UC Irvine
UC Irvine Previously Published Works

Title
Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK.

Permalink
https://escholarship.org/uc/item/9hf8k3c7

Journal
Science, 371(6530)

Authors
du Plessis, Louis
McCrone, John
Zarebski, Alexander
et al.

Publication Date
2021-02-12

DOI
10.1126/science.abf2946

Peer reviewed

Establishment and 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[10,11], Isaac I. Bogoch[12,13], Kamran Khan[10,11,12],
COVID-19 Genomics UK (COG-UK) Consortium†, David M. Aanensen[14,15],
Moritz U. G. Kraemer[1]‡, Andrew Rambaut[2]‡§, Oliver G. Pybus[1,16]‡§**

[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, Cardiff, UK. [7]Pathogen Genomics Unit, Public Health Wales NHS Trust, Cardiff, UK. [8]MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, 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]Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK. [15]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [16]Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, UK.

*These authors contributed equally to this work.

†See supplementary materials for list of consortium members and affiliations.

‡These authors contributed equally to this work.

§Corresponding author. Email: a.rambaut@ed.ac.uk (A.R.); oliver.pybus@zoo.ox.ac.uk (O.G.P.)

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 tended to be larger and more dispersed. Lineage importation and regional lineage diversity declined after lockdown, while lineage elimination was size-dependent. We discuss the implications of our genetic perspective on transmission dynamics for COVID-19 epidemiology and control.

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. While 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 unprecedent-

ed, 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) (*17*). 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 (*18*) we infer virus genomes were generated for 0.66% (*95%* CI = 0.46-0.95%) of all UK infections by 5th May (Fig. 1B).

Genetic structure and lineage dynamics of the UK epidemic from January to June

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. 2, A and B). Sources of statistical uncertainty in lineage assignation were taken into account (*17*). 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 (*19*) or incomplete or uneven genome sampling (*20, 21*). Therefore we expect this number to be an underestimate (*17*). In our phylogenetic analysis 1650 (*95%* HPD 1611-1783) UK genomes could not be allocated to a UK transmission lineage (singletons). Had more genomes been sequenced, it is likely that many of these singletons would have been assigned to a UK transmission lineage. Further, many singleton importations are likely to be unobserved.

Most transmission lineages are small and 72.4% (*95%* HPD 69.3-72.9%) contain <10 genomes (Fig. 2C). However the lineage size distribution is strongly skewed and follows a power-law distribution (Fig. 2C, inset), such that the 8 largest UK transmission lineages contain >25% of all sampled UK genomes (Fig. 2D; figs. S2 to S5 show further visualizations). 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). Since our dataset comprises only a small fraction of all UK infections, these observed lineage sizes will underestimate true lineage size. However, the true distribution of relative lineage sizes will closely match our observation, and its power-law shape indicates that almost all unobserved lineages will be small. All 8 largest lineages were first detected before the UK national lockdown was announced on 23 March and, as expected, larger lineages were observed for longer (Pearson's r = 0.82; *95%* CI = 0.8-0.83; fig. S7). The sampling frequency of lineages of varying sizes differed over time (Fig. 3A and figs. S8 and S9); while 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. 3A).

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. 3B). The per-genome rate of appearance of new lineages was initially high, then declined throughout March and April (Fig. 3C), 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. 1, A and B, and Fig. 3C). 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.

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 (data S1). Although genomes were not collected randomly (some lineages and regions will be over-represented due to targeted investigation of local outbreaks; e.g., (*22*) the number of UK lineages detected in each region correlates with the number of genomes sequenced (Fig. 4A, Pearson's r = 0.96, *95%* CI = 0.95-0.98) and the number of reported cases (fig. S10, Pearson's r = 0.53, 95% CI = 0.35-0.67, data S2) 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. 4B; 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. 4C and data S3). 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 (23). Locations with the three lowest nonzero SIs were in Scotland (Stirling = 0.96, Aberdeenshire = 1.04, Inverclyde = 1.32; Fig. 4C). We speculate that regional differences in transmission lineage diversity may be related to the level of connectedness to other regions.

To illustrate temporal trends in transmission lineage diversity, we plot SI through time for each of the UK's national capital cities (Fig. 4D). Lineage diversities in each peaked in late March and declined after the UK national lockdown, congruent with Fig. 3, C and D. Greater London's epidemic was the most diverse and characterised by an early, rapid rise in SI (Fig. 4D), consistent with epidemiological trends there (16, 24). Belfast's lineage diversity was notably lower (data S4 shows other locations).

We observe variation in the spatial range of individual UK transmission lineages. Although some lineages are widespread, most are more localized and the range size distribution is right-skewed (fig. S11), congruent with an observed abundance of small lineages (Figs. 2C and 4B) and biogeographic theory [e.g., (25)]. 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. 4E and fig. S12). The national distribution of cases therefore arose from the aggregation of multiple heterogeneous lineage-specific patterns.

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., (26)]. 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 (20), we developed a new approach that combines virus phylogenetics with epidemiological and travel data. First, we estimated the TMRCA (time of the most recent common ancestor) of each UK transmission lineage (17). 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 tend to be larger and longer-lived than those whose TMRCAs postdate the national lockdown (Fig. 5A and fig. S15).

Due to incomplete sampling, TMRCAs best represent the date of the first inferred transmission event in a lineage, not its importation date (Fig. 2B). 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 (17) worldwide. Before March, the UK received ~1.75m inbound travellers per week (school holidays explain the end-February ~10% increase; Fig. 5B). 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. 5B). 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 (17). Since both travel volumes and epidemic incidence fluctuate rapidly over orders of magnitude, the EII is robust to other sources of variation in the relative importation risk among countries (17). The EII peaks in mid-March, when high UK inbound travel volumes coincided with growing numbers of infectious cases elsewhere (Fig. 5, B and C).

Crucially, the EII's temporal profile closely matches, but precedes, that of the TMRCAs of UK transmission lineages (Fig. 5, A and C). The difference between the two represents the "importation lag", the time elapsed between lineage importation and the first detected local transmission event (Fig. 2B). Using a statistical model (17), 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. 2B) (17). We use this model to impute an importation date for each UK transmission lineage (Fig. 5D). 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 (tables S2 and 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. This assignment is statistical and does not take the effects of superspreading events into account. As with the rate of importation (Fig. 5A), the relative contributions of arrivals from different countries were dynamic (Fig. 5D). 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 and table S4). These large-scale trends were not apparent from individual-level travel histories; routine collection of such data ceased on 12 March (*27*).

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 (Fig. 5B and table S5) led to the establishment and co-circulation of >1000 identifiable UK transmission lineages (Fig. 5A), jointly contributing to accelerated epidemic growth that quickly exceeded national contact tracing capacity (*27*). 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., (*28–30*)].

Earlier lineages were larger, more dispersed, and harder to eliminate, highlighting the importance of rapid or preemptive interventions in reducing transmission [e.g., (*31–33*)]. The high heterogeneity in SARS-CoV-2 transmission at the individual level (*34–36*) appears to extend to whole transmission lineages, such that >75% of sampled viruses belong to the top 20% of lineages ranked by size. While the national lockdown coincided with limited importation and reduced regional lineage diversity, its impact on lineage extinction was size-dependent (Fig. 3, B and C). The over-dispersed nature of SARS-CoV-2 transmission likely exacerbated this effect (*37*), thereby favoring, as R_t declined, greater survival of larger widespread lineages and faster local elimination of lineages in low prevalence regions. The de-

gree to which the surviving lineages contributed to the UK's ongoing second epidemic, including the effect of specific mutations on lineage growth rates [e.g., (*11*)], 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 AND NOTES

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

2. O. Faye, P. Y. Boëlle, E. Heleze, O. Faye, C. Loucoubar, N. Magassouba, B. Soropogui, S. Keita, T. Gakou, 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). [doi:10.1016/S1473-3099(14)71075-8](doi:10.1016/S1473-3099(14)71075-8) Medline

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). doi:10.1016/j.jhin.2016.10.008 Medline

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). doi:10.1073/pnas.1109314108 Medline

5. G. J. Baillie, 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 H1N1/2009 influenza virus lineages revealed by whole-genome analysis. *J. Virol.* **86**, 11–18 (2012). [doi:10.1128/ JVI.05347-11](doi:10.1128/JVI.05347-11) Medline

6. G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, 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. Malboeuf, 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). doi:10.1038/nature22040 Medline

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 4th, K. N. Hogan, A. C. Cannons, R. Jean, A. J. Monaghan, 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). [doi:10.1038/nature22400](#) [Medline](#)

8. A. F. Y. Poon, R. Gustafson, P. Daly, L. Zerr, 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). [doi:10.1016/S2352-3018(16)00046-1](#) [Medline](#)

9. J. Thomas, N. Govender, K. M. McCarthy, L. K. Erasmus, T. J. Doyle, M. Allam, A. Ismail, N. Ramalwa, P. Sekwadi, G. Ntshoe, A. Shonhiwa, V. Essel, N. Tau, S. Smouse, H. M. Ngomane, B. Disenyeng, N. A. Page, N. P. Govender, A. G. Duse, R. Stewart, T. Thomas, D. Mahoney, M. Tourdjman, O. Disson, P. Thouvenot, M. M. Maury, A. Leclercq, M. Lecuit, A. M. Smith, L. H. Blumberg, Outbreak of Listeriosis in South Africa Associated with Processed Meat. *N. Engl. J. Med.* **382**, 632–643 (2020). [doi:10.1056/NEJMoa1907462](#) [Medline](#)

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). [doi:10.1038/s41467-019-11216-7](#) [Medline](#)

11. E. M. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. 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, COG-UK Consortium, I. Goodfellow, N. J. Loman, O. G. Pybus, D. L. Robertson, E. C. Thomson, A. Rambaut, T. R. Connor, Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 1–12 (2021). [doi:10.1016/j.cell.2020.11.020](#)

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). [doi:10.1093/molbev/msx195](#) [Medline](#)

13. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019). [doi:10.1038/s41564-018-0296-2](#) [Medline](#)

14. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). [doi:10.2807/1560-7917.ES.2017.22.13.30494](#) [Medline](#)

15. COVID-19 Genomics UK (COG-UK) Consortium, An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020). [doi:10.1016/S2666-5247(20)30054-9](#) [Medline](#)

16. U.K. Government, Coronavirus (COVID-19) in the UK; https://coronavirus.data.gov.uk/cases.

17. See supplementary materials.

18. 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, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, S. Bhatt, Imperial College COVID-19 Response Team, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020). [doi:10.1038/s41586-020-2405-7](#) [Medline](#)

19. C. J. Villabona-Arenas, W. P. Hanage, D. C. Tully, Phylogenetic interpretation during outbreaks requires caution. *Nat. Microbiol.* **5**, 876–877 (2020). [doi:10.1038/s41564-020-0738-5](#) [Medline](#)

20. 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* **370**, 564–570 (2020). [Medline](#)

21. 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 [20127738](#) [preprint]. 12 June 2020.

22. Welsh Government, Genomic analysis of Covid-19 lineages in Wales; [https://gov.wales/genomic-analysis-covid-19-lineages-wales](#).

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

24. C. Angus, CoVid Plots and Analysis. University of Sheffield (2020); [https://doi.org/10.15131/shef.data.12328226](#).

25. K. J. Gaston, F. He, The distribution of species range size: A stochastic process. *Proc. R. Soc. B* **269**, 1079–1086 (2002). [doi:10.1098/rspb.2002.1969](#) [Medline](#)

26. G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, Mathematical models to characterize early epidemic growth: A review. *Phys. Life Rev.* **18**, 66–97 (2016). [doi:10.1016/j.plrev.2016.07.005](#) [Medline](#)

27. C. Baraniuk, Covid-19 contact tracing: A briefing. *BMJ* **369**, m1859 (2020). [Medline](#)

28. 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 [20168930](#) [preprint]. 20 August 2020.

29. J. Lu, J. 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.e9 (2020). [doi:10.1016/j.cell.2020.04.023](#) [Medline](#)

30. T. Seemann, C. R. Lane, N. L. Sherry, S. Duchene, A. Gonçalves 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). [doi:10.1038/s41467-020-18314-x](#) [Medline](#)

31. C. Dye, R. C. H. Cheng, J. S. Dagpunar, B. G. Williams, The scale and dynamics of COVID-19 epidemics across Europe. *R. Soc. Open Sci.* **7**, 201726 (2020). [doi:10.1098/rsos.201726](#)

32. 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**, 638–642 (2020). [doi:10.1126/science.abb6105](#) [Medline](#)

33. 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). [doi:10.1016/S0140-6736(20)30746-7](#) [Medline](#)

34. 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.* **26**, 1714–1719 (2020). [doi:10.1038/s41591-020-1092-0](#) [Medline](#)

35. A. Endo, S. Abbott, A. J. Kucharski, S. Funk, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020). [doi:10.12688/wellcomeopenres.15842.3](#) [Medline](#)

36. L. Wang, X. Didelot, J. Yang, G. Wong, Y. Shi, W. Liu, G. 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). [doi:10.1038/s41467-020-18836-4](#) [Medline](#)

37. 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). [doi:10.1038/nature04153](#) [Medline](#)

38. "Establishment and lineage dynamics of the SARS-Cov-2 epidemic in the UK" GitHub repository (https://github.com/COG-UK/uk-intros-analyses, DOI: 10.5281/zenodo.4311597).

39. S. M. Nicholls *et al.*, MAJORA: Continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. bioRxiv 328328 [preprint]. 7 October 2020.

40. J. Quick, nCoV-2019 sequencing protocol v3 (LoCost) V.3 (2020); www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye.

41. J. B. Singer, R. J. Gifford, M. Cotton, D. L. Robertson, CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. Preprints 202006025 (18 June 2020).

42. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline

43. 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.* **5**, 1403–1407 (2020). doi:10.1038/s41564-020-0770-5 Medline

44. 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 (2020); https://cov-lineages.org/pangolin.

45. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). doi:10.1371/journal.pone.0009490 Medline

46. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). doi:10.1093/ve/vey016 Medline

47. M. A. R. Ferreira, M. A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36**, 355–369 (2008). doi:10.1002/cjs.5550360302

48. 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). doi:10.1093/molbev/mss265 Medline

49. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). doi:10.1093/sysbio/syy032 Medline

50. 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). doi:10.1093/oxfordjournals.molbev.a025892 Medline

51. E. M. Volz, S. D. W. Frost, Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3**, vex025 (2017). doi:10.1093/ve/vex025

52. 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). doi:10.1093/nar/gky783 Medline

53. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). doi:10.1093/ve/vex042 Medline

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

55. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* **5**, e1000520 (2009). doi:10.1371/journal.pcbi.1000520 Medline

56. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008). doi:10.1007/s00285-007-0120-8 Medline

57. 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, Y. Taha, C. J. A. Duncan, M. L. Schmid, PHE Incident Team, Airborne HCID Network, Novel coronavirus disease (Covid-19): The first two patients in the UK with person to person transmission. *J. Infect.* **80**, 578–606 (2020). doi:10.1016/j.jinf.2020.02.020 Medline

58. 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). doi:10.1016/S1473-3099(20)30120-1 Medline

59. United Nations, Department of Economic and Social Affairs, Population Division, *World Population Prospects 2019: Methodology of the United Nations Population Estimates and Projections* (2019); https://population.un.org/wpp/Publications/Files/WPP2019_Methodology.pdf.

60. X. He, E. H. Y. Lau, P. Wu, X. Deng, 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. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020). doi:10.1038/s41591-020-0869-5 Medline

61. 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, Y. Luo, Y. Liu, G. Shao, H. Li, 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. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline

62. R. Verity, L. C. Okell, I. Dorigatti, 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. Boonyasiri, 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. Donnelly, A. C. Ghani, N. M. Ferguson, Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020). doi:10.1016/S1473-3099(20)30243-7 Medline

63. 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). doi:10.1016/j.ijid.2020.03.020 Medline

64. L. Roques, E. K. Klein, J. Papaïx, A. Sar, S. Soubeyrand, Using Early Data to Estimate the Actual Infection Fatality Ratio from COVID-19 in France. *Biology* **9**, 97 (2020). doi:10.3390/biology9050097 Medline

65. 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**, 2000256 (2020).

66. 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**, 2000180 (2020). doi:10.2807/1560-7917.ES.2020.25.10.2000180 Medline

67. M. Day, Covid-19: Four fifths of cases are asymptomatic, China figures indicate. *BMJ* **369**, m1375 (2020). doi:10.1136/bmj.m1375 Medline

68. M. J. Sanderson, How Many Taxa Must Be Sampled to Identify the Root Node of a Large Clade? *Syst. Biol.* **45**, 168–173 (1996). doi:10.1093/sysbio/45.2.168

SUPPLEMENTARY MATERIALS
science.sciencemag.org/cgi/content/full/science.abf2946/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S5
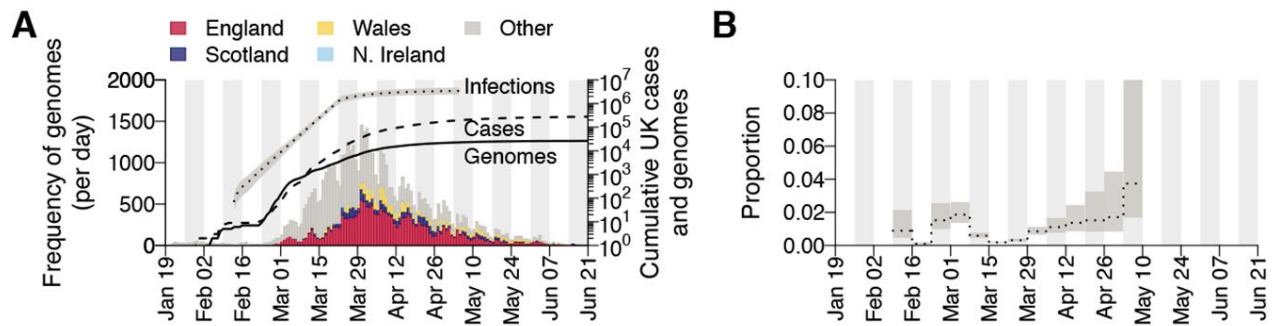Data S1 to S5
References (*39–68*)

Fig. 1. Genomic sequence data. (A) Collection dates of the 50,887 genomes analyzed here (left-hand axis). Genomes are colored 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 (18); 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) Proportion of weekly estimated UK infections (18) included in our genome sequence dataset.
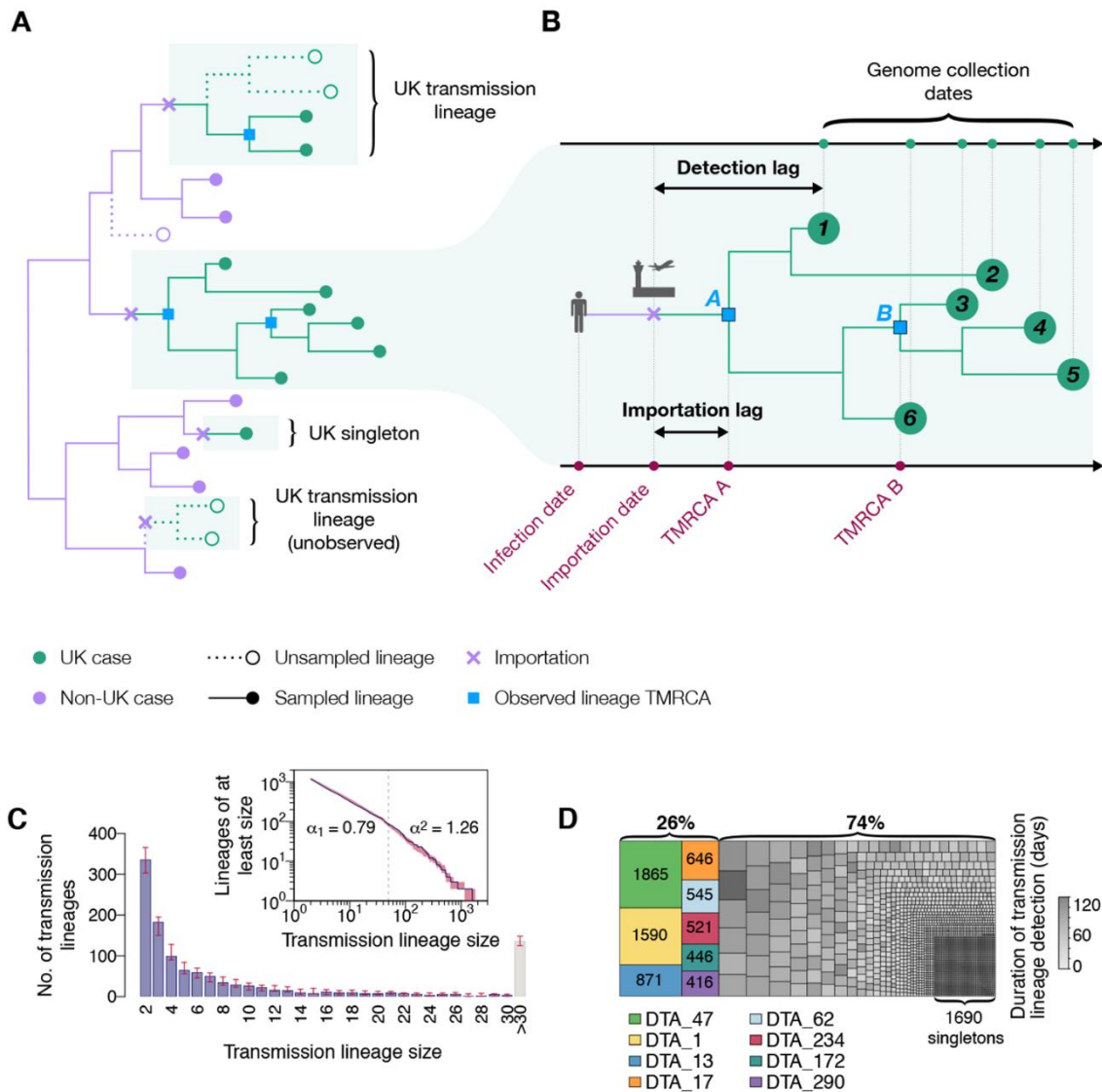
Fig. 2. Structure of UK transmission lineages detected through genome sampling. (A) Figurative illustration of the international context of UK transmission lineages. Note only half of the cases in the top UK transmission lineage are observed and the bottom UK transmission lineage is unobserved. To be detected, a UK transmission lineage must contain two or more sampled genomes; singletons are not classified here as UK transmission lineages. (B) Detailed view of one of the UK transmission lineages from (A), used to illustrate the terms TMRCA, detection lag, and importation lag. 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. (C) 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. (D) Partition of 26,181 UK genomes into UK transmission lineages and singletons, colored 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. The sizes of the 8 largest lineages are also shown in the figure.
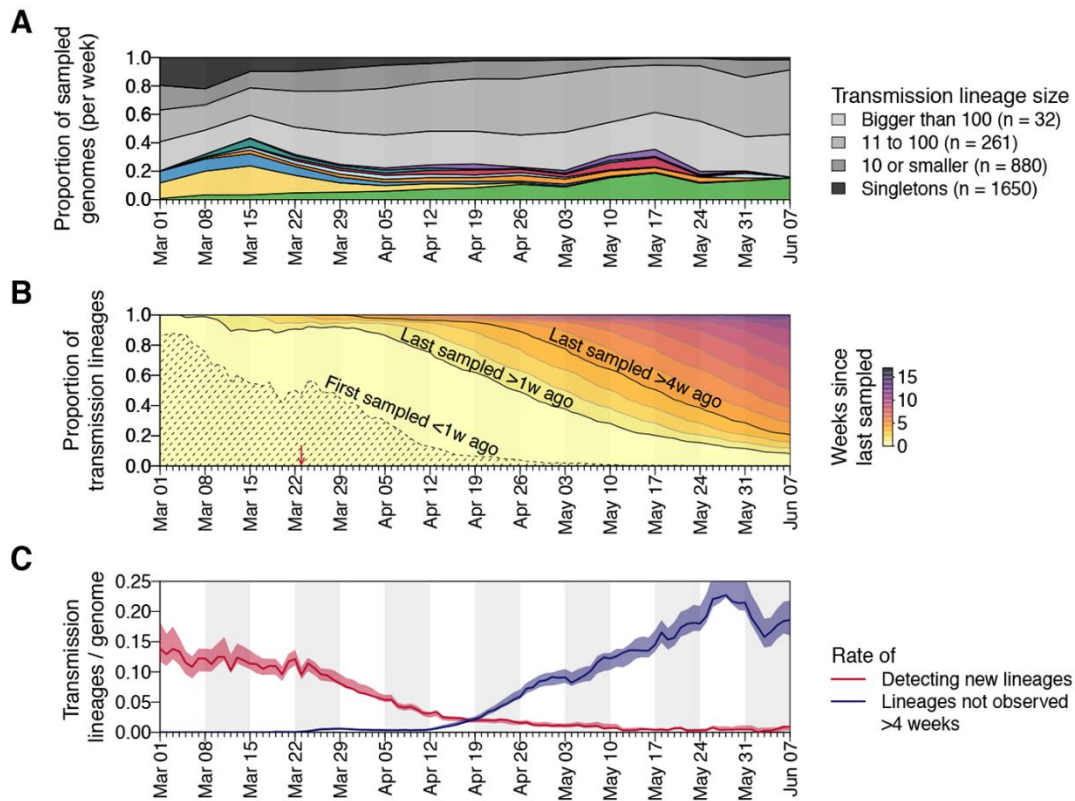
Fig. 3. Dynamics of UK transmission lineages. (A) Lineage size breakdown of UK genomes collected each week. Colors of the 8 largest lineages are as depicted in Fig. 2D. (B) Trends through time in the detection of UK transmission lineages. For each day, all lineages detected up to that day are colored 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. (C) Red line = daily rate of detecting new transmission lineages. Blue line = rate at which lineages have not been observed for >4 weeks, shading = 95% HPD across the posterior distribution of trees.
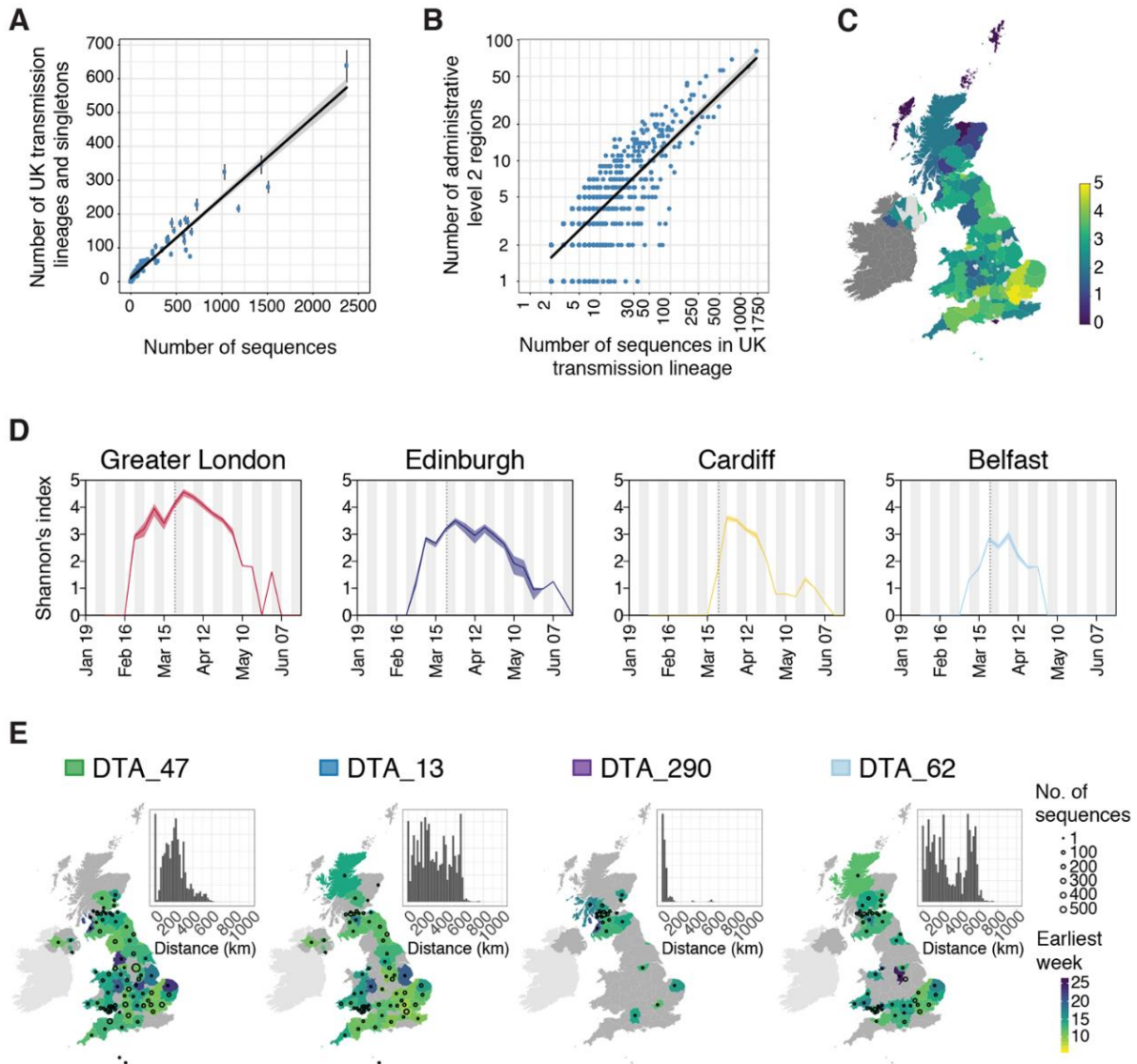
Fig. 4. Spatial distribution of UK transmission lineages. (A) Correlation between the number of transmission lineages detected in each region (points = median values, bars = 95% HPD intervals) and the number of UK virus genomes from each region (Pearson's r = 0.96, 95% CI = 0.95-0.98). (B) Correlation 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,) (C) Map showing Shannon's index (SI) for each region, calculated across the study period (2nd Feb-26th Jun). Yellow colors indicate higher SI values and darker colors lower values. (D) SI through time for the UK national capital cities. The dotted lines indicate the start of the UK national lockdown. (E) Illustration of the diverse spatial range distributions of UK transmission lineages. Colors 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 data S4 and fig. S12 for further details). Colored boxes next to lineage names are as depicted in Fig. 2D.
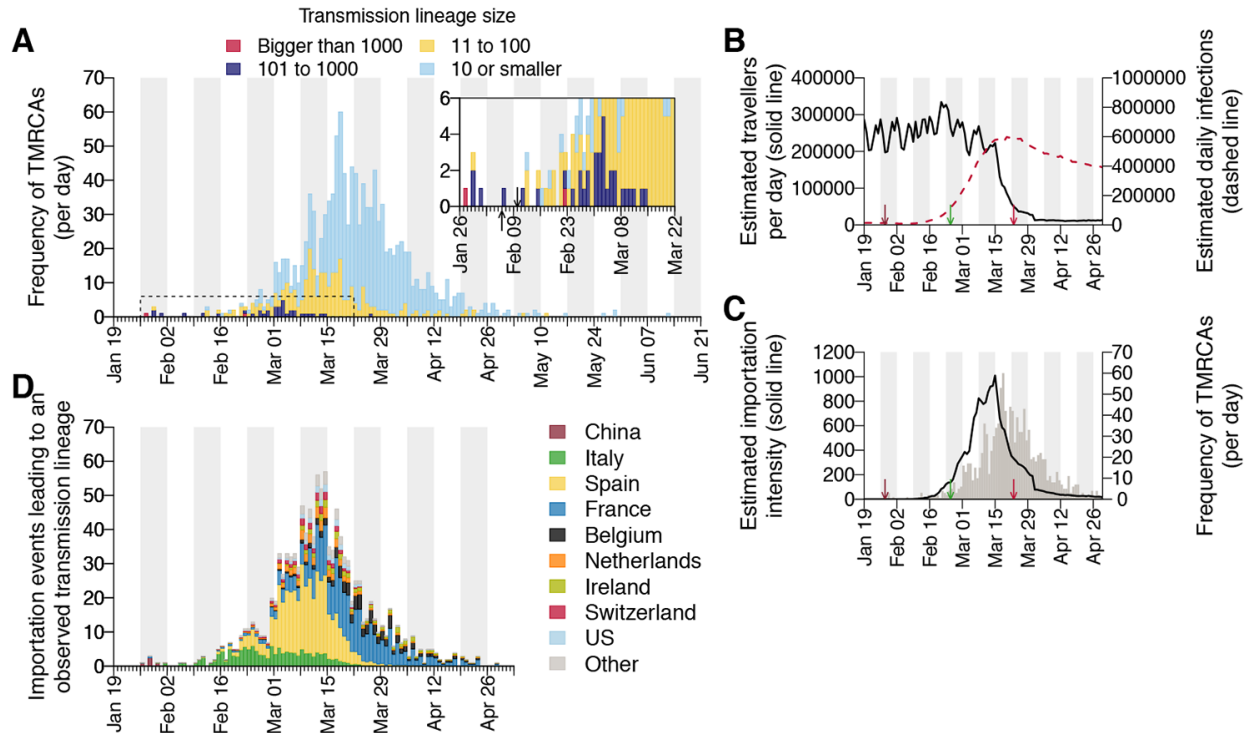
Fig. 5. Dynamics of UK transmission lineage importation. (A) Histogram of lineage TMRCAs, colored by lineage size. Inset: expanded view of the days prior to UK lockdown. Left-hand arrow = collection date of the UK's first laboratory-confirmed case; right-hand arrow = collection date of the earliest UK 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). Arrows here show, from left to right, dates of the first self-isolation advice for returning travellers from China, Italy, and of the start of the UK national 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. Colors show the proportion attributable each day to inbound travel from various countries (see table S4 and figs. S19 and S20). This assignment is statistical, i.e., we cannot ascribe a specific source location to any given lineage.