bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Science

The impact of super-spreaders in COVID-19: mapping genome variation worldwide

Alberto Gómez-Carballa[1,2,3,#], Xabier Bello[1,2,3,#], Jacobo Pardo-Seco[1,2,3], 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 Investigaciones Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), 15706, Galicia, Spain;

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

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

[#]Equally contributed

*Correspondence: antonio.salas@usc.es

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The human pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the major pandemic of the 21st century. We analyzed >4,700 SARS-CoV-2 genomes and associated meta-data retrieved from public repositories. SARS-CoV-2 sequences have a high sequence identity (>99.9%), which drops to >96% when compared to bat coronavirus. We built a mutation-annotated reference SARS-CoV-2 phylogeny with two main macro-haplogroups, A and B, both of Asian origin, and >160 sub-branches representing virus strains of variable geographical origins worldwide, revealing a uniform mutation occurrence along branches that could complicate the design of future vaccines. The root of SARS-CoV-2 genomes locates at the Chinese haplogroup B1, with a TMRCA dating to 12 November 2019 - thus matching epidemiological records. Sub-haplogroup A2a originates in China and represents the major non-Asian outbreak. Multiple bottleneck episodes, most likely associated with super-spreader hosts, explain COVID-19 pandemic to a large extent.

Keywords: COVID-19; coronavirus, SARS-CoV-2; genetic drift; bottlenecks; super-spreaders; natural selection

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Since the beginning of the COVID-19 pandemic, there has been a growing interest in exploring genetic variation in the genome of severe acute respiratory syndrome coronavirus (SARS-CoV-2). Identifying patterns of genomic variation can help understand the origin and spread of the pandemic and facilitate the development of future vaccines. The amount of genome data deposited in public repositories in a such a reduced timeframe offers a unique opportunity for a detailed phylogenetical characterization of SARS-CoV-2, as well as the geographic mapping of the different clades spreading worldwide, and of the impact of the outbreaks on the genome variability of the virus. Initial analyses so far used a limited number of SARS-CoV-2 genomes only, and focused mostly on various evolutionary aspects of the SARS-CoV-2 genomes (*1-4*).

We built a solid phylogenetic skeleton of SARS-CoV-2 genomes that allows to investigate sequence variation in a large number of genomes (>4,700; **Supplementary Material**) deposited in the Global Initiative on Sharing Avian Influenza Data GISAID (*5*), explore site-specific mutational instability, investigate phylogeographic patterns of variation worldwide, and clarify the role of super-spreader hosts in the pandemic.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identity of SARS-CoV-2 to other closely related species

Human SARS-CoV-2 genomes have a within sequence identity of 99.98% (**Table 1**); and are much more identical to bat coronavirus than to pangolin coronavirus, although the values vary substantially depending on the specimen, 93.44%–96.17% (**Table 1**). When compared to pangolin coronavirus, the range of genome identities drops to 85.24%–92.35%.

Between 1,699 and 3,727 substitution variants separate the pangolin coronavirus genomes from the SARS-CoV-2 reference sequence, and this range drops to 1,105 to 1,369 (**Table 1**) when compared to bat coronavirus. The bat #412976 coronavirus genome is conflictive because it has an unusual amount of mutational differences with respect to the SARS-CoV-2 reference and has an abnormally low sequence identity with human coronavirus (76.87%), comparable to pangolin coronavirus. This genome is problematic in the sequence alignment and should be avoided in future comparative analyses.

Inter- and intra-specific phylogeny and the root of SARS-CoV-2

An inter-specific Maximum Likelihood (ML) tree was built using pangolin, SARS, and bat coronavirus genomes as outgroups to investigate their phylogenetic relationships with SARS-CoV-2 (**Supplementary material**). The tree depicts the SARS coronavirus genome occupying the most external branch. Next, all the pan genomes cluster separately from bat and human coronavirus, which also group separately. In line with its very low identity with SARS-CoV-2, bat-412976 behaves as an outlier in the tree. Overall, the clustering pattern in the tree is in very good agreement with sequence identity values (**Table 1**).

We next focused our attention on the root for all existing SARS-CoV-2 genomes, assuming the bat coronavirus as its closest coronavirus relative. We built a new ML tree including all SARS-CoV-2 genomes sequenced up to 29 February 2020 (n = 621); almost all of them are of Asian origin and this group should contain the Most Recent Common Ancestor (MRCA), as it is evident from phylogenetics and epidemiology that the origin of the pandemic is in China and more particularly within haplogroup B (see below and **Supplementary Material**). The ML tree unequivocally reveals that the root of SARS-CoV-2 is located in the basal B1 haplogroup (B1 genomes that do not belong to derived

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B1 sub-clades; **Figure 1**), and therefore points to B1 as the clade at the origin of the pandemic.

SARS-CoV-2 mutation rate, as inferred from the ML tree, is 5.42×10^{-4} (Bootstrap 2.5% – 75% confidence interval: 4.29×10^{-4}–8.02×10^{-4}) according to an uncorrelated relaxed-clock method; a slightly higher mutation rate of 6.05×10^{-4} (Bootstrap 2.5% – 75% confidence interval: 4.46×10^{-4}–8.22×10^{-4}) was obtained assuming and strick-clock model.

According to a relaxed-clock model mutation rate the TMRCA for all SARS-CoV-2 genomes dates to 12th November 2019 (Bootstrap 2.5% – 75% confidence interval: 7th August 2019 to 8 December 2019), fully matching epidemiological dates; estimates using an strick clock mutation rate varied very little: 7th November 2019 (Bootstrap 2.5% – 75% confidence interval: 18 August 2019 to 2 December 2019).

The most parsimonious tree (fully developed in **Supplementary Material Figure S4**; see also skeleton in **Figure a**) shows that the two very stable transitions C8782T and T28144C (3 and 1 total occurrences in the phylogeny, respectively) separate SARS-CoV-2 variation into two main clades, A and B, both originating in China. Sub-haplogroups emerging from these main clades are mainly supported by single mutations, most of them being very solid along the phylogeny (**Table S1**), and therefore granting the robustness of the different clades. It is notable that the structure of the branches in the parsimonious tree fully agrees with the skeleton shown in the ML tree.

Haplogroup B (19.65% of the genomes in the database; *n* = 664) is present in all continents; being more prevalent in North America (46.35%), South America (25.93%) and Asia (22.33%), and having the lowest frequencies in Africa (8.33%) and Europe (3.74%) (see frequency interpolated maps in **Supplementary Material**). B1 is separated from B by a single transition (C18060T) and it is by far the most numerous B subclade (*n* = 424; 63.86% of all B). The main proportion of B1 lineages worldwide is present in North America (*n* = 365; 86.04% of all existing B1 genomes). Most of the B1 genomes belong to the subclade B1a1 (B1a[A17857G]>B1a1[C1774T]); this contains at least 11 minor sub-clades, each defined by characteristic single mutations. In consonance with the root of SARS-CoV-2 being within B, we observed that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

basal haplogroup B is more prevalent in Asia (70%) than in anywhere else, and it is the only region containing genomes belonging to all first level B-subclades (B1, B2, B3, etc; perhaps with the exception of the minor clade B9). It is noticeable however that, within B, the fourth level sub-clade B1a1 is the most frequent haplogroup in the database (399/664; 60.09%) and it appears mainly in North America (accounting for 357 [333 in USA] out of the 399 B1a1 counts in the database; 89.47%), while it is absent in Asia; **Figure 2**. In Europe, the main B sub-clade is B3a (61.02%), which is particularly prevalent in Spain, one of the main European epicenters of COVID-19. It is most likely that most of these B3a representatives arrived in South America from Spain (where it represents 71.43% of all B genomes) given the high connectively between the two regions; **Figure 2**. The high B3 frequency observed in Spain marks a notable difference with respect to other European countries; 32 out of the 37 (86.49%) B3 genomes in Europe are located in Spain.

Haplogroup A (*n* = 2,715), with complementary frequencies to B, is the predominant clade all over the world (**Figure 2**), although with heterogeneous distributions (**Supplementary Material**). It reaches its highest frequencies in Europe (97.1%) and Africa (93.1%), is relatively high in Asia (76.3%) and Oceania (77.4%), and it has the lowest frequencies in South America (68.2%) and North America (53.1%). By far, the most frequent sub-clade of haplogroup A is A2a (*n* = 1,849; 68.10% of all A genomes), which is the main representative of the non-Asian outbreak, followed by A1a (*n* = 287; 10.57%). Even though A2a is mostly present in Europe (*n* = 1,199; 64.85% of all A2a sequences), and North America (*n* = 370; 20.01%), it most likely originated in Asia (**Supplementary Material**). A3 is mainly found in Asia, and especially in the Middle East (81.25%); its sub-clade A3a is also highly prevalent in the same region (31.94%) but shows even higher frequency in Oceania (44.44%). Other minor clades are found in more restricted areas; for instance, A4a (*n* = 39; 1.44% of A) is only found in Wales (Europe), while A5, A7 and A9b appear only in Asia.

Phylogeographic information allows reconstructing dynamics of (sub)haplogroups worldwide (**Figure 2**). The main clades emerged in Asia (mainly in China), while some minor ones appeared outside Asia (next section; **Supplementary Material**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The number of sequences belonging to clade A and its main sub-clades increased exponentially during the outbreak occurring outside Asia at the end of February 2020, while the frequency of haplogroup B genomes increased more slowly at that time (**Supplementary Material**). Nucleotide diversity is almost homogeneous in all the different geographical regions for the main haplogroups; however, haplotype diversity (HD) values vary more substantially among haplogroups, probably indicating the weight of sequence founders in this index (see next section on super-spreaders); **Supplementary Material**.

Super-spreaders and founder effect

It is remarkable that a few haplotypes are disproportionally represented in continental regions or in particular countries (**Supplementary Material**; **Figure S10**), appearing abruptly in a few days' period. This pattern is compatible with super-spreader hosts arriving to certain geographic locations and giving rise to severe founder effects (**Figure 3A**). Haplotypes #H1, #H2, #H3, and #H4 (ID's as in **Table S8**) are the most frequently repeated ones. H1 (*n* = 163; haplogroup A2a4), represents one of the main haplotypes responsible for the introduction or the pandemic in Europe (104/163; 63.80%), with particular frequency in the UK (35/163; 21.47%) and Belgium (23/163; 14.11%); it is also prevalent in North America (with a one week delay; 15/163 13.50%), and Australia (18/163; 11.04%). H2 (*n* = 133; A2a2a) occurs also in Europe at high frequency (75/133; 56.39%; 34 times in Iceland) and in North America (48/133; 35.82%; mostly in USA with 45 occurrences). H3 (*n* = 132; B1a1) appears at remarkably high frequency and almost exclusively in USA (126/132; 95.45%; B1a1). H4 (*n* = 78; A) corresponds to the reference sequence (GenBank acc. nº MN908947.3) and it reaches the highest frequency in Asia (60/78; 76.92%), particularly in China (49/78; 62.82%); the frequency of H4 increased in two pulses, one coinciding with its first appearance in China at the end of December 2019, and the next coinciding with the large Asian outbreak in mid-February 2020; later, H4 moved to other non-Asian locations, e.g. USA (10/78; 12.82%).

There are additional examples of SARS-CoV-2 super-spreader hosts (**Table S8**) appearing in restricted geographical areas. For instance, H8 (*n* = 33; A3) appears at high frequency in Japan (28/33; 84.85%). In Iceland, founder haplotypes represent a large proportion of all existing haplotypes on the island

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

e.g. H7 exists only in Iceland (*n* = 37), and together with H2 (*n* = 34 in Iceland) and other four haplotypes, makes up 39.18% of all the haplotypes in this country. In USA, H3 occurs 126 times, and H2 45 times; together with other five haplotypes, they make up 31.75% of all genomes in this country. In the UK, eight haplotypes make-up 28.95% of the total haplotypes. H9 (*n* = 26; B3a) and H14 (*n* = 22; A2a5) are probably the main haplotypes responsible for the Spanish outbreak; H9 (21/26 in Spain; 80.77%) is particularly interesting because it belongs to haplogroup B3a, while almost all European haplotypes belong to haplogroup A (**Supplementary Material**; **Figure S9**); H14 appears 9 times in Spain (9/22; 40.91%).

Common haplotypes are frequently shared between neighboring countries, an observation mirroring the easy spread of the virus over short geographic distances; for instance, H33 (*n* = 9; of which 7 are in Portugal and 2 in Spain) or H45 (*n* = 7; of which 4 are in Portugal and 1 in Spain).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evolution of effective population size of SARS-CoV-2

Extended Bayesian Skyline Plot (EBPS) analysis undertaken on genomes sampled until the end of February (see **Supplementary Material**) reflects with great precision the main COVID-19 epidemiological episodes. If we consider the estimated TMRCA for SARS-CoV-2 to 12th November 2019 and allow 14-24 days of disease incubation (until approximately the 6th of December), this leaves a period of two or three weeks of silent local transmission of the virus until the first case is reported in Wuham on 30th December 2019. From this moment, N_e begins to slightly increase for a couple of weeks (**Figure 3B**), followed by exponential growth from 20th January 2020, coinciding with the Asian outbreak. The peak is reached on 30th January 2020, matching the Asian lockdown. Consequently, N_e drops remarkably for the next couple of weeks, but starts to grow progressively again from 12th February 2020, coinciding with the beginning of the non-Asian outbreak.

By overlapping COVID-19 incidence (officially reported cases per day worldwide; https://ourworldindata.org) with the EBPS plot, we observed comparable shape distributions, but with a remarkable 14–15 days' delay in reported cases per day worldwide with respect to the EBPS distribution (**Figure 3B**).

Discussion

We have undertaken a large-scale study on SARS-CoV-2 genomes considering a sample that is more than an order of magnitude higher than those of previous analyses. By focusing on high-quality (HQ) genomes, we devoted great effort to elucidate the most parsimonious phylogeny of SARS-CoV-2. This effort has allowed us to present novel phylogeographic inferences on the origin and dynamics of CoV-2 strains. In particular, we discovered a few dozen genomes (representing > 1/3 of the total database) that played a fundamental role as super-spreaders of COVID-19 disease. These SARS-CoV-2 strains (belonging to different haplogroups), occur with remarkable frequency in the dataset and became founders in restricted regions or countries in short time periods (of a few days).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SARS-CoV-2 genomes show very high identity among themselves (>99%) and lower to bat coronaviruses (>96%; BatCoV RaTG13); these values are very identical to earlier estimates based on a limited number of SARS-CoV-2 genomes (*6*). The pangolin coronavirus genome, initially proposed as the original host of SARS-CoV-2, shows significantly lower identity. The high identity observed between SARS-CoV-2 genomes and other betacoronaviruses adds support to its zoonotic origin from a bat relative (*6*). The differences found between SARS-CoV-2 and their most related coronaviruses in horseshoe bat indicate that large number of mutational jumps were needed to generate these differences from a common ancestor which could have existed in a time frame between 1948-1982 (*7*). Divergent genomes could have been incubated in animal reservoirs before the zoonotic jump to humans in the shape of a B1 genome, in a process similar to that observed for palm civet as intermediary in other SARS coronavirus (*8*). These new coronaviruses would be able to use human *ACE2* receptor to infect patients. Patterns of variation observed in SARS-CoV-2 could be explained assuming a unique index case, which would already contain the very specific and well-conserved PFCS insertion. This original B1 genome would then start to diverge very soon in Wuham in two directions of the phylogeny, giving rise to its most frequent sub-lineage B1a1 and, almost simultaneously, to other B lineages and the large haplogroup A (timeline in **figure 3C**).

According to our inferences, the TMRCA for all SARS-CoV-2 genomes would be 12th November 2019. Assuming a maximum incubation time in humans of up to 24 days (*9*), the virus could have been infecting the first citizens from Hubei in a silent mode of transmission until the end of November 2019; and started to be noticed by Chinese health authorities in early to mid-December. The EBSP distribution suggests that the N_e of SARS-CoV-2 could have started to grow significantly from 30th December 2019, i.e. only two-three weeks after the initial cases reported and probably favored by super-spreader hosts (e.g. genomes like the reference sequence played a special role in the beginning of the Asiatic epidemic). Subsequently, it followed an exponential growth that marked the beginning of the Asian outbreak on the 20th January 2020 and lasted until the end of this month. Next, N_e experienced a notable drop coinciding with human intervention and quarantine implemented in Asia on

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

30th January 2020. Finally, the beginning of a second wave of expansion outside Asia starting around 12th-27th February 2020 is also well-recorded on the SARS-CoV-2 genomes (**Figure 3C**).

The two-week delay between the dates suggested by the EBSP distribution and the official documented incidence of COVID-19 in Asia could be due to the mean incubation time of the disease, but also to the number of cases officially declared being well below the real incidence.

With the data available in GISAID, we were not able to detect association between main haplogroups and age and sex of carriers. Further research is needed to investigate the possible differential effect of strains (haplogroups) with the disease outcome.

Evidence of natural selection acting on SARS-CoV-2 genomes needs further investigation (**Supplementary Material**), although the data suggest purifying selection acting on most of the SARS-CoV-2 genes when explored at an inter-specific level, and weaker intra-specific purifying selection. In agreement with this latter observation is the recent report indicating a 81 deletion at gene *ORF7a* that would convert the coronavirus in a less virulent pathogen with reduced short-term selective advantage (*10*). None of the HQ genomes investigated in our report carry this deletion.

In contrast to the weak (or null) action of positive selection on COVID-19 spread, there is strong evidence pointing to the role of genetic drift occurring in many continental regions and restricted locations, especially outside China. Phylogeographic analysis allowed us to investigate pandemic dynamics worldwide. The high incidence of a few lineages outside Asia was more probably due to drift and not selective advantages. The main non-Asian sub-clade A2a was probably among the first ones to leave Asia before this region established a severe population lockdown. In good agreement with epidemiological data, we observed multiple worldwide introductions of SARS-CoV-2 coming from Asia. Super-spreader hosts were probably the main responsible for genetic drift episodes. We detected >48 haplotypes in our dataset that most likely represent genomes transmitted by super-spreaders. These haplotypes have three differential features: they reached high to moderate frequencies in the population, they are characteristic of specific continental regions or even individual countries, and they appeared in a very

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

short time period of only a few days. The data suggest that these genomes have played a fundamental role in COVID-19 spreading; they alone represent 34.61% of the total genomes in the database. The role of super-spreader hosts is well reported in previous pandemics, including SARS, MERS and Ebola (*11, 12*).

We found the 12bp polybasic furin cleavage site (PFCS) in all SARS-CoV-2 genomes with only two substitutions in two different genomes (belonging to different haplogroups; **Supplementary Material**). This segment is therefore highly mutationally stable. A BLAST search (https://blast.ncbi.nlm.nih.gov/) of the PFCS indicates that this sequence segment is absolutely specific of SARS-CoV-2. The fact that the PFCS has been found universally in all SARS-CoV-2 suggests that this insertion was acquired before the zoonotic event and not after (*3*). The virulence conferred by this deletion to the coronavirus constitutes the focus of several studies (*13*).

The origin of SARS-CoV-2 has become a very popular question. The results of the present study (TMRCA dating of SARS-CoV-2, EBSP plot, and phylogeny) are compatible with an index case living in Wuhan-China, belonging to basal haplogroup B1, and most likely existing not before the beginning of November 2019. Subsequently, the coronavirus was transmitted from a living animal to a human host and then it started to spread from human to human. By analyzing stored biological samples from cases occurring at the beginning of the epidemy in Wuhan, it would be possible to narrow the search for patient zero among those belonging to the root of B1. The phylogeny built in the present study would be compatible with a single patient zero initiating the epidemic. Identifying the index case would help better understand how and when the spread of the pandemic begun, a lesson that would be useful in future pandemics. In agreement with previous studies (*3*), the theory of SARS-CoV-2 originating artificially in a lab finds no support in the results of the present study, in the sense that variation (within and between other species), and the step-wise mutational evolution observed at SARS-CoV-2 genomes is as expected for a RNA virus in nature.

This study warrants further expansion to clarify the role of super-spreaders in COVID-19 by investigating epidemiological data locally. Detecting and analyzing the genome of super-spreaders might shed light on the specific

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

host genetic background contributing to their increased propensity to transmit the pathogen, as well as to understand the mechanisms of infection and transmission of the pathogen. Moreover, the phylogenic precision to which we classified SARS-CoV-2 genomes will also serve disease studies aimed at understanding the potential role of different pathogen strains in disease outcomes, and how these correlate to, and interact with, host genomic susceptibility.

Acknowledgements

This study received support from 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.

We gratefully acknowledge GISAID and contributing laboratories for giving us access to the SARS-CoV-2 genomes used in the present study.

References

1. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A*, (2020).
2. Z. Shen *et al.*, Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *Clin Infect Dis*, (2020).
3. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat Med* **26**, 450-452 (2020).
4. X. Li *et al.*, Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *Journal of medical virology*, (2020).
5. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. C. Ceraolo, F. M. Giorgi, Genomic variance of the 2019-nCoV coronavirus. *Journal of medical virology* **92**, 522-528 (2020).

7. M. F. Boni *et al.*, Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *bioRxiv* **doi: https://doi.org/10.1101/2020.03.30.015008**, (2020).

8. B. Hu *et al.*, Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS pathogens* **13**, e1006698 (2017).

9. W. J. Guan *et al.*, Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*, (2020).

10. L. A. Holland *et al.*, An 81 nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (Jan-Mar 2020). *Journal of virology*, (2020).

11. R. A. Stein, Super-spreaders in infectious diseases. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* **15**, e510-513 (2011).

12. G. Wong *et al.*, MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. *Cell Host Microbe* **18**, 398-401 (2015).

13. S. Y. Lau *et al.*, Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. *Emerging microbes & infections* **9**, 837-842 (2020).

14. M. A. Marra *et al.*, The Genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399-1404 (2003).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Legend to the Figures

Figure 1. Inter-specific ML tree indicating the root of all existing SARS-CoV-2 genomes.

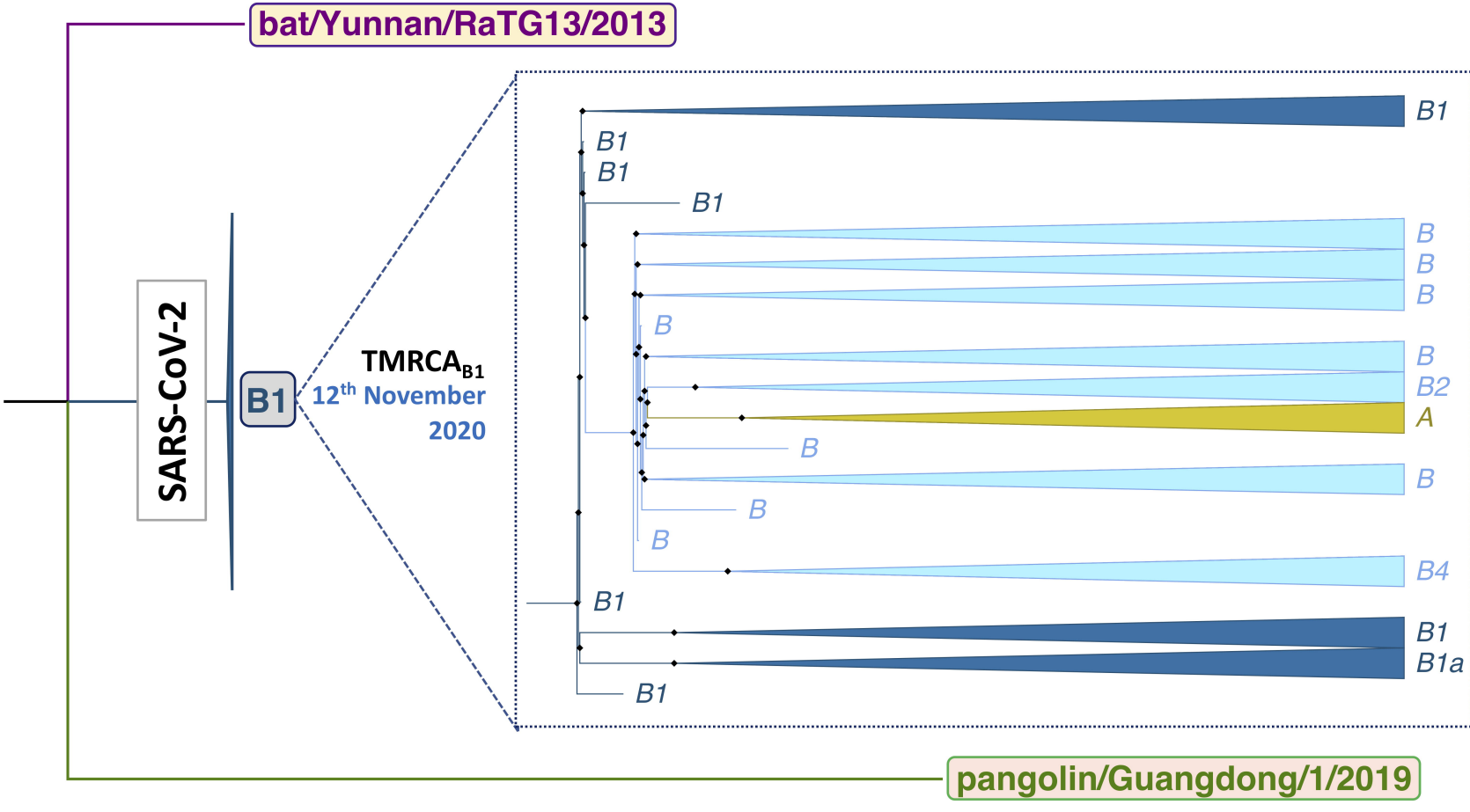
Figure 2. Map showing the worldwide spread of the main SARS-CoV-2 clades. Circle areas are not proportional to frequencies, and the arrows indicate just an approximate reconstruction of the phylo-dynamics of SARS-CoV-2 for the beginning of the Asian outbreak to the non-Asian spread of the pathogen based on meta-data (indicating the sampling origin and dates) and the classification of genomes into haplogroups according to the phylogeny in **Supplementary Material Figure S4**.

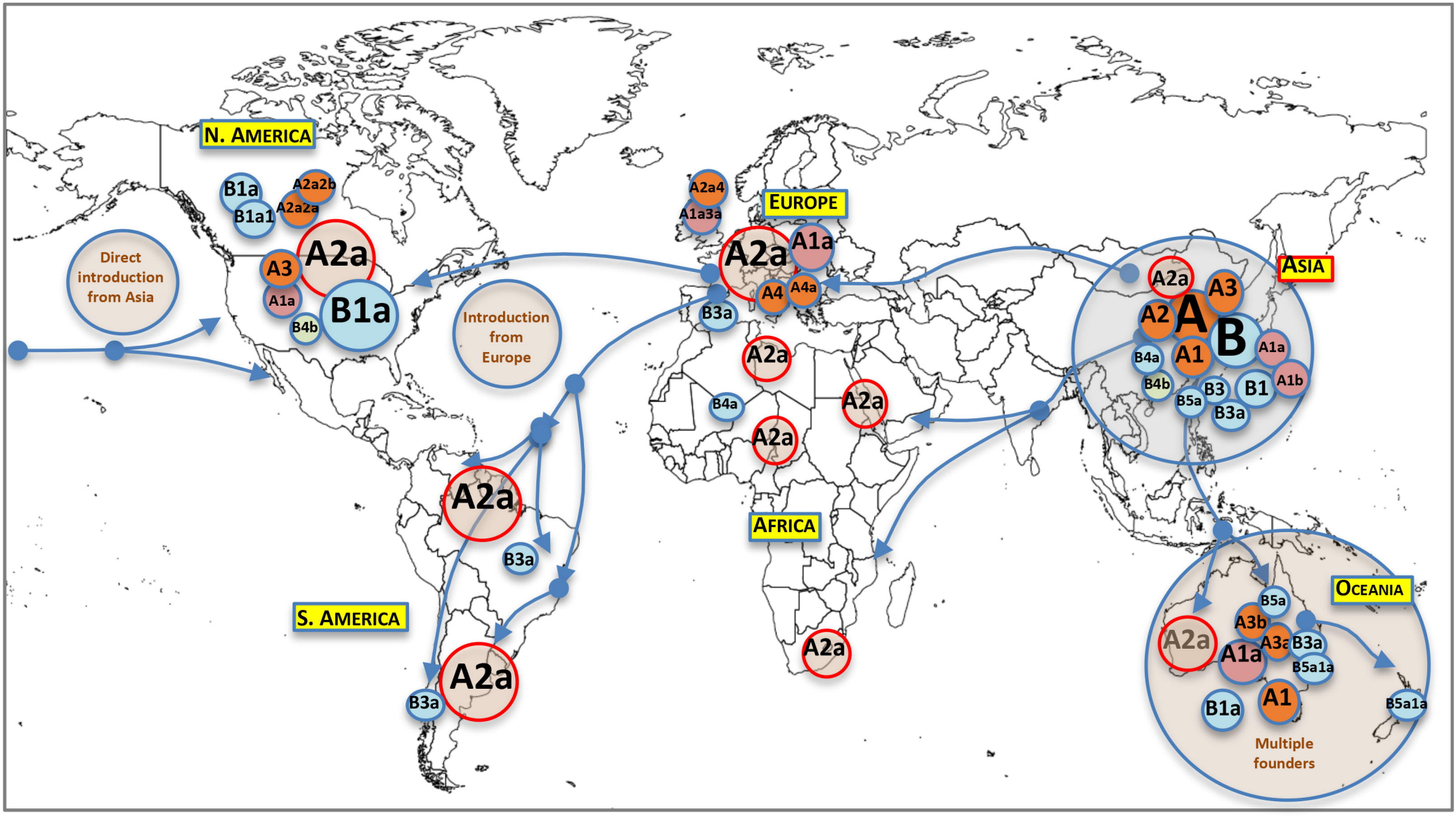
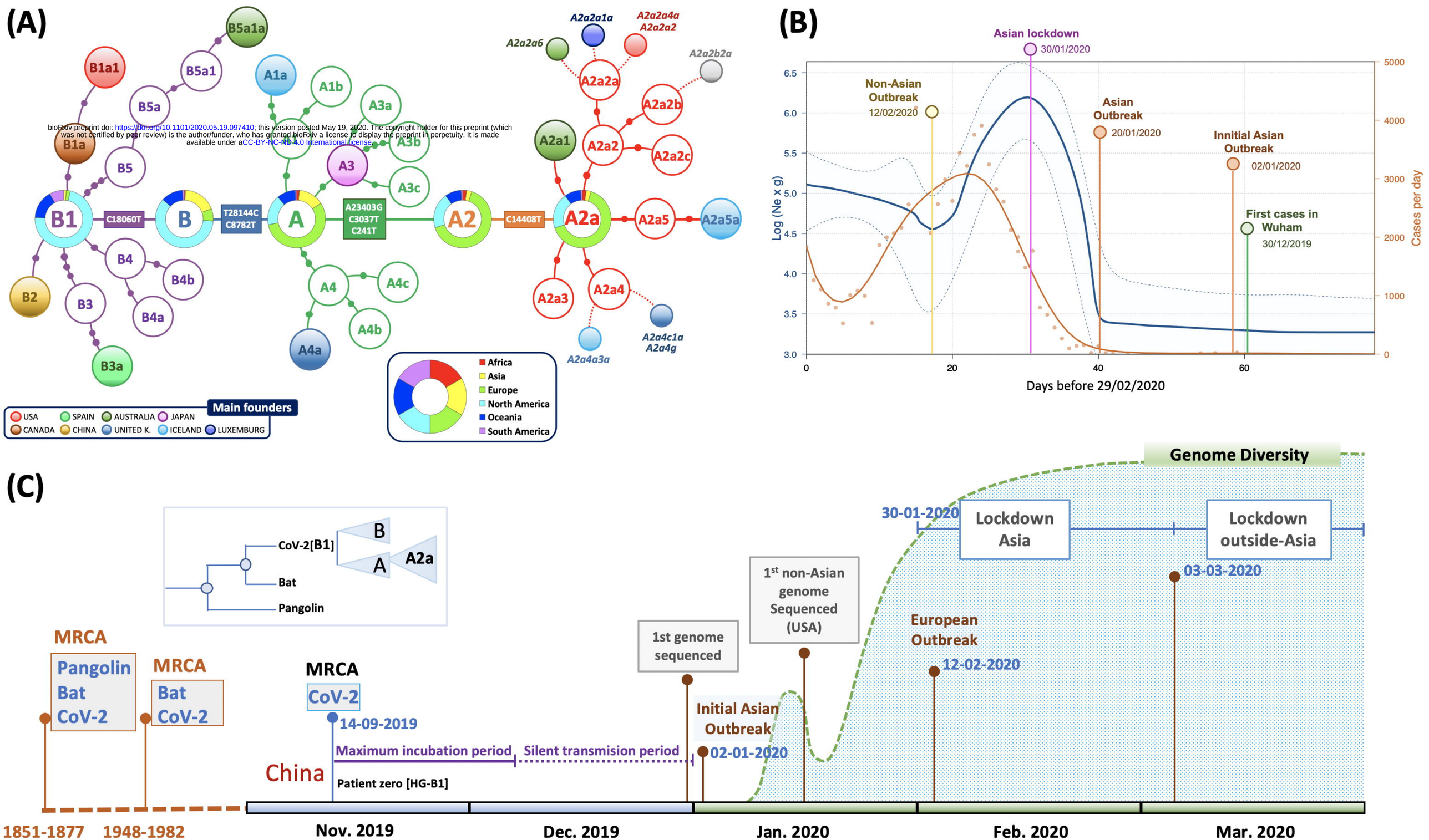
Figure 3. (A) Simplified SARS-CoV-2 phylogeny (see **Supplementary Material Figure S4** for the complete tree) illustrating the main branches and the main outbreaks occurring in Asia and outside Asia. The tree also shows the sub-clades that were mainly founded by a few super-spreaders. (B) EBPS based on genomes sampled from the beginning of the pandemic until the end of February 2020 (*n*=621). The orange distribution shows the real number of cases per day as recorded in https://ourworldindata.org for the same time period (we disregarded the abnormal peak occurring on 13th February 2020, since more than 15,000 new cases were reported in China in just one day, most likely representing not confirmed cases); and (C) Timeline of the main events occurring during the pandemic, and indicating the MRCA of all SARS-CoV-2 genomes; the dotted area is a schematic representation of the real diversity values reported in **Supplementary Material Figure S2**. Divergence dates between SARS-CoV-2 and bat sarbecovirus reservoir and between bat and pangolin coronavirus were taken from (*7*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Inter-specific comparisons of sequence identities between different species, including pangolin (*Manis javanica*) and bat (*Rhinolophus affinis*) against the HQ SARS-CoV-2 dataset. Comparisons involved 3,478 SARS-CoV-2 genomes against the coronaviruses indicated in the table. ID refers to identity number in GISAID (GS; omitting the prefix "EPI-ISL-") and GenBank (GB). NC_004718.3 corresponds to the reference SARS Coronavirus genome (*14*). The genome #402131 corresponds to RaTG13 that has been used in the literature as bat coronavirus reference. GISAID 414518 and 420923 correspond to coronavirus analyzed from a dog and a tiger (*Panthera tigris jacksoni*) that where infected by human SARS-CoV-2. Abbreviations are as follows: Time: refers to the collection year of the specimen. Dif: Mutational differences of the coronavirus indicated when compared to the SARS-CoV-2 references sequence (MN908947.3). Id: average identity of the HQ SARS-CoV-2 against the corresponding coronavirus in the table. SD: standard deviation of Dif values. Max and Min: maximum and minimum identities shown by a SARS-CoV-2 genome with the other coronaviruses. "N" means ambiguity.

Species	ID	Place	year	Dif	Id (%)	SD	Max	Min	C8782T	C18060T	T28144C
SARS	GB: NC_004718.3	Toronto; Canada	2004	4576	79.26	0.05	79.67	78.64	False	True	False
Pangolin	GS: 410544	Guangdong; China	2019	1699	92.35	0.03	92.51	91.47	True	False	True
Pangolin	GS: 410721	Guangdong; China	2020	2599	90.21	0.03	90.44	89.56	True	False	True
Pangolin	GS: 412860	China	2019	2320	90.12	0.03	90.39	89.63	True	False	True
Pangolin	GS: 410539	Guangxi; China	2017	3720	85.35	0.04	85.59	84.74	True	True	True
Pangolin	GS: 410538	Guangxi; China	2017	3720	85.36	0.03	85.59	84.46	True	True	True
Pangolin	GS: 410543	Guangxi; China	2017	3495	85.24	0.04	85.47	84.55	True	True	N
Pangolin	GS: 410542	Guangxi; China	2017	3727	85.34	0.04	85.58	84.71	True	True	True
Pangolin	GS: 410541	Guangxi; China	2017	3721	85.35	0.04	85.58	84.72	True	True	True
Pangolin	GS: 410540	Guangxi; China	2017	3716	85.36	0.04	85.60	84.74	True	True	True
Bat	GS: 402131	Yunnan; China	2013	1105	96.17	0.02	96.37	95.53	True	True	True
Bat	GS: 412977	Yunnan; China	2019	1369	93.44	0.04	93.75	92.80	True	True	True
Bat	GS: 412976	Yunnan; China	2019	3827	76.87	0.05	77.31	76.69	False	False	False
Canine	GS: 414518	Hong Kong	2020	11	99.95	0.07	99.99	96.15	False	False	False
Tiger	GS: 420293	New York; USC	2020	7	99.97	0.07	100.00%	96.17	False	False	False
Human	GB: MN908947.3	Shangai; China	2020	0	99.98	0.07	100	96.18	False	False	False


bat/Yunnan/RaTG13/2013
SARS-CoV-2
B1
TMRCA$_{B1}$
12th November
2020
B1
B1
B1
B
B
B
B
B
B2
A
B
B
B
B
B4
B1
B1
B1a
B1
pangolin/Guangdong/1/2019


N. AMERICA
B1a
B1a1
A2a2b
A2a2a
A3
A2a
A1a
B1a
B4b
Direct introduction from Asia
Introduction from Europe
EUROPE
A2a4
A1a3a
A2a
A1a
A4
A4a
B3a
A2a
B4a
A2a
A2a
AFRICA
A2a
ASIA
A2a
A3
A2
A
B
A1
A1a
B4a
B4b
B3
B1
A1b
B5a
B3a
A2a
B3a
S. AMERICA
A2a
B3a
OCEANIA
B5a
A3b
A2a
A1a
A3a
B3a
B5a1a
B1a
A1
B5a1a
Multiple founders

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.097410; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.