

SARS-CoV-2 European resurgence foretold: interplay of introductions and persistence by leveraging genomic and mobility data

Philippe Lemey (✉ philippe.lemey@kuleuven.be)

KU Leuven <https://orcid.org/0000-0003-2826-5353>

Nick Ruktanonchai

Virginia Tech

Samuel Hong

KU Leuven <https://orcid.org/0000-0001-6354-4943>

Vittoria Colizza

Inserm <https://orcid.org/0000-0002-2113-2374>

Chiara Poletto

INSERM

Frederik Van den Broeck

KU Leuven

Mandev Gill

KU Leuven

Xiang Ji

Tulane University <https://orcid.org/0000-0002-7243-0865>

Anthony Levasseur

Aix-Marseille Université, Méditerranée Infection

Adam Sadilek

Google Inc <https://orcid.org/0000-0003-0784-7098>

Shengjie Lai

University of Southampton <https://orcid.org/0000-0001-9781-8148>

Andrew Tatem

University of Southampton <https://orcid.org/0000-0002-7270-941X>

Guy Baele

K. U. Leuven

Marc Suchard

Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles

<https://orcid.org/0000-0001-9818-479X>

Simon Dellicour

ULB

Biological Sciences - Article

Keywords: COVID-19, SARS-CoV-2, Europe, second wave, phylogeography, international mobility

Posted Date: February 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-208849/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature on June 30th, 2021. See the published version at <https://doi.org/10.1038/s41586-021-03754-2>.

1 **SARS-CoV-2 European resurgence foretold: interplay of introductions and persistence**
2 **by leveraging genomic and mobility data**

3
4
5 Philippe Lemey^{1,2}, Nick Ruktanonchai^{3,4}, Samuel L. Hong¹, Vittoria Colizza⁵, Chiara Poletto⁵,
6 Frederik Van den Broeck^{1,6}, Mandev S. Gill¹, Xiang Ji⁷, Anthony Levasseur⁸, Adam Sadilek⁹, Shengjie
7 Lai³, Andrew J. Tatem³, Guy Baele¹, Marc A. Suchard^{10,11,12}, Simon Dellicour^{1,13}.

8
9
10 ¹Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven,
11 Belgium.

12 ²Global Virus Network (GVN), Baltimore, MD, USA.

13 ³WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton
14 SO17 1BJ, UK.

15 ⁴Population Health Sciences, Virginia Tech, Blacksburg, VA, USA.

16 ⁵INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique IPLESP, F75012
17 Paris, France.

18 ⁶Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium.

19 ⁷Department of Mathematics, School of Science & Engineering, Tulane University, New Orleans, LA, USA

20 ⁸Microbes, Evolution, Phylogeny and Infection, Aix-Marseille Université and Marseille Institut Universitaire
21 de France, Marseille, France.

22 ⁹Google, Mountain View, CA, USA.

23 ¹⁰Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles,
24 Los Angeles, CA 90095, USA.

25 ¹¹Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los
26 Angeles, CA 90095, USA.

27 ¹²Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles,
28 Los Angeles, CA 90095, USA.

29 ¹³Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP160/12, 50 av. FD Roosevelt, 1050
30 Bruxelles, Belgium.

31 Abstract

32 Following the first wave of SARS-CoV-2 infections in spring 2020, Europe experienced a
33 resurgence of the virus starting late summer that was deadlier and more difficult to contain.
34 Relaxed intervention measures and summer travel have been implicated as drivers of the second
35 wave. Here, we build a phylogeographic model to evaluate how newly introduced lineages, as
36 opposed to the rekindling of persistent lineages, contributed to the COVID-19 resurgence in
37 Europe. We inform this model using genomic, mobility and epidemiological data from 10 West
38 European countries and estimate that in many countries more than 50% of the lineages circulating
39 in late summer resulted from new introductions since June 15th. The success in onwards
40 transmission of these lineages is predicted by SARS-CoV-2 incidence during this period. Relatively
41 early introductions from Spain into the United Kingdom contributed to the successful spread of
42 the 20A.EU1/B.1.177 variant. The pervasive spread of variants that have not been associated with
43 an advantage in transmissibility highlights the threat of novel variants of concern that emerged
44 more recently and have been disseminated by holiday travel. Our findings indicate that more
45 effective and coordinated measures are required to contain spread through cross-border travel.

46

47 **Keywords:** COVID-19, SARS-CoV-2, Europe, second wave, phylogeography, international mobility

48 Introduction

49 Upon successfully curbing transmission in spring 2020, many European countries witnessed a
50 resurgence in COVID-19 cases in late summer. The number of COVID-19 infections increased
51 rapidly, and by the end of October, it was clear that the continent was deep into a second epidemic
52 wave. In England for example, the number of infections was doubling every nine days and the
53 reproduction number was estimated to have risen to 1.6¹. By early December, it was clear that the
54 second wave had become deadlier than the first as the number of Europeans that died of
55 COVID-19 in November had surpassed the total number in April². This forced governments to
56 reimpose new lockdowns and social restrictions in an effort to contain the second wave. While
57 these measures have again reduced infection rates across Europe³, this may only be temporary.
58 Different countries have witnessed a stabilization at relatively high levels or even a new surge in
59 infections. In the United Kingdom (UK), the surge can be largely attributed to the rapid spread of a
60 new variant (B.1.1.7, Variant of Concern 202012/01 or 20I/501Y.V1⁴), which appears to be more
61 transmissible across all age groups and therefore more challenging to contain⁵.

62

63 Already early in the pandemic, experts warned about secondary waves and modelling studies
64 informed by the seasonal variation of endemic coronaviruses predicted a larger winter peak in the
65 Northern hemisphere⁶. By mid April, the European Commission constructed a roadmap to lifting
66 coronavirus containment measures⁷, recommending a cautious and coordinated manner to revive
67 social and economic activities. However, the early start of the devastating second wave
68 demonstrated that, in practice, there was insufficient adherence to these measured
69 recommendations. Cross-border travel, and mass tourism in particular, has been implicated as a
70 major instigator of the second wave. In Belgium for example, which suffered a harsh spring wave
71 and witnessed one of the earliest incidence rises in summer, millions were returning from holidays
72 without testing or enforced quarantine demands. Genomic surveillance demonstrated that a new
73 variant (lineage B.1.177⁸, 20A.EU1 [nextstrain.org]) that emerged in Spain in early summer has
74 spread to multiple locations in Europe⁹. While this variant quickly grew into the dominant
75 circulating SARS-CoV-2 strain in different countries (e.g. the UK and the Netherlands) and
76 illustrated intensive transmission dynamics across countries, it did not appear to be associated
77 with a higher intrinsic transmissibility⁹.

78

79 Although it appears clear that travel had a significant impact on the second wave in Europe, it
80 remains challenging to assess how it may have restructured and reignited the epidemic in the
81 different European countries. Even without resuming travel, relaxing containment measures when
82 low-level transmission is ongoing risks the proliferation of locally circulating strains. Hodcroft et
83 al. (2020) demonstrated that genomic analyses provide important insights into the spread of new
84 variants underlying the resurgence dynamics. Specifically, the authors documented how the
85 spread of B1.177/20A.EU1 has impacted the genetic make-up of SARS-CoV-2 in European
86 countries in different ways. Phylodynamic analyses may provide further detail on the relative
87 importance of persistence versus the introduction of new lineages, but such analyses are
88 complicated for SARS-CoV-2 for different reasons. Phylogenetic reconstructions may be poorly
89 resolved due to the relatively limited substitutions accumulating in SARS-CoV-2 over time. This is
90 further confounded by the degree of mixing that can be expected from unrestricted travel prior to
91 the lockdowns in spring 2020. Here, we perform a phylodynamic analysis to evaluate the relative

92 importance of persistence versus new introductions in causing the resurgence in different
93 European countries. To maximize the resolution of our reconstructions, we incorporate
94 epidemiological data and measures of human mobility across countries. Using this approach, we
95 uncover the degree to which introductions over the summer period differentially contributed to
96 the second wave in various European countries.

97

98 **Results**

99 *Mobility data predicts phylogeographic patterns of SARS-CoV-2 spread*

100 We analyzed SARS-CoV-2 B.1 (20A) genomes from 10 European countries for which a minimal
 101 number of genomes from the second wave were already available on November 3rd, 2020. By
 102 subsampling relative to total case counts, we first compiled a data set of close to 3,000 genomes
 103 sampled up to October 20th, 2020, and subsequently updated this data set on January 5th, 2021,
 104 to close to 4,000 genomes sampled up to October 30th, 2020 (cfr. Methods, Extended Data Table
 105 1). Due to relatively low sequence diversity, phylogenetic reconstructions of SARS-CoV-2 may be
 106 poorly resolved ¹⁰. In order to achieve maximum resolution, we constructed a Bayesian
 107 time-measured phylogeographic model that integrates mobility and epidemiological data. Using
 108 this model, we first tested to what extent viral flow across countries can be predicted by mobility
 109 or connectivity measures. Specifically, we considered international air transportation data, the
 110 Google COVID-19 Aggregated Mobility Research Dataset (also referred to here as ‘mobility data’
 111 for short), as well as Facebook’s Social Connectedness Index (SCI), as covariates of
 112 phylogeographic spread (Extended data Figure 1). The Google mobility data contains anonymized
 113 mobility flows aggregated over users who have turned on the Location History setting, which is off
 114 by default (cfr. Methods). The Social Connectedness Index reflects the structure of social networks
 115 and has been suggested to correlate with the geographic spread of COVID-19 ¹¹. To help inform
 116 the phylogenetic coalescent time distribution, we parameterized the viral population size
 117 trajectories through time as a function of incidence data for the 10 countries under investigation.
 118

Model		Parameter estimates
Time-homogenous spatial diffusion	coalescent GLM	$\alpha = 2.44 [2.35, 2.53]$, $\beta = 1.77 [1.54, 1.92]$
	spatial GLM	air travel: $E[\delta] = 0.01$, $(\beta \delta=1) = -0.03 [-0.16, 0.09]$ SCI: $E[\delta] = 0.01$, $\beta(\delta=1) = -0.15 [-0.27, -0.02]$ mobility: $E[\delta] > 0.99$, $\beta(\delta=1) = 0.36 [0.24, 0.49]$
Time-inhomogeneous spatial diffusion	spatial GLM, constant inclusion probabilities	air travel: $E[\delta] = 0.05$, $\beta(\delta=1) = -0.07 [-0.21, 0.07]$ SCI: $E[\delta] = 0.16$, $\beta \delta=1 = -0.13 [-0.29, 0.01]$ mobility: $E[\delta] > 0.99$, $\beta(\delta=1) = 0.35 [0.22, 0.49]$
	spatial GLM, time-variable inclusion probabilities	air travel: $E[\delta_h] = 0.03$, $\beta (\delta_h=1) = 0.10 [-0.24, 0.12]$ SCI: $E[\delta_h] = 0.12$, $\beta \delta_h=1 = -0.14 [-0.35, 0.01]$ mobility: $E[\delta_h] = 0.93$, $\beta(\delta_h=1) = 0.39 [0.24, 0.57]$
	spatial GLM time-variable rate scalar GLM	mobility: $\beta = 0.26 [0.12, 0.41]$ mobility: $\alpha = 0.67 [0.54, 0.78]$, $\beta = 0.47 [0.32, 0.62]$

119

120 **Table 1. Parameter estimates for the various Bayesian time-measured phylogeographic models applied to**
 121 **the 3,959 genome data set.** The coalescent generalized linear model (GLM) parameterizes bi-weekly
 122 effective population sizes as a log-linear function of COVID-19 incidence data, with α and β representing the
 123 log intercept and log regression coefficient. In the time-inhomogeneous spatial diffusion models, no

124 coalescent prior was used as these models were fitted onto posterior trees inferred from the
125 time-homogeneous model (cfr. Methods). For the spatial GLM model, we report inclusion probability
126 estimates through the expectations of the boolean indicators (δ) associated with each predictor and log
127 conditional effect sizes (the regression coefficient conditional on the predictor being included in the model,
128 $\beta(|\delta=1)$). SCI = Social Connectedness Index, based on Facebook data. For the model with time-variable
129 inclusion probabilities, we report the parameters at the hierarchical level (δ_h and $\beta|\delta_h$, cfr. Methods). In the
130 model with a time-variable rate scalar, we parameterize this rate scalar as a log-linear function of the overall
131 between-country mobility, with α and β representing the log intercept and log regression coefficient.

132

133 Using a time-homogeneous model of spatial diffusion, we estimate a maximum inclusion
134 probability and positive log effect size for mobility data whereas air transportation data and SCI
135 offer no predictive value (Table 1). We also estimate a significantly positive association between
136 viral population size change through time and COVID-19 incidence (Table 1). We further confirm
137 the support for the mobility covariate in a time-inhomogeneous spatial model that incorporates
138 monthly mobility measures, with either constant or time-variable inclusion probabilities (Table 1).
139 The fact that mobility data encompassing both air and land-based transport are required to
140 explain COVID-19 spread highlights the need to consider both types of transport in containment
141 strategies. Having associated mobility with phylogeographic dispersal, we focus on this predictor
142 and include its bi-weekly variation in subsequent reconstructions. In addition to parameterizing
143 the relative rates of spread between countries according to this covariate, we extend our
144 time-inhomogeneous approach to also model bi-weekly variation in the overall rate of spread
145 between countries as a function of mobility measures (Fig. 1). This approach estimates a positive
146 association between the overall rate of spatial spread and mobility data (Table 1). Finally, we add
147 time-homogeneous random effects to the phylogeographic transition rates parameterized
148 according to mobility data across epochs in order to account for potential consistent biases in the
149 ability of mobility to predict phylogeographic spread. While posterior mean estimates for the
150 random effects vary, only very few indicate that individual phylogeographic transition rates
151 significantly deviate from the mobility data (Extended Data Figure 2).

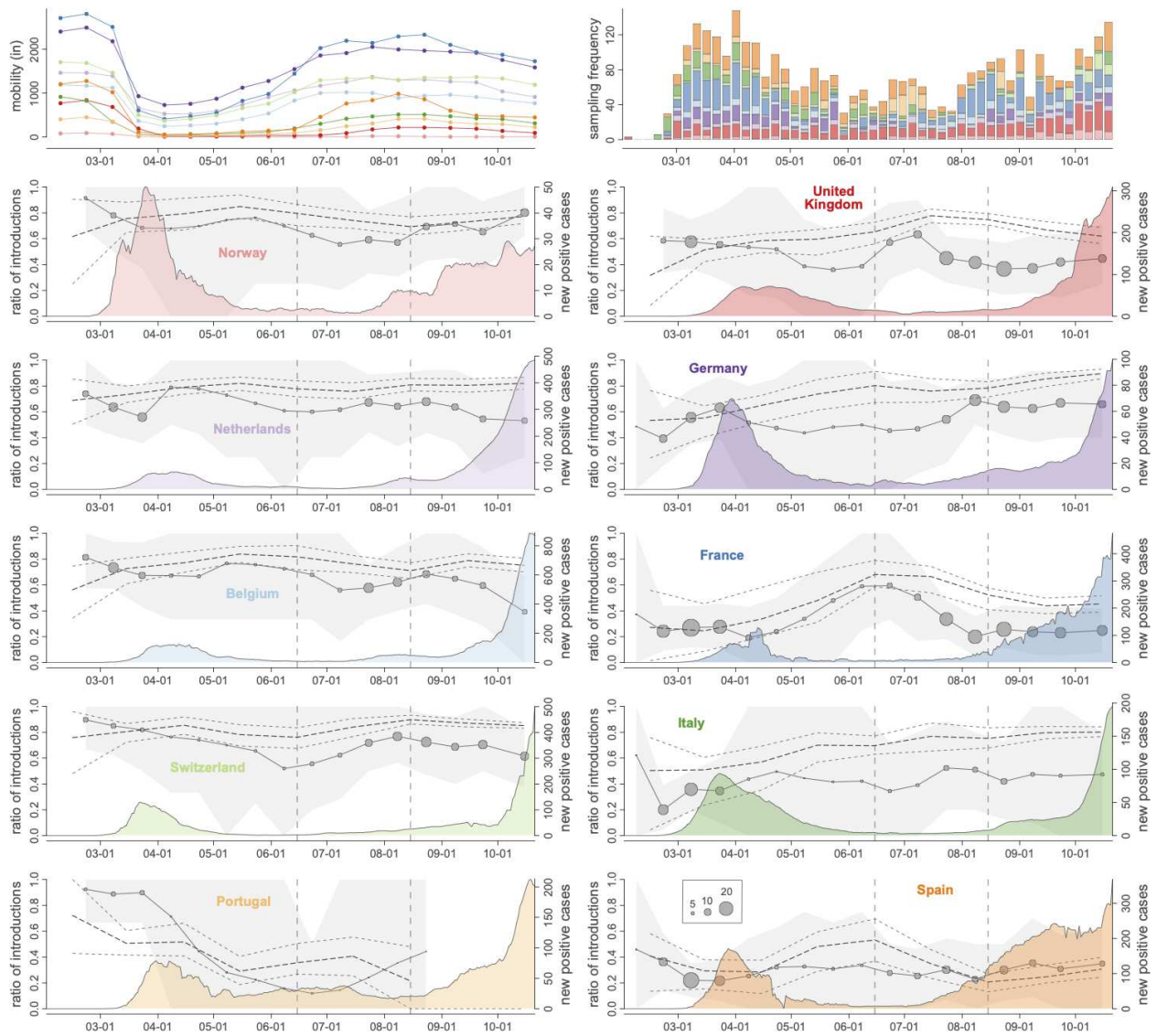
152

153 *A high degree of genetic mixing and dynamic cross-country transmission through time*

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

167

168 According to the proportion of introductions, we estimate more viral import than export events
169 for Norway, the Netherlands, Belgium and Switzerland throughout most of the time period under
170 investigation. According to the estimated phylogenetic entropy, these countries also experienced
171 many independent transmission chains from the beginning of the epidemic spread in Europe. This
172 is consistent with country-specific studies; in Belgium for example, about 331 individual
173 introductions were estimated in the ancestry of a limited sample of 740 genomes¹². For Portugal,
174 we also estimate higher proportions of introductions early in the first wave but with a subsequent
175 decline to predominantly export events. France, Italy and Spain on the other hand are
176 characterized by a relatively high viral export during the first wave. The proportion of
177 introductions remains relatively low for Italy and Spain following the first wave, while in France
178 these proportions are high from mid June till the end of July. The absolute number of transitions in
179 our sample are however low during this time period. These countries also have comparatively
180 lower entropy values early in the epidemic, with an increase for France by the start of summer and
181 a more gradual increase over time for Italy. In Spain however, the genetic complexity of
182 SARS-CoV-2 transmission chains remains limited. In the UK and Germany, the viral flow in and out
183 of the country is initially relatively balanced. A recent large-scale genomic analysis in the UK
184 indicates that this can imply very high absolute numbers of cross-country transmissions as more
185 than 2,800 independent introduction events were identified from the analysis of 26,181 genomes
186¹³. Although our sample is limited compared to this analysis, our reconstructions also recover
187 major influx from Spain, France and Italy during the first wave in the United Kingdom (Extended
188 Data Figure 3). We estimate the highest proportion of introductions for the United Kingdom
189 during the first half of July, indicating an important viral import relative to export around this time.
190 The phylogenetic entropy also peaks around this time. In Germany, the proportions peak
191 somewhat later in summer with a concomitant rise in phylogenetic entropy. We subsequently
192 focus on the time period between the two waves to determine the role of introductions versus
193 persistence in seeding the second wave.



194

195 **Figure 1. Mobility, genome sampling, case counts and phylogeographic summaries through time for 10 West European countries.**

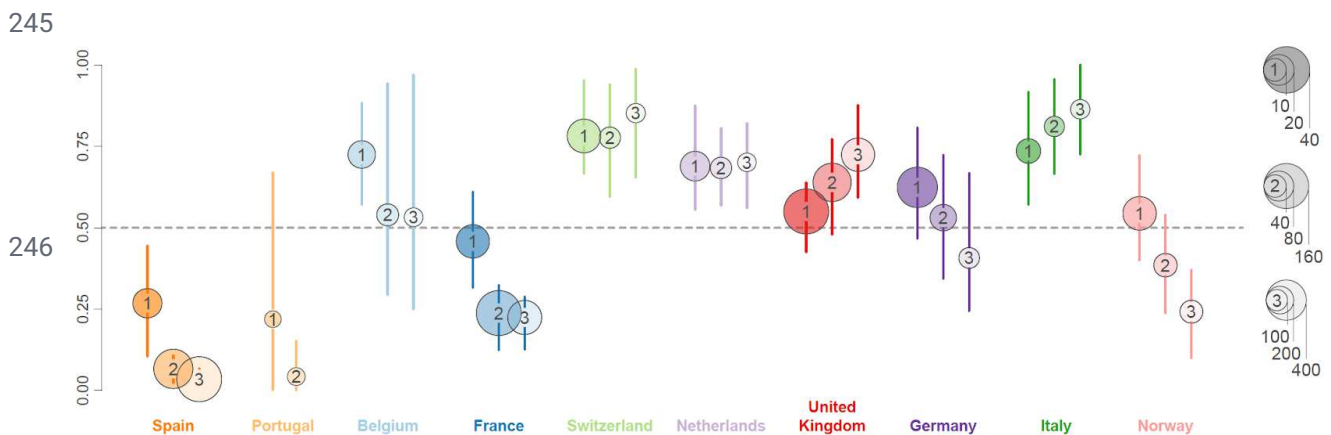
196 The upper left panel summarizes the Google mobility influx by country from the other 10 countries for two-week intervals, while the
 197 upper right panel depicts the weekly genome sampling by country used in the phylogeographic analysis. In the remaining panels, we
 198 plot for each country the ratio of introductions over the total viral flow from and to that country (for two-week intervals) and a monthly
 199 normalized entropy measure summarizing the phylogenetic structure of country-specific transmission chains. The posterior mean
 200 ratios of introductions are depicted with circles that have a size proportional to the total number of transitions from and to that country
 201 and the grey surface represents the 95% highest posterior density (HPD) intervals. The posterior mean normalized entropies and 95%
 202 HPD intervals are depicted by dotted lines. These normalized entropy measures indicate how phylogenetically structured the epidemic
 203 is in each country, and ranges from 0 (perfectly structured, e.g a single country-specific cluster) to 1 (unstructured interspersed
 204 country-specific sequences across the entire SARS-CoV-2 phylogeny). The introduction ratios and normalized entropy measures are
 205 superimposed over the number of COVID-19 cases reported for each country through time (coloured density plot). The two vertical
 206 dashed lines represent the summer time interval (June 15 and August 15, 2020) for which we subsequently evaluate introductions
 207 versus persistence (Figure 2).

208

209 *A high proportion of summer introductions is modulated by local incidence*

210 To assess the impact of summer travel on the second wave in the different countries, we use our
 211 genomic-mobility reconstruction to estimate both the number of lineages persisting in each
 212 country and the number of newly introduced lineages, and how these proliferated early in the
 213 second wave. We focus on a two-month time period between June 15th, on which many EU and

214 Schengen-area countries opened their borders to other countries, and August 15th, before which
 215 the majority of holiday return travel is expected for many countries. We identify the number of
 216 lineages circulating in each country on August 15th, and determine whether they result from a
 217 lineage that persisted since June 15th or from a unique introduction after this date, so independent
 218 of the number of descendants for this lineage on August 15th (Extended Data Figure 4). In Figure 2,
 219 we plot i) the ratio of these unique introductions over the total unique lineages (unique
 220 introductions and persisting lineages), ii) the proportion of descendant lineages on August 15th
 221 that resulted from the unique introductions over the total descendants circulating on this date and
 222 iii) the proportion of descendant tips (sampled genomes) after August 15th that resulted from the
 223 unique introductions over the total number of descendant tips (cfr. Methods and Extended Data
 224 Figure 4). The latter two proportions provide an assessment of how the unique introductions and
 225 persisting lineages evolved up to, and after, August 15th. We estimate a posterior mean proportion
 226 of unique introductions that is close to or higher than 0.5 except for Spain and Portugal. This
 227 indicates that by August 15th a relatively large fraction of circulating lineages in each country
 228 resulted from new introductions over the summer. Because we compare introductions and
 229 persistence relatively early (and irrespective of their number of descendants), and because the
 230 major variant involved appears not to be associated with increased transmissibility⁹, we consider
 231 the newly introduced lineages to be additional transmission chains that did not necessarily have to
 232 compete with persistent transmission chains for susceptibles. However, the two proportions of
 233 descendants from these introductions on August 15th and after this date measure their relative
 234 success compared to persisting lineages, indicating considerable variation in onwards
 235 transmission. The country estimates are ordered according to decreasing average incidence
 236 during the June 15 - August 15 time period, suggesting that incidence may shape the outcome of
 237 the introductions. In countries that experienced relatively high summer incidence, e.g. Spain,
 238 Portugal, Belgium and France, the introductions lead to comparatively fewer descendants on
 239 August 15th or after. Although the introductions in Norway also lead to fewer descendants despite
 240 the country having the lowest incidence, we find a significant overall association between
 241 incidence and the difference in the logit proportion of unique introductions and the logit
 242 proportion of their descendants on August 15th ($p = 0.01$). Norway may to some extent be an
 243 outlier because persistent lineages in this country could in fact be introductions from other
 244 Scandinavian countries that are not represented in our genome sample.



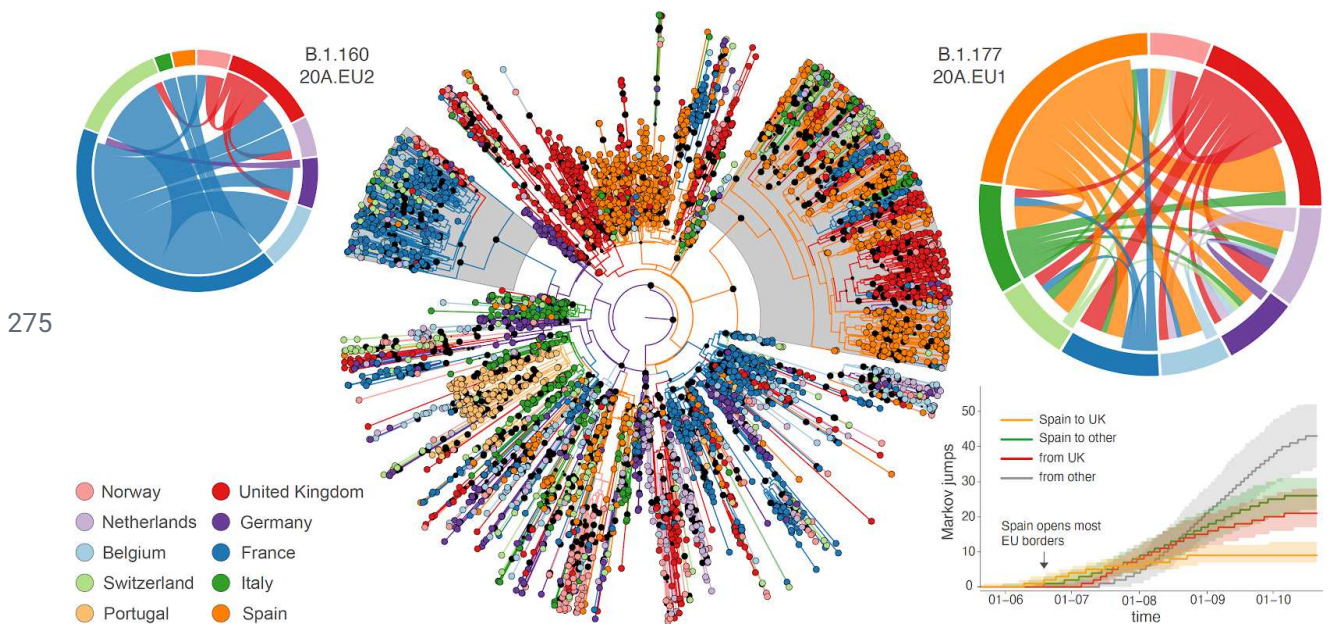
247 **Figure 2. Posterior estimates for relative importance of lineage introduction events among West European countries.** For each
 248 country, we report three summaries (posterior mean and 95% HPD intervals): (1) the ratio of unique introductions over the total
 249 number of unique persisting lineages and unique introductions between June 15 and August 15, 2020, (2) the ratio of descendant

250 lineages from these unique introduction events over the total number of descendants circulating on August 15, 2020, and (3) the ratio
 251 of descendant taxa from these unique introductions over the total number of descendant taxa sampled after August 15, 2020 (cfr.
 252 Extended Data Figure 4). The dot sizes are proportional to: (1) the total number of unique lineage introductions identified between
 253 June 15 and August 15, 2020, (2) the total number of lineages inferred on August 15, 2020, and (3) the total number of descendant
 254 sequences after August 15, 2020. The third ratio is not included for Portugal due to insufficient sequences sampled after August 15,
 255 2020.

256

257 Our estimates show a marked increase in descendants from introductions in the United Kingdom
 258 (Figure 2), with a considerable fraction of introductions originating from Spain (Extended Data
 259 Figure 5) likely reflecting the spread of B.1.177/20A.EU1 that rapidly became the most dominant
 260 strain in the United Kingdom⁹. Our analysis captures the expansion of this variant as well as that
 261 of B.1.160/20A.EU2, which together account for more than 25% of the genomes in our sample.
 262 While Spain is indeed inferred to be the origin of B.1.177/20A.EU1, the United Kingdom also
 263 considerably contributed to its spread (Figure 3). The earliest introduction from Spain to the
 264 United Kingdom is estimated around the time Spain opened most EU borders (June 21st, Figure 3).
 265 While introductions from Spain to other countries soon followed, we estimate a similar rate and
 266 amount of spread from the United Kingdom to other countries before these other countries also
 267 disseminate the virus. Whereas our sample remains limited, it illustrates a dynamic pattern of
 268 spread and the importance of the early establishment of B.1.177/20A.EU1 in the United Kingdom
 269 that served as an important secondary center of dissemination. While the United Kingdom is also
 270 to some extent involved in the spread of B.1.160/20A.EU2, this variant has been largely
 271 disseminated from France. The simple fact this variant expanded later in France and subsequently
 272 started to spread later than B.1.177/20A.EU1 (Extended Data Figure 6) may explain why the latter
 273 spread more successfully.

274



276 **Figure 3. Phylogeographic estimates of SARS-CoV-2 spread in western Europe.** The radial tree in the center represents the maximum
 277 clade credibility tree summary of the Bayesian inference. Colors correspond to the countries in the legend. Two clades corresponding
 278 to B.1.177/20A.EU1 and B.1.160/20A.EU2 are highlighted in grey. The circular migration flow plots for these variants are based on the
 279 posterior expectations of the Markov jumps. In these plots, migration flow out of a particular location starts close to the outer ring for
 280 that origin location whereas migration flow into a particular location ends more distant from the outer ring for that destination
 281 location. For B.1.177/20A.EU1, we summarize phylogeographic transitions as mean estimates with 95% HPD intervals over time for 4

282 types of Markov jumps: i) from Spain to the United Kingdom, ii) from Spain to other countries, iii) from the United Kingdom, and iv) from
283 other countries.

284 Discussion

285 In this study, we gain insight into the dynamics of SARS-CoV-2 spread through phylogeographic
286 analyses, specifically focusing on the impact of European travel during the summer of 2020.
287 Because such analyses may suffer from a lack of resolution offered by SARS-CoV-2 genomic data,
288 we integrate epidemiological and mobility data to help shape the phylodynamic process. Our
289 model supports mobility data, including both air and land transportation, as a predictor of viral
290 flow between countries. The resulting reconstructions show that the composition of lineages
291 circulating towards the end of the summer was to a significant extent shaped by introductions in
292 most of the European countries. Interestingly, the relative success of onwards transmission of the
293 introduced lineages appears to be shaped by the average summer COVID-19 incidence. In
294 countries that maintained a relatively high incidence, e.g. Spain, Portugal, Belgium and France,
295 these introductions resulted in comparably less onwards transmission over relatively short-term
296 than for the lineages that persisted in the countries over summer.

297

298 As documented by Hodcroft et al. (2020)⁹, SARS-CoV-2 spread during summer in Europe involved
299 to a large extent the B.1.177/20A.EU1 variant. This variant became the dominant strain relatively
300 early in the UK⁹, facilitating its further spread over the second half of summer. A limited sample
301 offers only a limited view of the transmission history of this variant; in depth analyses of close to
302 20,000 20A.EU1 genomes indicate hundreds of introductions to countries across Europe⁹.
303 Importantly however, there is no evidence of increased transmissibility of this variant and its
304 success is currently attributed to repeated introductions upon resuming travel with insufficient
305 effective containment strategies⁹.

306

307 Our results should be interpreted in light of several important limitations. Although about 4,000
308 genomes constitute a large dataset for Bayesian phylodynamic inference, it remains a limited
309 sample to infer detailed COVID-19 transmission dynamics. In addition, the genome data do not
310 cover all Western European countries, implying that we are missing transmission events that
311 involve unsampled countries. This may be particularly important for Norway for example, which
312 according to our mobility data, is largely connected to other Scandinavian countries. Also, the
313 mobility data are subject to limitations as these may not be representative for the population as
314 whole and their representativeness may vary by location.

315

316 The significant mixing and rapid spread to high frequencies of variants that are not associated with
317 higher transmissibility underscores the risk for pervasive spread of variants that are more
318 transmissible, like B.1.1.7 (Variant of Concern 202012/01, 20I/501Y.V1) in the United Kingdom, or
319 could be more transmissible, like B.1.351 (20H/501Y.V2) in South Africa and variant P1 in Brazil.
320 Different from the rise in frequency of 20A.EU1 due to repeated introductions in the UK, B.1.1.7
321 appears to have emerged from within the UK. Preliminary findings demonstrate a consistent
322 repeated pattern of faster epidemic growth of B.1.1.7 and provide evidence for a significant
323 transmission advantage over prior lineages⁵. The consequences for control of COVID-19 have
324 become clear for the United Kingdom. As of January 19, 2021, approximately 16,800 B.1.1.7 cases
325 have been identified in the UK and approximately 2,000 cases have been identified in 60 other
326 countries¹⁴. More intense genomic surveillance is needed across Europe to track this variant, and
327 control of COVID-19 in general would benefit from strengthened coordination. Disruptive border

328 closures (e.g. between France and UK, on December 21-22, 2020) can only serve as a short-term
329 emergency measure to put into place better strategies to prevent between-country transmission.
330 No matter how much more transmissible SARS-CoV-2 variants may be, quarantining, testing and
331 social distancing remain effective when adhered to and they will be required for some time even as
332 vaccination programs are being rolled out.

333 Methods

334 *Sequence data and subsampling*

335 We used a two-step genome data collection procedure. We first evaluated the available genomes
336 from European countries in GISAID¹⁵ on November 3, 2020. We selected genomes from Belgium,
337 France, Germany, Italy, Netherlands, Norway, Portugal, Spain, Switzerland and the United
338 Kingdom primarily based on the availability of genome data from both the first and second wave at
339 that time but also because of their high ratio of genomes to positive cases. Portugal represented
340 an exception because data for this country were limited to the first wave at that time, but we
341 included genomes from Portugal because of its potential importance as a summer travel location.

342

343 We aligned the genomes from each country using MAFFT v7.453¹⁶ and trimmed the 5' and 3'
344 ends and only retained unique sequences from each location. To further mitigate the disparities in
345 sampling, we subsampled each country proportionally to the cumulative number of cases on
346 October 21st (the most recently sampled sequence at the time) by setting an arbitrary threshold of
347 6.5 sequences per 10,000 cases, with a minimum number of 100 sequences per country. To
348 maximize the temporal and spatial coverage in each country, we binned genomes by epi-week and
349 sampled as evenly as possible, sampling from a different region within the country when available.
350 Only sequences from the B.1 lineage with the D614G mutation and exact sampling dates were
351 selected for the analyses. From the final aligned sequence set, we removed 12 potential outliers,
352 based on a root-to-tip regression on TempEst v1.5.3¹⁷ on a maximum-likelihood tree inferred with
353 IQTREE v2.0.3¹⁸, yielding a data set of 2,909 genomes (Extended Data Table 1).

354

355 Because of the nature of genome sequence accumulation, fewer recently sampled genomes were
356 available for most countries on November 3rd (relative to the case counts at this time). Because
357 our primary goal was to assess the persistence and introduction of lineages leading up to the
358 second wave, we sought to augment our data set with more recent genomes, having already
359 performed analyses on the initial data set. In the section on Bayesian evolutionary
360 reconstructions, we outline how we update these analyses accordingly. On January 5, 2021, we
361 updated our dataset by adding over 1,000 non-identical sequences collected between August 1st
362 and October 31st. For Portugal, we extended this period back to June 22nd (the most recent
363 sampling date for the previous Portuguese selection). We downloaded all new B.1 sequences with
364 the D614G mutation collected during the selected time period from GISAID and performed the
365 following subsampling. The number of genomes to add by country was obtained by raising the
366 threshold ratio of sequences/cases to 8.5 and increasing the minimum number of sequences to
367 200. To bias the temporal coverage towards more recent samples, the genomes from each country
368 were binned by week and sampled such that the number of sequences added by week was
369 proportional to an exponential function of the form $e^{t/4}$, where $t=0$ represents August 1st and $t=13$
370 is October 31st. For Portugal, we did not use this preferential sampling as we needed to include
371 close to all available genomes to raise the number of genomes to 200. The sampled sequences
372 were then deduplicated and outliers were removed as described in the previous section. With the
373 additional selection of 1,050 genomes, we arrived at a data set of 3,959 genomes (Extended Data
374 Table 1).

375

376 *Mobility data*

377 We analysed four different mobility/connectivity metrics: air traffic flows, a social connectedness
378 index provided by Facebook, as well as aggregate Google and Facebook international mobility
379 data. Air traffic flow data were obtained from the International Air Transport Association
380 (<http://www.iata.org>) and based on the number of origin-destination tickets while also taking into
381 account connections at intermediate airports (Gilbert et al. 2020). We used monthly air traffic
382 data between the 10 western European countries under investigation for the time period between
383 January 2020 and October 2020. The social connectedness index (SCI) is an anonymized snapshot
384 of active Facebook users and their friendship networks to measure the intensity of social
385 connectedness between countries (<https://data.humdata.org/>). In practice, the SCI measures the
386 relative probability of a Facebook friendship link between two users of the application in different
387 countries. We used the SCI calculated for the 10 Western european countries as of August 2020.
388

389 The Google COVID-19 Aggregated Mobility Research Dataset contains anonymized mobility
390 flows aggregated over users who have turned on the Location History setting, which is off by
391 default. To produce this dataset, machine learning is applied to logs data to automatically segment
392 it into semantic trips ¹⁹. To provide strong privacy guarantees, all trips were anonymized and
393 aggregated using a differentially private mechanism ²⁰ to aggregate flows over time (see
394 <https://policies.google.com/technologies/anonymization>). This research was done on the resulting
395 heavily aggregated and differentially private data. No individual user data was ever manually
396 inspected, only heavily aggregated flows of large populations were handled. All anonymized trips
397 were processed in aggregate to extract their origin and destination location and time. For example,
398 if users traveled from location a to location b within time interval t , the corresponding cell (a, b, t)
399 in the tensor would be $n \pm \eta$, where η is Laplacian noise. The automated Laplace mechanism adds
400 random noise drawn from a zero-mean Laplace distribution and yields (ϵ, δ) -differential privacy
401 guarantee of $\epsilon = 0.66$ and $\delta = 2.1 \times 10^{-29}$ per metric. The parameter ϵ controls the noise intensity
402 in terms of its variance, while δ represents the deviation from pure ϵ -privacy. The closer they are
403 to zero, the stronger the privacy guarantees. We used aggregated mobility flows between the 10
404 western European countries and summarized by two-week or monthly time periods between
405 January 2020 and October 2020.

406

407 Finally, we also considered international mobility data from Facebook mobility data as an
408 alternative to Google mobility data. These data are based on numbers of Facebook users moving
409 over large distances, like air or train travel. Counts of international travel patterns are updated
410 daily based only on users who have opted into sharing precise location data from their device with
411 the Facebook mobile app through location services. Also in this case, we used aggregated mobility
412 flows between the 10 western European countries summarized by month between January 2020
413 and October 2020. Because international aggregate mobility data obtained from Google and
414 Facebook are highly correlated (monthly Spearman correlation ranging from 0.84 to 0.92;
415 Extended Figure 7), we only included the Google aggregate mobility data as a covariate in the
416 phylogeographic analyses.

417

418 *Bayesian evolutionary reconstructions*

419 - Joint sequence-trait inference with a time-homogeneous GLM diffusion model

420 We performed Bayesian evolutionary reconstruction of timed phylogeographic history using
421 BEAST 1.10²¹ incorporating genome sequences, their country and date of sampling,
422 epidemiological and mobility/connectivity data. Because of the relatively low degree of resolution
423 offered by the sequence data, our full probabilistic model specification focuses on i) relatively
424 simple model specifications and ii) informing parameters by additional non-genetic data sources.
425 We modeled sequence evolution using an HKY85 nucleotide substitution model with
426 gamma-distributed rate variation among sites and a strict molecular clock model. Our genome set
427 includes three genomes from an early outbreak in Bavaria, which was caused by an independent
428 introduction from China^{22,23}. We therefore constrained these genomes as an outgroup in the
429 analysis, which according to root-to-tip regression plots as a function of sampling time resulted in
430 a better correlation coefficient/R-squared compared to the best-fitting root under the heuristic
431 mean residual squared criterion (Extended Figure 8)¹⁷.

432

433 As a coalescent tree prior, we modeled the effective population size trajectory as a piecewise
434 constant function that changes values at pre-specified times^{following 24}, with log population sizes
435 modelled as a deterministic function of log COVID-19 case counts^{following 25}. This reduces the
436 nonparametric skygrid parameterization to a generalized linear model (GLM) formulation with an
437 estimable regression intercept and coefficient. Specifically, we used two-week intervals and
438 specified as a covariate the total case counts over these time intervals for the 10 countries of
439 sampling. The earliest interval with non-zero cases counts was from 2020-01-14 to 2020-01-28;
440 before 2020-01-14, the log-transformed and standardized case count covariate was set to the
441 equivalent of 1 case. Case count data were obtained from
442 <https://www.ecdc.europa.eu/en/covid-19/data>.

443

444 Similar to sequence evolution, we modelled the process of transitioning through discrete location
445 states (countries of sampling) according to a continuous-time Markov chain (CTMC)²⁶. We
446 employed a parameterization that models the log transition rates as a log linear function of
447 mobility/connectivity covariates²⁷. As covariates we considered Facebook's SCI, air
448 transportation data and mobility data. For the two time-variable mobility measures, we used the
449 average of the log-transformed and standardized monthly mobility measures as a single covariate
450 in our time-homogeneous phylogeographic GLM model. In addition to estimating the contribution
451 (effect size) of each covariate in this GLM, we also estimated their inclusion probabilities through a
452 spike-and-slab procedure.

453

454 We performed inference under the full model specification using Markov chain Monte Carlo
455 (MCMC) sampling and used the BEAGLE library v3²⁸ to increase computational performance. We
456 specified standard transition kernels on all parameters, except for the regression coefficients of
457 the piecewise-constant coalescent GLM model. For these parameters, we implemented new
458 Hamiltonian Monte Carlo (HMC) transition kernels to improve sampling efficiency. These kernels
459 use principles from Hamiltonian dynamics and their approximate energy conserving properties to
460 reduce correlation between successive sampled states, but require computation of the gradient of
461 the model log-posterior with respect to the parameters of interest, in addition to efficient
462 evaluation of the log-posterior that BEAGLE provides. To accomplish this, we extended our
463 previous analytic derivation of the gradient of the log-density from the skygrid coalescent model

464 with respect to the log-population-sizes ²⁹ to now be with respect to the regression coefficients
465 using the chain rule and their regression design matrix.

466

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

473

474 - Data augmentation through online BEAST

475 As we updated our dataset following initial analyses of the 2,909 genome collection using the
476 approach discussed in the previous subsection, we sought to capitalize on these efforts to limit the
477 burn-in for subsequent analyses of the 3,959 dataset. Specifically, we adopted the distance-based
478 procedure to insert new taxa into a time-measured phylogenetic tree sample as implemented in
479 the BEAST framework for online inference ³⁰. We subsequently use the augmented tree as starting
480 tree for the analyses of the updated dataset.

481

482 - Time-inhomogeneous reconstructions

483 To accommodate the time-variability of the mobility measures, we constructed epoch model
484 extensions of the discrete phylogeography approach that allow specifying arbitrary intervals over
485 the evolutionary history and associating them with different model parameterizations ³¹. As a
486 complement to testing covariates of spatial diffusion using a time-homogeneous model, we used
487 the epoch extension to specify monthly intervals allowing us to incorporate monthly mobility
488 matrices (air transportation data was only available as monthly numbers), but assuming
489 time-homogeneous effect sizes and inclusion probabilities. Monthly covariates were again
490 log-transformed and standardized after adding a pseudocount to each entry in the monthly
491 matrices.

492

493 In addition, we performed another analysis in which we relaxed the constant-through-time
494 inclusion probability of the covariates. In this model specification, each interval is associated with
495 a specific set of indicator variables to represent the inclusion/exclusion of covariates, but we pool
496 information about predictor inclusion across the intervals using hierarchical graph modelling ³².
497 This approach uses a set of indicator variables to model covariate inclusion at the hierarchical
498 level but allows interval-specific inclusion or predictors to diverge from the hierarchical level with
499 a non-zero probability (with the number of differences modelled as a binomial distribution, ³²),
500 which set to 0.10 in our case. We estimated hierarchical and interval-level inclusion using
501 spike-and-slab.

502

503 Finally, we performed an analysis using the time-inhomogeneous model in which the
504 interval-specific transition rates are modelled as a function of the single covariate that is
505 supported by the analyses above leveraging aggregate mobility. We incorporated more variability
506 through time by specifying two-week intervals (similar to the coalescent GLM interval
507 specification). The time-inhomogeneous GLM approach we employ allows modelling relative

508 differences in transition rates, but also the overall rate of migration between countries varies
509 through time and likely more than the relative preferences of migration. Therefore, we further
510 extended this model by incorporating a time-inhomogeneous overall CTMC rate scaler and
511 parameterize it as a log linear function of the total monthly between-country log-transformed and
512 standardized mobility. To generate realisations of the discrete location CTMC process and obtain
513 estimates of the transitions (Markov jumps) between states under this model, we employed
514 posterior inference of the complete Markov jump history through time^{27,33}.

515

516 While the epoch model allows us to flexibly accommodate time-variable spatial dynamics, it
517 considerably increases the computational burden associated with likelihood evaluations. In order
518 to efficiently draw inference under this model for our large data set, we fit the
519 time-inhomogeneous spatial diffusion process to a set of trees inferred under the
520 time-homogeneous GLM diffusion model described above. Although likelihood evaluations remain
521 computationally expensive, even with the speed-up offered by GPU computation with BEAGLE,
522 eliminating simultaneous tree estimation tremendously reduces parameter-space, requiring only
523 modest MCMC chain lengths to adequately explore it.

524

525 - Posterior Summaries

526 We assessed MCMC mixing (e.g. using ESSs) and summarized continuous parameter estimates
527 using Tracer v1.7.1³⁴. Credible intervals were computed as 95% HPD intervals. Trees were
528 visualized using FigTree v1.4.4 (available at <https://github.com/rambaut/figtree/releases>). In terms
529 of phylogeographic estimates, we mainly focused on i) transitions to each location and from each
530 location (based on Markov jump estimates) instead of pairwise transitions, ii) ratios of these
531 transitions and iii) how these transitions structured transmission chains in individual countries.
532 Transitions to each and from each location avoid drawing conclusions about direct migration
533 between countries, which can be tenuous given the incomplete genomes coverage of Europe,
534 while their ratios avoid using absolute numbers of transitions, which are highly sample-dependent.
535 Phylogeographic inference is limited to reconstructing the transitions in the ancestral history of a
536 sample of sequences, which will only be a small fraction of the actual migration events especially
537 when these events result in insufficient onwards transmission to be captured in our limited
538 sample. In addition, SARS-CoV-2 genome data can be poorly resolved and identical genomes in
539 different locations are consistent with hypotheses that involve both a sparse and a rich number of
540 virus flows between these locations. As the data hold little information to distinguish these
541 hypotheses, we only consider sparse scenario's by including only unique sequences for each
542 location. A joint inference of sequence evolution and discrete spatial diffusion would err on the
543 side of sparse hypotheses anyway because it will tend to cluster identical sequences that share a
544 location. Despite the general underestimation of spatial dispersal, a phylogeographic inference is
545 still likely to capture the transition events with important onward transmission, and evaluating the
546 importance of such events relative to persistence is a major focus of this study.

547

548 We provide three new tree sample tools in the BEAST codebase available at
549 <https://github.com/beast-dev/beast-mcmc> to obtain posterior summaries of location transition
550 histories using posterior tree distributions annotated with Markov jumps:

551

552 ● *TreeMarkovJumpHistoryAnalyzer* allows collecting Markov jumps and their timings from a
553 posterior tree distribution annotated with Markov jumps histories in a .csv file for further
554 analyses.

555
556 ● *TreeStateTimeSummarizer* decomposes the total tree time into the times associated with
557 contiguous partitions of a tree associated with a particular location state, with the
558 partitions determined by the Markov jumps. An arbitrary lower and upper time boundary
559 can be specified to restrict the summary to a particular time interval in the evolutionary
560 history. We use the time estimates for the separate partitions associated with each state to
561 calculate an entropy measure that summarizes the genetic make-up of country-specific
562 transmission chains. Specifically, we use for each location a normalized Shannon entropy:

563
$$- \frac{1}{\ln(n)} \sum_i^n p_i \ln(p_i) , \quad (1)$$

564 Where p_i is the proportion of time associated with that location for partition i of a
565 phylogeographic tree and n represents the number of partitions for that location in the
566 tree.

567
568 ● *PersistenceSummarizer* also uses posterior tree distributions annotated with Markov jumps
569 to summarize the number of lineages at a particular point in time (evaluation time, T_e , cfr.
570 Extended Figure 5), which location states they are associated with, since what time point in
571 the past they have maintained that state and how many sampled descendants they have
572 after time T_e (Extended Figure 5). In addition, it allows identifying how long these lineages
573 have circulated independently prior to T_e , so before sharing common ancestry with other
574 lineages that maintained the same location state. This information allows us to determine
575 how many lineages are circulating at T_e that stem either from a unique persistent lineage
576 (maintaining the same location states) or unique introduction event since a particular time
577 prior to T_e . The association between incidence and the difference in the logit proportion of
578 unique introductions and the logit proportion of their descendants on August 15th was
579 evaluated using p -value obtained by a linear regression analysis.

580 **Data availability**

581 BEAST XML input files are available at

582 https://github.com/phylogeography/SARS-CoV-2_EUR_PHYLOGEOGRAPHY

583

584 The SARS-CoV-2 genome data required for running these xmls can be downloaded from
585 <https://www.gisaid.org>. The Google COVID-19 Aggregated Mobility Research Dataset used for
586 this study is available with permission from Google LLC. The Facebook mobility data can be
587 requested from Facebook (<https://dataforgood.fb.com/>). COVID-19 incidence data was obtained
588 from <https://www.ecdc.europa.eu/en/covid-19/data>.

589

590 **Code availability**

591 The code for running BEAST analyses is available in the hmc-develop branch of the BEAST
592 codebase available at <https://github.com/beast-dev/beast-mcmc>. The tools
593 *TreeMarkovJumpHistoryAnalyzer*, *TreeStateTimeSummarizer* and *PersistenceSummarizer* are available
594 from the master branch in the same codebase.

595

596 **Acknowledgments**

597 We would like to thank all the authors who have kindly shared genome data on GISAID, and we
598 have included a table (Extended Table 3) acknowledging the authors and institutes involved.

599

600 The research leading to these results has received funding from the European Research Council
601 under the European Union's Horizon 2020 research and innovation programme (grant agreement
602 no. 725422-ReservoirDOCS) and from the European Union's Horizon 2020 project MOOD (grant
603 agreement no. 874850), and the Bill & Melinda Gates Foundation (OPP1094793 and
604 INV-024911). The Artic Network receives funding from the Wellcome Trust through project
605 206298/Z/17/Z. PL acknowledges support by the Research Foundation - Flanders ('Fonds voor
606 Wetenschappelijk Onderzoek - Vlaanderen', G066215N, GOD5117N and GOB9317N). GB
607 acknowledges support from the 'Interne Fondsen KU Leuven' / Internal Funds KU Leuven under
608 grant agreement C14/18/094, and the Research Foundation - Flanders ('Fonds voor
609 Wetenschappelijk Onderzoek - Vlaanderen', GOE1420N). MAS acknowledges support from
610 National Institutes of Health grant U19 AI135995 and R01 AI153044. SD is supported by the
611 *Fonds National de la Recherche Scientifique* (FNRS, Belgium). We also gratefully acknowledge
612 support from NVIDIA Corporation with the donation of parallel computing resources used for this
613 research.

614

615 **Author contributions**

616 P.L. & S.D. designed the study, performed analyses and drafted the manuscript. V.C., C.P. and A.S.
617 provided and analyzed data. S.H., F.V., N.R., S.L. & A.T. compiled and analyzed data. A.L. contributed
618 data. G.B. performed data analyses. M.S.G., X.J. and M.A.S. developed statistical inference
619 methodology. All authors contributed to interpreting and reviewing the manuscript.

620

621 **Competing Interests**

622 The authors declare no competing interests.

623

624 **Materials and correspondence**

625 philippe.lemey@kuleuven.be & simon.dellicour@ulb.ac.be

626

627 References

628

- 629 1. Riley, S. et al. High prevalence of SARS-CoV-2 swab positivity and increasing R number in England
630 during October 2020: REACT-1 round 6 interim report. *bioRxiv* (2020)
631 doi:10.1101/2020.10.30.20223123.
- 632 2. Data on 14-day notification rate of new COVID-19 cases and deaths.
633 [https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-](https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19)
634 [19](https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19) (2021).
- 635 3. COVID-19 situation update for the EU/EEA, as of week 3, updated 28 January 2021.
636 <https://www.ecdc.europa.eu/en/cases-2019-ncov-eueea>.
- 637 4. Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli,
638 Tom Connor, Tom Peacock, David L Robertson, Erik Volz, on behalf of COVID-19 Genomics
639 Consortium UK (CoG-UK). Preliminary genomic characterisation of an emergent SARS-CoV-2
640 lineage in the UK defined by a novel set of spike mutations. *virological.org*
641 [https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-](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)
642 [e-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563](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).
- 643 5. Volz, E. et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking
644 epidemiological and genetic data. *bioRxiv* (2021) doi:10.1101/2020.12.30.20249034.
- 645 6. Neher, R. A., Dyrdak, R., Druelle, V., Hodcroft, E. B. & Albert, J. Potential impact of seasonal
646 forcing on a SARS-CoV-2 pandemic. *Swiss Med. Wkly* 150, w20224 (2020).
- 647 7. McKee, M. A European roadmap out of the covid-19 pandemic. *BMJ* 369, m1556 (2020).
- 648 8. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic
649 epidemiology. *Nat Microbiol* 5, 1403–1407 (2020).
- 650 9. Hodcroft, E. B. et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the
651 summer of 2020. *medRxiv* (2020) doi:10.1101/2020.10.25.20219063.
- 652 10. Morel, B. et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular Biology and*
653 *Evolution* (2020) doi:10.1093/molbev/msaa314.
- 654 11. Kuchler, T., Russel, D. & Stroebel, J. The Geographic Spread of COVID-19 Correlates with the
655 Structure of Social Networks as Measured by Facebook. (2020) doi:10.3386/w26990.
- 656 12. Dellicour, S. et al. A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History
657 and Dynamics of SARS-CoV-2 Lineages. *Mol. Biol. Evol.* (2020) doi:10.1093/molbev/msaa284.
- 658 13. du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK.
659 *Science* (2021) doi:10.1126/science.abf2946.
- 660 14. European Centre for Disease Prevention and Control (ECDC). Risk related to the spread of new
661 SARS-CoV-2 variants of concern in the EU/EEA – first update. (2021).
- 662 15. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to
663 reality. *Euro Surveill.* 22, (2017).
- 664 16. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods*
665 *Mol. Biol.* 537, 39–64 (2009).
- 666 17. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of
667 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2, vew007 (2016).
- 668 18. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
669 Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534 (2020).

- 670 19. Bassolas, A. et al. Hierarchical organization of urban mobility and its connection with city
671 livability. *Nat. Commun.* 10, 4817 (2019).
- 672 20. Wilson, R. J. et al. Differentially Private SQL with Bounded User Contribution. *Proceedings on*
673 *Privacy Enhancing Technologies 2020*, 230–250.
- 674 21. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.
675 *Virus Evol* 4, vey016 (2018).
- 676 22. Böhmer, M. M. et al. Investigation of a COVID-19 outbreak in Germany resulting from a single
677 travel-associated primary case: a case series. *Lancet Infect. Dis.* 20, 920–928 (2020).
- 678 23. Worobey, M. et al. The emergence of SARS-CoV-2 in Europe and North America. *Science* 370,
679 564–570 (2020).
- 680 24. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model
681 for multiple loci. *Mol. Biol. Evol.* 30, 713–724 (2013).
- 682 25. Faria, N. R. et al. Genomic and epidemiological monitoring of yellow fever virus transmission
683 potential. *Science* 361, 894–899 (2018).
- 684 26. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its
685 roots. *PLoS Comput. Biol.* 5, e1000520 (2009).
- 686 27. Lemey, P. et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global
687 Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog.* 10, e1003932 (2014).
- 688 28. Ayres, D. L. et al. BEAGLE 3: Improved performance, scaling, and usability for a high-performance
689 computing library for statistical phylogenetics. *Syst. Biol.* 68, 1052–1061 (2019).
- 690 29. Baele, G., Gill, M. S., Lemey, P. & Suchard, M. A. Hamiltonian Monte Carlo sampling to estimate
691 past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics
692 framework. *Wellcome Open Res* 5, 53 (2020).
- 693 30. Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A. & Baele, G. Online Bayesian Phylodynamic
694 Inference in BEAST with Application to Epidemic Reconstruction. *Mol. Biol. Evol.* 37, 1832–1842
695 (2020).
- 696 31. Bielejec, F., Lemey, P., Baele, G., Rambaut, A. & Suchard, M. A. Inferring heterogeneous
697 evolutionary processes through time: from sequence substitution to phylogeography. *Syst. Biol.*
698 63, 493–504 (2014).
- 699 32. Cybis, G. B., Sinsheimer, J. S., Lemey, P. & Suchard, M. A. Graph hierarchies for phylogeography.
700 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368, 20120206 (2013).
- 701 33. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Philos.*
702 *Trans. R. Soc. Lond. B Biol. Sci.* 363, 2985–2995 (2008).
- 703 34. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in
704 Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* 67, 901–904 (2018).
- 705

706

707 **Extended Data.**

708 **Extended Data Table 1. Genome sampling by country, collected on Nov. 3rd, 2020, and updated**
709 **on Jan 5th, 2021.**

country	genomes (Nov. 3rd, 2020)	genomes (Jan 5th, 2021)	total
Belgium	183	53	236
France	600	167	767
Germany	246	75	321
Italy	281	75	356
The Netherlands	159	47	206
Norway	100	92	192
Portugal	100	100	200
Spain	647	191	838
Switzerland	100	98	198
The United Kingdom	493	152	645
total	2909	1050	3959

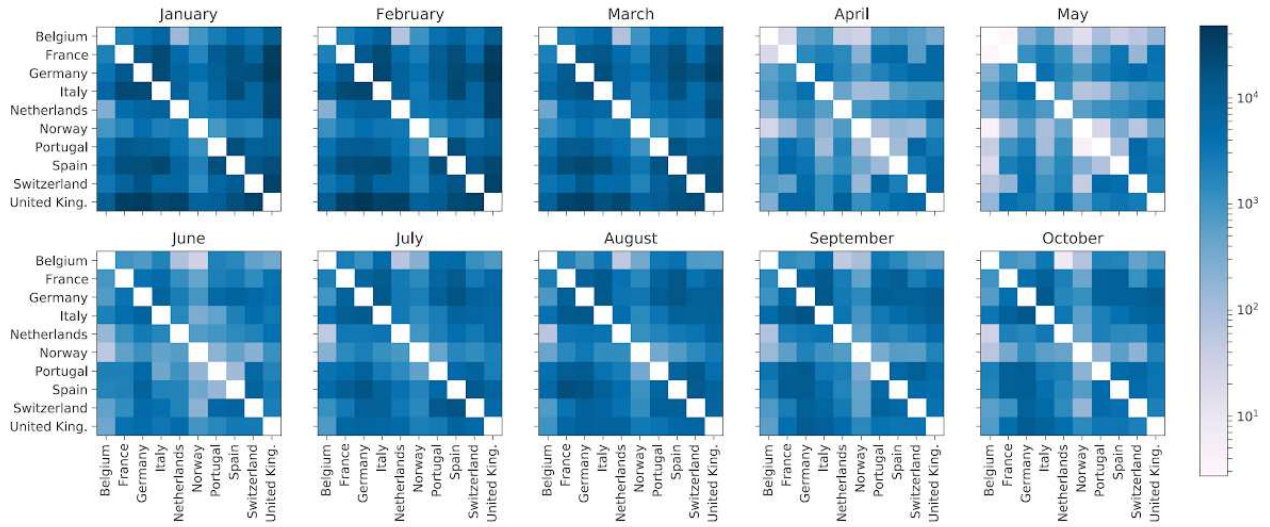
710

711 **Extended Data Table 2. mobility to or from each country within our 10 country sample as the**
712 **percentage of the total between-country mobility within Europe.**

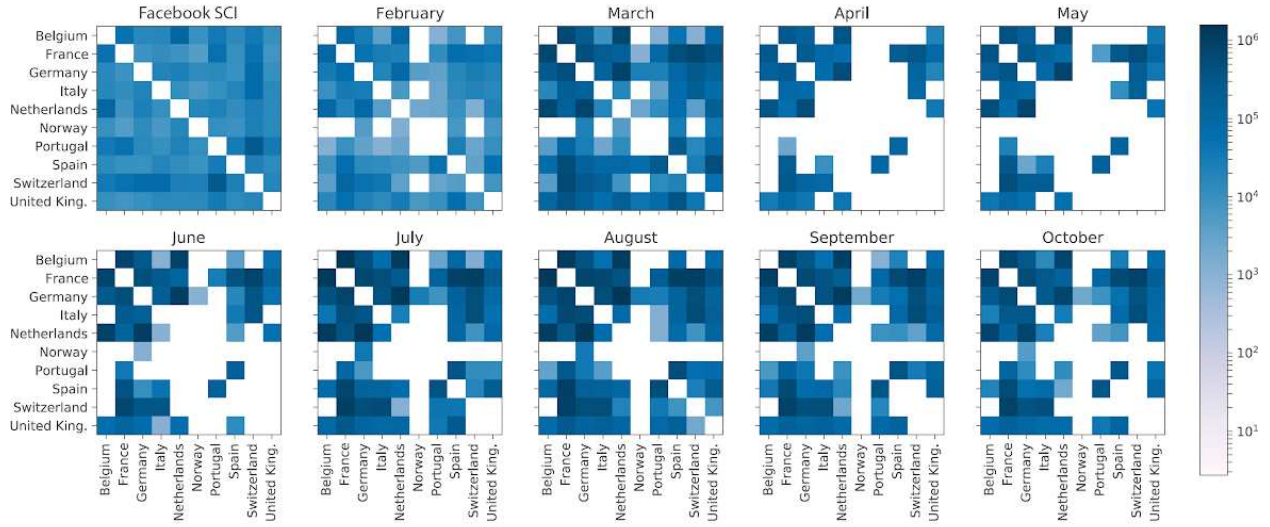
country	Mobility percentage
Belgium	87.2
France	89.5
Germany	63.9
Italy	64.8
The Netherlands	93.2
Norway	27.1
Portugal	94.0
Spain	90.3
Switzerland	84.8

The United Kingdom	48.6
--------------------	------

A. International air traffic data

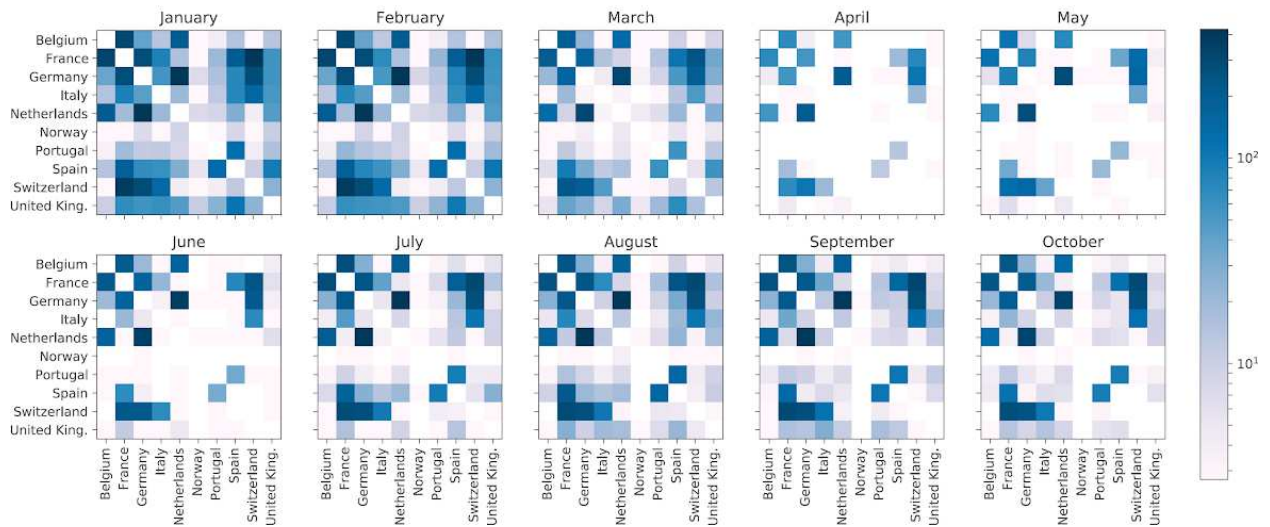


B. International Facebook mobility data (and SCI)



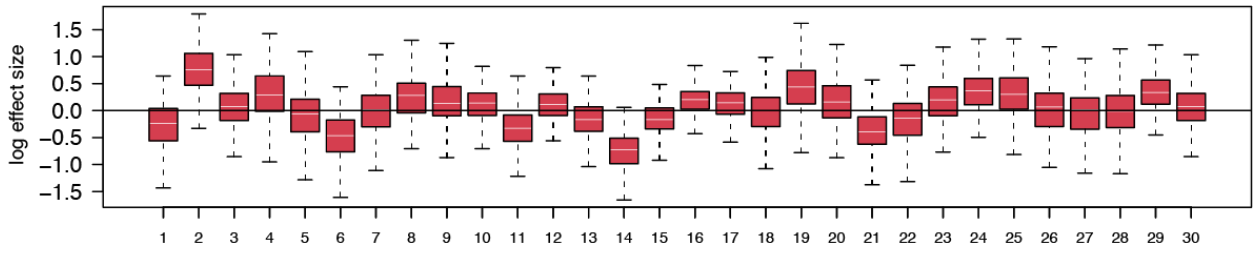
714

C. International Google mobility data

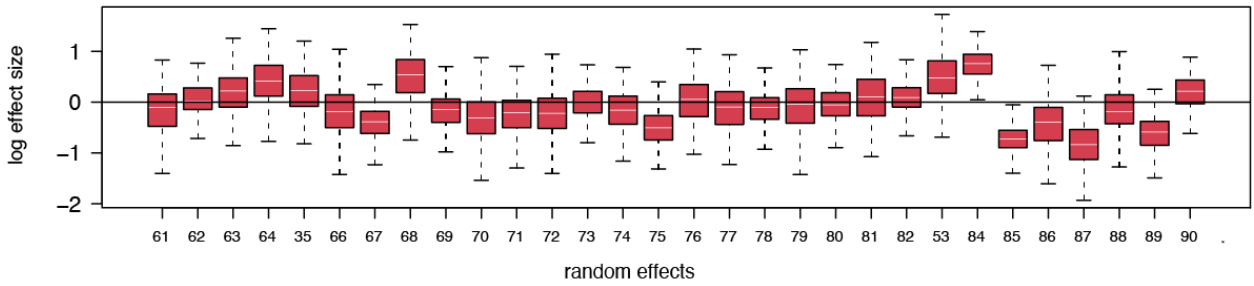
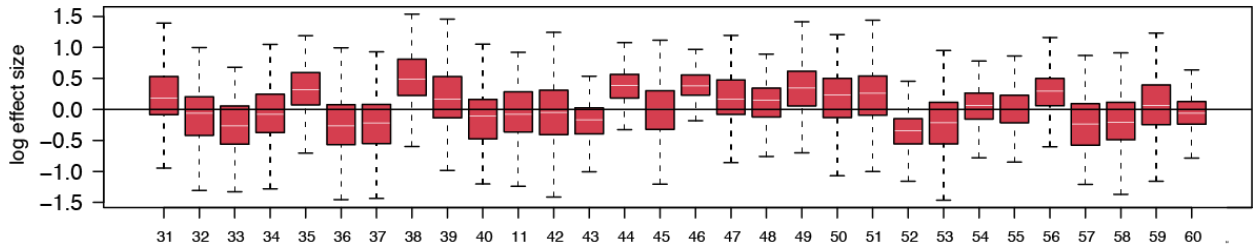


715 **Extended Data Figure 1. Monthly international mobility data matrices: international air traffic data, international Facebook mobility**
 716 **data, and international mobility data. For Facebook data, we also report the single social connectedness index matrix (SCI, B).**

717

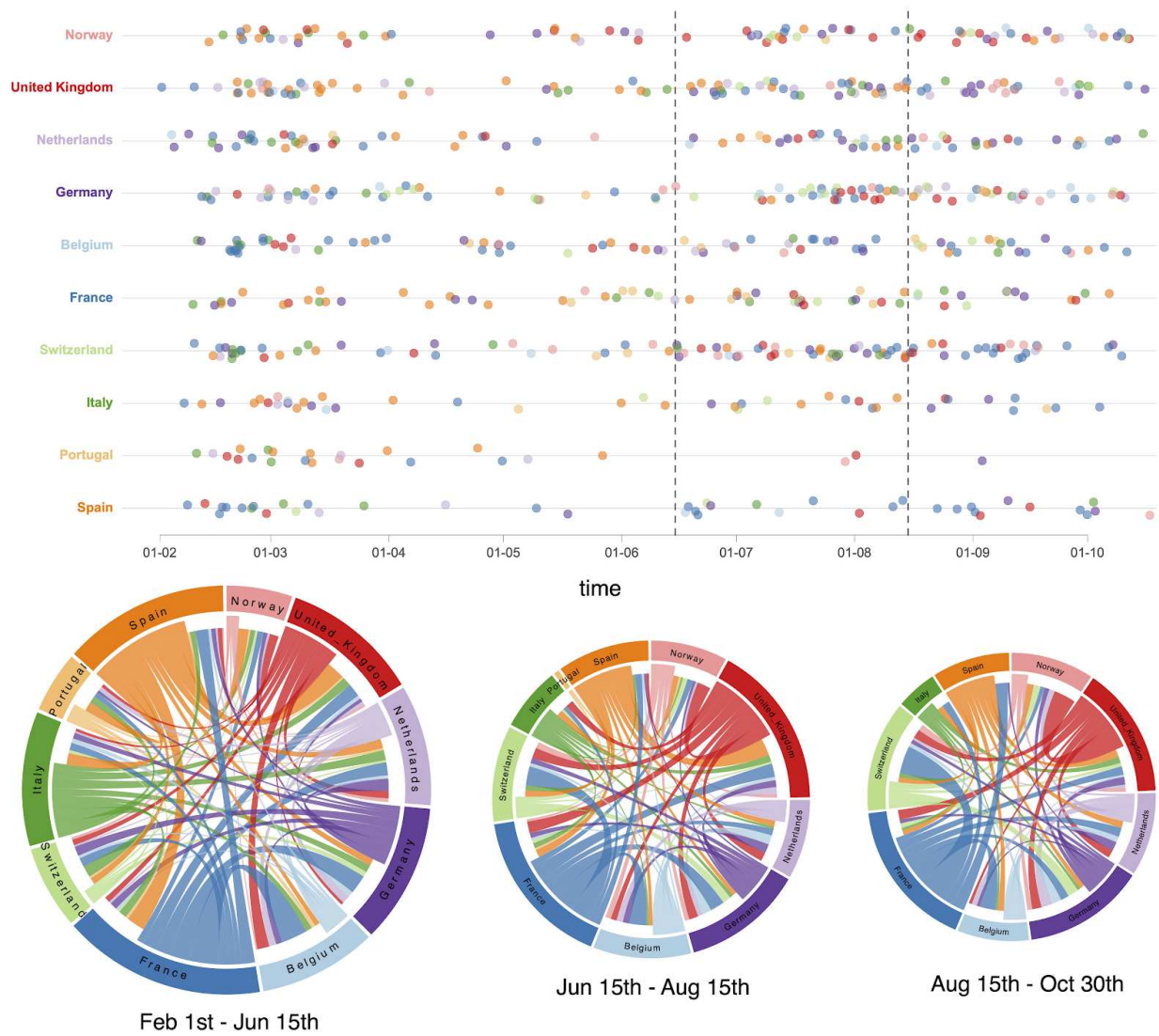


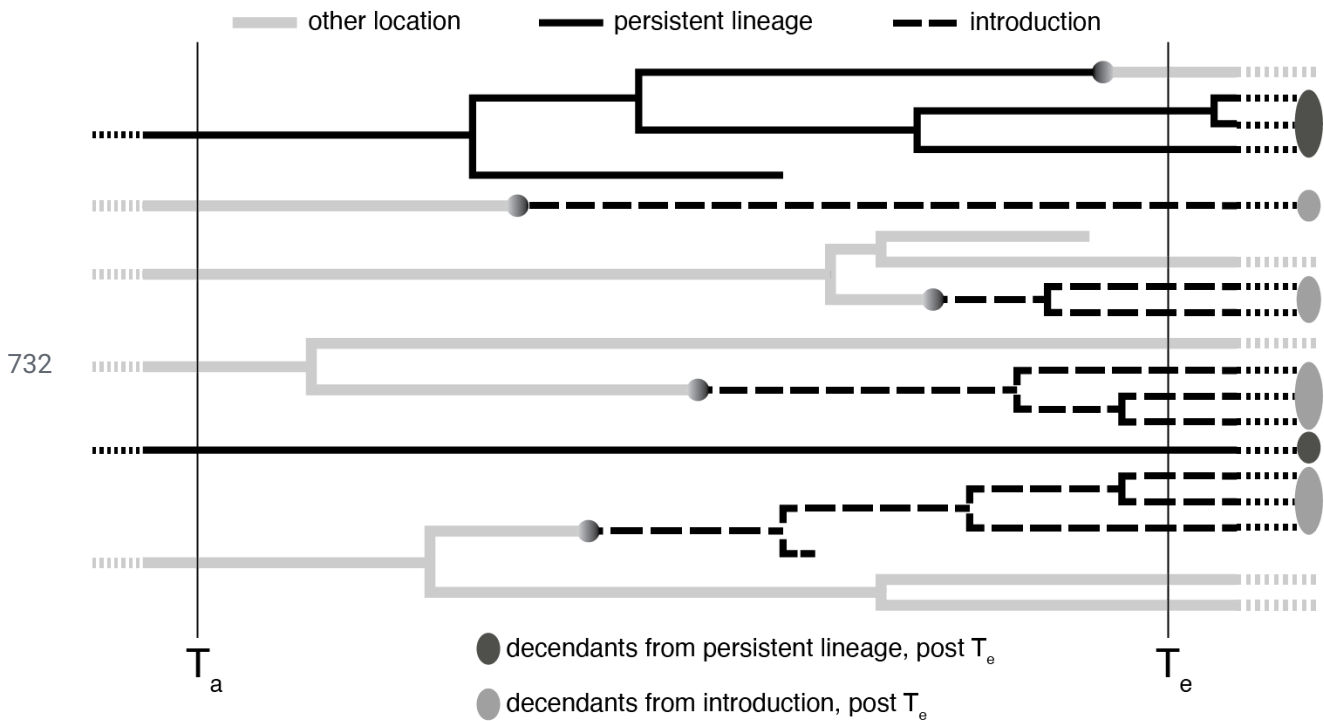
718



719 **Extended Data Figure 2. Posterior summary of the GLM random effects.** The posterior distribution for each random effect in log space
720 is summarized as an error bar plot. The mean effect size is represented by a white horizontal line while the whiskers represent the 95%
721 HPD intervals.

722



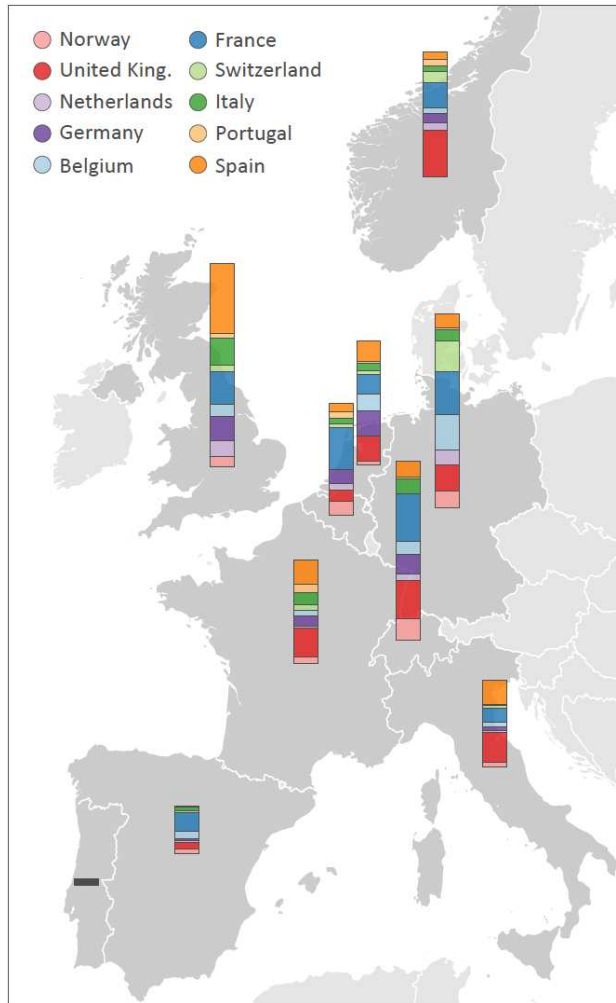


733

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

743

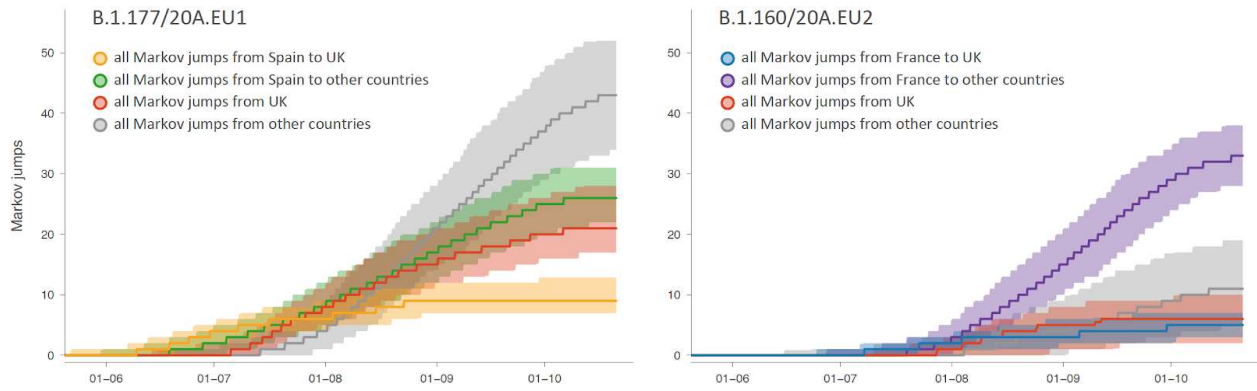
744



745
746
747
748

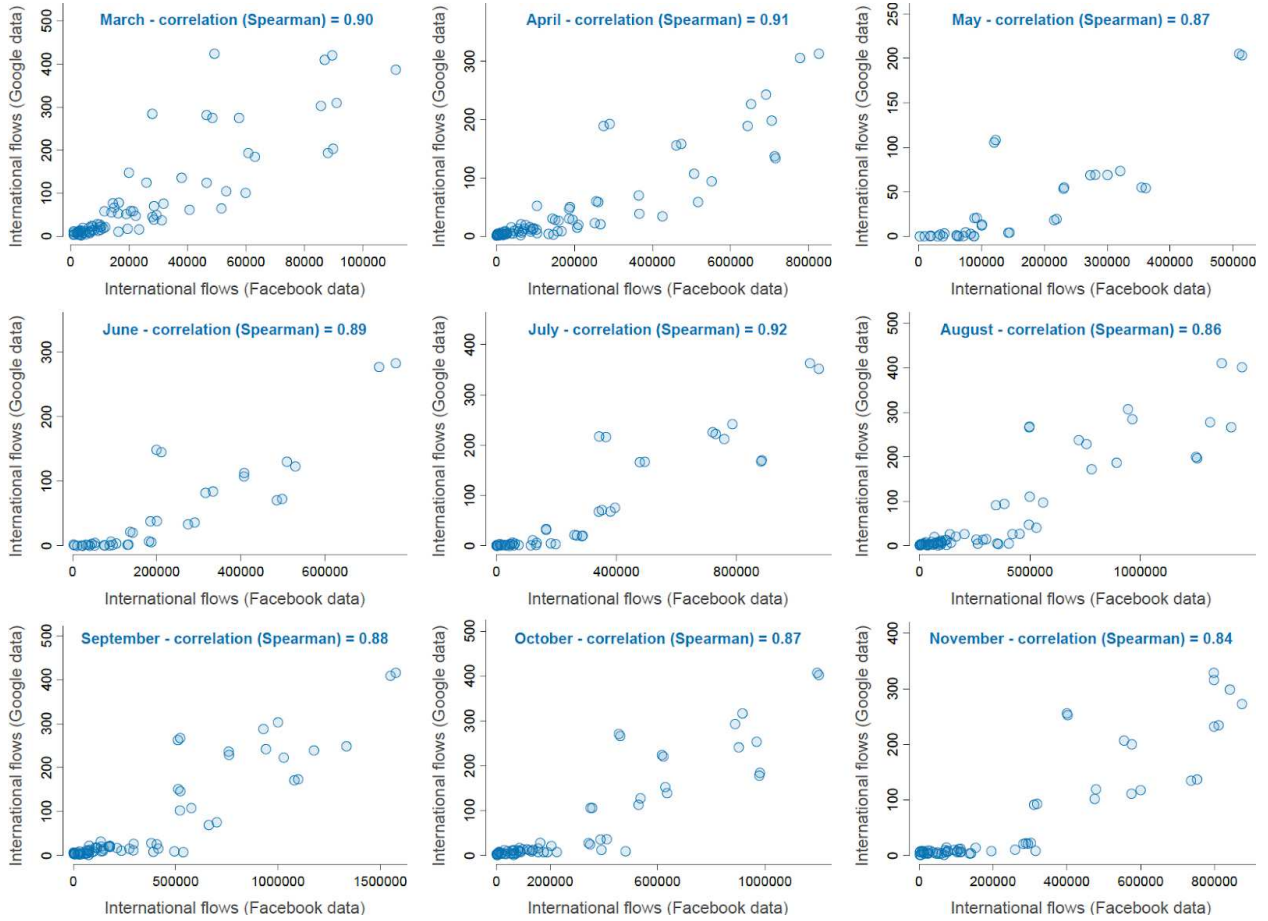
Extended Data Figure 5. Estimated geographic origin of viral influx over the summer n(June 15th - August 15th, 2020) in each country. Each barplot summarizes the posterior Markov jump estimates into a specific country.

749



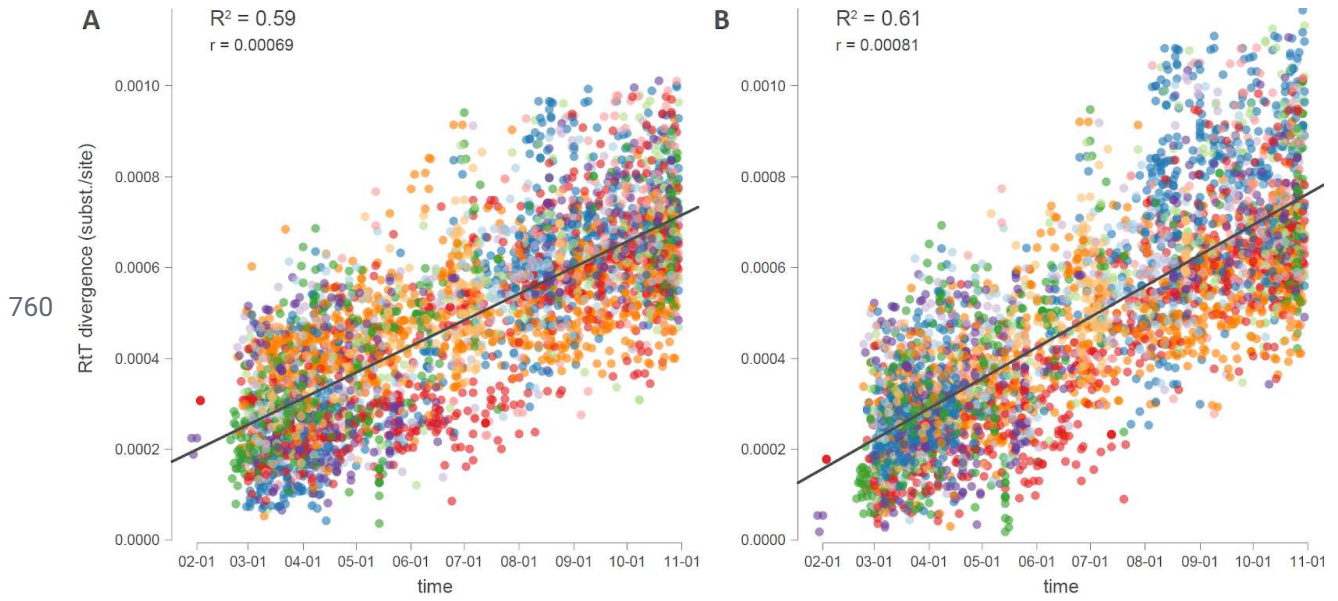
750 **Extended Data Figure 6. Phylogeographic transitions for lineages B1.1777/20A.EU1 and B1.160/20A.EU2.** Cumulative
751 phylogeographic transitions are summarized as posterior mean estimates with 95% HPD intervals over time for 4 types of Markov
752 jumps. For B1.1777/20A.EU1: i) from Spain to the United Kingdom (UK), ii) from Spain to other countries, iil) from the UK, and iv) from
753 other countries; For B1.160/20A.EU2: i) from France to the UK, ii) from France to other countries, iil) from the UK, and iv) from other
754 countries.

755



756

757 **Extended Data Figure 7. Comparison between Google and Facebook aggregate international mobility data.** We summarize monthly
758 correlations using scatter plots and Spearman's rank correlation. Each dot in the scatter plots corresponds to a specific pair of
759 European countries considered in our study.



761 **Extended Data Figure 8. Root-to-tip divergence as a function of sampling time for the 3959 genome data set with a different rooting**
 762 **of the same maximum likelihood tree. A.** Tree rooted according to the best-fitting root under the heuristic residual mean squared
 763 **criterion. B.** Tree rooted along the branch leading to the cluster of 3 Bavarian genomes that resulted from an independent introduction
 764 **into Europe.**

Figures

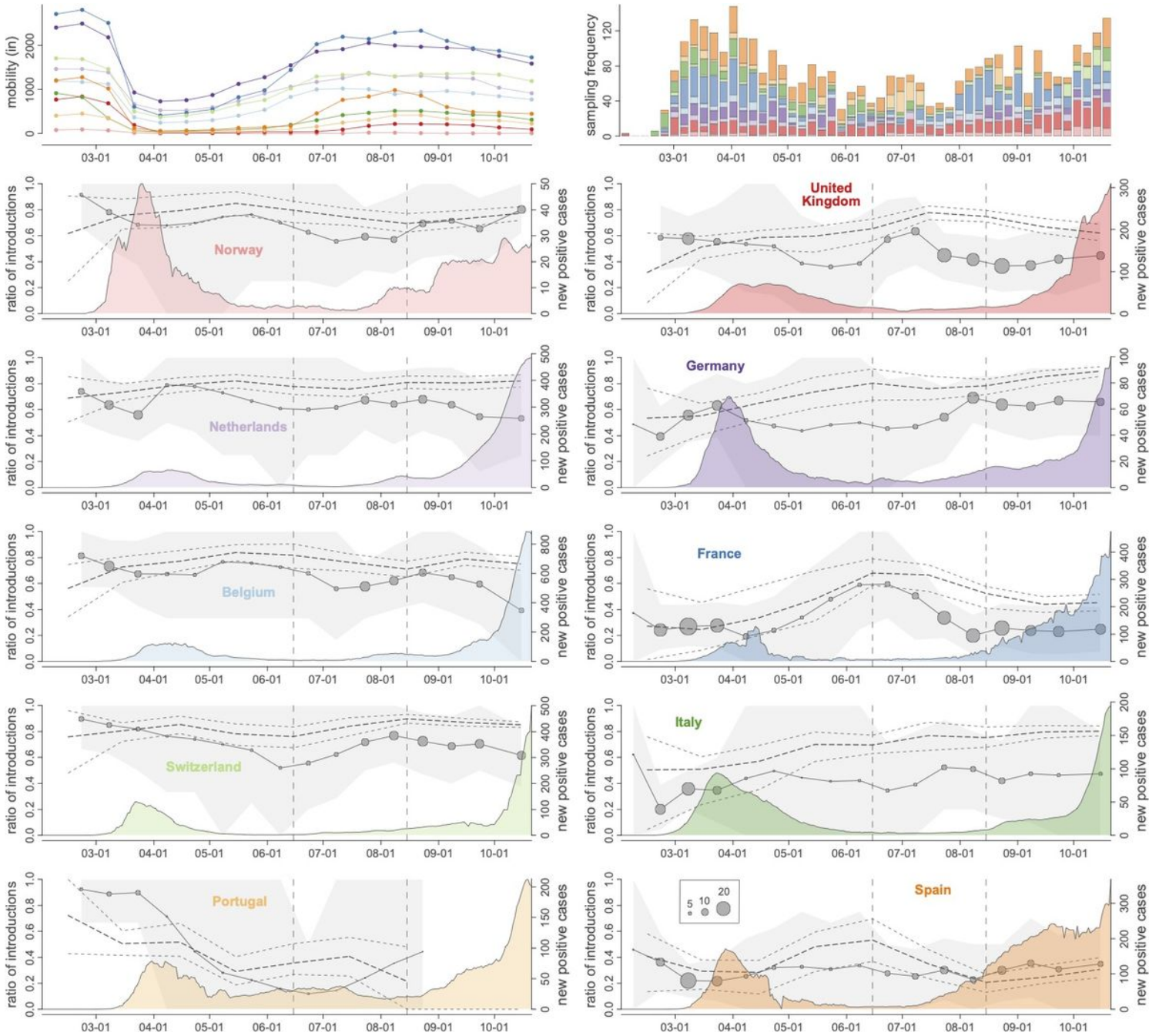


Figure 1

Mobility, genome sampling, case counts and phylogeographic summaries through time for 10 West European countries. (see full figure caption in Manuscript file)

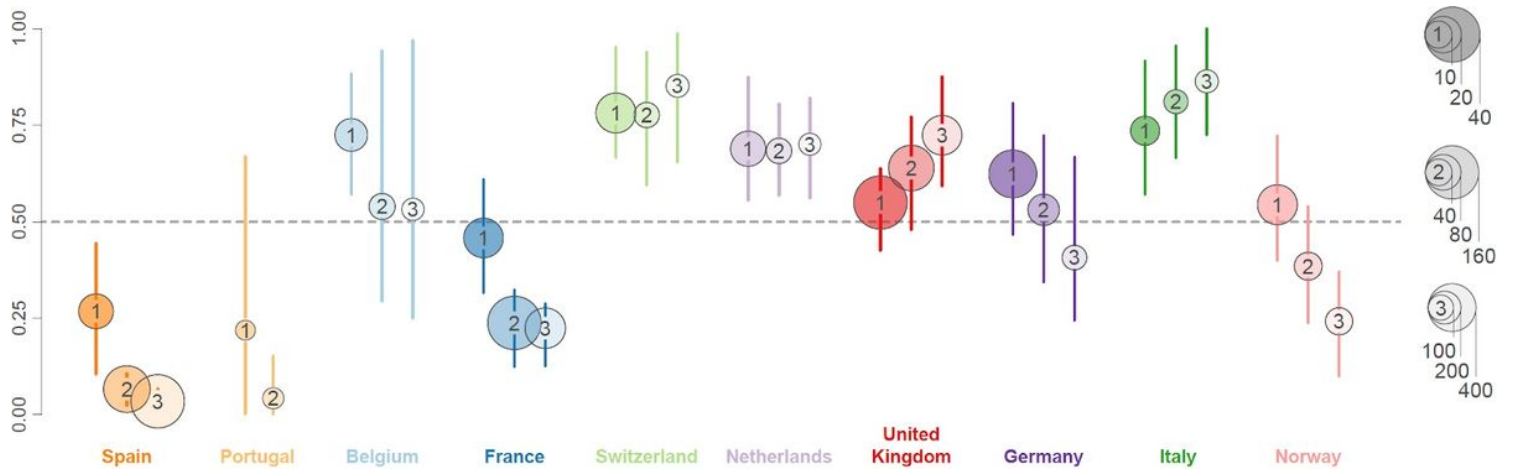


Figure 2

Posterior estimates for relative importance of lineage introduction events among West European countries. (see full figure caption in Manuscript file)

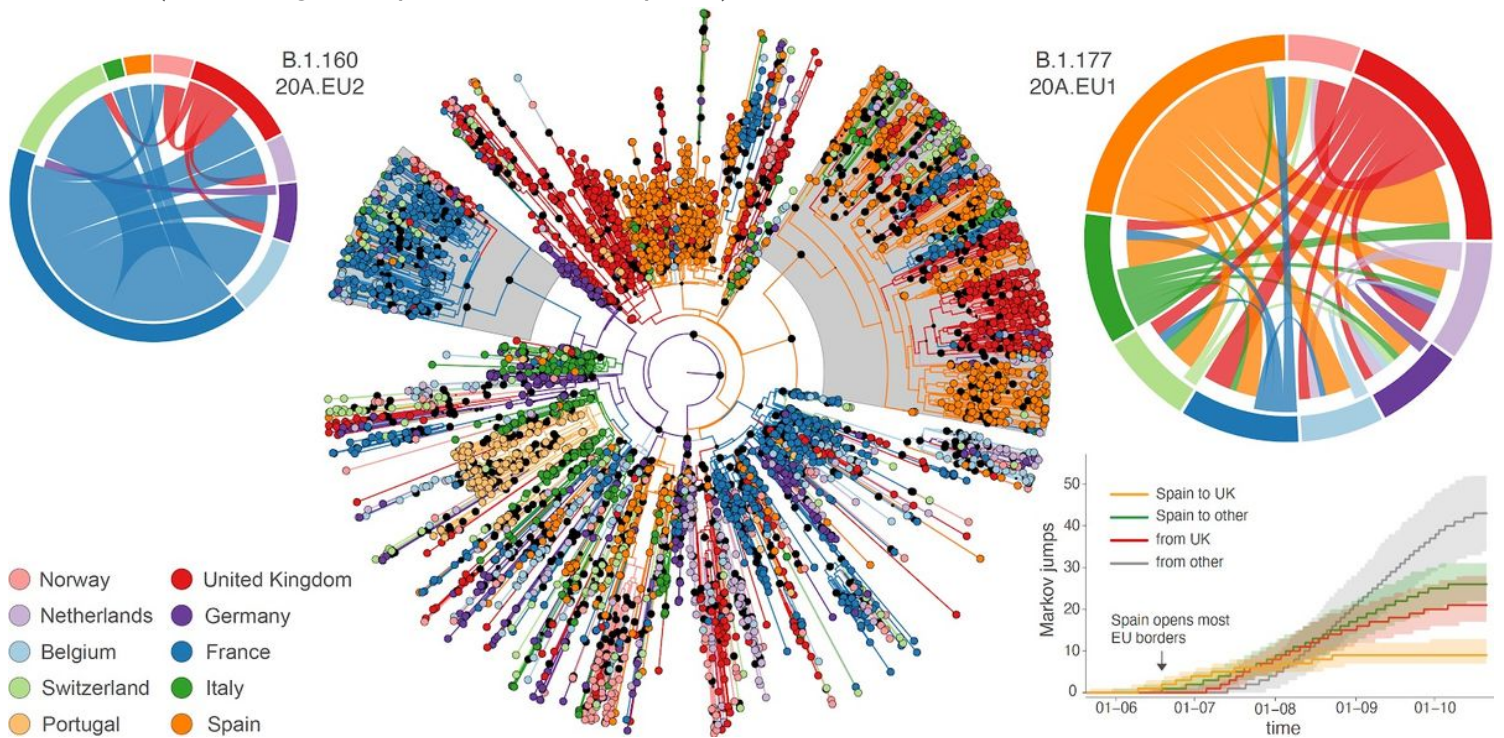


Figure 3

Phylogeographic estimates of SARS-CoV-2 spread in western Europe. (see full figure caption in Manuscript file)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [acknowledgementtable.pdf](#)