ZOOLOGICAL RESEARCH

Phylogeography of SARS-CoV-2 pandemic in Spain: a story of multiple introductions, micro-geographic stratification, founder effects, and super-spreaders

Alberto Gómez-Carballa[1,2,3,#], Xabier Bello[1,2,3,#], Jacobo Pardo-Seco[1,2,3,#], María Luisa Pérez del Molino[4], Federico Martinón-Torres[2,3], Antonio Salas[1,2,3,*]

[1] *Unidade de Xenética*, *Instituto de Ciencias Forenses (INCIFOR)*, *Facultade de Medicina*, *Universidade de Santiago de Compostela*, *and GenPoB Research Group*, *Instituto de Investigación Sanitaria (IDIS)*, *Hospital Clínico Universitario de Santiago (SERGAS)*, *Galicia* 15706, *Spain*

[2] *Genetics*, *Vaccines and Pediatric Infectious Diseases Research Group (GENVIP)*, *Instituto de Investigación Sanitaria de Santiago (IDIS) and Universidade de Santiago de Compostela (USC)*, *Galicia* 15706, *Spain*

[3] *Translational Pediatrics and Infectious Diseases*, *Department of Pediatrics*, *Hospital Clínico Universitario de Santiago de Compostela (SERGAS)*, *Galicia* 15706, *Spain*

[4] *Servicio de Microbiología y Parasitología*, *Complejo Hospitalario Universitario de Santiago de Compostela*, *Santiago de Compostela*, *Galicia*, 15706 *Spain*

ABSTRACT

Spain has been one of the main global pandemic epicenters for coronavirus disease 2019 (COVID-19). Here, we analyzed >41 000 genomes (including >26 000 high-quality (HQ) genomes) downloaded from the GISAID repository, including 1 245 (922 HQ) sampled in Spain. The aim of this study was to investigate genome variation of novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and reconstruct phylogeographic and transmission patterns in Spain. Phylogeographic analysis suggested at least 34 independent introductions of SARS-CoV-2 to Spain at the beginning of the outbreak. Six lineages spread very successfully in the country, probably favored by super-spreaders, namely, A2a4 (7.8%), A2a5 (38.4%), A2a10 (2.8%), B3a (30.1%), and B9 (8.7%), which accounted for 87.9% of all genomes in the Spanish database. One distinct feature of the Spanish SARS-CoV-2 genomes was the higher frequency of B lineages (39.3%, mainly B3a+B9) than found in any other European country. While B3a, B9, (and an important sub-lineage of A2a5, namely, A2a5c) most likely originated in Spain, the other three haplogroups were imported from other European locations. The

Received: 26 August 2020; Accepted: 16 September 2020; Online: 16 September 2020

Foundation items: This study was supported by the Instituto de Salud Carlos III: project GePEM (Instituto de Salud Carlos III(ISCIII)/PI16/01478/Cofinanciado FEDER), DIAVIR (Instituto de Salud Carlos III(ISCIII)/DTS19/00049/Cofinanciado FEDER; Proyecto de Desarrollo Tecnológico en Salud) and Resvi-Omics (Instituto de Salud Carlos III(ISCIII)/PI19/01039/Cofinanciado FEDER) and project BI-BACVIR (PRIS-3; Agencia de Conocimiento en Salud (ACIS)—Servicio Gallego de Salud (SERGAS)—Xunta de Galicia; Spain) given to A.S.; and project ReSVinext (Instituto de Salud Carlos III(ISCIII)/PI16/01569/Cofinanciado FEDER) and Enterogen (Instituto de Salud Carlos III(ISCIII)/ PI19/01090/Cofinanciado FEDER) given to F.M.-T.

[#]Authors contributed equally to this work

*Corresponding author, E-mail: antonio.salas@usc.es

DOI: 10.24272/j.issn.2095-8137.2020.217

B3a strain may have originated in the Basque Country from a B3 ancestor of uncertain geographic origin, whereas B9 likely emerged in Madrid. The time of the most recent common ancestor (TMRCA) of SARS-CoV-2 suggested that the first coronavirus entered the country around 11 February 2020, as estimated from the TMRCA of B3a, the first lineage detected in the country. Moreover, earlier claims that the D614G mutation is associated to higher transmissibility is not consistent with the very high prevalence of COVID-19 in Spain when compared to other countries with lower disease incidence but much higher frequency of this mutation (56.4% in Spain vs. 82.4% in rest of Europe). Instead, the data support a major role of genetic drift in modeling the micro-geographic stratification of virus strains across the country as well as the role of SARS-CoV-2 super-spreaders.

Keywords: Covid-19; SARS-CoV-2; Genomics; Phylogeny; Phylogeography

INTRODUCTION

Coronavirus disease 2019 (COVID-19) results from infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This pathogen originated from a zoonotic event in mid-November 2019 (Ceraolo & Giorgi, 2020; Gómez-Carballa et al., 2020; Wong et al., 2020). The first COVID-19 case was reported in Wuhan (Hubei Province, China) at the end of 2019, with the World Health Organization (WHO) declaring the disease a global pandemic in early 2020 (WHO, 2020).

Spain has been severely stricken by the COVID-19 pandemic (https://ourworldindata.org), with more than 282 641 infected cases and 28 441 deaths (29 July 2020; https://www.mscbs.gob.es). The first documented Spanish case was a German tourist in La Gomera (Canary Islands) recorded on 31 January 2020, and the second case was from a British tourist in Palma de Mallorca (Balearic Islands). The first cases reported in continental Spain were not detected until 24 February. Subsequently, the number of patients increased very rapidly through community transmission, starting in mid-February, with important local outbreaks occurring in the Mediterranean Valencian Community (South East Spain), Catalonia (North East Spain), and Madrid (capital city, Center Spain). The Spanish government imposed a lockdown on 14 March 2020, followed by an immediate state of alarm on 15 March. The lockdown was accompanied by severe travel restrictions with other countries. The national epidemiological peak occurred on 26 March.

Analysis of SARS-CoV-2 genomes has helped establish a solid phylogeny of worldwide variation (Gómez-Carballa et al., 2020), estimate global genomic diversity patterns, and reconstruct movements of main coronavirus strains across the world (Forster et al., 2020a; Gómez-Carballa et al., 2020; Shen et al, 2020; Tang et al., 2020; Van Dorp et al., 2020; Yu et al., 2020). The same phylogeographic approach can be applied to study the dynamics of SARS-CoV-2 at a more local scale. This may help clarify the variable impact of the pandemic in affected countries and how the pandemic has evolved at a micro-geographic level, i.e., within a country. In Spain, for instance, Madrid and surrounding provinces of Castilla-La Mancha have experienced a concentrated proportion of national cases; however, it is unknown if these strains are different or identical to those in other important centers of infection such as the Basque Country, Valencian Community, or Catalonia. Thus, analyses of the different strains are mandatory in order to establish any potential associations with the seemingly disproportionate impacts of the pandemic and virus spread.

With the availability of SARS-CoV-2 genomes in public repositories, in particular GISAID (https://www.gisaid.org; (Shu & Mccauley, 2017)), we now have the unique opportunity to explore the phylodynamics of the pandemic at both the global and regional scale. Furthermore, in a pandemic scenario, it is useful to investigate genomes from outside the regions of interest. In particular, the wide availability of genomes sampled from other countries offers an excellent framework to reconstruct the different ways a given virus strain has entered such countries and its relationship with strains in other regions. Furthermore, we can obtain precise chronologies of the main events and pathways of these transmissions.

The present study aimed to explore high-quality (HQ) SARS-CoV-2 genomes (characterized by high coverage and >29 kb in length; downloaded from GISAID) sampled in Spanish patients (n>1 200), using worldwide genome sequences (n>41 000) as a framework for phylogeographic interpretation. Based on these data, we detected multiple spatial-temporal introductions of the virus to the country, estimated the effective population size of the main coronavirus lineages in Spain, and inferred the role of genetic drift in virus spread as well as the potential role of super-spreaders in Spanish transmission patterns. To the best of our knowledge, the substantial number of genomes processed in this study, aimed at reconstructing phylogeographic patterns of the coronavirus in a specific region, has no precedent in the literature.

MATERIALS AND METHODS

Samples

A total of 41 362 SARS-CoV-2 genomes, including 26 506 HQ genomes, were downloaded from GISAID (https://www.epicov.org/epi3/frontend) on 12 June 2020. To reduce noise in the analyses originating from sequencing errors, we only used HQ genomes in all calculations and inferences, unless otherwise specified. Sequence ambiguities were eliminated to further reduce background noise. To make phylogenetic patterns clearer, we also removed insertions and multi-nucleotide polymorphisms (MNPs) from the analyses. Thirteen additional genomes (nine HQ and four low-quality (LQ)

genomes) had no information on country or continent exposure and were thus excluded.

The genomes obtained from GISAID represented samples collected across 90 countries (87 for HQ data) from the following continental regions: Africa (*n*=395 (188 HQ); 12 countries (11 in HQ genomes)); Asia (*n*=3 080 (2 674 HQ); 31 countries (30 in HQ genomes)); Europe (*n*=27 188 (14 953 HQ); 32 countries (32 in HQ genomes)); North (and Meso) America (*n*=8 125 (6 701 HQ); five countries (five in HQ genomes)); Oceania (*n*=2 069 (1 621 HQ); three countries); South America (*n*=492 (360 HQ); seven countries (six in HQ genomes)).

Spain was represented in this large dataset with 1 245 samples (922 HQ). Genomes were analyzed from patients sampled in the following Spanish regions: Andalusia (*n*=219 (139 HQ)); Asturias (*n*=6 (6 HQ)); Balearic Islands (*n*=4 (3 HQ)); Basque Country (*n*=274 (223 HQ)); Canary Islands (*n*=10 (10 HQ)); Castilla La Mancha (*n*=3 (3 HQ)); Castilla y Leon (*n*=4 (4 HQ)); Catalonia (*n*=35 (34 HQ)); Galicia (*n*=45 (37 HQ)); La Rioja (*n*=12 (12 HQ)); Madrid (*n*=161 (146 HQ)); Melilla (*n*=6 (6 HQ)); Navarra (*n*=63 (42 HQ)); and the Valencian Community (*n*=404 (245 HQ)). A few Spanish genomes were sampled at the beginning of the pandemic (end of February; *n*=10), but most were sampled from mid to late March, around the peak of the pandemic in Spain (26 March) (https://cnecovid.isciii.es; Supplementary Figure S1). All Spanish samples were considered together when carrying out analyses related to the country; however, areas with sample sizes below 34 (size of the Catalonian dataset) were not considered for regional analyses due to their low values.

All genomes were aligned against a reference sequence from GenBank (accession No. MN908947.3; GISAID ID #402125). Meta-data containing information on the geographic location of the sample (city, country, territory, area, and continental region) and sampling date were downloaded from https://gisaid.org.

Alignment of FASTA sequences was carried out manually. The genomes were trimmed for the 5' and 3' untranslated regions to retain a 29 774 bp consensus sequence that ran from position 55 to position 29 829 in the SARS-CoV-2 genome.

Statistical analysis

We computed two diversity indices from the SARS-CoV-2 genome data: i.e., average number of nucleotide differences per site between DNA sequences or nucleotide diversity (π) (Nei & Li, 1979), and sequence/haplotype diversity (*HD*). Maps of haplogroup frequencies were built using SAGA v.7.6.2 (http://www.saga-gis.org/) (Conrad et al., 2015) and the ordinary Kriging method. When the sampling size was below 10, we collapsed neighboring geographic points whenever possible.

We constructed Extended Bayesian Skyline Plots (EBSPs) (Heled & Drummond, 2008) to infer the demography of SARS-CoV-2 genomes of the main Spanish haplogroups using BEAST v.2.6.2 (Drummond & Rambaut, 2007). EBSP analysis can infer effective population size (N_e) through time. Analyses were carried out as per recent research (Gómez-Carballa et al., 2020). The time of the most recent common ancestor (TMRCA) of the main Spanish lineages was estimated using a coalescent model with exponential growth in BEAST, with a reference sequence from Wuhan (GISAID ID #402125) used as an outgroup. We excluded samples carrying reverse mutations from analysis to avoid problems with the location of ancestral haplotypes. For both EBPS and TMRCA analyses, we used a strict-clock and an evolution rate of 0.80×10^{-3} (0.14×10^{-3}–1.31×10^{-3}) s/s/y (substitutions per site per year) (http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356; comparable to the one reported by others, e.g., Gómez-Carballa et al., 2020), with two independent Markov chain-Monte Carlo runs (200 000 000 steps, samples taken every 1 000 steps, and 10% discarded as burn-in). We used Tracer v.1.6 (Drummond & Rambaut, 2007) to explore distribution convergence; the runs were combined using *LogCombiner* v.1.8.2 (Drummond & Rambaut, 2007).

The maximum clade credibility tree was visualized and edited using *FigTree* v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenetic analysis

A maximum parsimony tree was built using the phylogenetic skeleton and haplogroup nomenclature from Gómez-Carballa et al. (2020) (Figure 1); thus, samples were classified into clades and sub-clades by comparing genome variants with diagnostic sites of haplogroups in the reference classification tree. New haplogroups (38 within haplogroup A and 29 within haplogroup B; Figure 1) were created when at least one mutation difference existed between a minimum of two samples belonging to the same haplogroup.

We analyzed topological features of the phylogenetic trees for the main Spanish haplogroups (Colijn & Gardy, 2014) to distinguish patterns of super-spreader transmissions from those of homogeneous and chain transmissions. Thus, the best candidates for genomes transmitted by super-spreaders corresponded to those haplotypes that reached high frequency in particular cities or localities in a short time period (Gómez-Carballa et al., 2020).

For this purpose, we built phylogenetic trees from sequence alignments of haplogroups represented by the super-spreader candidates using *SplitsTree5* software (Huson & Bryant, 2006). The R library *phyloTop* (Kendall et al., 2018) was then used to calculate tree features that showed high values in transmission models mediated by super-spreaders (Colijn & Gardy, 2014), including normalized average ladder (mean size of ladders in tree/(*n*–2)), *Colless* index or imbalance, *IL-number* (portion of internal nodes with a single leaf descendant), maximum height (maximum number of steps from root), *Sackin index* or *imbalance* (mean path length from tip to root), staircase-ness 1 (portion of imbalanced nodes), as well as other indices that showed low values in this model of transmission, including *cherry number*/*n* (number of cherries over number of leaves), *pitchforks* (number of nodes with

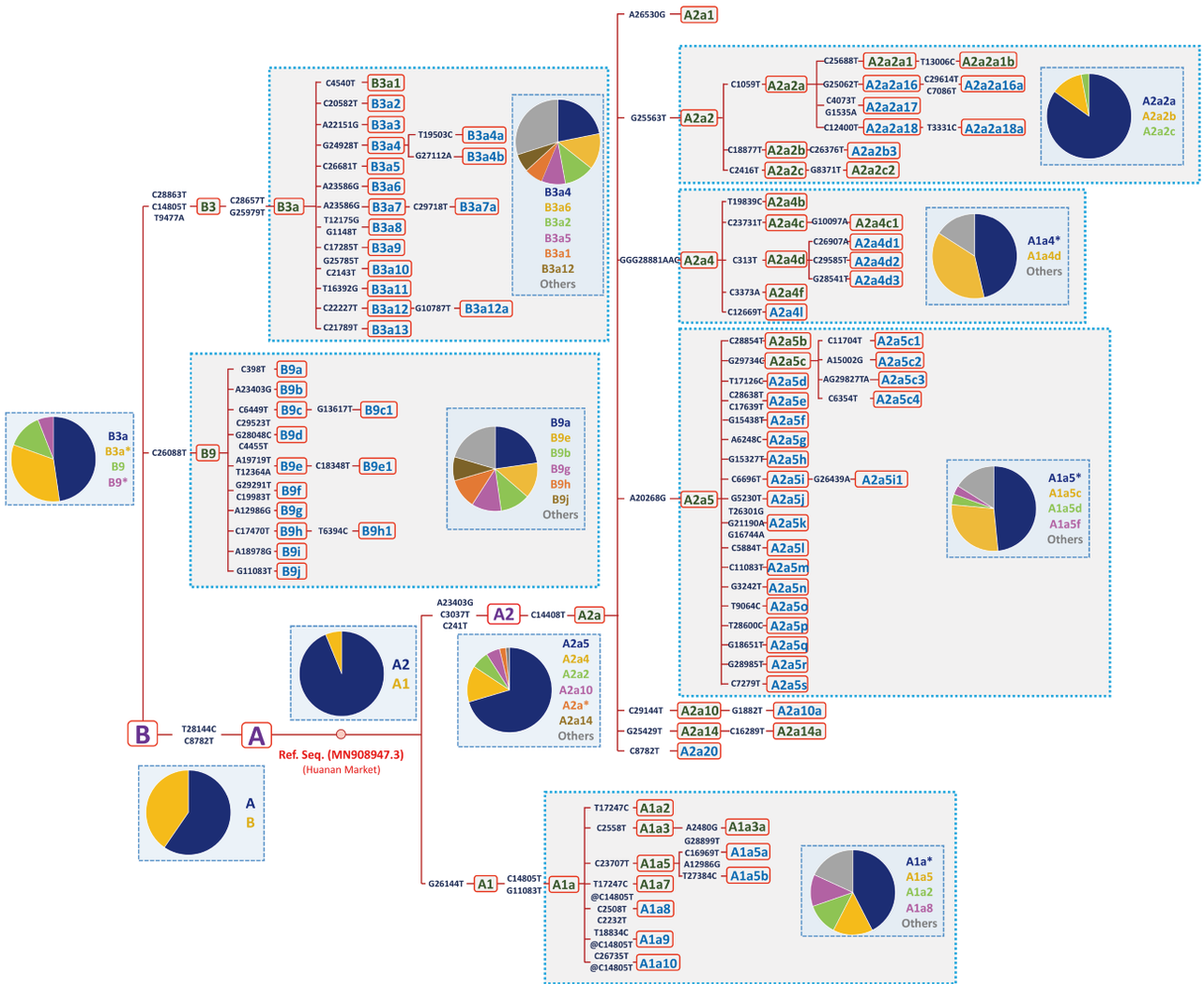

Figure 1 Skeleton of maximum parsimony tree of Spanish genomes

Pie charts show haplogroup frequency distribution for main clades. Haplogroup labels in green refer to previously described clades (Gómez-Carballa et al, 2020), in blue are newly defined in present study. Mutations along branches are nucleotide changes against reference sequence. Mutations with a @ symbol indicate reversions. The * symbol after the name of the haplogroup in the piecharts refers to those sequences that can not be clasified into any of the sub-haplogrups included by the specified haplogroup.

three tip descendants (Metzig et al., 2019)), and *staircase-ness 2* (defined as the portion of imbalance nodes; average of $min(r_i,s_i)/\max(r_i,s_i)$ over internal nodes) (Norström et al., 2012). To summarize these indices into a single score, considering that all range from 0 to 1, we used a simple approach that involved: (*i*) calculating the average of indices characteristically high (*H*) and low (*L*) in super-spreader host transmissions, respectively, and then (*ii*) computing the *H/L* ratio, which will be higher (close to 1) in scenarios of super-spreader transmissions.

Finally, we also built median joining networks (Bandelt et al., 1999) using *POPart* (Leigh & Bryant, 2015) for super-spreader candidates to better visualize the star-like shape that is characteristic of this mode of transmission, which differs from the pattern generated by homogeneous and chain transmissions (Colijn & Gardy, 2014). In a super-spreader model of transmission, the basal node would be occupied by the SARS-CoV-2 strain mainly transmitted by the super-spreader host(s); in other words, the super-spreader and people infected in primary transmissions (i.e., directly infected by the super-spreader) would tend to share the same viral genome sequence. On the other hand, the derivative sequence genomes emerging from this central node (making the star-like shape of the phylogenies) would correspond to secondary transmissions (those transmitted by primary infected people): due to the incubation time needed for the development of COVID-19 symptoms (according to WHO (8 November 2020), 5–6 days, but up to 14 days), coupled with

the fact that SARS-CoV-2 accumulates, on average, one mutation every two weeks (according to its evolution rate), haplotypes in the secondary transmissions would tend to have a one step-mutation difference with respect to the node haplotype. Last but not least, according to the super-spreader theory, primary transmissions (identical haplotypes) should occur over a very short time period (about one week), whereas homogeneous initial transmissions would require at least two or more weeks. For this reason, we also estimated the lifespan of a strain using, as a proxy, the chronological ages of identical genomes in the basal nodes.

RESULTS

Molecular diversity of Spanish SARS-CoV-2 genomes

Of the 41 362 genomes downloaded from GISAID, 26 506 (64.1%) were labeled as HQ and 14 856 (35.9%) as LQ. The HQ samples had, on average, slightly higher genome length, and fewer genomic islands, ambiguities, and indels than the LQ genomes (Supplementary Figure S2). For these HQ genomes, we observed a total of 1 378 insertions, 763 MNPs, and 14 291 substitutions (Table 1).

There were 1 259 Spanish genomes in the database, 922 of which were labeled as HQ (74.3%). We observed 40 indels, 12 MNPs, 216 ambiguities, and 1 000 (769 unique) substitutions (Table 1). There were 5 184 different haplotypes; 438 of which were found only once in the database, 44 twice, and 36 more than three times. A few were found at very high frequencies (Table 2).

All haplotypes were classified into haplogroups according to the phylogeny described in Gómez-Carballa et al. (2020) and the extended nomenclature of the present study. The most frequent haplogroups in Spain were A2a5 (*n*=354; 38.4% of Spanish haplotypes; diagnostic variants: C241T–C3037T–C14408T–A20268G–A23403G), A2a4 (*n*=72; 7.8%; diagnostic variants: C241T–C3037T–C14408T–A23403G plus characteristic MNP: GGG28881AAC), A2a10 (*n*=26; 2.8%; diagnostic variants: C241T–C3037T–C14408T–A23403G–C29144T), B3a (*n*=278; 30.2%; diagnostic variants: C8782T–T9477A–C14805T–G25979T–T28144C–C28657T–C28863T), and B9 (*n*=80; 8.7%; diagnostic variants: C8782T–C26088T–T28144C); together, these haplogroups made up 87.9% of the whole database (*n*=810). Spain was the only European country where haplotypes belonging to macro-haplogroup B reached very high frequency (*n*=362; 39.3%), in comparison to the European average (Europe: *n*=514 B haplotypes (*n*=151 without Spain); 3.4% (1.1% without Spain)), where macro-haplogroup A was much more predominant.

Within Spain, the distribution of the main haplogroups was far from homogeneous. For instance, haplogroup A2a5 made up 88.2% of samples from Catalonia, but only 16.6% of those from Basque Country, whereas haplogroup B3a made up 67.3% in Basque Country but only 2.9% in Catalonia and 4.8% in Navarra. Moreover, while some haplogroups, e.g., A2a5, were highly dispersed across the Iberian Peninsula, others were highly concentrated in a particular region (e.g., 96.2% of all A2a10 haplotypes occurred in the Valencian Community) (Supplementary Figures S3–S10).

Table 1 Characteristics of SARS-CoV-2 genome database used in present study

	Worldwide	Spain
Total genomes in database		
n	41 362	1 259
Ambiguities	12 646	553
Indels	1 378	57
MNPs	763	17
Different haplotypes	19 968	680
Singleton haplotypes	15 934	565
Substitutions	14 291	1 000
Singleton substitutions	7 189	746
HQ genomes in database		
n	26 506	922
Ambiguities	4 936	216
Indels	795	40
MNPs	318	12
Different haplotypes	13 559	518
Singleton haplotypes	10 942	438
Substitutions	10 663	788
Singleton substitutions	5 313	609
LQ genomes in database		
n	14 856	337
Ambiguities	9 340	351
Indels	873	21
MNPs	474	6
Different haplotypes	8 141	209
Singleton haplotypes	6 709	178
Substitutions	8 408	353
Singleton substitutions	4 802	275

HQ: High-quality; LQ: Low-quality.

Multiple introductions of SARS-CoV-2 to Spain

In total, 97 different haplogroups and sub-haplogroups have been identified in the Spanish territory. Some emerged *de novo* within Spain by accumulation of a mutation on top of an ancestral haplotype (e.g., B3(Spain?)>B3a(Spain); see below). We obtained a simple rough estimate of the minimum number of haplogroups with a likely geographic origin in Spain: i.e., among the 97 possible categories, we considered those with their first representative outside Spain (according to the known sampling chronology of genomes), with the exception of B3a, which likely originated in Spain despite two earlier representatives in France (see below). Such an approach is not without risks due to known variability, e.g., incubation and sampling periods, and potential errors in GISAID, e.g., reported sampling dates; however, it can provide an indicative figure. According to this estimation, there were 34 independent introductions of SARS-CoV-2 to Spain from abroad, while 63 lineages likely originated within country

Table 2 Regional distribution of most frequent haplotypes sampled in Spain

ID	n_T	n_H	#H1	#H2	#H3	#H4	#H5	#H6
Haplogroup			B3a	A2a5	A2a5c	A2a10	B9	A2a4
Spanish region								
Andalusia	139	31 (22.3)	4 (2.9)	18 (12.9)	7 (5.0)	–	2 (1.4)	–
Basque Country	3	1 (33.3)	–	1 (33.3)	–	–	–	–
Balearic Island	223	92 (41.3)	75 (33.6)	10 (4.5)	1 (0.4)	–	6 (2.7)	–
Catalonia	33	9 (27.3)	–	7 (21.2)	2 (16.1)	–	–	–
Canary Islands	9	2 (22.2)	1 (11.1)	–	–	–	–	1 (11.1)
Castilla Leon	4	2 (50.0)	2 (1.4)	–	–	–	–	–
Castilla La Mancha	3	1 (33.3)	–	–	–	1 (33.3)	–	–
Galicia	37	10 (27.0)	–	8 (21.6)	–	–	–	2 (5.4)
La Rioja	12	2 (16.7)	2 (16.7)	–	–	–	–	–
Madrid	143	31 (21.7)	–	14 (9.8)	16 (11.2)	–	1 (0.7)	–
Navarra	42	11 (26.2)	–	5 (11.9)	5 (11.9)	–	1 (2.4)	–
Valencian Community	244	83 (34.0)	29 (11.9)	16 (6.6)	13 (5.3)	15 (6.1)	4 (1.6)	7 (2.9)
Total	920	277 (30.1)	113 (12.3)	80 (8.7)	44 (4.8)	16 (1.7)	14 (1.5)	10 (1.1)
Continental region								
Africa	188	11 (5.9)	–	–	–	–	–	11 (5.9)
Asia	2 674	69 (2.6)	9 (0.3)	4 (0.1)	2 (0.1)	–	–	54 (2.0)
Europe (excluding Spain)	14 031	1 068 (7.6)	13 (0.1)	91 (0.6)	27 (0.2)	8 (0.1)	4 (0.0)	925 (6.2)
North America	6 701	108 (1.6)	2 (0.0)	7 (0.1)	1 (0.0)	–	3 (0.0)	95 (1.4)
Oceania	1 621	73 (4.5)	–	7 (0.4)	1 (0.1)	–	1 (0.1)	64 (3.9)
South America	360	43 (11.9)	7 (1.9)	19 (5.3)	2 (0.6)	1 (0.3)	–	14 (3.9)
Total (excluding Spain)	25 575	1 372 (5.4)	31 (0.1)	128 (0.5)	33 (0.1)	9 (0.0)	7 (0.0)	1 163 (4.5)
Total (including Spain)	26 497	1 485 (5.6)	144 (0.5)	208 (0.8)	33 (0.3)	25 (0.1)	22 (0.1)	1 173 (4.4)

Haplotype IDs refers to: #H1=C8782T–T9477A–C14805T–G25979T–T28144C–C28657T–C28863T; #H2=C241T–C3037T–C14408T–A20268G–A23403G; #H3=C241T–C3037T–C14408T–A20268G–A23403G–G29734C; #H4=C241T–C3037T–C14408T–A23403G–C29144T; #H5=C8782T–C26088T–T28144C; and #H6 =C241T–C3037T–C14408T–A23403G. Other abbreviations: n_T: Total sample size; n_H: Sample size of unique haplotypes. Brackets show frequency of corresponding haplotype in each region. Total samples from Spain do not match sum of values of Spanish regions because 12 sequences did not have state information. –: No sequences available.

based on the accumulation of "domestic" mutational changes.

By examining genome chronologies of Spanish lineages, 55 of the 97 (56.7%) different haplogroups appeared from 25 February (see below) to the Spanish lockdown on 15 March, whereas 42 appeared after. In addition, 27 out of 42 genomes emerging after 15 March (64.3%) originated within Spain, with the remaining 15 (35.7%) potentially introduced from abroad.

It is important to note that the 15 new clades presumably arriving to Spain after the lockdown represented only 5.2% (*n*=48) of total Spanish haplotypes. Therefore, it is most likely that the impact of the new introductions after the lockdown was well below this figure of 5.2% if we assume that many of the genomes belonging to these clades evolved within country after their first entrance (summing to a total of 48 genomes).

Basque Country (Spain) is best candidate region for origin of haplogroup B3a

There were 495 B3a genomes worldwide in the database, most of which were sampled in Europe (354/495; 71.5%), representing 2.4% of lineages in the continent, and Oceania (99/495; 20.0%), representing 6.1% of genomes in this geographic area. The first B3a haplotype worldwide was sampled in Europe on 25 February 2020 and outside Europe on 3 March (at which time there were 23 B3a genomes in the European dataset). In the Asian dataset, the first B3a genome appeared about a week after the first European appearance (8 March; at which time Europe had 76 instances in the dataset), and the second genome more than two weeks later (20 March, with 275 instances in the European dataset).

The first and second B3a genome emerged in France (Grand East) on 25 and 26 February 2020, respectively, and almost simultaneously in Spain, with two genomes on 27 February 2020. This haplogroup did not appear in the French dataset again; meanwhile, Spain accumulated 278 B3a genomes, especially in the period from 5 to 19 March 2020 (*n*=222 cases), and reached the highest frequency worldwide (30.8% of total Spanish cases), with the exception of Kazakhstan (35.8% of all Kazakhstani cases; however, its first B3a genome was sampled on 25 March and most cases appeared from 14 April to 14 May) (Figures 2, 3). Of note, the two French cases were identical and were one-step mutation derivatives of the basal B3a lineage (transition C25553T on top of B3a sequence motif; Figure 4), whereas the first Spanish genomes matched the root of this haplogroup (see

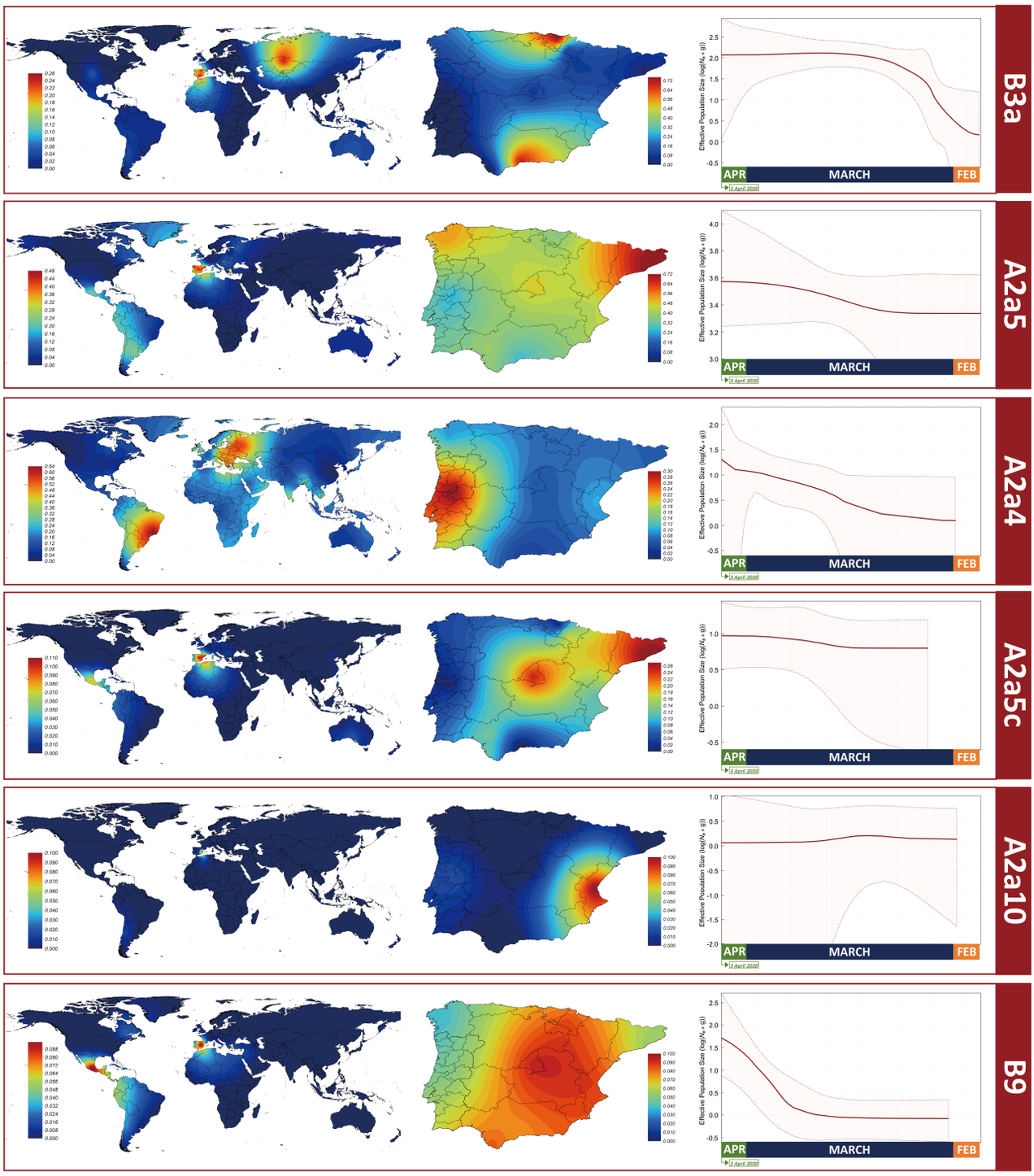

Figure 2 Interpolation maps of haplogroup frequencies worldwide and for Iberia, and extended Bayesian skyline plots (EBSPs) for main Spanish haplogroups

Sampling points considered for interpolation are given in map of Supplementary Figure S11.

below). B3a spread very little across Europe, with only a few instances in Denmark, France, Greece, Luxembourg, the Netherlands, Poland, Portugal, and the UK (together *n*=74), all appearing during or after the main B3a outbreak in Spain. Most European B3a examples may have been imported from Spain, e.g., those from neighboring Portugal (*n*=9; three initial

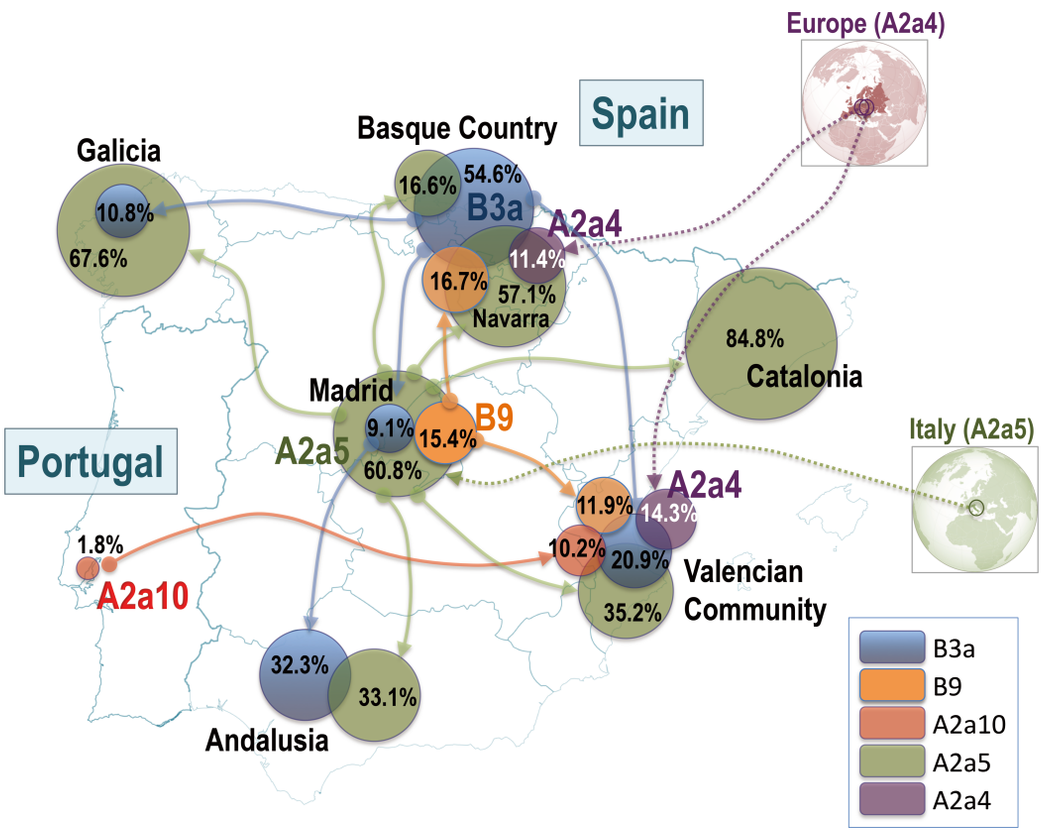

Figure 3 Schematic representation of reconstructed movements of main Spanish SARS-CoV-2 clades in Iberian Peninsula according to phylogeographic information inferred from phylogenies and genome chronologies

Areas of circles are proportional to frequencies. Note, frequency of each haplogroup in different regions should not be interpreted as a proportion of genomes received from any particular region because frequency of each haplogroup in a given region may or may not have been favored by local super-spreading events and therefore strongly depends on regional circumstances.

ones on 13 March). In Australia, the first B3a sequence occurred on 16 March, followed by an important outbreak from 20 to 25 March (65 genomes in only five days); in Australia, 6.7% of genomes belonged to B3a; at the time when the first B3a genome was detected in this country, only Spain was experiencing an important outbreak of this lineage. There is evidence of one instance of B3a appearing in Asia (Vietnam; *n*=1) dated to 20 March. Spain may also have exported a few B3a representatives to Latin America due to its high connection with Spanish and Portuguese-speaking Latin American countries: e.g., Brazil (*n*=2), Chile (*n*=10), Colombia (*n*=1), Mexico (*n*=1), and Uruguay (*n*=1).

We further investigated the most likely origin of B3a within Spain according to chronologies. The highest frequency of B3a occurred in the Basque Country (*n*=150 instances; 54.0% of all B3a Spanish genomes), followed by Andalusia (*n*=45; 12.3%), and the Valencian Community (*n*=51; 18.3%) (Figures 2, 3). The first three B3a genomes in the database appeared in the Valencian Community on 27–29 March. After examining the LQ B3a genomes, we observed other haplotypes also appearing in the Valencian Community on 26 and 27 March. Three instances of B3a appeared in the Basque Country two days after the first Valencian case (29 March); at this time, the Valencian Community had accumulated six B3a representatives (five with the basal sequence B3a motif and one with an extra mutation). Although the genomes from the Basque Country (three HQ and one LQ) appeared in the database two days later than the Valencian ones, remarkably, one (GISAID #455350; HQ) had five mutations on top of the B3a root and another (GISAID #455351; HQ) had one extra mutation to the basal motif, apart from representatives of the basal motif. This suggests that B3a could have been circulating in the Basque Country before it reached the Valencian Community; a lead time of a few weeks would be needed to explain the generation of such a high level of variation. In agreement with this hypothesis, the Basque Country accumulated B3a genomes very quickly and some instances appeared almost simultaneously in the neighboring regions of La Rioja and Castilla-Leon (Burgos). Furthermore, the Valencian Community had fewer instances during the first two weeks (49 and eight genomes in the Basque Country and Valencian Community, respectively, on 8

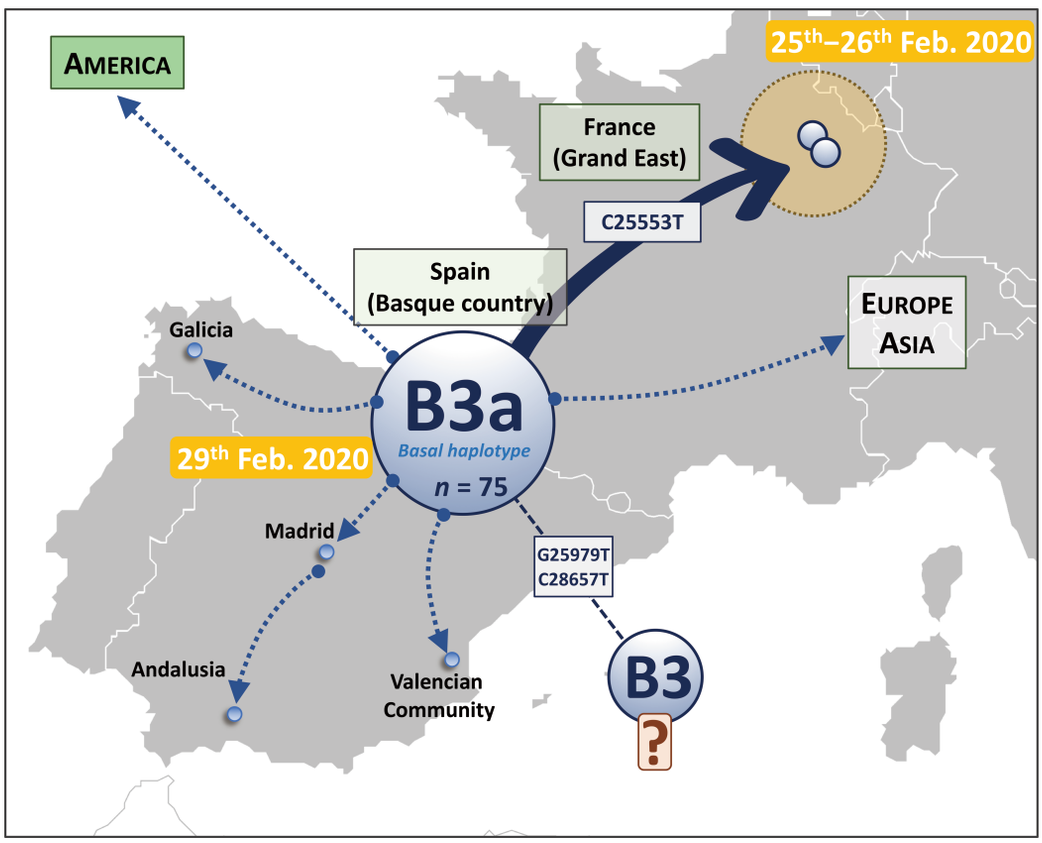

Figure 4 Origin and spread of Spanish haplogroup B3a

We highlight arrow connecting Basque Country (Spain) with two identical SARS-CoV-2 genomes from France because these two were sampled before the Spanish ones but were derivatives of basal B3a haplotype (see main text). Question mark on B3 indicates doubts on its origin. From a likely origin of B3a in the Basque Country, B3a moved to other Spanish regions, as well as to other American, European, and Asian locations.

March). This suggests that B3a may have been circulating more widely in the former than the latter.

Important outbreaks of B3a occurred in the Basque Country during 5 to 14 March (*n*=92 cases) and 16 to 19 March (*n*=44 genomes), whereas the main outbreak occurred later in the Valencian Community at around 9 to 10 March (*n*=25 genomes). The epicenter of the Basque Country B3a dispersion was around the city of Vitoria (see below). Most Andalusian cases accumulated during 11 to 17 March (27 out of 41), and the few cases occurring in Galicia (*n*=4) appeared in parallel to the initial Basque Country and Valencian Community outbreaks (9 to 10 March). The incidence of B3a in Madrid was minor compared to other lineages, with only 13 instances between March and April.

The basal motif of B3 only appeared in a single sequence sampled in Madrid on 3 March (GISAID: #417981), i.e., after the first sampling of B3a. There were a few sequences with a partial sequence motif of B3, one of which was sampled in Shanghai on 1 February 2020. There are, therefore, reasonable doubts regarding the origin of B3, but we know that at least one member of B3 was living together with B3a members in Spain. Basal B3 representatives could have disappeared from Spain very quickly, but not before producing one of the most successful lineages in the country: i.e., B3a. Thus, this sub-clade could have originated in Spain from a B3 ancestor. More precisely, and notwithstanding some uncertainty, B3a could have originated in the Basque Country (with the Valencian Community being another good candidate for its geographic origin).

The ancestral haplotype of B3a (#H1: C8782T–T9477A–C14805T–G25979T–T28144C–C28657T–C28863T; Table 2) occurred 113 times in Spain (first introduction on 27 March), 75 of which occurred in the Basque Country. Conversely, there were only 31 instances outside Spain, the first of these in Chile (3 and 6 March) and the USA (4 March).

EBSP analysis indicated that B3a experienced the earliest and most rapid and explosive growth among all SARS-CoV-2 lineages in Spain. It started at the end of February, coinciding with its first appearance in the Basque Country, and spread for at least 10 days until reaching a plateau around 15 March when the Spanish lockdown started. The rapid initial growth fits well with the proposition of an outbreak starting in the Basque Country, which overlapped with later outbreaks emerging in the Valencian Community and subsequently in

Andalusia (Figures 2, 3).

Most likely Spanish origin of haplogroup B9

B9 is a minor clade, mostly present in Europe (122 out of 151 total B9 genomes in the database; 80.8%), with low frequency in South America (17/151; 11%). The initial two B9 instances appeared in Spain on 28 February, growing to a total of 80 (65.6% of all European cases). The B9 genomes observed in Portugal (8 out of 501 total genomes (1.6%)) and Latin America (Uruguay: 6/13 (46.2%); Mexico 3/22 (13.6%); Colombia: 4/85 (4.7%); Chile: 7/153 (4.6%)); Figures 2, 3) may have been imported from Spain.

Almost simultaneously, sporadic instances of B9 appeared in regions such as Andalusia, Madrid, and neighboring Castilla-La Mancha (from 28 February to 3 March). Madrid accumulated a total of 22 cases (22 out of 77 B9 Spanish genomes) (Figures 2, 3) faster and sooner than other regions, including Navarra (*n*=7; 16.7% of total genomes in the region) and especially the Valencian Community (*n*=29; 11.9%), where all B9 genomes were sampled from 16 March onwards. The root haplotype of B9 (#H5: C8782T–C26088T–T28144C; Table 2) appeared in the Spanish database 14 times (first in Madrid) and nine times in other locations abroad (e.g., 23 February in Australia, and 10 and 13 March in Mexico).

Overall, the data suggest a Spanish origin for B9, with Madrid as the best candidate region. It seems clear from the data that B9 incubated in Madrid more than in other regions at the beginning of its expansion, followed by spread to the rest of the country and an important outbreak in the Valencian Community from 16 March onwards.

The effective population size of the B9 genome in Spain was relatively constant for a long period, spanning the end of February to mid-March, as indicated by EBSP analyses (Figures 2, 3). This phase would fit with an incubation period of B9 in Madrid. However, this clade experienced rapid and continuous growth, which coincided with the Valencian Community outbreak occurring from 16 March to the end of that month (Figures 2, 3).

Possible importation of haplogroup A2a5 from Italy to Spain

A2a5 appeared for the first time in the database in Europe on 26 February 2020, with five instances by 28 February when the first sample emerged in another continent (North America). Subsequently, it started to spread very quickly in Europe from 7 March to the end of April (*n*=1 105 genomes) and more cases appeared until the end of May (*n*=123), making up 8.3% of total genomes sampled in Europe. The main A2a5 occurrences outside Europe roughly coincided with the initial European outbreak. It reached a higher frequency in South America (14.4%; e.g., 28.6% of total genomes in Mexico; 22.7% in Argentina; 19% in Chile) compared to that in Oceania (3.8%) and North America (1.0%). The first Asian A2a5 case emerged on 19 March, which was most likely imported from Europe to Asia (19 total cases across several countries, e.g., China, Japan, Singapore, and Thailand) (Figures 2, 3).

A2a5 representatives occurred in 44 different countries, mostly in Europe. The first instances of this clade in the database were recorded in Italy on 26 February, with six cases in Italy by 4 March (3.4% of all lineages in Italy). In this period, only sporadic A2a5 genomes appeared in Switzerland (two on 27 February), Belgium (one on 2 March), the Netherlands (two on 1 March), and USA (two on 28 February and 1 March). The first Spanish case appeared on 4 March, and it was here that A2a5 underwent the most important outbreak worldwide from 4 March to 15 April (*n*=345 cases), with peaks on 25 to 28 March (83 genomes in four days) and 1 to 2 April (45 genomes sampled in two days). The A2a5 genomes represented 38.3% of total genomes sampled in the country. This haplogroup also successfully spread to Iceland (95 out of 359 total country cases (26.5%); 80 occurring between 12 to 20 March) and Chile (29 out of 153 total Chilean genomes, with more scattered occurrence from 9 March to 4 April). In the UK, this lineage accumulated continuously over a long period from 17 March (when the A2a5 outbreak was already set in Spain) to 25 May, reaching a total of 540 cases (6.8% of all lineages in UK). The A2a5 genomes observed in Portugal (*n*=37 from 13 March to 3 April; 7.4% of all cases in Portugal) could have been imported from Spain, taking into account the geographic proximity and high incidence of this strain in the Spanish territory.

Within Spain, the first A2a5 genomes were sampled in Madrid (4 March), which increased to 26 from 4 to 12 March. During this period, five cases also appeared in the Valencian Community. In Madrid, A2a5 represented 88 out of 146 total cases in the region (60.8%). Considering the sampling chronology, this sub-lineage could have arrived at other Spanish regions later, accounting for a higher proportion of total cases in Catalonia (85.6%; mainly from 15 to 28 March) and Galicia (67.6%; mostly appearing on 7 to 15 April), but lower proportion in Navarra (57.1%; all occurring from 12 to 23 March), Valencian Community (35.5%; from 6 March to 1 April), and Andalusia (33.1%; 12 March to 1 April) (Figures 2, 3).

Overall, A2a5 may have evolved from a European A2a lineage, potentially from Italy because it is (i) the first country where the first A2a5 genome was sampled, (ii) the earliest country within Europe suffering considerable impact from COVID-19, (iii) highly inter-connected with Spain, and (iv) the most likely origin of its immediate phylogenetic ancestral node, A2a, by far the most prevalent European clade (Gómez-Carballa et al., 2020). A2a5 moved to different regions in Europe, including to Madrid at the end of February. The first Italian A2a5 genome corresponding to the root haplotype (#H2: C241T–C3037T–C14408T–A20268G–A23403G; Table 2) appeared on 26 February. Almost simultaneously, it was also sampled in other European regions (e.g., Switzerland on 27 February), with particularly high prevalence in the Spanish dataset (80 genomes, starting on 4 March). As the first genomes were sampled in Madrid (and subsequently in Galicia, followed by the Valencian Community), Madrid probably constitutes the main focus of A2a5 dispersion to

other regions in the country and to neighboring Portugal (*n*=9). A2a5 also spread to other regions in South/Meso- America that are highly connected with Spain, namely, Panama (*n*=1), Argentina (*n*=7; 26.9%), Chile (*n*=29; 19.0%), Colombia (*n*=16; 18.8%), and Mexico (*n*=6; 27.3%).

According to EBSP analysis, A2a5 initially appeared in Spain with a very high effective population size (suggesting external introduction) and maintained almost constant growth for a long period of about one month. It most likely spread from an initial focus in Madrid, then reached the highest N_e among all Spanish SARS-CoV-2 lineages (Figures 2, 3).

Haplogroup A2a5c most likely originated in Madrid (Spain)

Although the first recorded instances of A2a5c appeared simultaneously in Europe and North America (5 March), A2a5c was clearly circulating in Europe beforehand. As many as 266 out of 294 A2a5c cases in the worldwide database (90.5%) were found on this continent (compared to six cases each in Asia, North America, and Oceania, and four in South America) (Figures 2, 3).

A2a5c appeared mostly in Europe, representing 1.8% of total lineages in the dataset. The first European case occurred in Spain on 5 March. Taking into account that its immediate ancestral node, A2a5, was already rooted in Spain at the beginning of March, it seems plausible that this lineage originated here (and not in Italy, the likely origin of A2a5, but without A2a5c representatives). It spread very quickly in Spain, where it reached its highest prevalence (100 out of 294 A2a5c genomes in the database; 34.0%); these Spanish A25ac genomes appeared almost continuously from 5 March to mid-April. Subsequently, this sub-lineage began to spread gradually to other European countries, e.g., the Netherlands (*n*=61; 4.4% of country total) and UK (*n*=73; 0.9% of country total).

Within Spain, the first A2a5c genome was sampled in Madrid on 05 March (9 genomes from 5 to 12 March); this lineage could have moved to other regions a few days after. It reached the highest frequency in Catalonia (26.5% of all genomes in this region; first case on 16 March) followed by Madrid (24.7%) and Navarra (19.0%; first case on 12 March) (Figures 2, 3). The A2a5c root haplotype (i.e., with substitution G29734C on top of A2a5 sequence motif; #H3: C241T–C3037T–C14408T–A20268G–A23403G–G29734C; Table 2) reached its highest incidence in Spain (*n*=44), and appeared only sporadically in other countries (e.g., Colombia, Chile, Portugal, UK).

The EBSP results for A2a5c agreed with the analysis carried out on its ancestral node A2a5. It started with a significant N_e slightly later than its ancestor and underwent discrete growth for a month (Figures 2, 3).

Haplogroup A2a10 could have originated in Portugal

A2a10 is a minor clade worldwide (*n*=39), predominantly present in Europe (*n*=37), with only two cases exported to South America. A2a10 has been recorded in four countries only, mostly in Iberia: i.e., 26 in Spain (2.9% of total cases in the country) and nine in Portugal (1.8% of total cases). It was sampled almost simultaneously in Portugal (1 March) and Spain (2 March), but more rooted in the former: e.g., nine cases accumulated in Portugal (mostly the basal sequence motif) from 1 to 9 March compared to one case in Spain during the same period. The second Spanish case in the database appeared on 10 March, and 25 out of 26 cases were recorded from 10 to 31 March. Almost all Spanish A2a10 genomes appeared in the Valencian Community (25 out of 26 genomes in the Spanish dataset) (Figures 2, 3).

A2a10 most likely evolved from an A2a European ancestor that mutated into A2a10 in Portugal in mid to late February and spread in this country first before moving to Spain (probably to the Valencian Community). The core haplotype of A2a10 (#H4: C241T–C3037T–C14408T–A23403G–C29144T; Table 2) was observed 16 times in Spain, eight times in Portugal, and one time in Chile.

Effective population size analysis of this haplogroup showed very subtle population changes, probably mirroring a minor role of this lineage in the Spanish pandemic (Figures 2, 3).

European origin of Spanish A2a4 haplogroup

Haplogroup A2a4 is one of the most successful clades worldwide, with important representation in South America (36.1%), Europe (32.7%), Africa (15.4%), Oceania (12.5%), and Asia (9.7%), and minor presence in North America (3.0%). The A2a4 genome first appeared in the European database on 24 February, whereas the first genome in North America was recorded on 27 February. A2a4 genomes have occurred in 63 different countries, areas, and territories. From 24 to 29 February, the database recorded A2a4 representatives in Central and Western Europe, e.g., Austria (*n*=4), Denmark (*n*=1), Italy (*n*=124), the Netherlands (*n*=4), Spain (*n*=1), Switzerland (*n*=13), and UK (*n*=2), and beyond, e.g., USA (*n*=3) and Mexico (*n*=2). It reached the highest frequency in Russia (*n*=144; 66.1%), but the first genome in this country was recorded on 11 March, i.e., later than that in Central and Western Europe. The most important initial A2a4 European outbreak occurred in Switzerland (*n*=21 instances from 27 February to 3 March). Outside Western Russia, the highest frequency of European A2a4 genomes occurred in Greece (59.4% of all samples of this genome collected from 9 to 31 March), Czech Republic ((60.0%)), Serbia (54.5%), Latvia (52%), Poland (49.3%), Romania (50.0%), Italy (47.0%), Portugal (46.7%), and UK (40.1%). This genome also reached a very high frequency in several South American, Asian, and African countries, e.g., Brazil (80.0%), Argentina (50.0%), Chile (22.9%), Bangladesh (66.7%), Vietnam (60.0%), Morocco (47.4%), Oman (43.9%), and Jordan (40.0%) (Figures 2, 3).

With the current dataset, it is not possible to infer the European origin of the Spanish A2a4 genomes (7.7% of total Spanish genomes). The first recorded genome was sampled in the Canary Islands on 29 February, and the next in continental Spain (Valencian Community) on 2 March (with

>40 cases across different European countries in the same period). A2a4 appeared in Spain and Portugal at almost the same time. It accumulated in Spain mainly from 11 to 28 March (59 out of 68 A2a4 genomes), parallel to the outbreak of this clade in Portugal (117 cases from 7 to 31 March, with a peak of 74 from 17 to 21 March) where it reached a high frequency (46.9% of total genomes). By this time, A2a4 was already well rooted in Europe, e.g., Sweden, the Netherlands, Belgium, Italy, UK, Austria, Denmark, and Portugal (with 95, 67, 58, 41, 40, 32, 18, and 14 cases accumulated by 10 March, respectively).

Within Spain, A2a4 reached 7.5% of total genomes, with the main outbreak occurring in the Valencian Community (35 out of 69 Spanish cases from 2 March to 1 April; 14.3% of total genomes in the region). It appeared later in Navarra (five cases from 15 to 23 March; 11.94% in the region) and the Basque Country (all 13 cases from 26 to 28 March) (Figures 2, 3).

It is highly likely that A2a4 emerged in Europe in mid-February, but its origin within the continent remains uncertain. Its phylogeny does not provide further insight, as the root haplotype appeared almost simultaneously in many European countries. Considering chronology and frequency, we speculate an origin in Switzerland, and its arrival to Spain could have come from the Netherlands or Austria, coinciding with the A2a4 outbreaks in these countries.

EBSP analysis of A2a4 indicated a constant population growth starting at the beginning of March and reaching an effective population size similar to that of B9 at the beginning of April. This pattern coincided with the recorded chronology of the individual genomes, indicating almost continuous growth of this lineage in this period, especially in the Valencian Community, with a minor outbreak occurring in the Basque Country at the end of March, as also visible in the EBSP curve (Figures 2, 3).

Characterizing super-spreader transmissions in Spain

The six most frequent haplotypes (#H1 to #H6; Table 2) accounted for 275 cases (30.4%) of total Spanish cases in the database. Their frequencies were remarkably high when compared to that of other Spanish haplotypes. Obviously, these genomes spread more rapidly than others and increased their frequency in particular geographic locations, thus suggesting that their transmission could have been mediated by super-spreaders. These lineages emerged in different Spanish regions relatively early and experienced sudden growth, suggesting that they were the main initiators of the pandemic in the country. The impact of these common haplotypes left a clear signature in the diversity indices: i.e., the repetition of the same sequence a number of times resulted in a remarkable decrease in nucleotide and sequence diversities of the regions more affected by these haplotypes (e.g., Basque Country and Valencian Community; Supplementary Figure S12).

To investigate transmission patterns, we built phylogenetic networks of haplogroups that could have been transmitted by super-spreaders. As shown in Figure 5, many displayed a star-like phylogeny, characteristic of super-spreader host transmissions (B3a in Vitoria and Valencian Community; A2a5 in Madrid and Valencian Community). See Supplementary Figure S13 for the network of all Spanish genomes where clusters suggestive of super-spreader transmissions are joined by branches mirroring homogeneous and/or chains of transmission.

We determined several indices that summarize the topology of the phylogenetic trees from the main lineages, which are indicative of the potential mode of transmissions (Table 3). The minor clade A2a4 behaved more as a super-spreading scenario in the Valencian Community than in Donostia (Basque Country; (*H/L* index=4.92 vs. 1.24, respectively). A2a5 displayed the clearest imprint of super-spreading in Madrid (*H/L*=12.47) and the Valencian Community (*H/L*=10.43), whereas the signature was much more moderate in other regions. A2a5c also showed a super-spreading signature in Madrid (*H/L*=9.54) and the Valencian Community (*H/L*=9.24), and more moderately in Barcelona (Catalonia; *H/L*=3.28). A2a10 also showed values consistent with the presence of super-spreaders (*H/L*=6.2). Similarly, B3a had a clear star-like phylogeny in Vitoria and the Valencian Community, and topological features that favored a super-spreading transmission scenario for these clades, most notably in Vitoria (*H/L*=8.75) than in the Valencian Community (*H/L*=4.38). Finally, despite the low sample size of B9*, it showed a discrete super-spreading signature in Donostia and the Valencian Community (*H/L*=2.60 in both regions), but this was almost negligible in Madrid (*H/L*=1.53).

In agreement with the star-like networks of super-spreader host candidates and topological index values or phylogenetic trees is the very low average lifespan of identical genomes in SARS-CoV-2 (e.g., those representing basal nodes in the networks), with a mean of 6.8 days and median of 4.0 days (worldwide: mean=9.8 days; median=6 days) (Supplementary Figure S14). This short time is compatible with the rapid spread of the virus by one (or a few) super-spreader(s) represented in the basal nodes of these networks.

Mutation D614G in Spain

Recent reports have discussed a potential link between amino-acid mutation D614G and SARS-CoV-2 infectivity, pointing to a higher advantage of mutation carriers in viral dispersion (Korber et al., 2020). The nucleotide change leading to the amino-acid mutation D614G is a diagnostic variant of haplogroup A2, namely, A23403T. Here, however, it showed moderate recurrence in worldwide phylogeny, appearing sporadically in 12 haplogroup contexts, although mostly as a diagnostic variant of A2. Due to the high dispersion of A2 outside Asia (Supplementary Figure S15), this mutation appeared frequently in Africa (86.2%), South America (86.1%), Europe (80.8%), North America (70.2%), and Oceania (61.1%), and less so in Asia (35.3%).

Spain is, by far, the European country with the lowest frequency of this mutation (56.4%). Within Spain, and

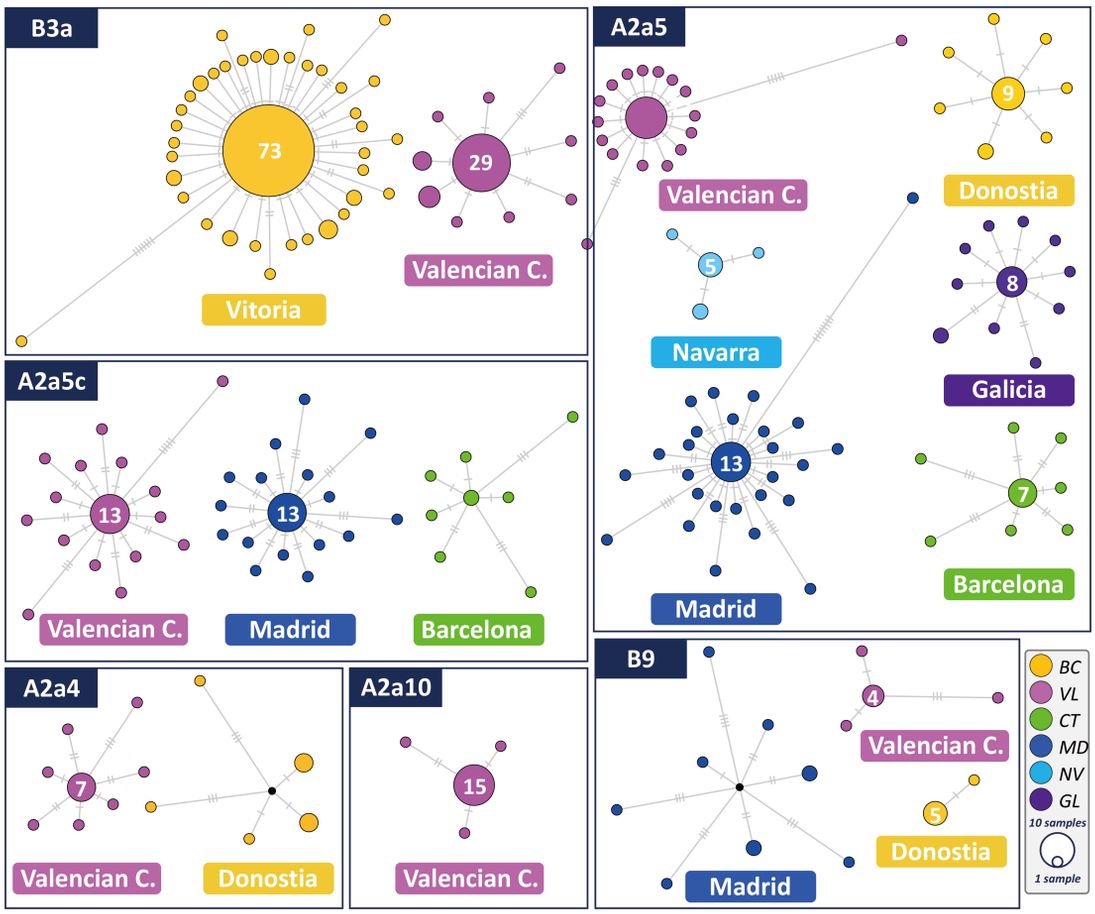

Figure 5 Networks of main Spanish super-spreader candidates by region

BC: Basque Country; VL: Valencian Community; CT: Catalonia; MD: Madrid; NV: Navarra; and GL: Galicia. We used two political divisions in this figure, namely, cities and autochthonous communities (Galicia and Navarra); this is due to the difficulty in finding a consistent geographic/political designation that fits with sampling. For instance, most Galician cases occurred around the Santiago de Compostela city, which is in the political boundaries of several provinces of the Galician region. In Navarra, most cases were sampled in the city or area close to Pamplona.

considering only regions with significant sample sizes, Catalonia (97.1%) and Galicia (89.2%) had the highest frequencies, while Andalusia (49.6%) and the Basque Country (26.9%) had the lowest values (Supplementary Figure S15).

DISCUSSION

Spain has been severely affected by the COVID-19 pandemic. The pandemic was controlled during its first epidemic wave only after human intervention involving severe preventive social measures. In this study, we showed that COVID-19 in Spain cannot be explained by the single introduction of the virus to the country, but by at least 34 independent events. This figure is probably an underestimation of the real number due to the limited sample size of the Spanish dataset and total number of different haplogroups we identified (n=97). We observed five different clades represented in 810 out of 922 genomes sampled (88.8%). In particular, the presence of haplogroup B (at a frequency of 39.2%) was a very distinct aspect of Spain when compared to the rest of Europe (B haplotype frequency of only 11.0%, with haplogroup A accounting for the rest). The phylogeny of these highly frequent clades was basically star-like, with the epicenter occupied by their basal haplogroup motif. To a large extent, these phylogenies are compatible with a viral transmission pattern favored by super-spreaders. This proposition was also supported by the high frequency of these haplotypes in geographically constrained areas and over narrow time windows.

The phylogeny of SARS-CoV-2 Spanish genomes, with clusters representing infections mediated by super-spreaders separated by patterns mirroring homogeneous and chain transmissions, suggests strong action of genetic drift and multiple founder events. This pattern seems to be characteristic of the pandemic at a more global scale (Gómez-Carballa et al., 2020). It supports a scenario where a single founder, such as B3a, if successful in a super-spreading event, can initiate a local outbreak in a specific location, which

Table 3 Normalized phylogenetic features of potential super-spreader candidate phylogenies in Spanish COVID-19 outbreak

HG	AL	*n*	n_1	n_2	AL	CH	CI	IL	MH	PF	SI	SN_1	SN_2	H	L	H/L
A2a4	VL	14	7	7	1.00	0.14	1.00	1.00	1.00	0.21	1.00	0.92	0.24	0.99	0.20	4.92
A2a4	DN	9	0	9	0.43	0.44	0.71	0.71	0.75	0.67	0.91	0.63	0.55	0.69	0.56	1.24
A2a5	VL	32	15	17	1.00	0.06	1.00	1.00	1.00	0.09	1.00	0.97	0.13	0.99	0.10	10.43
A2a5	MD	39	13	26	1.00	0.05	1.00	1.00	1.00	0.08	1.00	0.97	0.11	1.00	0.08	12.47
A2a5	DN	17	9	8	0.43	0.24	0.90	0.87	0.94	0.18	0.93	0.88	0.27	0.82	0.23	3.61
A2a5	BC	14	7	7	1.00	0.14	1.00	1.00	1.00	0.21	1.00	0.92	0.24	0.99	0.20	4.92
A2a5	GL	19	8	11	0.88	0.21	0.88	0.88	0.94	0.16	0.92	0.89	0.25	0.90	0.21	4.36
A2a5	NV	9	5	4	0.57	0.44	0.75	0.71	0.88	0.33	0.89	0.75	0.47	0.76	0.42	1.83
A2a5c	MD	29	13	16	1.00	0.07	1.00	1.00	1.00	0.10	1.00	0.96	0.14	0.99	0.10	9.54
A2a5c	VL	28	13	15	1.00	0.07	1.00	1.00	1.00	0.11	1.00	0.96	0.14	0.99	0.11	9.24
A2a5c	BN	9	2	7	1.00	0.22	1.00	1.00	1.00	0.33	1.00	0.88	0.34	0.98	0.30	3.28
A2a10	VL	18	15	3	1.00	0.11	1.00	1.00	1.00	0.17	1.00	0.94	0.20	0.99	0.16	6.20
B3a	VT	117	73	44	0.18	0.12	0.91	0.90	0.94	0.05	0.92	0.94	0.10	0.80	0.09	8.75
B3a	VL	43	29	14	0.29	0.14	0.81	0.90	0.88	0.21	0.84	0.93	0.18	0.78	0.18	4.38
B9	DN	6	5	1	1.00	0.29	1.00	1.00	1.00	0.43	1.00	0.83	0.41	0.97	0.37	2.60
B9	VL	7	4	3	1.00	0.29	1.00	1.00	1.00	0.43	1.00	0.83	0.41	0.97	0.37	2.60
B9	MD	10	0	10	0.33	0.55	0.76	0.56	0.80	0.27	0.89	0.70	0.50	0.67	0.44	1.53

Analysis was carried out in indicated haplogroups after subtracting corresponding nested sub-clades. *n*: Total sample size; n_1: sample size of principal node (only for super-spreader candidate); n_2: sample size of derived haplotypes (only for super-spreader candidate). For Spanish regions: BN: Barcelona (Catalonia); DN: Donostia (Basque Country); GL: Galicia; MD: Madrid; NV: Navarra; VL: Valencian Community; VT: Vitoria (Basque Country); Statistical indices: AL: Normalized average ladder; CH: Cherries; CI: *Colless index*; IL: IL number; MH: Maximum height; PF: Pitchforks; SC: Sackin index; SN_1 and SN_2: Staircase-ness 1 and 2, respectively. H: Average of AL, CI, IL, MH, and SI, SN_1; L: Average of CH, PF, and SN_2.

subsequently facilitates the spread of both core and derived strains to other regions.

We found no evidence to support previous claims suggesting increased transmissibility of SARS-CoV-2 strains carrying the amino-acid mutation D614G (basically all A2 haplotypes; see phylogeny of Figure 1 and (Gómez-Carballa et al., 2020)). In fact, the low frequency of this mutation in Spain is difficult to reconcile with the high incidence of the disease in the country, compared to other countries with lower disease incidence but a higher frequency of D614G. Furthermore, it is apparent that Spanish regions with the highest frequency of this mutation (e.g., Galicia in the northwest) have had some of the lowest incidences of the disease within the country. However, it is important to state that inferences based exclusively on frequencies must be regarded with caution; they should be based on proper functional analyses and supported by clinical evidence. Moreover, interpretations in regard to this and other mutations of interest should not be isolated from epidemiological evidence; for instance, the pandemic arrived later to some regions in Spain (e.g., Galicia), with lockdown likely preventing more effective dissemination.

By combining phylogenetics with the chronology and geographic location of viral genome sampling, we established the most parsimonious origin for the main strains responsible for COVID-19 in Spain (Figure 3). Five lineages (six if we consider A2a5c, which is a sub-lineage of A2a5) were responsible for nearly 90% of genomes in Spain. The first occurrences of all of them were recorded between 26 February and 5 March, an important incubation period right before the rapid growth of COVID-19 cases starting in mid-March, which triggered the national alarm and lockdown (Figure 6). Most lineages were sampled during this period, although a few emerged after the lockdown. Of the lineages appearing after 15 March, only a small proportion appear to have been introduced by people returning to Spain from abroad, indicating that the lockdown was very effective at preventing new entrances into the country (Figure 6).

We estimated the TMRCA of the main clades observed in the Spanish database (Table 4; Figure 6; Supplementary Figures S16–S21). Phylogeographic inferences suggest that B3a, B9, and A2a5c may have originated in Spain. The most recent TMRCA for these autochthonous lineages was for B3a, dated to 11 February 2020 (HPD 95% CI: 30 January to 20 February), followed by B9 (22 February). The other non-Spanish clades arrived later to Spain according to their TMRCAs (between 20 February and 3 March) (Table 4; Figure 6). This suggests that the coronavirus was already circulating as B3a in Spain in mid-February at least, but probably not before the end of January. Therefore, the introduction of SARS-CoV-2 in the country could have taken place in the form of a B3 genome of unknown geographic origin, which mutated in Spain into B3a around 11 February. The five main haplogroups (and sub-lineage A2a5c) incubated in several regions of the Spanish territory before the lockdown (Figure 6).

The present study has several limitations, which have been detailed previously in a global study using the same methodology by Gómez-Carballa et al. (2020) and other related studies (Forster et al., 2020a, 2020b; Yu et al., 2020).

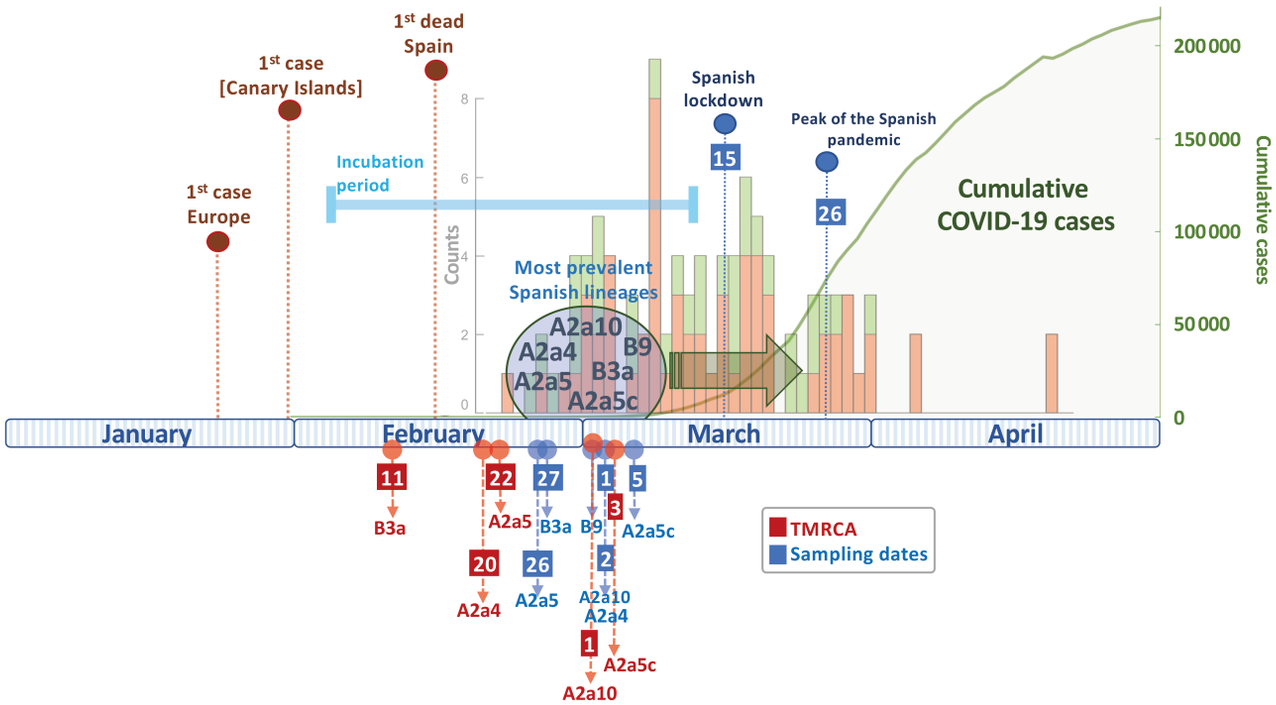

Figure 6 Timeline of main episodes related to spread of SARS-CoV-2 in Spain

Histogram in background indicates number of transmission lineages in Spain; red indicates ones generated *de novo* in Spain (*n*=64) from "domestic" mutations, green indicates lineages introduced from abroad (*n*=30). Indicated below schematic line representing months are dates when first genomes were sampled for main haplogroups, and TMRCA of these clades. Distribution of COVID-19 cases in Spain was built with data obtained from https://ourworldindata.org.

Table 4 TMRCA of main Spanish clades

Haplogroup	*n*	TMRCA	HPD	Origin
A2a4	68	2020/02/20	2020/02/11–2020/02/27	Europe
A2a5	315	2020/02/22	2020/02/12–2020/02/29	Italy
A2a5c	95	2020/03/03	2020/02/28–2020/03/05	Spain
A2a10	27	2020/03/01	2020/02/29–2020/03/02	Portugal
B3a	266	2020/02/11	2020/01/30–2020/02/20	Spain
B9	74	2020/02/22	2020/02/15–2020/02/28	Spain

HPD: 95% highest posterior density (HPD) confidence interval. Origin refers to most likely origin of haplogroups inferred as most favorable scenario that considers both genome sampling chronology and regional genome variation. Time of the most recent common ancestor (TMRCA)

Briefly, some inferences carried out in the present study rely (although not exclusively) on sampling origin and dating records in GISAID, the reference database for SARS-CoV-2 genomes. Sampling of patients cannot be considered random; a non-random sample may overestimate the effect of super-spreading but also underestimate the variability observed in the country, as well as the number of different independent introductions. The regions in Spain were heterogeneously represented in the database; for instance, Catalonia was underrepresented compared to the Basque Country or Valencian Community. This could affect the frequency of clades and chronology of the events as recorded. Furthermore, with the currently available SARS-CoV-2 genomic data, it is impossible to separate at the individual level those genomes transmitted via horizontal or chain mode from those transmitted by a single or a few super-spreaders, even though the overall network phylogenies and topological statistical features, coupled with the short time period of all genomes in the basal nodes and their restricted geographic location, are all compatible with the existence of super-spreaders in these transmissions. Future studies, including epidemiological data, could help verify this hypothesis. For the time being, however, it is important to highlight that a number of cases reported in the literature and echoed by the media and governments worldwide (Kupferschmidt, 2020) point to super-spreaders of COVID-19 as one of the main drivers of the pandemic (Gómez-Carballa et al., 2020). In addition, future research should examine what makes a COVID-19

patient a super-spreader, e.g., the super-spreader condition may depend on particular circumstances, environments, and/or biological features.

In summary, the present study represents one of the most detailed phylogeographic analyses of SARS-CoV-2 focusing on a particular country and provides a methodological framework that could be applied to other regions. The phylogeographic reconstruction of the main haplogroups that have successfully spread in Spain was possible thanks to the establishment of a solid phylogeny for SARS-CoV-2 genomes (Gómez-Carballa et al., 2020). We demonstrated that genetic drift, most likely powered by the presence of super-spreaders, has played a fundamental role in the Spanish COVID-19 pandemic. This is in good agreement with the highly heterogeneous transmission observed according to epidemiological evidence, and the very low value of *k* dispersion index observed for the SARS-CoV-2 pathogen (Endo et al., 2020).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

A.S., F.M.-T., and M.L.P.-M.-B. conceived the study. A.S., A.G.-C., X.B., and J.P.-S. carried out the phylogenetic and statistical analyses. A.S. prepared the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We gratefully acknowledge GISAID's EpiFluTM (www.gisaid.org) Database and contributing laboratories (Supplementary Table S1) for giving us access to the SAR-CoV-2 genomes used in the present study.

REFERENCES

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**(1): 37−48.

Ceraolo C, Giorgi FM. 2020. Genomic variance of the 2019-nCoV coronavirus. *Journal of Medical Virology*, **92**(5): 522−528.

Colijn C, Gardy J. 2014. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*, **2014**(1): 96−108.

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, et al. 2015. System for automated geoscientific analyses (SAGA) v.2.1.4. *Geoscientific Model Development*, **8**(7): 1991−2007.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**: 214.

Endo A, Abbott S, Kucharski AJ, Funk S. 2020. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, **5**: 67.

Forster P, Forster L, Renfrew C, Forster M. 2020a. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(17): 9241−9243.

Forster P, Forster L, Renfrew C, Forster M. 2020b. Reply to Sanchez-Pacheco et al., Chookajorn, and Mavian et al.: explaining phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(23): 12524−12525.

Gómez-Carballa A, Bello X, Pardo-Seco J, Martinón-Torres F, Salas A. 2020. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Research*, doi: 10.1101/gr.266221.120.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**: 289.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**(2): 254−267.

Kendall M, Boyd M, Colijn C. 2018-02-21. Calculating topological properties of phylogenies. https://cran.r-project.org/web/packages/phyloTop/.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 Virus. *Cell*, **182**(4): 812−827.

Kupferschmidt K. 2020-05-19. Why do some COVID-19 patients infect many others, whereas most don't spread the virus at all? https://www.sciencemag.org/news/2020/05/why-do-some-covid-19-patients-infect-many-others-whereas-most-don-t-spread-virus-all.

Leigh JW, Bryant D. 2015. POPART: full‐feature software for haplotype network construction. *Methods in Ecology and Evolution*, **6**(9): 1110−1116.

Metzig C, Ratmann O, Bezemer D, Colijn C. 2019. Phylogenies from dynamic networks. *PLoS Computational Biology*, **15**(2): e1006761.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**(10): 5269−5273.

Norström MM, Prosperi MCF, Gray RR, Karlsson AC, Salemi M. 2012. PhyloTempo: a set of R scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. *Evolutionary Bioinformatics*, **8**: 261−269.

Shen ZJ, Xiao Y, Kang L, Ma WT, Shi LS, Zhang L, et al. 2020. Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clinical Infectious Diseases*, **71**(15): 713−720.

Shu YL, McCauley J. 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, **22**(13): 30494.

Tang XL, Wu CC, Li X, Song YH, Yao XM, Wu XK, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, **7**(6): 1012−1023.

van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, **83**: 104351.

WHO. 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020.

Wong G, Bi YH, Wang QH, Chen XW, Zhang ZG, Yao YG. 2020. Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important?. *Zoological Research*, **41**(3): 213−219.

Yu WB, Tang GD, Zhang L, Corlett RT. 2020. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zoological Research*, **41**(3): 247−257.