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

Maríana G. López1,111, Álvaro Chiner-Oms1,111, Dario García de Viedma2,3,4, Paula Ruiz-Rodríguez5, Maria Alma Bracho6,7, Irving Cancino-Muñoz1, Giuseppe D’Auria7,8, Griselda de Marco8, Neris García-González6, Galo Adrian Goig9, Inmaculada Gómez-Navarro1, Santiago Jiménez-Serrano1, Lluís Martínez-Priego8, Paula Ruiz-Hueso8, Lidia Ruiz-Roldán6, Manuela Torres-Puente1, Juan Alberola10,11,12, Eliseo Albert13, Maitane Aranzamendi Zaldumbide14,15, María Pilar Bea-Escudero16, Jose Antonio Boga17,18, Antoni E. Bordoy19, Andrés Canut-Blasco20, Ana Carvajal21, Gustavo Cilla Eguiluz22, Maria Luz Cordón Rodríguez20, José J. Costa-Alcalde23, María de Toro24, Inmaculada de Toro Peinado24, Jose Luis del Pozo25, Sebastián Duchéne26, Jovita Fernández-Pinero27, Begoña Fuster Escrivá12,28, Concepción Gimeno Cardona28, Verónica González Galán29, Nieves Gonzalo Jiménez30, Silvia Hernández Crespo29, Marta Herranz2,3,4, José Antonio Lepe29, Carla López-Causapé31, José Luis López-Hontangas32, Vicente Martín1,7,33, Elisa Martró19, Ana Milagro Beamonte34, Milagrosa Montes Ros22, Rosario Moreno-Muñoz35, David Navarro13,12, José María Navarro-Mar26,37, Anna Not19, Antonio Oliver31,38, Begoña Palop-Borrás24, Mónica Parra Grande39, Irene Pedrosa-Corral36,37, Maria Carmen Pérez-González40, Laura Pérez-Lago2,3, Mercedes Pérez-Ruiz24, Luis Piñeiro Vázquez22, Nuria Rabella41,42,43, Antonio Rezusta34,44,45, Lorena Robles Fonseca46, Ángel Rodríguez-Villodres29, Sara Sanbonmatsu-Gámez26,37, Jon Sicilia2,3, Alex Soriano47, María Dolores Tirado Balaguer35, Ignacio Torres3, Alexander Tristancho34,44, José María Marimón22, SeqCOVID-Spain consortium*, Minea Coscolla5,5, Fernando González-Candelas6,7, and Iñaki Comas1,7,12

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 fatalities3. Furthermore, a nationwide seroprevalence study showed that only one in ten cases of infection by SARS-CoV-2 were diagnosed and reported in that period1, 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
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. 8).

Despite the high incidence of infection across the country, some regions had substantially higher incidence than others. Genomic epidemiology and phylogenetics8–10 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 GISAID11, 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 evolves5,12,13, which limits the resolution of phylogenetic and phylodynamic analyses14. 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 (https://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 effect of lockdown measures in controlling the virus spread and eliminating 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 GISAID11 (Supplementary Table 1). Spain showed a particular viral population structure with a higher proportion of lineage A sequences compared with other European countries18 (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 countries8–10. 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
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 Pvalue « 0.01, Supplementary Note).

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 visualization⁵. 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%.
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
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). 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 95%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 SECs23. 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 suggests the reduction in the rate of increase of confirmed cases and that is in agreement with estimates from epidemiological models and data\textsuperscript{11,12}.

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\textsuperscript{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\textsuperscript{34}. 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\textsuperscript{11,12,23,24}.

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\textsuperscript{29} and in transmission studies\textsuperscript{9,36}. Consistently, our analysis supports the observation\textsuperscript{36} 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 doubtful\textsuperscript{36}, suggesting that its impact, if any, on epidemic transmission was minor. In the case of Spain, 614D 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, 614D 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\textsuperscript{35}. 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\textsuperscript{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\textsuperscript{35}. The most successful SECs during the first wave were probably those that arrived early, several times, and to diverse locations. Thus, as suggested elsewhere\textsuperscript{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\textsuperscript{37}. 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\textsuperscript{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\textsuperscript{36}. 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. By contrast, new variants with different transmissibility and immunogenicity profiles started to emerge 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, 202001164 (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 seroprevalence study. Lancet 396, 535–544 (2020).
5. Ministerio de la Presidencia, R. C. las C. y. M. D. BOE-A-2020-3692
6. Instituto de Salud Carlos III, Spanish Government. COVID-19 reported cases in Spain https://covid19.isciii.es/covid19/#cua (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/Aactualizacion_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. Voz, 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. Vaysseix, 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, 236–239 (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. Gunthegger, 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. Candel, 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. Gryafi, A. & Moríña, 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. Bohmer, 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. Nature 592, 116–121 (2021).
26. Voel, E. M. et al. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. Cell 184, 64–75.e11 (2021).
27. Barbot, 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).
28. van Dorp, L. et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat. Commun. 11, 5986 (2020).
29. To, V. et al. COVID-19 community mobility reports. Google https://www.google.com/covid19/mobility/ (2021).
30. Russell, T. W. et al. Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study. Lancet Public Health 6, e12–e20 (2021).
31. 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).
32. Lemieux, J. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, 6529 (2021).
33. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature 595, 707–712 (2021).
34. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb. Genomics 2, e00093 (2016).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
SeqCOVID-Spain consortium

Álvaro Chiner-Oms1,111, Irving Cancino-Muñoz1, Mariana G. López1,111, Manuela Torres-Puente1, Inmaculada Gómez-Navarro1, Santiago Jiménez-Serrano1, Jordi Pérez-Tur1, Darío García de Viedma2,3,4, Laura Pérez-Lago2,3, Marta Herranz2,3,4, Jon Sicilia2,3, Pilar Catalán-Alonso2,3,4, Julia Suárez González2, Patricia Muñoz2,3, Mireia Coscolla5, Paula Ruiz-Rodríguez5, Fernando González-Candelas6,7, Iñaki Comas1,7, Lidia Ruiz-Roldán6, María Alma Bracho6,7, Neris García-González6, Lluícia Martínez Priego6, Inmaculada Galán-Vendrell6, Paula Ruiz-Hueso8, Griselda De Marco8, María Loreto Ferrús-Abad8, Sandra Carbó-Ramírez8, Giuseppe D'Auria7,8, Galo Adrian Goig9, Juan Alberola10,11,12, Jose Miguel Nogueira10,11,12, Juan José Camarena10,11,12, David Navarro12,13, Eliseo Albert13, Ignacio Torres13, Maitane Aranzamendi Zaldumbide14,15, Óscar Martínez Expósito14,15, Nerea Antona Urieta14,15, María de Toro16, Marla Pilar Bea-Escudero16, Jose Antonio Boga17,18, Santiago Melón17,18, Elisa Martró7,19, Darío García de Viedma2,3,4, Laura Pérez-Lago2,3, Marta Herranz2,3,4, Jon Sicilia2,3, Pilar Catalán-Alonso2,3,4, Julia Suárez González2, Patricia Muñoz2,3, Mireia Coscolla5, Paula Ruiz-Rodríguez5, Fernando González-Candelas6,7, Iñaki Comas1,7, Lidia Ruiz-Roldán6, María Alma Bracho6,7, Neris García-González6, Lluícia Martínez Priego6, Inmaculada Galán-Vendrell6, Paula Ruiz-Hueso8, Griselda De Marco8, María Loreto Ferrús-Abad8, Sandra Carbó-Ramírez8, Giuseppe D'Auria7,8, Galo Adrian Goig9, Juan Alberola10,11,12, Jose Miguel Nogueira10,11,12, Juan José Camarena10,11,12, David Navarro12,13, Eliseo Albert13, Ignacio Torres13, Maitane Aranzamendi Zaldumbide14,15, Óscar Martínez Expósito14,15, Nerea Antona Urieta14,15, María de Toro16, Marla Pilar Bea-Escudero16, Jose Antonio Boga17,18, Santiago Melón17,18, Elisa Martró7,19, Antoni E. Bordoy19, Anna Not19, Adrián Antuori19, Anabel Fernández-Navarro19, André Canut-Blasco20, Silvia Hernández Crespo20, Maria Luz Cordón Rodríguez20, Maria Concepción Lecaroz Agara20, Carmen Gómez-González20, Amaia Aguirre-Quinonero20, José Israel López-Mirones20, Marina Fernández-Torres20, Maria Rosario Almela-Ferrer20, Ana Carvajal21, Juan Miguel Fregeneda-Grande23, Héctor Argüello21, Gustavo Cilla Eguíluz22, Milagrosa Montes Ros22, Luis Piñeiro Vázquez22, Ana Sorarrain22, José María Marimón22, José J. Costa-Alcalde23, Rocio Trastoy23, Gema Barbeito Castiñeiras23, Amparo Coira23, María Luisa Pérez del Molino23, Antonio Aguilera23, Begoña Palop-Borràs24, Inmaculada de Toro Peinado24, Maria Concepción Mediavilla Gradolph24, Mercedes Pérez-Rúiz24, Mirian Fernández-Alonso25, Jose Luis del Pozo25, Oscar González-Recio27, Mónica Gutiérrez-Rivas27, Jovita Fernández-Pinero27, Miguel Ángel Jiménez Claver27, Begoña Fuster Escrivá28, Concepción Gimeno Cardona28, María Dolores Ootec Mochón28, Rafael Medina-González28, José Antonio Lepe29, Verónica González Galán29, Ángel Rodríguez-Villodres29, Nieves Gonzalo Jiménez30, Jordi Reina31, Carla López-Causapé31, Maria Dolores Gómez-Ruiz32,
Eva M. Gonzalez-Barbera32, José Luis López-Hontangas32, Vicente Martín33, Antonio J. Molina33, Tania Fernandez-Villa33, Ana Milagro Beamonte34, Nieves Felisa Martínez-Cameo34, Yolanda Gracia-Grataloup34, Rosario Moreno-Muñoz35, Maria Dolores Tirado Balaguer35, José María Navarro-Mari36,37, Irene Pedrosa-Corral36,37, Sara Sanbonmatsu-Gámez36,37, Antonio Oliver31,38, Mónica Parra Grande39, Bárbara Gómez Alonso39, Francisco José Arjona Zaragoz39, María Carmen Pérez González40, Francisco Javier Chamizo López40, Ana Bordes-Benítez40, Núria Rabella41,42,43, Ferran Navarro41,42,43, Elisís Maria Miró41,42,43, Antonio Rezusta34,44,45, Alexander Tristancho34,44, Encarnación Simarro Córdoba46, Julia Lozano-Serra46, Lorena Robles Fonseca46, Álex Soriano47, Francisco Javier Roig Sena48, Hermelinda Vanaclocha Luna48, Isabel Sanmartín49, Daniel García-Souto50,51,52, Ana Pequeño-Valtierra50, Jose M. C. Tubio50,51, Javier Temes50,51, Jorge Rodríguez-Castro50, Martín Santamarina García50, Manuel Rodríguez-Iglesias53,54,55, Fátima Galán-Sánchez53,54,55, Salud Rodríguez-Pallares53,54, José Manuel Azcona-Gutiérrez56, Miriam Blasco-Alberdi56, Alfredo Mayor57,58, Alberto L. García-Basteiro57,58, Gemma Moncunill57, Carlota Dobaño57, Pau Cistera57, Oriol Mitjà59,60, Camila González-Beiras59, Martí Vall-Mayans59, Marc Carbó-Monné59, Andrea Alemany59, Cristina Muñoz-Cuevas61,62, Guadalupe Rodríguez-Rodríguez61,62, Rafael Benito63,64, Sonia Algarate63, Jessica Bueno63, Andrea Vergara-Gómez64, Miguel J. Martínez65,66, Jordi Vila67,68, Elisa Rubio67,68, Aida Peiró-Mestres67,68, Jessica Naveiro-Castillejos67,68, David Posada67,68,69, Diana Valverde67,68,69, Nuria Estévez67, Iria Fernández-Silva67,68, Loretta de Chiara67,68, Pilar Gallego-García67, Nair Varela67, Ulises Gómez-Pinedo70, Mónica Gozalo-Margüello71, Maria Eliecer Cano García71, José Manuel Méndez-Legaza71, Jesús Rodríguez-Lozano71, María Siller71, Daniel Pablo-Marcos71, Maria Montserrat Ruiz-García72,73, Antonio Galiana73, Judith Sánchez-Almendro73, María Isabel Gascón Ros74, Cristina Juana Torregrosa-Hetland74, Eva María Pastor Boix74, Paloma Cascales Ramos74, Pedro Luis Gardínuñ Enríquez74, Salvador Raga Borja74, Julia González Cantó75, Olalla Martínez Macías75, Adolfo de Salazar76, Laura Viñuela González76, Natalia Chueca76, Federico García76, Cristina Gómez-Camarasa76, Amparo Farga Martí77, Rocío Falcón77, Victoria Domínguez-Márquez77, Anna M. Planas78, Israel Fernández-Cádenas79, María Àngeles Marcos80, Carmen Ezepeleta81,82, Ana Navascués81,82, Ana Miqueleiz Zapatero81, Manuel Segovia83,84, Antonio Moreno-Docón83,84, Esther Viedma85, Raúl Recio Martínez85, Irene Muñoz-Gallego85, Sara Gonzalez-Bodi85, Maria Dolores Folgueira85, Jesús Mingorance86, Elias Dahdou86, Fernando Lázaro-Perona86, María Rodríguez-Tejedor86, María Pilar Romero-Gómez86, Julio García-Rodríguez86, Juan Carlos Galá87, Mario Rodríguez-Dominguez87,88, Laura Martínez-García87,88,89, Melanie Abreu Di Berardino87,88, Manuel Ponce-Alonso87,88,89, Jose Maria González-Alba87,88, Ivan Sanz-Muño89,90, Diana Pérez San José90,91, Maria Gil Fortuño91, Juan B. Bellido-Blasco91, Alberto Yagüe Muñoz91, Noelia Hernández Pérez91, Helena Buj Jordá91, Óscar Pérez Olaso91, Alejandro González Praetorius92, Nora Mariela Martínez Ramírez92, Aida Ramírez Marinero93, Eduardo Padilla León93, Alba Vilas Bush93, Mireia Canal Aranda93, Albert Bernet Sánchez94, Alba Bellés Bellés94, Eric López González94, Iván Prats Sánchez94, Merec García-González94, Miguel José Martínez-Lirola95, Manuel Ángel Rodríguez Maresca95, Maria Teresa Cabezas Fernández95, María Eugenia Carrillo Gil95, Maria Paz Vento Martin96, Carmen Molina Pardines96, Nieves Orta Mira97, María Navarro Cots97, Inmaculada Vidal Catalá97, Isabel García Nava97, Soledad Illescas Fernández-Bermejo98,99, José Martínez-Alarcón98,99,
Marta Torres-Narbona98, Cristina Colmenarejo98, Lidia García-Agudo98, Jorge A. Pérez García98, Martín Yago López100, María Ángeles Goberna Bravo100, Victoria Simón García101, Gonzalo Llop Furquet101, Agustín Irazo Tatay101, Sandra Moreno-Marro101, Noelia Lozano Rodríguez101, Amparo Broseta Tamarit102, Juan José Badiola Díez103, Amparo Martínez-Ramírez104, Ana Dopazo105, Sergio Callejas105, Alberto Benguría105, Begoña Aguado106, Antonio Alcamí106, Marta Bermejo Bermejo107, Ricardo Ramos-Ruiz108, Victor Manuel Fernández Soria108, Fernando Simón Soria109 and Mercedes Roig Cardells110

98Servicio de Vigilancia y Control Epidemiológico. Dirección General de Salud Pública y Adicciones, Conselleria de Sanitat Universal i Salut Pública, Generalitat Valenciana, Valencia, Spain. 99Real Jardín Botánico, Consejo Superior de Investigaciones Científicas, Madrid, Spain. 100Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. 101Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. 102Cancer Ageing and Somatic Mutation Programme, Welcome Sanger Institute, Cambridge, UK. 103Servicio de Microbiología, H.U. Puerta del Mar, Cádiz, Spain. 104INIBICA, Instituto de Investigación Biomédica de Córdoba, Córdoba, Spain. 105Departamento de Biomedicina, Biotecnología y Salud Pública. Facultad de Medicina, Universidad de Cádiz, Cádiz, Spain. 106Laboratorio de Microbiología, Hospital San Pedro, Logroño, Spain. 10715Global, Barcelona Institute for Global Health, Hospital Clinic – Universitat de Barcelona, Barcelona, Spain. 108Centro de Investigación en Salud de Mancha (CISM), Maputo, Mozambique. 109Fight AIDS and Infectious Diseases Foundation, Hospital Germans Trias i Pujol, Barcelona, Spain. 110Llibre Medical Centre-International SOS, Lihir Island, Papua New Guinea. 111Servicio de Microbiología Clínica, Hospital San Pedro de Alcántara, Cáceres, Spain. 112Servicio Extremo de Salud, Sevilla, Spain. 113Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain. 114Servicio de Microbiología & CORE de Biología Molecular, CDB, Hospital Clínico, Barcelona, Spain. 115Department of Microbiology - CDB, Hospital Clinic de Barcelona, Barcelona, Spain. 116University of Barcelona, Barcelona, Spain. 117CINBIO, Universidad de Vigo, Vigo, Spain. 118Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, Vigo, Spain. 119Hospital General de Reina, Valencia, Spain. 120Hospital Universitario La Paz, IDiPaz, Madrid, Spain. 121Servicio de Microbiología, Hospital Arnau de Vilanova, Valencia, Spain. 122Biomolecular Research Institute of Barcelona (IIBB), Spanish National Research Council (CSIC), Barcelona, Spain. 123Hospital Universitario La Ribera, Alzira, Valencia, Spain. 124Hospital Universitario San Cecilio, Granada, Spain. 125Servicio de Microbiología, Hospital Arnau de Vilanova, Valencia, Spain. 126Biomedical Research Institute of Barcelona (IIBB), Spanish National Research Council (CSIC), Barcelona, Spain. 127Servicio de Microbiología, Hospital Arnau de Vilanova, Valencia, Spain. 128Hospital Universitario San José, Pontevedra, Spain. 129Servicio de Microbiología, Hospital Universitario Ramón y Cajal, Madrid, Spain. 130Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain. 131Red Española de Investigación en Patología Infectiosa (REIPI), Madrid, Spain. 132Centro Nacional de Gripe, Valladolid, Spain. 133Hospital Universitari de La Plana, Vila-real, Spain. 134Hospital Universitario de Guadalajara, Guadalajara, Spain. 135Laboratorio de Referencia de Cataluña, Barcelona, Spain. 136Hospital Universitari Arnau de Vilanova de Lleida, Lleida, Spain. 137Complejo Hospitalario Universitario Torrecárdenas, Almería, Spain. 138Servicio de Microbiología, Hospital General Universitario de Alicante. Instituto de Investigación Sanitaria y Biomática de Alicante (ISABIAL), Alicante, Spain. 139Hospital Francisco de Borja, Sección Microbiología, Valladolid, Spain. 140Hospital General Universitario de Ciudad Real, Ciudad Real, Spain. 141Facultad de Medicina de Ciudad Real. UCLM, Ciudad Real, Spain. 142Hospital General de Requena, Valencia, Spain. 143Laboratorio de Biología Molecular, Servicio de Análisis Clínicos y Microbiología, Hospital de Sagunto, Valencia, Spain. 144Sintlab - Hospital de Manises, Valencia, Spain. 145Centro de Enfermedades Transmisibles Emergentes, Facultad de Veterinaria, Universidad de Zaragoza, Zaragoza, Spain. 146Sección de Genómica Servicio Central de Soporte a la Investigación Experimental (SCSIIE), Universitat de València, Valencia, Spain. 147Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain. 148Centro de Biología Molecular Severo Ochoa (CBMSO) (CSIC-UAM), Nicolás Cabrera 1, Cantoblanco, Madrid, Spain. 149Unidad de Vigilancia de Salud y Medicina del Trabajo CSIC, Serrano 113p, Madrid, Spain. 150Fundación Parque Científico de Madrid, Madrid, Spain. 151Centro de Coordinación de Alertas y Emergencias Sanitarias, Ministerio de Sanidad, Madrid, Spain. 152Laboratorio de microbiología del Hospital Comarcal de Vinarós, Vinarós, 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 extracted 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 Ct 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 using the V3 multiplex primers scheme. 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. Whole genome sequencing was carried out using the MiSeq platform (2 × 200 cycles per sample).

The sequences obtained went through a bioinformatic pipeline based on IVARe, 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 Kraken²; (2) filtering of the fastq files using fastp v0.26.1 (ref. 1); (3) indexing the reads with the SARS-CoV-2 reference genome SARS-CoV-2 mapping. In short, the pipeline goes through the following steps: (1) Removal of the human reads with Kraken²; (2) filtering of the fastq files using fastp v0.26.1 (ref. 1); (3) indexing the reads with the SARS-CoV-2 reference genome SARS-CoV-2 v.1.2 (variant calling cut-offs: minimum quality for SNP calling = 20, minimum position depth = 20, 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.

Global alignment and phylogenetic reconstruction. To build the global alignment, we downloaded and concatenated all non-Spanish sequences present in GISAID on 21 June 2020 that passed strict filtering criteria: (1) sequences should be more than 29,000 bases 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 using MAFFT6. Specific positions that have been reported to be problematic for phylogenetic reconstruction were masked, following the procedure described by Lancianò using the Blasdan alignment script.

Finally, a maximum-likelihood (ML) phylogenetic reconstruction was performed using IQTREEm 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. We used 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 preceding 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, 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 was used for phylogenetic visualization.

SEC detailed analysis. To get more detail on SEC phylogenetic structure, and to evaluate if mobility restrictions were effective to hinder SEC transmission, we used the original SEC 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.

SECs potential superspreading events were defined as groups of more than ten sequences, having at least one SNP in common and having a within-region 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 Vicențy (ellipsoidal) method. Genetic diversity was calculated with two different methods: (1) genetic distance between each pair of samples in terms 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, 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.3.14.16-16P1, 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. 3). We implemented a coalescent Bayesian exponential growth model available in Beast 1.4 with the HKY + Γ model of nucleotide substitution. Tree priors were defined as follows: for effective population size we used a lognormal distribution (mean(μ) = 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/stopping stone analysis. Clock priors were defined as: ucld.mean: lognormal distribution with mean in real space = 1.4 × 10⁻¹⁵ subs per site per year and s.d. of the uncorrelated lognormal relaxed clock ucl.sdsvd = 5 × 10⁻¹⁵. Parameters were estimated using Markov chain Monte Carlo (MCMC) Bayesian inference, with 5 × 10⁶ steps-long chains with exception of SE7 and SE8, for which longer chains were run (1 × 10⁶); in all cases a total of 10⁶ steps were sampled in the log files. For all analyses, three independent runs were performed from different starting trees. 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. 4) 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
Phylogenetic analysis to estimate $R_t$. To estimate discrete changes in $R_t$ through the two largest epidemic clades SE7C, SE8C 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$^8$ implemented in BEAST v.2.6 (ref. 3). 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 $R_t$ is constant and estimated the date with most evidence for a change in $R_t$. To this end, we set a uniform prior distribution. $R_t$ 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 uninfected, $\delta = 36.5$ years$^{-1}$) in accordance with consistent global estimates of an infectious period of 10 days$^8$. 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/skilinetools/wiki). For this analysis, 1 x 10$^4$ and 4 x 10$^8$ steps-long chains were used for SEC7/B.1/B.1.1 and SE8C, respectively. Results were inspected with Tracer (v.1.7.1)$^9$ 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 ($\gamma = (R_t - \delta)$ and doubling time $= \ln(2)/\gamma$).

Statistical analysis. All statistical analyses were carried out using the R statistical language$^{10}$. Packages ape$^{11}$, treeio$^{12}$, doParallel$^{13}$ and foreach$^{14}$ were used for phylogenetic manipulation and analysis. We additionally used packages geosphere$^{15}$, igwgeom$^{15}$, sp$^{16}$, sf$^{17}$ and rgeos$^{18}$ to calculate the geographic distances between samples and the geographical representation in the data. The ggplot2 R package$^{19}$ 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$^{20}$ 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/fsabio-ngs/sars-cov-2-mapping.

References
38. Quick, J. nCoV-2019 sequencing protocol. protocols.io https://doi.org/10.17504/protocols.io.bmu6kw6 (2020).
39. arcit-network. arcit-network-arcit-n cov2019. github.com https://github.com/arcit-network-arcit-n cov2019 (2019).
40. Grubaugh, N. D. et al. An ampiclon-based sequencing framework for accurately measuring intrahost virus diversity using PrimaiSeq 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, X., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, 884−889 (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, X. & Wangnavirao 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).
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.I., 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.
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.
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 B.1.1). 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).
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-rgs/sars-cov2-mapping](https://gitlab.com/fisabio-rgs/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 (v.0.20.1), 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/](https://www.gisaid.org/) and the accession numbers can 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-list.pdf](http://nature.com/documents/nr-reporting-summary-list.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 accessible and the accession numbers can be found in Supplementary 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 SEVs, 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, 608 were in range 35-55, 615 were in range 55-75, 580 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 approved 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 possibilities. This were traduced into an heterogeneous sampling for each hospital, independently of the epidemic impact suffered in this hospital’s region. This heterogeneity 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.