

CORONAVIRUS

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 Raghwan¹, Jordan Ashworth², Rachel Colquhoun², Thomas R. Connor^{6,7}, Nuno R. Faria^{1,8}, Ben Jackson², Nicholas J. Loman⁹, Áine O'Toole², Samuel M. Nicholls⁹, Kris V. Parag⁸, Emily Scher², Tetyana I. Vasylieva¹, Erik M. Volz⁸, 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‡§}

The United Kingdom's COVID-19 epidemic during early 2020 was one of world's largest and was unusually well represented by virus genomic sampling. We determined the fine-scale genetic lineage structure of this epidemic through analysis of 50,887 severe acute respiratory syndrome coronavirus 2 (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 quantified the size, spatiotemporal 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, whereas 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, or how key properties such as epidemic size and duration arise from their combined action. Although 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 instead be 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 researchers to 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 severe acute respiratory syndrome corona-

virus 2 (SARS-CoV-2) genomes shared online by 1 October 2020 (14). About half of these represent infections in the United Kingdom 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 after first positive test (16)]. Here, we combine this large genomic dataset with epidemiolog-

ical and travel data to provide a full characterization 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 represent 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 that virus genomes were generated for 0.66% [95% confidence interval (CI), 0.46 to 0.95%] of all UK infections by 5 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 assignment were taken into account (17). We identified a total of 1179 [95% highest posterior density (HPD), 1143 to 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 as a result of limited SARS-CoV-2 genetic diversity (19) or incomplete or uneven genome sampling (20, 21). Therefore we expect this number to be an

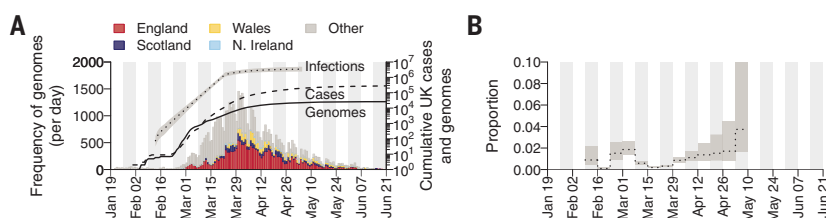


Fig. 1. Genomic sequence data. (A) Collection dates of the 50,887 genomes analyzed here (left axis). Genomes are colored by sampling location (red, England; dark blue, Scotland; yellow, Wales; light blue, Northern Ireland; gray, elsewhere). The solid line shows the cumulative number of UK virus genomes (right 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); gray shading denotes the 95% CI. As a result of 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.

¹Department of Zoology, University of Oxford, Oxford, UK. ²Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. ³Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, UK. ⁴Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. ⁵School of Biological and Environmental Sciences, Universidad San Francisco de Quito, Quito, Ecuador. ⁶School of Biosciences, Cardiff University, Cardiff, UK. ⁷Pathogen Genomics Unit, Public Health Wales NHS Trust, Cardiff, UK. ⁸MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK. ⁹Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. ¹⁰Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Canada. ¹¹BlueDot, Toronto, Canada. ¹²Department of Medicine, University of Toronto, Toronto, Canada. ¹³Divisions of General Internal Medicine and Infectious Diseases, University Health Network, Toronto, Canada. ¹⁴Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK. ¹⁵Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ¹⁶Department 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.)

underestimate (17). In our phylogenetic analysis, 1650 (95% HPD, 1611 to 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 to 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 eight 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 to 2133 and 1342 to 2011 genomes, respectively). Because our dataset constitutes 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 eight 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 to 0.83; fig. S7). The sampling frequency of lineages of varying sizes differed over time (Fig. 3A and figs. S8 and S9); whereas 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 characterized 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 1 May, 96.2% of sampled genomes belonged to transmission lineages that were first observed >7 days previously. By 1 June, a growing number of lineages (>73%) had not been detected by genomic sampling for >4 weeks, which suggests that they were rare or had gone extinct; this result 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 nonpharmaceutical interventions (NPIs) was followed by the apparent extinction of lineages in a size-dependent manner.

Transmission lineage diversity and geographic range

We also characterized 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 overrepresented because of 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 to 0.98) and the number of reported cases (fig. S10; Pearson's $r = 0.53$; 95% CI, 0.35 to 0.67; see also 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 six or seven regions (Fig. 4B; Pearson's $r = 0.8$; 95% CI, 0.78 to 0.82). Thus, bigger regional epidemics comprised a greater diversity of

Fig. 2. Structure of UK transmission lineages detected through genome sampling. (A) Figurative illustration of the international context of UK transmission lineages. Note that 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 to 6 are sampled, and TMRCA B is observed if only genomes 3 to 5 are sampled. (C) Distribution of UK transmission lineage sizes. Blue bars show the number of transmission lineages of each size; error bars are 95% HPDs of these sizes across the posterior tree distribution. The inset shows the corresponding cumulative frequency distribution of lineage size (blue line) on double logarithmic axes; red shading denotes the 95% HPD of this distribution across the posterior tree distribution. Values to either side of the 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 eight 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 eight largest lineages are also shown.

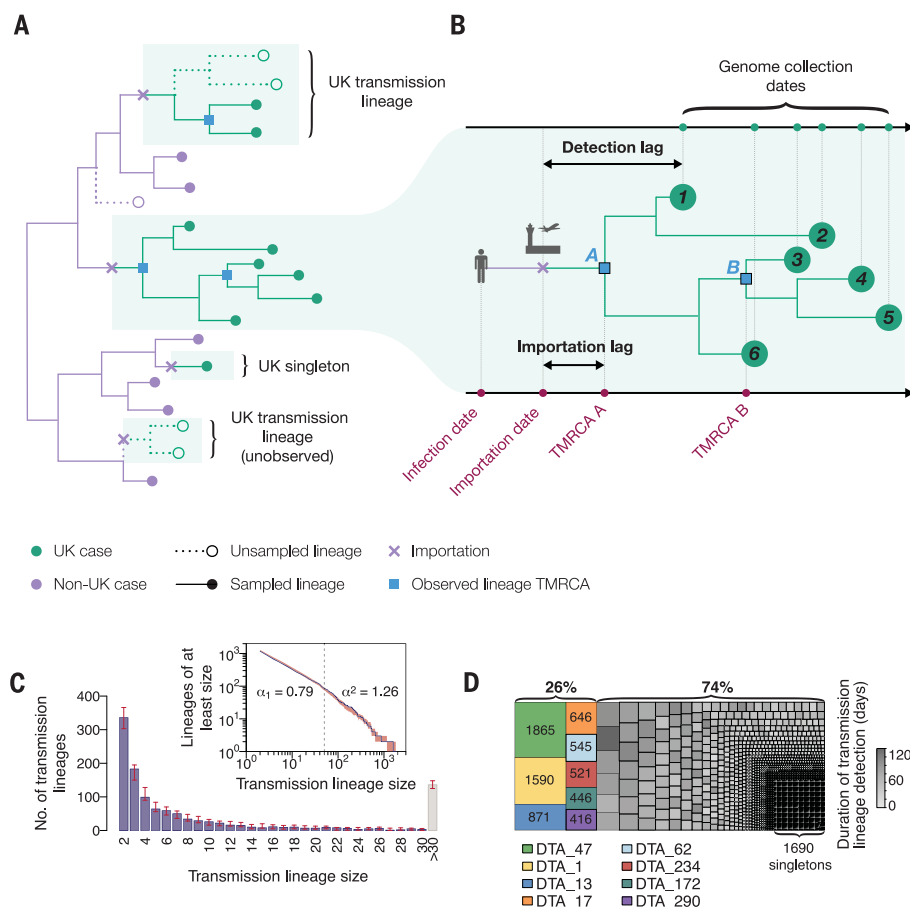


Fig. 3. Dynamics of UK transmission lineages.

(A) Lineage size breakdown of UK genomes collected each week. Colors of the eight 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 denotes transmission lineages that were first sampled less than 1 week ago. The red arrow indicates the start of the UK lockdown. (C) The daily rate of detecting new transmission lineages (red line) and the rate at which lineages have not been observed for >4 weeks (blue line); shading denotes the 95% HPD across the posterior distribution of trees.

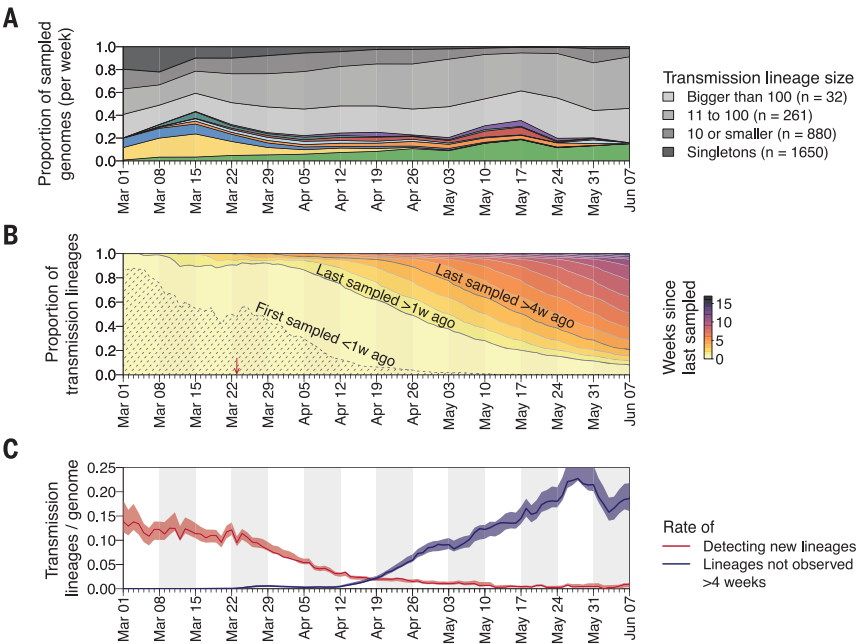


Fig. 4. Spatial distribution of UK transmission lineages.

(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 to 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 to 0.82). (C) Map showing Shannon's index (SI) for each region, calculated across the study period (2 February to 26 June). Yellow colors indicate higher SI values; 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. Histograms (bottom row) show the distribution of geographic distances for all sequence pairs within each lineage (see data S4 and fig. S12 for further details). Colored boxes next to lineage names are as depicted in Fig. 2D.

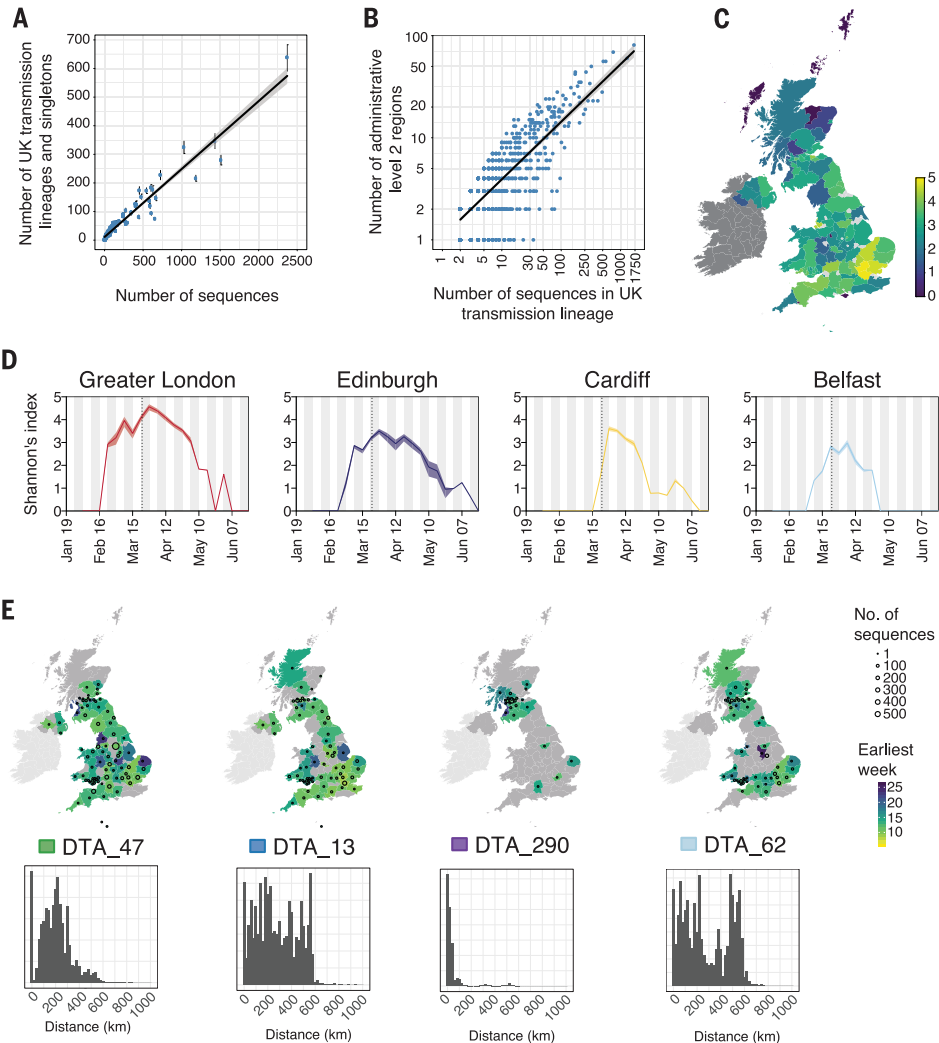


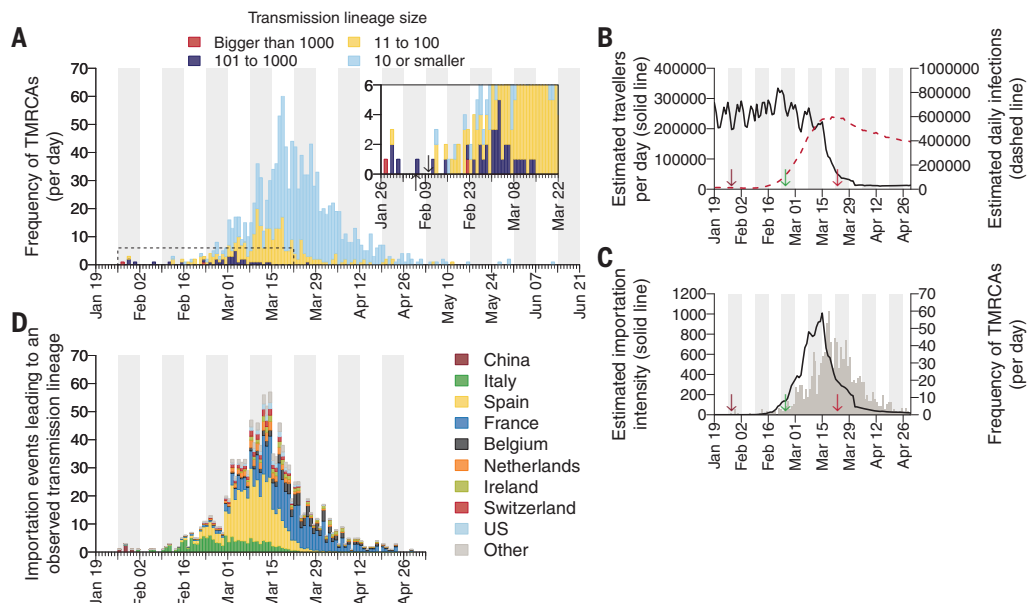
Fig. 5. Dynamics of UK transmission lineage importation. (A) Histogram of lineage TMRCA, colored by lineage size. The inset is an expanded view of the days prior to UK lockdown;

the upward arrow indicates the collection date of the UK's first laboratory-confirmed case, and the downward arrow shows the collection date of the earliest UK virus genome in our dataset.

(B) Estimated number of inbound travelers to the UK per day (black line) and estimated number of infectious cases worldwide (dashed red line).

Arrows show, from left to right, the dates of the first self-isolation advice for returning travelers from China, the same for Italy, and the start of the UK national lockdown. (C) Estimated importation intensity (EII) curve (black) and histogram of lineage TMRCA (gray).

(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).



transmission lineages, and larger lineages were more geographically widespread. These observations indicate substantial dissemination of a subset of lineages across the UK and suggest that 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 characterized by frequent commuter travel to or 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 plotted 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 was characterized 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 >234 km apart), whereas DTA_290 is strongly local (95% of sequence pairs sampled <100 km 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. Because 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 TMRCA of most UK lineages are dated to March and early April [median = 21 March; interquartile range (IQR) = 14 to 29 March]. UK lineages with earlier TMRCA tend to be larger and longer-lived than those whose TMRCA postdate the national lockdown (Fig. 5A and fig. S15).

Because of incomplete sampling, TMRCA best represent the date of the first inferred

transmission event in a lineage, not its importation date (Fig. 2B). To infer the latter and to quantify the delay between importation and onward within-UK transmission, we generated daily estimates of the number of travelers arriving in the UK and of global SARS-CoV-2 infections (17) worldwide. Before March, the UK received ~1.75 million inbound travelers 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). Because 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 peaked 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 TMRCA 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 to 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 to 18) and declined through time (tables S2 and S3).

To investigate country-specific contributions to virus importation, we generated separate 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). The earliest importations were most likely from China or elsewhere in Asia but were rare relative to those from Europe. Over our study period, we infer that ~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. Before the 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 warrant further investigation. We expect that 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. Although 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 the epidemic reproduction number (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, including the effect of specific mutations on lineage growth rates [e.g., (17)], 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

- Centers for Disease Control and Prevention, *MMWR Morb. Mortal. Wkly. Rep.* **52**, 405–411 (2003).
- O. Faye et al., *Lancet Infect. Dis.* **15**, 320–326 (2015).
- K. H. Kim, T. E. Tandi, J. W. Choi, J. M. Moon, M. S. Kim, *J. Hosp. Infect.* **95**, 207–213 (2017).
- J. Bahl et al., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19359–19364 (2011).
- G. J. Baillie et al., *J. Virol.* **86**, 11–18 (2012).
- G. Dudas et al., *Nature* **544**, 309–315 (2017).
- N. D. Grubaugh et al., *Nature* **546**, 401–405 (2017).
- A. F. Y. Poon et al., *Lancet HIV* **3**, e231–e238 (2016).
- J. Thomas et al., *N. Engl. J. Med.* **382**, 632–643 (2020).
- M. A. Beale et al., *Nat. Commun.* **10**, 3255 (2019).
- E. M. Volz et al., *Cell* **184**, 1–12 (2021).
- L. M. Li, N. C. Grassly, C. Fraser, *Mol. Biol. Evol.* **34**, 2982–2995 (2017).
- N. D. Grubaugh et al., *Nat. Microbiol.* **4**, 10–19 (2019).
- Y. Shu, J. McCauley, *Euro Surveill.* **22**, 30494 (2017).
- COVID-19 Genomics UK (COG-UK) Consortium, *Lancet Microbe* **1**, e99–e100 (2020).
- U.K. Government, Coronavirus (COVID-19) in the UK; <https://coronavirus.data.gov.uk/cases>.
- See supplementary materials.
- S. Flaxman et al., *Nature* **584**, 257–261 (2020).
- C. J. Villabona-Arenas, W. P. Hanage, D. C. Tully, *Nat. Microbiol.* **5**, 876–877 (2020).
- M. Worobey et al., *Science* **370**, 564–570 (2020).
- S. A. Nadeau, T. G. Vaughan, J. Sciré, J. S. Huisman, T. Stadler, *medRxiv* 20127738 [preprint], 12 June 2020.
- Welsh Government, Genomic analysis of Covid-19 lineages in Wales; <https://gov.wales/genomic-analysis-covid-19-lineages-wales>.
- Greater London Authority Intelligence and Analysis Unit, “Census Information Scheme: Commuting in London” (2014); <https://londondatastore-upload.s3.amazonaws.com/Zho%3Dttw-flows.pdf>.
- C. Angus, CoVid Plots and Analysis. University of Sheffield (2020); <https://doi.org/10.15131/shef.data.12328226>.
- K. J. Gaston, F. He, *Proc. R. Soc. B* **269**, 1079–1086 (2002).

- G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, *Phys. Life Rev.* **18**, 66–97 (2016).
- C. Baranik, *BMJ* **369**, m1859 (2020).
- J. L. Geoghegan et al., *medRxiv* 20168930 [preprint], 20 August 2020.
- J. Lu et al., *Cell* **181**, 997–1003.e9 (2020).
- T. Seemann et al., *Nat. Commun.* **11**, 4376 (2020).
- C. Dye, R. C. H. Cheng, J. S. Dagpunar, B. G. Williams, *R. Soc. Open Sci.* **7**, 201726 (2020).
- H. Tian et al., *Science* **368**, 638–642 (2020).
- K. Leung, J. T. Wu, D. Liu, G. M. Leung, *Lancet* **395**, 1382–1393 (2020).
- D. C. Adam et al., *Nat. Med.* **26**, 1714–1719 (2020).
- A. Endo, S. Abbott, A. J. Kucharski, S. Funk, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, *Wellcome Open Res.* **5**, 67 (2020).
- L. Wang et al., *Nat. Commun.* **11**, 5006 (2020).
- J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, *Nature* **438**, 355–359 (2005).
- “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).

ACKNOWLEDGMENTS

We are grateful to everyone worldwide involved in generating the virus genome data shared on GISAID. We thank S. Bhatt, P. Lemey, and C. Dye for insightful discussion. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. **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. Also supported by BBSRC grant BB/M010996/1 (V.H.); Fondation Botnar Research Award programme grant 6063 and UK Cystic Fibrosis Trust Innovation Hub Award 001 (C.R.); UKRI GCRF One Health Poultry Hub grant BB/S011269/1 (J.R.); a Branco Weiss Fellowship and EU grant 874850 MOOD (M.U.G.K.); WT fellowship 204311/Z/16/Z and MRC-FAESP awards MR/S0195/1 and 18/14389-0 (N.R.F.); a Branco Weiss Fellowship (T.I.V.); Canadian Institutes for Health Research grant 02179-000 (I.I.B.); WT Collaborators Award 206298/Z/17/Z (J.T.M., R.M.C., N.J.L., and A.R.); ERC grant 725422 (A.R. and E.S.); NIHR Global Health Research Unit grant 16/136/111 (D.M.A.); and the Oxford Martin School (O.G.P., M.U.G.K., L.d.P., and A.E.Z.). **Author contributions:** Study design: L.d.P., J.T.M., M.U.G.K., A.R., O.G.P. Methods development/programming: L.d.P., J.T.M., M.U.G.K., A.R., O.G.P., A.E.Z., V.H., C.R., J.A., R.C., T.C., B.J., N.J.L., A.O., S.N., D.M.A., E.S. Data analysis: L.d.P., M.U.G.K., A.R., O.G.P., A.E.Z., V.H., C.R., D.M.A., J.T.M., A.W., I.I.B., K.K., B.G., K.V.P., E.S., T.I.V. Wrote paper: L.d.P., J.R., M.U.G.K., O.G.P. Edited paper/figure creation: L.d.P., V.H., A.E.Z., M.U.G.K., J.R., C.R., B.G., T.I.V., N.R.F., E.M.V. **Competing interests:** K.K. is the founder of BlueDot, a social enterprise that develops digital technologies for public health. A.W. and I.I.B. 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 from the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB37886 (www.ebi.ac.uk/ena/browser/view/PRJEB37886). Non-UK genomes were obtained from gisaid.org. Raw data, code, and analysis files for this work are in supplementary materials or available from GitHub at <https://github.com/COG-UK/uk-intros-analyses> (DOI: 10.5281/zenodo.4311597), which also contains a list of sequence accession numbers. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6530/708/suppl/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S5
Data S1 to S5
References (39–68)

22 October 2020; accepted 18 December 2020
Published online 8 January 2021
10.1126/science.abr2946

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

Louis du Plessis, John T. McCrone, Alexander E. Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghvani, Jordan Ashworth, Rachel Colquhoun, Thomas R. Connor, Nuno R. Faria, Ben Jackson, Nicholas J. Loman, Aine O'Toole, Samuel M. Nicholls, Kris V. Parag, Emily Scher, Tetyana I. Vasylyeva, Erik M. Volz, Alexander Watts, Isaac I. Bogoch, Kamran Khan, COVID-19 Genomics UK (COG-UK) Consortium, David M. Aanensen, Moritz U. G. Kraemer, Andrew Rambaut and Oliver G. Pybus

Science **371** (6530), 708-712.
DOI: 10.1126/science.abf2946originally published online January 8, 2021

Lineage dynamics

The scale of genome-sequencing efforts for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is unprecedented. The United Kingdom has contributed more than 26,000 sequences to this effort. This volume of data allowed du Plessis *et al.* to develop a detailed picture of the influxes of virus reaching U.K. shores as the pandemic developed during the first months of 2020 (see the Perspective by Nelson). Before lockdown, high travel volumes and few restrictions on international travel allowed more than 1000 lineages to become established. This accelerated local epidemic growth and exceeded contact tracing capacity. The authors were able to quantify the abundance, size distribution, and spatial range of the lineages that were transmitted. Transmission was highly heterogeneous, favoring some lineages that became widespread and subsequently harder to eliminate. This dire history indicates that rapid or even preemptive responses should have been used as they were elsewhere where containment was successful.

Science, this issue p. 708; see also p. 680

ARTICLE TOOLS

<http://science.sciencemag.org/content/371/6530/708>

SUPPLEMENTARY MATERIALS

<http://science.sciencemag.org/content/suppl/2021/01/07/science.abf2946.DC1>

RELATED CONTENT

<http://stm.sciencemag.org/content/scitransmed/12/559/eabc3103.full>
<http://stm.sciencemag.org/content/scitransmed/12/573/eabe2555.full>
<http://stm.sciencemag.org/content/scitransmed/12/564/eabd5487.full>
<http://stm.sciencemag.org/content/scitransmed/12/556/eabc7075.full>
<http://science.sciencemag.org/content/sci/371/6530/680.full>

REFERENCES

This article cites 63 articles, 6 of which you can access for free
<http://science.sciencemag.org/content/371/6530/708#BIBL>

PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works