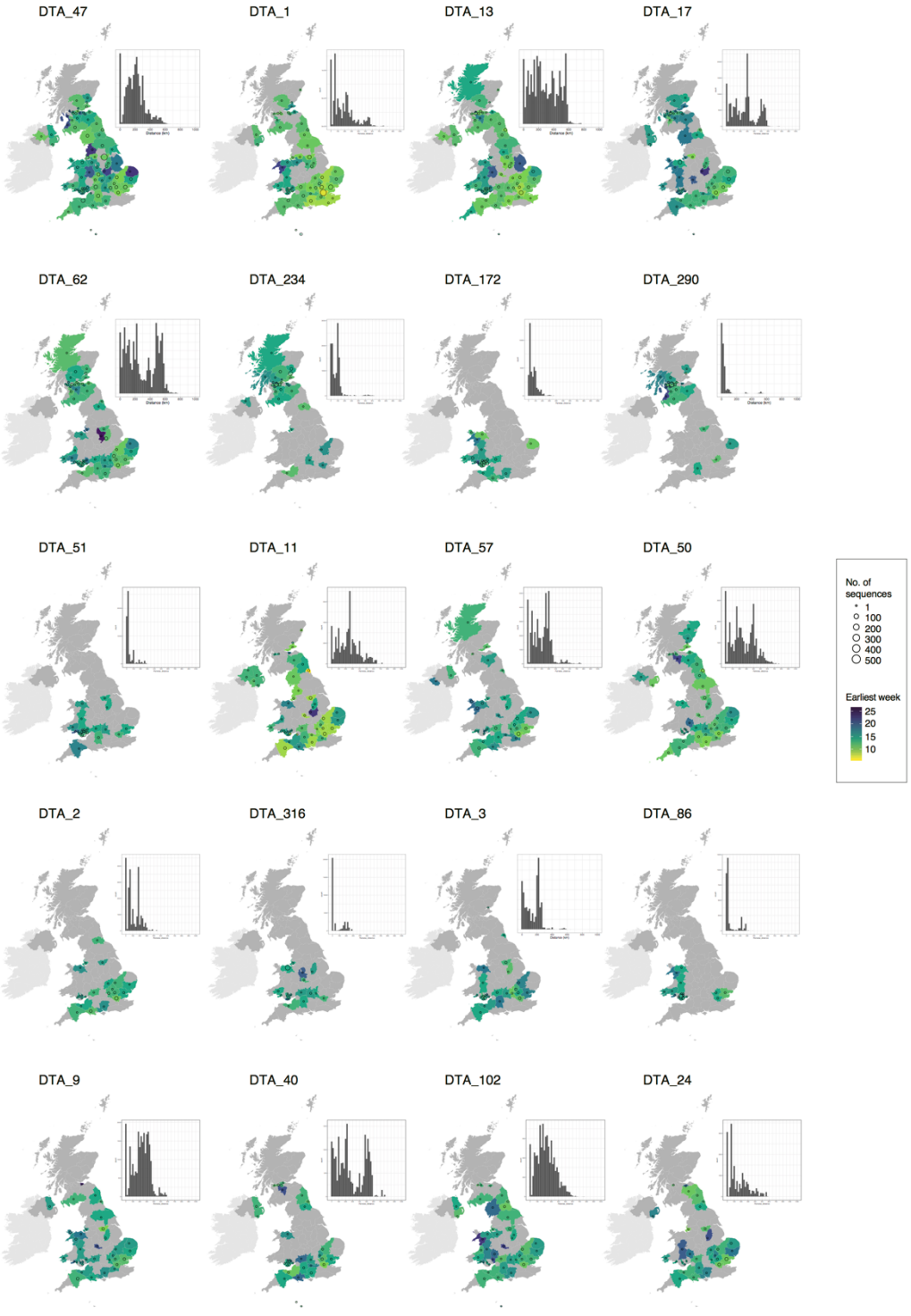

Fig. S12. *Spatial distribution of the twenty largest UK transmission lineages. Colours represent the week of the first detected genome in the transmission lineage in each location. Circles show the number of sampled genomes per location. Insets show the distribution of geographic distances for all sequence pairs within the lineage.*

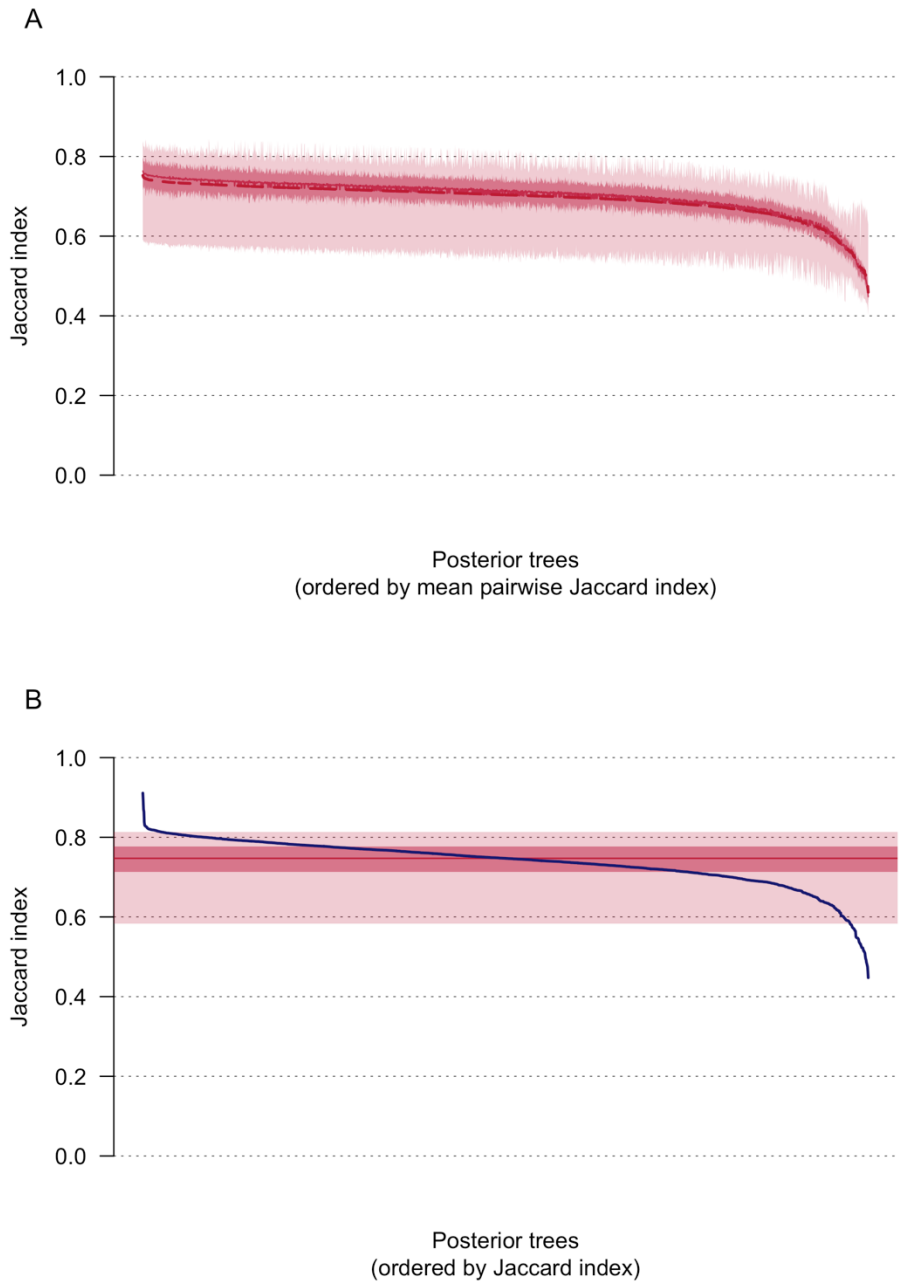

Fig. S13. (A) *Median (solid line) and mean (dashed line) Jaccard indices comparing the classification of UK genomes into transmission lineages and singletons on each of the 2000 posterior trees to the 1999 other trees. Dark shading shows the interquartile range and lighter shading the 95% CI.* ***(B)*** *Jaccard indices comparing the classification of UK genomes into transmission lineages and singletons on the MCC trees to each of the 2000 posterior trees (blue line). The solid red line indicates the median Jaccard index, dark shading the interquartile range and lighter shading the 95% CI.*

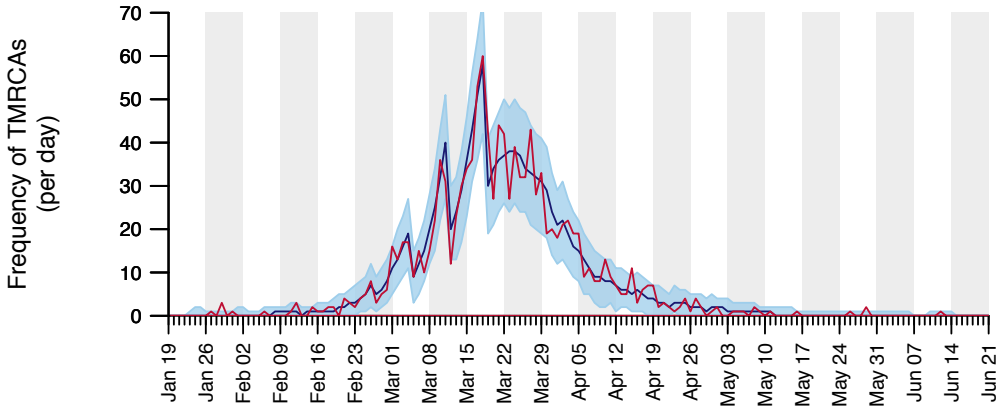

Fig. S14. *Comparison between the number of UK transmission lineage TMRCAs on each date in the MCC trees (red line) and across the 2000 posterior trees (median = blue line, 95% HPD interval = blue shading). Unevenness in this distribution is mostly likely caused by the phylogenetic constraints imposed by the sequence sampling times.*

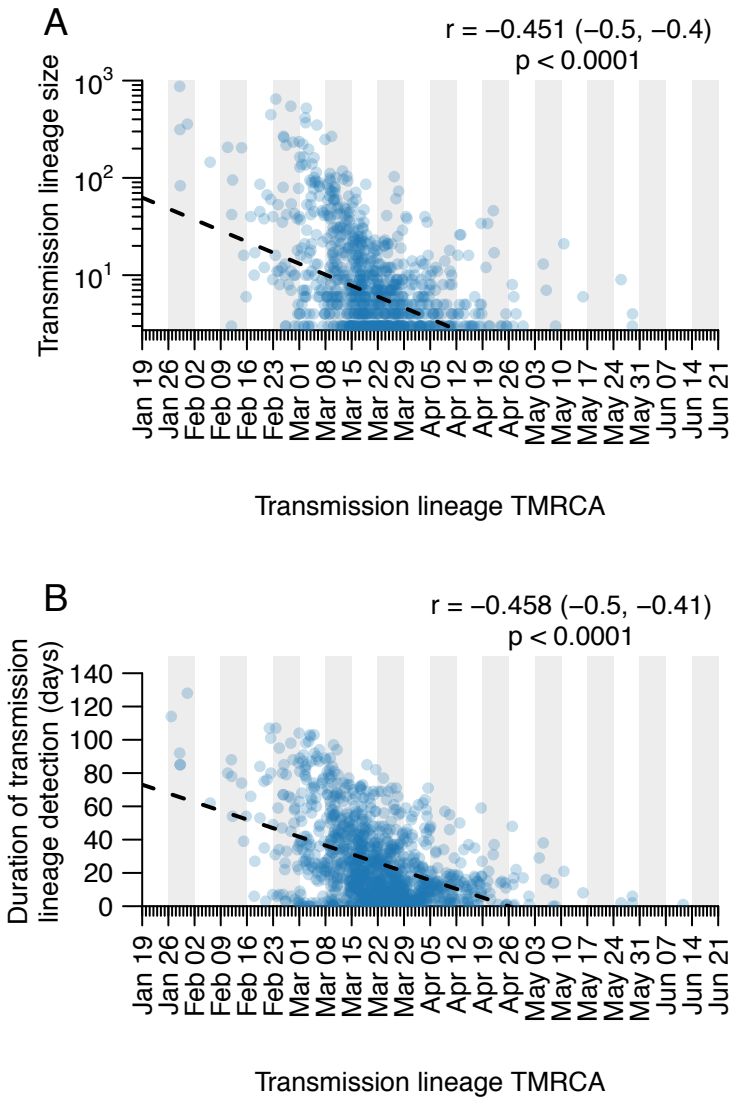

Fig. S15. *Scatterplots showing the relationship between (A) UK transmission lineage size and lineage TMRCA and between (B) UK transmission lineage sampling duration and lineage TMRCA. Pearson correlation coefficients, 95% CIs and p-values are shown in the top-right corners.*

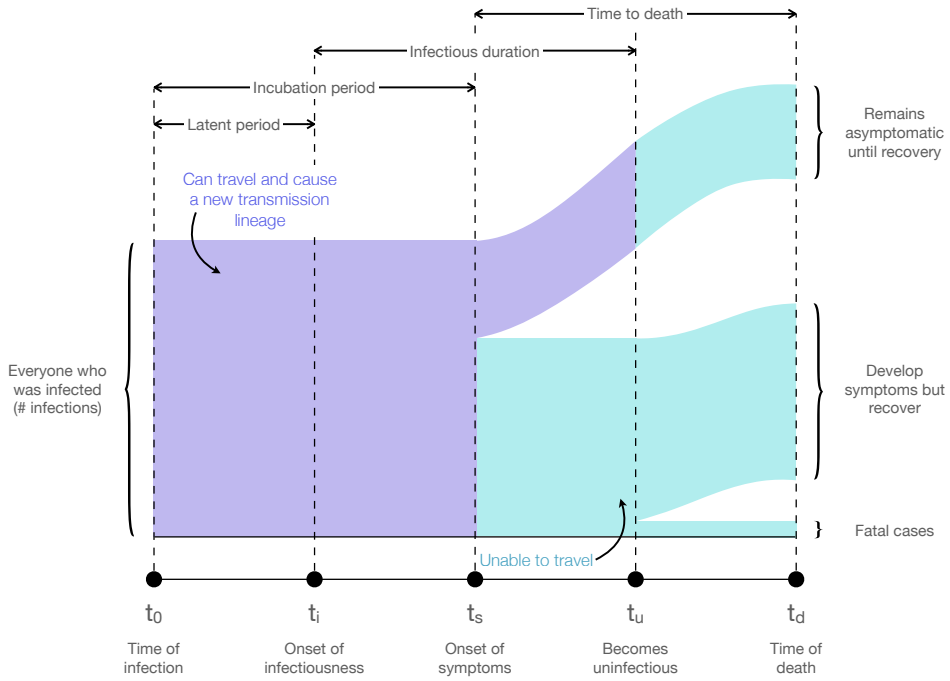

Fig. S16. *Sankey diagram showing the assumptions about the natural progression of a SARS-CoV-2 infection used in the estimation of global infectious cases. Infected individuals in the purple areas are potential initiators of a transmission lineage (PITL), but once they have progressed to the cyan areas they are assumed to no longer be capable of initiating a transmission lineage. We used the proportional flow through this diagram to estimate the total number of PITL through time given the number of COVID-19 associated deaths on each day.*

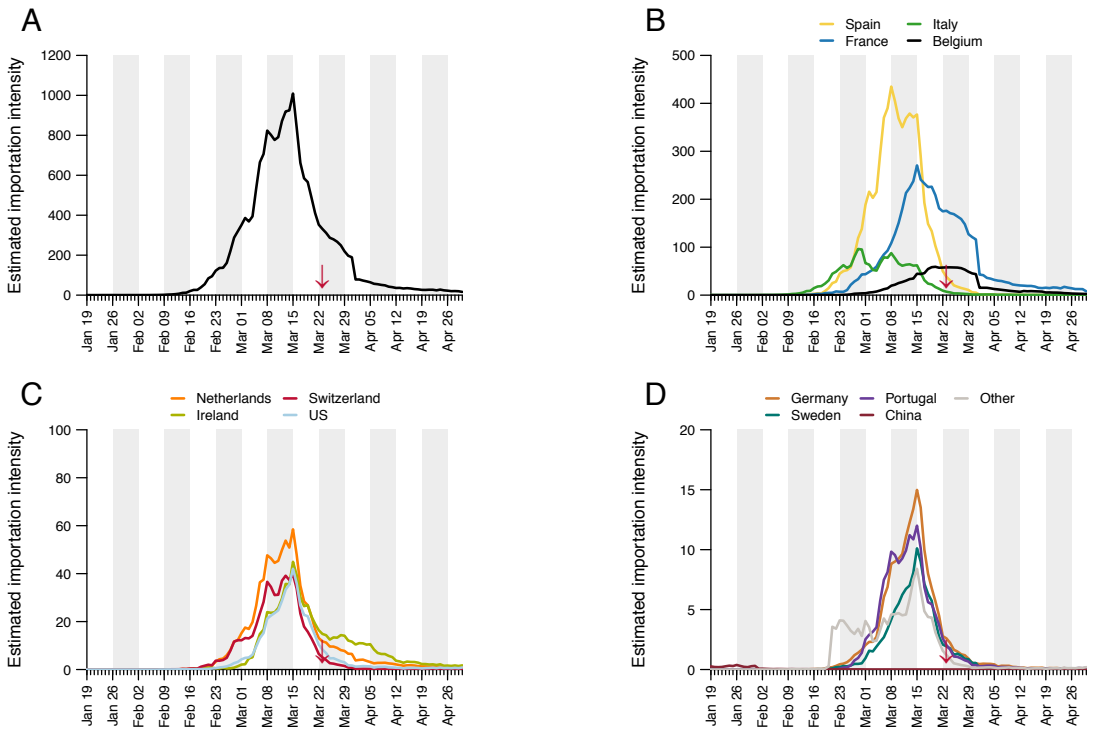

Fig. S17. *Estimated importation intensity (EII) curves for the 12 countries estimated to have contributed the most importations to the UK epidemic (see* ***Table S4****). Panel A shows the EII for all countries. The red arrows indicate the start of the UK lockdown.*

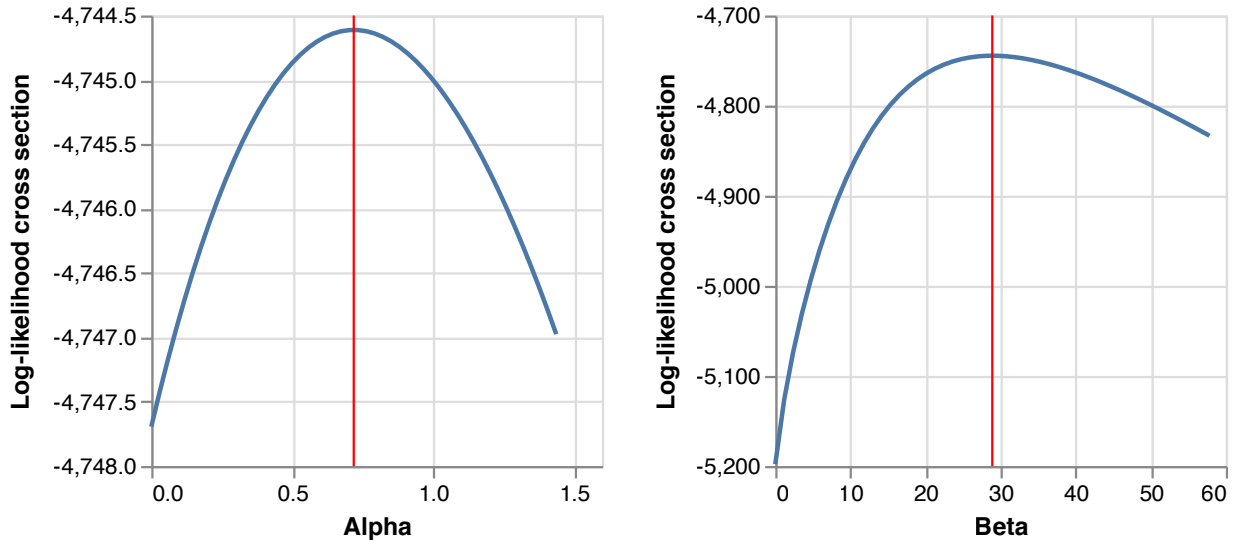

Fig S18. *Log-likelihood function cross-section plots for possible parameter values of α and β inferred from genomic data, conditional on daily importation probabilities derived from the estimated importation intensities (EIIs). The maximum likelihood estimate (MLE) for each parameter is shown in red.*

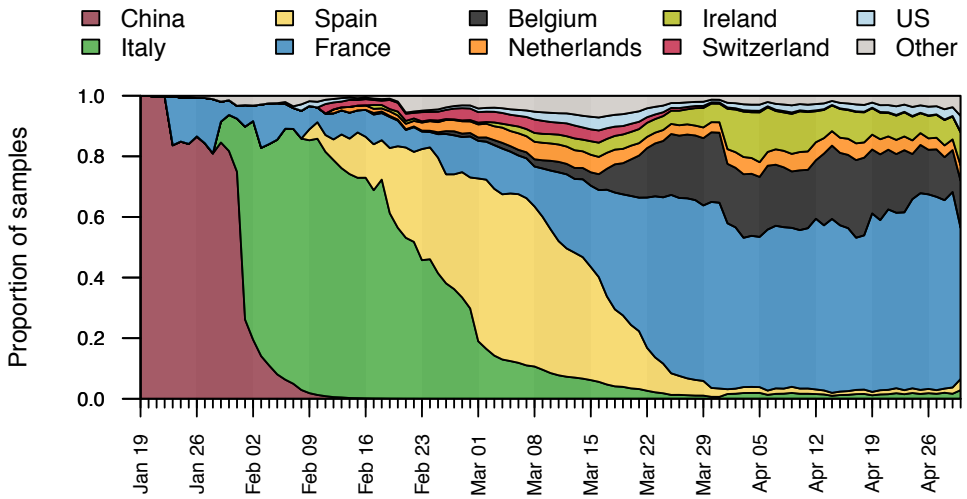

Fig. S19. *The estimated proportion of importation events that are attributable to inbound travellers from each of several source countries over time.*

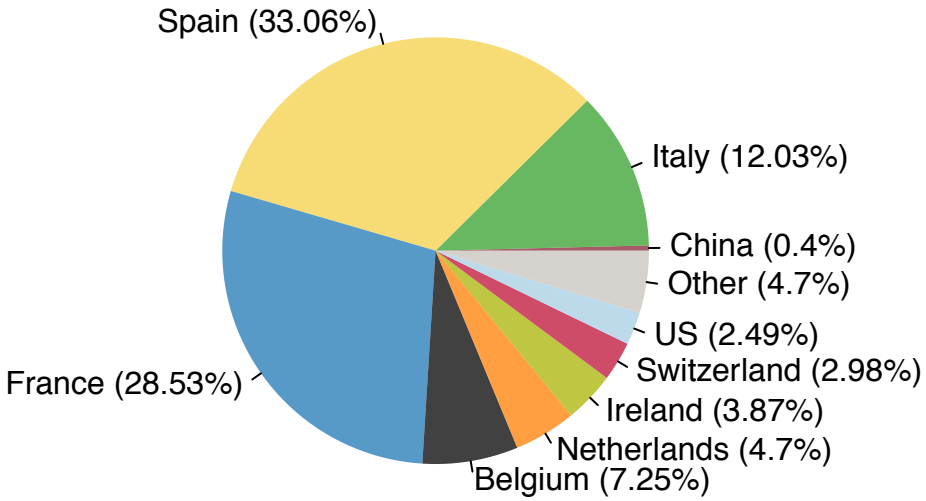

Fig. S20. *The estimated total fraction of importation events that are attributable to inbound travellers from each country.*

Table S1. The number of location state transitions (non-UK to UK and vice-versa) taken across the set of 2000 posterior trees, as well as the total number of transmission lineages and singletons inferred across the set of 2000 posterior trees and the MCC trees. Numbers are given for the whole dataset and for each individual subtree.

Lineages	Non-UK to UK state transitions (median, 95% HPD)	UK to non-UK state transitions (median, 95% HPD)	Transmission lineages and singletons (median, 95% HPD)	Transmission lineages and singletons in MCC tree
Total	2968 [2829-3103]	1468 [1362-1566]	2918 [2773-3048]	2829
A	74 [67-80]	3 [0-7]	74 [67-80]	69
B	372 [333-419]	446 [402-485]	365 [326-409]	360
B.1.1	1143 [1023-1258]	666 [584-749]	1115 [992-1230]	1074
B.1.pruned	977 [925-1026]	283 [245-321]	964 [914-1015]	943
B.1.X	398 [374-422]	70 [49-93]	396 [370-419]	383

HPD – highest posterior density interval

Table S2. Estimated importation lags for UK transmission lineages of different sizes. Importation lag is the waiting time between importation date and the TMRCA of the sampled genomes in the transmission lineage (see **Fig. S2**). Detection lag is the waiting time from the importation date to the sampling time of the oldest (first) sampled genome in the transmission lineage (see **Fig. S2**).

Lineages of size	No. of lineages	Importation lag (mean ± SD)	Importation lag (median, IQR)	Detection lag (mean ± SD)	Detection lag (median, IQR)
All	1179	8.22 ± 5.21	7.95 [3.35-15.18]	14.13 ± 5.61	14 [10-18]
2 to 10	880	10.37 ± 4.24	10.36 [6.5-15.18]	15.49 ± 5	16 [12-18]
11 to 100	261	2.07 ± 0.74	2.03 [1.41-2.65]	9.96 ± 4.92	9 [6-13]
101 to 1000	36	0.87 ± 0.08	0.86 [0.81-0.93]	11.08 ± 8.03	8.5 [5.75-15]
> 1000	2	0.74 ± 0	0.74 [0.74-0.74]	12.5 ± 2.12	12.5 [11.75-13.25]

SD – standard deviation
IQR – interquartile range

Table S3. Estimated importation and detection lags for UK transmission lineages ordered by importation date and aggregated by epi-week. Importation lag is the waiting time between importation date and the TMRCA of the sampled genomes in the transmission lineage (see **Fig. S2**). Detection lag is the waiting time from the importation date to the sampling time of the oldest (first) sampled genome in the transmission lineage (see **Fig. S2**). All statistics show means and standard deviations computed from the MCC trees.

Week starting	Epi-week	Estimated no. of importations	Lineage sizes (mean ± SD)	Importation lag (mean ± SD)	Detection lag (mean ± SD)
Jan-05	2	0	-	-	-
Jan-12	3	0	-	-	-
Jan-19	4	0	-	-	-
Jan-26	5	6	536.33 ± 598.96	2.42 ± 3.89	20.83 ± 11.81
Feb-02	6	2	73.5 ± 101.12	8.05 ± 10.08	20 ± 2.83
Feb-09	7	14	42.36 ± 73.45	8.72 ± 6.81	19.07 ± 4.41
Feb-16	8	45	63.4 ± 282.84	8.92 ± 5.8	14.27 ± 5.28
Feb-23	9	80	44.05 ± 108.48	7.82 ± 5.85	13.43 ± 6.46
Mar-01	10	206	26.53 ± 67.98	9.07 ± 5.64	14.34 ± 6.65
Mar-08	11	335	14.14 ± 25.04	7.87 ± 5.18	14.11 ± 5.39
Mar-15	12	235	8.32 ± 10.59	8.14 ± 4.91	13.54 ± 4.88
Mar-22	13	120	9.26 ± 13.88	7.78 ± 4.67	13.47 ± 4.78
Mar-29	14	71	5.96 ± 6.45	8.92 ± 4.62	14.77 ± 5.35
Apr-05	15	31	6.87 ± 6.59	7.92 ± 4.09	15.06 ± 5.83
Apr-12	16	15	6.6 ± 8.58	9.38 ± 4.67	15.73 ± 5.01
Apr-19	17	10	12.4 ± 15.49	7.9 ± 5.74	13.9 ± 3.84
Apr-26	18	3	7.67 ± 5.03	6.05 ± 3.85	15.33 ± 3.06
May-03	19	1	21	2.1	16
May-10	20	1	6	5.54	15
May-17	21	3	5.33 ± 3.21	7.41 ± 3.25	14.67 ± 2.89
May-24	22	1	2	15.18	19
May-31	23	0	-	-	-
Jun-07	24	0	-	-	-
Jun-14	25	0	-	-	-

Table S4. Number of observed importations in our dataset and the percentage of the total (1179) that can be attributed to the 40 countries inferred to be sources for the most importations.

Country	Observed importations	Percentage
Spain	387.12	33.066
France	334.04	28.532
Italy	140.83	12.029
Belgium	84.88	7.25
Netherlands	55.08	4.705
Ireland	45.3	3.869
Switzerland	34.91	2.982
US	29.16	2.491
Germany	10.85	0.927
Portugal	9.56	0.817
Sweden	6.71	0.573
China	4.64	0.397
Denmark	3.84	0.328
Austria	3.48	0.297
Romania	2.24	0.191
Norway	1.95	0.167
Poland	1.28	0.109
Canada	1.08	0.093
Turkey	0.96	0.082
Hungary	0.95	0.081
Czechia	0.65	0.056
Greece	0.55	0.047
United Arab Emirates	0.3	0.026
Israel	0.27	0.023
Finland	0.25	0.022
Iran	0.22	0.019
South Korea	0.2	0.017
Morocco	0.18	0.015
Brazil	0.17	0.015
Dominican Republic	0.16	0.013
Mexico	0.08	0.007
Serbia	0.07	0.006
Japan	0.07	0.006
Egypt	0.06	0.005
Malaysia	0.05	0.004
Pakistan	0.05	0.004
Moldova	0.04	0.004
Philippines	0.04	0.003
Russia	0.03	0.003
Ecuador	0.03	0.003
Other	8.39	0.717

Table S5. Summary of travel advice in the United Kingdom related to the COVID-19 pandemic from January to March.

Date	Type of notification or travel advice	Type of self-isolation recommended	Countries/regions affected by travel advice	Sources
24/01/2020	Considerations if returning from Wuhan City (14 days prior)	None	China (Wuhan City)	https://web.archive.org/web/20200124231713/https://www.gov.uk/guidance/wuhan-novel-coronavirus-information-for-the-public
26/01/2020	FCO advises against all travel to Hubei, China	14 days (all travellers returning from Hubei 14 days prior)	China (Hubei Province)	https://www.fitfortravel.nhs.uk/news/newsdetail.aspx?id=23664 https://web.archive.org/web/20200126084226/https://www.gov.uk/foreign-travel-advice/china
28/01/2020	FCO advises against all travel to Hubei, China; against all non-essential travel to Continental China	14 days (all travellers returning from Hubei 14 days prior)	China (Continental)	https://www.fitfortravel.nhs.uk/news/newsdetail.aspx?id=23665 https://web.archive.org/web/20200128151730/https://www.gov.uk/guidance/wuhan-novel-coronavirus-information-for-the-public
04/02/2020	FCO advises against all travel to Hubei, China; against all non-essential travel to Continental China	14 days (if symptomatic); 14 days (all travellers returning from Hubei)	China (Continental)	https://www.gov.uk/government/news/coronavirus-and-travel-to-china-foreign-secretarys-statement-4-february-2020 https://web.archive.org/web/20200204143029/https://www.gov.uk/guidance/wuhan-novel-coronavirus-information-for-the-public
06/02/2020	FCO advises against all travel to Hubei, China; against all non-essential travel to Continental China; self-quarantine if returning from countries at risk (see Countries/regions affected by travel advice column)	14 days (if symptomatic); 14 days (all travellers returning from Hubei)	China (Continental); Hong Kong; Japan; Macao; Malaysia; South Korea; Singapore; Taiwan; Thailand	https://www.fitfortravel.nhs.uk/news/newsdetail.aspx?id=23675 https://web.archive.org/web/20200206023753/https://www.gov.uk/guidance/wuhan-novel-coronavirus-information-for-the-public
25/02/2020	FCO advises against all travel to Hubei,	14 days (if symptomatic);	China (Continental); Hong Kong; Japan;	https://www.fitfortravel.nhs.uk/news/newsdetail.aspx?id=2

	China; against all non-essential travel to Continental China; against all non-essential travel to Daegu and Cheongdo, South Korea self-quarantine if returning from countries at risk (see Countries/regions affected by travel advice column)	14 days (all travellers returning from Hubei 14 days prior; all travellers returning from Daegu or Cheongdo, northern Italy or Iran 6 days prior)	Macao; Malaysia; South Korea; Singapore; Taiwan; Thailand; Vietnam; Cambodia; Laos; Myanmar; Iran; Italy (north of Pisa/Florence/Rimini)	3695 https://web.archive.org/web/20200225231202/https://www.gov.uk/guidance/wuhan-novel-coronavirus-information-for-the-public
13/03/2020	FCO advises against all travel to Hubei, China; against all non-essential travel to various countries at risk (see Countries/regions affected by travel advice column)	7 days (if symptomatic, applied to the general population and not just returning travellers)	China (Continental); Hong Kong; Japan; Macao; Malaysia; South Korea; Singapore; Taiwan; Thailand; Vietnam; Cambodia; Laos; Myanmar; Iran; Italy; Spain; Denmark; Norway; Czech Republic; Cyprus; Romania; Lebanon; South Africa; Peru; Kenya; Jamaica; Poland; Slovakia; Argentina; Malta; Albania; Kosovo; Estonia; San Marino; Equatorial Guinea; Liberia; Lithuania; Latvia; Mongolia; Philippines; Sierra Leone; Portugal (Madeira and Azores); Ecuador; Sri Lanka; Paraguay; Guatemala; Honduras; United States of America	https://web.archive.org/web/20200315234341/https://www.gov.uk/guidance/coronavirus-covid-19-information-for-the-public; https://www.itv.com/news/2020-03-15/coronavirus-outbreak-foreign-office-advises-against-all-but-essential-travel-fco-advice; https://web.archive.org/web/20200313135510/https://www.gov.uk/guidance/travel-advice-novel-coronavirus
23/03/2020	FCO advises British people travelling abroad to return to the UK if commercial flights are available	General stay-at-home order	All residents in the UK and all returning travellers regardless of destination	https://www.fitfortravel.nhs.uk/news/newsdetail.aspx?id=23713 https://www.theguardian.com/uk-news/2020/mar/23/boris-johnsons-address-to-the-nation-in-full