Article

Untangling introductions and persistence in COVID-19 resurgence in Europe

https://doi.org/10.1038/s41586-021-03754-2

Received: 4 February 2021

Accepted: 22 June 2021

Published online: 30 June 2021

Check for updates

Philippe Lemey[1,2]✉, Nick Ruktanonchai[3,4], Samuel L. Hong[1], Vittoria Colizza[5], Chiara Poletto[5], Frederik Van den Broeck[1,6], Mandev S. Gill[1], Xiang Ji[7], Anthony Levasseur[8], Bas B. Oude Munnink[9], Marion Koopmans[9], Adam Sadilek[10], Shengjie Lai[3], Andrew J. Tatem[3], Guy Baele[1], Marc A. Suchard[11,12,13] & Simon Dellicour[1,14]✉

After the first wave of SARS-CoV-2 infections in spring 2020, Europe experienced a resurgence of the virus starting in late summer 2020 that was deadlier and more difficult to contain[1]. Relaxed intervention measures and summer travel have been implicated as drivers of the second wave[2]. Here we build a phylogeographical model to evaluate how newly introduced lineages, as opposed to the rekindling of persistent lineages, contributed to the resurgence of COVID-19 in Europe. We inform this model using genomic, mobility and epidemiological data from 10 European countries and estimate that in many countries more than half of the lineages circulating in late summer resulted from new introductions since 15 June 2020. The success in onward transmission of newly introduced lineages was negatively associated with the local incidence of COVID-19 during this period. The pervasive spread of variants in summer 2020 highlights the threat of viral dissemination when restrictions are lifted, and this needs to be carefully considered in strategies to control the current spread of variants that are more transmissible and/or evade immunity. Our findings indicate that more effective and coordinated measures are required to contain the spread through cross-border travel even as vaccination is reducing disease burden.

Upon successfully curbing transmission in spring 2020, many European countries witnessed a resurgence in cases of COVID-19 in the late summer. The number of COVID-19 infections increased rapidly, and by the end of October, it was clear that the continent was deep into a second epidemic wave. This forced governments to reimpose lockdowns and social restrictions in an effort to contain the resurgence. Although these measures reduced infection rates across Europe[3], several countries witnessed a stabilization at high levels or even a new surge in infections. The spread of more transmissible variants, in particular B.1.1.7 (Alpha variant or 20I (V1)), which was first identified in the UK[4], has considerably exacerbated the challenge to contain COVID-19.

Already early on in the pandemic, modelling studies warned about new waves due to partial relaxation of restrictions[5] or seasonal variations[6]. By mid-April, the European Commission constructed a roadmap to lifting coronavirus containment measures[7], recommending a cautious and coordinated manner to revive social and economic activities. However, the early start of the devastating second wave demonstrated that there was insufficient adherence to these measured recommendations. Cross-border travel, and mass tourism in particular, has been implicated as a major instigator of the second wave. Genomic surveillance demonstrated that a new variant (lineage B.1.177[8], 20E (EU1) (https://nextstrain.org/), which emerged in Spain in early summer, has spread to multiple locations in Europe[2]. Although this variant quickly grew into the dominant circulating SARS-CoV-2 strain in several countries, it did not appear to be associated with a higher intrinsic transmissibility[2].

Although it appears clear that travel considerably contributed to the second wave in Europe, it remains challenging to assess how it may have restructured and reignited the epidemic in the different European countries. Even without resuming travel, relaxing containment measures when low-level transmission is ongoing risks the proliferation of locally circulating strains. Phylodynamic analyses may provide insights into the relative importance of persistence versus the introduction of new lineages, but such analyses are complicated for SARS-CoV-2 for different reasons. Phylogenetic reconstructions may be poorly resolved owing to the relatively limited SARS-CoV-2 sequence diversity[9]. This is further confounded by the degree of genetic mixing that can be expected from unrestricted travel before the lockdowns in spring 2020.

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [2]Global Virus Network (GVN), Baltimore, MD, USA. [3]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, UK. [4]Population Health Sciences, Virginia Tech, Blacksburg, VA, USA. [5]INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique IPLESP, Paris, France. [6]Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium. [7]Department of Mathematics, School of Science & Engineering, Tulane University, New Orleans, LA, USA. [8]UMR MEPHI (Microbes, Evolution, Phylogeny and Infections), Aix-Marseille Université (AMU) and Institut Universitaire de France (IUF), Marseille, France. [9]Department of Viroscience, WHO Collaborating Centre for Arbovirus and Viral Hemorrhagic Fever Reference and Research, Erasmus MC, Rotterdam, The Netherlands. [10]Google, Mountain View, CA, USA. [11]Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. [12]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, USA. [13]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. [14]Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Bruxelles, Belgium. ✉e-mail: philippe.lemey@kuleuven.be; simon.dellicour@ulb.ac.be

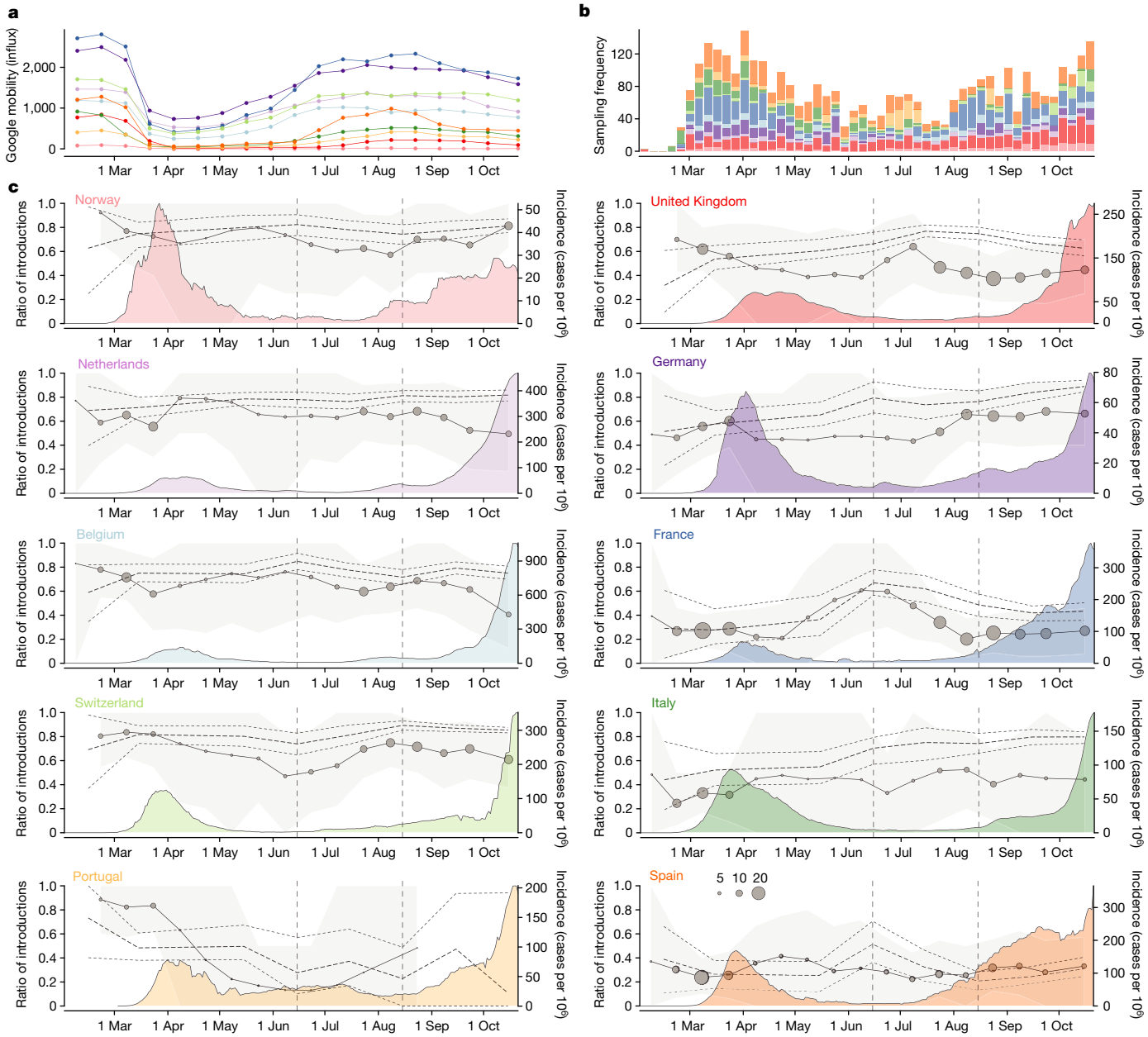


Fig. 1 | Mobility, genome sampling, case counts and phylogeographical summaries through time for 10 European countries. a, The country-specific Google mobility influx in the 10 countries during 2-week intervals. **b**, The weekly genome sampling by country used in the phylogeographical analysis. **c**, For each country, the ratio of introductions over the total viral flow from and to that country (in 2-week intervals) and a monthly normalized entropy measure summarizing the phylogenetic structure of country-specific transmission chains are shown. The posterior mean ratios of introductions are depicted with circles that have a size proportional to the total number of transitions from and to that country and the grey surface represents the 95% highest posterior density (HPD) intervals. The posterior mean normalized entropies and 95% HPD intervals are depicted by dotted lines. These normalized entropy measures indicate how phylogenetically structured the epidemic is in each country, and ranges from 0 (perfectly structured, for example, a single country-specific cluster) to 1 (unstructured interspersion of country-specific sequences across the entire SARS-CoV-2 phylogeny). The introduction ratios and normalized entropy measures are superimposed on the incidence of COVID-19 (daily cases per 10^6 people) reported for each country through time (coloured density plot). The two vertical dashed lines represent the summer time interval (15 June and 15 August 2020) for which we subsequently evaluate introductions versus persistence (see Fig. 2).

Mobility data predicts SARS-CoV-2 spread

We analysed SARS-CoV-2 B.1 (20A) genomes from 10 European countries for which a minimal number of genomes from the second wave were already available on 3 November 2020. Using a two-step procedure that relied on subsampling relative to country-specific case counts (see Methods), we compiled a dataset of close to 4,000 genomes sampled between 29 January and 31 October 2020 (Extended Data Table 1). To achieve maximum resolution in our evolutionary reconstructions, we constructed a Bayesian time-measured phylogeographical model that integrates mobility and epidemiological data. Our approach simultaneously infers phylogenetic history and ancestral movement throughout this history while also identifying the drivers of spatial spread[10]. We used the latter functionality to determine the most appropriate mobility or connectivity measure. Specifically, we considered international air transportation data, the Google COVID-19 Aggregated Mobility Research Dataset (also referred to here as 'mobility data'), and

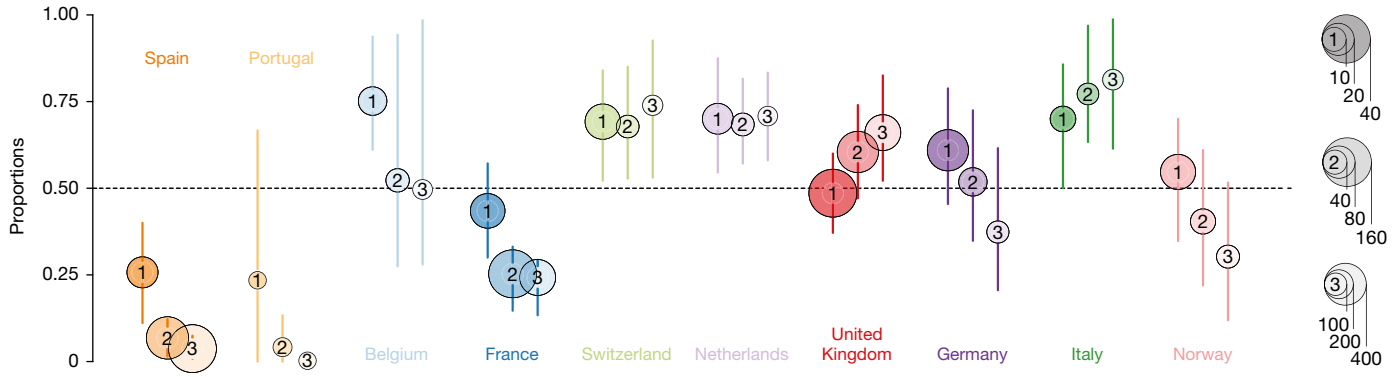


Fig. 2 | Posterior estimates for the relative importance of lineage introduction events in 10 European countries. We report three summaries (posterior mean and 95% HPD intervals) for each country: the ratio of unique introductions over the total number of unique persisting lineages and unique introductions between 15 June and 15 August 2020 (1), the ratio of descendant lineages from these unique introduction events over the total number of descendants circulating on 15 August 2020 (2) and the ratio of descendant taxa from these unique introductions over the total number of descendant taxa sampled after 15 August 2020 (3) (see Extended Data Fig. 4). The dots are numbered and the sizes are proportional to: (1) the total number of unique lineage introductions identified between 15 June and 15 August 2020; (2) the total number of lineages inferred on 15 August 2020; and (3) the total number of descendant tips after 15 August 2020.

the social connectedness index of Facebook, as covariates of phylogeographical spread (Extended Data Fig. 1). The Google mobility dataset contains anonymized mobility flows aggregated over users who have turned on the location history setting, which is off by default (see Methods). The social connectedness index reflects the structure of social networks and has been suggested to correlate with the geographical spread of COVID-19[11]. To help to inform the phylogenetic coalescent time distribution, we parameterized the viral population size trajectories through time as a function of epidemiological case count data for the countries under investigation.

Analyses using both time-homogeneous and time-inhomogeneous models offered strong support for mobility data as a predictor of spatial diffusion whereas air transportation data and the social connectedness index offered no predictive value (Extended Data Table 2). The fact that mobility data encompassing both air and land-based transport are required to explain COVID-19 spread highlights the need to consider both types of transport in containment strategies. To ensure that containment strategies were accommodated by our reconstructions, we further extended our time-inhomogeneous approach to model biweekly variation in the overall rate of spread between countries as a function of mobility (see Methods and Extended Data Table 2).

Dynamic viral transmission through time

We use our probabilistic model of spatial spread informed by genomic, mobility and epidemiological data to characterize the dynamics of spread throughout the epidemic in Europe. We first focus on the ratio of introductions over the total viral flow in and out of each country over time and the genetic structure of country-specific transmission chains (Fig. 1). For the latter, we use a normalized entropy measure that quantifies the degree of phylogenetic interspersion of country-specific transmission chains in the SARS-CoV-2 phylogeny (see Methods). Although estimates for individual dispersal between pairs of countries can also be obtained (Extended Data Fig. 2), we remain cautious in interpreting these as direct pathways of spread because the genome sampling only covers a restricted set of European countries. The mobility to and from each country within our 10-country sample covers between 64% and 96% of the mobility of these countries to and from all countries within Europe (Extended Data Table 3 and Extended Data Fig. 3), except for Norway (27%), for which other Scandinavian countries account for considerable mobility connections (61%), and the UK (49%), for which Ireland accounts for a large fraction of mobility connections (38%).

According to the proportion of introductions, we estimate more viral import than export events for Switzerland, Norway, the Netherlands and Belgium throughout most of the time period under investigation. According to the estimated phylogenetic entropy, these countries also experienced many independent transmission chains since the epidemic started to unfold. This is consistent with country-specific studies; for the first wave in Belgium, for example, about 331 individual introductions were estimated in the ancestry of a limited sample of 740 genomes[12]. For Portugal, we also estimate higher proportions of introductions early in the first wave but with a subsequent decline to predominantly export events. France, Italy and Spain, on the other hand, are characterized by a relatively high viral export during the first wave. The proportion of introductions remained relatively low for Italy and Spain after the first wave, whereas in France these proportions were high from mid-June until the end of July. However, the absolute number of transitions in our sample are low during this time period. These countries also had comparatively lower entropy values early in the epidemic, with an increase for France by the start of summer and a more gradual increase over time for Italy. In Spain, however, the genetic complexity of the SARS-CoV-2 transmission chains remained limited. In the UK and Germany, the viral flow in and out of the country was initially relatively balanced. A recent large-scale genomic analysis in the UK indicates that this can imply very high absolute numbers of cross-country transmissions, as more than 2,800 independent introduction events were identified from the analysis of 26,181 genomes[13]. Although our sample is limited compared to this UK-focused analysis[13], our reconstructions also recover major influx from Spain, France and Italy during the first wave in the UK (Extended Data Fig. 2). We estimate an increase in the proportion of introductions for the UK from mid-June, indicating an important viral import relative to export around this time. The phylogenetic entropy also peaked around this time. In Germany, the proportions increased slightly later in the summer with a concomitant rise in phylogenetic entropy.

Introductions thrive in low incidence

To assess the effect of summer travel on the second wave in the different countries, we use our genomic–mobility reconstruction to estimate both the number of lineages persisting in each country and the number of newly introduced lineages, and how these proliferated early in the second wave. We focus on a 2-month time period between 15 June 2020—when many EU and Schengen-area countries opened

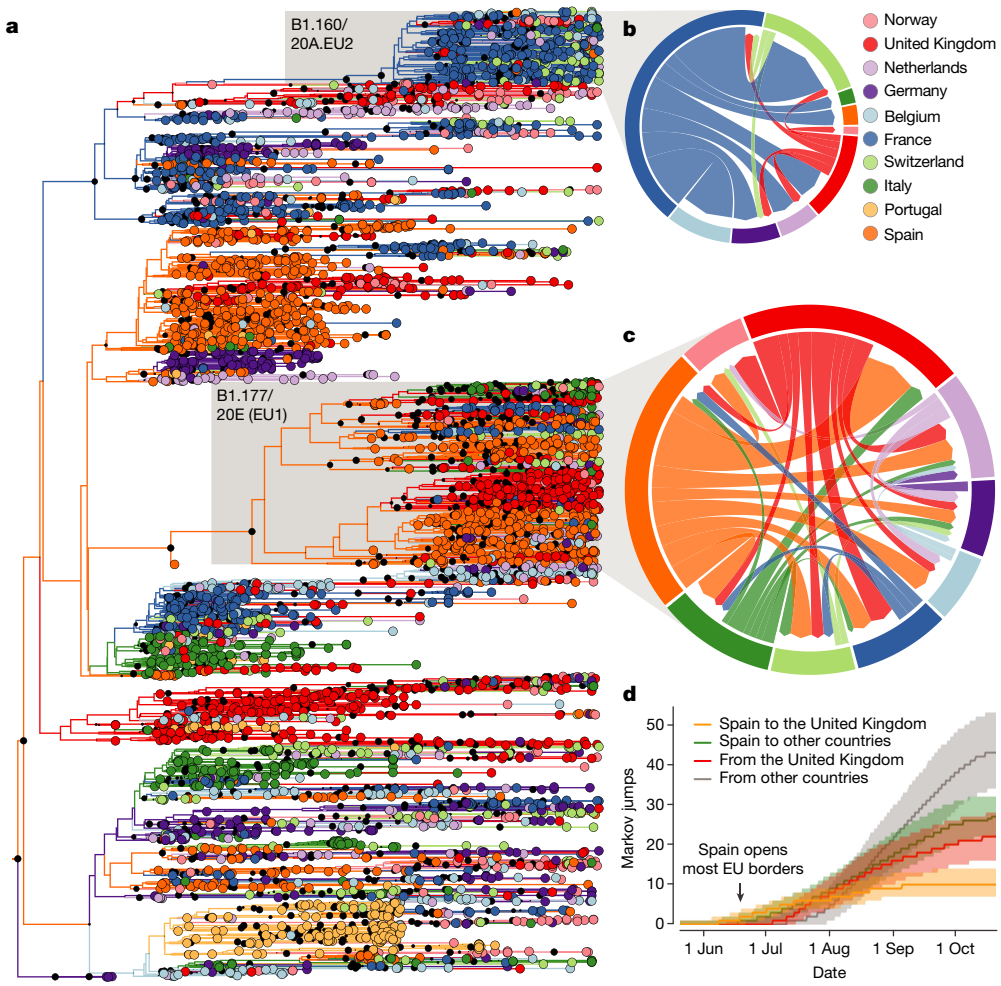


Fig. 3 | Phylogeographical estimates of SARS-CoV-2 spread in 10 European countries. a, The maximum clade credibility tree summary of the Bayesian inference. Colours correspond to the countries in the legend. The two clades corresponding to B1.160/20A.EU2 and B1.177/20E (EU1) are highlighted in grey. **b**, **c**, Circular migration flow plots for B1.160/20A.EU2 (**b**) and B1.177/20E (EU1) (**c**) based on the posterior expectations of the Markov jumps. In these plots, migration flow out of a particular location starts close to the outer ring and ends with an arrowhead more distant from the destination location. **d**, Posterior mean estimates with 95% HPD intervals over time for four types of Markov jumps for B1.177/20E (EU1): from Spain to the UK, from Spain to other countries, from the UK and from other countries.

their borders to other countries—and 15 August 2020, before which the majority of holiday return travel is expected for many countries. We identify the number of lineages circulating in each country on 15 August and determine whether they result from a lineage that persisted since 15 June or from a unique introduction after this date (independent of the number of descendants for this lineage on 15 August; Extended Data Fig. 4). In Fig. 2, we plot (1) the ratio of these unique introductions over the total unique lineages (unique introductions and persisting lineages); (2) the proportion of descendant lineages on 15 August that resulted from the unique introductions over the total descendants circulating on this date; and (3) the proportion of descendant tips (sampled genomes) after 15 August that resulted from the unique introductions over the total number of descendant tips (see Methods and Extended Data Fig. 4). We estimate a posterior mean proportion of unique introductions that is close to or higher than 0.5 except for Spain and Portugal. This indicates that by 15 August, a relatively large fraction of circulating lineages in each country was produced by new introductions over summer. Because the B.1.177/20E (EU1) variant that was predominantly disseminated through summer travel does not appear to be particularly more transmissible[2], this is unlikely to be due to strong intrinsic advantages of the newly introduced viruses.

The two proportions of descendants from these introductions on 15 August and after this date measure the relative success of newly introduced lineages compared to persisting lineages, indicating considerable variation in onward transmission. In Fig. 2, the country estimates are ordered according to decreasing average incidence during the 15 June–15 August time period, suggesting that incidence may shape the outcome of the introductions. In countries that experienced relatively high summer incidence (for example, Spain, Portugal, Belgium and France), the introductions lead to comparatively fewer descendants on 15 August or after. We find a significant overall association between the incidence and the difference in the logit-scaled proportion of unique introductions and the logit-scaled proportion of their descendants on 15 August ($P = 0.007$) as well as between the incidence and the difference in the logit-scaled proportion of unique introductions and the logit-scaled proportion of descendant tips after 15 August ($P = 0.019$) (Extended Data Fig. 5). With comparatively few descendants from introductions (Fig. 2), Norway may to some extent be an outlier because lineages estimated as persisting in this country could in fact be introductions from other Scandinavian countries that are not represented in our genome sample. We recover qualitatively similar, but more variable and statistically unsupported associations between the success of introductions and incidence for the 2-month time periods before and after the 15 June–15 August time period (Extended Data Fig. 5). This indicates that the comparatively higher proportion of introductions as well as the more stable and lower incidence between 15 June and

15 August provided the ideal conditions for a process of genetic drift by which introductions were able to fuel transmission.

Our estimates show that introductions in the UK particularly benefited from the conditions for successful onward transmission (Fig. 2), with a considerable fraction of introductions originating from Spain (Extended Data Fig. 6), reflecting the spread of B.1.177/20E (EU1), which rapidly became the most dominant strain in the UK[2]. Our analysis captures the expansion of this variant as well as that of B.1.160/20A.EU2, which together account for more than 25% of the genomes in our dataset. Although Spain was indeed inferred to be the origin of B.1.177/20E (EU1), the UK also considerably contributed to its spread (Fig. 3). The earliest introduction from Spain to the UK was estimated around the time Spain opened most EU borders (21 June) (Fig. 3). Although introductions from Spain to other countries soon followed, we estimate a similar rate and amount of spread from the UK to other countries before these other countries also further disseminated the virus. Although inferred from a limited sample, this illustrates a dynamic pattern of spread and the importance of the early establishment of B.1.177/20E (EU1) in the UK that probably served as an important secondary centre of dissemination. We note, however, that this pattern may be affected by the intensive and continuous genomic surveillance in the UK, which may also be reflected in our subsample of the available data. Although the UK is also involved in the spread of B1.160/20A.EU2, this variant has been largely disseminated from France (Fig. 3). The fact that this variant expanded later in France and subsequently also started to spread later compared to B.1.177/20E (EU1) (Extended Data Fig. 7) may explain why the latter spread more successfully.

Discussion

Our Bayesian phylogeographical approach builds on a rich history of identifying drivers of spatial spread, with applications to various pathogens at different spatial scales, ranging from air transportation for influenza at a global scale[10] to gravity model transmission for Ebola in West Africa[14]. Such studies use a relatively limited genomic sample to gain insights into viral transmission dynamics. This is also the case in our application to SARS-CoV-2 in Europe for which we further extend the phylodynamic data integration approach to confront the lack of resolution offered by SARS-CoV-2 genomic data. A concerted effort in containing international spread further sets apart the COVID-19 pandemic from these earlier events. For this reason, we have now incorporated variation in mobility over time to account for the effect of these measures. Our reconstructions show that the composition of lineages circulating towards the end of the summer was to an important extent shaped by introductions in most of the European countries. The relative success of onward transmission of the introduced lineages appears to be shaped by the local incidence of COVID-19 during summer.

Our results should be interpreted in light of several important limitations. In addition to a limited overall size, the genome data only cover a selection of European countries, suggesting that we are missing transmission events that involve unsampled countries. This may be important for Norway, for example, which according to our mobility data, is largely connected to other Scandinavian countries. We also lack sampling from eastern Europe, which was to a large extent spared by border controls and lockdowns during the first wave, but witnessed high excess mortality rates during the second wave. The emergence of more transmissible variants has led to more intensified genomic surveillance, so similar phylodynamic reconstructions may now be performed on a wider scale.

The pandemic exit strategy offered by vaccination programmes is a source of optimism that also sparked proposals by EU member states to issue vaccine passports in a bid to revive travel and rekindle the economy. In addition to implementation challenges and issues of fairness, there are risks associated with such strategies when immunization is incomplete, as probably will be the case for the European population this summer. A recent modelling study for the UK suggests that vaccination in adults alone is unlikely to completely halt the spread of cases of COVID-19 and that lifting containment measures early and suddenly can lead to a large wave of infections[15]. A gradual release of restrictions was shown to be critical for minimizing the infection burden[15]. We believe that travel policies may be a key consideration in this respect because similar conditions may arise to the ones that we demonstrated to provide fertile ground for viral dissemination and resurgence in 2020. This may now also involve the spread of variants that are more transmissible and/or evade the immune responses triggered by vaccines and previous infections. Well-coordinated European strategies will therefore be required to manage the spread of SARS-CoV-2 and reduce future waves of infection, with hopefully a more unified implementation than hitherto observed.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03754-2.

1. European Centre for Disease Prevention and Control. *Data on 14-Day Notification Rate of New COVID-19 Cases and Deaths* (2021); https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19
2. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* https://doi.org/10.1038/s41586-021-03677-y (2021).
3. European Centre for Disease Prevention and Control. *COVID-19 Situation Update for the EU/EEA, as of Week 3, Updated 28 January 2021* (2021); https://www.ecdc.europa.eu/en/cases-2019-ncov-eueea
4. Rambaut, A. et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).
5. Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y. & Colizza, V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Med*. **18**, 240 (2020).
6. Neher, R. A., Dyrdak, R., Druelle, V., Hodcroft, E. B. & Albert, J. Potential impact of seasonal forcing on a SARS-CoV-2 pandemic. *Swiss Med. Wkly*. **150**, w20224 (2020).
7. McKee, M. A European roadmap out of the COVID-19 pandemic. *Br. Med. J*. **369**, m1556 (2020).
8. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol*. **5**, 1403–1407 (2020).
9. Morel, B. et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol*. **38**, 1777–1791 (2020).
10. Lemey, P. et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog*. **10**, e1003932 (2014).
11. Kuchler, T., Russel, D. & Stroebel, J. *The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook*, NBER Working Paper 26990 (National Bureau of Economic Research, 2020).
12. Dellicour, S. et al. A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Mol. Biol. Evol*. **38**, 1608–1613 (2021).
13. du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
14. Dudas, G. et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
15. Moore, S., Hill, E. M., Tildesley, M. J., Dyson, L. & Keeling, M. J. Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. *Lancet Infect. Dis*. **21**, 793–802 (2021).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sequence data and subsampling

We used a two-step genome data collection procedure. We first evaluated the available genomes from European countries in GISAID[16] on 3 November 2020. We selected genomes from Belgium, France, Germany, Italy, the Netherlands, Norway, Portugal, Spain, Switzerland and the UK primarily based on the availability of genome data from both the first and second wave at that time but also because of their high ratio of genomes to positive cases. A total of 39,812 genomes were available for these countries on 3 November 2020; the available number of genomes per country is listed in Extended Data Table 1. Portugal represented an exception because data for this country were limited to the first wave at that time, but we included genomes from Portugal because of its potential importance as a summer travel location.

We aligned the genomes from each country using MAFFT v.7.453[17] and trimmed the 5′ and 3′ ends and only retained unique sequences from each location. To further mitigate the disparities in sampling, we subsampled each country proportionally to the cumulative number of cases on 21 October 2020 (the most recently sampled sequence at the time) by setting an arbitrary threshold of 6.5 sequences per 10,000 cases, with a minimum number of 100 sequences per country. To maximize the temporal and spatial coverage in each country, we binned genomes by epi-week and sampled as evenly as possible, sampling from a different region within the country when available. Only sequences from the B.1 lineage with the D614G substitution and exact sampling dates were selected for the analyses. From the final aligned sequence set, we removed 12 potential outliers, based on a root-to-tip regression applying TempEst v.1.5.3[18] to a maximum-likelihood tree inferred with IQTREE v.2.0.3[19], yielding a dataset of 2,909 genomes (Extended Data Table 1).

Because of the nature of genome sequence accumulation, fewer recently sampled genomes were available for most countries on 3 November 2020 (relative to the case counts at this time). Because our primary goal was to assess the persistence and introduction of lineages leading up to the second wave, we sought to augment our dataset with more recent genomes, having already performed analyses on the initial dataset. In the section on Bayesian evolutionary reconstructions, we outline how we updated these analyses accordingly. On 5 January 2021, we updated our dataset by adding more than 1,000 non-identical sequences collected between 1 August and 31 October (out of a total of 56,395 available genomes; the available and selected numbers of genomes per country are listed in Extended Data Table 1). For Portugal, we extended this period back to 22 June (the most recent sampling date for the previous Portuguese selection). We downloaded all new B.1 sequences with the D614G substitution collected during the selected time period from GISAID and performed the following subsampling. The number of genomes to add per country was obtained by raising the threshold ratio of sequences/cases to 8.5 and increasing the minimum number of sequences to 200. To bias the temporal coverage towards more recent samples, the genomes from each country were binned by week and sampled such that the number of sequences added per week was proportional to an exponential function of the form $e^{t/4}$, where $t = 0$ represents 1 August and $t = 13$ is 31 October. For Portugal, we did not use this preferential sampling as we needed to include close to all available genomes to increase the number of genomes to 200. The selected sequences were deduplicated and outliers were removed as described in the previous paragraph. With the additional selection of 1,050 genomes, we obtained a dataset of 3,959 genomes (Extended Data Table 1).

Mobility data

We analysed four different mobility and connectivity measures: air traffic flows, a social connectedness index provided by Facebook, as well as aggregated Facebook[20] and Google international mobility data. Air traffic flow data were obtained from the International Air Transport Association (http://www.iata.org) and based on the number of origin–destination tickets while also taking into account connections at intermediate airports[21]. We used monthly air traffic data between the 10 European countries under investigation for the time period between January 2020 and October 2020. The social connectedness index (SCI) is an anonymized snapshot of active Facebook users and their friendship networks to measure the intensity of social connectedness between countries (https://data.humdata.org/)[22]. In practice, the SCI measures the relative probability of a Facebook friendship link between two users of the application in different countries. We used the SCI calculated for the 10 European countries represented in our genomic sample as of August 2020.

The Google COVID-19 Aggregated Mobility Research Dataset contains anonymized mobility flows aggregated over users who have turned on the location history setting (on a range of platforms[23]), which is off by default. To produce this dataset, machine learning is applied to logs data to automatically segment it into semantic trips[24]. To provide strong privacy guarantees, all trips were anonymized and aggregated using a differentially private mechanism[25] to aggregate flows over time (see https://policies.google.com/technologies/anonymization). This research was done on the resulting heavily aggregated and differentially private data. No individual user data was ever manually inspected, only heavily aggregated flows of large populations were handled. All anonymized trips were processed in aggregate to extract their origin and destination location and time. For example, if users travelled from location a to location b within time interval t, the corresponding cell (a, b, t) in the tensor would be $n \pm \eta$, where η is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero-mean Laplace distribution and yields (ε, δ)-differential privacy guarantee of $\varepsilon = 0.66$ and $\delta = 2.1 \times 10^{-29}$ per metric. Specifically, for each week W and each location pair (A,B), we compute the number of unique users who took a trip from location A to location B during week W. To each of these metrics, we add Laplace noise from a zero-mean distribution of scale 1/0.66. The parameter ε controls the noise intensity in terms of its variance and δ represents the deviation from pure ε privacy. The closer these parameters are to zero, the stronger the privacy guarantees. We used aggregated mobility flows between the 10 European countries and summarized them by 2-week or monthly time periods between January 2020 and October 2020.

Finally, we also considered international mobility data from Facebook mobility data as an alternative to Google mobility data. These data are based on the numbers of Facebook users moving over large distances, such as air or train travel. Counts of international travel patterns are updated daily based only on users who have opted to share precise location data from their device with the Facebook mobile app through location services. Also in this case, we used aggregated mobility flows between the 10 European countries and summarized them by month between January 2020 and October 2020. Because international aggregate mobility data obtained from Google and Facebook are highly correlated (monthly Spearman correlation ranging from 0.84 to 0.92) (Supplementary Fig. 1), we only included the Google aggregate mobility data as a covariate in the phylogeographical analyses. We note that the mobility data are subject to limitations as these may not be representative of the population as whole and their representativeness may vary by location.

Bayesian evolutionary reconstructions

Joint sequence–trait inference with a time-homogeneous generalized linear model of discrete trait diffusion. We performed a Bayesian evolutionary reconstruction of timed phylogeographical history using BEAST 1.10[26], incorporating the genome sequences, their

country and date of sampling, epidemiological and mobility and/or connectivity data. Because of the relatively low degree of resolution offered by the sequence data, our full probabilistic model specification focuses on (1) relatively simple model specifications and (2) informing parameters by additional non-genetic data sources. We modelled sequence evolution using an HKY85 nucleotide substitution model with a gamma-distributed rate variation among sites and a strict molecular clock model. Our genome set includes three genomes from an early outbreak in Bavaria, which was caused by an independent introduction from China[27,28]. We therefore constrained these genomes as an outgroup in the analysis, which according to root-to-tip regression plots as a function of sampling time resulted in a better correlation coefficient and R^2 compared to the best-fitting root under the heuristic mean residual squared criterion[18] (Supplementary Fig. 2).

As a coalescent tree prior, we modelled the effective population size trajectory as a piecewise constant function that changes values at pre-specified times (following a previously published study[29]), with log-transformed population sizes modelled as a deterministic function of log-transformed counts of cases of COVID-19 (following a previous publication[30]). This reduces the nonparametric skygrid parameterization to a generalized linear model (GLM) formulation with an estimable regression intercept (α) and coefficient (β). In this parameterization, a coefficient estimate centred around 0 would imply constant population size dynamics through time. We specified 2-week intervals and summarized as a covariate the total case counts over these time intervals for the 10 countries of sampling (obtained from https://www.ecdc.europa.eu/en/covid-19/data). The earliest interval with non-zero cases counts was from 14 January 2020 to 28 January 2020; before 14 January 2020, the log-transformed and standardized case count covariate was set to the equivalent of 1 case. We also tested whether a lag time was required for the case count covariate using marginal likelihood estimation[31]. Specifically, we shifted the case counts by 1, 2, 3 and 4 weeks before summarizing them according to 2-week intervals and estimated the model fit of these covariates against case counts without lag time (Supplementary Table 1). To mitigate the computational burden associated with the marginal likelihood estimation procedure, we performed these analyses on a subset of 1,000 genomes (obtained using the Phylogenetic Diversity Analyzer tool[32]). We estimated the highest (log-transformed) marginal likelihood for a two-week lag time (Supplementary Table 1) and used this for the case count covariate in our analyses.

Similar to sequence evolution, we modelled the process of transitioning through discrete location states (countries of sampling) according to a continuous-time Markov chain (CTMC)[33]. We used a parameterization that models the log-transformed transition rates as a log-linear function of mobility and connectivity covariates[10]. The Bayesian implementation of this model simultaneously estimates the phylogenetic history, ancestral movement and the contribution of covariates to the movement patterns[10]. Although we mainly use this approach to obtain well-informed phylodynamic estimates, we also make use of its capacity to identify the most-relevant mobility measure to inform our reconstructions. As covariates we considered the SCI of Facebook, air transportation data and mobility data. For the two time-variable mobility measures, we used the average of the log-transformed and standardized monthly mobility measures as a single covariate in our time-homogeneous phylogeographical GLM model. In this GLM formulation, we estimate the positive effect sizes for each covariate as well as their inclusion probability through a spike-and-slab procedure[10]. Although we subsampled the number of SARS-CoV-2 genomes per country in proportion to case counts, they do not fully correspond because we used a minimum number of genomes for countries with low case counts. We therefore evaluated whether this resulted in signal for sampling bias by including an origin and destination covariate in the GLM based on the residuals for a regression analysis between genomes and case counts (following a previously published study[14]). We performed this analysis using a set of empirical trees (see 'Time-inhomogeneous reconstructions') applying both a time-homogeneous and time-inhomogeneous model, but found no support for these additional covariates (Supplementary Table 2).

We performed inference under the full model specification using Markov chain Monte Carlo (MCMC) sampling and used the BEAGLE library v.3[34] to increase computational performance. We specified standard transition kernels on all parameters, except for the regression coefficients of the piecewise-constant coalescent GLM model. For these parameters, we implemented new Hamiltonian Monte Carlo transition kernels to improve sampling efficiency. These kernels use principles from Hamiltonian dynamics and their approximate energy conserving properties to reduce correlation between successive sampled states, but require computation of the gradient of the model log-posterior with respect to the parameters of interest, in addition to efficient evaluation of the log-posterior that BEAGLE provides. To accomplish this, we extended our previous analytic derivation of the gradient of the log-transformed density from the skygrid coalescent model with respect to the log-transformed population sizes[35] to now be with respect to the regression coefficients using the chain rule and their regression design matrix.

Owing to the dataset size, MCMC burn-in takes up considerable computational time. We therefore iterated through a series of BEAST inferences, initially only considering sequence evolution and subsequently adding the location data, to arrive at a tree distribution from which trees were taken as starting trees in our final analyses. The latter was composed of multiple independent MCMC runs that were run sufficiently long to ensure that their combined posterior samples achieved effective sample sizes larger than 100 for all continuous parameters.

Data augmentation through online BEAST. As we updated our dataset after the initial analyses of the 2,909 genome collection using the approach discussed (see 'Bayesian evolutionary reconstructions'), we sought to capitalize on these efforts to limit the burn-in for subsequent analyses of the 3,959 dataset. Specifically, we adopted the distance-based procedure to insert new taxa into a time-measured phylogenetic tree sample as implemented in the BEAST framework for online inference[36]. We subsequently use the augmented tree as the starting tree for the analyses of the updated dataset.

Time-inhomogeneous reconstructions. To accommodate the time variability of the mobility measures, we constructed epoch model extensions of the discrete phylogeography approach that allow specifying arbitrary intervals over the evolutionary history and associating them with different model parameterizations[37]. As a complement to testing covariates of spatial diffusion using a time-homogeneous model, we used the epoch extension to specify monthly intervals, enabling us to incorporate monthly mobility matrices (air transportation data were available only as monthly numbers), but assuming time-homogeneous effect sizes and inclusion probabilities. Monthly covariates were again log-transformed and standardized after adding a pseudo-count to each entry in the monthly matrices.

In addition, we performed another analysis in which we relaxed the constant-through-time inclusion probability of the covariates. In this model specification, each interval is associated with a specific set of indicator variables to represent the inclusion or exclusion of covariates, but we pool information about predictor inclusion across the intervals using hierarchical graph modelling[38]. This approach uses a set of indicator variables to model covariate inclusion at the hierarchical level but enables interval-specific inclusion or predictors to diverge from the hierarchical level with a non-zero probability (with the number of differences modelled as a binomial distribution[38]), which was set to 0.10 in our study. We estimated hierarchical and interval-level inclusion using the spike-and-slab procedure[38].

Finally, we performed an analysis using the time-inhomogeneous model in which the interval-specific transition rates are modelled as

a function of the single covariate that is supported by the analyses above leveraging aggregate mobility. We incorporated more variability through time by specifying 2-week intervals (similar to the coalescent GLM interval specification). In addition, we add time-homogeneous random effects to the phylogeographical transition rate parameterization to account for potential biases in the ability of mobility to predict phylogeographical spread. Although the posterior mean estimates for these random effects vary, only very few indicate that individual phylogeographical transition rates significantly deviate from the mobility data (Supplementary Fig. 3). The time-inhomogeneous GLM approach that we use enables the modelling of relative differences in transition rates, but also the overall rate of migration between countries varies through time, and importantly, this is strongly affected by intervention strategies. To accommodate these dynamics, we further extend this model by incorporating a time-inhomogeneous overall CTMC rate scaler and parameterize it as a log-linear function of the total monthly between-country log-transformed and standardized mobility (time-variable rate scalar GLM in Extended Data Table 2). To generate realizations of the discrete location CTMC process and obtain estimates of the transitions (Markov jumps) between states under this model, we used posterior inference of the complete Markov jump history through time[10,39].

Although the epoch model enables us to flexibly accommodate time-variable spatial dynamics, it considerably increases the computational burden associated with likelihood evaluations. To efficiently draw inferences under this model for our large dataset, we fit the time-inhomogeneous spatial diffusion process to a set of trees inferred under the time-homogeneous GLM diffusion model described above. Although likelihood evaluations remain computationally expensive, even with the speed-up offered by GPU computation with BEAGLE, eliminating simultaneous tree estimation tremendously reduces the parameter space, requiring only modest MCMC chain lengths to adequately explore it. Model and inference specifications for the different analyses are available as BEAST XML input files on GitHub (https://github.com/phylogeography/SARS-CoV-2_EUR_PHYLOGEOGRAPHY) and Zenodo (https://doi.org/10.5281/zenodo.4876442).

Posterior summaries. We assessed MCMC mixing (for example, using effective sample sizes) and summarized continuous parameter estimates using Tracer v.1.7.1[40]. Credible intervals were computed as 95% HPD intervals. Trees were visualized using FigTree v.1.4.4 (available at https://github.com/rambaut/figtree/releases). In terms of phylogeographical estimates, we mainly focused on (1) transitions to each location and from each location (based on Markov jump estimates) instead of pairwise transitions; (2) ratios of these transitions and (3) how these transitions structured transmission chains in individual countries. Transitions to and from each location avoid drawing conclusions about direct migration between countries, which can be tenuous given the incomplete genome coverage of Europe, while their ratios avoid using absolute numbers of transitions, which are highly sample-dependent. Phylogeographical inference is limited to reconstructing the transitions in the ancestral history of a sample of sequences, which will only be a small fraction of the actual migration events especially when these events result in insufficient onward transmission to be captured in our limited sample. In addition, SARS-CoV-2 genome data can be poorly resolved and identical genomes in different locations are consistent with hypotheses that involve both a sparse and a rich number of virus flows between these locations. As the data hold little information to distinguish these hypotheses, we only consider sparse scenarios by including only unique sequences for each location. A joint inference of sequence evolution and discrete spatial diffusion would err on the side of sparse hypotheses anyway because it will tend to cluster identical sequences that share a location. Despite the general underestimation of spatial dispersal, a phylogeographical inference is still likely to capture the transition events with important onward transmission, and evaluating the importance of such events relative to persistence is a major focus of this study. Cryptic transmission also complicates the ability to reconstruct spatial dispersal, but we expect this to be equally likely for introductions and persistence and therefore focus on their ratio for each location.

We provide three new tree sample tools in the BEAST codebase on GitHub (https://github.com/beast-dev/beast-mcmc) to obtain posterior summaries of location transition histories using posterior tree distributions annotated with Markov jumps:

(1) TreeMarkovJumpHistoryAnalyzer enables the collection of Markov jumps and their timings from a posterior tree distribution annotated with Markov jumps histories in a .csv file for further analyses.

(2) TreeStateTimeSummarizer decomposes the total tree time into the times associated with contiguous partitions of a tree estimated to be in a particular location state, with the partitions determined by the Markov jumps. An arbitrary lower- and upper-time boundary can be specified to restrict the summary to a particular time interval in the evolutionary history. We use the time estimates for the separate partitions associated with each state to calculate an entropy measure that summarizes the genetic make-up of country-specific transmission chains. Specifically, we use for each location a normalized Shannon entropy:

$$-\frac{1}{\ln(n)}\sum_{i}^{n} p_i \ln(p_i), \quad (1)$$

where p_i is the proportion of time associated with that location for partition i of a phylogeographical tree and n represents the number of partitions for that location in the tree.

(3) PersistenceSummarizer also uses posterior tree distributions annotated with Markov jumps to summarize the number of lineages at a particular point in time (evaluation time (T_e); see Extended Data Fig. 5), which location states they are associated with, since what time point in the past they have maintained that state and how many sampled descendants they have after time T_e (Extended Data Fig. 5). In addition, it enables summarizing how long these lineages have circulated independently before T_e, so before sharing common ancestry with other lineages that maintained the same location state. This information allows us to determine how many lineages are circulating at T_e that stem either from a unique persistent lineage (maintaining the same location states) or unique introduction event since a particular time before T_e (ancestral time (T_a) in Extended Data Fig. 5). The association between incidence and the difference in the logit proportion of unique introductions and the logit proportion of their descendants on 15 August was evaluated using a P value obtained by a linear regression analysis.

Reporting summary

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

Data availability

BEAST XML input files are available on GitHub (https://github.com/phylogeography/SARS-CoV-2_EUR_PHYLOGEOGRAPHY) and Zenodo (https://doi.org/10.5281/zenodo.4876442). The SARS-CoV-2 genome data required to run these XML files can be downloaded from https://www.gisaid.org; all GISAID accession numbers are listed in the GISAID acknowledgements table (Supplementary Table 3).

The Google COVID-19 Aggregated Mobility Research Dataset and the Facebook mobility data are not publicly available owing to stringent licensing agreements. Information on the process of requesting access to the Google mobility data are available from A.S. (sadilekadam@google.com) and the COVID-19 Community Mobility Reports that were derived from the Google data are publicly available at https://www.google.com/covid19/mobility/. The Facebook mobility data are made available through the Data for Good programme (https://dataforgood.fb.com) under the terms of a data license agreement that defines the

allowed terms of use by partners (contact: disastermaps@fb.com). Once a request for access from a partner institution is vetted and an appropriate data license agreement is signed, then access is granted through Facebook's web-based spatial visualization tool called GeoInsight. Air travel data were obtained from the International Air Transport Association (http://www.iata.org).

log-transformed and standardized among-country mobility and air travel data are specified in the available BEAST XML files (https://github.com/phylogeography/SARS-CoV-2_EUR_PHYLOGEOGRAPHY and https://doi.org/10.5281/zenodo.4876442). COVID-19 incidence data were obtained from https://www.ecdc.europa.eu/en/covid-19/data.

Code availability

The code to run the BEAST analyses is available in the hmc_develop branch of the BEAST codebase on GitHub (https://github.com/beast-dev/beast-mcmc) and Zenodo (https://doi.org/10.5281/zenodo.4895235). The tools TreeMarkovJumpHistoryAnalyzer, TreeStateTimeSummarizer and PersistenceSummarizer are available from the master branch in the same codebase.

16. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. **22**, 30494 (2017).
17. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol*. **537**, 39–64 (2009).
18. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. **2**, *vew007* (2016).
19. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol*. **37**, 1530–1534 (2020).
20. Maas, P. Facebook disaster maps: aggregate insights for crisis response & recovery. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, 2019).
21. Gilbert, M. et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *Lancet* **395**, 871–877 (2020).
22. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. Social connectedness: measurement, determinants, and effects. *J. Econ. Perspect*. **32**, 259–280 (2018).
23. Kraemer, M. U. G. et al. Mapping global variation in human mobility. *Nat. Hum. Behav*. **4**, 800–810 (2020).
24. Bassolas, A. et al. Hierarchical organization of urban mobility and its connection with city livability. *Nat. Commun*. **10**, 4817 (2019).
25. Wilson, R. J. et al. Differentially private SQL with bounded user contribution. *Proc. Priv. Enhanc. Technol*. **2020**, 230–250 (2020).
26. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. **4**, vey016 (2018).
27. Böhmer, M. M. et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. *Lancet Infect. Dis*. **20**, 920–928 (2020).
28. Worobey, M. et al. The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
29. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol*. **30**, 713–724 (2013).
30. Faria, N. R. et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* **361**, 894–899 (2018).
31. Baele, G., Lemey, P. & Suchard, M. A. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. *Syst. Biol*. **65**, 250–264 (2016).
32. Chernomor, O. et al. Split diversity in constrained conservation prioritization using integer linear programming. *Methods Ecol. Evol*. **6**, 83–91 (2015).
33. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLOS Comput. Biol*. **5**, e1000520 (2009).
34. Ayres, D. L. et al. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol*. **68**, 1052–1061 (2019).
35. Baele, G., Gill, M. S., Lemey, P. & Suchard, M. A. Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework. *Wellcome Open Res*. **5**, 53 (2020).
36. Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A. & Baele, G. Online Bayesian phylodynamic inference in BEAST with application to epidemic reconstruction. *Mol. Biol. Evol*. **37**, 1832–1842 (2020).
37. Bielejec, F., Lemey, P., Baele, G., Rambaut, A. & Suchard, M. A. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Syst. Biol*. **63**, 493–504 (2014).
38. Cybis, G. B., Sinsheimer, J. S., Lemey, P. & Suchard, M. A. Graph hierarchies for phylogeography. *Phil. Trans. R. Soc. Lond. B* **368**, 20120206 (2013).
39. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Phil. Trans. R. Soc. Lond. B* **363**, 3985–3995 (2008).
40. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol*. **67**, 901–904 (2018).

Acknowledgements We thank all of the authors who have shared genome data on GISAID, and we have included a table (Supplementary Table 3) acknowledging the authors and institutes involved. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (725422-ReservoirDOCS) and the Bill & Melinda Gates Foundation (OPP1094793 and INV-024911). This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD 005. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z. P.L. acknowledges support by the Research Foundation–Flanders ('Fonds voor Wetenschappelijk Onderzoek–Vlaanderen', G066215N, G0D5117N and G0B9317N). G.B. acknowledges support from the 'Interne Fondsen KU Leuven'/Internal Funds KU Leuven under grant agreement no. C14/18/094, and the Research Foundation–Flanders ('Fonds voor Wetenschappelijk Onderzoek–Vlaanderen', G0E1420N and G098321N). M.A.S. acknowledges support from National Institutes of Health grants U19 AI135995 and R01 AI153044. S.D. is supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium). We acknowledge support from NVIDIA Corporation through the donation of parallel computing resources used for this research and thank AMD for the donation of critical hardware and support resources from its HPC Fund that made this work possible.

Author contributions P.L. and S.D. designed the study, performed analyses and drafted the manuscript. V.C., C.P. and A.S. provided and analysed data. S.L.H., F.V.d.B., N.R., S.L. and A.J.T. compiled and analysed data. A.L., B.B.O.M. and M.K. contributed data. G.B. performed data analyses. M.S.G., X.J. and M.A.S. developed statistical inference methodology. All authors contributed to interpreting and reviewing the manuscript.

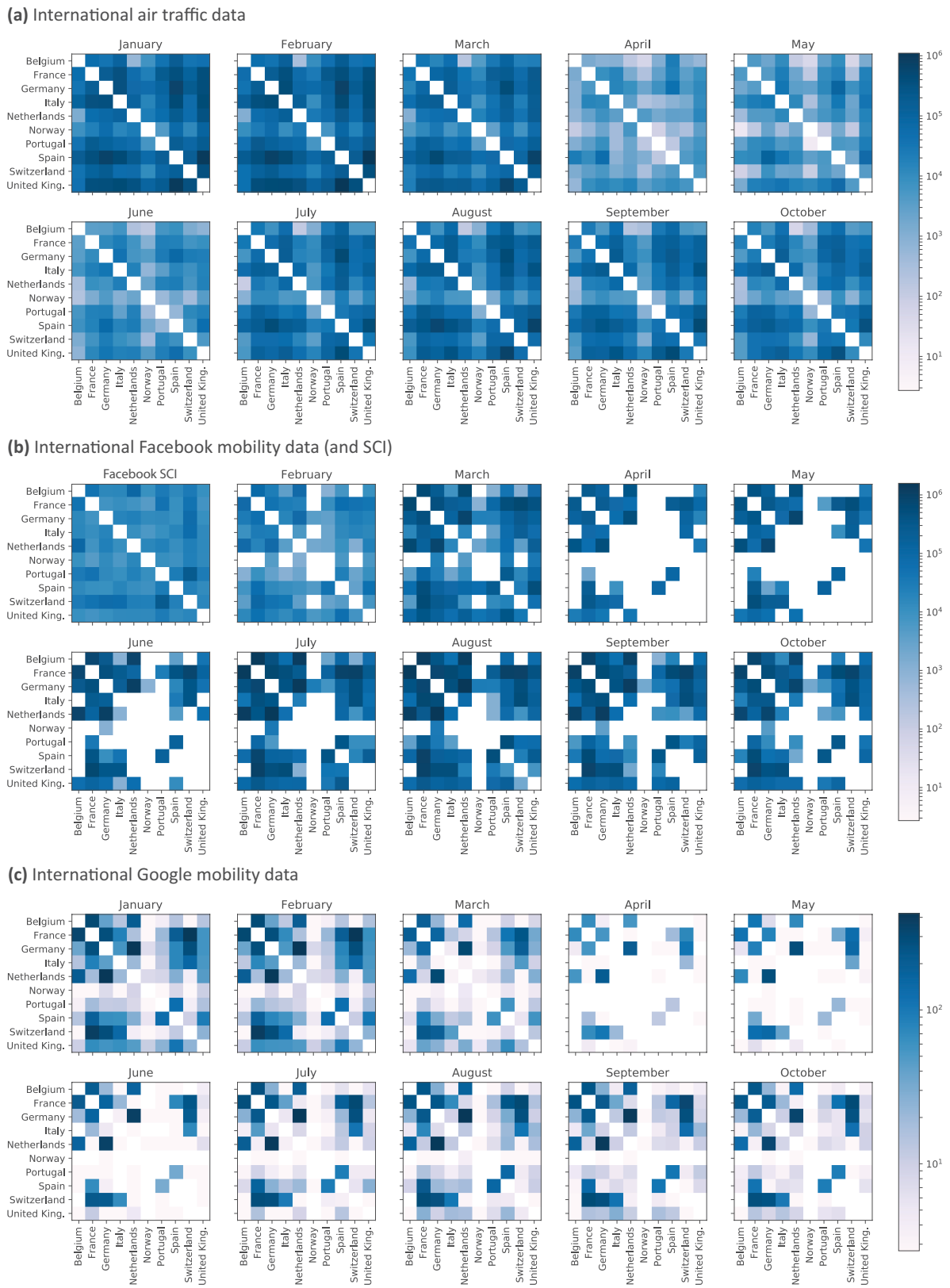
Competing interests The authors declare no competing interests.

Additional information

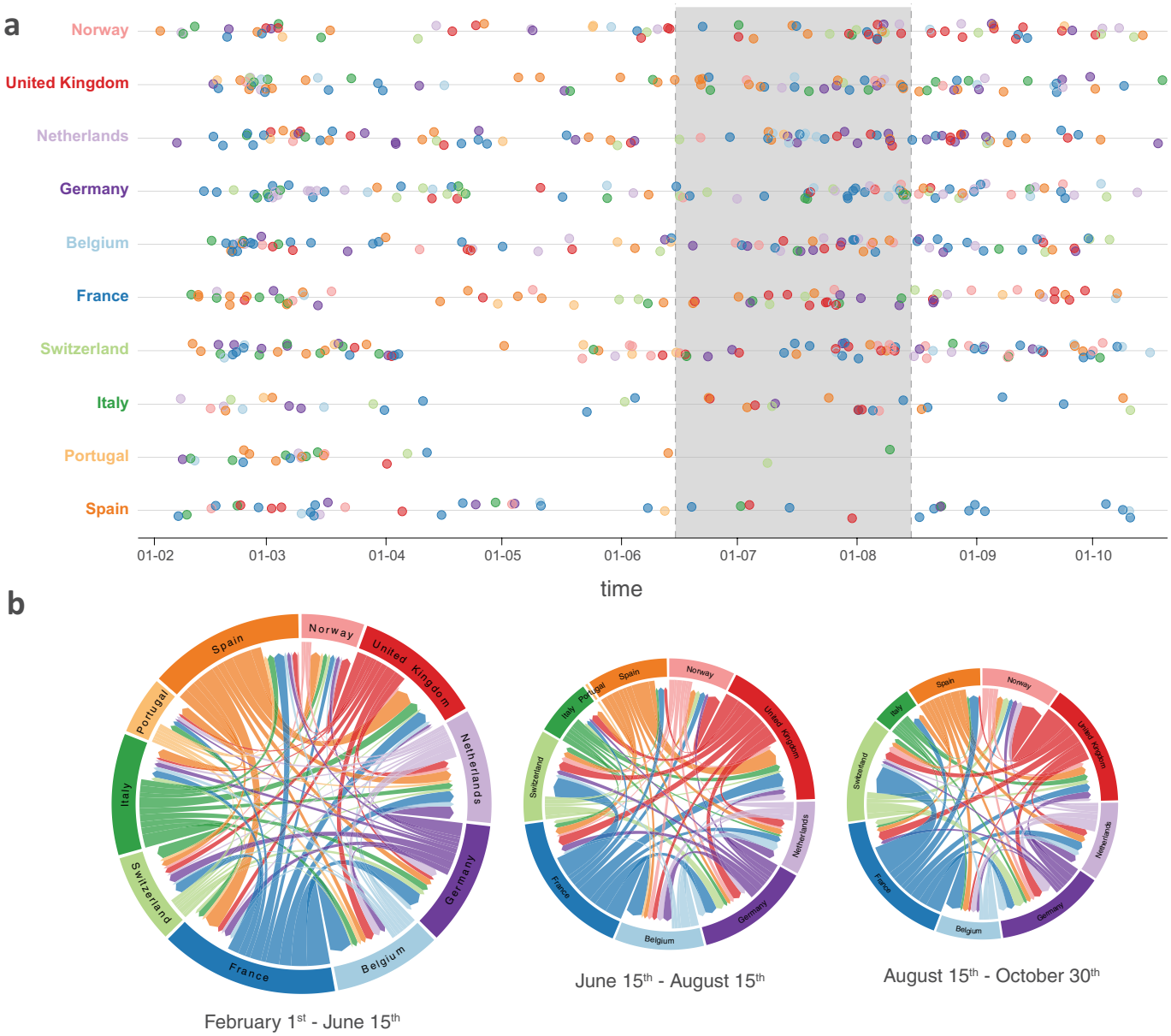
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03754-2.

Correspondence and requests for materials should be addressed to P.L. or S.D.

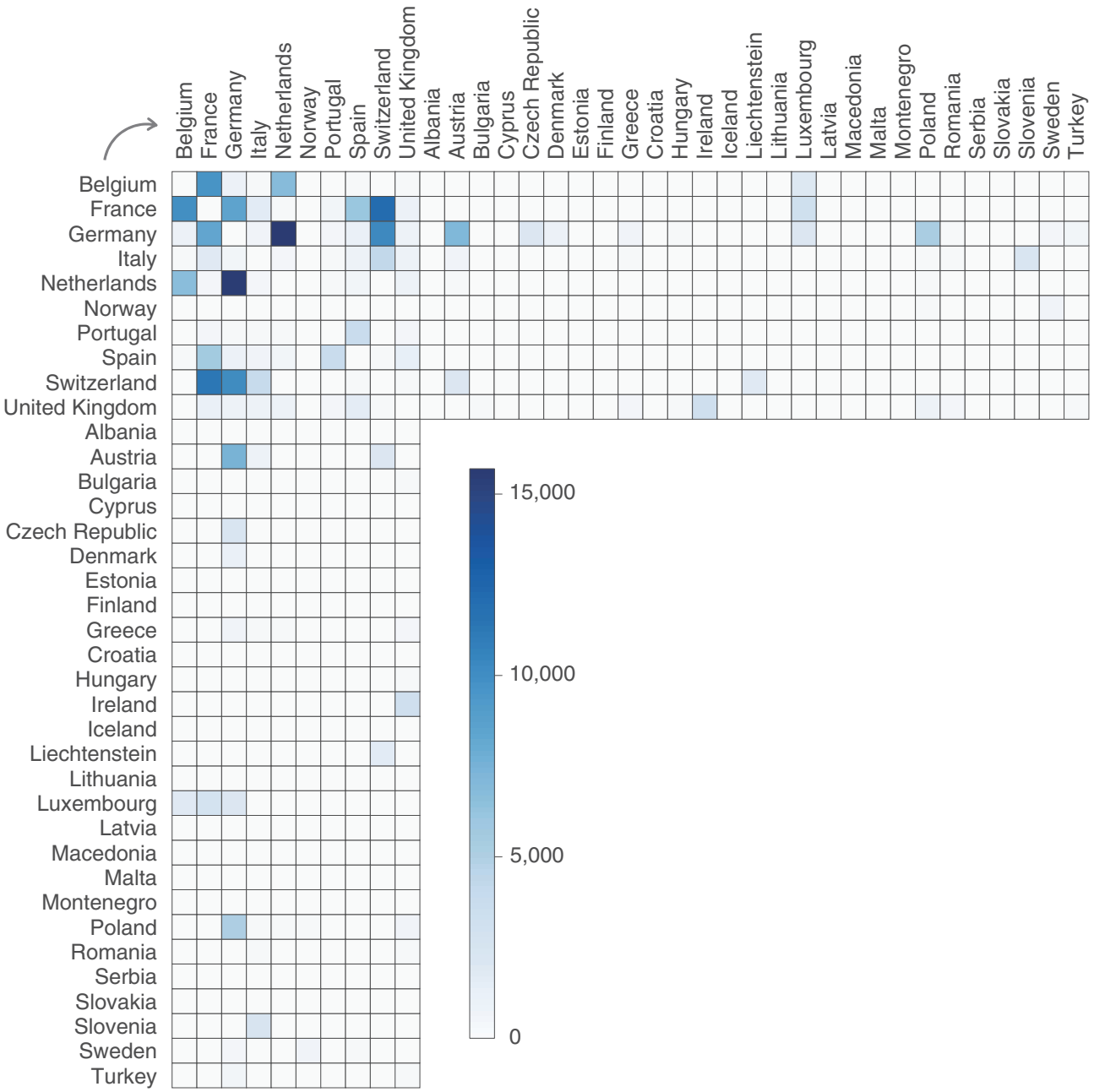
Peer review information *Nature* thanks Jason Thomas Ladner, Matthew Scotch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.



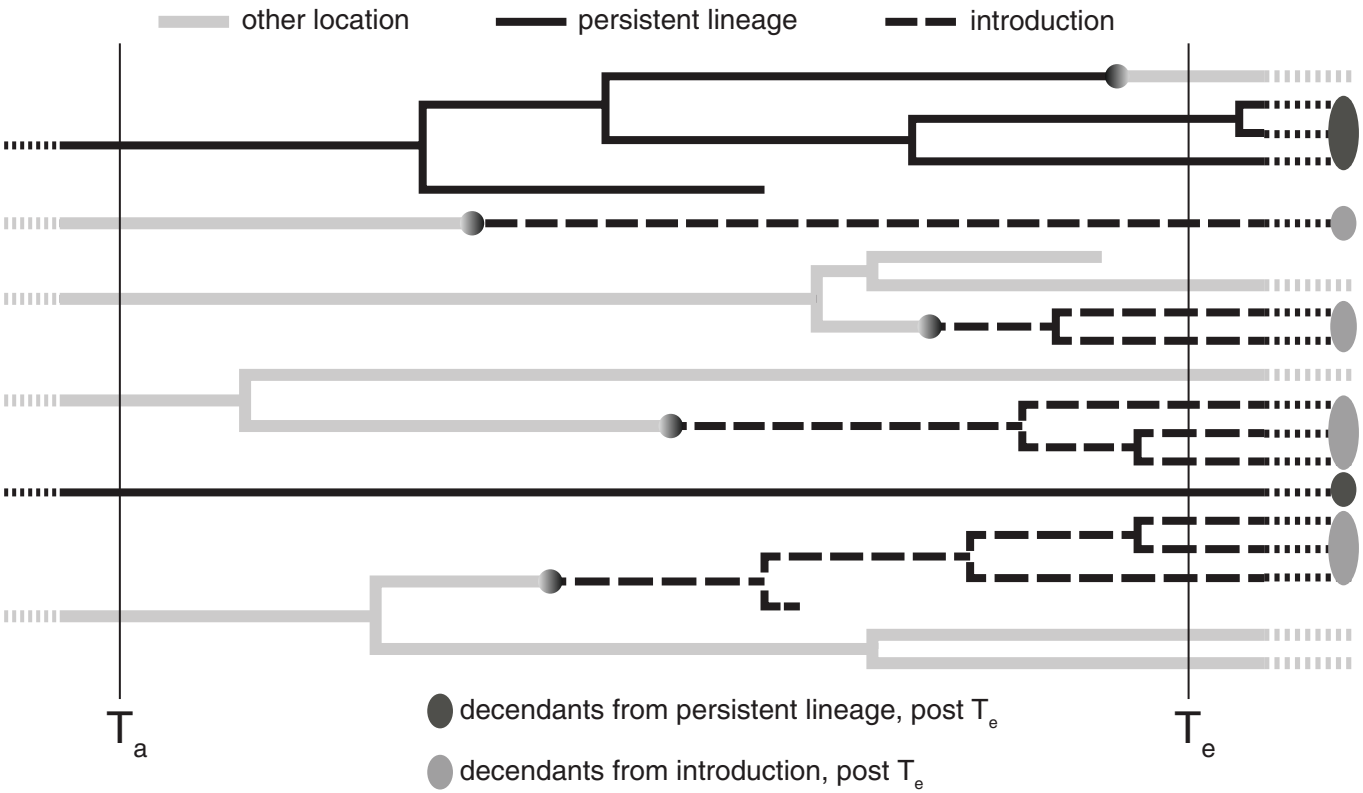
Extended Data Fig. 1 | Monthly international mobility data matrices for air traffic and Google and Facebook mobility data. a–c, International air traffic data (**a**), international Facebook mobility data (**b**), and international Google mobility data (**c**). For Facebook data (**b**), we also report the single SCI matrix.



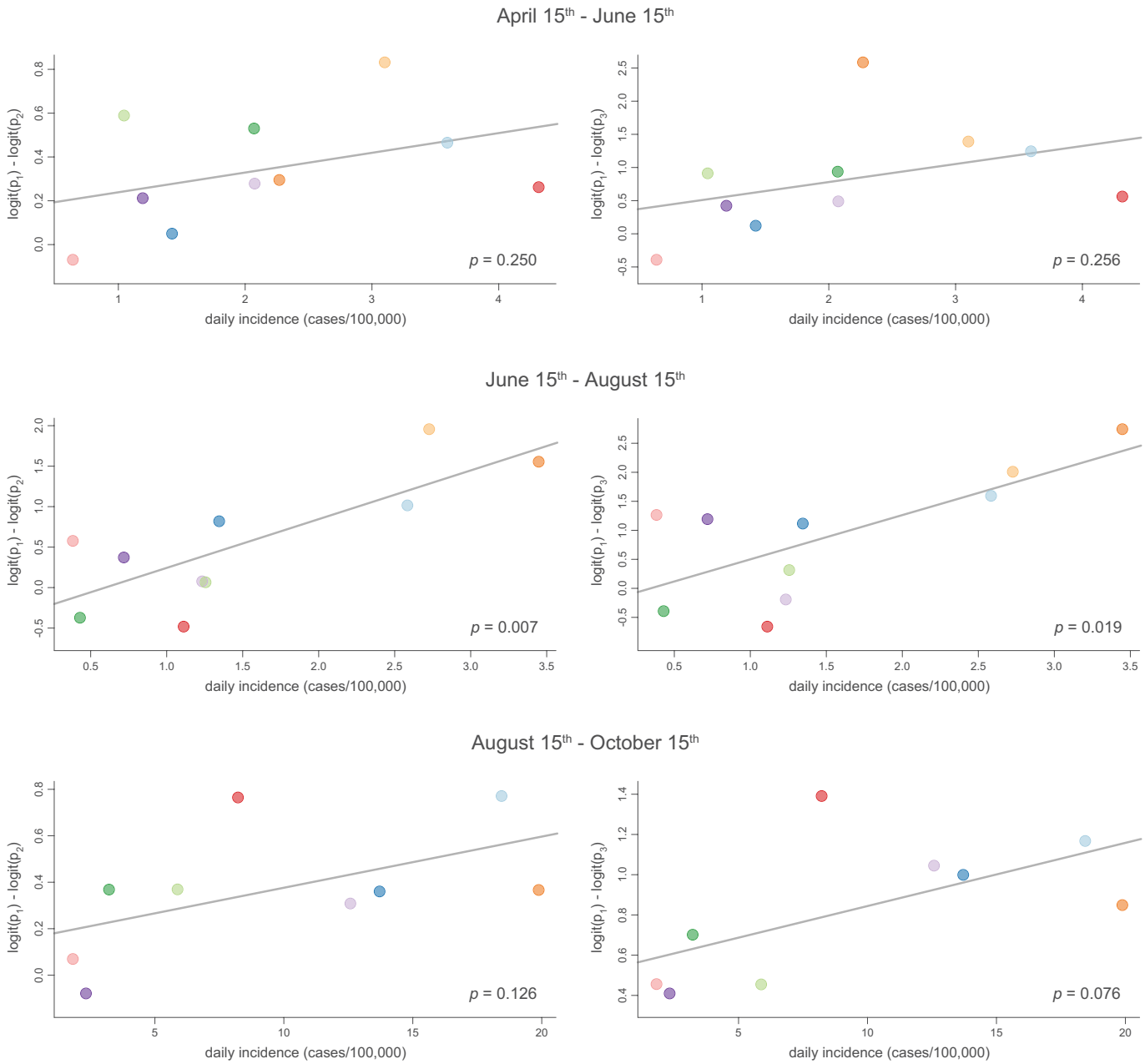
Extended Data Fig. 2 | Estimated introductions through time in the 10 European countries and circular migration flow plots summarizing the estimated transitions between the countries for different time intervals throughout the evolutionary history of SARS-CoV-2. a, The introductions through time serve as an illustration and are based on the Markov jump history in the maximum clade credibility tree. We note that the posterior distribution of trees is accompanied by considerable uncertainty about the location of origin, destination and timing of the transitions that is difficult to appropriately visualize. The grey box represents the time period from 15 June to 15 August 2020. **b**, The circular migration flow plots are based on the posterior expectations of the Markov jumps. The sizes of the plots reflect the total number of transitions for each period. In these plots, migration flow out of a particular location starts close to the outer ring and ends with an arrowhead more distant from the destination location.



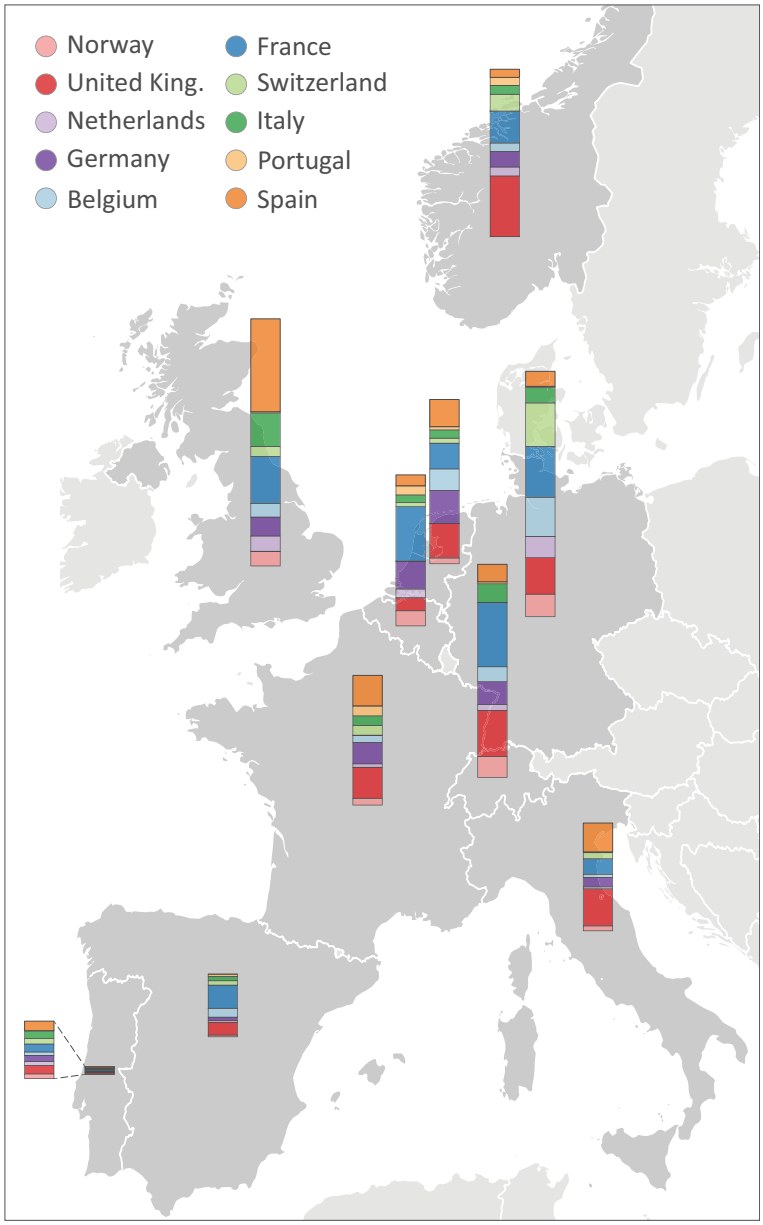
Extended Data Fig. 3 | Pairwise mobility data among the 10 countries included in the phylogeographical analysis and other European countries. Heat-map cells are coloured according to international Google mobility data for the time period between January and October 2020.



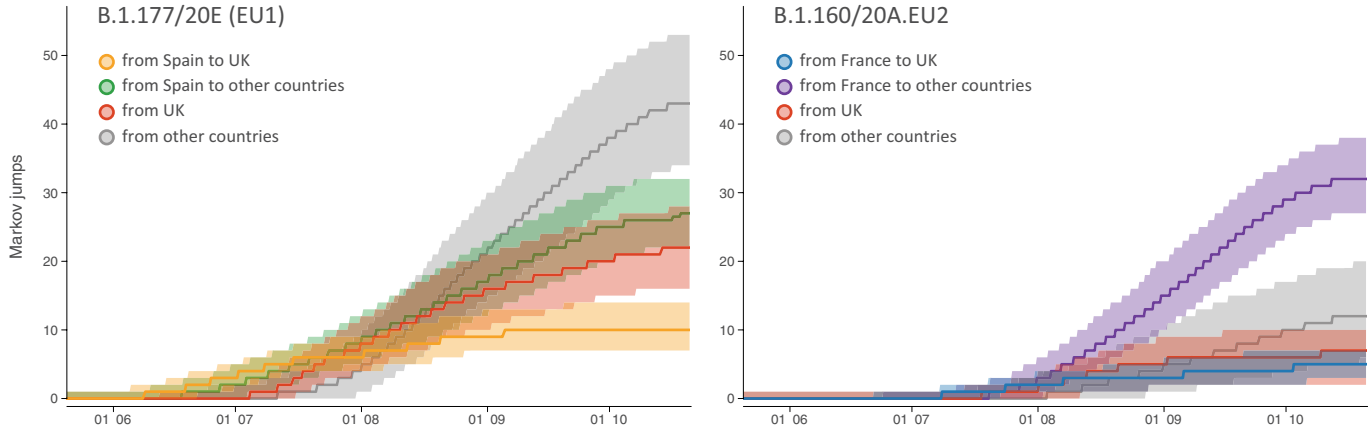
Extended Data Fig. 4 | Conceptual representation of persistent lineages and introductions during the time interval delineated by the evaluation time and the ancestral time. At evaluation time (T_e), we evaluate how many lineages are circulating in the location of interest; in this case, 12 lineages (lineages in other locations are represented by thick grey branches). We subsequently identify whether these lineages maintained this location up to ancestral time (T_a) in their ancestry or whether they result from an introduction event in the time interval of interest. By determining whether other lineages circulating in the location of interest at T_e are descendants of the same persistent lineage or whether they share an introduction event, we identify the unique persistent lineages or introductions, in this case 2 and 4 lineages, respectively. In addition to the proportion of unique introductions ($p_1 = 4/6$), we also summarize the proportion of their descendants at T_e ($p_2 = 9/(9 + 3)$ in this case) and the proportion of their descendants in terms of sampled tips after T_e (p_3). Those tips are not shown here but are conceptually represented for both introductions and persistent lineages by ovals.



Extended Data Fig. 5 | Scatter plots of the difference in the logit proportion of unique introductions and the logit proportion of their descendants on 15 August against the incidence and the difference in the logit proportion of unique introductions and the logit proportion of descendant tips after 15 August against incidence. Left, the difference in logit proportions of unique introductions (p_1) and their descendants (p_2). Right, the difference in logit proportions of unique introductions (p_1) and descendant tips (p_3). Data are shown for the periods between 15 April and 15 June, between 15 June and 15 August, and between 15 August and 15 October. The *P* values in the lower right corner of the plots are the p-values for the hypothesis tests based on the t-statistic evaluating whether the regression coefficient in a linear regression model is different from 0.



Extended Data Fig. 6 | Estimated geographical origin of viral influx over the summer (15 June–15 August 2020) in each country. Each bar plot summarizes the posterior Markov jump estimates into a specific country. For the bar representing a low number of introductions into Portugal, a magnified view is provided.



Extended Data Fig. 7 | Phylogeographical transitions for lineages B1.177/20E (EU1) and B1.160/20A.EU2. Cumulative phylogeographical transitions are summarized as posterior mean estimates with 95% HPD intervals over time for four types of Markov jumps. For B1.177/20E (EU1), jumps from Spain to the UK, from Spain to other countries, from the UK and from other countries are included. For B1.160/20A.EU2 jumps from France to the UK, from France to other countries, from the UK and from other countries are included.

Extended Data Table 1 | Genome sampling by country, collected on 3 November 2020 and updated on 5 January 2021

country	genomes (Nov. 3rd, 2020)	genomes (Jan 5th, 2021)	total
Belgium	183 (1,091)	53 (957)	236
France	600 (1,441)	167 (762)	767
Germany	246 (486)	75 (482)	321
Italy	281 (795)	75 (257)	356
The Netherlands	159 (2,387)	47 (1,032)	206
Norway	100 (414)	92 (482)	192
Portugal	100 (1,370)	100*	200
Spain	647 (2,443)	191 (827)	838
Switzerland	100 (3,019)	98 (1,421)	198
The United Kingdom	493 (26,366)	152 (50,175)	645
total	2,909	1,050	3,959

The numbers in between brackets represent the number of available genomes that were subsampled. *For Portugal, almost all available genomes were included to increase the number of genomes to 200.

Extended Data Table 2 | Parameter estimates for the various Bayesian time-measured phylogeographical models

Model		Parameter estimates
Time-homogenous spatial diffusion	coalescent GLM	α = 2.604 [2.487,2.735], β = 1.711 [1.603,1.898]
	spatial GLM	air travel: E[δ] = 0.018, (β|δ=1) = 0.044 [0.001,0.123] SCI: E[δ] = 0.004, β(|δ=1) = 0.013 [0.003,0.032] mobility: E[δ] > 0.999, β(|δ=1) = 0.358 [0.258,0.456]
Time-inhomogeneous spatial diffusion	spatial GLM, constant inclusion probabilities	air travel: E[δ] = 0.018, β(|δ=1) = 0.029 [0.001,0.105] SCI: E[δ] = 0.008, β|δ=1 = 0.024 [0.001,0.078] mobility: E[δ] > 0.999, β(|δ=1) = 0.333 [0.229,0.438]
	spatial GLM, time-variable inclusion probabilities	air travel: E[δ_h] = 0.010, β|(δ_h=1) = 0.047 [0.002,0.139] SCI: E[δ_h] = 0.012, β|δ_h=1 = 0.018 [0.000,0.062] mobility: E[δ_h] = 0.949, β(|δ_h=1) = 0.357 [0.230,0.503]
	spatial GLM	mobility: β = 0.271 [0.118,0.444]
	time-variable rate scalar GLM	mobility: α = 0.740 [0.618,0.856], β = 0.504 [0.350,0.646]

The coalescent GLM parameterizes biweekly effective population sizes as a log-linear function of COVID-19 incidence data, with α and β representing the log intercept and log regression coefficient. In the time-inhomogeneous spatial diffusion models, no coalescent prior was used as these models were fitted onto posterior trees inferred from the time-homogeneous model (see Methods). For the spatial GLM model, we report inclusion probability estimates through the expectations of the Boolean indicators (δ) associated with each predictor and log conditional effect sizes (the regression coefficient conditional on the predictor being included in the model, $\beta|\delta = 1$). The SCI is based on Facebook data. For the model with time-variable inclusion probabilities, we report the parameters at the hierarchical level (δ_h and $\beta|\delta_h$, see Methods). In the model with a time-variable rate scalar, we parameterize this rate scalar as a log-linear function of the overall between-country mobility, with α and β representing the log intercept and log regression coefficient.

Using a time-homogeneous model of spatial diffusion, we estimate a maximum inclusion probability for the mobility data whereas air transportation data and SCI offer no predictive value. We also estimate a strong positive association between the change in the viral population size through time and COVID-19 incidence in the coalescent GLM. We further confirm the support for the mobility covariate in a time-inhomogeneous spatial model that incorporates monthly mobility measures, with either constant or time-variable inclusion probabilities. In addition to parameterizing the relative rates of spread between countries according to this covariate, we extend our time-inhomogeneous approach to also model biweekly variation in the overall rate of spread between countries as a function of mobility measures (time-variable rate scalar GLM). This approach estimates a positive association between the overall rate of spatial spread and mobility data.

Extended Data Table 3 | Mobility percentage to or from each country within our 10-country sample

country	Mobility percentage
Belgium	87.2
France	89.5
Germany	63.9
Italy	64.8
The Netherlands	93.2
Norway	27.1
Portugal	94.0
Spain	90.3
Switzerland	84.8
The United Kingdom	48.6

For each country, the mobility to or from each country within in our dataset is listed as a percentage of the total between-country mobility within Europe. The pairwise mobility measures summarized in this table are shown in Extended Data Fig. 3.

nature research

Corresponding author(s): Philippe Lemey, Simon Dellicour

Last updated by author(s): Jun 2, 2021

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Not applicable
Data analysis	hmc_develop branch of BEAST available in the codebase at https://github.com/beast-dev/beast-mcmc. TreeMarkovJumpHistoryAnalyzer v1.0.0, TreeStateTimeSummarizer v1.0.0 and PersistenceSummarizer v1.0.0 are part of the BEAST codebase available at https://github.com/beast-dev/beast-mcmc (DOI: 10.5281/zenodo.4895235). FigTree v1.4.4., Tracer v1.7.1, TempEst v1.5.3, IQTREE v2.0.3, MAFFT v.7.453, BEAGLE v3.1.2, Phylogenetic Diversity Analyzer v0.5 available at http://www.cibiv.at/software/pda/web-pda/

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

BEAST XML input files are available at https://github.com/phylogeography/SARS-CoV-2_EUR_PHYLOGEOGRAPHY (DOI: 10.5281/zenodo.4876442). The SARS-CoV-2 genome data required for running these XML files can be downloaded from https://www.gisaid.org; all GISAID accession numbers are listed in the GISAID acknowledgments table (Supplementary Table 3).

The Google COVID-19 Aggregated Mobility Research Dataset and the Facebook mobility data are not publicly available owing to stringent licensing agreements. Information on the process of requesting access to the Google mobility data is available from A.S. (sadilekadam@google.com) and the COVID-19 Community Mobility Reports that were derived from the Google data are publicly available at https://www.google.com/covid19/mobility/. The Facebook mobility data are made available through the Data for Good program (https://dataforgood.fb.com) under the terms of a data license agreement which defines the allowed terms of use by partners (contact: disastermaps@fb.com). Once a partner institution's request for access is vetted and an appropriate data license agreement is signed, then access is granted through a Facebook's web-based spatial visualization tool called GeoInsight. Air travel data were obtained from the International Air Transport Association (http://www.iata.org).
Log-transformed and standardized among country mobility and air travel data are specified in the available XML files. COVID-19 incidence data was obtained from https://www.ecdc.europa.eu/en/covid-19/data.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	phylogeographic reconstruction of the spread of SARS-CoV-2 based on 3,959 genomes
Research sample	The research sample consists of 3,959 SARS-CoV-2 B.1 genomes with a sampling date between January and October 2020. The genomes were downloaded from GISAID (www.gisaid.org); GISAID accession numbers can be found in the GISAID acknowledgments table.
Sampling strategy	The number of SARS-CoV-2 genomes by country in our sampling strategy was proportional to the cumulative number of cases, with a minimum of 200 in order to ensure that each country is well-represented. The ratio of sequences over cases was selected in order to achieve a large genomic data set that could still be analyzed with complex models in a Bayesian phylogeographic analysis in a reasonable time frame (aiming at about 4,000 genomes).
Data collection	We used a two-step genome data collection procedure. Genomes were downloaded by S.L.H. and P.L. using from EpiCoV database in GISAID. In the first step, we assembled 2,909 genomes from 10 European countries with sufficient numbers of genomes available from the first wave an the beginning of the second wave on November 3, 2020. To maximize the temporal and spatial coverage in each country, we binned genomes by epi-week and sampled as evenly as possible, sampling from a different region within the country when available. To increase the representation recently sampled genomes and include genomes up till the end of the month October, we updated our dataset on January 5, 2021 by adding over 1,000 non-identical sequences collected between August 1st and October 31st. To bias the temporal coverage towards more recent samples, the genomes from each country were binned by week and sampled such that the number of sequences added by week was proportional to an exponential function of the form et/4, where t=0 represents August 1st and t=13 is October 31st. For Portugal, we did not use this preferential sampling as we needed to include close to all available genomes to raise the number of genomes to 200.
Timing and spatial scale	From the globally sampled SARS-CoV-2 genomes available in GISAID (www.gisaid.org), we selected B.1 genomes from Belgium, France, Germany, Italy, Netherlands, Norway, Portugal, Spain, Switzerland and the United Kingdom sampled from January to October 2020. We focused on these countries based on the availability of genome data from both the first and second wave on November 3, 2020 (the date at which we initiated the study), and because of their relatively high ratio of genomes to positive cases. The date range encompasses a time period from first documented spread of SARS-CoV-2 B.1 in Europe to the initial rise in cases during the second COVID-19 wave in Europe. On November 3, 2020, SARS-CoV-2 B.1 genomes were available up to October 21. In the second data augmentation step (cfr. 'Data collection'), this was extended to October 31 to cover the full month.
Data exclusions	From the final aligned sequence set, we removed 12 potential outliers, based on a root-to-tip regression on TempEst v1.5.3 on a maximum-likelihood tree inferred with IQTREE v2.0.3 18.
Reproducibility	Analyses are reproducible through the xml files complemented with the genomic data from GISAID
Randomization	Phylogenetic inference seeks to infer evolutionary history from the complete collection of gene sequences and does not involve comparing an intervention group to a control group, where randomization is needed to assign individuals to these groups.
Blinding	Phylogenetic inference provides a statistical estimate of evolutionary history that is not subject to observer bias as can occur in randomized controlled trials for example. So, no blinding was needed.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging