Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020

Emma B. Hodcroft,[1,2] Moira Zuber,[1] Sarah Nadeau,[3,2] Iñaki Comas,[4,5,6] Fernando González Candelas,[7,5,6] SeqCOVID-SPAIN consortium,[8] Tanja Stadler*,[3,2] and Richard A. Neher*[1,2]
[1] *Biozentrum, University of Basel, Basel, Switzerland*
[2] *Swiss Institute of Bioinformatics, Basel, Switzerland*
[3] *D-BSSE, ETHZ, Basel, Switzerland*
[4] *Tuberculosis Genomics Unit, Biomedicine Institute of Valencia (IBV-CSIC), Valencia, Spain*
[5] *CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain*
[6] *on behalf or the SeqCOVID-SPAIN consortium*
[7] *Joint Research Unit "Infection and Public Health" FISABIO-University of Valencia, Institute for Integrative Systems Biology (I2SysBio), Valencia, Spain*
[8] *SeqCOVID-SPAIN consortium*

A variant of SARS-CoV-2 emerged in early summer 2020, presumably in Spain, and has since spread to multiple European countries. The variant was first observed in Spain in June and has been at frequencies above 40% since July. Outside of Spain, the frequency of this variant has increased from very low values prior to 15th July to 40-70% in Switzerland, Ireland, and the United Kingdom in September. It is also prevalent in Norway, Latvia, the Netherlands, and France. Little can be said about other European countries because few recent sequences are available. Sequences in this cluster (20A.EU1) differ from ancestral sequences at 6 or more positions, including the mutation `A222V` in the spike protein and `A220V` in the nucleoprotein. We show that this variant was exported from Spain to other European countries multiple times and that much of the diversity of this cluster in Spain is observed across Europe. It is currently unclear whether this variant is spreading because of a transmission advantage of the virus or whether high incidence in Spain followed by dissemination through tourists is sufficient to explain the rapid rise in multiple countries.

CAVEATS:

- **This variant rose in frequency in multiple countries, but we have no direct evidence that it spreads faster. The rise in frequency could also be due to epidemiological factors.**
- **There are currently no data to evaluate whether this variant affects the severity of the disease.**
- **While dominant in some countries, 20A.EU1 has not taken over everywhere and diverse variants of SARS-CoV-2 continue to circulate across Europe.**

INTRODUCTION

Following its emergence in Wuhan in late 2019 (WHO Emergency Committee, 2020; Zhu et al., 2020), SARS-CoV-2 has caused a global pandemic resulting in unprecedented efforts to reduce transmission and develop therapies and vaccines. The spread of the virus across the world has been tracked with phylogenetic analysis of viral genome sequences (Worobey et al., 2020; Hadfield et al., 2018; Pybus et al., 2020) which were and still are generated at a rate far greater than for any other pathogen. More than 157,000 full genomes are available in GISAID as of October 2020 (Shu and McCauley, 2017).

In addition to tracking the viral spread, these genome sequences have been used to monitor mutations which might change the transmission, pathogenesis, or antigenic properties of the virus. One mutation in particular, D614G in the spike protein, has received much attention. This variant (Nextstrain clade 20A) seeded large outbreaks in Europe in early 2020 and subsequently dominated the outbreaks in the Americas, thereby largely replacing previously circulating lineages. This rapid rise has led to the suggestion that this variant is more transmissible (Korber et al., 2020; Volz et al., 2020).

While the virus spread globally in early 2020 before borders were closed and viral variants circulating were distributed across the world, intercontinental travel remained suppressed through the summer of 2020. The paucity of intercontinental travel allowed continent-specific variants to emerge. Within Europe, however, travel resumed in the summer of 2020. Here we report on a novel SARS-CoV-2 variant 20A.EU1 (S:A222V) that emerged in early summer 2020, presumably in Spain, and subsequently spread to multiple locations in Europe. Over the summer, it rose in frequency in parallel in multiple countries. As we report here, this variant, 20A.EU1, and a second variant 20A.EU2 with mutation `S:S477N` in the spike protein account for the majority of recent sequences in Europe. It is unclear at present whether the rapid spread of either variant is due to association with particular demographics, properties of the virus, or

chance but the dynamics of both should be carefully monitored.

METHODS

Phylogenetic analysis

We use the Nextstrain pipeline for our phylogenetic analyses `https://github.com/nextstrain/ncov/` (Hadfield et al., 2018). Briefly, we align sequences using mafft (Katoh et al., 2002), subsample sequences (see below), add sequences from the rest of the world for phylogenetic context based on genomic proximity, reconstruct a phylogeny using IQ-Tree (Minh et al., 2019) and infer a time scaled phylogeny using TreeTime (Sagulenko et al., 2018). For computational feasibility, ease of interpretation, and to balance disparate sampling efforts between countries, the Nextstrain-maintained runs subsample the available genomes across time and geography, resulting in final builds of ∼4,000 genomes each.

Sequences were downloaded from GISAID using the nextstrain/ncov workflow. A table acknowledging the invaluable contributions by many labs is available as a supplement. We focus in particular on the epidemics in the UK, Switzerland (dataset described in Nadeau et al. (2020)) and Spain for which we have most sequencing data.

Defining the 20A.EU1 Cluster

The cluster was initially identified as a monophyletic group of sequences stemming from the larger 20A clade with amino acid substitutions at positions `S:A222V`, `ORF10:V30L`, and `N:A220V` or `ORF14:L67F` (overlapping reading frame with N), corresponding to nucleotide mutations `C22227T`, `C28932T`, and `G29645T`. In addition, sequences in 20A.EU1 differ from their ancestors by the synonymous mutations `T445C`, `C6286T`, and `C26801G`. There are currently 7,906 sequences in the cluster by this definition.

The sub-sampling of the standard Nextstrain analysis means that we are not able to visualise the true size or phylogenetic structure of the cluster in question. To specifically analyze this cluster using all available sequences, we designed a specialized build which focuses on cluster-associated sequences and their most genetically similar neighbours.

We identify sequences in the cluster based on the presence of nucleotide substitutions at positions 22227, 28932, and 29645 and use this set as a 'focal' sample in the nextstrain/ncov pipeline. This selection will exclude any sequences with no coverage or reversions at these positions, but the similarity-based sampling during the Nextstrain run will identify these, as well as any other nearby sequences, and incorporate them into the dataset. We used these three mutations as they included the largest number of sequences that are distinct to the cluster. By this criterion, there are currently 7,864 sequences in the cluster – slightly fewer than above because of missing data at these positions.

Variant	Representative Mutations	Spike Substitution
20A.EU1	C22227T, C28932T, G29645T	A222V
20A.EU2	C4343T, G5629T, G22992A	S477N
S:S98F	C21855T, A25505G, G25996T	S98F
S:D80Y	C3099T, G21800T, G27632T	D80Y

TABLE I Representative mutations of 20A.EU1 (the focus of this study) and other notable variants.

To visualise the changing prevalence of the cluster over time, we plotted the proportion of sequences identified by the four substitutions described above as a fraction of the total number of sequences submitted, per ISO week. Frequencies of other clusters are identified in an analogous way. Case data for Fig S2 were obtained from ECDC (European Center for Disease Control, 2020).

Phylogeny and Geographic Distribution

The size of the cluster and number of unique mutations among individual sequences means that interpreting overall patterns and connections between countries is not straightforward. We aimed to create a simplified version of the tree that focuses on connections between countries and de-emphasizes onward transmissions within a country. As our focal build contains 'background' sequences that do not fall within the cluster, we used only the monophyletic clade containing the four amino-acid changes and three synonymous nucleotide changes that identify the cluster. Then, subtrees that only contain sequences from one country were collapsed into the parent node. The resulting phylogeny contains only mixed-country nodes and single-country nodes that have mixed-country nodes as children. Nodes in this tree thus represent ancestral genotypes of subtrees: sequences represented within a node may have further diversified within their country, but share a set of common mutations. We count all sequences in the subtrees towards the geographic distribution represented in the pie-charts in Fig. 3.

This tree allows us to infer lower bounds for the number of introductions to each country, and to identify plausible origins of those introductions. It is important to remember that, particularly for countries other than the United Kingdom, the full circulating diversity of the variant is probably not being captured, thus intermediate transmissions cannot be ruled out. In particular, the closest relative of a particular sequence will often have

been sampled in the UK simply because sequencing efforts in the UK exceed most other countries by orders of magnitude. It is, however, not our goal to identify all introductions but to investigate large scale patterns of spread in Europe.

Growth rate estimates

To estimate the growth rate of the cluster, we fit logistic curves to countries with ≥ 150 sequences available. In order to examine more trends, we split the United Kingdom into its constituent countries. To model the variant frequency over time t, we fit a logistic curve $f(t) = e^{s(t-t_{50})}/(1+e^{s(t-t_{50})})$ with growth rate s and inflection point t_{50} to the number of sequences k_i in week i that fall into 20A.EU1 and those $n_i - k_i$ that do not by maximizing the log-likelihood

$$C + \sum_i \left[k_i \log(f(t_i)) + (n_i - k_i)\log(1 - f(t_i))\right] \quad (1)$$

where C is a constant. The logistic frequency trajectory was clamped between 0.01 and 0.99 to decrease sensitivity to outliers and increase robustness of the fit.

Estimating confidence intervals of the parameters is tricky because the sequences are most likely not independent samples. To nevertheless get a sense of confidence, we resample the data by week with replacement to produce 100 bootstrap estimates and report the interquartile range of these bootstraps.

All code used for the above analyses is available at github.com/emmahodcroft/cluster_scripts. The code used to run the cluster builds is available at github.com/emmahodcroft/ncov_cluster.

RESULTS

Figure 1 shows a time scaled phylogeny of sequences sampled in Europe and their global context colored by the amino-acids at position 222, 477, and 614 in the spike protein. Clade 20A and its daughter clades 20B and 20C have variant `S:D614G` and are colored in yellow. A cluster of sequences in clade 20A has an additional mutation `S:A222V` colored in blue. We designate this cluster as 20A.EU1. Another prominent cluster (20A.EU2) with the substitution `S:S477N` in the spike protein is common in France, but appears less prevalent than 20A.EU1 in other countries where both variants have circulated (Fig. S3).

Our analysis here focuses on the variant 20A.EU1 with substitution S:A222V. The substitution `S:A222V` is in the spike protein's domain A (also referred to as the NTD), and is not in a portion of spike that is known to play a direct role in receptor binding or membrane fusion, though mutations can sometimes mediate long-range effects on protein conformation or stability. However, 20A.EU1 has increased rapidly in size over the summer and now accounts for a large fraction of sequences in several European countries. When all lineages circulating in Switzerland since 1 May are considered, the notable rise and expansion of 20A.EU1 is clear (see Fig S4).

Updated phylogenies of SARS-CoV-2 of Europe and individual European countries are provided at nextstrain.org/groups/neherlab. The page also includes links to analysis of the individual clusters discussed in here (the cluster build for 20A.EU1 can be found at nextstrain.org/groups/neherlab/ncov/20A.EU1).

Earliest Sequences in the 20A.EU1 cluster

The earliest sequences identified date from the 20th of June, when 7 Spanish sequences and 1 Dutch sequence were sampled. The next non-Spanish sequence was from the UK (England) on the 18th July, with a Swiss sequence sampled on the 22nd and an Irish sampled on the 23rd. By the end of July, samples from Spain, the UK (England, Northern Ireland), Switzerland, Ireland, Belgium, and Norway were identified as being part of the cluster. By the 22nd August, the cluster also included sequences from France, more of the UK (Scotland, Wales), Germany, Latvia, Sweden, and Italy. Two sequences from Hong Kong and nine sequences from New Zealand, presumably exports from Europe, were first detected in mid-August and mid-September, respectively.

The proportion of sequences from each country which fall into the cluster, by ISO week, is plotted in Fig 2. Here, we included countries with at least 20 sequences from the cluster. By plotting the weekly proportion of sequences submitted by a country that fall into the cluster, we can see how the cluster-associated sequences have risen in frequency (Fig 2). As might be expected from the much earlier first sample date, the cluster first rises in frequency in Spain, initially jumping to around 60% prevalence within a month of the first sequence being detected. In the United Kingdom, France, Ireland, and Switzerland we observe a gradual rise starting in mid-July. In Wales and Scotland the variant was at 80% in mid-September, whereas frequencies in Switzerland and England were around 50% at that time. In contrast, Norway observed a sharp peak in early August, but few sequences are available for later dates. The date ranges and number of sequences observed in this cluster are summarized in Table II.

To quantify the growth of this cluster more precisely, we fit a logistic growth curve to the observed frequency of 20A.EU1 in Switzerland, Spain, England, Scotland and Wales, see Fig. S1. The estimated growth rate varies between around $s = 30 - 40\%$ per week in all five countries.

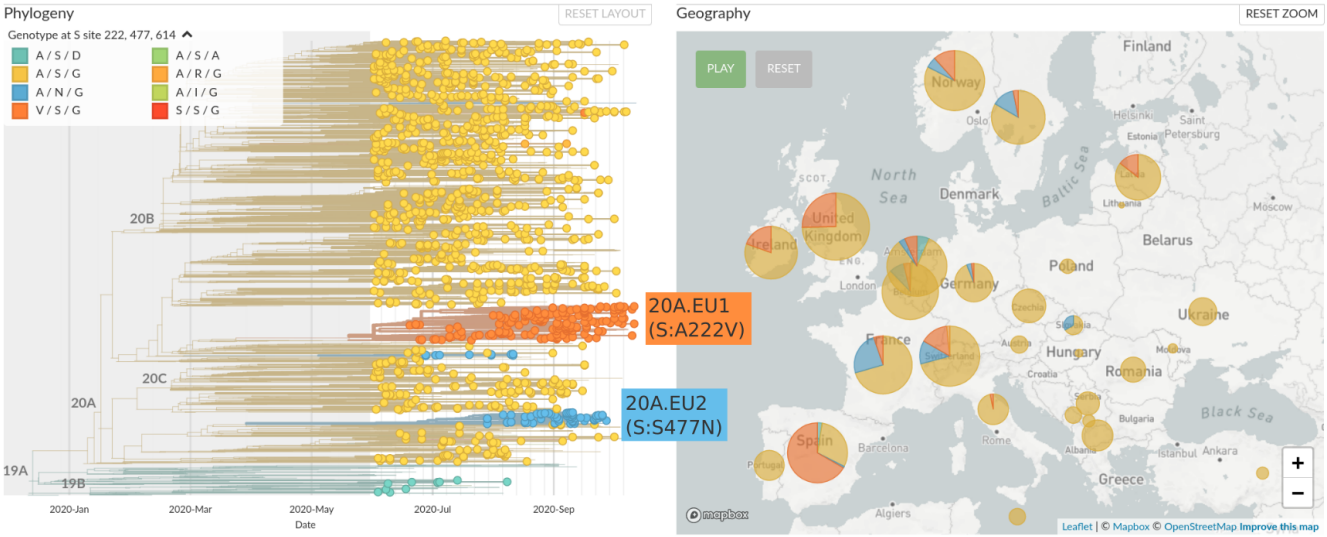

FIG. 1 Phylogenetic overview of SARS-CoV-2 in Europe. The tree shows a representative sample of isolates from Europe colored by the amino acid at positions 222, 477, and 614 of the spike protein. A novel rapidly spreading variant (blue; 20A.EU1) with mutation `S:A222V` on a `S:D614G` background emerged in early summer and is common in most countries with recent sequences. A separate variant (20A.EU2) with mutation `S:S477N` is prevalent in France. On the right, the proportion of sequences belonging to each variant (through the duration of the pandemic) is shown per country. Tree and visualization were generated using the Nextstrain platform (Hadfield et al., 2018) as described in methods.

Country	First Observation	# Sequences	Last Observation	Frequency in Sept & Oct
Netherlands	2020-06-20	49	2020-09-28	0.21
Spain	2020-06-20	256	2020-09-20	0.88
United Kingdom	2020-07-18	7207	2020-10-13	0.43
England	2020-07-18	4732	2020-10-13	0.36
Northern Ireland	2020-07-21	41	2020-09-23	0.14
Scotland	2020-08-01	926	2020-10-11	0.66
Wales	2020-08-03	1508	2020-10-10	0.74
Switzerland	2020-07-22	179	2020-09-10	0.37
Ireland	2020-07-23	56	2020-09-16	0.51
Norway	2020-07-29	59	2020-08-19	0
Belgium	2020-07-29	10	2020-09-23	0.11
France	2020-08-01	21	2020-09-16	0.12
Sweden	2020-08-14	3	2020-09-16	0.17
Hong Kong	2020-08-15	2	2020-08-17	0
Germany	2020-08-17	2	2020-08-19	0
Latvia	2020-08-22	10	2020-08-25	0
Italy	2020-08-25	1	2020-08-25	0
New Zealand	2020-09-22	9	2020-10-08	0.16

TABLE II Summary of sequences observed in the 20A.EU1 cluster

Cluster Source and Number of Introductions across Europe

Fig. 3 shows a collapsed phylogeny, as described in Methods, indicating the observations of different genotypes across Europe. The prevalence of early samples in Spain, diversity of the Spanish samples, and prominence of the cluster in Spanish sequences suggest Spain as the likely origin for the cluster. Throughout July and August 2020, Spain had a higher per capita incidence than most other European countries (see Fig S5) further supporting Spain as origin of the cluster.

Since it is unlikely that diversity and phylogenetic patterns sampled in multiple countries arose independently, it is reasonable to assume that the majority of mutations within the cluster arose once and were carried (possibly multiple times) between countries. We use this rationale

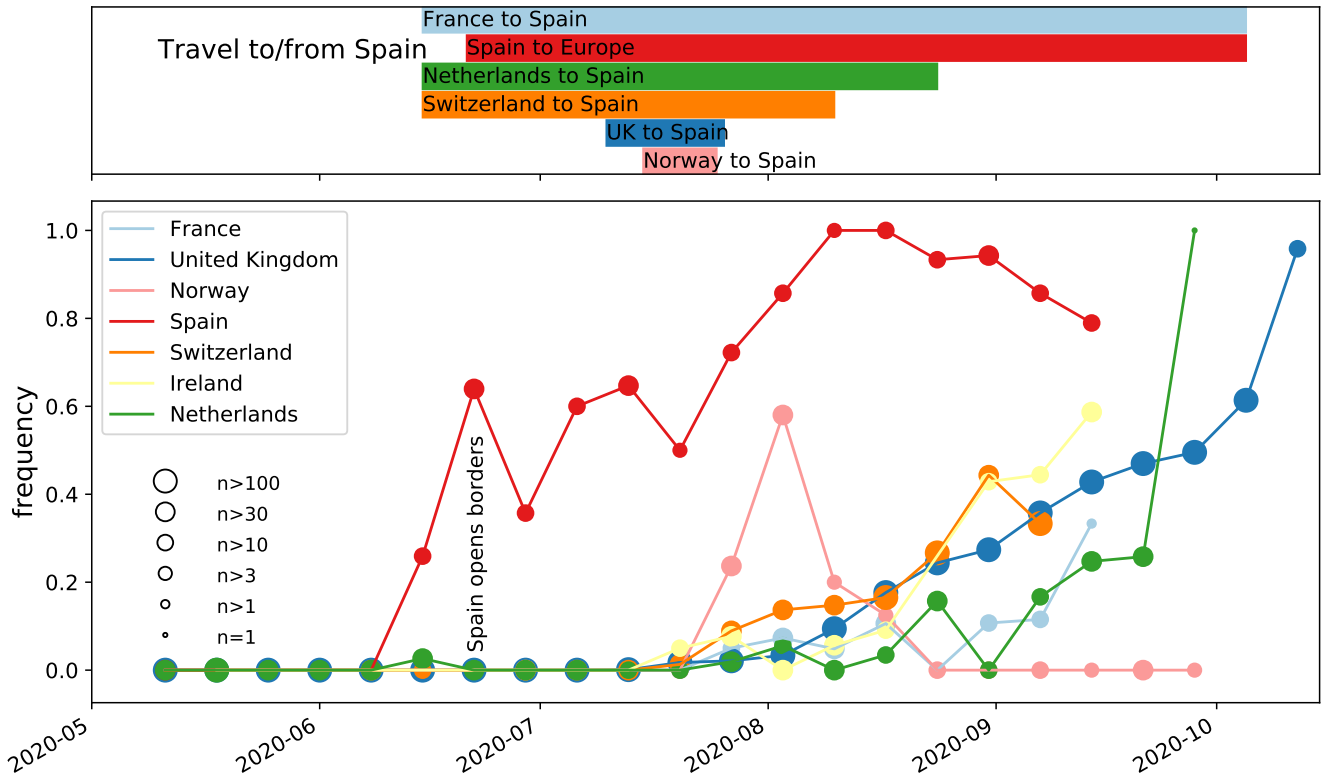

FIG. 2 **Frequency of submitted samples that fall within the cluster, with quarantine-free travel dates shown above.** We include the seven countries which have at least 20 sequences from the considered cluster. The symbol size indicates the number of available sequence by country and time point in a non-linear manner. Travel restrictions are shown to/from Spain, as this is the possible origin of the cluster. Most European countries allowed quarantine-free travel to other (non-Spanish) countries in Europe for a longer period.

below to provide lower bounds on the number of introductions to different countries.

The 256 sequences in the cluster from Spain likely do not represent the full diversity. Variants found only outside of Spain may reflect diversity acquired since being transmitted to another country, or may represent diversity not sampled in Spain. In particular, as the UK sequences much more than any other country in Europe, it is not unlikely they may have sampled diversity that exists in Spain but has not yet been sampled there. Despite limitations in sampling, Fig. 3 clearly shows that most major genotypes in this cluster were distributed to multiple European countries.

Per-Country Inferences

From inspecting the phylogenetic relationships between samples (Fig 3), we can estimate the number of times the 20A.EU1 cluster has been introduced into each country. In some cases countries have only one introduction, but many countries have indications of multiple separate introductions, and these we will cover in more detail below.

There are only eleven non-European samples in the cluster, from Hong Kong and New Zealand. Both are likely exports from Europe: the Hong Kong sequences indicate a single introduction, whereas the New Zealand samples are from at least three separate transmissions from Europe.

In order to estimate the number of introductions and try to infer possible sources, the phylogenetic connections between countries (Fig 3) were linked with the date of the country's first sample in the cluster and travel restrictions. As well as the dates when countries reopened their borders, we considered periods during which quarantine-tree travel was possible and travel volume to and from Spain (Fig. S8). Many EU and Schengen-area countries, including Switzerland, the Netherlands, and France, opened their borders to other countries in the bloc on 15th June, though the Netherlands kept the United Kingdom on their 'orange' list (NL Times, 2020; Occupational Health & Safety and Environmental Protection Unit at CERN, 2020; The Federal Council of the Swiss Confederacy, 2020). Spain opened its borders to EU member states (except Portugal, at Portugal's request) and associated countries on 21st June (McMurtry,

2020).

Norway, Latvia, Germany, Italy, Sweden: The sequences from Norway, Latvia, Germany, and Italy all indicate single introduction events, whereas Sweden's three sequences each indicate a separate introduction. Italy, Germany, and Sweden have only a very small number of sequences – one, two, and three, respectively – meaning that many introductions might have been missed. Norway and Latvia's larger sequence counts form two clear separate monophyletic groups within the 20A.EU1 cluster. The Norwegian samples seem likely to be a direct introduction from Spain, as they cluster tightly with Spanish sequences and the first sample (29th July) was just after quarantine-free travel to Spain was stopped. In Latvia, quarantine-free travel to Spain was only allowed until the 17th July - a month before the first sequence was detected on 22nd August. Latvia allowed quarantine-free travel to other European countries for a longer period, and this introduction may therefore have come via a third country. Though the Latvian sequences cluster most closely with diversity found in the UK, the over-representation of UK sequences may explain this, and the introduction may have come from unsampled diversity in another country in Europe.

Spain: Linked epidemiological data from Spain indicates the earliest sequences in the cluster are associated with two known outbreaks in the north-east of the country. The cluster variant seems to have initially spread among agricultural workers in Aragon and Catalonia, then moved into the local population, where it was able to travel to the Valencia Region and on to the rest of the country. This initial expansion of the 20A.EU1 cluster via a suspected super-spreading event that spread into the local population and moved across the country may have been critical in increasing the cluster's prevalence in Spain just before borders re-opened.

Switzerland: Quarantine-free travel to Spain was possible from 15th June to 10th August. The majority of holiday return travel is expected from mid-July to mid-August towards the end of school holidays. The first sequence identified as part of the cluster was taken on 22nd July.

To estimate introductions, we consider 13 nodes where Swiss sequences share diversity with Spanish sequences, or have parental nodes that contain Spanish sequences, suggesting an introduction into Switzerland, possibly via a third country. Additionally, we see 6 nodes where a genotype was observed in Switzerland and in another country, suggesting either an additional import from Spain, a third country, or a transmission between Switzerland and the other country. Three of the 19 putative introductions involve more than twenty sequences, and seem to have grown rapidly, consistent with the growth of the overall cluster.

For those nodes that don't directly or through their parents share diversity with Spanish sequences, the Swiss sequences are most closely related to diversity found in the UK, France, and the Netherlands, suggesting possible transmission from other EU countries to Switzerland or diversity in Spain that was not sampled.

In a second Switzerland-specific analysis we compared the size of transmission chains started by sequences from within the 20A.EU1 cluster to outside the 20A.EU1 cluster. To do so, we identified introductions into Switzerland and downstream Swiss transmission chains through considering a tree of all available Swiss sequences combined with foreign sequences with high similarity to Swiss sequences (procedure described in Nadeau et al. (2020)). Overall, we estimate between 5-71 independent introductions of the 20A.EU1 variant into Switzerland are plausible based on available sequences. Introductions within the 20A.EU1 cluster tend to cause larger transmission chains compared to introductions of non-20A.EU1 strains (Fig. S7). In particular, although a similar percentage of introductions of 20A.EU1 variants and non-20A.EU1 strains give rise to local transmission chains, 17-40% of 20A.EU1 variant transmission chains contain more than 20 sequences, compared to 0-4% of non-20A.EU1 transmission chains introduced around the same time. The given ranges represent two different definitions of a transmission chain.

Belgium: Along with many European countries, Belgium reopened to EU and Schengen Area countries on the 15th June. Belgium employed a regional approach to travel restrictions, meaning that while travellers returning from some regions of Spain were subject to quarantine from the 6th of August, it was not until the 4th September that most of Spain was subject to travel restrictions. Belgian sequences share diversity with sequences from Spain, the UK, Switzerland, the Netherlands, and France, and stem from at least three separate introductions.

France: France has had no restrictions on EU and Schengen-area travel since it re-opened borders on the 15th of June. France's 21 sequences have stemmed from at least four introductions: two cluster directly with Spanish sequences and one stems directly from a parent with Spanish sequences. The remaining introduction is genetically further from the diversity sampled in Spain, and may indicate an introduction from another country, possibly the United Kingdom or Switzerland.

Netherlands: The Netherlands began imposing a quarantine on travellers returning from some regions of Spain on the 28th July (Dutch News, 2020a), increasing the areas from which travellers must quarantine until the whole of Spain was included on the 25th August (Dutch News, 2020b). The distribution of sequences from the Netherlands suggests at least six introductions. Four introductions share diversity with Spanish sequences, suggesting direct importations from Spain. In the remaining introductions, the sequences and their parent nodes share no diversity with Spanish sequences, but instead

share genotypes with sequences from the UK, Ireland, Switzerland, and Germany.

The UK and Ireland: The first sequences in the UK (England) which associate with the cluster are from the 18th July, in the middle of the period from the 10th to 26th July when quarantine-free travel to Spain was allowed for England, Wales, and Northern Ireland. The first Irish sequences to associate with the cluster were taken a short time later, on the 23rd of July.

The prevalence of sequences from the United Kingdom make introductions harder to quantify. To allow a conservative estimate, we considered each node with both sequences from Spain and the UK as a possible introduction, but considered nodes containing sequences from the UK that descend from these as expansion of that introduction in the UK. This method gives an estimate of at least 21 introductions from Spain into the UK. Many of those introductions are represented by dozens to hundreds of genomes, while one genotype present in the UK, carrying the `21614T` mutation, is responsible for almost a two-thirds of the sequences associated with the cluster in the country.

The fifty-six sequences that fall in the cluster from Ireland indicate at least five separate introductions. In three nodes, Irish sequences either share diversity with Spanish sequences or have parents that do, with the remaining Irish sequences clustering most closely with the diversity present in the UK. However, as mentioned before, the diversity in Spain is likely not fully represented in the tree, so direct transmission cannot be ruled out.

Differing Travel Restrictions in the UK and Ireland: Interestingly, while quarantine-free travel was allowed in England, Wales, and Northern Ireland from the 10th–26th July, Scotland refrained from adding Spain to the list of 'exception' countries until the 23th July (meaning there were only 4 days during which returnees did not have to quarantine). On the other hand, Ireland never allowed quarantine-free travel to Spain, but did allow quarantine-free travel from Northern Ireland, though general movement restrictions were in place until May. Similarly, Scotland allowed quarantine-free travel to and from England, Wales, and Northern Ireland. Despite having only a very short or no period where quarantine-free travel was possible to Spain, both Scotland and Ireland have cases linked to the cluster.

DISCUSSION

Real-time genomic epidemiology allows us to track the spread of a pathogen through the mutations that accumulated in the genome during viral replication. The great majority of these mutations are of little functional relevance and merely serve as neutral markers that we can use to link related variants. Some mutations, however, are adaptive and increase in frequency because they increase the rate a which the virus transmits. Such adaptations are expected after a zoonosis when a pathogen is not yet fully adapted to its new host (Diehl et al., 2016; van Dorp et al., 2020) or in endemic pathogens that escape preexisting immunity, as is common for example in seasonal influenza viruses (Rambaut et al., 2008).

During a dynamic outbreak, it is particularly difficult to unambiguously tell whether a particular variant is increasing in frequency because it has an intrinsic advantage, or because of epidemiological factors (Grubaugh et al., 2020). In fact, it is a tautology that every novel big cluster must have grown recently and multiple lines of independent evidence are required in support of an intrinsically elevated transmission potential.

The cluster we describe here – 20A.EU1 (S:A222V) – was dispersed across Europe by travelers to and from Spain and repeated imports might be sufficient to explain the rapid rise in frequency and the displacement of other variants. In July and August Spain reported approximately 20 and 90 cases per 100k inhabitants per week, respectively. Taking reported incidence at face value and assuming that returning tourists have a similar incidence, we expect more than 50 introductions to the UK in July and more than 200 in August (see Table SI for tourism summaries (Instituto Nacional de Estadistica, 2020), total travel volume is about 2-3 fold higher, see Fig. S8). Similarly, Switzerland would expect around 15 and 40 introductions in July and August. While these numbers are only rough order of magnitude estimates that don't account for underreporting, stochastic establishment of new transmission chains, and ascertainment bias, they nevertheless indicate that wide-spread dispersal of this cluster across Europe is plausible. If these introductions occurred in demographics that were more likely to engage in risky behavior in Spain and continue to engage in such behavior at home, epidemiological factors alone could explain the rise of the 20A.EU1.

On the other hand, we observe a consistent and rapid growth of this variant in multiple countries and its frequency is above 50% in some localities. In countries where more than one introduction has occurred, it appears that multiple introductions have expanded. Its frequency in the UK has continued to increase even after quarantine-free travel was discontinued and the main summer travel period ended. Thus this variant might transmit faster than competing variants. Notably, case numbers across Europe started to rise rapidly around the same time the 20A.EU1 variant started to dominate in Spain, Switzerland and the UK (limited sequence data are available for other countries and we can not draw clear conclusions), see Fig. S2. However, this acceleration of transmission also coincided with the arrival of fall and seasonal factors are an alternative or compounding explanation (Neher et al., 2020).

In contrast to the early stages of the pandemic, many countries in Europe have experienced sustained low

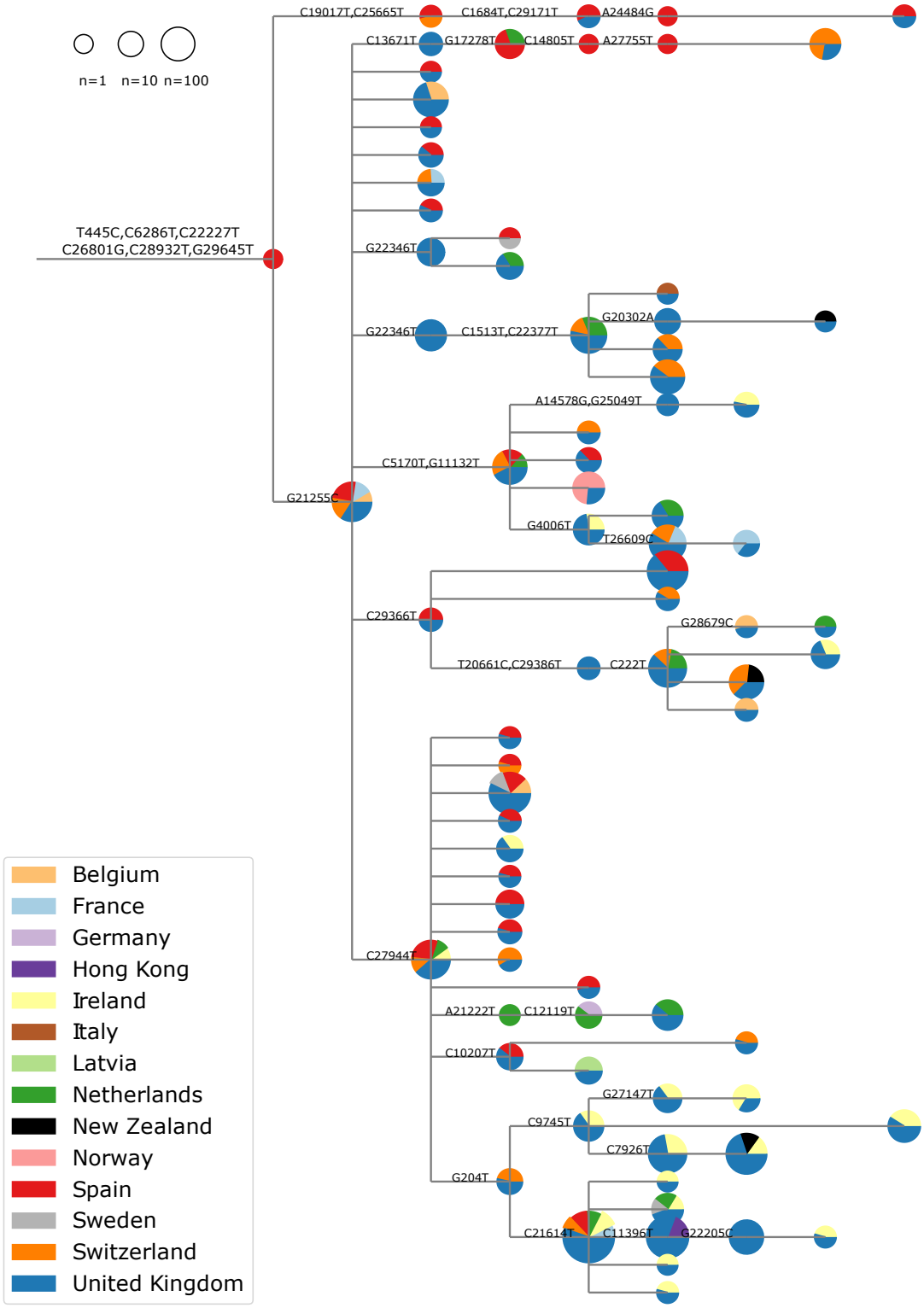

FIG. 3 **Collapsed genotype phylogeny.** The phylogeny shown is the subtree of the 20A.EU1 cluster, with sequences carrying all six defining mutations. Pie charts show the representation of sequences from each country at each node. Size of the pie chart indicates the total number of sequences at each node. Pie chart fractions scale non-linearly with the true counts (fourth root) to ensure all countries are visible.

level community transmission following the first wave. Founder effects and importations are expected to be less impactful if local circulation is higher. Comparatively high incidence over the summer (e.g. Belgium, see Figs. S5 and S6) might explain why 20A.EU1 remains at low frequencies in some countries despite high-volume travel to Spain.

In addition to the 20A.EU1 cluster we describe here, an additional variant (20A.EU2) with several amino acid substitutions, including `S:S477N` and mutations in the nucleocapsid protein, has become common in some European countries (see Fig. 1 and Fig. S3). The `S:S477N` substitution has arisen multiple times independently, for example in a variant in clade 20B that has dominated the recent outbreak in Oceania. The position 477 is close to the receptor binding site, and deep mutational scanning studies indicate that `S:S477N` slightly increases the receptor binding domain's affinity for ACE2 (Starr et al., 2020). Thus, the spread of this variant should also be closely monitored.

The rapid rise of these variants in Europe highlights the importance of genomic surveillance of the SARS-CoV-2 pandemic. If 20A.EU1 and/or the 20A.EU2 variant do increase the transmissibility of the virus, previously effective infection control measures might not longer be sufficient. Along similar lines, it is imperative to understand whether novel variants impact the severity of the disease. So far, we have no evidence for any such effect: The low mortality over the summer in Europe was predominantly explained by a marked shift in the age distribution of confirmed cases and this variant was not yet prevalent enough in July and August to have had a big effect. As sequences and clinical outcomes for patients infected with this variant become available, it will be possible to better infer whether this lineage has any impact on disease prognosis.

It is only through multi-country genomic surveillance that it has been possible to detect and track this cluster. However, the absence of consistent and uniform sequencing across Europe has still limited our efforts: we do not know the role this cluster may be playing in many countries' outbreaks because no or very few sequences have been shared. This work underscores the importance of a coordinated and regular sequencing effort, particularly in closely-associated communities like Europe, in order to detect, track, and analyze emerging SARS-CoV-2 variants.

Finally, our analysis highlights that countries should carefully consider their approach to travellers from areas with high SARS-CoV-2 incidence when travel resumes across Europe. Whether the 20A.EU1 cluster identified here has rapidly spread due to a transmission advantage or due to epidemiological factors alone, its observed introduction and rise in prevalence in multiple countries implies that the summer travel guidelines and restrictions were generally not sufficient to prevent onward transmission of introductions. While long-term travel restrictions and border closures are not tenable or desirable, identifying better ways to reduce the risk of introducing variants, and ensuring that those which are introduced do not go on to spread widely, will help countries maintain often hard-won low levels of SARS-CoV-2 transmission.

Acknowledgements

We are gratefully to researchers, clinicians, and public health authorities for making SARS-CoV-2 sequence data available in a timely manner. We also wish to thank the COVID-19 Genomics UK consortium for their notable sequencing efforts, which have provided more than half of the sequences currently publicly available. Jesse Bloom provided valuable feedback on the manuscript. This work was supported by the SNF through grant numbers 31CA30_196046 (to RAN, EBH) and 31CA30_196267 and core funding by the University of Basel. SeqCOVID-SPAIN is funded by the Instituto de Salud Carlos III project COV20/00140, Spanish National Research Council and ERC StG 638553 to IC.

Transparency declaration

The authors are not aware of any conflicts of interest.

Authors' contribution

EBH identified the cluster, led the analysis, and drafted the manuscript. RAN analysed data and drafted the manuscript. MZ and SN analysed data and created figures. IC and FGC interpreted the origin of the cluster and contributed data. All authors contributed to and approved the final manuscript.

REFERENCES

WHO Emergency Committee, Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov), 2020.

N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, Brief Report: A Novel Coronavirus from Patients with Pneumonia in China, 2019, The New England Journal of Medicine 382 (2020) 727. Publisher: NEJM Group.

M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America, Science (2020). Publisher: American Association for the Advancement of Science Section: Research Article.

J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics (2018).

O. Pybus, A. Rambaut, et al, Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages, 2020.

Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality, Eurosurveillance 22 (2017) 30494. Publisher: European Centre for Disease Prevention and Control.

B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus, Cell 182 (2020) 812–827.e19.

E. M. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, A. O'Toole, J. A. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, S. M. Rey, S. M. Nicholls, R. M. Colquhoun, A. d. S. Filipe, N. Pacchiarini, M. Bull, L. Geidelberg, I. Siveroni, I. G. Goodfellow, N. J. Loman, O. Pybus, D. L. Robertson, E. C. Thomson, A. Rambaut, T. R. Connor, T. C.-. G. U. Consortium, Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity, medRxiv (2020) 2020.07.31.20166082. Publisher: Cold Spring Harbor Laboratory Press.

K. Katoh, K. Misawa, K.-i. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research 30 (2002) 3059–3066.

B. Q. Minh, H. Schmidt, O. Chernomor, D. Schrempf, M. Woodhams, A. v. Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era, bioRxiv (2019) 849372.

P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis, Virus Evolution 4 (2018).

S. Nadeau, C. Beckmann, I. Topolsky, T. Vaughan, E. Hodcroft, T. Schar, I. Nissen, N. Santacroce, E. Burcklen, P. Ferreira, K. P. Jablonski, S. Posada-Cespedes, V. Capece, S. Seidel, N. S. de Souza, J. M. Martinez-Gomez, P. Cheng, P. Bosshard, M. P. Levesque, V. Kufner, S. Schmutz, M. Zaheri, M. Huber, A. Trkola, S. Cordey, F. Laubscher, A. R. Goncalves, K. Leuzinger, M. Stange, A. Mari, T. Roloff, H. Seth-Smith, H. Hirsch, A. Egli, M. Redondo, O. Kobel, C. Noppen, N. Beerenwinkel, R. A. Neher, C. Beisel, T. Stadler, Quantifying sars-cov-2 spread in switzerland based on genomic sequencing data, medRxiv (2020).

European Center for Disease Control, COVID-19 situation update worldwide, as of 19 October 2020, 2020.

NL Times, Dutch travel bans, domestic rules to ease up on Monday, 2020.

Occupational Health & Safety and Environmental Protection Unit at CERN, Travel restrictions in Switzerland and France (Covid-19), 2020.

The Federal Council of the Swiss Confederacy, Coronavirus: Switzerland to lift COVID restrictions regarding all EU/EFTA states, Federal Council Press Release (2020).

A. McMurtry, Spain opens borders as cases inch up again, AA (2020).

Dutch News, Barcelona no-go area for Dutch tourists as official risk rises to orange, DutchNews.nl (2020a).

Dutch News, All of Spain, more parts of France are code orange, as coronavirus cases rise, DutchNews.nl (2020b).

W. E. Diehl, A. E. Lin, N. D. Grubaugh, L. M. Carvalho, K. Kim, P. P. Kyawe, S. M. McCauley, E. Donnard, A. Kucukural, P. McDonel, S. F. Schaffner, M. Garber, A. Rambaut, K. G. Andersen, P. C. Sabeti, J. Luban, Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic, Cell 167 (2016) 1088–1098.e6.

L. van Dorp, M. Acman, D. Richard, L. P. Shaw, C. E. Ford, L. Ormond, C. J. Owen, J. Pang, C. C. S. Tan, F. A. T. Boshier, A. T. Ortiz, F. Balloux, Emergence of genomic diversity and recurrent mutations in SARS-CoV-2, Infection, Genetics and Evolution 83 (2020) 104351.

A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, E. C. Holmes, The genomic and epidemiological dynamics of human influenza A virus, Nature 453 (2008) 615–619.

N. D. Grubaugh, W. P. Hanage, A. L. Rasmussen, Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear, Cell 182 (2020) 794–795.

Instituto Nacional de Estadistica, Hotel Industry and Tourism – Tourist Movement on Borders Survey Frontur, 2020.

R. A. Neher, R. Dyrdak, V. Druelle, E. B. Hodcroft, J. Albert, Potential impact of seasonal forcing on a SARS-CoV-2 pandemic, Swiss Medical Weekly 150 (2020). Publisher: EMH Media.

T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding, Cell 182 (2020) 1295–1310.e20.

SUPPLEMENTARY MATERIAL

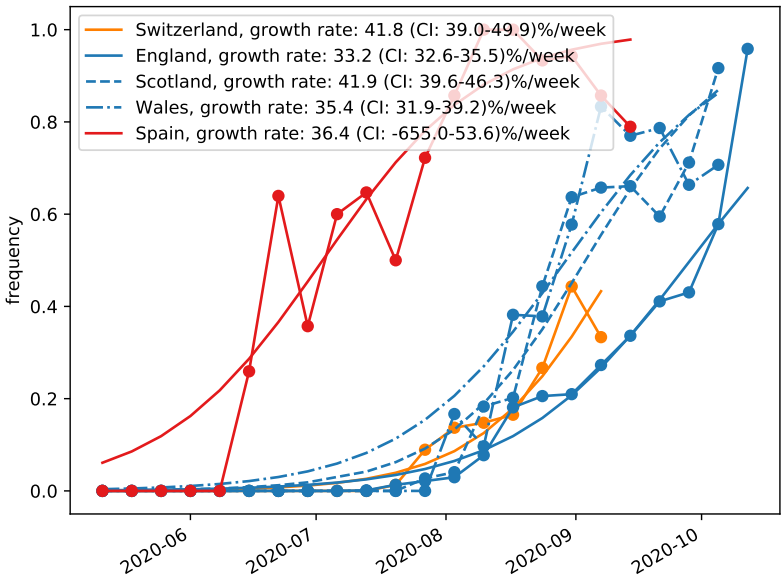

Figure S 1 The relative frequency of the novel variant has grown with a rate between 30 and 70% per week. These estimates were obtained by fitting a logistic curve to the number of sequences per week that are or are not part of the cluster for countries with more than 100 sequences in 20A.EU1. Confidence intervals indicate inter-quartile ranges obtained by bootstrapping data by week (see legend).

Country of residence	May	June	July	August	Total
Total	0	204926	2464441	2442999	5112366
France	0	64895	597244	863665	1525804
Germany	0	33740	432302	298217	764259
United Kingdom	0	8473	377886	256528	642887
Rest of Europe	0	21330	177896	234043	433269
Netherlands	0	12321	189995	151308	353624
Belgium	0	9608	154826	119284	283718
Italy	0	10426	103650	137978	252054
Portugal	0	0	90022	112767	202789
Other countries	0	0	70523	76879	147402
Nordic Countries	0	3965	95263	47990	147218
Switzerland	0	3610	83860	47578	135048
Rest of America	0	0	40822	55385	96207
Ireland	0	0	31323	25758	57081
United States of America	0	0	14943	12498	27441

Table S I Arrival statistics of tourists in Spain over the Summer 2020 (Instituto Nacional de Estadistica, 2020)

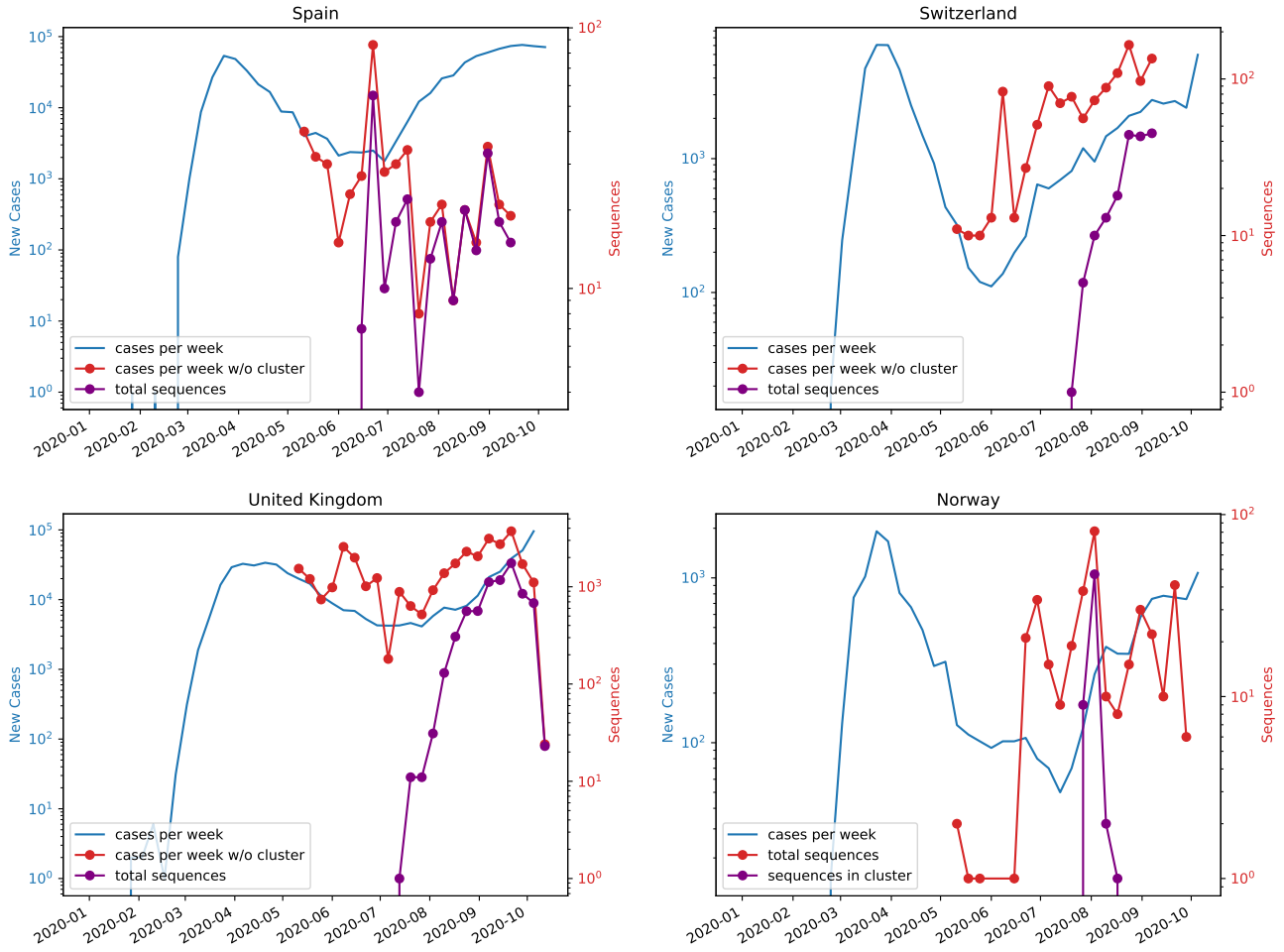

Figure S 2 Sequence availability varies by country, but Spain, Switzerland, and the United Kingdom have provided data until September. In the United Kingdom and Switzerland, the novel variant was first detected in July and rose quickly through August and September.

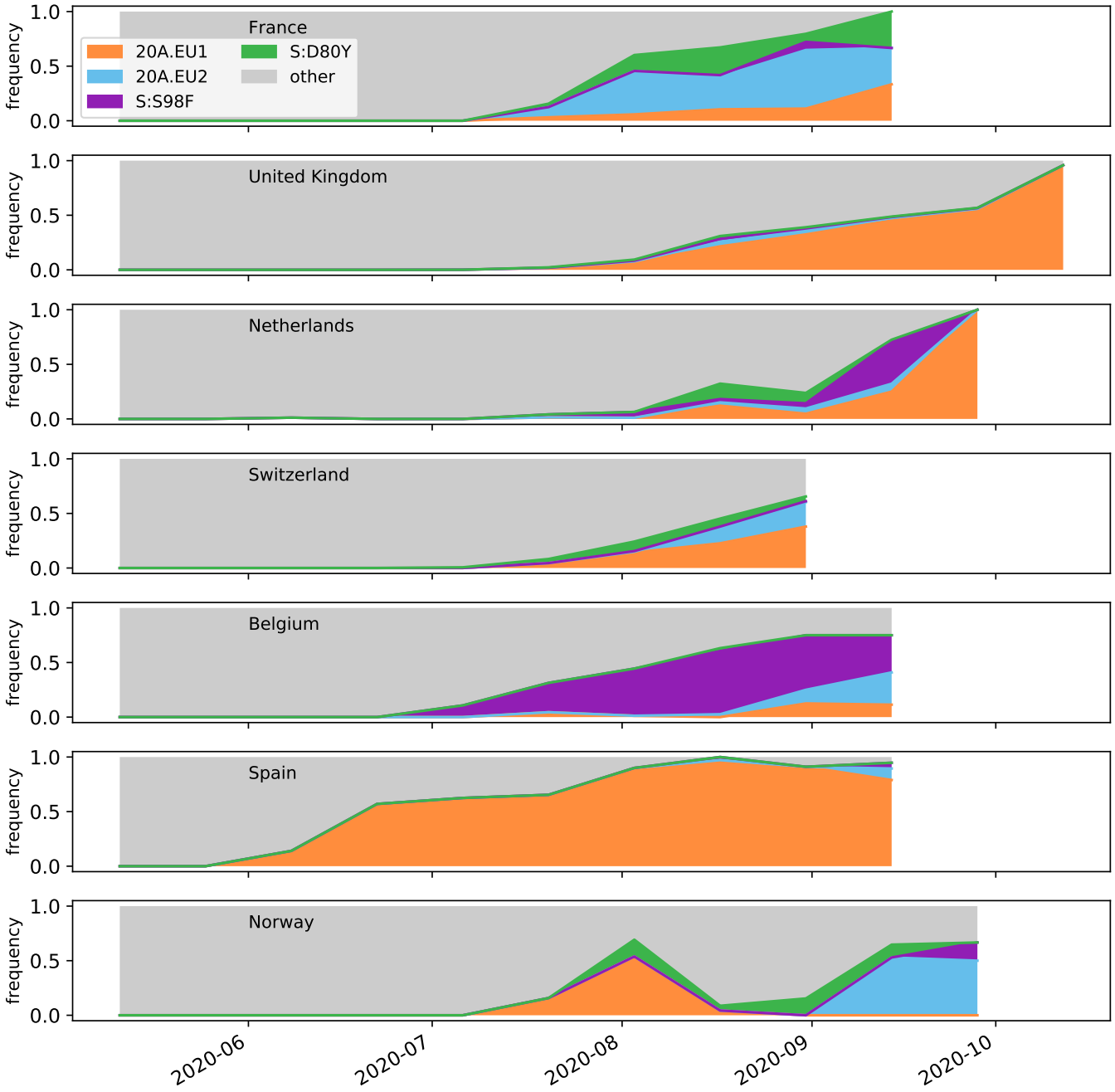

Figure S 3 In countries with at least ten sequences that fall into the 20A.EU1 cluster and into the 20A.EU2 cluster, the proportion of sequences per ISO week that fall into each cluster (or neither), as well as two other clusters that had some prominence in Europe, `S:D80Y` and `S:S98F`, is shown.

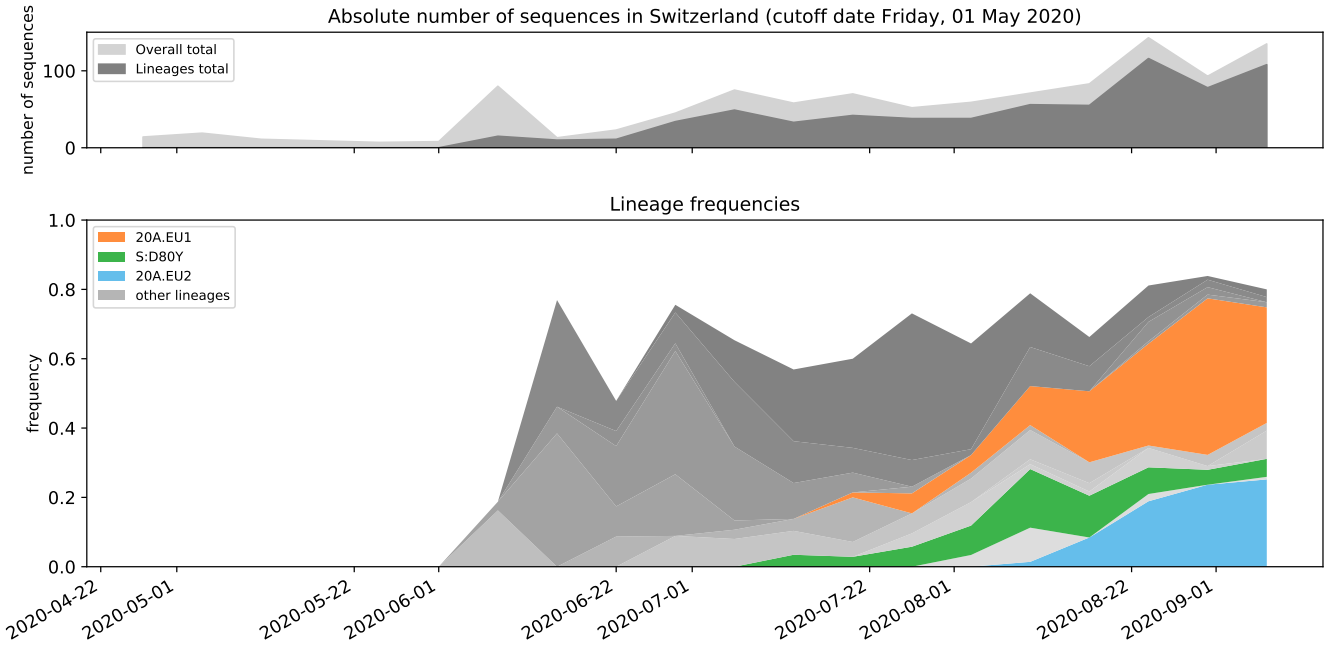

Figure S 4 Lineages found in a Swiss-focused Nextstrain build. A lineage is defined as a node present in the tree after the cutoff date of 1 May 2020 with at least 10 Swiss sequences as children. The 20A.EU1, 20A.EU2 and `S:D80Y` clusters are labelled.

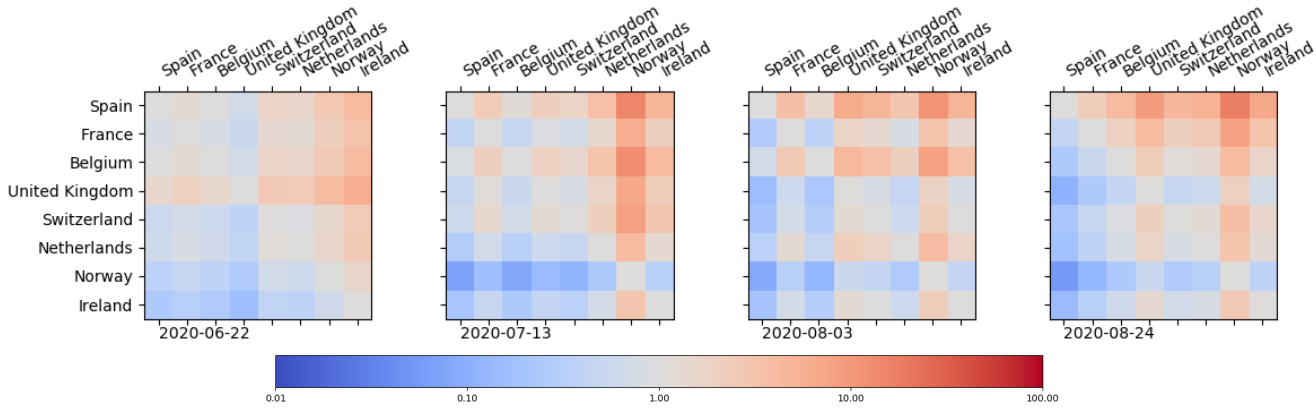

Figure S 5 Ratios of incidence in various countries at different time points over the summer. Countries listed down the left side are shown relative to countries listed across the top. Darker blue colors indicate the country on the left being more likely to be a 'sink' for the 'source' on the top. Darker red colors indicate the inverse.

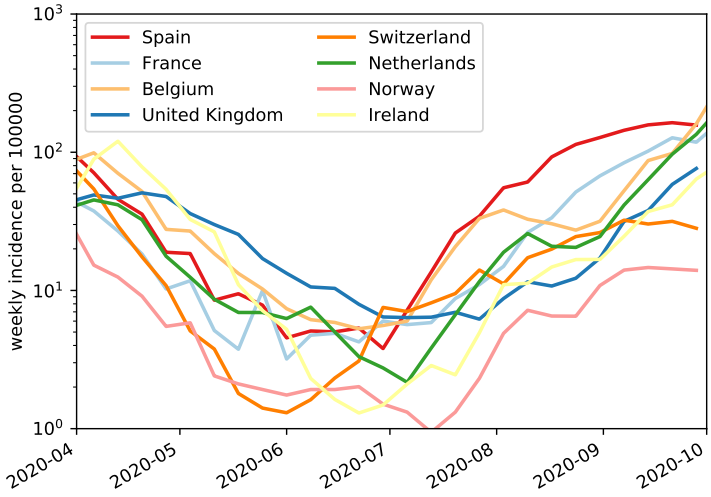

Figure S 6 Incidence in various countries over the summer. Spain and Belgium had relatively higher incidence from the start of July compared with other countries in Europe.

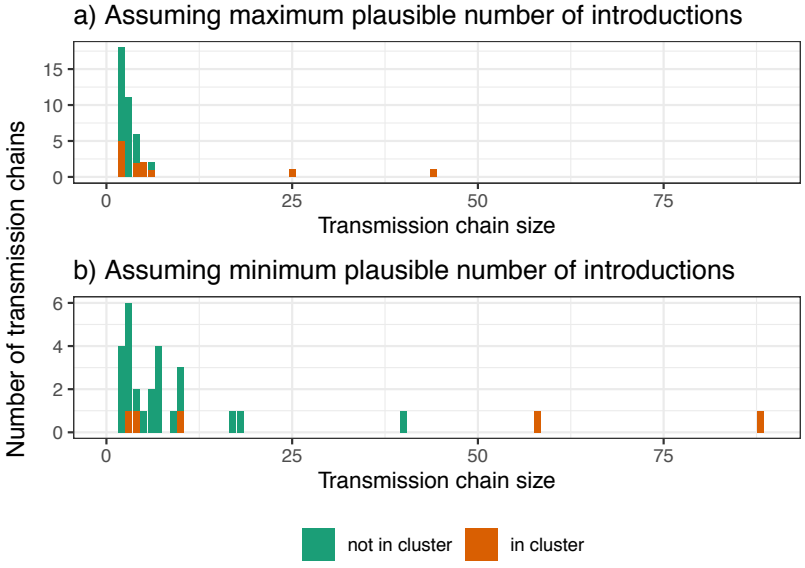

Figure S 7 The size of transmission chains caused by introductions into Switzerland.

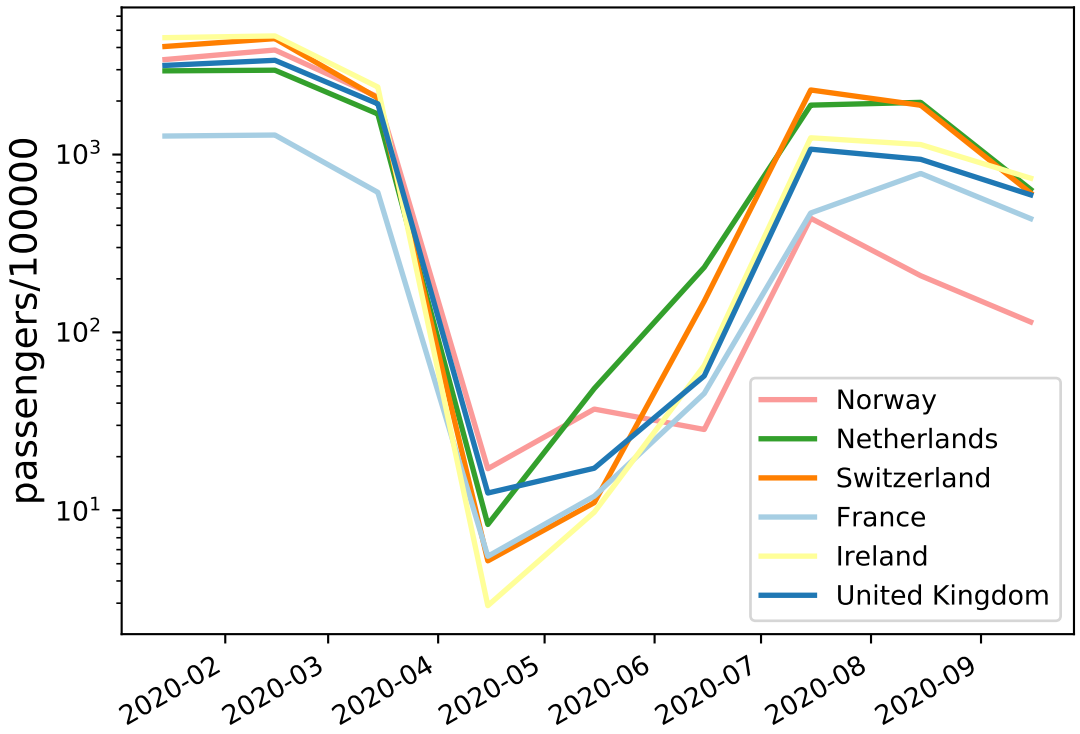

Figure S 8 Passenger volume to and from Spain in 2020.