The phylodynamics of SARS-CoV-2 during 2020 in Finland — Disappearance and re-emergence of introduced strains.

Phuoc Truong Nguyen
University of Helsinki https://orcid.org/0000-0002-4830-4113
Ravi Kant
University of Helsinki https://orcid.org/0000-0003-3878-9775
Frederik Van den Broeck
KU Leuven https://orcid.org/0000-0003-2542-5585
Maija T. Suvanto
University of Helsinki https://orcid.org/0000-0003-4883-5755
Hussein Alburkat
University of Helsinki https://orcid.org/0000-0001-7847-3762
Jenni Virtanen
University of Helsinki https://orcid.org/0000-0001-6518-3088
Ella Ahvenainen
University of Helsinki
Robert Castren
University of Helsinki
Samuel L. Hong
KU Leuven https://orcid.org/0000-0001-6354-4943
Guy Baele
KU Leuven https://orcid.org/0000-0002-1915-7732
Maarit J. Ahava
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0001-6653-3238
Hanna Jarva
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0002-9154-354X
Suvi Tuulia Jokiranta
University of Helsinki https://orcid.org/0000-0003-4988-185X
Hannimari Kallio-Kokko
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0002-9773-9586
Eliisa Kekäläinen
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0001-6045-108X
Vesa Kirjavainen

University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0001-6664-8406
Elisa Kortela
University of Helsinki https://orcid.org/0000-0001-9843-8841
Satu Kurkela
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0003-2911-3212
Maija Lappalainen
University of Helsinki and Helsinki University Hospital https://orcid.org/0000-0001-5400-1250
Hanna Liimatainen
University of Helsinki and Helsinki University Hospital
Marc A. Suchard
UCLA https://orcid.org/0000-0001-9818-479X
Sari Hannula
Institute for Molecular Medicine Finland (FIMM)
Pekka Ellonen
Institute for Molecular Medicine Finland (FIMM) https://orcid.org/0000-0001-6072-0489
Tarja Sironen
University of Helsinki https://orcid.org/0000-0002-2344-2755
Philippe Lemey (philippe.lemey@kuleuven.be)
KU Leuven https://orcid.org/0000-0003-2826-5353
Olli Vapalahti (olli.vapalahti@helsinki.fi)
University of Helsinki https://orcid.org/0000-0003-2270-6824
Teemu Smura (teemu.smura@helsinki.fi)
University of Helsinki https://orcid.org/0000-0002-9187-3151

Research Article

Keywords: SARS-CoV-2, transmission, turnover, Variants of Concern, Finland, Nextstrain clade 20C

Posted Date: July 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-753457/v1

Abstract

Finland has had a low incidence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) infections as compared to most European countries. Here we report the origins and turnover of SARS-CoV-2 lineages circulating in Finland in 2020. SARS-CoV-2 introduced to Finland in January 2020 and spread rapidly across southern Finland during spring. We observed rapid turnover among Finnish lineages during this period. Clade 20C became the most prevalent among sequenced cases and was replaced by other strains in fall 2020. Bayesian phylogeographic reconstructions suggested 42 independent introductions into Finland during spring 2020, mainly from Italy, Austria, and Spain, which might have been the source for a third of cases. The investigations of the original introductions of SARS-CoV-2 to Finland during the early stages of the pandemic and of the subsequent lineage dynamics could be utilized to assess the role of transboundary movements and effects of early intervention and public health measures.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel betacoronavirus (genus *Betacoronavirus*) that is responsible for the current, socially, and economically devastating COVID-19 pandemic. The virus has infected more than 184 million people in 221 countries and has caused over 3.9 million deaths as of July 5, 2021 [1]. SARS-CoV-2 emerged in Wuhan, China, from a horseshoe bat reservoir in late 2019 possibly through a currently uncharacterized intermediate host animal reservoir [2], and causes COVID-19, a respiratory infection with severe cases leading to respiratory failure and multiorgan manifestations in humans. The genome of SARS-CoV-2 is approximately 30 kb in length and is composed of 13–15 open reading frames (ORFs). The ORFs contain 11 genes that encode for 12 proteins, including ORF1a, ORF1b, spike (S), ORF3a, envelope (E), membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N), and ORF10 [3]. The virus is similar to other betacoronaviruses in terms of infectivity and evolutionary rate (9.8×10^{-4} substitutions per site per year) [4]. This has led to the emergence of multiple viral lineages circulating the globe. Viral lineages may become more common in a given host population due to selective advantages or by chance (e.g., due to founder effect or genetic drift). Despite there being currently a plethora of viral lineages, only a small proportion of these are classified as variants of concern (VOCs), i.e. are considered to have enhanced transmissibility, pathogenicity, evasion of immune responses, or resistance to vaccines [5]. Currently, these include the lineage B.1.1.7 (Alpha), first detected in the United Kingdom (UK) [6], the lineage B.1.351 (Beta) first detected in South Africa [7], the lineage P.1 (Gamma) first detected in Brazil [8] and the lineage B.1.617.2 (Delta) first detected in India [9].

The first Finnish SARS-CoV-2 case was detected January 29, 2020, from a tourist from Wuhan, China [10]. No onward transmission was detected from this case. Two positive cases were later reported in week 9 (February 25 and 27, 2020) within the Hospital District of Helsinki and Uusimaa (HUS) area (Fig. 1). This area covers the capital region (includes Helsinki, Vantaa, Espoo and Kauniainen), which is the most

inhabited (ca. 1.2 million residents of the ca. 5.5 million total population) and the most densely populated area (ca. 1,680 residents per km^2 on average) in Finland (Fig. 2) [11]. The first epidemic wave began in week 9 (end of February), when the number of weekly diagnosed cases increased to six. The number of reported cases subsequently nearly tripled to 16 cases in week 10 (beginning of March) and the number of diagnosed cases reached 196 on March 13, 2020 (week 11). The government declared a national emergency and closed all public schools and government buildings starting March 16, 2020 (week 12) [12].

Despite these responses the number of cases continued to rise. Travel restrictions to and from the Uusimaa region were imposed by the Finnish parliament on March 27, 2020 (week 13), and subsequently lifted on April 15, 2020 (week 16). The number of weekly cases spiked during April (weeks 14–17) with over 600 cases. The number of tests performed increased rapidly from less than 100 per day to over 10,000 per day during this period (weeks 9–17, i.e. from end of February to mid-April). The test positivity rate dropped during the same time from ca. 21.4% (six detections from 28 tests) to 5.6% (611 positive cases from 10,853 tests). As the number of positive SARS-CoV-2 cases started to decrease, kindergartens and elementary schools were reopened on May 14 (week 20). In the beginning of June (week 23), restrictions on social gatherings were partially loosened to allow larger gatherings of 10–50 people in public spaces (e.g., restaurants and sport events). The rate of weekly detections steadily decreased to less than 100 during this time, and the number of tests per week also dropped to half during summer until week 26 (end of June) [14]. International travel restrictions were lifted between European countries with low infection rates (less than 25 per 100,000) [15] on July 8, 2020 (week 28). Weekly SARS-CoV-2 cases in the HUS area remained low with less than 50 cases per week for seven weeks until week 31 (end of July). The number of weekly tests performed steadily increased over the rest of summer and fall.

The second epidemic wave began in week 32 (beginning of August) with approximately 100 new cases each week until week 37 (beginning of September). After this point, the number of positive cases increased by on average 100 cases per week until peaking at approximately 1,600 cases per week. During this time, the rates of SARS-CoV-2 detections were well below the set national epidemic threshold until mid-September (week 38), when the number of cases started to rise.

The peak of the second SARS-CoV-2 wave lasted from week 47 to 50 (from mid-November to the beginning of December). Over 51,000 tests were performed during the peak in week 48 (end of November). Compared to the previous peak in spring, the winter peak had nearly 2.5 times more positive detections. However, the detection rate at the time was 4.0% (1,662 detections from 41,769 tests), lower than in spring peak. Nearing the end of the year, the number of weekly positive findings then sharply dropped to approximately 900, almost half compared to the peak of the second wave.

Here we report the origins and turnover of SARS-CoV-2 lineages circulating in Finland during the year 2020. We will describe the detected viral strains and variations in their compositions among cases across the year. The aim of this study is to provide information about the SARS-CoV-2 strains that were circulating during the pandemic in Finland in 2020. This includes strain composition at different time periods, identifying their countries of origin, and comparing spring and fall sequences to separate

endemic strains from newly imported ones. These results could be utilized to assess and determine the role of transboundary movements and early intervention and public health measures during ongoing pandemic.

Results

Lineage distribution

During the year 2020, there were 37,145 laboratory confirmed COVID-19 cases in Finland [13]. 21,730 (58.5%) of these cases were diagnosed in the HUS Diagnostic Center [13]. We sequenced a total of 1,597 SARS-CoV-2 genomes from the year 2020, which accounts for over 7.4% of all positive samples from the HUS area and represents 4.3% of positive samples from Finland in 2020.

By the week 11 (mid-March), all major clades of SARS-CoV-2 (i.e. GISAID clades G, GR, L and V, which correspond Nextstrain clades 19A, 20A, 20B and 20C, and Pango lineages B, B.1 and B.1.1 respectively [16]) had been introduced to Finland. However, by week 16 (mid of April), the lineages that contain the D614G substitution in the spike protein (20A, 20B and 20C) became dominant (Fig. 3). Out of these, Nextstrain clade 20C grew relatively fast into the dominant clade, starting from week 14 (beginning of April), when the number of 20C detections among the weekly cases nearly tripled from 17 (23%) to 59 (54%) from the previous week (Fig. 3). The second dominant clade, 20A, is notable due to its likely introduction from Spain and being highly prevalent among cases during the second wave in fall.

The majority of genomes from the dominant 20C clade contained the D936Y spike mutation. The prevalence of this mutation increased concurrently also in Sweden. In addition, the mutation has been detected mostly in Wales (as early as March 15, 2020) in association with the D614G mutation, as well as in England, and with low frequencies in Denmark, Poland and the United States [17]. The emergence and rapid spread of this mutation might be caused by periodic positive selection pressures [18] despite its destabilizing effect on post-fusion spike protein assembly due to a loss of a salt bridge between monomers [17]. While the number of cases (and sequences) were low during June and July, all of them represented this cluster. Further, two detections were recorded in August; one likely originating from a Finnish springtime strain and the other from October introduced from abroad (Supplementary tree file). The clade was subsequently replaced by lineages 20A and 20B later in autumn.

By August 2020, there was sufficient global genetic diversity in SARS-CoV-2 for more fine grained analysis using Pango lineage classification (Fig. 4). Notably, throughout fall two lineages, B.1.36.22 and B.1.463 that consist almost exclusively of viruses sequenced from Finland, formed the majority of detected lineages. Both of these lineages have been detected only sporadically in other countries (Norway, Denmark, Latvia and Canada for B.1.36.22 and Denmark for B.1.463) [20]. The third most common lineage B.1.160 is a large European lineage found in many countries [20]

During the initial stages of the second wave of SARS-CoV-2 in fall (weeks 32–38), several lineages were detected approximately in equal proportions, e.g., in week 34 with many lineages constituting 14% of cases, until the number of cases rapidly increased in week 39 (end of September). During this time, three lineages, Finland-specific B.1.36.22 and B.1.463 as well as pan European B.1.160 grew in prominence and became dominant in Finland. Notably, 20E(EU1) (lineage B.1.177), despite being a widespread clade in Europe during summer of 2020 [21,22], was not detected in high numbers in fall (10 cases in August), suggesting that it did not contribute significantly to the rise of cases during the second wave in Finland.

Phylogenetic analyses

We traced the geographic sources of viral introductions into Finland using travel-aware Bayesian phylogeographic reconstructions [23,24] of dispersal patterns between pairs of 17 European countries during the first epidemic wave (Fig. 5A). Our analysis included 1,643 genome sequences with their country and date of sampling, google mobility data and travel history information for 13% of the sampled patients in Finland (see Methods). We identified a total of 42 individual introductions (95% highest posterior density, HPD, interval = [36–47]) in the ancestry of a sample of 333 genomes from Finland. This estimate of the relative contribution of external introductions in establishing local transmission chains is similar to the one observed in New York State (116 introductions in 828 sampled genomes) [25] but lower compared to Belgium (331 introductions in 740 genomes) [26]. The majority of introductions is estimated to have occurred during the second week of March and originated from Italy (12 introductions, 95% HPD interval = [9–16]), Austria (8 introductions, 95% HPD interval = [6–10]) and Spain (8 introductions, 95% HPD interval = [7–10]) (Fig. 5B). Germany, Sweden, Switzerland, France, the United Kingdom and Denmark each accounted for 1–3 introductions, while we did not identify any viral diffusion into Finland from Turkey, the Netherlands, Latvia, Estonia, Poland, Norway and Hungary (Fig. 5B). The pattern of viral introductions mainly from Italy, Austria and Spain during the first wave largely matched our epidemiological records with travel history data, as 40 out of 44 imported cases returned from these countries (Fig. 5B). Notably, Austria shows at least twice the number of travel history entries (n = 20) compared to viral introductions (n = 8) as estimated from our Bayesian phylogeographic analyses (Fig. 5B). Close inspection of the phylogenetic tree as obtained from our Bayesian reconstructions of revealed that 12 out of 20 cases returning from Austria clustered tightly within two subclades belonging to a larger cluster of predominantly Austrian genomes (Supplementary Figure S1), suggesting that these Finnish patients may have picked up the infections from the same source of in popular skiing resorts.

As part of the 43 independent introduction events, our analysis identified 35 introductions (81%) (95% HPD interval = [30–39]) resulting in relatively few sampled Finnish descendants (≤ 10), including 16 singleton introductions (95% HPD interval = [12–20]). Hence, the large majority of introductions account for a relatively small number of the lineages we sample, a pattern typically observed for all European countries (Fig. 6A). This highlights extensive heterogeneity in SARS-CoV-2 transmission dynamics underlying the establishment of local transmission chains. While the largest number of independent introduction events originated from Italy (Fig. 6B), we identified one introduction from Spain that gave rise to 119 (95% HPD interval = [100–134]) genomes sampled in Finland (Fig. 6C), indicating that one

third of our first wave sample traces back to a viral lineage originating from Spain. Descendant taxa from this single Spanish introduction belong to Nextstrain clade 20A, which is the second most dominant clade during the first wave epidemic in Finland (Fig. 3). The most dominant clade during the first epidemic wave in Finland, clade 20C with D936Y spike mutation, clustered together with Swedish sequences. However, the posterior probability for this clade was low and, therefore, the origin of this clade remained unresolved.

Our phylogeographic reconstructions based on 1,643 sampled genomes may potentially suffer from the impact of sampling bias. For instance, while one of the Finnish COVID-19 cases returned from neighbouring country Estonia, we inferred no viral movements to Finland directly from Estonia (Fig. 5B), a country that is severely underrepresented by viral genomes (n = 4) compared to most other countries. To explore the sensitivity of our phylogeographic reconstructions to sampling bias, we incorporated unsampled taxa for 6 locations (Estonia, Latvia, Norway, Hungary, Poland and Turkey) that were represented by less than 60 sequences (other European countries were represented by at least 100 sequences) (Supplementary Table S1), resulting in a dataset of 2,019 taxa. Tip ages ("sampling times") were specified as randomly sampled dates from the case count distributions per undersampled country. In addition, as Finland is severely oversampled according to case counts in spring (13.87% versus 0.13–5.79% for the 16 selected European countries) (Supplementary Table S1), we also performed an analysis including the unsampled taxa and in which Finnish genomes were downsampled from 333 to 100 sampled genomes, for a total of 1,786 genomes.

Results obtained from reconstructions without unsampled taxa (Fig. 6), with unsampled taxa (Supplementary Figure S2A) and with unsampled taxa and downsampled Finnish taxa (Supplementary Figure S2B) were largely similar in terms of the total number of introductions and the dominant contribution of viral introductions from Italy, Austria, and Spain. However, in contrast to the reconstructions with the full set of Finnish taxa (Fig. 1B and Supplementary Figure S2A), the reconstruction with downsampled Finnish taxa (Supp. Figure S2B) offered support for additional viral introductions from both Estonia (2 introductions, 95% HPD interval = [0–3]) and Latvia (2 introductions, 95% HPD interval = [0–3]). The largely similar results from reconstructions without and with unsampled tax, suggests limited impact of sampling bias on our Bayesian phylogeographic reconstructions.

Phylogenetic trees constructed using maximum-likelihood inference including Finnish SARS-CoV-2 sequences from Fall of 2020 (Fig. 7A) show that many of these sequences form three major clusters representing lineages B.1.36.22, B.1.463 and B.1.160. Both B.1.36.22 and B.1.463 from Finland form monophyletic clusters suggesting a single ancestor for these major lineages circulating in Finland during the autumn 2020. These clusters may have originated from a few earlier strains which were either already circulating (undetected) locally or were introduced from other countries during summer (Fig. 7B). In order to assess this, we included closest BLAST matches to these clades to the phylogenetic tree. However, the origin of these clades remained unresolved.

Discussion

Finland has had a low incidence of SARS-CoV-2 cases compared to most European countries, including neighboring countries Sweden and Russia. Intriguingly, during both spring and fall epidemic waves, the majority of infections were caused by a limited number of viral lineages. These included a subcluster of clade 20C with spike protein D936Y substitution during spring, and the predominance of lineages B.1.36.22 and B.1.463 during fall. This suggests that despite multiple introductions of the virus to an immunologically naive population, only few of these resulted in long transmission chains. This is consistent with the well-known super-spreading events that dominate the epidemiology of SARS-CoV-2, yet the major unanswered question is whether natural selection played any role in the lineage distribution or if this was due to epidemiological factors such as fluctuation in lineage frequencies due to the random transmission events, i.e., founder effects [27].

The Bayesian phylogeographic analyses (Fig. 5B) showed that there were at least twice the number of returning travelers from Austria compared to our estimated viral introductions. One possible explanation for this observation is that travelers from Austria may have picked up the same source of infections in popular skiing resorts, resulting in their viral genome clustering (Supplementary Figure S1).

Several genomic and environmental factors might explain the lineage turnover in Finland during 2020. The mutation rate of RNA viruses is considerably high, resulting in highly polymorphic virus populations. While the majority of mutations in the viral genome are either neutral or lead to viral attenuation, occasionally, mutations can result in higher fitness, such as more efficient transmission [28–30]. However, it is uncertain whether lineage turnover in Finland during 2020 was due to the higher fitness of introduced lineages. A more likely hypothesis would be that viral strains introduced to a new region with an immunologically naive population and relatively low incidence of infections become dominant due to the epidemiological factors. This is exemplified by the high heterogeneity in the frequency of onward transmission of imported viruses with limited genetic diversity, as well as rapid turnover of circulating viruses during August/September 2020. Regarding the latter, while there is some evidence that spike mutation D936Y may be positively selected, the dominant subcluster during late spring and summer containing this mutation was completely replaced by other lineages in August/September 2020. This is likely due to the low incidence of infections during late spring and summer. In such circumstances any lineage (B.1.36.22 and B.1.463 in this case) may become dominant due to the super-spreading events or other epidemiological factors. However, the potential biological factors affecting lineage turnover require further empirical investigation.

To conclude, several genomic and epidemiological factors might have contributed to the rapid turnover of prevalent lineages among Finnish SARS-CoV-2 cases during the first wave in spring and the second one in fall of 2020. Our data suggest that the observed heterogeneity of detected virus cases is likely due to independent introductions from several neighboring and distant European countries before imposing travel restrictions. In addition, we observed that the majority of circulating virus lineages were country-specific, mostly likely due to the high heterogeneity in the frequency of onward transmission of imported viruses.

Methods

Sequence and Google mobility data

Research data for this report consists of SARS-CoV-2 genomes (n = 1,597) that were sequenced from SARS-CoV-2 PCR positive patient samples with Illumina NovaSeq and MiSeq sequencing platforms in-house within the Department of Virology in the University of Helsinki and submitted to the GISAID database. Due to HUSLAB initially being the only clinical laboratory sequencing patient samples, some of the virus sequences originate from outside the HUS area. The collection period was from spring to fall 2020. In addition to the local sequence data, global SARS-CoV-2 genomes (n = 20,720) were acquired from the GISAID database (Supplementary Table S1).

In order to infer the geographic source(s) of SARS-CoV-2 lineages contributing to the first wave in Finland, we extended our dataset of Finnish genomes with genomes available for other European countries. A recent phylogeographic analysis demonstrated that SARS-CoV-2 spread in Europe was strongly predicted by Google mobility flows [31]. To inform our sampling, we therefore turned to the Google COVID-19 Aggregated Mobility Research Dataset containing anonymized mobility flows aggregated over users who have turned on the Location History setting (on a range of platforms). Aggregated mobility flows between Finland and all other European countries were summarized between January and April 2020, and we selected the following 16 countries that were responsible for 95% of international travels from and to Finland: Estonia, Latvia, Norway, Hungary, Poland, Turkey, Sweden, Netherlands, Austria, Denmark, Italy, Germany, Switzerland, Spain, France and the United Kingdom. For these countries, we downloaded the available SARS-CoV-2 genomes from GISAID on April 17, 2020. For six countries (Estonia, Latvia, Norway, Hungary, Poland and Turkey) represented by a relatively small number of genomes, we decided to augment our dataset with genomes from GISAID with a sampling date up to April 31, 2020.

We selected only sequences from the B.1 lineage with the D614G mutation for the analyses. We removed duplicate genomes for each country using SeqKit v0.11 [32]. For Finland, we retained duplicate genomes when these were sampled from cases with different travel histories. All genomes were aligned using MAFFT [33] and trimmed at the 5' and 3' ends. We then subsampled each country proportionally to the cumulative number of cases on April 17, 2020 by setting an arbitrary threshold of 7.5 genomes per 10,000 cases, with a minimum number of 100 sequences per country. For the 6 countries where the number of unique genomes was below 100, all genomes were included in the analysis. To maximize the spatial and temporal coverage of the subsampling, we partitioned each country's genome pool by week and sampled as evenly as possible, selecting sequences from a different region within the country when available. We checked the resulting dataset for potential outliers with a root-to-tip regression using TempEst v1.5.3 [34] on a maximum likelihood inferred using IQ-TREE v2.0.3 [35], and removed 9 genomes. The final dataset consisted of 1,643 genomes out of an initial 8,513 genomes. Total, unique and downsampled number of genomes by country are given in Supplementary Table S1. All genomes were associated with exact sampling dates, except for the four genomes from Estonia that were sampled in March 2020.

Bioinformatic analyses

Consensus sequence data for Finnish SARS-CoV-2 were computed and classified with HAVoC [36] and a modified pipeline consisting of Jovian [37] and pangolin [38]. Clade and lineage assignment for GISAID sequences was done Nextclade [19] and pangolin.

Maximum-likelihood phylogenetic trees

Viral sequences were aligned with MAFFT [33] and the trees were computed with a SARS-CoV-2 version IQ-TREE [35]. The phylogenetic trees were visualized in R with ggtree [39].

Bayesian time-measured phylogeographic analyses

We performed Bayesian evolutionary reconstruction of timed phylogeographic history using BEAST 1.10 [40] incorporating genome sequences, their country and date of sampling, Google mobility data and individual travel history [23,24]. We modelled sequence evolution using a strict molecular clock model and an HKY nucleotide substitution model [41] with gamma-distributed rate variation among sites [42]. We assumed an exponential growth coalescent model as the tree-generative process prior. Uncertainty in the sampling time for the four Estonian genomes was accommodated by sampling uniformly across the reported collection month in the Markov chain Monte Carlo (MCMC) analysis. Our phylogeographic model incorporated the country of sampling as discrete traits associated with the sampled genomes, and following a recent European SARS-CoV-2 phylogeographic analysis [31], we adopted a generalized linear model (GLM) specification to parametrize each rate of among-location movement as a log linear function of the total Google mobility flows for the period January-April 2020. Total mobility flows were log-transformed and standardized after adding a pseudocount to each entry in the matrix. The main goal of our GLM extension was to obtain well-informed phylodynamic estimates. As the ancestral reconstruction of locations depends on the availability of samples, over- or undersampling of sequences from a given location can greatly impact the estimated ancestral locations [23]. To mitigate sampling bias and improve the location-transition history reconstructions, we augmented our elementary phylogeographic model by incorporating travel history information obtained from 44 cases that returned to Finland from Austria (n = 20), Italy (n = 13), Spain (n = 7), Estonia (n = 1), Germany (n = 1), Switzerland (n = 1) and United Kingdom (n = 1).

We also investigated how unsampled diversity for six European countries or oversampling of Finnish SARS-CoV2 diversity may impact our phylogeographic reconstructions. Building on our extended phylogeographic model including sampling locations and individual travel histories, we incorporated unsampled taxa for the under-sampled countries Estonia (n = 96 taxa added), Latvia (n = 83), Norway (n = 56), Hungary (n = 54), Poland (n = 46) and Turkey (n = 41) to arrive at a minimum of 100 genomes for all countries. Unsampled taxa without observed sequence data were added with associated location and sampling times, for which we randomly sampled dates from case count distributions per country. For this

analysis, we also downsampled the Finnish genome dataset to 100 taxa, while ensuring we incorporated the 44 samples with known travel histories.

We performed inference under the full model specification using MCMC sampling while employing the BEAGLE library v3 [43] to increase computational performance. Because MCMC burn-in takes considerable computational time due to the size of our dataset, with the tree topology representing the most challenging parameter for convergence, we initially only considered sequence evolution to arrive at a tree distribution from which trees were taken as starting trees in our phylogeographic analyses. Multiple independent MCMC runs were run to ensure that their combined posterior samples achieved effective sample sizes (ESSs) larger than 100 for all continuous parameters. Transition histories were summarized using the tree sample tool, TreeMarkovJumpHistoryAnalyzer, implemented in BEAST to collect Markov jumps [44] and their timings from a posterior tree distribution annotated with Markov jumps histories [31].

References

1. Worldometer. COVID-19 Virus Pandemic. *Worldometer* at < https://www.worldometers.info/coronavirus/>
2. Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. & Hsueh, P.-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int. J. Antimicrob. Agents* **55**, 105924 (2020).
3. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
4. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
5. Mahase, E. Covid-19: Novavax vaccine efficacy is 86% against UK variant and 60% against South African variant. *BMJ* **372**, n296 (2021).
6. GOV.UK. Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01. (2020). at < https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>
7. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* (2020). doi:10.1101/2020.12.21.20248640
8. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological* (2021). at < https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586>
9. Lineage B.1.617.2. *PANGO lineages* at < https://cov-lineages.org/global_report_B.1.617.2.html>
10. Haveri, A. *et al.* Serological and molecular findings during SARS-CoV-2 infection: the first case study in Finland, January to February 2020. *Euro Surveill.* **25**, (2020).
11. Jarva, H. *et al.* Laboratory-based surveillance of COVID-19 in the Greater Helsinki area, Finland, February-June 2020. *Int. J. Infect. Dis.* **104**, 111–116 (2021).

12. Government, in cooperation with the President of the Republic, declares a state of emergency in Finland over coronavirus outbreak. *Finnish Government* at < https://valtioneuvosto.fi/-/10616/hallitus-totesi-suomen-olevan-poikkeusoloissa-koronavirustilanteen-vuoksi?languageId=en_US>
13. Finnish Institute for Health and Welfare (THL). COVID-19 cases in the infectious diseases registry. at < https://sampo.thl.fi/pivot/prod/en/epirapo/covid19case/fact_epirapo_covid19case?&row=hcdmunicipality2020-445193&column=dateweek20200101-509030>
14. Willberg, E., Järv, O., Väisänen, T. & Toivonen, T. Escaping from Cities during the COVID-19 Crisis: Using Mobile Phone Data to Trace Mobility in Finland. *ISPRS Int J Geoinf* **10**, 103 (2021).
15. Finnish Institute for Health and Welfare (THL). Traffic light model to help in the assessment of risks associated with foreign travel. at < https://thl.fi/en/web/infectious-diseases-and-vaccinations/what-s-new/coronavirus-covid-19-latest-updates/travel-and-the-coronavirus-pandemic/traffic-light-model-to-help-in-the-assessment-of-risks-associated-with-foreign-travel>
16. Alm, E. *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill.* **25**, (2020).
17. Cavallo, L. & Oliva, R. D936Y and Other Mutations in the Fusion Core of the SARS-Cov-2 Spike Protein Heptad Repeat 1 Undermine the Post-Fusion Assembly. *BioRxiv* (2020). doi:10.1101/2020.06.08.140152
18. Ling, J. *et al.* Spatio-Temporal Mutational Profile Appearances of Swedish SARS-CoV-2 during the Early Pandemic. *Viruses* **12**, (2020).
19. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
20. pangolin. PANGO lineages. at < https://cov-lineages.org/lineages.html>
21. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* (2021). doi:10.1038/s41586-021-03677-y
22. Lemey, P. *et al.* Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* (2021). doi:10.1038/s41586-021-03754-2
23. Lemey, P. *et al.* Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
24. Hong, S. L., Lemey, P., Suchard, M. A. & Baele, G. Bayesian phylogeographic analysis incorporating predictors and individual travel histories in BEAST. *Curr. Protoc.* **1**, e98 (2021).
25. Dellicour, S. *et al.* Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLoS Pathog.* **17**, e1009571 (2021).
26. 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).
27. Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**, 52–61 (2004).

28. McCarthy, K. R. *et al.* Natural deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *BioRxiv* (2020). doi:10.1101/2020.11.19.389916
29. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).
30. Santos, J. C. & Passos, G. A. The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. *BioRxiv* (2021). doi:10.1101/2020.12.29.424708
31. Lemey, P. *et al.* SARS-CoV-2 European resurgence foretold: interplay of introductions and persistence by leveraging genomic and mobility data. *Res. Sq.* (2021). doi:10.21203/rs.3.rs-208849/v1
32. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
33. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
34. 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).
35. 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).
36. Truong Nguyen, P. *et al.* HaVoC, a bioinformatic pipeline for reference-based consensus assembly and lineage assignment for SARS-CoV-2 sequences. *BioRxiv* (2021). doi:10.1101/2021.02.12.431018
37. Zwagemaker, F. *et al.* DennisSchmitz/Jovian: Release 1.2.01. *Zenodo* (2021). doi:10.5281/zenodo.4431861
38. O'Toole, Á. *et al.* pangolin: lineage assignment in an emerging pandemic as an epidemiological tool. at < https://github.com/cov-lineages/pangolin>
39. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. GGTREE: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* (2016). doi:10.1111/2041-210X.12628
40. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
41. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
42. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
43. Ayres, D. L. *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).
44. Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol* **56**, 391–412 (2008).

Declarations

ACKNOWLEDGEMENTS

This study was supported by the Academy of Finland (grant number 336490), VEO - European Union's Horizon 2020 (grant number 874735), and the Jane and Aatos Erkko Foundation, as well as Helsinki University Hospital Funds (TYH2018322 and TYH2021343). We thank Kerstin Ahlskog for her technical assistance.

AUTHOR CONTRIBUTIONS

Conceptualization: PTN RK TS TSi OV. Formal Analysis: PTN RK TS FVDB PL SH. Funding acquisition: TSi OV. Investigation: PTN RK HL TS. Methodology: PTN RK SH PE TS FVDB PL GB MAS. Project administration: RK TS OV. Resources: PTN RK HKK MTS HA JV EA RC SH HL SK HJ ML PE TS TSi OV. Validation: PTN RK HL SK ML PE TS TSi OV. Writing – original draft: PTN RK TS FVDB. Writing – review & editing: PTN RK FVDB HL HKK MTS HA JV EA RC SH GB MAS SH HL SK HJ ML PE TS PL TSi OV.

COMPETING INTERESTS

The authors declare that they have no competing interests.

DATA AVAILABILITY

All data in the main text or supplementary material will be publicly available.

Figures

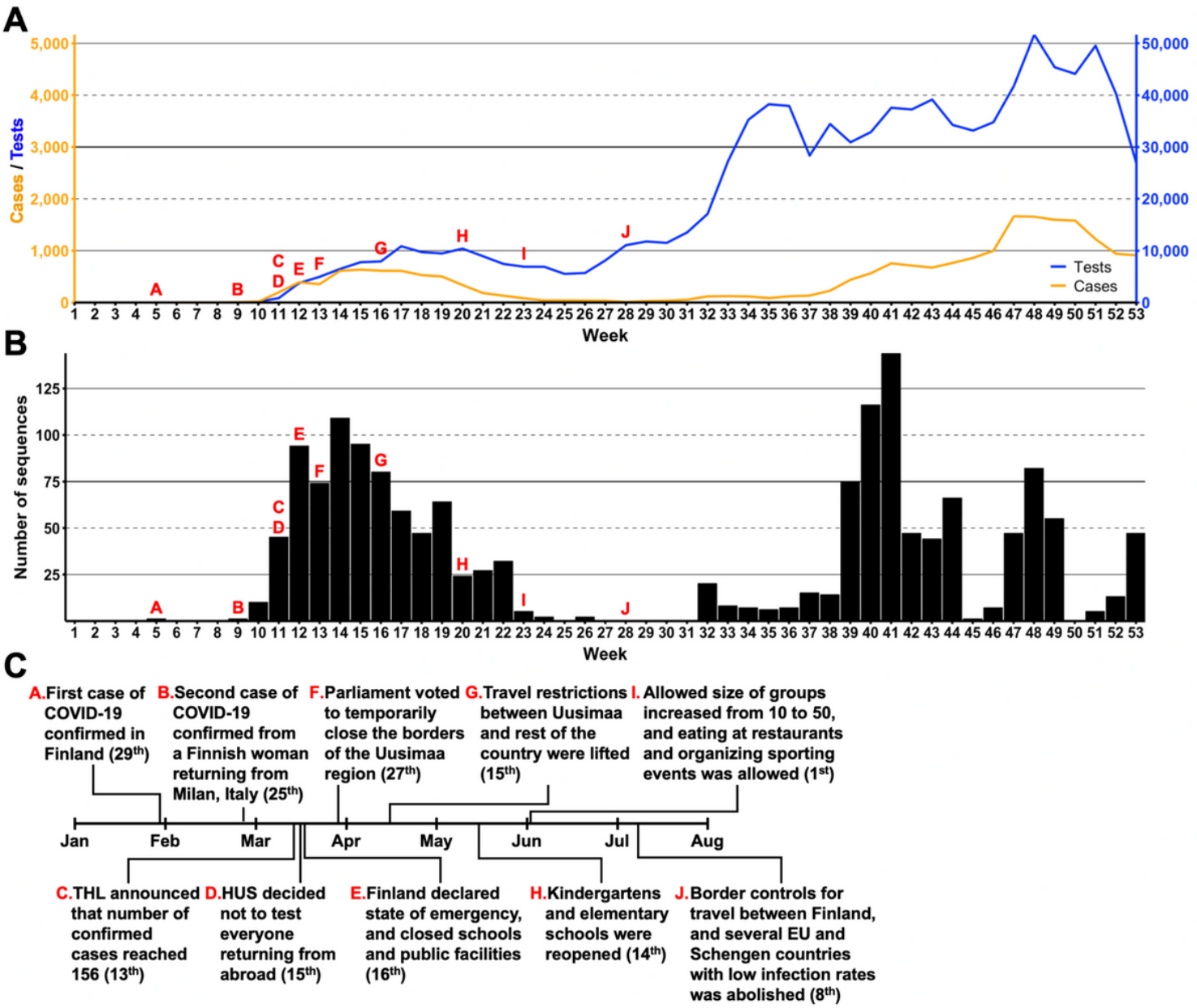

Figure 1

Weekly SARS-CoV-2 statistics and general timeline of Finland in 2020. The number of PCR tests (total n = 967,885) and positive findings (total n = 21,731) based on the COVID-19 infectious diseases registry of the 13 are shown in panel (A). The color of lines matches their respective axes, i.e. the axis indicating number of tests is on the right and number of positive cases on the left. The number of SARS-CoV-2 sequences submitted to GISAID (total n = 1,597) are displayed in panel (B). Panel (C) depicts the general timeline of the arrival of SARS-CoV-2 in Finland and the subsequent responses by the Finnish government and health authorities, which are indicated by letters A-J in panels (A) and (B). Exact dates for each response are mentioned within brackets. This information is based on public records by THL. HUS = Hospital District of Helsinki and Uusimaa.

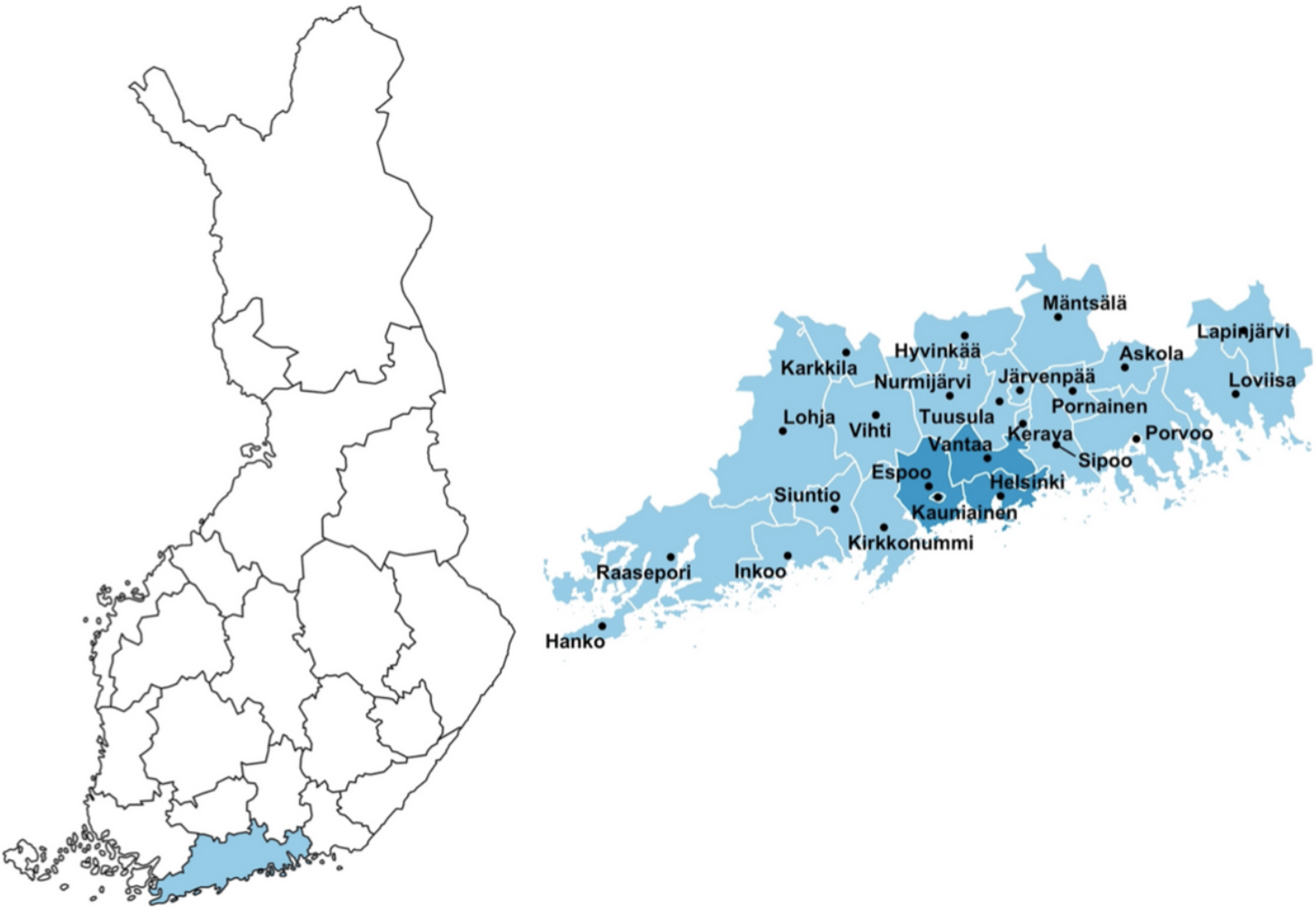

Figure 2

Map of the Hospital District of Helsinki and Uusimaa (HUS) and municipalities belonging to it in Finland. The capital region, which includes the municipalities of Helsinki, Espoo, Vantaa and Kauniainen, is highlighted in darker blue.

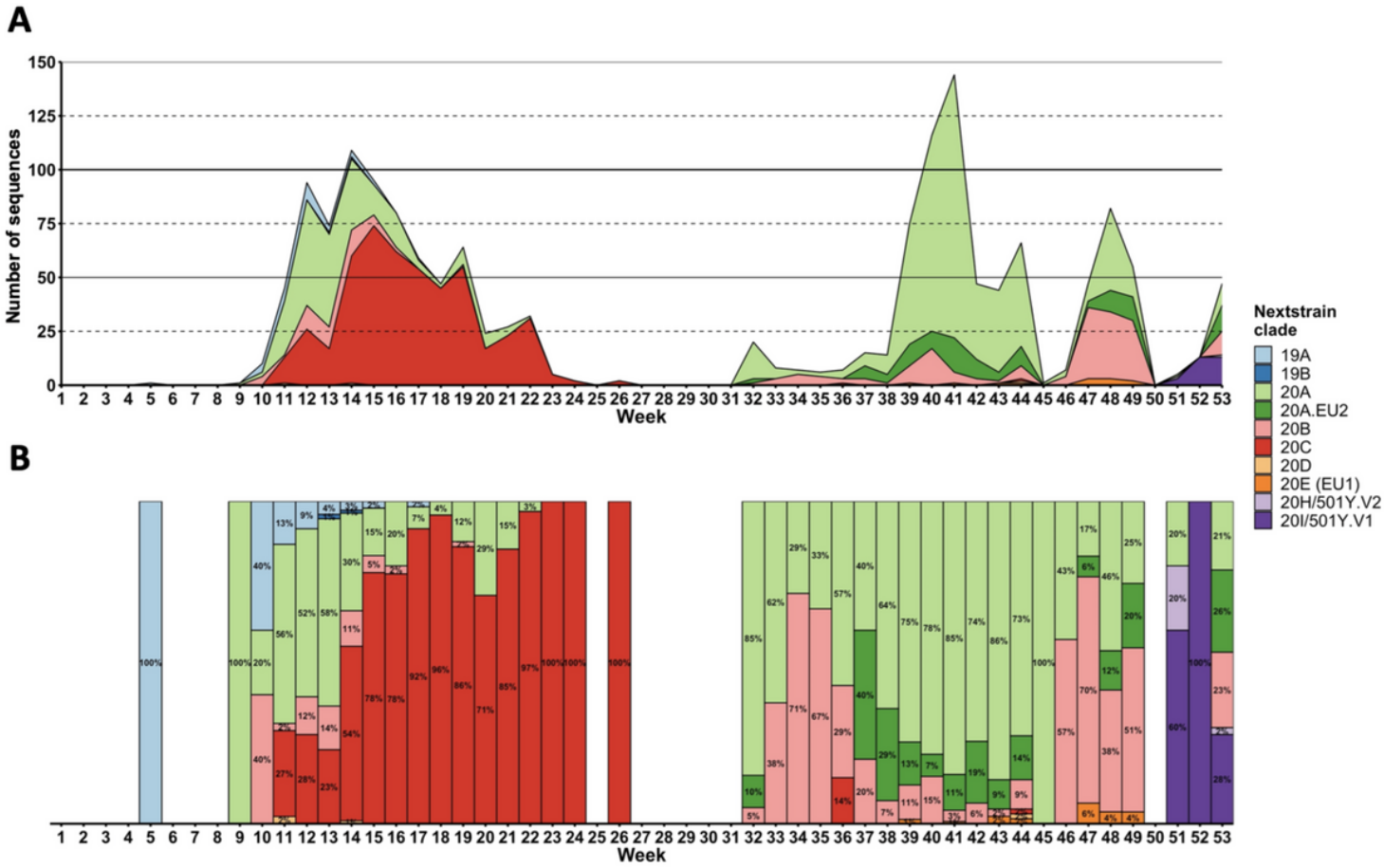

Figure 3

The number of sequences per week are displayed in panel (A) and clade proportions are shown in panel (B). The Nextstrain clade 20C stands out as it was the most prominent strain and was only detected during the first pandemic wave (weeks 9–24) in spring in 2020. Clade assignment was done with the Nextclade tool.

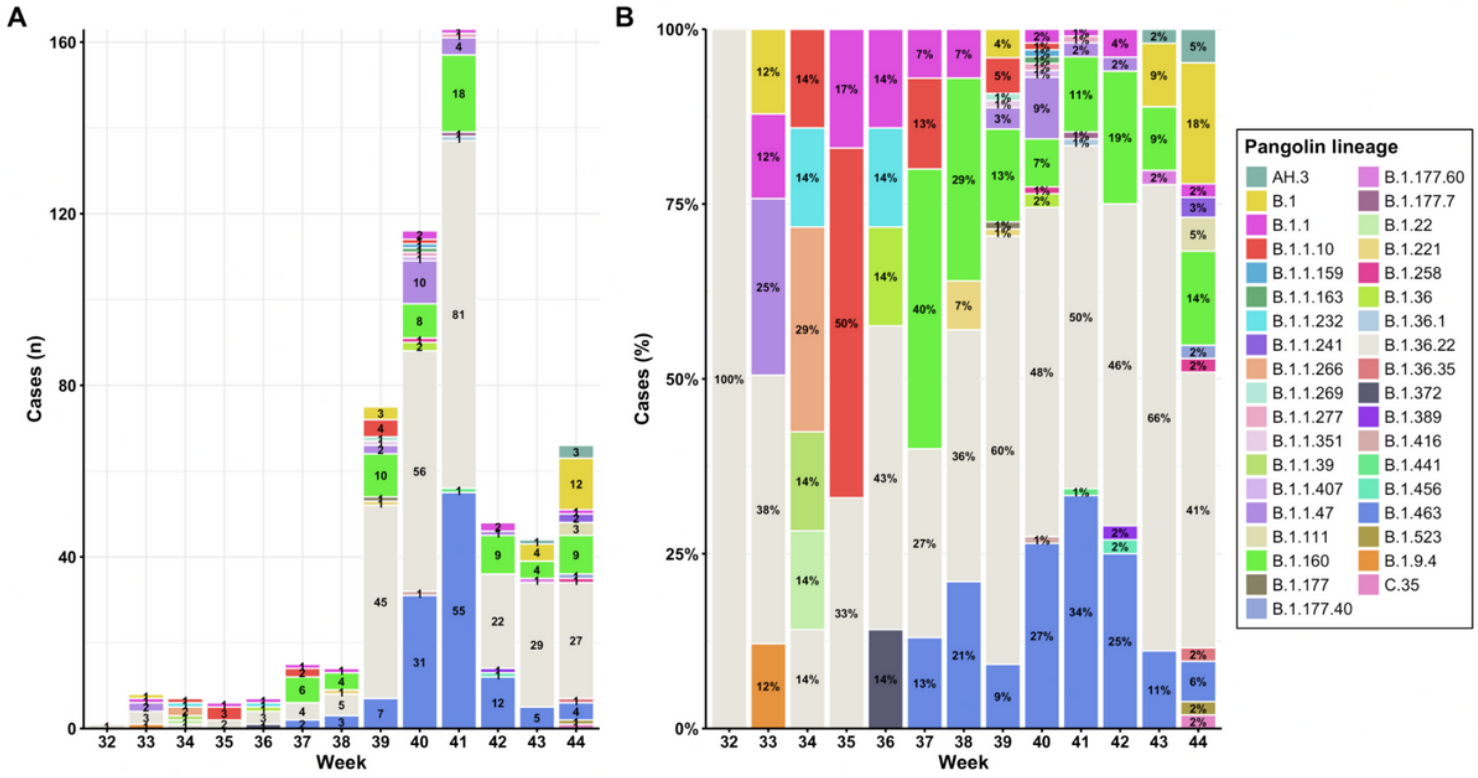

Figure 4

Turnover of Finnish SARS-CoV-2 lineages in fall of 2020 (weeks 32–44). During weeks 32–38 many lineages were detected among Finnish cases with relatively similar frequencies until B.1.36.22 and B.1.463 (both Finland-specific lineages) became the most prevalent lineages beginning in week 39. Panel (A) depicts the number of SARS-Cov-2 cases and lineages detected in the HUS area per week and panel (B) shows the proportions (%) of lineages per week.

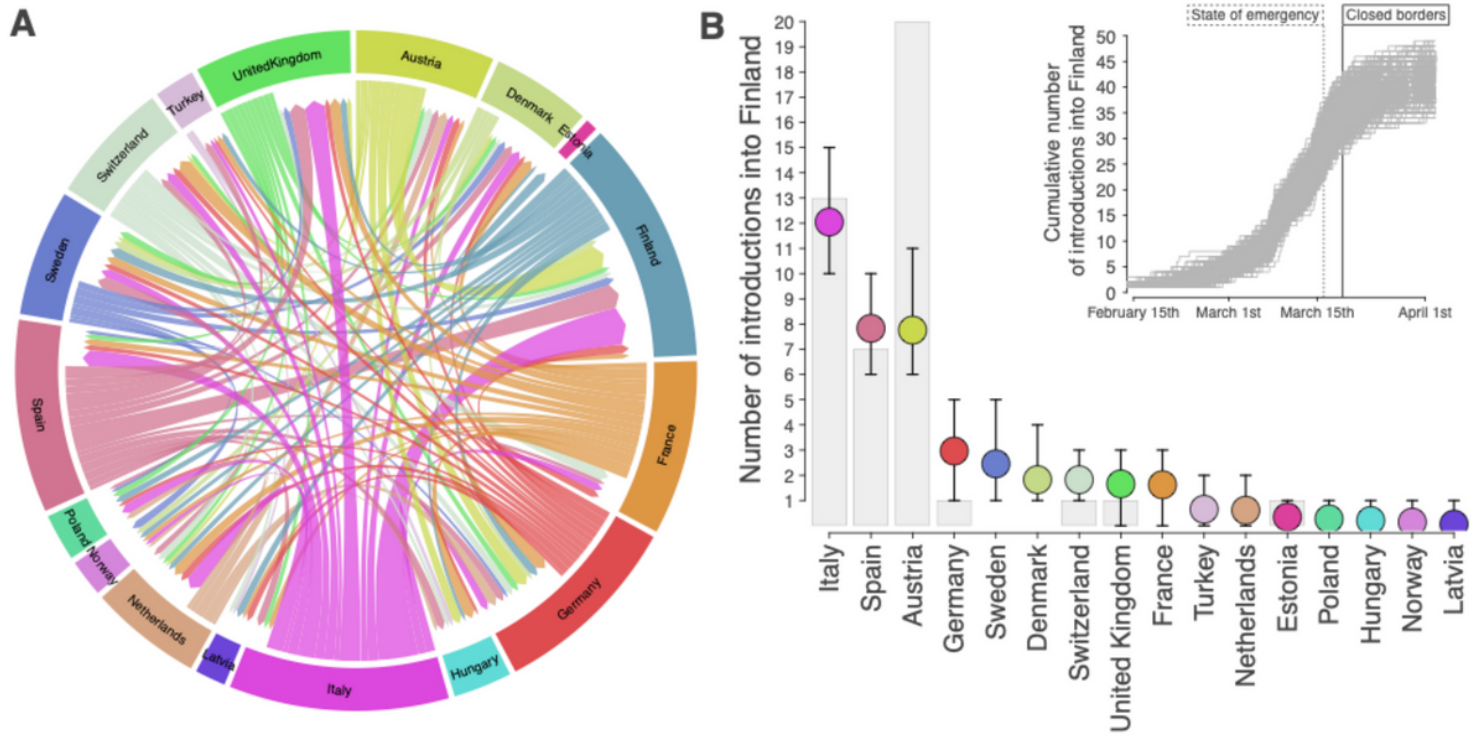

Figure 5

Discrete phylogeographic reconstruction of SARS-CoV-2 introductions into Finland during the first wave epidemic. (A) Circular migration flow plot based on the posterior expectations of the Markov jumps between 17 country-level locations, including Finland and 16 selected European countries. Migration flow out of a particular location starts close to the outer ring and ends with an arrowhead more distant from the destination location. (B) Mean and 95% highest posterior density (HPD) number of transitions to Finland from each of the 16 selected European countries, as estimated from 1,000 trees subsampled from the posterior distribution. Gray bars indicate the number of cases with travel history data returning from each country. Inset shows the cumulative number of introductions into Finland summarized from a posterior sample of phylogeographic trees. Dashed line indicates the day (16th of March) when the Finish authorities declared a state of emergency due to COVID-19. Full line indicates the day (19th of March) when the Finish authorities announced a restriction of passenger traffic at Finland's borders.

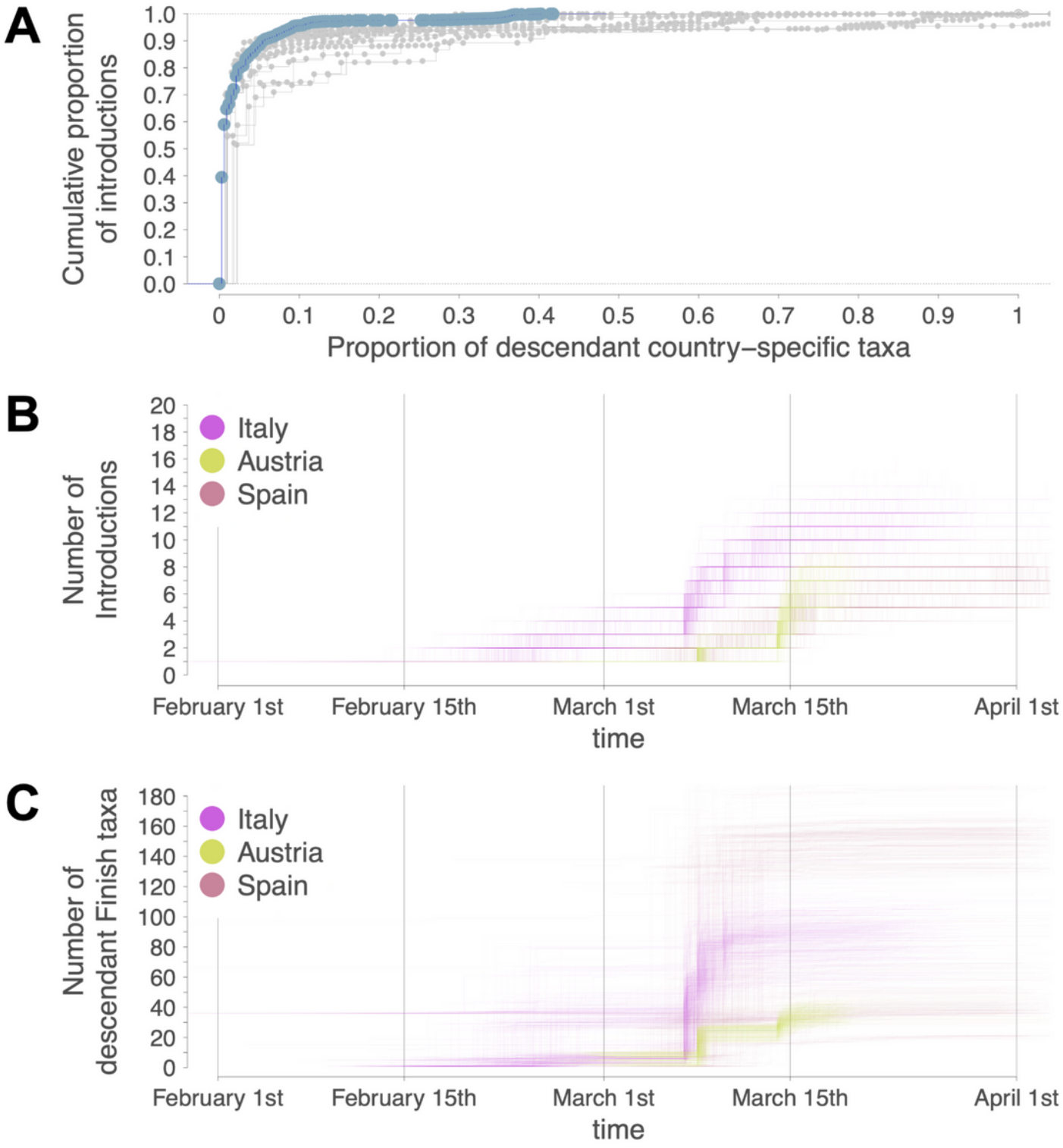

Figure 6

(A) Empirical cumulative distribution function plot for the proportion of descendant country-specific taxa, in blue for Finland and in gray for each of the 16 selected European countries excluding Latvia and Estonia for which few genomes were available. Cumulative number of phylogeographic transitions (B) and cumulative number of Finnish descendant state taxa (C) over time from Italy, Austria and Spain to Finland.

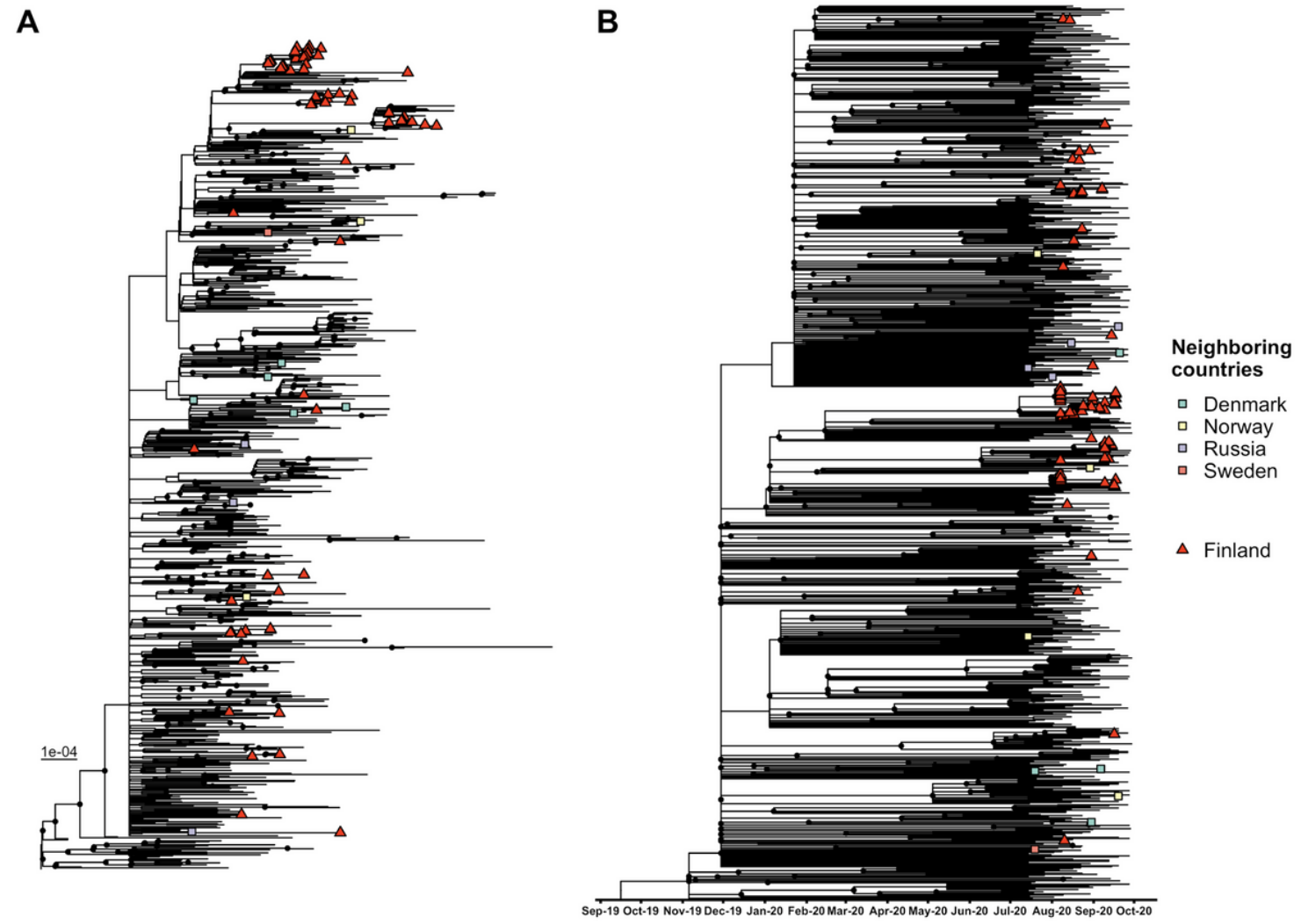

Figure 7

Phylogenetic trees of Finnish SARS-CoV-2 sequences from fall of 2020. The trees were constructed from Finnish sequences (n = 77) and a reference set of SARS-CoV-2 sequences from different countries around the globe (n = 746). Sequences from the neighboring countries of Finland were highlighted in the trees. Panel (A) shows a maximum-likelihood tree and panel (B) a time-scaled maximum likelihood tree. All sequences were obtained from the GISAID database and include complete genomes with full collection dates from July 15 to September 30, 2020. Finnish sequences were collected between August 8 and September 18, 2020 (weeks 32–38). The nucleotide sequences were aligned with MAFFT and computed with SARS-CoV-2 version of IQ-TREE with 1,000 bootstraps. The Wuhan reference strain (NC_045512.2) was used as an outgroup and the root. Black balls show nodes with bootstrap values above 80.

Supplementary Files

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

- Supplementarymaterial.zip