nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00936-6

The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant

Mariana G. López[1,111], Álvaro Chiner-Oms[1,111], Darío García de Viedma[2,3,4], Paula Ruiz-Rodriguez[5], Maria Alma Bracho[6,7], Irving Cancino-Muñoz[1], Giuseppe D'Auria[7,8], Griselda de Marco[8], Neris García-González[6], Galo Adrian Goig[9], Inmaculada Gómez-Navarro[1], Santiago Jiménez-Serrano[1], Llúcia Martinez-Priego[8], Paula Ruiz-Hueso[8], Lidia Ruiz-Roldán[6], Manuela Torres-Puente[1], Juan Alberola[10,11,12], Eliseo Albert[13], Maitane Aranzamendi Zaldumbide[14,15], María Pilar Bea-Escudero[16], Jose Antonio Boga[17,18], Antoni E. Bordoy[19], Andrés Canut-Blasco[20], Ana Carvajal[21], Gustavo Cilla Eguiluz[22], Maria Luz Cordón Rodríguez[20], José J. Costa-Alcalde[23], María de Toro[16], Inmaculada de Toro Peinado[24], Jose Luis del Pozo[25], Sebastián Duchêne[26], Jovita Fernández-Pinero[27], Begoña Fuster Escrivá[12,28], Concepción Gimeno Cardona[28], Verónica González Galán[29], Nieves Gonzalo Jiménez[30], Silvia Hernáez Crespo[20], Marta Herranz[2,3,4], José Antonio Lepe[29], Carla López-Causapé[31], José Luis López-Hontangas[32], Vicente Martín[7,33], Elisa Martró[7,19], Ana Milagro Beamonte[34], Milagrosa Montes Ros[22], Rosario Moreno-Muñoz[35], David Navarro[13,12], José María Navarro-Marí[36,37], Anna Not[19], Antonio Oliver[31,38], Begoña Palop-Borrás[24], Mónica Parra Grande[39], Irene Pedrosa-Corral[36,37], Maria Carmen Pérez González[40], Laura Pérez-Lago[2,3], Mercedes Pérez-Ruiz[24], Luis Piñeiro Vázquez[22], Nuria Rabella[41,42,43], Antonio Rezusta[34,44,45], Lorena Robles Fonseca[46], Ángel Rodríguez-Villodres[29], Sara Sanbonmatsu-Gámez[36,37], Jon Sicilia[2,3], Alex Soriano[47], María Dolores Tirado Balaguer[35], Ignacio Torres[13], Alexander Tristancho[34,44], José María Marimón[22], SeqCOVID-Spain consortium*, Mireia Coscolla[5] ✉, Fernando González-Candelas[6,7] ✉ and Iñaki Comas[1,7] ✉

The coronavirus disease 2019 (COVID-19) pandemic has affected the world radically since 2020. Spain was one of the European countries with the highest incidence during the first wave. As a part of a consortium to monitor and study the evolution of the epidemic, we sequenced 2,170 samples, diagnosed mostly before lockdown measures. Here, we identified at least 500 introductions from multiple international sources and documented the early rise of two dominant Spanish epidemic clades (SECs), probably amplified by superspreading events. Both SECs were related closely to the initial Asian variants of SARS-CoV-2 and spread widely across Spain. We inferred a substantial reduction in the effective reproductive number of both SECs due to public-health interventions ($R_e < 1$), also reflected in the replacement of SECs by a new variant over the summer of 2020. In summary, we reveal a notable difference in the initial genetic makeup of SARS-CoV-2 in Spain compared with other European countries and show evidence to support the effectiveness of lockdown measures in controlling virus spread, even for the most successful genetic variants.

The new coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) emerged in China in October–November 2019 (ref. [1]) and by the end of March 2020 it was present in most countries of the world. The World Health Organization (WHO) declared the new disease as a pandemic on 11 March 2020. Spain suffered a severe epidemic, with the first case reported on 29 January 2020 (ref. [2]), and an accumulated number of 249,659 cases by 1 July 2020, including 28,363 fatalities[3]. Furthermore, a nationwide seroprevalence study showed that only one in ten cases of infection by SARS-CoV-2 were diagnosed and reported in that period[4], suggesting that the total number of infections has been vastly underestimated. Spain ordered a series of nonpharmaceutical intervention measures, including a general lockdown on 14 March 2020 (ref. [5]), later applied by many

A full list of affiliations appears at the end of the paper.

other countries, and was successful in reducing infection rates by the end of May 2020 (ref. [6]). Despite these measures, almost 30,000 individuals died during the first wave of the epidemic (until 14 May 2020), and a second wave of COVID-19 was beginning to emerge by the beginning of July 2020 (ref. [7]).

Despite the high incidence of infection across the country, some regions had substantially higher incidence than others. Genomic epidemiology and phylodynamics[8–10] offer a unique opportunity to understand the early events of the epidemic at the global, regional and local levels, to track the evolution of the epidemic after its initial stages and to quantify the impact of lockdown measures on the genetic variants of the virus. However, there are challenges and caveats that prevent the use of pathogen genomes as the sole source of interpretation. While there is now a large number of SARS-CoV-2 sequences deposited in GISAID[11], there are still important unsampled areas of the world, including some that played an important role in the initial spread of the epidemic. In addition, the virus spreads faster than it evolves[12,13] which limits the resolution of phylogenetic and phylodynamic analyses[14]. Finally, despite important efforts by sequencing consortiums, only a fraction of the total number of infections has been sequenced. Nevertheless, genomic epidemiology has played an important role in understanding the global and local epidemiology of COVID-19 (refs. [15–17]).

After the pandemic was declared in Spain, we assembled the National Consortium of genomic epidemiology of SARS-CoV-2 (http://seqcovid.csic.es/). This established a unique network incorporating more than 50 hospitals and scientific institutions across the country to collect clinical samples and epidemiological information from COVID-19 cases. Here, we present the results of this nationwide effort. We were able to sequence 12% of the reported cases before the national lockdown, and 1% of the reported cases of the first wave when lockdown measures ended (14 May 2020), including samples of SARS-CoV-2 across Spain in the early months of the pandemic (February–May 2020). Using a combination of pathogen genomics, phylogenetic tools and clinical and epidemiological data, we have been able to dissect the very early events in the dispersion of SARS-CoV-2 throughout Spain, as well the evolution of the virus during the exponential phase and after the lockdown. We document simultaneous introductions into the country from multiple sources. We show that up to 40% of cases were caused by two SECs, named SEC7 and SEC8. Seven other SECs were detected but their role was minor, probably because they were introduced relatively close to the lockdown and, unlike the initial two clades, had no opportunity for a rapid exponential expansion. In contrast to clades from other European countries, these SECs belong to early lineages in the epidemic (A in Pango, 19B in NextStrain). We also show that the reproductive number, R_e, of the most successful SECs declined quickly after the implementation of lockdown measures, and they were completely absent from samples taken in July–September 2020. Our results suggest that the most successful variants were those associated with earlier introductions, but also that their success may have depended on the synergy between superspreading events and high mobility. These results also show the effectiveness of lockdown measures in controlling the virus spread and eliminating established successful epidemic clusters from circulation.

SARS-CoV-2 was introduced multiple times from multiple sources

Our dataset consists of 2,170 sequences from Spain, collected under ethical approval, from 25 February to 22 June 2020, coinciding with the initial phases of the COVID-19 pandemic in the country (Fig. 1a). The most populated Spanish regions were sampled, resulting in a dataset with sequences representing 16 of the 17 administrative regions into which the country is divided (Fig. 1b). Of the 2,170 (90.4%) samples analyzed here, 1,962 have been sequenced by the SeqCOVID consortium, while the remaining 208 have been generated by independent laboratories and downloaded from GISAID[11] (Supplementary Table 1). Spain showed a particular viral population structure with a higher proportion of lineage A sequences compared with other European countries[18] (Fig. 1c). Strains from patients in Spain were related more closely to strains from cases sequenced in China, and were the most abundant during the first weeks of the epidemic in Spain. They were later replaced by lineage B strains (Extended Data Fig. 1), which differ by at least six to seven substitutions from lineage A and that dominated the beginning of the pandemic in most European countries, in contrast to the patterns seen in Spain. In addition, we observed a heterogeneous distribution of the SARS-CoV-2 genetic diversity within Spain, both at regional and local levels. For example, our analysis shows that viral diversity was higher in some urban areas, and it declined with geographic distance from the city centers, as observed in Valencia (Supplementary Note).

Phylogenomic analyses suggest the existence of multiple independent entries of the virus into Spain, similar to what was seen for other countries[19,20]. To identify possible introductions, we inspected the placement of Spanish viral samples in a global phylogeny constructed with more than 30,000 sequences (Fig. 1d). Given the low genetic diversity of the virus, particularly at the beginning of the epidemic, we found many instances in which a Spanish sample was genetically identical to other variants circulating in the rest of the world. According to their phylogenetic placement, three different possibilities were considered for the phylogenetic position of Spanish sequences. A sequence was included in a 'candidate transmission cluster' when it was found in a monophyletic clade with other Spanish sequences; it was included in a 'zero-distance' group when it grouped with other genetically identical Spanish sequences but also with other foreign sequences; and it was denoted as 'unique' when no matching sequence in the Spanish dataset was identified and the sequence differed by more than one single nucleotide polymorphism (SNP) from other Spanish sequences (Supplementary Fig.1a; detailed definitions of the groups are in Methods and Extended Data Fig. 2). We detected 224 'candidate transmission clusters', comprising 827 sequences (~40% of the Spanish samples); 30 'zero-distance clusters', comprising 831 sequences, and 513 'unique' sequences (Supplementary Fig. 2). Next, we determined how many unique cases and clusters were compatible with an introduction before the general lockdown. We detected that 191 groups (165 'candidate transmission clusters' plus 26 'zero-distance clusters') and 328 unique sequences met this criterion, representing at least 519 independent introductions (distribution of dates in Fig. 1e, distribution of 'unique' sequences across regions in Supplementary Fig. 1b). This is probably an underestimate of the total number of entries because the number of sequences analyzed is a small subset of the total notified cases (Fig. 1a). Phylogenetic analysis suggests that the most probable introduction of cases with a clear phylogenetic link (Methods) came from Italy, the Netherlands, England and Austria (accounting for ~23%, ~20%, ~13% and 12% of the cases for which a probable country of origin can be inferred, respectively) (Fig. 1f). The observation that more than half of the introduction events detected are unique sequences illustrates the disparate outcomes after an introduction, as some events resulted in large epidemiological clusters, and others disappeared leaving almost no trace. A clear example is the case of the first described death in Spain, for which we have generated a partial viral sequence. The patient was infected in Nepal but there were no identifiable secondary cases in our dataset.

A few genetic variants dominated the first wave in Spain

To identify those introductions that resulted in sustained transmission and, therefore, the ones that were epidemiologically successful in the long term, we scanned the phylogeny for larger clades comprised mainly of Spanish samples (Methods). We identified

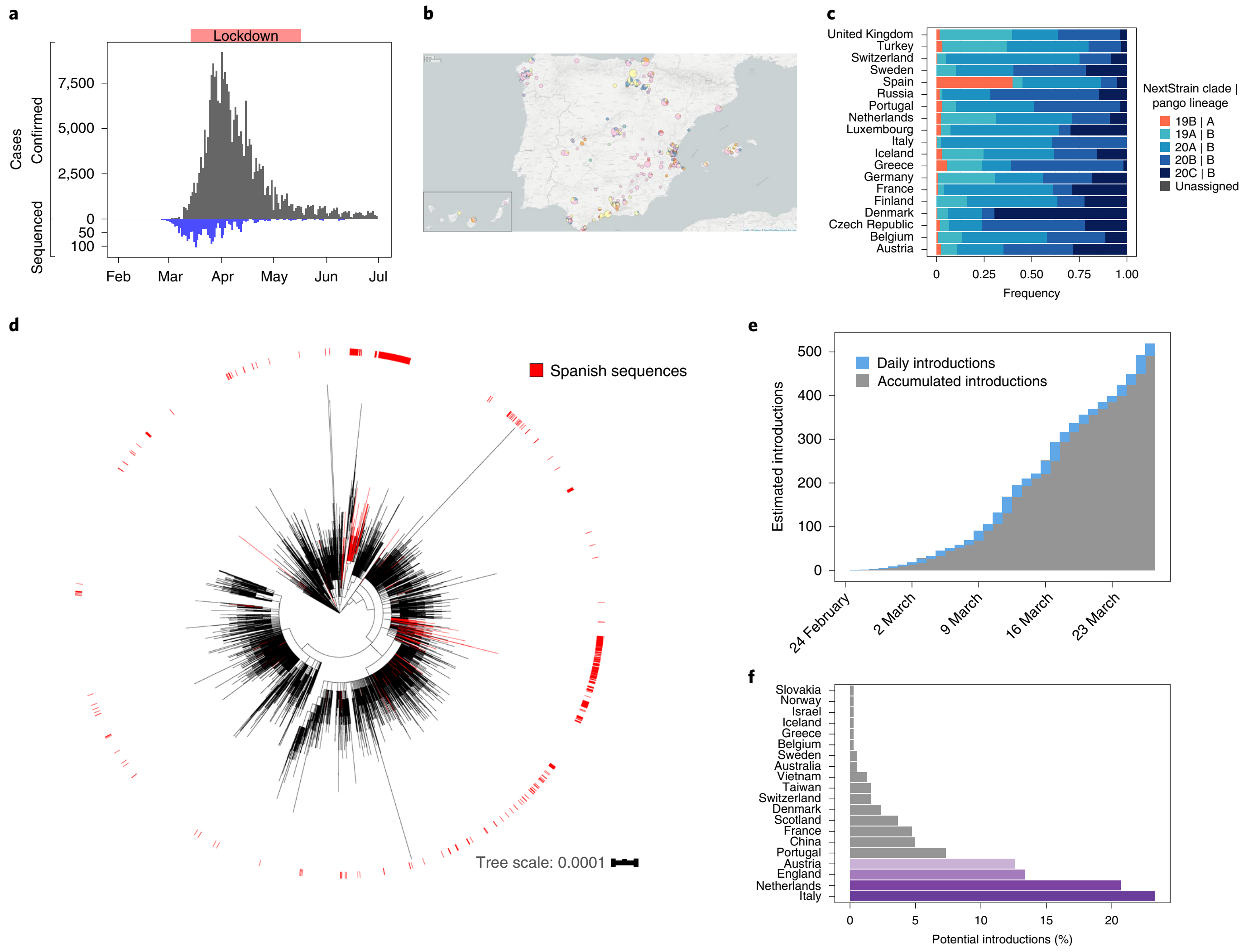


Fig. 1 | SARS-CoV-2 genomes sequenced from Spain. **a**, Distribution of sequenced samples (blue) versus confirmed cases in Spain (gray) by date. Country lockdown measures were in effect from 14 March to 14 May 2020. **b**, Distribution of the sequenced samples across Spain plotted in Microreact. These data can be explored with more detail in the Microreact webpage (https://microreact.org/showcase) loading the Supplementary Data 1 files. The size of each piechart correlates with the number of sequences collected in the corresponding area. Each color corresponds to a specific Pango lineage, as detailed in Extended Data Fig. 1 (light yellow and green correspond to lineage A, all the others are lineage B). The map image was extracted from the Microreact visualization[37]. **c**, Distribution of main SARS-CoV-2 clades during the first stages of the pandemic (before 1 April 2020), in those European countries with more than 50 sequences deposited in GISAID as of 13 November 2020. **d**, Global ML phylogeny constructed with 32,416 sequences, placement of Spanish samples is indicated in red. **e**, New and accumulated introductions to Spain. Lower-bound introduction estimates were defined as the probable date of the infection of the first case in a cluster (14 days before symptom onset). **f**, Estimated international origin of SARS-CoV-2 introductions based on phylogenetic data; in color, those countries with a probable contribution greater than 10%.

nine SECs distributed across the phylogeny, representing 46% of the total Spanish dataset analyzed (995 out of 2,170 samples) (Fig. 2a, Extended Data Figs. 3 and 4 and Supplementary Tables 1 and 2). We noticed first that only two SECs encompassed 30% and 10% of all Spanish samples (SEC8 and SEC7, respectively). This implies that the introduction of these two specific genetic variants explains a high proportion of the entire epidemic for the first wave in the country. In fact, they were responsible for 44% of the 'candidate transmission clusters' identified before the lockdown (Fig. 2b). We then estimated the time of introduction in Spain for the nine SECs using a Bayesian approach (Supplementary Table 2). As a conservative estimate we considered the time of introduction as any time between the age of the most recent common ancestor of the SEC and the date of the first Spanish sample (Fig. 2c). Thus, we assume that the ancestor of the SEC was not necessarily in Spain.

Our analysis shows that the earlier the introduction, the larger the size of the SEC (Supplementary Fig. 3). The larger clades, SEC7 and SEC8, were the first successful genetic variants introduced into Spain during late January–February 2020 (Fig. 2b). Both belong to lineage A (Pango nomenclature) and partially explain the particular population structure observed in Spain relative to other European countries (Fig. 1c). In addition, compared with other SECs, SEC7 and SEC8 were spread widely in the country, being present in at least 10 of the 17 administrative regions (Fig. 2b), and having a mean pairwise geographic distance between samples of more than 300 km, regardless of whether or not the Islas Canarias and Baleares are included (Extended Data Fig. 5). By contrast, SECs that were introduced later were smaller and showed a narrower geographic spread (between 0 and 58 km, analysis of variance (ANOVA)-adjusted P value « 0.01, Supplementary Note).

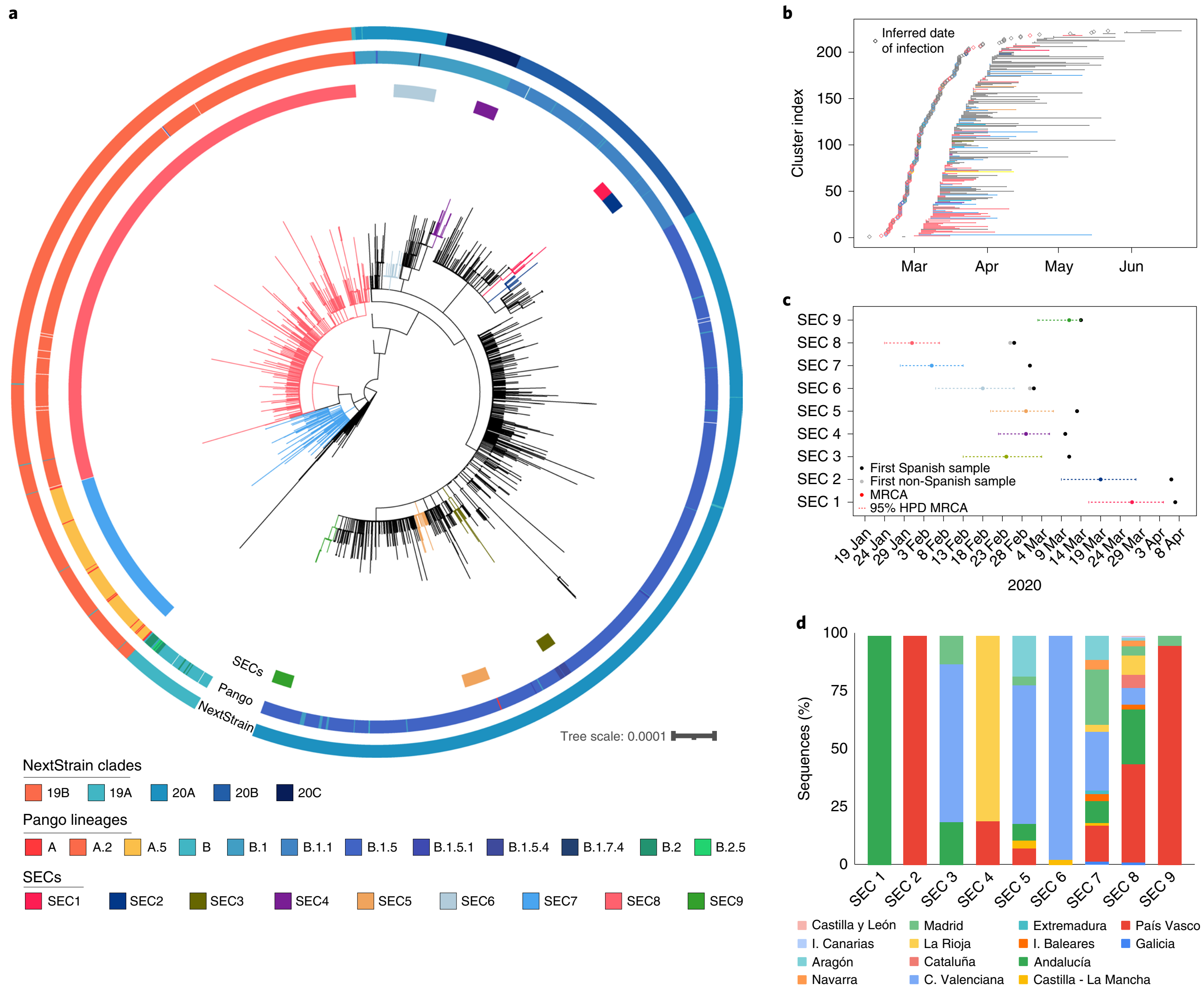


Fig. 2 | Inferred introduction times and expansion of SECs. a, ML phylogenetic tree of Spanish sequences indicating the identified SECs (inner circle), the Pango lineage (middle circle) and the NextStrain clade (outer circle). **b**, Range of dates for each 'candidate transmission cluster' identified within the SECs, and the most probable origin date (14 days before the first documented case). **c**, Time of the MRCA of each SEC is plotted, including the 95% HPD interval. First collected sample, and whether it is Spanish or non-Spanish, is indicated. **d**, SEC dispersion through the different regions of the country. Some SECs are restricted to one or two regions, while others have expanded through the complete territory.

Superspreading events and mobility led to success of SEC8

Why some genetic variants succeed over others cannot be answered from genomic sequence data alone. We must also consider the epidemiological dynamics in the country. There are data supporting a role for the spike protein mutation 614G in epidemiological success. However, SEC7 and SEC8 do not harbor that mutation, explaining why 614G was less frequent in Spain during the first weeks of the epidemic than in other countries (Extended Data Fig. 6). In addition, analysis of signature positions for both SECs did not lead to any probable genomic determinant of epidemiological success (Supplementary Table 3). Unfortunately, we had no access to linked epidemiological data for the complete dataset. However, we had access to detailed information from two superspreading events linked to SEC8. On the basis of the phylogenetic analysis and the linked epidemiological data, we are able to shed light on the early success of SEC8, the dynamics of which can be explained in three stages.

In the first stage, SEC8 was introduced at least twice from Italy to the city of Valencia (Fig. 3a). There is epidemiological evidence that the individual in both cases became infected in Italy, as they attended the Atalanta-Valencia Champions League football match on 19 February 2020, and that, upon returning to Valencia a few days later, one of them initiated a transmission chain of at least 24 cases according to Public Health sources. Of these 24 cases, 12 were sequenced (Fig. 3a, highlighted in orange). This epidemiological link strongly suggests that the SEC8 genetic variant was imported from Italy. This introduction occurred in agreement with the estimated time of entry of SEC8 into Spain (Supplementary Table 2). NextStrain tracking tools for viral spatial spread suggest additional SEC8-related early seedings in Madrid, País Vasco, Andalucía and La Rioja regions (Supplementary Data 2), which might have involved other countries, not exclusively Italy. Given the lack of genetic differentiation of the virus and scarce epidemiological information, there is no certainty on whether these infections resulted from independent

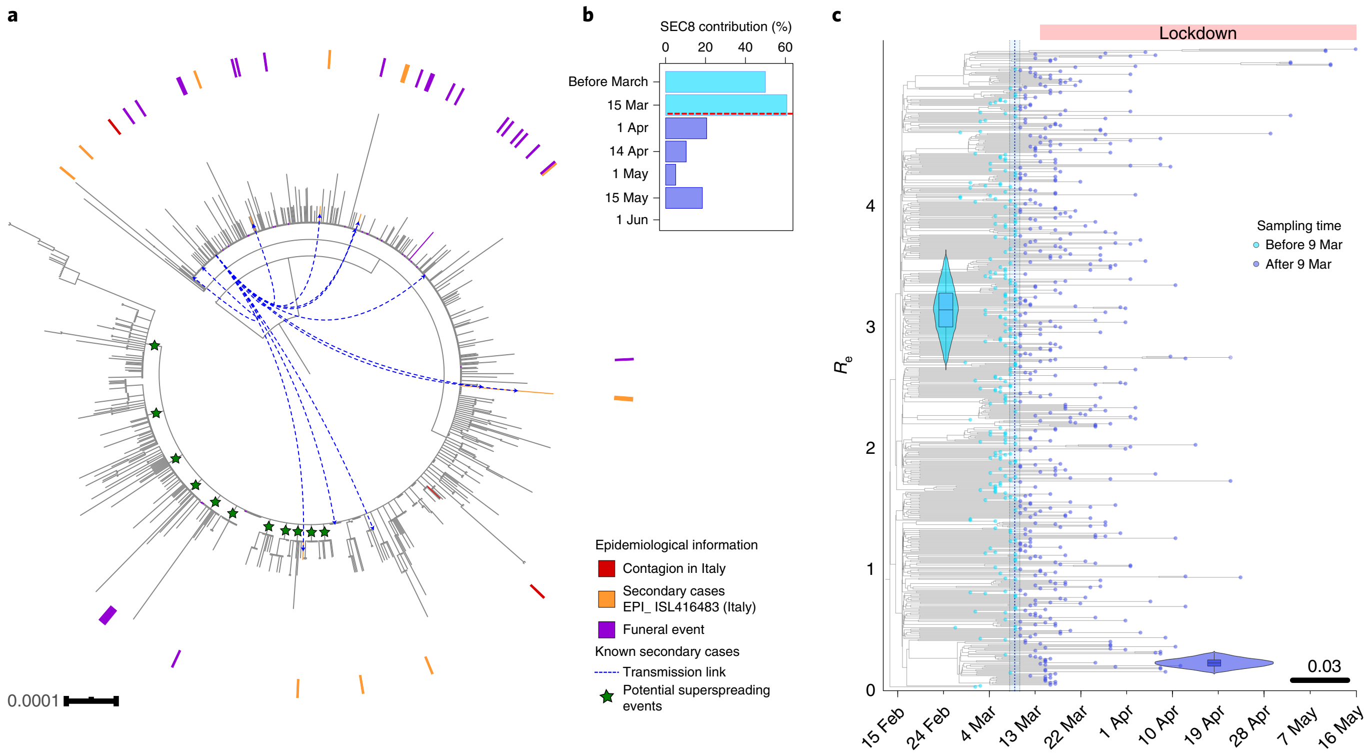


Fig. 3 | SEC8 epidemiological success and impact of mobility restrictions. a, ML phylogeny with all strains of SEC8. Samples with epidemiological evidence about their origin are marked in the tree. In red, cases imported from different events in Italy. In orange, secondary cases originated from one of the cases introduced from Italy (also marked with blue arrows). In purple, cases related to a large funeral in La Rioja. Green stars mark potential superspreading events of more than ten sequences sharing at least one clade-defining SNP. **b**, Contribution of SEC8 to the total of samples sequenced over time. The horizontal red line marks the start of the Spanish lockdown on 14 March 2020. **c**, Phylodynamic estimates of the reproductive number (R_e) of SEC8. The *x* axis represents time, from the origin of the sampled diversity of SEC8 to the date of the last collected genome on 16 May 2020. The blue dotted line shows the posterior value of the timing of the most important change in R_e, around 9 March 2020 [95% HPD: 8–10 March]. The *y* axis represents R_e, and the violin plots show the posterior distribution of this parameter before and after the change time in R_e, with a mean of 3.14 [95% HPD: 2.71–3.57] and 0.23 [95% HPD: 0.15–0.32] before and after the change time, respectively. The phylogenetic tree in the background is a maximum clade credibility tree with the tips colored according to whether they were sampled before or after 9 March 2020. The lower whisker, higher whisker, center and bounds of each boxplot refers to quartile 1–1.5 interquartile range, quartile 3 + 1.5 interquartile range, mean, first and third quartiles of the data. Individual points are outliers (values lower than quartile 1–1.5 interquartile range and higher than quartile 3 + 1.5 interquartile range). Boxplot was constructed with all the Spanish sequences in SEC8 (*N* = 636).

introductions from abroad or from internal migrations of infected persons, although the simultaneous detection in different regions favors the first option. Most of these multiple introductions occurred during the second half of February 2020, a period in which more than 11,000 daily entries of travelers from Italy were recorded.

In a second stage, SEC8 was fueled by superspreading events (Supplementary Data 2). On the basis of the topology of the phylogenetic tree (Fig. 2a), there were multiple clades within SEC8 involving a large number of very closely related sequences (1–3 SNPs) (Fig. 3a). Of special relevance was a funeral on 23 February 2020, with attendees from the País Vasco and La Rioja regions (Public Health officers estimated 800 attendees, resulting in 36 confirmed symptomatic cases) from which 25 samples were sequenced successfully (Fig. 3a, highlighted in purple). Importantly, although they did not differ by more than two SNPs, these sequences are spread across the SEC8 phylogeny, suggesting the existence of many more nonsampled secondary cases across the country (Fig. 3a). In a third stage, SEC8, after reaching high frequencies locally, was redistributed across the country and, in less than 2 weeks, it reached a prevalence of 60% among the sequenced genomes (Fig. 3b), being present in almost every region analyzed. All these stages occurred between the first known diagnosed SEC8 case on 25 February 2020 (Supplementary Table 2) and before the lockdown on 14 March 2020, highlighting the need for very early containment measures to stop the spread of SARS-CoV-2.

Effect of lockdown on the main clades

In the second half of March 2020, Spain imposed a strict lockdown on nonessential services and movements. Consistently, the number of cases for all SECs dropped quickly after the lockdown (Extended Data Fig. 7). A Bayesian birth–death skyline analysis allowed us to evaluate the effect of the lockdown on the effective reproductive number (R_e) of the most successful SECs. The analyses of SEC7 (Extended Data Fig. 8) and SEC8 (Fig. 3c) resulted in similar estimates for R_e before the lockdown (2.10 with 95% highest posterior density (HPD):1.67–2.62 and 3.14 HPD: 2.71–3.57, respectively) similar to the R_e estimated early in the epidemic for SARS-CoV-2 (ref. [21,22]). After the lockdown there was a substantial decrease to less than 0.5 in both cases (0.27 95% HPD: 0.06–0.47; 0.23 HPD: 0.15–0.32, respectively). The model also estimated that the date with highest support for a change in R_e coincides roughly with the start of the lockdown in Spain on 14 March 2020 (20 March HPD: 15–25 March; 9 March HPD: 8–10 March, respectively). In addition, we calculated the doubling time for both SECs[23]. Before the corresponding date of change for R_e, the doubling time for SEC7 was estimated at 6.3 days (95% HPD: 4.3–10.2 days) and that for SEC8 at 3.3 days (95% HPD: 2.7–4.1 days). R_e values after those dates had a posterior distribution that did not include 1.0 for both SECs (Supplementary Note), a result that supports the reduction in the rate of increase of confirmed cases and that is in agreement with estimates from epidemiological models and data[21,22].

In addition, all the viral variants not included in the SECs, mostly harboring the 614G mutation, displayed a similar decrease in the number of cases after the lockdown compared with the most successful SECs (Extended Data Fig. 9a). The impact of the measures implemented in the R_e was also evaluated in two representative Pango lineages carrying the 614G mutation, and a substantial decrease in R_e after the lockdown was observed, from an R_e greater than 1.5 to an R_e of ~0.25, similar to the results obtained for SEC7 and SEC8 (Extended Data Fig. 9b).

Discussion

Our analyses have revealed more than 500 independent introductions of SARS-CoV-2 into Spain between late January, coinciding with the first reported cases in our country[2,24], and mid-April 2020. The earliest entries corresponded to lineage A, matching the virus diversity profile reported for the country. This lineage was common in Asia but rare in the rest of Europe[25]. We observed that two genetic variants (SEC7 and SEC8) of this lineage dominated the first stages of the epidemic wave in Spain, contrary to what was observed in other European countries. In fact, most cases described in Europe at the beginning of the pandemic were lineage B, which makes the situation in Spain more unique. This highlights the importance of epidemiological data, from which we know that SEC8 was introduced at least from Italy despite not being the dominant lineage in the country at that time and illustrating the role of founder effects early in the pandemic[18,26,27].

Reasons why some variants dominate over others can be related to viral genetics, to founder events associated to particular variants and to the implementation of different public health measures over time, not necessarily in an exclusive manner. The variant distribution could also be explained partly by sampling bias. No mutation probably associated with epidemiological success has been identified in our analyses of SEC7 and SEC8 (Supplementary Table 3). In fact, neither SECs carry the 614G mutation in the spike protein, contrary to what is seen in most lineage B variants (Extended Data Fig. 6). The mutation 614G has been associated with increased viral shedding compared to the ancestral 614D variant in laboratory conditions[28] and in transmission studies[29,30]. Consistently, our analysis supports the observation[30] that 614G strains had higher associated viral loads measured as lower cycle threshold (Ct) values (Supplementary Fig. 4). However, one study reports that its actual role in the epidemic is doubtul[31], suggesting that its impact, if any, on epidemic transmission was minor. In the case of Spain, 614G did not account for the initial success of the epidemic because SEC7 and, in particular, SEC8 were much more common than other genetic variants until the lockdown (10% and 30% of cases, respectively). Notably, 614G strains were introduced and expanded later, and closer to the start of lockdown, than 614D (Extended Data Figs. 6 and 9a), explaining the particular lineage structure observed in Spain (Fig. 1c). By contrast, founder events seem to have played an important role for the two main SECs. Our analysis shows that these two SECs were the first variants introduced in the country and, at least in the case of SEC8, were linked to very early superspreading events that contributed to their success. However, an early introduction of lineage A variants also occurred in other European countries, but they did not take hold and were displaced sooner by lineage B. Despite the early adoption of strong nonpharmaceutical intervention measures, we hypothesize that epidemic control in the first wave in Spain was soon overwhelmed as compared with countries that controlled early outbreaks[15]. This was probably associated with a strict implementation of the case definition by the WHO, which allowed for a stealth dispersion of the first introductions. Spain implemented one of the strictest lockdowns in Europe, with a high compliance from the population as tracked by mobility data[32]. The efficacy of nonpharmaceutical intervention measures was evident a few weeks later, and it was reflected in the almost complete elimination of SEC7, SEC8, and most variants by the end of the first wave (Extended Data Figs. 7 and 9a). However, we do observe a replacement of lineage A (SEC7 and SEC8) just after the start of the lockdown by B.1 variants harboring D614G (Extended Data Fig. 9a). Contrary to that of lineage A, the spread of B.1 variants in Spain was represented by smaller SECs and more isolated cases. It is probable that these smaller clusters correspond to a new stage in the epidemic at the national level characterized by more limited mobility and social interactions.

This study has several limitations. Even though Spain is one of the countries with high contribution to public repositories, our dataset represents only a small subset of confirmed cases that occurred in the first COVID-19 wave (1% of cases). Moreover, sampling across the country was heterogeneous and the representation of each region in the dataset was not always proportional to the incidence during the studied period. Lack of genome data from countries with high disease burden, especially at the beginning of the pandemic, may have prevented a reliable identification of their probable sources based only on viral genome sequences. In addition, we did not have access to individual patient data for most cases. These caveats could have an impact on, for example, an exact quantification on the number of introductions, which will be always an estimate. However, our analysis already gives clues about the role of multiple introductions in the early days of the epidemic. Despite these limitations, we have been able to investigate some of the key cases and events that initiated the epidemic in Spain. This allowed us to understand the origin and early spread of SEC8, which would not have been possible based only on genome data. But we have also shown that genetic data can be used to accurately estimate relevant epidemiological parameters, such as R_e and doubling times, even when the proportion of sampling is low.

We believe that our results allow us to draw lessons for the control of this, as well as future, pandemics. First, we have shown how specific variants can be used to track the effectiveness of public health control measures. In February 2020, the number of SEC8 cases was just a few dozen and yet it ended up accounting for 60% of the sequenced samples in the first weeks of March 2020. Second, the closure of borders to countries with high incidence is relevant to reduce simultaneous and multiple imports of the virus, but the efficacy of these restrictions also depends on the internal incidence of the disease[33]. The most successful SECs during the first wave were probably those that arrived early, several times, and to diverse locations. Thus, as suggested elsewhere[36], founder effects are important for the success of certain variants. Third, SEC7 and SEC8 spread across Spain in a matter of days. Controlling mobility is essential when the level of community transmission is high, as demonstrated by the important decrease in R_e for these high-transmission genetic variants after the lockdown. As a comparison, before the lockdown, R_e values were 50% higher in Spain (3.3 for SEC8) than in Australia (1.63), and they underwent a reduction down to 7% of the original value (0.23) as a result of the containment measures, compared to a reduction to 30% (0.48) in Australia[17]. From a public health perspective, our results add to the evidence that the success of specific genetic variants, with no intrinsic biological difference, is fueled by superspreading events that rapidly increase the prevalence of the virus[34]. Subsequently, coupled to the high mobility of our connected world, a variant may end up dominating the epidemic in a particular geographic location. This is what occurred with SEC8 and what, at a local level, has been described in Boston[35]. In fact, we have recently described a new variant in Europe, the prevalence of which was growing rapidly in several countries during the summer of 2020, which is also linked to initial superspreading events[36]. By contrast, new variants with different transmissibility and immunogenicity profiles started to merge at the end of 2020. Some of these variants, such as B.1.1.7 (alpha) were able to replace existing variants in a matter of months. For the first wave in Spain, the conclusion is that early diagnosis and notification of cases would have helped timely implementation of effective contact tracing that, coupled with earlier mobility closures and maybe tighter border control, could probably have delayed by a few days the expansion of genetic variants, such SEC8, during the early stages of the epidemic. Whether this might have changed the global shape of the epidemic in the country or whether other genetic variants would have performed this role instead, leading to a similar outcome, cannot be ascertained. The comparison with other countries leads us to suspect that the difference would have been minimal.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00936-6.

Received: 22 December 2020; Accepted: 11 August 2021;
Published online: 30 September 2021

References

1. Zhu, N. et al. Brief report: a novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727 (2020).
2. Spiteri, G. et al. First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. *Eurosurveillance* **25**, PMC7068164 (2020).
3. Centro de Coordinación de Alertas y Emergencias Sanitarias. *Enfermedad por el coronavirus, situación en España* https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Actualizacion_153_COVID-19.pdf (2020).
4. Pollán, M. et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *Lancet* **396**, 535–544 (2020).
5. Ministerio de la Presidencia, R. C. las C. y. M. D. BOE-A-2020-3692 https://www.boe.es/eli/es/rd/2020/03/14/463/con (2020).
6. Instituto de Salud Carlos III, Spanish Government. COVID-19 reported cases in Spain https://cnecovid.isciii.es/covid19/#ccaa (2020).
7. Centro de Coordinación de Alertas y Emergencias Sanitarias, Ministerio de Sanidad. Enfermedad por el coronavirus (COVID-19) https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Actualizacion_268_COVID-19.pdf (2020).
8. Grenfell, B. T. et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
9. Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J. & Frost, S. D. W. Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–1430 (2009).
10. Vasylyeva, T. I. et al. Phylodynamics helps to evaluate the impact of an HIV prevention intervention. *Viruses* **12**, 469 (2020).
11. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, 30494 (2017).
12. Callaway, E. The coronavirus is mutating – does it matter? *Nature* **585**, 174–177 (2020).
13. Kupferschmidt, K. The pandemic virus is slowly mutating. But does it matter? *Science* **369**, 238–239 (2020).
14. Morel, B. et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* **38**, 1777–1791 (2020).
15. Worobey, M. et al. The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
16. Geoghegan, J. L. et al. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat. Commun.* **11**, 6351 (2020).
17. Seemann, T. et al. Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* **11**, 4376 (2020).
18. Alm, E. et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance* **25**, 2001410 (2020).
19. Oude Munnink, B. B. et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).
20. Candido, D. S. et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
21. Guirao, A. The Covid-19 outbreak in Spain. A simple dynamics model, some lessons, and a theoretical framework for control response. *Infect. Dis. Model.* **5**, 652 (2020).
22. Hyafil, A. & Moriña, D. Analysis of the impact of lockdown on the reproduction number of the SARS-Cov-2 in Spain. *Gac. Sanit.* **35**, 453–458 (2021).
23. Lurie, M. N., Silva, J., Yorlets, R. R., Tao, J. & Chan, P. A. Coronavirus disease 2019 epidemic doubling time in the United States before and during stay-at-home restrictions. *J. Infect. Dis.* **222**, 1601–1606 (2020).
24. Böhmer, M. M. et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. *Lancet Infect. Dis.* **20**, 920–928 (2020).
25. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
26. Lai, A. et al. Molecular tracing of SARS-CoV-2 in Italy in the first three months of the epidemic.*Viruses* **12**, 798 (2020).
27. Hodcroft, E. & Neher, R. Phylogenetic analysis of SARS-CoV-2 diversity in Europe (Italy) https://nextstrain.org/groups/neherlab/ncov/italy (2021).
28. Plante, JA. et al. Spike mutation D614G alters SARS-CoV-2 fitness.*Nature* **592**, 116–121 (2021).
29. Volz, EM. et al. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75.e11 (2021).
30. Korber, B. et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827 (2020).
31. van Dorp, L. et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
32. COVID-19 community mobility reports. *Google* https://www.google.com/covid19/mobility/ (2021).
33. Russell, TW. et al. Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study.*Lancet Public Health* **6**, E12–E20 (2021).
34. Popa, A. et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2.*Sci. Transl. Med.* **12**, eabe2555 (2020).
35. Lemieux, J. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, 6529 (2021).
36. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
37. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* **2**, e000093 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Instituto de Biomedicina de Valencia (IBV-CSIC), Valencia, Spain. [2]Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain. [3]Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain. [4]CIBER Enfermedades Respiratorias (CIBERES), Bunyola, Spain. [5]Instituto de Biología Integrativa de Sistemas, I2SysBio (CSIC-Universitat de València), Valencia, Spain. [6]Joint Research Unit Infection and Public Health FISABIO-University of Valencia I2SysBio, Valencia, Spain. [7]Ciber en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [8]FISABIO, Servicio de Secuenciación, València, Spain. [9]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland. [10]Servicio de Microbiología. Hospital Dr Peset, Valencia, Spain. [11]Conselleria de Sanitat i Consum, Generalitat Valenciana, Valencia, Spain. [12]Departamento Microbiología, Facultad de Medicina, Universitat de València, Valencia, Spain. [13]Microbiology Service, Hospital Clínico Universitario, INCLIVA Research Institute, Valencia, Spain. [14]Servicio de Microbiología, Hospital Universitario Cruces, Bilbao, Spain. [15]Grupo de Microbiología y Control de Infección, Instituto de Investigación Sanitaria Biocruces Bizkaia, Barakaldo, Spain. [16]Plataforma de Genómica y Bioinformática, Centro de Investigación Biomédica de La Rioja (CIBIR), Logroño, Spain. [17]Servicio de Microbiología, Hospital Universitario Central de

Asturias, Oviedo, Spain. [18]Grupo de Microbiología Traslacional, Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Asturias, Spain. [19]Servicio de Microbiología, Laboratori Clínic Metropolitana Nord, Hospital Universitari Germans Trias i Pujol, Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol (IGTP), Badalona, Barcelona, Spain. [20]Servicio de Microbiología, Hospital Universitario de Álava, Osakidetza-Servicio Vasco de Salud, Vitoria-Gasteiz (Álava), Spain. [21]Animal Health Department, Universidad de León, León, Spain. [22]Servicio de MicrobiologíaBiodonostia, Osakidetza, Hospital Universitario Donostia, San Sebastián, Spain. [23]Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain. [24]Servicio de Microbiologia, Hospital Regional Universitario de Málaga, Málaga, Spain. [25]Servicio de Enfermedades Infecciosas y Microbiología clínica, Clínica Universidad de Navarra, Pamplona, Spain. [26]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia. [27]Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, O.A., M.P. – INIA, Madrid, Spain. [28]Servicio de Microbiología, Consorcio Hospital General Universitario de Valencia, Valencia, Spain. [29]Servicio de Microbiología UCEIMP, Hospital Universitario Virgen del Rocío, Sevilla, Spain. [30]Servicio Microbiología, Departamento de Salud de Elche-Hospital General, Elche, Alicante, Spain. [31]Servicio de Microbiología, Hospital Universitario Son Espases, Palma de Mallorca, Spain. [32]Hospital Universitario y Politécnico La Fe, Servicio de Microbiología, Valencia, Spain. [33]Research Group on Gene-Environment Interactions and Health. Institute of Biomedicine (IBIOMED), Universidad de León, León, Spain. [34]Servicio de Microbiología Clínica, Hospital Universitario Miguel Servet, Zaragoza, Spain. [35]Hospital General Universitario de Castellón, Castellón, Spain. [36]Servicio de Microbiología, Hospital Universitario Virgen de las Nieves, Granada, Spain. [37]Hospital Universitario Virgen de las Nieves, Instituto de Investigación Biosanitaria ibs, Granada, Spain. [38]Instituto de Investigación Sanitaria de las Islas Baleares, Palma, Spain. [39]Laboratorio de Microbiología, Hospital Marina Baixa, Villajoyosa, Spain. [40]Hospital Universitario de Gran Canaria Dr. Negrin, Las Palmas de Gran Canaria, Spain. [41]Servei de Microbiologia, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. [42]CREPIMC, Institut d'Investigació Biomèdica Sant Pau, Barcelona, Spain. [43]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Cerdanyola, Spain. [44]Instituto de Investigación Sanitaria de Aragón, Centro de Investigación Biomédica de Aragón (CIBA), Zaragoza, Spain. [45]Facultad de Medicina, Universidad de Zaragoza, Zaragoza, Spain. [46]Hospital General Universitario de Albacete, Albacete, Spain. [47]Servicio de Enfermedades Infecciosas, Hospital Clínic de Barcelona, Barcelona, Spain. [111]These authors contributed equally: Mariana G. López, Álvaro Chiner-Oms. *A full list of members and affiliations appears at the end of the paper.
✉e-mail: mireia.coscolla@uv.es; fernando.gonzalez@uv.es; icomas@ibv.csic.es

SeqCOVID-Spain consortium

Álvaro Chiner-Oms[1,111], Irving Cancino-Muñoz[1], Mariana G. López[1,111], Manuela Torres-Puente[1], Inmaculada Gómez-Navarro[1], Santiago Jiménez-Serrano[1], Jordi Pérez-Tur[1], Darío García de Viedma[2,3,4], Laura Pérez-Lago[2,3], Marta Herranz[2,3,4], Jon Sicilia[2,3], Pilar Catalán-Alonso[2,3,4], Julia Suárez González[3], Patricia Muñoz[2,3,4], Mireia Coscolla[5], Paula Ruiz-Rodríguez[5], Fernando González-Candelas[6,7], Iñaki Comas[1,7], Lidia Ruiz-Roldán[6], María Alma Bracho[6,7], Neris García-González[6], Llúcia Martínez Priego[8], Inmaculada Galán-Vendrell[8], Paula Ruiz-Hueso[8], Griselda De Marco[8], María Loreto Ferrús-Abad[8], Sandra Carbó-Ramírez[8], Giuseppe D'Auria[7,8], Galo Adrian Goig[9], Juan Alberola[10,11,12], Jose Miguel Nogueira[10,11,12], Juan José Camarena[10,11,12], David Navarro[12,13], Eliseo Albert[13], Ignacio Torres[13], Maitane Aranzamendi Zaldumbide[14,15], Óscar Martínez Expósito[14,15], Nerea Antona Urieta[14,15], María de Toro[16], María Pilar Bea-Escudero[16], Jose Antonio Boga[17,18], Cristian Castelló-Abietar[17,18], Susana Rojo-Alba[17,18], Marta Elena Álvarez-Argüelles[17,18], Santiago Melón[17,18], Elisa Martró[7,19], Antoni E. Bordoy[19], Anna Not[19], Adrián Antuori[19], Anabel Fernández-Navarro[19], Andrés Canut-Blasco[20], Silvia Hernáez Crespo[20], Maria Luz Cordón Rodríguez[20], Maria Concepción Lecaroz Agara[20], Carmen Gómez-González[20], Amaia Aguirre-Quiñonero[20], José Israel López-Mirones[20], Marina Fernández-Torres[20], Maria Rosario Almela-Ferrer[20], Ana Carvajal[21], Juan Miguel Fregeneda-Grandes[21], Héctor Argüello[21], Gustavo Cilla Eguiluz[22], Milagrosa Montes Ros[22], Luis Piñeiro Vázquez[22], Ane Sorarrain[22], José María Marimón[22], José J. Costa-Alcalde[23], Rocío Trastoy[23], Gema Barbeito Castiñeiras[23], Amparo Coira[23], María Luisa Pérez del Molino[23], Antonio Aguilera[23], Begoña Palop-Borrás[24], Inmaculada de Toro Peinado[24], Maria Concepción Mediavilla Gradolph[24], Mercedes Pérez-Ruiz[24], Mirian Fernández-Alonso[25], Jose Luis del Pozo[25], Oscar González-Recio[27], Mónica Gutiérrez-Rivas[27], Jovita Fernández-Pinero[27], Miguel Ángel Jiménez Clavero[27], Begoña Fuster Escrivá[12,28], Concepción Gimeno Cardona[28], María Dolores Ocete Mochón[28], Rafael Medina-Gonzalez[28], José Antonio Lepe[29], Verónica González Galán[29], Ángel Rodríguez-Villodres[29], Nieves Gonzalo Jiménez[30], Jordi Reina[31], Carla López-Causapé[31], Maria Dolores Gómez-Ruiz[32],

Eva M. Gonzalez-Barbera[32], José Luis López-Hontangas[32], Vicente Martín[7,33], Antonio J. Molina[33], Tania Fernandez-Villa[33], Ana Milagro Beamonte[34], Nieves Felisa Martínez-Cameo[34], Yolanda Gracia-Grataloup[34], Rosario Moreno-Muñoz[35], Maria Dolores Tirado Balaguer[35], José María Navarro-Marí[36,37], Irene Pedrosa-Corral[36,37], Sara Sanbonmatsu-Gámez[36,37], Antonio Oliver[31,38], Mónica Parra Grande[39], Bárbara Gómez Alonso[39], Francisco José Arjona Zaragozí[39], Maria Carmen Pérez González[40], Francisco Javier Chamizo López[40], Ana Bordes-Benítez[40], Núria Rabella[41,42,43], Ferran Navarro[41,42,43], Elisenda Miró[41,42], Antonio Rezusta[34,44,45], Alexander Tristancho[34,44], Encarnación Simarro Córdoba[46], Julia Lozano-Serra[46], Lorena Robles Fonseca[46], Álex Soriano[47], Francisco Javier Roig Sena[48], Hermelinda Vanaclocha Luna[48], Isabel Sanmartín[49], Daniel García-Souto[50,51,52], Ana Pequeño-Valtierra[50], Jose M. C. Tubio[50,51], Javier Temes[50,51], Jorge Rodríguez-Castro[50], Martín Santamarina García[50], Manuel Rodríguez-Iglesias[53,54,55], Fátima Galán-Sanchez[53,54,55], Salud Rodríguez-Pallares[53,54], José Manuel Azcona-Gutiérrez[56], Miriam Blasco-Alberdi[56], Alfredo Mayor[7,57,58], Alberto L. García-Basteiro[57,58], Gemma Moncunill[57], Carlota Dobaño[57], Pau Cisteró[57], Oriol Mitjà[59,60], Camila González-Beiras[59], Martí Vall-Mayans[59], Marc Corbacho-Monné[59], Andrea Alemany[59], Cristina Muñoz-Cuevas[61,62], Guadalupe Rodríguez-Rodríguez[61,62], Rafael Benito[45,63], Sonia Algarate[63], Jessica Bueno[63], Andrea Vergara-Gómez[64], Miguel J. Martínez[57,65,66], Jordi Vila[57,65], Elisa Rubio[57,65], Aida Peiró-Mestres[57,65], Jessica Navero-Castillejos[57,65], David Posada[67,68,69], Diana Valverde[67,68,69], Nuria Estévez[67], Iria Fernández-Silva[67,68], Loretta de Chiara[67,68], Pilar Gallego-García[67], Nair Varela[67], Ulises Gómez-Pinedo[70], Mónica Gozalo-Margüello[71], Maria Eliecer Cano García[71], José Manuel Méndez-Legaza[71], Jesus Rodríguez-Lozano[71], María Siller[71], Daniel Pablo-Marcos[71], Maria Montserrat Ruiz-García[30,72], Antonio Galiana[73], Judith Sánchez-Almendro[73], Maria Isabel Gascón Ros[74], Cristina Juana Torregrosa-Hetland[74], Eva María Pastor Boix[74], Paloma Cascales Ramos[74], Pedro Luis Garcinuño Enríquez[74], Salvador Raga Borja[74], Julia González Cantó[75], Olalla Martínez Macias[75], Adolfo de Salazar[76], Laura Viñuela González[76], Natalia Chueca[76], Federico García[76], Cristina Gómez-Camarasa[76], Amparo Farga Martí[77], Rocío Falcón[77], Victoria Domínguez-Márquez[77], Anna M. Planas[78], Israel Fernández-Cádenas[79], Maria Ángeles Marcos[80], Carmen Ezpeleta[81,82], Ana Navascués[81,82], Ana Miqueleiz Zapatero[81], Manuel Segovia[83,84], Antonio Moreno-Docón[83,84], Esther Viedma[85], Raúl Recio Martínez[85], Irene Muñoz-Gallego[85], Sara Gonzalez-Bodi[85], Maria Dolores Folgueira[85], Jesús Mingorance[86], Elias Dahdouh[86], Fernando Lázaro-Perona[86], María Rodríguez-Tejedor[86], María Pilar Romero-Gómez[86], Julio García-Rodríguez[86], Juan Carlos Galán[87], Mario Rodríguez-Dominguez[7,87,88], Laura Martínez-García[7,87,88], Melanie Abreu Di Berardino[87,88], Manuel Ponce-Alonso[87,88,89], Jose Maria González-Alba[87,88], Ivan Sanz-Muñoz[90], Diana Pérez San José[90], Maria Gil Fortuño[91], Juan B. Bellido-Blasco[91], Alberto Yagüe Muñoz[91], Noelia Hernández Pérez[91], Helena Buj Jordá[91], Óscar Pérez Olaso[91], Alejandro González Praetorius[92], Nora Mariela Martínez Ramírez[92], Aida Ramírez Marinero[93], Eduardo Padilla León[93], Alba Vilas Basil[93], Mireia Canal Aranda[93], Albert Bernet Sánchez[94], Alba Bellés Bellés[94], Eric López González[94], Iván Prats Sánchez[94], Mercè García-González[94], Miguel José Martínez-Lirola[95], Manuel Ángel Rodríguez Maresca[95], Maria Teresa Cabezas Fernández[95], María Eugenia Carrillo Gil[95], Maria Paz Ventero Martín[96], Carmen Molina Pardines[96], Nieves Orta Mira[97], María Navarro Cots[97], Inmaculada Vidal Catalá[97], Isabel García Nava[97], Soledad Illescas Fernández-Bermejo[98,99], José Martínez-Alarcón[98,99],

Marta Torres-Narbona[98], Cristina Colmenarejo[98], Lidia García-Agudo[98], Jorge A. Pérez García[98], Martín Yago López[100], María Ángeles Goberna Bravo[100], Victoria Simón García[101], Gonzalo Llop Furquet[101], Agustín Iranzo Tatay[101], Sandra Moreno-Marro[101], Noelia Lozano Rodríguez[101], Amparo Broseta Tamarit[102], Juan José Badiola Díez[103], Amparo Martínez-Ramírez[104], Ana Dopazo[105], Sergio Callejas[105], Alberto Benguría[105], Begoña Aguado[106], Antonio Alcamí[106], Marta Bermejo Bermejo[107], Ricardo Ramos-Ruíz[108], Víctor Manuel Fernández Soria[108], Fernando Simón Soria[109] and Mercedes Roig Cardells[110]

[48]Servicio de Vigilancia y Control Epidemiológico. Dirección General de Salud Pública y Adicciones, Conselleria de Sanitat Universal i Salut Pública. Generalitat Valenciana, Valencia, Spain. [49]Real Jardín Botánico, Consejo Superior de Investigaciones Científicas, Madrid, Spain. [50]Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [51]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [52]Cancer Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Cambridge, UK. [53]Servicio de Microbiología, H.U. Puerta del Mar, Cádiz, Spain. [54]INIBICA, Instituto de Investigación Biomédica de Cádiz, Cádiz, Spain. [55]Departamento de Biomedicina, Biotecnología y Salud Pública. Facultad de Medicina, Universidad de Cádiz, Cádiz, Spain. [56]Laboratorio de Microbiología, Hospital San Pedro, Logroño, Spain. [57]ISGlobal, Barcelona Institute for Global Health, Hospital Clínic – Universitat de Barcelona, Barcelona, Spain. [58]Centro de Investigação em Saúde de Manhiça (CISM), Maputo, Mozambique. [59]Fight AIDS and Infectious Diseases Foundation, Hospital Germans Trias i Pujol, Barcelona, Spain. [60]Lihir Medical Centre-InternationalSOS, Lihir Island, Papua New Guinea. [61]Servicio de Microbiología Clínica, Hospital San Pedro de Alcántara, Cáceres, Spain. [62]Servicio Extremeño de Salud, Sevilla, Spain. [63]Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain. [64]Servicio de Microbiología & CORE de Biología Molecular, CDB, Hospital Clínic, Barcelona, Spain. [65]Department of Microbiology – CDB, Hospital Clínic de Barcelona, Barcelona, Spain. [66]University of Barcelona, Barcelona, Spain. [67]CINBIO, Universidade de Vigo, Vigo, Spain. [68]Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, Vigo, Spain. [69]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain. [70]IdISSC/Hospital Clínico San Carlos, Madrid, Spain. [71]Hospital Marqués de Valdecilla – IDIVAL, Santander, Spain. [72]Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Elche, Spain. [73]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana: Elche, Alicante, ES, (Hospital General Universitario de Elche, Microbiologia), Elche, Spain. [74]Laboratorio de Microbiología, Hospital General Universitario de Elda. Elda, Alicante, Spain. [75]Laboratorio Biología Molecular, Área de Diagnóstico Biológico, Hospital Universitario La Ribera, Alzira, Valencia, Spain. [76]Hospital Universitario San Cecilio, Granada, Spain. [77]Servicio de Microbiología, Hospital Arnau de Vilanova, Valencia, Spain. [78]Biomedical Research Institute of Barcelona (IIBB), Spanish National Research Council (CSIC), Barcelona, Spain. [79]Sant Pau Hospital Research Institute, Barcelona, Spain. [80]Microbiology Department, Hospital Clinic I Provincial de Barcelona, Institut of Global Health of Barcelona (ISGlobal), Barcelona, Spain. [81]Servicio de Microbiología Clínica, Complejo Hospitalario de Navarra (Pamplona, Navarra), Pamplona, Spain. [82]Instituto de Investigación Sanitaria de Navarra (IdiSNA), Pamplona, Spain. [83]Servicio de Microbiología, Hospital Clínico Universitario Virgen de la Arrixaca, El Palmar, Spain. [84]Departamento de Genética y Microbiología, Universidad de Murcia, Carretera Madrid-Cartagena sn, El Palmar, Murcia, Spain. [85]Hospital Universitario 12 de Octubre, Madrid, Spain. [86]Hospital Universitario La Paz, IdiPAZ, Madrid, Spain. [87]Servicio de Microbiología, Hospital Universitario Ramón y Cajal, Madrid, Spain. [88]Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain. [89]Red Española de Investigación en Patología Infecciosa (REIPI), Madrid, Spain. [90]Centro Nacional de Gripe, Valladolid, Spain. [91]Hospital Universitari de La Plana, Vila-Real, Spain. [92]Hospital Universitario de Guadalajara, Guadalajara, Spain. [93]Laboratori de Referència de Catalunya, Barcelona, Spain. [94]Hospital Universitari Arnau de Vilanova de Lleida, Lleida, Spain. [95]Complejo Hospitalario Universitario Torrecárdenas, Almería, Spain. [96]Servicio de Microbiología, Hospital General Universitario de Alicante. Instituto de Investigación Sanitaria y Biomédica de Alicante (ISABIAL), Alicante, Spain. [97]Hospital Francesc de Borja, Sección Microbiología, Valladolid, Spain. [98]Hospital General Universitario de Ciudad Real, Ciudad Real, Spain. [99]Facultad de Medicina de Ciudad Real. UCLM, Ciudad Real, Spain. [100]Hospital General de Requena, Valencia, Spain. [101]Laboratorio de Biología Molecular, Servicio de Análisis Clínicos y Microbiología, Hospital de Sagunto, Valencia, Spain. [102]Synlab - Hospital de Manises, Valencia, Spain. [103]Centro de Encefalopatías y Enfermedades Transmisibles Emergentes, Facultad de Veterinaria, Universidad de Zaragoza, Zaragoza, Spain. [104]Sección de Genómica Servei Central de Suport a la Investigació Experimental (SCSIE), Universitat de València, València, Spain. [105]Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain. [106]Centro de Biología Molecular Severo Ochoa (CBMSO) (CSIC-UAM), Nicolás Cabrera 1, Cantoblanco, Madrid, Spain. [107]Unidad de Vigilancia de Salud y Medicina del Trabajo CSIC, Serrano 113p, Madrid, Spain. [108]Fundación Parque Científico de Madrid, Madrid, Spain. [109]Centro de Coordinación de Alertas y Emergencias Sanitarias, Ministerio de Sanidad, Madrid, Spain. [110]Laboratorio de microbiología del Hospital Comarcal de Vinaròs, Vinaròs, Spain.

Methods

SeqCOVID sampling and sequencing. RNA samples were received from different hospitals, and confirmed as SARS-Cov-2-positive by PCR with reverse transcription (RT-PCR) by Microbiological Services. Samples consisted of the remaining RNA extracts from naso- and oropharyngeal clinical specimens used for diagnosis. The use of such samples has been approved by the ethics committee Comité Ético de Investigación de Salud Pública y Centro Superior de Investigación en Salud Pública (CEI DGSP-CSISP) N° 20200414/05.

In general, we applied the following criteria for selecting the samples that underwent sequencing: (1) only one sample per patient, (2) diagnostic PCR should have a *C*t under 30, (3) prioritize samples from poorly sampled regions and hospitals to maximize geographic diversity, (4) prioritize samples according to their diagnosis date, to maximize sampling of high incidence periods. These criteria were adopted weeks before the beginning of the project, after analyzing the firsts sets of sequences, so, in the initial weeks, we did not preselect the samples for sequencing.

In the SeqCOVID consortium webpage (http://seqcovid.csic.es/), the number of samples received, sequenced and uploaded to public repositories are shown and updated periodically.

RNA was retrotranscribed into cDNA and SARS-CoV-2 complete genome amplification was conducted in two multiplex PCRs, accordingly to an openly available protocol developed by the ARTIC network[38] using the V3 multiplex primers scheme[39]. Two resulting amplicon pools were combined and used for library preparation. Genomic libraries were constructed with the Nextera DNA Flex Sample Preparation kit (Illumina Inc.) according to the manufacturer's protocol, with five cycles for indexing PCR. Whole genome sequencing was carried out in the MiSeq platform (2 × 200 cycles paired-end run; Illumina).

The sequences obtained went through a bioinformatic pipeline based on IVAR[40], which is open source and can be accessed at https://gitlab.com/fisabio-ngs/sars-cov2-mapping. In short, the pipeline goes through the following steps: (1) Removal of the human reads with Kraken[41]; (2) filtering of the fastq files using fastp v.0.20.1 (ref. [42]) (arguments:–cut tail,–cut-window-size,–cut-mean-quality, -max_len1,-max_len2); (3) mapping and variant calling using bwa and IVAR v.1.2 (variant calling cut-offs: minimum quality for SNP calling = 20, minimum frequency to call a SNP = 0.05, minimum depth for calling a SNP = 20, consensus construction cut-offs: minimum quality for consensus calling = 20, minimum frequency to consider fixed a SNP = 0.8, minimum position depth = 30 (ambiguous base otherwise)) and (4) quality control assessment with MultiQC[43].

Global alignment and phylogenetic reconstruction. To build the global alignment, we downloaded and concatenated all non-Spanish sequences present in GISAID[11] on 21 June 2020 that passed strict filtering criteria: (1) sequences should be more than 29,000 bp in length, (2) verified insertions/deletions and (3) less than 1% of Ns and less than 0.05% of unique amino acid mutations (compared with other sequences in GISAID).

Later, we added all Spanish sequences deposited in GISAID up to 29 July 2020. The final alignment constructed included 32,914 sequences. The accession numbers of the sequences used in this study can be found in Supplementary Table 1.

Sequences were aligned against the SARS-CoV-2 reference genome[44] using MAFFT[45]. Specific positions that have been reported to be problematic for phylogenetic reconstruction[46] were masked, following the procedure described by Lanfear[47], using the mask_alignment.sh script.

Finally, a maximum-likelihood (ML) phylogeny was reconstructed using IQTREE[48] with the GTR model and based on the complete masked genome alignment. This phylogeny was rooted to the SARS-CoV-2 sequence obtained in Wuhan on 24 December 2019 (GISAID ID: EPI_ISL_402123).

Identification of introductions and transmission clusters. We identified transmission groups between Spanish sequences by inspecting the global phylogeny (32,914 leaves) and searching for Spanish sequences (or groups of) that were embedded within sequences with other geographic origins. Given the general low diversity among sequences, most phylogenetic nodes ended up being polytomic in the ML tree. Because of this, we defined three different transmission scenarios: (1) strains that represent introductions in Spain but differ from those from other countries and form well defined transmission groups ('candidate transmission clusters'); (2) strains introduced into Spain that are equal to other Spanish sequences and that are also equal to sequences from other countries ('zero-distance clusters') and (3) Spanish sequences found within groups of sequences from other countries and which are not phylogenetically near any other Spanish sequences ('unique'). The 'candidate transmission clusters' were identified as monophyletic groups of sequences composed exclusively by Spanish sequences in the phylogeny. The 'zero-distance clusters' were identified as Spanish sequences that share a common ancestor and that are at zero SNP distance from each other. Finally, the 'unique' sequences were identified as those sequences that do not share their most recent ancestor with any other Spanish sequence.

Next, we inferred how many of these transmission groups have a potential contagion date for their first case that predates the start of mobility restrictions, on 14 March 2020, by subtracting 14 days from the diagnosis date.

Finally, we wanted to investigate the international origin of these introductions. For each of the identified groups or 'unique' sequences with an inferred contagion date before 14 March 2020, we looked for the closest non-Spanish sequence in the phylogeny with a diagnosis date predating the first case of the transmission group. As the current consensus is that the pandemic began in Asia and later it moved to Europe, we considered only those sequences with an Asian or European origin as potential sources of introductions.

SEC alignment and phylogeny. Using the global phylogeny, we identified nodes that had at least 20 leaves and in which at least 50% of these correspond to Spanish sequences. Next, for each of these nodes or clades we reconstructed an alignment of the complete masked genomes including: (1) the sequences that belong to the identified clade; (2) 11 basal sequences from Wuhan acting as an 'anchor' for the phylogeny (Supplementary Table 1) and (3) a subset of 51 representative sequences, each one from a different pangolin lineage, selected to maximize the global SARS-CoV-2 genetic diversity (Supplementary Table 1, downloaded from GISAID on 20 July 2020).

For each of these alignments we inferred a ML phylogeny, using IQTREE[48], with the model GTR, 1,000 fast-bootstrap replicates and rooted to the Wuhan sequence (EPI_ISL_402123). Then, in the resulting phylogeny, we identified less inclusive nodes embedded within the above identified clades and had a bootstrap support value > 80. These clades were named as potential SECs. The iTOL tool[49] was used for phylogenetic visualization.

SEC8 detailed analysis. To get more detail on SEC8 phylogenetic structure, and to evaluate if mobility restrictions were effective to hinder SEC8 transmission, we enriched the original SEC8 phylogenetic tree with all the isolates of this clade sampled from February to October 2020 by the SeqCOVID consortium (959 sequences in total). Later, epidemiological information was included and plotted in the tree using the iTOL tool[49].

SEC8 potential superspreading events were defined as groups of more than ten sequences, having at least one SNP in common and having a within-sequence median distance from one to three SNPs.

Population genetics and differentiation geography. Geographic distance between sequences were computed using the GPS coordinates of the patient residence city and applying the Vicenty (ellipsoid) method. Genetic diversity was calculated with two different methods: (1) genetic distance between each pair of samples in number of substitutions (SNPs), and (2) number of base substitutions per site averaged over all sequence pairs in a group of sequences. Both values have been estimated using the MEGA software[50], skipping one position when a gap is found in the two compared sequences.

Demographic data for all Spanish regions and municipalities were downloaded from INE (https://www.ine.es/), and had been updated on 1 January 2020.

The genetic diversity heatmap of the Comunidad Valenciana autonomous region was generated with QGIS v.3.14.16-Pi[51], using the inverse distance weighting (IDW) algorithm to interpolate the mean genetic diversity of each municipality for which we had at least two sequences.

To compare the genetic and geographic distance distribution between the different SECs, we used a one-way ANOVA test, followed by multiple pairwise-comparisons of the between-groups mean with a Tukey HSD (honestly significant difference) test.

Dating analyses. To estimate the most recent common ancestor (MRCA) of each of the nine SECs defined above, a multisequence alignment was performed including the 11 samples belonging to basal phylogenetic clades and the 51 representative sequences from different lineages (Supplementary Table 1). Before phylogenetic dating, root-to-tip regression of genetic divergence against sampling dates was performed to investigate the molecular clock signal of SECs using TempEst v.1.5.3 (ref. [52]). We implemented a coalescent Bayesian exponential growth model available in Beast 2.6 (ref. [53]) with the HKY + Γ model of nucleotide substitution. Tree priors were defined as follows: for effective population size we used a lognormal distribution (mean(M) = 1, standard deviation (S) = 2) and for growth rate a Laplace distribution ($M = 0$, $S = 100$). The uncorrelated lognormal relaxed clock was selected as the best fitting clock model using Bayes Factor comparisons of strict and relaxed clocks based on path sampling/stepping stone analysis[54]. Clock priors were defined as: ucld.mean: lognormal distribution with mean in real space = 1.4×10^{-3} subs per site per year and s.d. of the uncorrelated lognormal relaxed clock ucld.stdev = 5×10^{-2}. Parameters were estimated using Markov chain Monte Carlo (MCMC) Bayesian inference, with 5×10^{7} steps-long chains with exception of SEC7 and SEC8, for which longer chains were run (1×10^{8}); in all cases a total of 10^{5} steps were sampled in the log files. For all analysis, three independent runs starting from different seeds were conducted to ensure convergence, then combined with LogCombiner v.2.6.3 after removing the initial 10% of the MCMC as burn-in. Adequate mixing of parameters and convergence among runs were assessed using Tracer v.1.7.1 (ref. [55]) by verifying that each parameter reached an effective sampling size (ESS) above 200 and that traces showed stationarity and good mixing. The final posterior distribution contained a total of 9,000 trees, annotated with Treeannotator v.2.6.3 and visualized in FigTree v.1.4.3 (ref. [56])

Phylodynamics analysis to estimate R_e. To estimate discrete changes in R_e through for the two largest epidemic clades SEC7, SEC8 and two Pango lineages harboring the 614G mutation (B.1, B.1.1), we used a Bayesian birth–death skyline model (BDSKY) with serial sampling[57] implemented in BEAST v.2.6 (ref. [53]). BDSKY uses an episodic, piecewise birth–death model in which the parameter is allowed to change at discrete points in time, with the magnitude and timing of changes estimated from the data. In our analysis, we set two intervals wherein R_e is constant and estimated the date with most evidence for a change in R_e. To this end, we set a uniform prior distribution. R_e was estimated before and after the changing time. The same parameters as above were used but fixing the clock rate and the recovery rate (become uninfectious rate, $\delta = 36.5$ years^{-1}) in accordance with consistent global estimates of an infectious period of 10 days[58]. To avoid bias in the model parameters due to constant sampling proportion assumed by BDSKY models, this parameter was set to zero before the first sample date using TreeSlicer (https://github.com/laduplessis/skylinetools/wiki). For this analysis, 1×10^{8} and 4×10^{8} steps-long chains were used for SEC7/B.1/B.1.1 and SEC8, respectively. Results were inspected with Tracer (v.1.7.1)[55] by verifying that every parameter had effective sampling sizes above 200 and good mixing was obtained. Doubling time was calculated from the parameters estimated by BDSKY model in which growth rate $(r) = (R_e \times \delta) - \delta$ and doubling time $= \ln(2) r^{-1}$.

Statistical analysis. All statistical analyses were carried out using the R statistical language[59]. Packages ape[60], treeio[60,61], doParallel[62] and foreach[63] were used for phylogenetic manipulation and analysis. We additionally used packages geosphere[64], lwgeom[65], sp[66], sf[67] and rgeos[68] to calculate the geographic distances between samples and the geographical representation in the data. The ggplot2 R package[69] was used extensively for analysis and data plotting.

Epidemic wave definitions. There is not a formal (or even official) definition for 'first' and 'second' wave, and the valley between both. We therefore used two lines of evidence to define the boundaries. One is the official number of total cases diagnosed by PCR[70] and the second is the end of the mobility restrictions. On the bases of these facts, we tentatively identify the following dates for the different waves:

First wave: February 2020–14 May 2020 (from the first case reported to the end of the national lockdown and start of lifting measures).

Second wave: July 2020–first week of December 2020 (from the first large outbreaks reported after the first wave, which were caused and led to the expansion of the 20E/EU1 variant across the country to the new increase of cases in December).

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

Data availability

All the genomic sequences used in the analyses are available in the GISAID database, and the accession numbers, originating and submission laboratories can be found in Supplementary Table 1. Sequencing data (fastq files) of the samples sequenced by the SeqCOVID consortium have been deposited to the European Nucleotide Archive (ENA), and the corresponding accession numbers can also be found in Supplementary Table 1.

Code availability

The analysis pipeline used to map and analyze the sequences is available at https://gitlab.com/fisabio-ngs/sars-cov2-mapping.

References

38. Quick, J. nCoV-2019 sequencing protocol. *protocols.io* https://doi.org/10.17504/protocols.io.bbmuik6w (2020).
39. artic-network. artic-network/artic-ncov2019. *github* https://github.com/artic-network/artic-ncov2019 (2019).
40. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
41. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
42. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
43. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
44. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
45. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
46. De Maio, N. et al. Issues with SARS-CoV-2 sequencing data. *virological.org* https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (2020).
47. Lanfear, R. A global phylogeny of SARS-CoV-2 sequences from GISAID. *Zenodo* https://doi.org/10.5281/ZENODO.3958883 (2020).
48. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
49. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
50. Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**, 2685–2686 (2012).
51. *QGIS Geographic Information System* (QGIS Development Team, 2020); https://qgis.org
52. 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).
53. Bouckaert, R. et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
54. Grummer, J. A., Bryson, R. W. & Reeder, T. W. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). *Syst. Biol.* **63**, 119–133 (2014).
55. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
56. Rambaut, A. *FigTree* http://tree.bio.ed.ac.uk/software/figtree/ (2016).
57. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* **110**, 228–233 (2013).
58. He, X. et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
59. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
60. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
61. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
62. Microsoft Corporation & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'Parallel' Package* (Comprehensive R Archive Network (CRAN), 2019).
63. Microsoft Corporation & Weston, S. *foreach: Provides Foreach Looping Construct* (CRAN, 2020).
64. Hijmans, R. J., Williams, E. & Vennes, C. *Spherical Trigonometry* (CRAN, 2019).
65. Pebesma, E., Rundel, C. & Teucher, A. *lwgeom: Bindings to Selected 'liblwgeom' Functions for Simple Features* (CRAN, 2020).
66. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R* (Springer Science & Business Media, 2013).
67. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439–446 (2018).
68. Bivand, R. et al. *rgeos: Interface to Geometry Engine – Open Source ('GEOS')* (CRAN, 2020).
69. Wilkinson, L. ggplot2: Elegant graphics for data analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).
70. Instituto de Salud Carlos III. *COVID-19: Data and Documentation* (Spanish Government, 2021).

Acknowledgements

This work was mainly funded by the Instituto de Salud Carlos III project COV20/00140, with additional funding by Spanish National Research Council project CSIC-COV19-021, Ministerio de Ciencia project PID2019-104477RB-100, ERC StG 638553 and ERC CoG 101001038 to I.C., and BFU2017-89594R to F.G.C. M.C. is supported by Ramón y Cajal program from Ministerio de Ciencia and grants RTI2018-094399-A-I00 and Generalitat Valenciana (Regional Government) project SEJI/2019/011. We gratefully acknowledge Hospital Universitari Vall d'Hebron, Instituto de Salud Carlos III, IrsiCaixa AIDS Research Lab and all the international researchers and institutions that submitted sequenced SARS-CoV-2 genomes to the GISAID's EpiCov Database (Supplementary Table 1), as an important part of our analyses has been made possible by the sharing of their work. We also thank Unidad de Bioinformática y Estadística, Centro de Investigación Príncipe Felipe, for allowing us to use the Computer Cluster to perform some of the bioinformatic analysis.

Author contributions

I.C., F.G.C. and M.C. conceived the work. G.A.G., G.D.A. and S.J.S. set up the bioinformatics environment and the analysis pipeline. M.G.L., A.C.O., P.R.R. and N.G.G. analyzed the data. A.C.O. and M.G.L. wrote the first version of the draft. A.O., E.M., A.E.B., A.N., D.G.V., L.P.L., M.H., J.S., C.L.C., M.T., M.P.B.E., N.G.J., G.M., L.M.P., P.R.H., M.A.B., N.G.G., L.R.R., M.T.P., I.G.N., J.F.P. and A.S. sequenced genomes.

G.C.E., M.M.R., L.P.V., J.M.M., R.M.M., M.D.T.B., J.A.L., V.G.G., A.R.V., D.N., E.A., I.T., A.C.B., S.H.C., M.L.C.R., A.R., A.T., A.M.B., J.M.N.M., I.P.C., S.S.G., B.F.E., C.G.C., B.P.B., I.T.P., A.C., V.M., M.P.G., L.R.F., J.L.P., J.A., J.J.C.A., M.C.P.G., J.A.B., N.R., J.L.L.H., M.A.Z. and M.P.R. provided samples. I.C., F.G.C., M.C., A.C.O., M.G.L., S.D. and D.G.V. critically reviewed and contributed to the final version of the paper. The remaining SeqCOVID-Spain consortium members provided samples, and clinical and epidemiological information.

Competing interests

The authors declare no competing interests.

Additional information

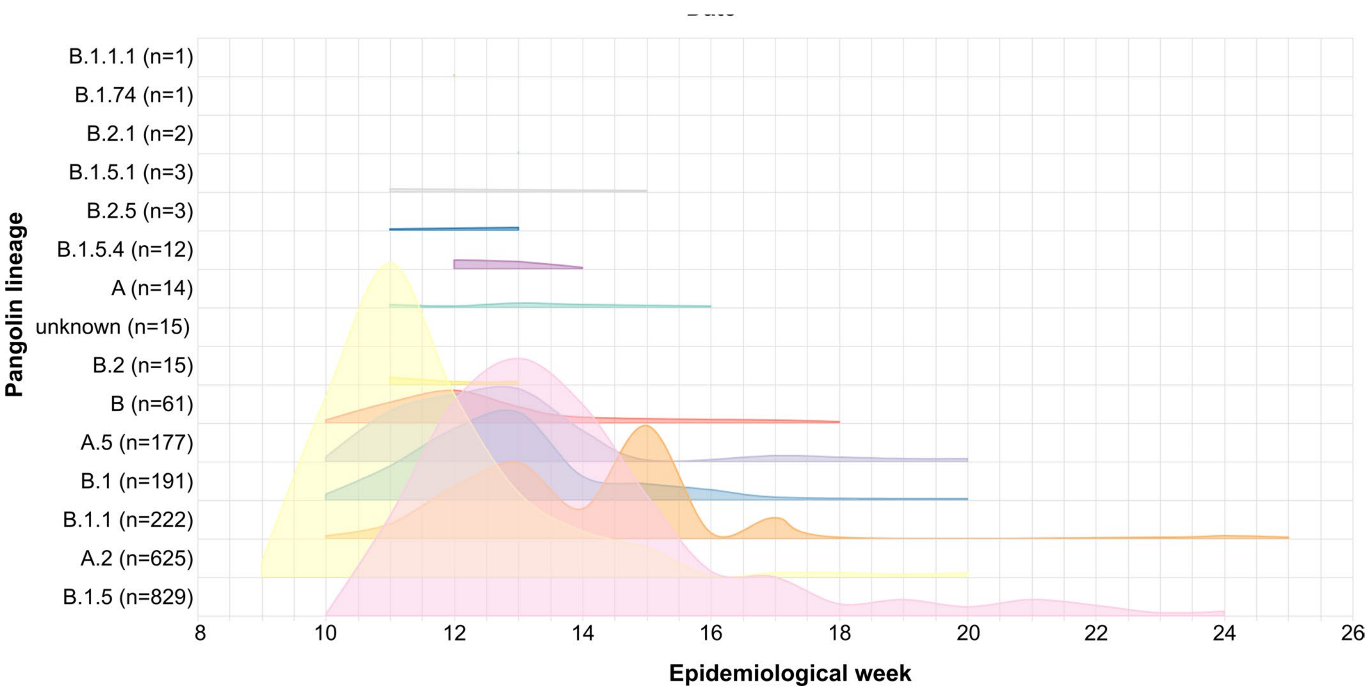
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00936-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00936-6.

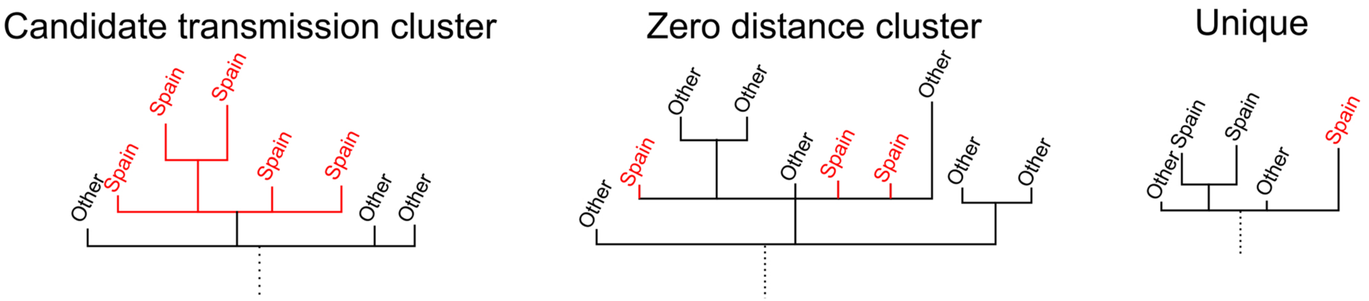
Correspondence and requests for materials should be addressed to Mireia Coscolla, Fernando González-Candelas or Iñaki Comas.

Peer review information *Nature Genetics* thanks R. Rockett and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

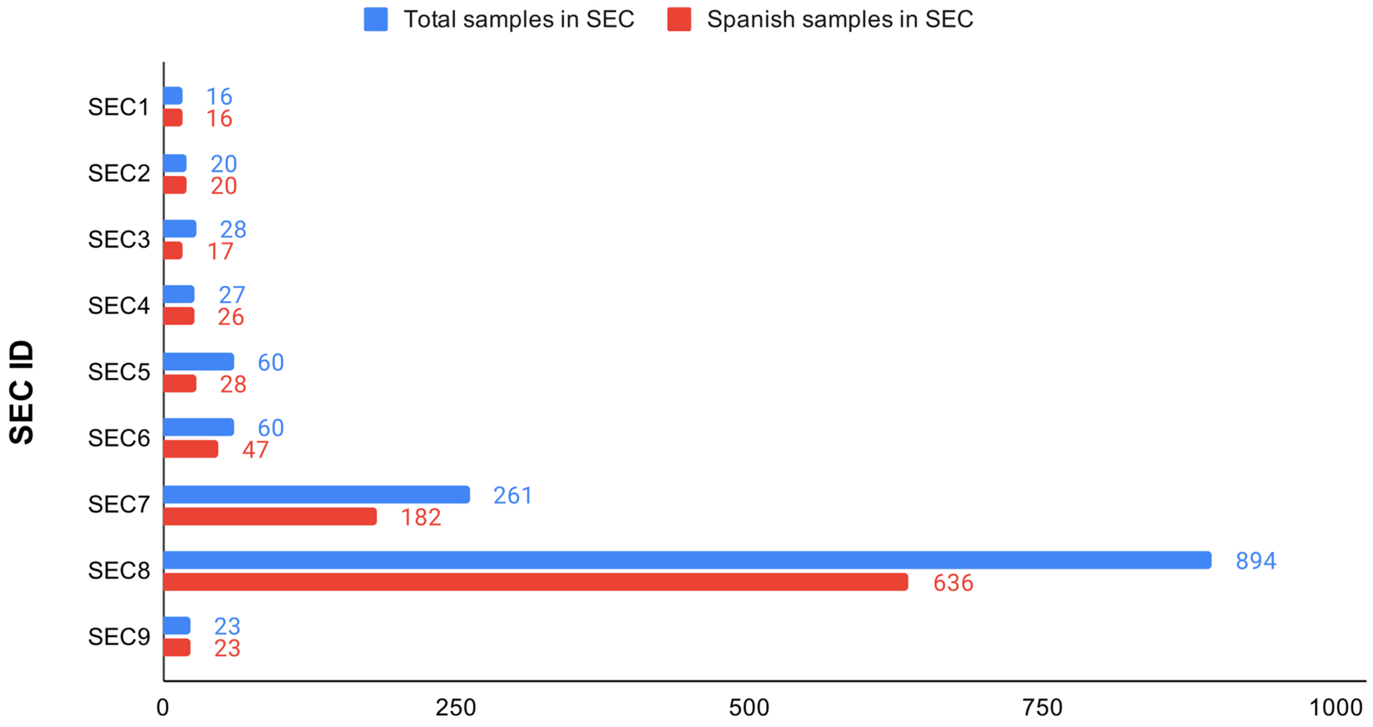
Reprints and permissions information is available at www.nature.com/reprints.



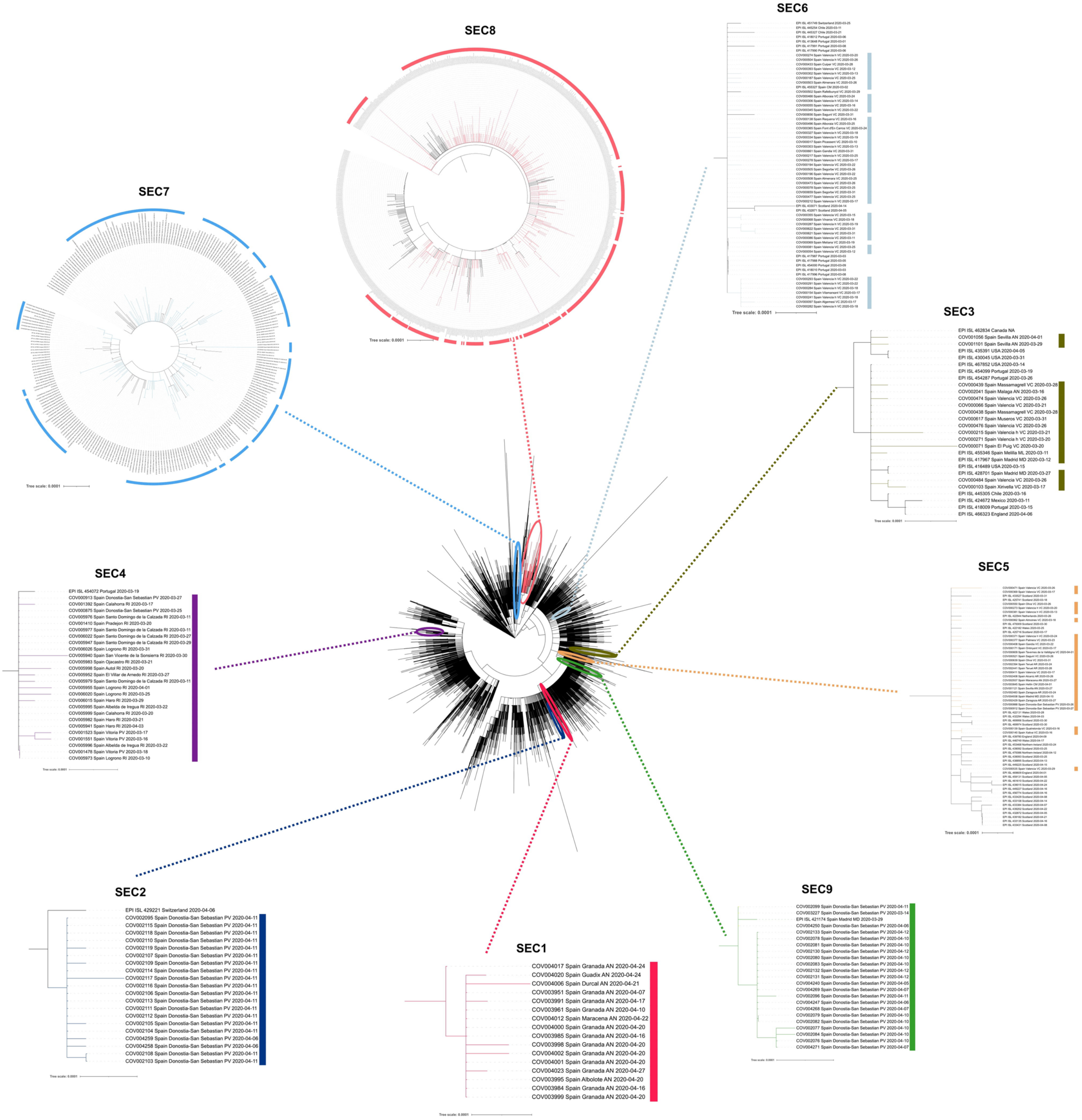
Extended Data Fig. 1 | Abundance of the different Pango lineages in the dataset. In the x-axis, the epidemiological week as plotted in Microreact.



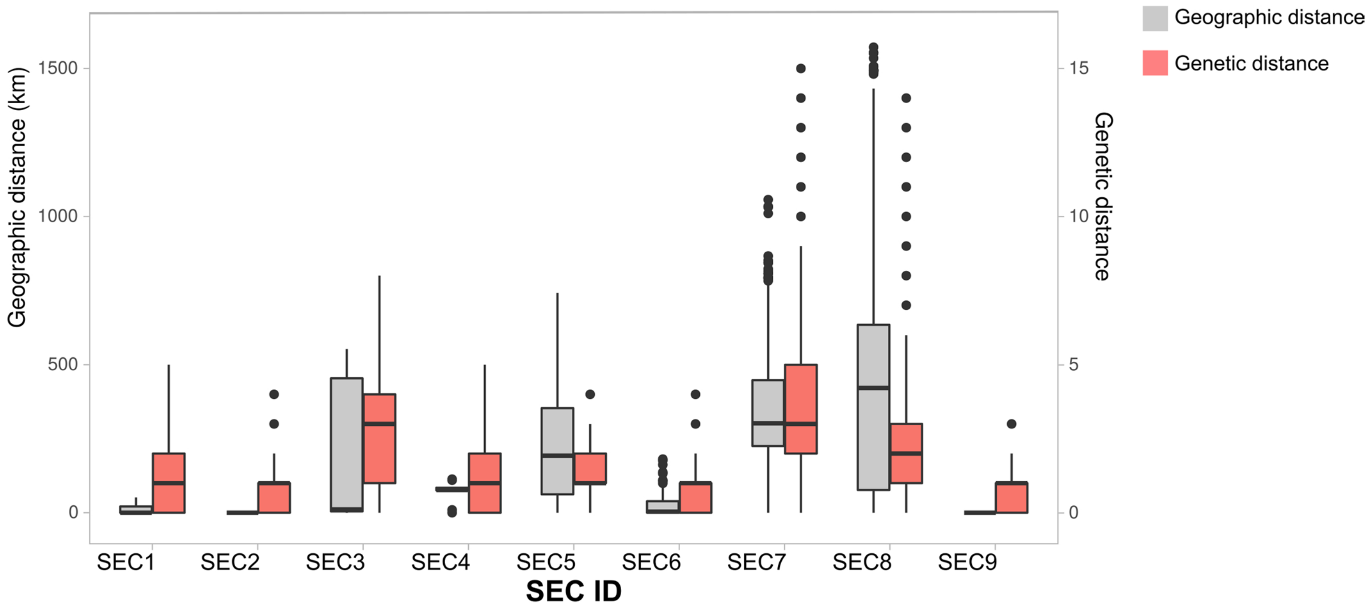
Extended Data Fig. 2 | Examples of the different groups of sequences identified. 'Candidate transmission clusters' are groups of Spanish sequences that form a clade. 'Zero distance clusters' are groups of Spanish sequences that are at zero distance from each other. Finally, the 'unique' sequences are Spanish sequences that are more than 1 SNP away from any other Spanish sequence and that do not share a most recent common ancestor (MRCA) node with other Spanish sequences.



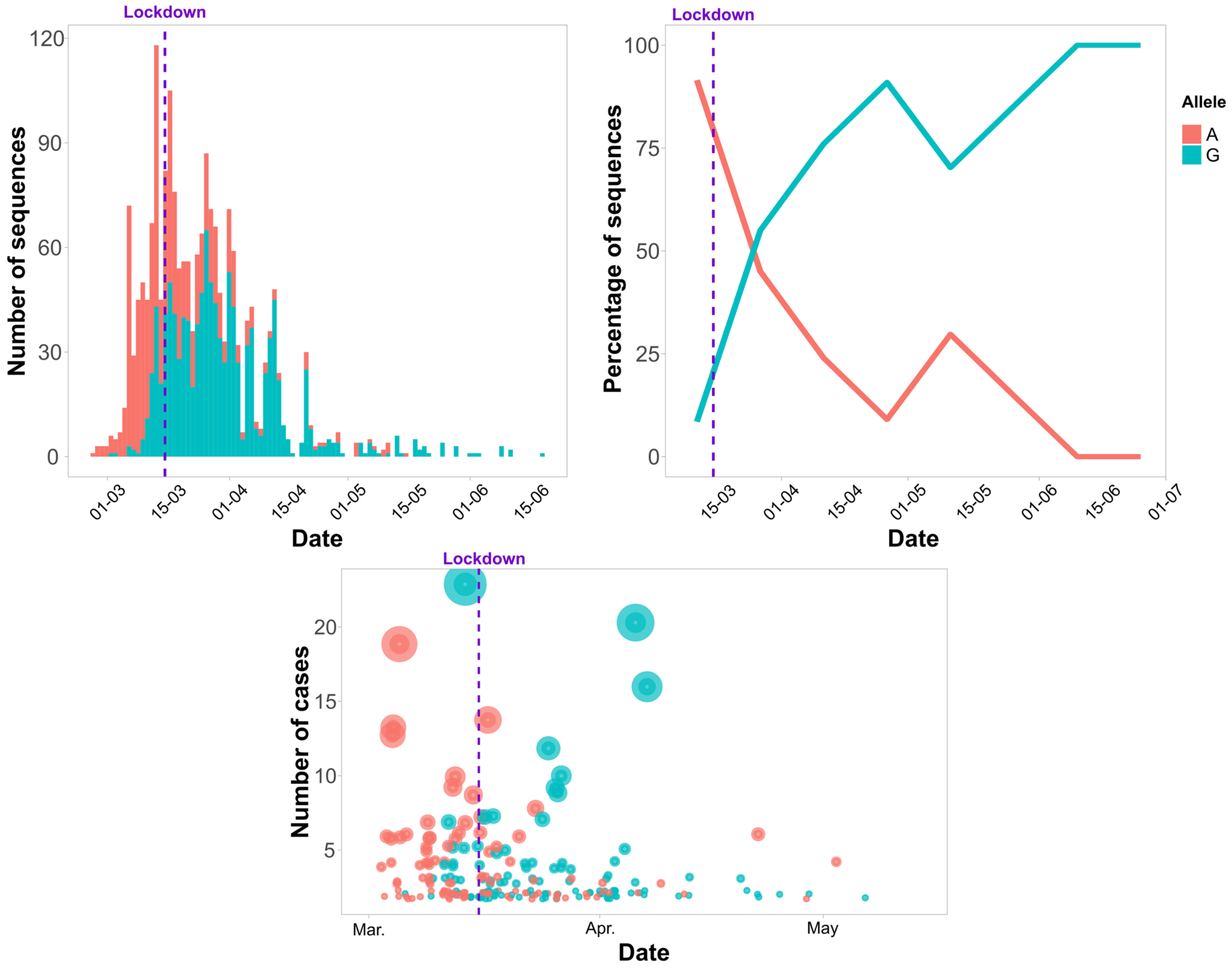
Extended Data Fig. 3 | Number of international and Spanish sequences in each SEC.


SEC8
SEC6
SEC7
SEC3
EPI ISL 462834 Canada NA
COV001056 Spain Sevilla AN 2020-04-01
COV001101 Spain Sevilla AN 2020-03-29
EPI ISL 435391 USA 2020-04-05
EPI ISL 430045 USA 2020-03-31
EPI ISL 467852 USA 2020-03-14
EPI ISL 454099 Portugal 2020-03-19
EPI ISL 454287 Portugal 2020-03-26
COV000439 Spain Massamagrell VC 2020-03-28
COV002041 Spain Malaga AN 2020-03-16
COV000474 Spain Valencia VC 2020-03-26
COV000066 Spain Valencia VC 2020-03-21
COV000438 Spain Massamagrell VC 2020-03-28
COV000617 Spain Museros VC 2020-03-31
COV000476 Spain Valencia VC 2020-03-26
COV000215 Spain Valencia h VC 2020-03-21
COV000271 Spain Valencia h VC 2020-03-20
COV000071 Spain El Puig VC 2020-03-20
EPI ISL 455346 Spain Melilla ML 2020-03-11
EPI ISL 417967 Spain Madrid MD 2020-03-12
EPI ISL 416489 USA 2020-03-15
EPI ISL 428701 Spain Madrid MD 2020-03-27
COV000484 Spain Valencia VC 2020-03-26
COV000103 Spain Xirivella VC 2020-03-17
EPI ISL 445305 Chile 2020-03-16
EPI ISL 424672 Mexico 2020-03-11
EPI ISL 418009 Portugal 2020-03-15
EPI ISL 466323 England 2020-04-06
SEC4
EPI ISL 454072 Portugal 2020-03-19
COV000913 Spain Donostia-San Sebastian PV 2020-03-27
COV001392 Spain Calahorra RI 2020-03-17
COV000875 Spain Donostia-San Sebastian PV 2020-03-25
COV005976 Spain Santo Domingo de la Calzada RI 2020-03-11
COV001410 Spain Pradejon RI 2020-03-20
COV005977 Spain Santo Domingo de la Calzada RI 2020-03-11
COV006022 Spain Santo Domingo de la Calzada RI 2020-03-27
COV005947 Spain Santo Domingo de la Calzada RI 2020-03-29
COV006026 Spain Logrono RI 2020-03-31
COV005940 Spain San Vicente de la Sonsierra RI 2020-03-30
COV005963 Spain Ojacastro RI 2020-03-21
COV005998 Spain Autol RI 2020-03-20
COV005952 Spain El Villar de Arnedo RI 2020-03-27
COV005979 Spain Santo Domingo de la Calzada RI 2020-03-11
COV005955 Spain Logrono RI 2020-04-01
COV006020 Spain Logrono RI 2020-03-25
COV006015 Spain Haro RI 2020-03-29
COV005995 Spain Albelda de Iregua RI 2020-03-22
COV005999 Spain Calahorra RI 2020-03-20
COV005982 Spain Haro RI 2020-03-21
COV005941 Spain Haro RI 2020-04-03
COV001523 Spain Vitoria PV 2020-03-17
COV001551 Spain Vitoria PV 2020-03-16
COV005996 Spain Albelda de Iregua RI 2020-03-22
COV001478 Spain Vitoria PV 2020-03-18
COV005973 Spain Logrono RI 2020-03-10
SEC5
SEC2
EPI ISL 429221 Switzerland 2020-04-06
COV002095 Spain Donostia-San Sebastian PV 2020-04-11
COV002115 Spain Donostia-San Sebastian PV 2020-04-11
COV002118 Spain Donostia-San Sebastian PV 2020-04-11
COV002110 Spain Donostia-San Sebastian PV 2020-04-11
COV002119 Spain Donostia-San Sebastian PV 2020-04-11
COV002107 Spain Donostia-San Sebastian PV 2020-04-11
COV002109 Spain Donostia-San Sebastian PV 2020-04-11
COV002114 Spain Donostia-San Sebastian PV 2020-04-11
COV002117 Spain Donostia-San Sebastian PV 2020-04-11
COV002116 Spain Donostia-San Sebastian PV 2020-04-11
COV002106 Spain Donostia-San Sebastian PV 2020-04-11
COV002113 Spain Donostia-San Sebastian PV 2020-04-11
COV002111 Spain Donostia-San Sebastian PV 2020-04-11
COV002112 Spain Donostia-San Sebastian PV 2020-04-11
COV002105 Spain Donostia-San Sebastian PV 2020-04-11
COV002104 Spain Donostia-San Sebastian PV 2020-04-11
COV004259 Spain Donostia-San Sebastian PV 2020-04-06
COV004258 Spain Donostia-San Sebastian PV 2020-04-06
COV002108 Spain Donostia-San Sebastian PV 2020-04-11
COV002103 Spain Donostia-San Sebastian PV 2020-04-11
SEC1
COV004017 Spain Granada AN 2020-04-24
COV004020 Spain Guadix AN 2020-04-24
COV004006 Spain Durcal AN 2020-04-21
COV003951 Spain Granada AN 2020-04-07
COV003991 Spain Granada AN 2020-04-17
COV003961 Spain Granada AN 2020-04-10
COV004012 Spain Maracena AN 2020-04-22
COV004000 Spain Granada AN 2020-04-20
COV003985 Spain Granada AN 2020-04-16
COV003998 Spain Granada AN 2020-04-20
COV004002 Spain Granada AN 2020-04-20
COV004001 Spain Granada AN 2020-04-20
COV004023 Spain Granada AN 2020-04-27
COV003995 Spain Albolote AN 2020-04-20
COV003984 Spain Granada AN 2020-04-16
COV003999 Spain Granada AN 2020-04-20
SEC9
COV002099 Spain Donostia-San Sebastian PV 2020-04-11
COV003227 Spain Donostia-San Sebastian PV 2020-03-14
EPI ISL 421174 Spain Madrid MD 2020-03-29
COV004250 Spain Donostia-San Sebastian PV 2020-04-06
COV002133 Spain Donostia-San Sebastian PV 2020-04-12
COV002078 Spain Donostia-San Sebastian PV 2020-04-10
COV002081 Spain Donostia-San Sebastian PV 2020-04-10
COV002130 Spain Donostia-San Sebastian PV 2020-04-12
COV002080 Spain Donostia-San Sebastian PV 2020-04-10
COV002083 Spain Donostia-San Sebastian PV 2020-04-10
COV002132 Spain Donostia-San Sebastian PV 2020-04-12
COV002131 Spain Donostia-San Sebastian PV 2020-04-12
COV004240 Spain Donostia-San Sebastian PV 2020-04-05
COV004269 Spain Donostia-San Sebastian PV 2020-04-07
COV002096 Spain Donostia-San Sebastian PV 2020-04-11
COV004247 Spain Donostia-San Sebastian PV 2020-04-06
COV004268 Spain Donostia-San Sebastian PV 2020-04-07
COV002079 Spain Donostia-San Sebastian PV 2020-04-10
COV002082 Spain Donostia-San Sebastian PV 2020-04-10
COV002077 Spain Donostia-San Sebastian PV 2020-04-10
COV002084 Spain Donostia-San Sebastian PV 2020-04-10
COV002076 Spain Donostia-San Sebastian PV 2020-04-10
COV004271 Spain Donostia-San Sebastian PV 2020-04-07
Tree scale: 0.0001

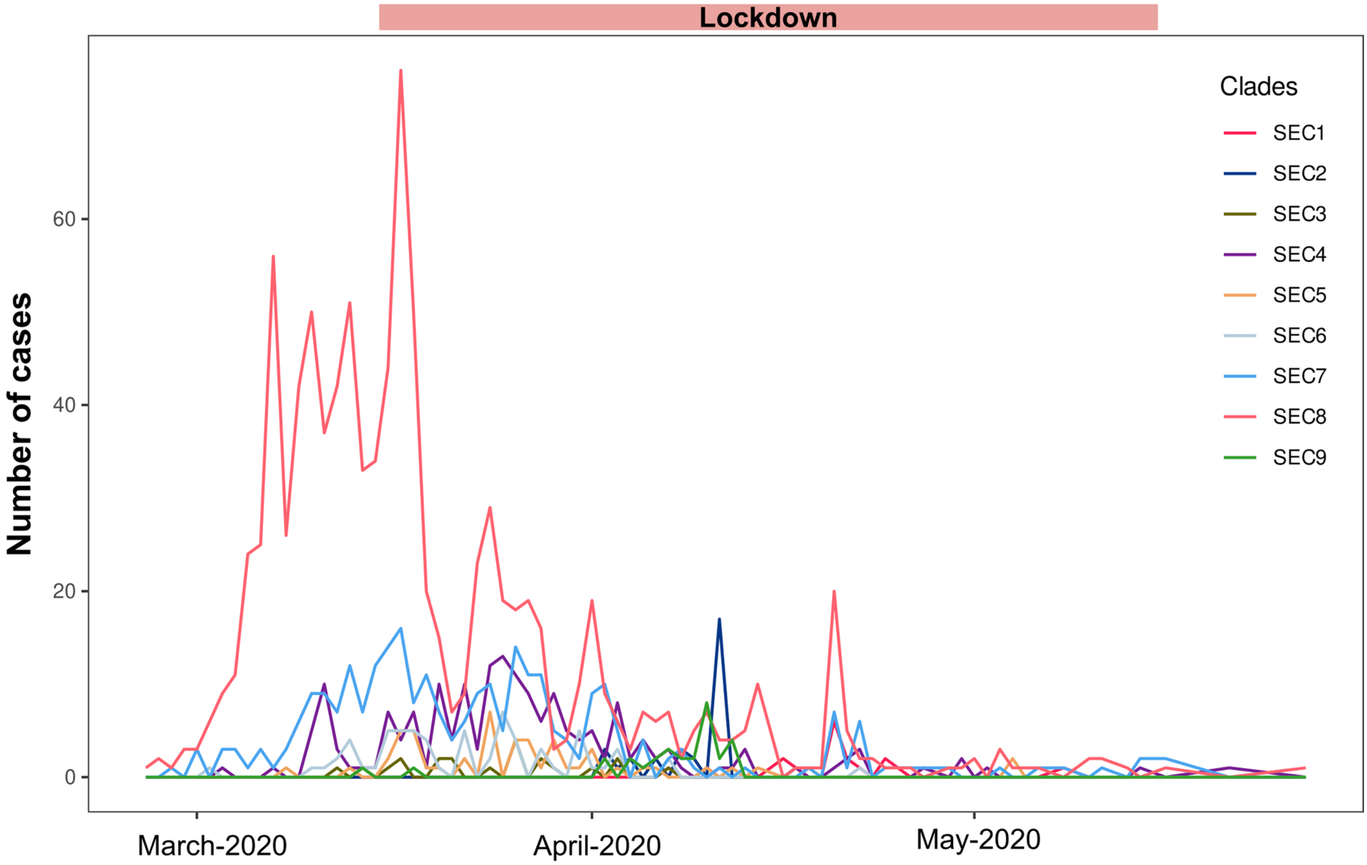
Extended Data Fig. 4 | Phylogenetic location of each SEC in the global SARS-CoV-2 phylogeny. Sequences from Spain are colored according to their SEC color (as indicated in Fig. 2a legend) while international sequences remain in black color.



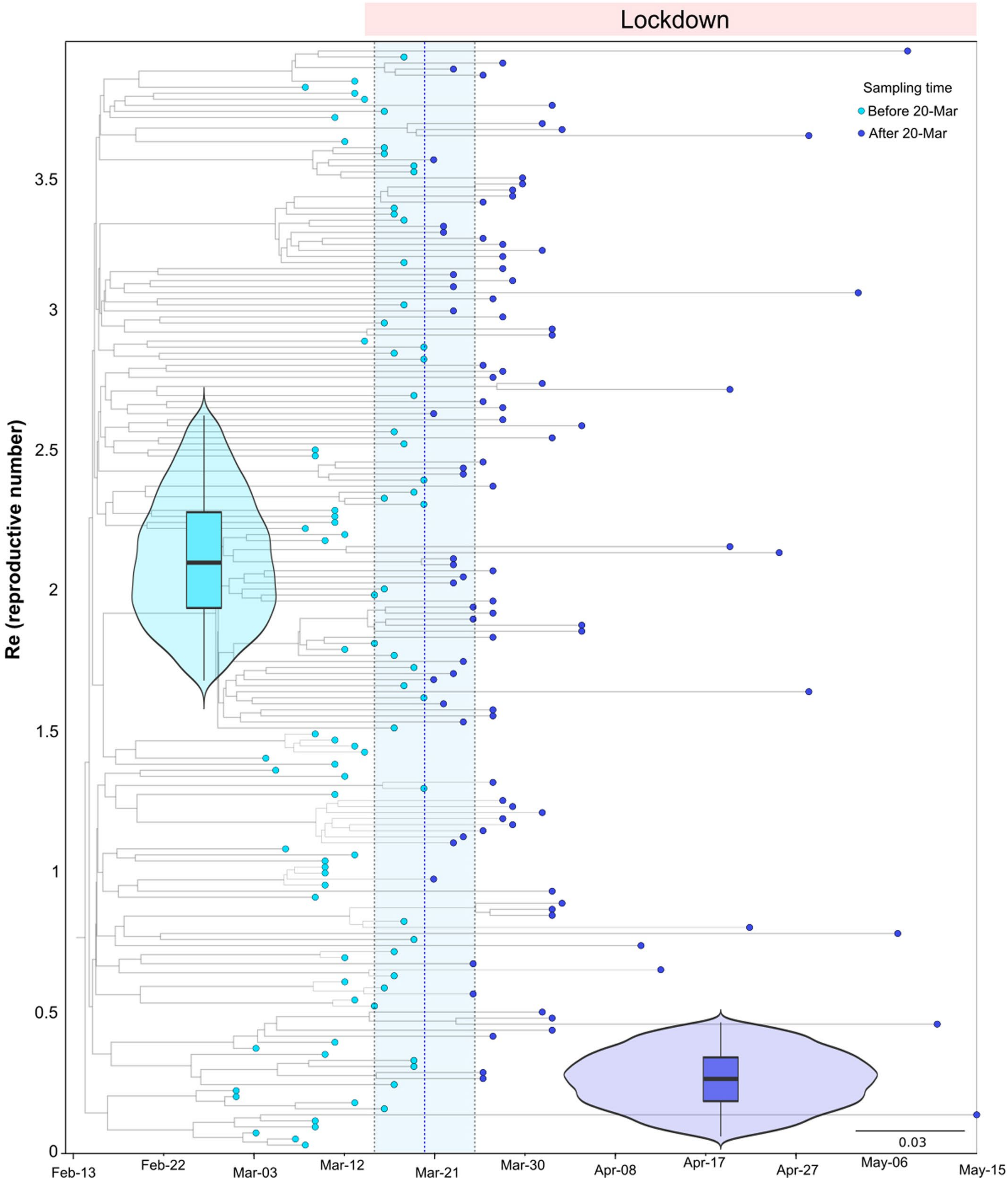
Extended Data Fig. 5 | Distribution of genetic (salmon) versus geographic (gray) distances within each pair of samples belonging to the same SEC. For each SEC we had the following comparisons (data points): SEC1 (N = 120), SEC2 (N = 190), SEC3 (N = 91), SEC4 (N = 325), SEC5 (N = 378), SEC6 (N = 990), SEC7 (N = 14,028), SEC8 (N = 178,503) and SEC9 (N = 231). The lower whisker, higher whisker, center and bounds of each boxplot refers to quartile 1–1.5 interquartile range, quartile 3 + 1.5 interquartile range, mean, first and third quartiles of the data. Individual points are outliers (values lower than quartile 1–1.5 interquartile range and higher than quartile 3 + 1.5 interquartile range).



Extended Data Fig. 6 | Distribution of sequences harboring the 614 G mutation (blue) versus the 614D mutation (salmon,wild-type) in the S gene for the spanish sequences in our dataset. In the left panel, a histogram of samples sorted by date of sequencing. At right, frequency of both mutations in the sequenced samples by date. The national lockdown event is marked by a purple vertical line. At the bottom, 'candidate transmission clusters' by date and size, colored according to the allele found at the 614 position of the S gene.

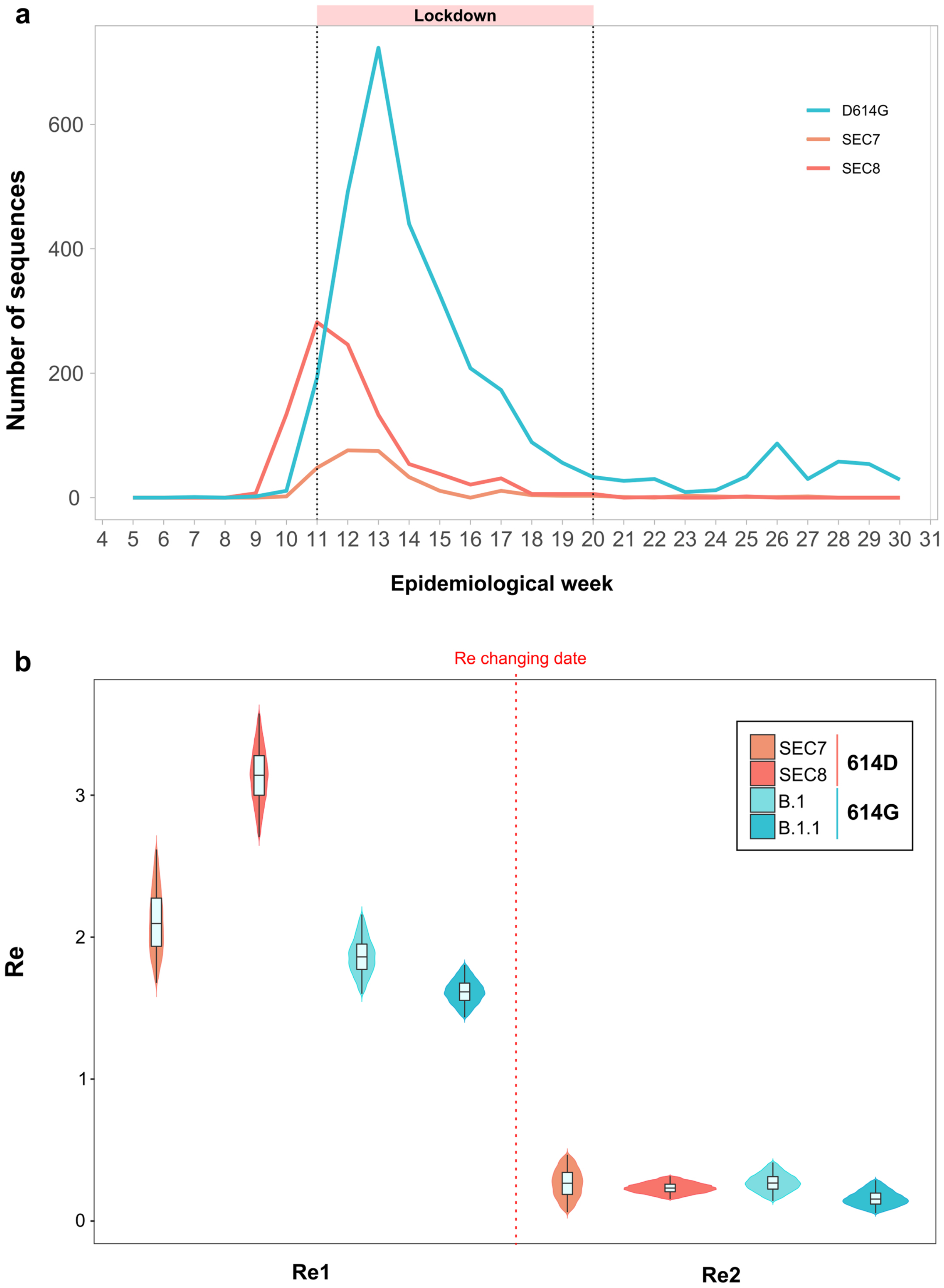


Extended Data Fig. 7 | Cases sequenced during the period that includes the first wave until the end of the lockdown (14th May, 2020). Lines represent the number of cases belonging to different Spanish Epidemic Clades (SECs).



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Phylodynamic estimates of the effective reproductive number (Re) of Spanish SEC7. A birth–death skyline (BDSKY) model was implemented in Beast v.2, allowing for piecewise changes in Re, with the time and magnitude estimated from the data. The X axis represents time, starting with the MRCA of all sampled diversity within SEC7 and ending with the date of the most recently sequenced genome from 15th May. The blue dotted line indicates the posterior value of the timing of a most important decrease in Re, around 20th March 2020 [95% HPD: 15–25th March]. The Y axis represents Re, and the violin plots show the posterior probability distribution for this parameter before and after the change time in Re; with a mean of 2.10 [95% HPD: 1.67–2.62] and 0.27 [95% HPD:0.06–0.47] before and after this time, respectively. The phylogenetic tree in the background is the maximum clade credibility tree from the BDSKY analysis, with the tips colored according to whether they were sampled before or after 20th March. The lower whisker, higher whisker, center and bounds of each boxplot refers to quartile 1–1.5 interquartile range, quartile 3 + 1.5 interquartile range, mean, first and third quartiles of the data. Individual points are outliers (values lower than quartile 1–1.5 interquartile range and higher than quartile 3 + 1.5 interquartile range). Boxplot was constructed with all the Spanish sequences in SEC7 (N = 182).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Comparison between strains carrying 614D and 614 G mutations. a. Number of sequences belonging to SEC7, SEC8 (614D) and those having the 614 G allele. **b**. Phylodynamic estimates of the effective reproductive number (Re) of clades harboring 614D mutation (SEC7 and SEC8) and Pango lineages with 614 G mutation (B.1 and B1.1). A birth-death skyline (BDSKY) model was implemented in Beast v.2, allowing for piecewise changes in Re, with the time and magnitude estimated from the data. The violin plots show the posterior probability distribution (HPD) interval for Re parameter before (Re1) and after (Re2) the changing time estimates (dotted line). For SEC7, Re1: 2.10 [95% HPD: 1.67–2.62] and Re2: 0.27 [95% HPD:0.06–0.47]; changing time 20th March 2020 [95% HPD: 15–25th March]. For SEC8, Re1: 3.14 [95% HPD: 2.71–3.57] and Re2: 0.23 [95% HPD: 0.15–0.32]; changing time 9th March 2020 [95% HPD: 8–10th March]. For B.1, Re1: 1.86 [1.6–2.16] and Re2: 0.26 [0.14–0.41]; changing time 23rd March 2020 [95% HPD: 21–25th March]. For B.1.1, Re1: 1.62 [1.44–1.80] and Re2: 0.15 [0.05–0.29]; changing time 10th April 2020 [95% HPD: 9–12th April]. The lower whisker, higher whisker, center and bounds of each boxplot refers to quartile 1–1.5 interquartile range, quartile 3 + 1.5 interquartile range, mean, first and third quartiles of the data. Individual points are outliers (values lower than quartile 1–1.5 interquartile range and higher than quartile 3 + 1.5 interquartile range). Boxplots were constructed with all the Spanish sequences in SEC7 (N = 182), SEC8 (N = 636), B.1 (N = 191) and B.1.1 (N = 223).

nature portfolio

Corresponding author(s): Mireia Coscollá, Fernando González-Candelas, Iñaki Comas

Last updated by author(s): Jun 29, 2021

Reporting Summary

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

Statistics

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

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☒ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection: No software was used for data collection

Data analysis: The custom pipeline analysis used to analyze the data is fully accessible at (https://gitlab.com/fisabio-ngs/sars-cov2-mapping). We have used the following softwares (and versions). bwa (v.0.7.17), iVAR (v.1.2), Kraken (v.0.10.5-beta), fastp (v0.20.1), MultiQC (v.1.8), MAFFT (v.7.471), IQTREE (v.1.6.12), iTOL (v.5), MEGA (v.10.0.5), QGIS (v.3.14.16-Pi), TempEst (v.1.5.3), Beast (V.2.6), LogCombiner (v.2.6.3), Treeannotator (v.2.6.3), FigTree (v.1.4.3), Tracer (v.1.7.1), R (v.4.0.4).

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the genomic sequences used in the analyses are available in the GISAID database (https://www.gisaid.org/) and the accession numbers can be found in Supplementary Table 1.

Sequencing data (fastq files) of the samples sequenced by the SeqCOVID-consortium have been deposited to ENA (https://www.ebi.ac.uk/ena/browser/home), and the corresponding accession numbers can also be found in Supplementary Table 1.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We used all the SARS-CoV-2 samples sequenced by the SeqCOVID-Spain consortium by July 2020. In addition, we added all the Spanish samples deposited in the GISAID database at this time. All the sequences used are freely accesible and the accession numbers can be found in Suplementary Table 1.
Data exclusions	No data was excluded from the analyses.
Replication	Findings can be replicated by downloading the sequences listed in Supplementary Table 1, and applying the approaches described in the Methods section (alignment, phylogenetic reconstruction, identification of clusters and SECs, dating...). We used those approaches regularly and we have not noticed any replicability issue.
Randomization	This is not relevant in our study as we have not gone through any randomization step in our analyses.
Blinding	This is not relevant in our study, as we have not performed any clinical trial, clustering or any other assessment in which blinding was required.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Samples came from patients that were SARS-CoV-2 positive at Spanish hospitals during the study period. This is a convenience sample so no bias regarding age and sex have been applied for the selection of samples. For the 2171 samples, 1111 were female, 948 were male and 112 were unknown/not reported. For the 2171 samples, 13 were under 15 years old, 236 were in range 15-35, 600 were in range 35-55, 615 were in range 55-75, 582 were older than 75 y.o. and 125 were unknown
Recruitment	Samples were the remaining RNA extracts from naso- and oropharyngeal clinical specimens employed for diagnosis at the Microbiological services from participating hospitals as part of the routine SARS-CoV-2 diagnosis. The use of such samples have been aproved by the ethics comittee. We did not demand a minimum/maximum number of samples per hospital. Hence, hospitals sent a variable number of samples depending on their technical and personal posibilities. This were traduced into an heterogeneous sampling for each hospital, independently of the epidemic impact suffered in this hospital's region. This hetereogeneinty could have affected some of the analysis, as we have pointed out in the Discussion although it is unlikely.
Ethics oversight	The use of such samples have been approved by the ethics comittee Comité Ético de Investigación en Salud Pública y Centro Superior de Salud Pública (CEI DGSP-CSISP) Nº 20200414/05.

Note that full information on the approval of the study protocol must also be provided in the manuscript.