Supplemental Material

Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders

INDEX

Supplemental Data/Results

References

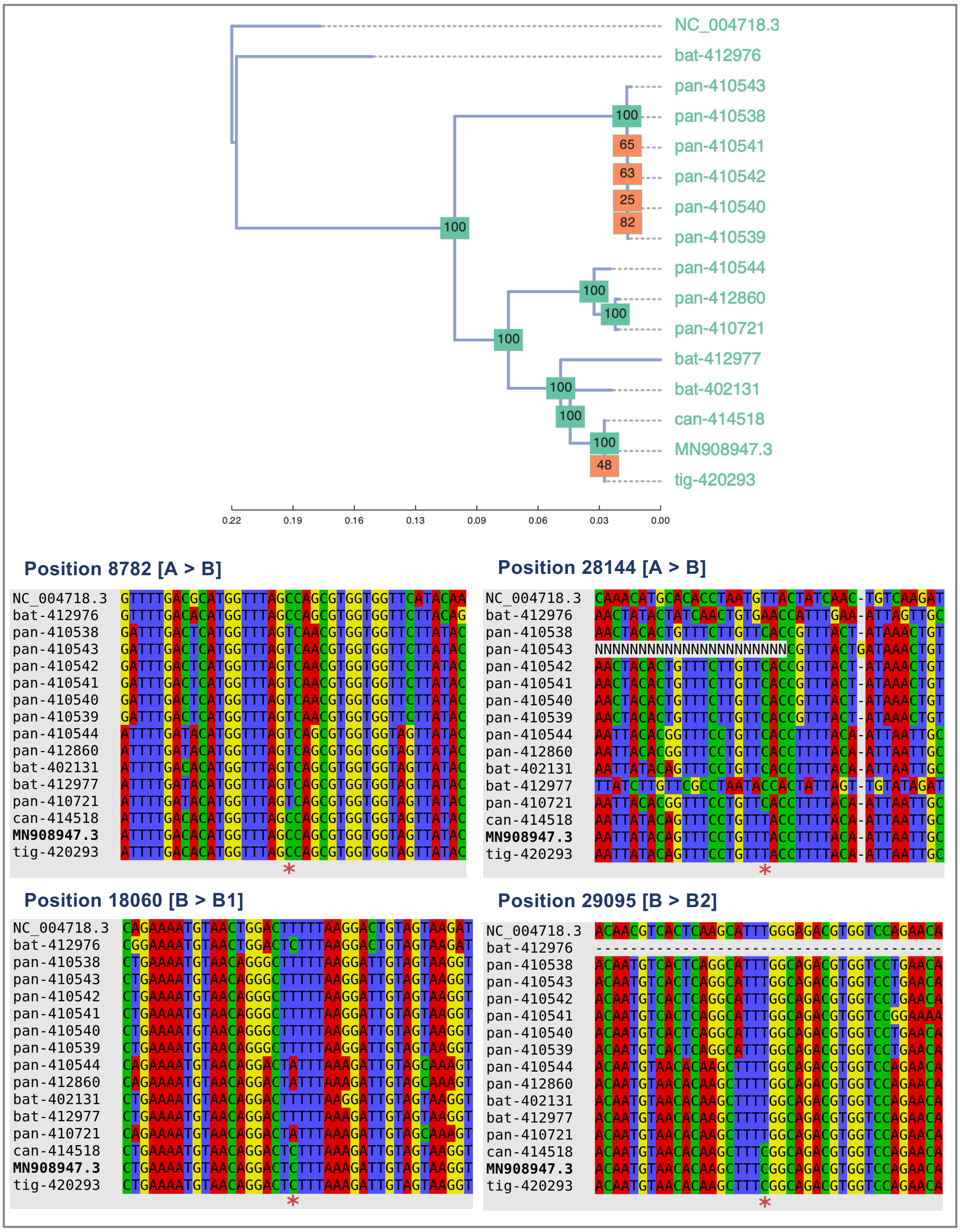
Supplemental Data/Results

The root of SARS-CoV-2 genomes

An initial ML tree was built using pangolin, SARS and bat coronavirus genomes as outgroups of SARS-CoV-2 genomes (see main text and **Figure S1**). We also constructed a ML tree using most related bat and pangolin sequences in order to investigate the root of SARS-CoV-2 genomes. This tree depicts B1 genomes most closely related to bat coronaviruses (**Figure 1** in main text).

In agreement with the ML tree (**Figure 1** in main text), we observed that non-human coronavirus sequences from pangolin and bat carry the three transitions C8782T–C18060T–T28144C that identify haplogroup B1 (see below on maximum parsimony tree). There are two exceptions in this alignment. First, the SARS genome does not present two of the three mutations (**Figure S1**). It is worth mentioning that there is a large divergence time (which means plenty of time for reversions) between SARS and SARS-CoV-2 genomes, and this poses difficulties for a correct homologous sequence alignment. Second, the genome sequenced from bat #412976 also lacks the three characteristic mutations, but this sequence is of very low quality, as can be visually appreciated in the alignment (**Figure S1**) and also in its atypical low identity value to SARS-CoV-2 when compared to the other bat genomes (**Table 1** in main text).

When using genomes sampled before 15th January 2020 (thus B1 representatives are not included in the dataset, see main text), the new root inferred from a new ML tree is set to haplogroup B2 (**Figure S2**). Note that non-human coronavirus sequences from pangolin and bat also carry the transition C29095T that lead from B to B2 (see alignment in **Figure S1**).



Position 8782 [A > B]

```
NC_004718.3  GTTTTGACGCATGGTTTAGCCAGCGTGGTGGTTCATACAA
bat-412976   GTTTTGACACATGGTTTAGCCAGCGTGGTGGTTCTTACAG
pan-410538   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410543   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410542   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410541   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410540   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410539   GATTTGACTCATGGTTTAGTCAACGTGGTGGTTCTTATAC
pan-410544   ATTTTGATACATGGTTTAGTCAGCGTGGTGGTAGTTATAC
pan-412860   ATTTTGATACATGGTTTAGTCAGCGTGGTGGTAGTTATAC
bat-402131   ATTTTGACACATGGTTTAGTCAGCGTGGTGGTAGTTATAC
bat-412977   ATTTTGATACATGGTTTAGTCAGCGTGGTGGTAGTTATAC
pan-410721   ATTTTGATACATGGTTTAGTCAGCGTGGTGGTAGTTATAC
can-414518   ATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATAC
MN908947.3   ATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATAC
tig-420293   ATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATAC
                                 *
```

Position 28144 [A > B]

```
NC_004718.3  CAAACATGCACACCTAATGTTACTATCAAC-TGTCAAGAT
bat-412976   AACTATACTATCAACTGTGAACCATTTGAA-ATTAGTTGC
pan-410538   AACTACACTGTTTCTTGTTCACCGTTTACT-ATAAACTGT
pan-410543   NNNNNNNNNNNNNNNNNNNNNNCGTTTACTGATAAACTGT
pan-410542   AACTACACTGTTTCTTGTTCACCGTTTACT-ATAAACTGT
pan-410541   AACTACACTGTTTCTTGTTCACCGTTTACT-ATAAACTGT
pan-410540   AACTACACTGTTTCTTGTTCACCGTTTACT-ATAAACTGT
pan-410539   AACTACACTGTTTCTTGTTCACCGTTTACT-ATAAACTGT
pan-410544   AATTACACGGTTTCCTGTTCACCTTTTACA-ATTAATTGC
pan-412860   AATTACACGGTTTCCTGTTCACCTTTTACA-ATTAATTGC
bat-402131   AATTATACAGTTTCCTGTTCACCTTTTACA-ATTAATTGC
bat-412977   TTATCTTGTTCGCCTAATACCACTATTAGT-TGTATAGAT
pan-410721   AATTACACGGTTTCCTGTTCACCTTTTACA-ATTAATTGC
can-414518   AATTATACAGTTTCCTGTTTACCTTTTACA-ATTAATTGC
MN908947.3   AATTATACAGTTTCCTGTTTACCTTTTACA-ATTAATTGC
tig-420293   AATTATACAGTTTCCTGTTTACCTTTTACA-ATTAATTGC
                                 *
```

Position 18060 [B > B1]

```
NC_004718.3  CAGAAAATGTAACTGGACTTTTTAAGGACTGTAGTAAGAT
bat-412976   CGGAAAATGTAACTGGACTCTTTAAGGACTGTAGTAAGAT
pan-410538   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410543   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410542   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410541   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410540   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410539   CTGAAAATGTAACAGGGCTTTTTAAGGATTGTAGTAAGGT
pan-410544   CAGAAAATGTAACAGGACTATTTAAAGATTGTAGCAAAGT
pan-412860   CAGAAAATGTAACAGGACTATTTAAAGATTGTAGCAAAGT
bat-402131   CTGAAAATGTAACAGGACTTTTTAAGGATTGTAGTAAGGT
bat-412977   CTGAAAATGTAACAGGACTTTTTAAAGATTGTAGTAAGGT
pan-410721   CAGAAAATGTAACAGGACTATTTAAAGATTGTAGCAAAGT
can-414518   CTGAAAATGTAACAGGACTCTTTAAAGATTGTAGTAAGGT
MN908947.3   CTGAAAATGTAACAGGACTCTTTAAAGATTGTAGTAAGGT
tig-420293   CTGAAAATGTAACAGGACTCTTTAAAGATTGTAGTAAGGT
                                 *
```

Position 29095 [B > B2]

```
NC_004718.3  ACAACGTCACTCAAGCATTTGGGAGACGTGGTCCAGAACA
bat-412976   ----------------------------------------
pan-410538   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCTGAACA
pan-410543   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCTGAACA
pan-410542   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCTGAACA
pan-410541   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCGGAAAA
pan-410540   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCTGAACA
pan-410539   ACAATGTCACTCAGGCATTTGGCAGACGTGGTCCTGAACA
pan-410544   ACAATGTAACACAAGCTTTTGGCAGACGTGGTCCAGAACA
pan-412860   ACAATGTAACACAAGCTTTTGGCAGACGTGGTCCAGAACA
bat-402131   ACAATGTAACACAAGCTTTTGGCAGACGTGGTCCAGAACA
bat-412977   ACAATGTAACACAAGCTTTTGGCAGACGTGGTCCAGAACA
pan-410721   ACAATGTAACACAAGCTTTTGGCAGACGTGGTCCAGAACA
can-414518   ACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACA
MN908947.3   ACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACA
tig-420293   ACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACA
                                 *
```

Figure S1. Sequence alignments for the segments containing the variants (stars located at the bottom of the alignments) that characterized the candidate roots in B (C8782T and T28144C), B1 (C18060T) and B2 (C29095T) of all SARS-CoV-2 genomes according to ML analysis (see also main text), and a phylogenetic tree that relates the SARS-CoV-2 human coronavirus reference (MN908947.3) to other coronaviruses in bat, pangolin ("pan") and SARS (NC_004718.3). We included other human coronavirus sequence segments that were sequenced from an infected tiger and a dog. Alignment is complicated for NC_004718.3 with SARS-CoV-2 due to the reduced identity between them, and this is particularly evident for the segment around position 28144.

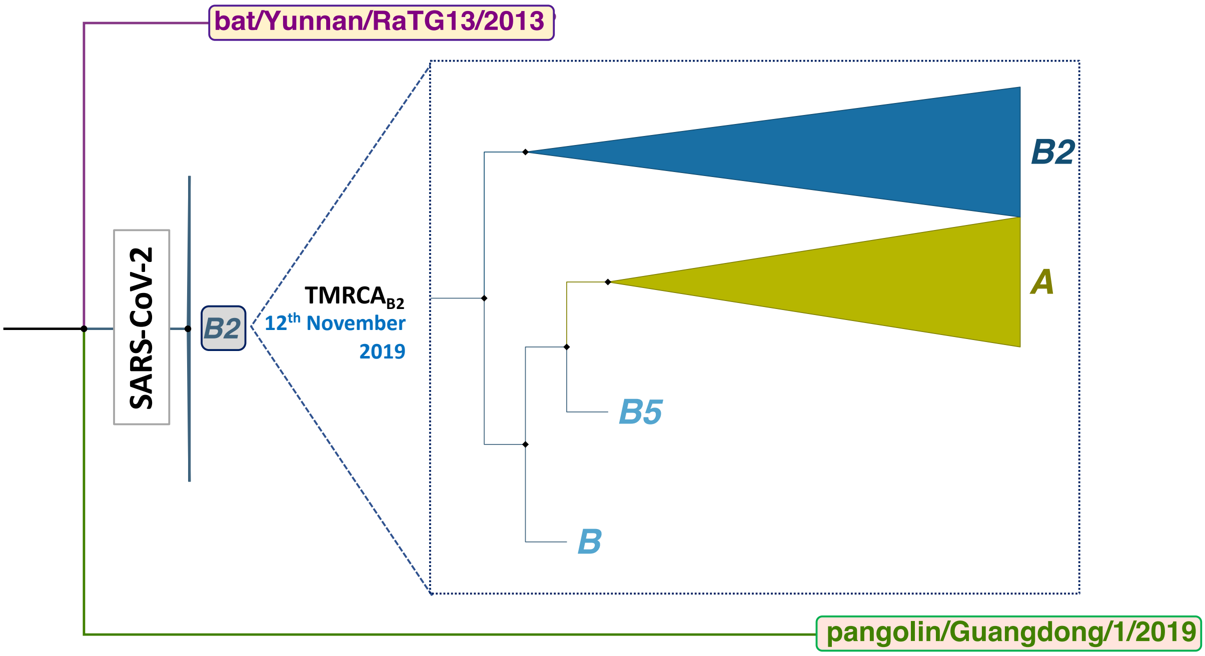


Figure S2. Inter-specific ML tree indicating the root of all existing SARS-CoV-2 genomes in B2 using genomes sampled in GISAID before 15 January 2020.

Intra-specific phylogeny of SARS-CoV-2

We aimed at reconstructing a solid phylogenetic skeleton for SARS-CoV-2 genomes that would allow the identification of diagnostic mutations for main branches. First, we used a parsimony approach which has been demonstrated to be very useful to build one of the better known phylogenies, namely, the human mtDNA tree (van Oven and Kayser 2009); **Figure 3** in main text.

Next, a ML tree of all SARS-CoV-2 genomes in the database was built as in **Figure 1** (main text) in order to evaluate the robustness of the most parsimonious tree built manually. As sown in **Figure S3**, the ML phylogeny reproduces the branching patterns observed using parsimony.

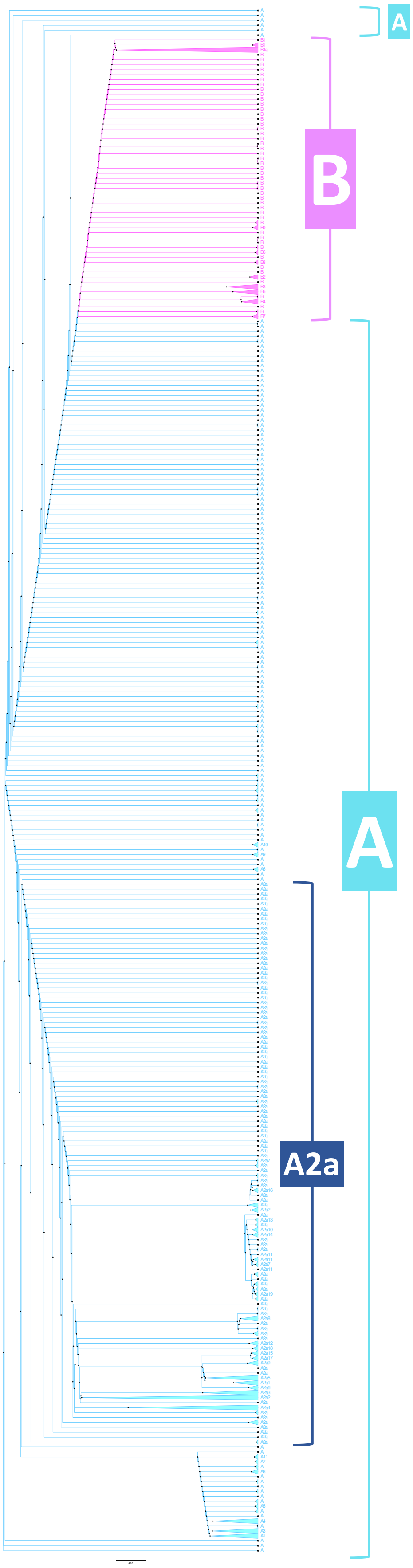


Figure S3. ML tree of SARS-CoV-2 all HQ genomes in the dataset

To further evaluate the robustness of the parsimonious tree, we built a MDS on unique SARS-CoV-2 genomes using the average pairwise discrepancy values for (a) all mutational variants (**Figure S4A**) and (b) only the diagnostic positions of the parsimonious tree (**Figure S4B**). The patterns for both analyses in the MDS plots resemble those inferred from phylogenetic trees.

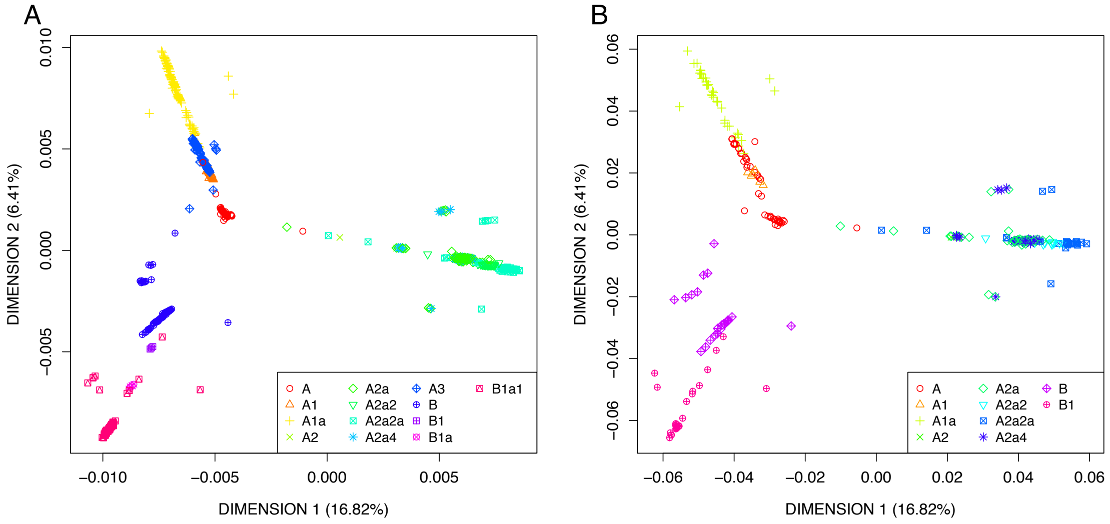


Figure S4. MDS plot of SARS-CoV-2 genomes (A) using all mutational variants in the genomes, and (B) using only the diagnostic sites of the phylogeny.

A number of studies agree with the definition of two main branches separated by only two transitions: C8782T–T28144C; although authors use different names for these branches, we follow the labeling system initiated by Nextstrain, given its popularity among virologists and other specialists. The vast majority of the genomes analyzed in the present study (99.29%) could be unambiguously classified into one of the 164 (sub)haplogroups defined. Only five genomes closely related to the reference sequence could not be classified into either A or B, and there are 19 additional genomes that have a problem of resolution but only at the tips of the phylogeny. The mutational pathways defining each of these clades are described in **Supplemental Table S2**.

By counting the occurrence of mutations along the tree branches and those at the tips of the phylogeny (see Methods), it is possible to detect the positions that are mutationally stable from those that are mutational hotspots

(**Supplemental Table S1; Figure S5**). The pattern of occurrences can be summarized as follows:

a) There are 2,147 substitutions (**Supplemental Table S1**), of which 1,749 (81.46%) occurred only once, and 284 (13.23%) twice in the SARS-CoV-2 genomes; therefore, 94.69% of the mutations occurred no more than twice in SARS-CoV-2 genomes.

b) Mutations along branches in the phylogenetic tree are very stable: there are 185 diagnostic sites in the tree branches (**Supplemental Table S1**), 110 (59.46%) were singletons, and 154 occurred twice at most (83.24%).

c) There are a few mutational hotspots in the phylogeny; the more important hotspots are C575T (15 hits), G11083T (15 hits), T13402G (13 hits), and A4050C. Note that some rapid mutated mutations, although unstable, can still be diagnostic variants of sub-haplogroups when accompanying other more stable variants (otherwise these changes alone should not be used to define new sub(clades). This is the case of e.g. mutation G11083T with 15 total hits in the phylogeny (three of them accompanying another diagnostic mutation in the tree; this is one of the two most unstable positions in the phylogeny); in combination with the diagnostic positions of A1+C14805T, it is diagnostic for e.g. haplogroup A1a: out of the 269 genomes having the sequence motif A1+C14805T, 257 also have G11083T and 19 have only the G11083T mutation.

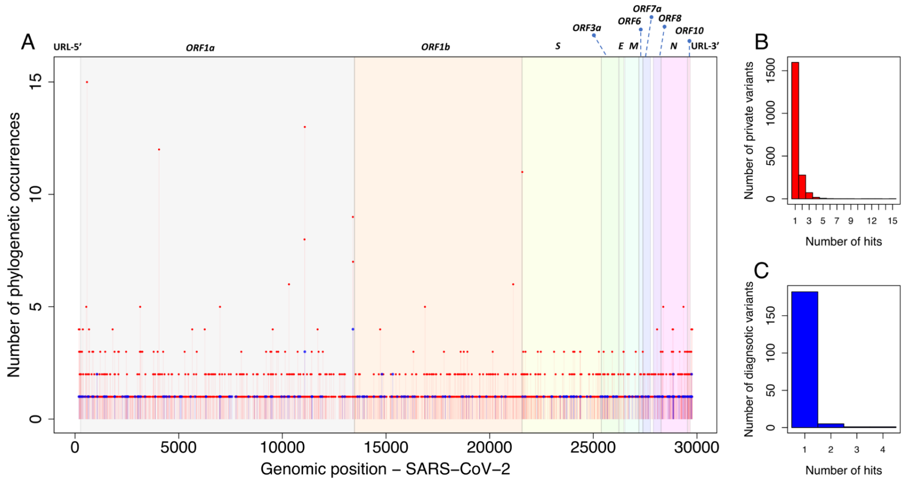


Figure S5. (A) Spectra of mutation occurrences along the maximum parsimony tree of **Figure S4**. (B) Number of private variants. (C) Number of mutations located along the branches of the tree (diagnostic variants).

Mutational changes are distributed quite homogeneously across the SARS-CoV-2 genomes. In fact, there are very short gaps in the genome of the virus that were not hit by mutations (**Figure S6**).

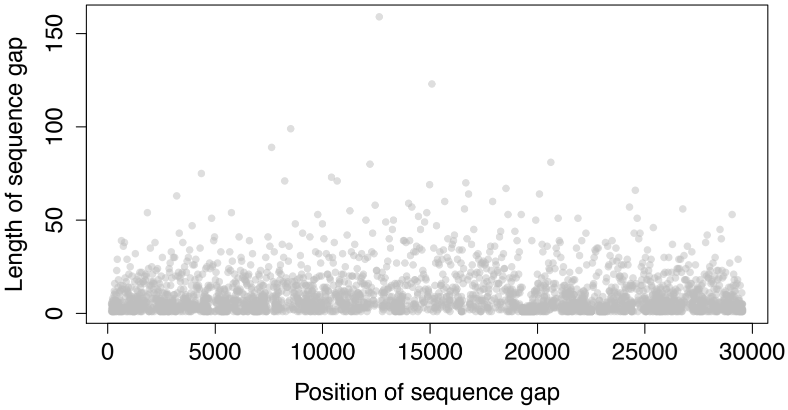


Figure S6. Location of the gaps that were not hit by mutations (grey dots) and their length in the genome of SARS-CoV-2.

The number of sequences added to the GISAID database increased with the growing impact of the pandemic worldwide. The main contribution of genomes to GISAID coincides with the outbreak outside Asia, leading to an exponential growth of the number of sequences belonging to haplogroup A (**Figure S7**).

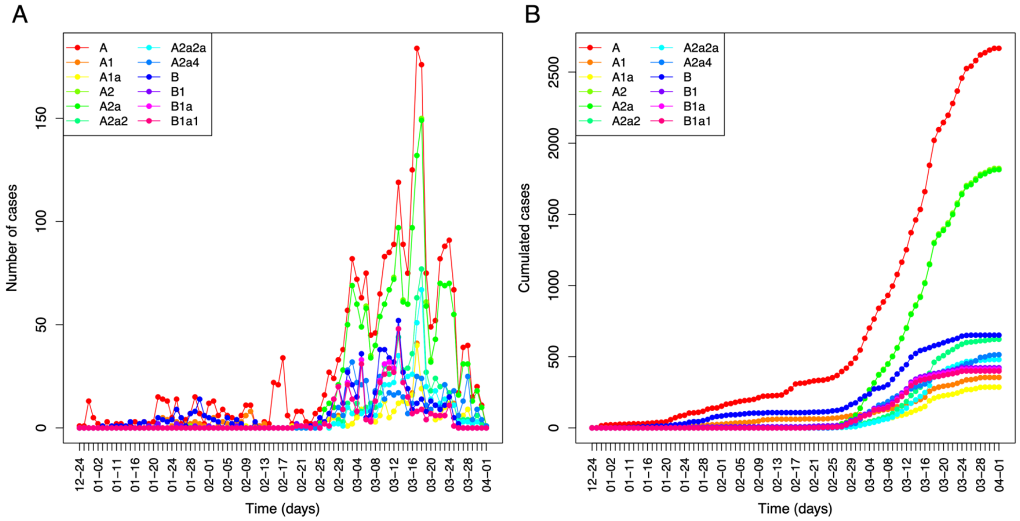


Figure S7. Evolution of the number of sequences over time and classified by main (sub)haplogroups.

Analysis of diversity indices was also performed on main haplogroups. While π values are almost the same in all regions and haplogroups, HD values are more variable between regions (**Figure S8**).

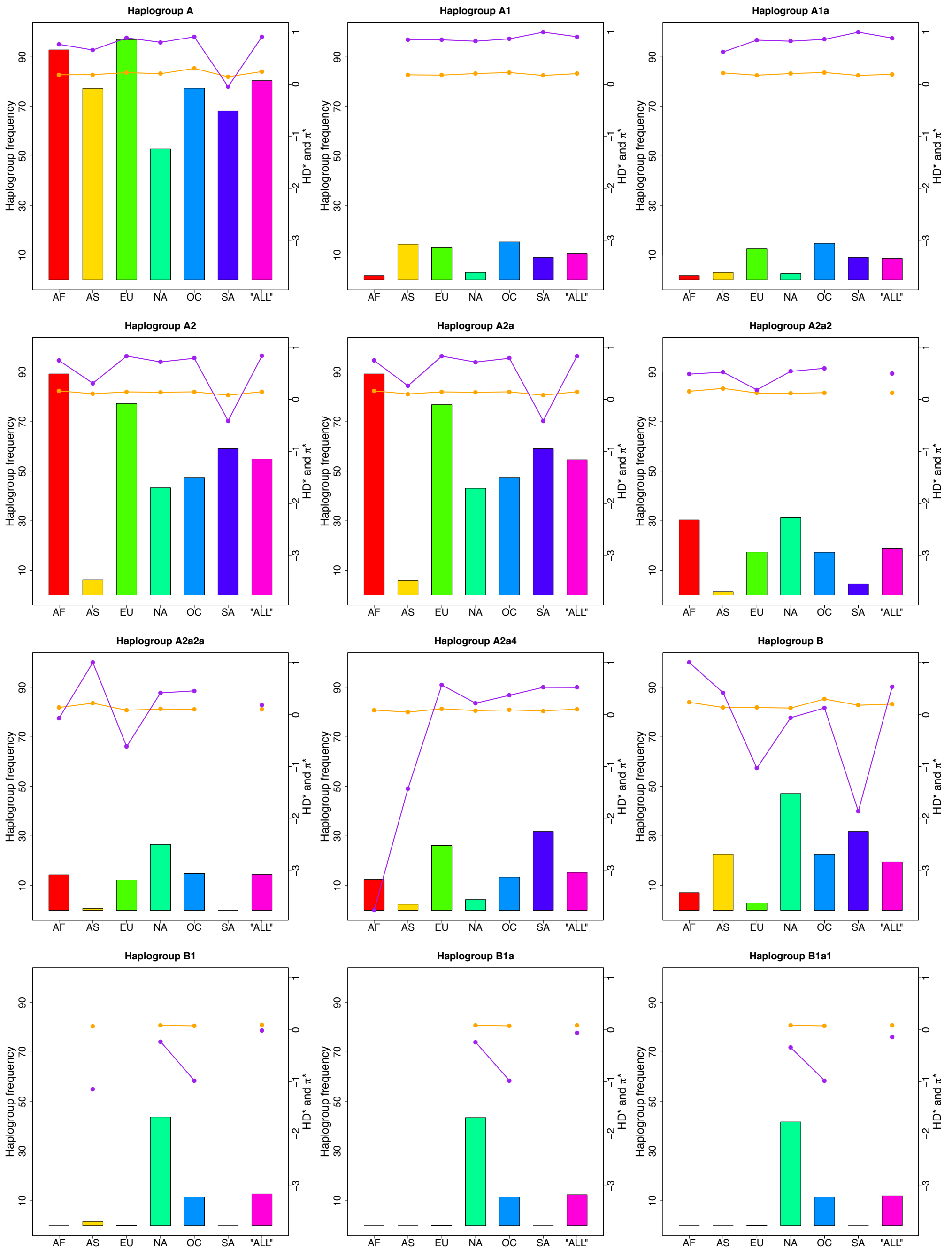


Figure S8. Diversity indices computed for main haplogroups. Purple lines connecting dots indicate sequence diversity, and orange lines nucleotide diversity. In order to present both indices together, we depicted values indicated of π/1000 (π*) and *HD*×10^{-9} (*HD**). Haplogroup frequencies in the main regions are also represented as bar colors.

Considerations on the parsimonious phylogeny

We built a reference parsimonious phylogeny for SARS-CoV-2 genomes. It reveals that: a) most of the mutations are phylogenetically stable, showing only one or two mutational occurrences in the phylogeny; b) there are only a few mutational hotspots that should not be used to define phylogenetic branches unless accompanied by other more stable mutations; the data available in the present study do not allow to infer if these substitution hotspots relate to different transmission rates and/or infectivity of SARS-CoV-2; c) mutational instability is detected in some positions located at the tips of the phylogeny; further research is needed in order to determine if this phylogenetic noise is due to mutational instability, sequencing errors or recombination. In this regard, it is important to note that there is an important number of ambiguities in the HQ dataset. Moreover, many different sequencing platforms have been used to sequence SARS-CoV-2 genomes, each being prone to specific sequencing artifacts.

It can be anticipated that the quite uniform distribution of mutations along the genome of SARS-CoV-2 (**Figure S6**) might be a challenge for the development and efficiency of future vaccines.

In addition, the basal node of B1 or B2 (according to ML) or A (according to genome chronology; see main text) are the candidate roots of SARS-CoV-2; meaning that the index patient most likely carried one of these virus strains.

A practical application of the SARS-CoV-2 tree built in the present study is to facilitate classification of genomes into clades, which might facilitate the work of epidemiologists and other specialists aimed at establishing potential correlations between different clade members and the different clinical phenotypes observed in COVID-19, disease severity and differential spread of the disease worldwide. The phylogeny presented is scalable, and nomenclature works in a hierarchical way similar to that demonstrated to be successful in other research areas such as human population genetics (e.g. mtDNA studies).

Further considerations on phylogeographic patterns of SARS-CoV-2 genomes

According to GISAID, the first genome to be sequenced was sampled in a patient from China (#402123) on 24 December 2019. A few weeks later, more than a dozen several other genomes had been obtained from Chinese patients from the Hubei province. The first genome sequenced outside Asia corresponds to a sample extracted from a USA patient in Washington (#404895; 19 February 2020). Soon, many other genomes were sequenced from USA, Oceania, Europe, etc. In the database used in the present study, there are genomes sequenced from >62 countries representing the main continental locations: Africa (1.7%), Asia (15.3%), Europe (46.4%), North America (25.3%), and South America (0.6%).

Worldwide patterns of some haplogroup frequencies are summarized in the maps in **Figure S9**.

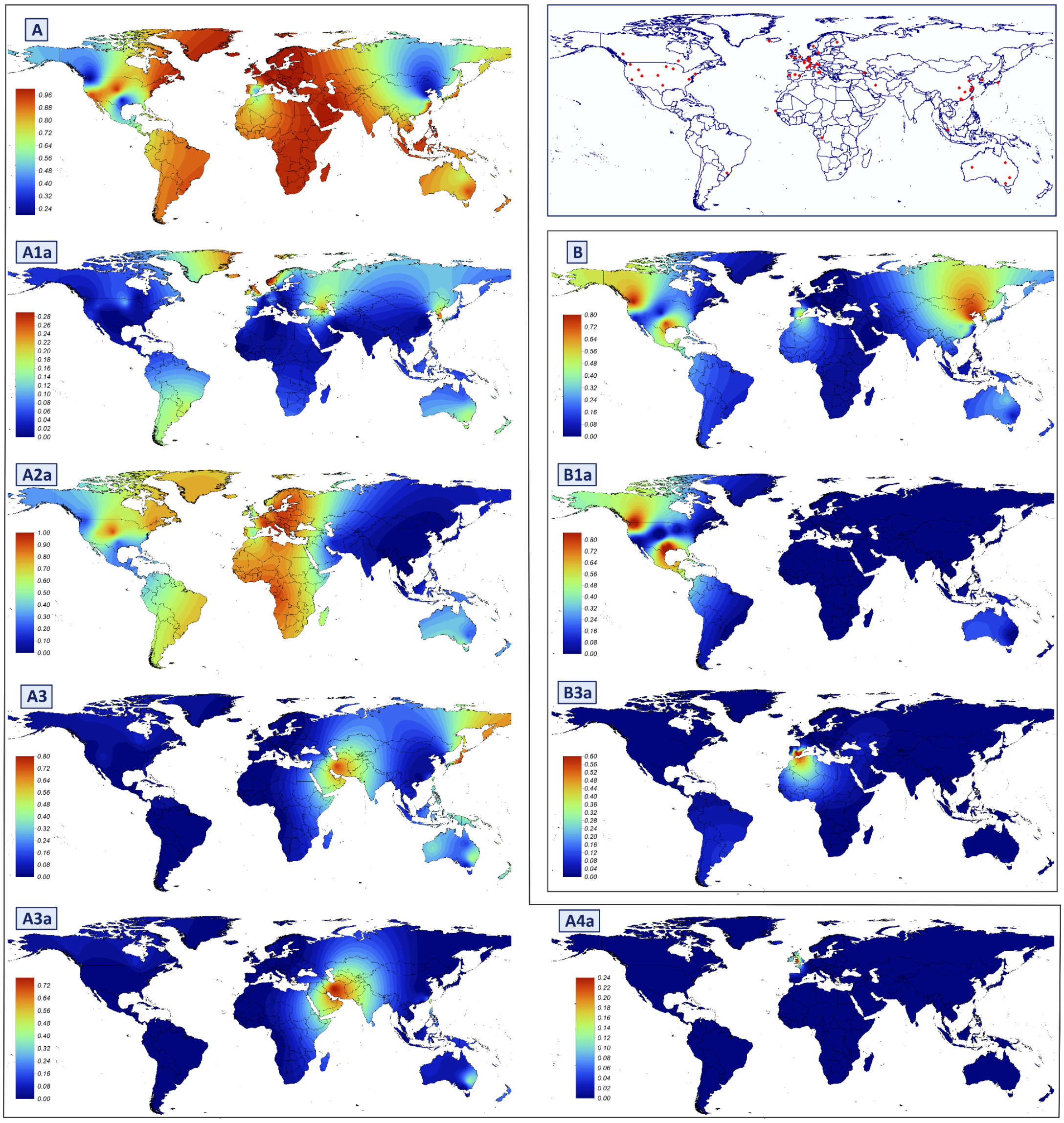


Figure S9. Worldwide maps of interpolated haplogroup frequencies. Only the most common (sub)clades are depicted.

Molecular variation of SARS-CoV-2 genomes

We identified a large number of indels in our dataset: 2,209 insertions and 25 deletions. We also counted 2,159 substitutions and 49 Multi-Nucleotide position Polymorphisms (MNPs). From the data available, it is not possible to reconstruct the mutational process leading to MNP variants; for instance, GGG28881AAC could be interpreted as a single mutational event (MNPs) or e.g. as the concatenation of three independent mutational events: G28881A+G28881A+G28881C. Here we opted for the most parsimonious solution, namely considering them as single mutational events (in contrast to e.g. CNCB, which favors the latter option). Although the parsimonious interpretation entails the risk of under-estimating overall and allele-specific mutation rates, however, the few MNPs existing in the database would only increase the number of mutations from 49 to maximum 195 events. In this regard, there is evidence to suggest that mutations in MNPs are not independent (Rosenfeld et al. 2010). Unless indicated, indels and MNPs were disregarded in most of the analyses.

The substitutions observed in SARS-CoV-2 genomes show the following pattern: (a) transitions (*n* = 1,532), with 72% being pyrimidine transitions (C <–> T) and 28% purine transitions (A <–> G); and (b) transversions (*n* = 649), with 76% being purine to pyrimidine, and 24% pyrimidine to purine. The *ts*/*tv* ratio is 2.37 (**Supplemental Table S7**). In a comparison between humans and chimpanzees, Ebersberger et al. (2002) found a *ts*/*tv* ratio of 2.4 – in agreement with our *ts*/*tv* ratio. Further, according to DePristo et al. (2011), the *ts*/*tv* ratio is expected to be 2.1 for whole-genome sequencing and 2.6-3.3 for exome sequencing in human studies, while Zook et al. (2014) have stated that very low ratios (0.5) generally point to sequence artifacts. Therefore, our ratio falls within the expected range for a good quality dataset in terms of substitutions. The *ts*/*tv* ratio was similar when considering the main clades of SARS-CoV-2 (according to the phylogeny described below): namely, 2.44 for haplogroup A, 2.74 for haplogroup B, and 2.42 for the main sub-haplogroup in the database, A2a (**Supplemental Table S7**). We also investigated the pattern of mutations by genes (**Supplemental Table S8**). The highest *ts*/*tv* ratio occurs in genes ORF1a (3.28) and *E* (3.25), and the lowest one in gene *ORF6* (1.11) and the intergenic regions (1.03); these differences are statistically significant under a Fisher's exact test for the comparisons involving *ORF1a versus S*, *ORF3a* and Intergenic (*P*-value < 0.0001) and considering a strict Bonferroni adjustment.

There is a much higher number of non-synonymous changes (*n* = 1,293, 62.17%; including missense, start lost and stop gain) compared to synonymous changes (*n* = 787); the non-synonymous/synonymous changes ratio is 0.62. This ratio was also similar when partitioning by main haplogroups; namely, 0.63 for haplogroup A, 0.60 for haplogroup B, and 0.62 for haplogroup A2a (**Supplemental Table S7**). There is a notable difference between the ratio of non-synonymous/synonymous mutations when analyzed by gene: the maximum value was obtained for gene *ORF6* (0.84), and the minimum for gene *M* (0.41); the most significant difference was noted between genes *ORF6* and *M*. However, a Fisher's exact test did not reach significance when adjusted for multiple testing. We also tested for signals of natural selection in SARS-CoV-2 genes by studying K_a/K_s index in an interspecific context that includes genomes from pangolin, bat, and SARS coronavirus. For almost all genes, values are below 1, with *ORF1b* being the gene with the lowest value, suggesting the action of purifying selection on these genes. Only *ORF10* genes have a value above 1, suggesting the action of positive selection operating on this gene (**Supplemental Table S6**).

We counted 1,741 unique genomes in the database. It is remarkable that a number of sequences were present at high frequencies; for instance, the reference sequence belonging to haplogroup A, is repeated 78 times; most of them sampled in Asia (76.8%; mostly in China, 60.8%; see below).

Haplotype and nucleotide diversities, as well as Tajima's *D* test values were computed by geographic regions, haplogroups and genes (**Supplemental Table S3**). Diversity was very similar in the different continental regions, with very minor differences with respect to the global sample; the only exception is Oceania (represented by Australia and Oceania), which shows notable high diversity values (π = 3.63E-04; *HD* = 9.90E-01). Tajima's *D* values are statistically significant (below -2) in all regions with the exception of South America. Diversity indices and Tajima's *D* statistics are very similar when computed by main haplogroups (**Supplemental Table S3**). However, diversity values are particularly different when comparing genes. Thus, for instance, *ORF8* shows 11.8 times higher nucleotide diversity (π = 1.07E-03) compared to *E* (π = 9.07E-05); while *ORF1a* shows 47.9 times higher sequence diversity (*HD* = 9.30E-01) than *E* (*HD* = 1.94E-02). *ORF8* and *ORF1a* are the most diverse genes, while *E* shows the lowest diversity. In contrast, Tajima's *D* values show minor differences between genes.

We investigated mutation variation at two notable sequence features of all SARS-CoV-2 genomes. First, the receptor binding domain (RBD) located in the spike protein, which has been reported to be the most variable part of the coronavirus genome (Andersen et al. 2020; Wan et al. 2020); these studies indicated that there are six amino acids that are critical for binding to the angiotensin-converting enzyme 2 (*ACE2*) receptor, namely, L455 (pos. 22925-22927), F486 (pos. 23018-23020), Q493 (pos. 23039-23041), S494 (pos. 23042-23044), N501 (pos. 23063-23065), and Y505 (pos. 23075-23077). Second, we also examined variation at the 12 characteristic nucleotide insertion (amino acid sequence PRRA; pos. 23606-23620), which constitutes a polybasic furin cleavage site (PFCS) that is also related to three adjacent predicted O-linked glycan residues: S673 (pos. 23579-23581), T678 (pos. 23594-23596), and S687 (pos. 23621-23623). The low diversity found in these regions is noteworthy. We only found two mutations at the cleavage site (which is present in all the SARS-CoV-2 genomes), namely G23607A (as a non-synonymous change [CGG>CAG or R>Q] private variant in haplogroup A1; GISAID #415709) and 23611G>A (as private synonymous variant of haplogroup A2a2c in GISAID #418390).

We also analyzed the evolution of the diversity indices with time (**Figure S10**; considering only time-points with a minimum accumulated number of 10 sequences). The first genomes sequenced correspond to those sampled in China in late December 2019, followed by genomes from other Asian countries. As expected, sequence and nucleotide diversity experienced an initial rapid growth until 19 January 2020. However, diversity values experienced a very remarkable drop (especially in π; **Figure S10A**) starting on 21 January 2020 and persisting for the next 4 days. Afterwards, diversity values progressively grow again from 25 January until reaching the highest values and then a plateau. In non-Asian regions, the pandemic had a delay of a few days (from about 28 January in North America and Europe), but it followed a similarly continuous increase that overtook Asian diversity values. The very high diversity values observed in Oceania are particularly striking (**Figure S10B**) compared to those of other continental regions. Tajima's *D* values are significantly negative in Asia from the initial outbreak; the growth of this index values is slower in other regions, but it eventually reaches similar values everywhere (see also **Supplemental Table S3**).

For the beginning of the pandemic, there are only Asian SARS-CoV-2 genomes available in public repositories; however, the number of sequences increases progressively in every region, with Europe being by far the region that has contributed more genomes to the GISAID database, followed by North America (**Figure S10D**).

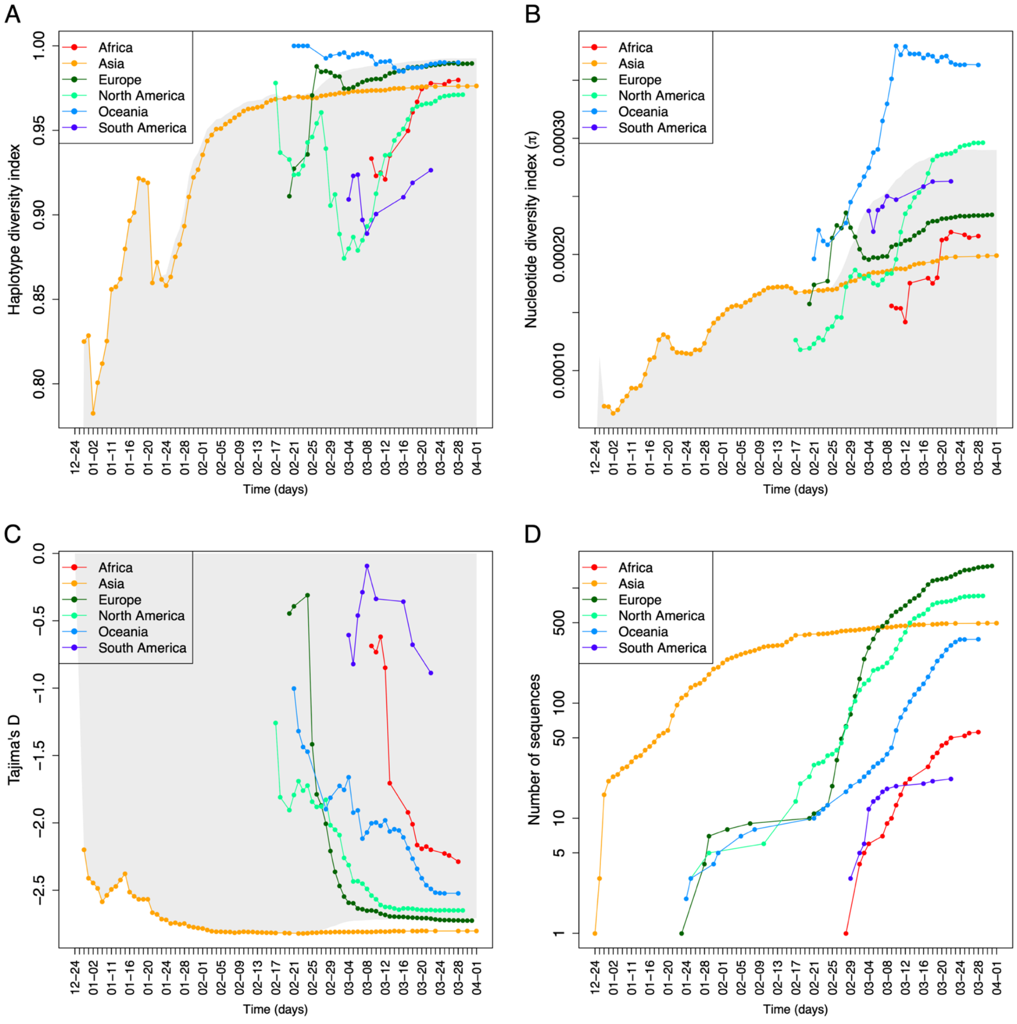


Figure S10. Accumulated diversity indices by sampling date in the main regions for the main SARS-CoV-2 phylogenetic branches. The shadowed background area represents the corresponding values for the whole sampling dataset. In the Y-axis are represented Haplotype diversity (**A**), Nucleotide Diversity (**B**), Tajima *D* (**C**) and Number of sequences (**D**) values.

Further considerations on super-spreaders and founder effect

A total of 49 haplotypes appear at least 7 times in the database (**Table S4**). Almost all of them, with the exception of the reference sequence (#H4), which originated at the beginning of the pandemic in China, appear several weeks after the beginning of the Asian pandemic (**Figure S11**), coinciding with the non-Asian outbreak and indicating their important role in the rapid spread of the pandemic.

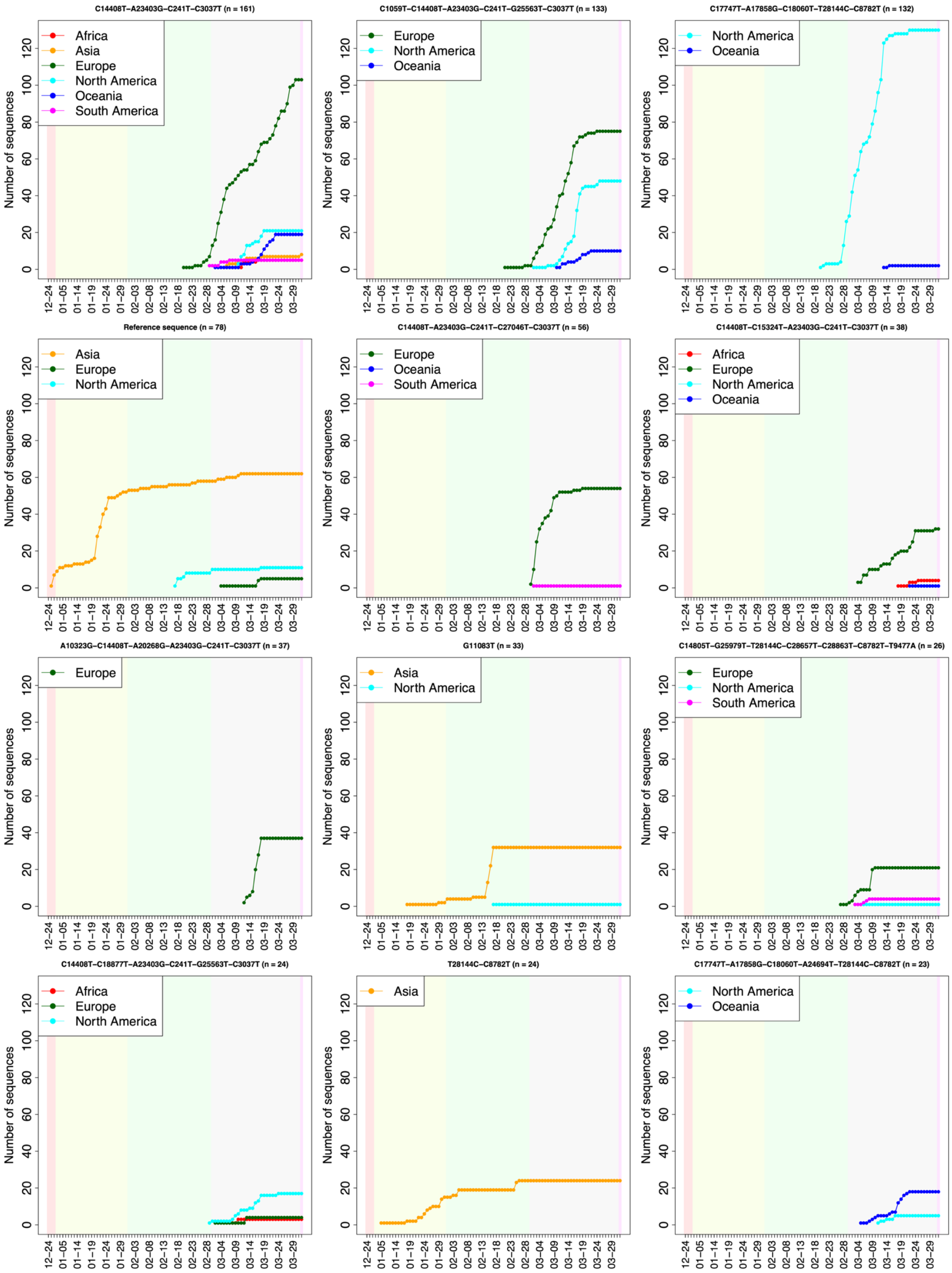


Figure S11. Occurrence of the nine most common haplotypes in the SARS-CoV-

2 dataset by regions and sampling date. Vertical background colors denote the sampling months.

We carried out network analyses of genome SARS-CoV-2 variation for the super-spreading event occurring in the Diamond Princess shipboard (see main text for more information) (**Figure S12**).

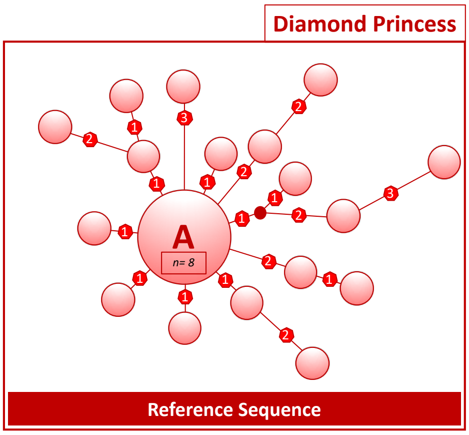


Figure S12. Network analysis of the Diamond Princess SARS-CoV-2 genomes. See legend of **Figure 5** from main text for further details.

Considerations on natural selection acting on SARS-CoV-2 genomes

We explored if patterns of SARS-CoV-2 spread worldwide could be explained by natural selection forces. First, we compared patterns of variation over time in different continental regions (**Figure S10**). Diversity values increased rapidly in China during the first outbreak. Next, diversity values drop significantly (especially true for the HD) probably coinciding with the role of the 'super-spreader' haplotype #H4 (reference sequence; see main text) coupled with human intervention in China, which slowed-down further diversification of lineages.

The outbreak outside China in other Asian countries and continents made the curve of cumulative diversity increase to high values of diversity. Again, human intervention worldwide led this exponential increase to reach a plateau. Values of diversity outside Asia overtook those in Asia, because control of the pandemic by other countries was less efficient than in Asia, as observed from the epidemiological data worldwide. Haplogroups A and B where the main lineages responsible for the initial outbreaks, while A2 (more specifically, A2a) sparked the main outbreak outside Asia. The behavior of the curve of Tajima's *D* index is most paradoxical. Tajima's *D* shows very low values from the very beginning and displays the same behavior, albeit with delay, in the other regions. Significant negative values of this index (below -2) are suggestive of purifying natural selection; however, negative values could also be compatible with heterogeneity of mutation rates and population expansion, both variables being present in the SARS-CoV-2 pandemic. At the same, human intervention on the disease could also mimic purifying selection.

We further explored the possible action of natural selection by investigating patterns of intra-specific K_a/K_s values by genes in groups of SARS-CoV-2 genomes that represent two poles of the pathogen genome variation i.e. haplogroups B1 and A2a. These values are suggestive of purifying selection acting on gene *ORF1a* (ω = 0.19742) and especially on gene *N* (ω = 0.0001); **Supplemental Table S6**. Only gene *S* shows a moderately positive value (ω = 1.04877), which could be suggestive of slightly positive selection acting on SARS-CoV-2.

The mutational changes differentiating haplogroups A and B do not seem to have relevant pathogenic effects; for instance, C18060T (B > B1), T28144C and C8782T (A > B) are all synonymous changes with low predicted severity (**Supplemental Table S9**).

Overall, there is suggestive evidence for a role of purifying selection operating during the spread of SARS-CoV-2 worldwide.

Association test of haplogroups with sex and age

Patterns of age and sex were analyzed by regions and main haplogroups (**Figure S13**). The median age was very similar for the different haplogroup pairwise comparisons considered (**Supplemental Table S5**) with the exception of non-A4a (50 [IQR: 35.0-63.0]) *versus* haplogroup A4a patients (76 [48.0-87.0)], non-B3a (50 [35.0-63.0]) and B3a (61 [50.5-82.0]. The highest female proportion was observed for haplogroup A3 (58.8), and the lowest for B3a (35.4).

Association tests were carried out to evaluate if a particular age group was more severely hit by specific SARS-CoV-2 lineages. An initial test was carried out for the main sub(haplogroups) and age, resulting in two significant associations under a Bonferroni correction, namely, haplogroup A4a (*n* = 39), and B3a (*n* = 51); in both cases, Mann-Whitney *U* test: *P*-value<0.001. Note however that haplogroups A4a and B3a are the haplogroups with the highest median age of all the ones compared but also those with the largest difference with the median age of the comparable groups. In order to account for the different haplogroup frequency and age patterns existing in each region, we carried out another test considering all the sampling and accounting by regions (*n* = 2,409); this analysis did not yield significant association (Kruskal-Wallis test: *P*-value = 0.6887).

Additionally, we tested if haplogroups could be related to sex. The most significant finding was observed for haplogroup A3 when compared to non-A3 lineages (n_{A3} = 184; Fisher's Exact Test, *P*-value<0.009); note however that this could be a false positive as A3 carriers are mainly from Asia and Oceania while non-A3 carriers appear mainly in other continents. More importantly, the observed *P*-value for this association did not surpass Bonferroni-corrected significance.

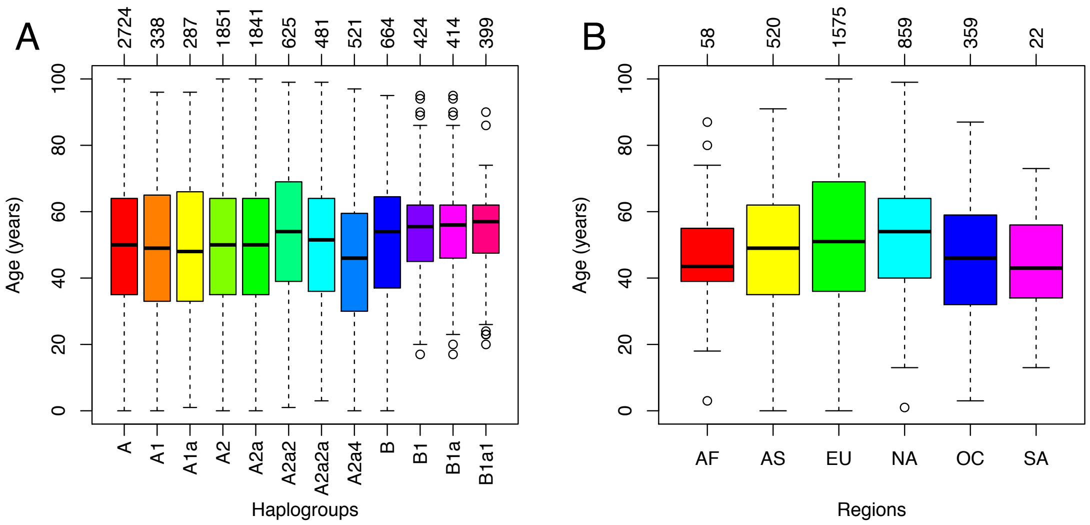


Figure S13. Distribution of age by haplogroups and main geographic regions. Numbers on the upper side of each panel indicate sample sizes.

The SARS-CoV-2 genome database

We initially investigated 4,716 SARS-CoV-2 genomes; a proportion of them (71.9%; *n* = 3,393) were identified by GISAID as having "<1% of NNN" and being complete (referred herein as the HQ dataset). We first explored the differences between the full dataset and the HQ sequences. An indirect indication of the quality of the sequences is the sequence length, which is on average 25,636 bp for the HQ dataset, compared to a significantly lower value for the full database, namely 23,736 bp. Moreover, the average number of ambiguities (including all kinds of IUPAC codes) per sequence in the full database is 6.5x the number of ambiguities in the HQ database (249.3 *versus* 37.9). The large number of ambiguities and indels in the low quality (LQ) data distorts sequence alignment and, consequently, a correct sequence annotation and phylogenetic reconstruction. Still, we counted an average of 43.5 ambiguities per genome in the HQ dataset. It is also important to mention that different platforms were used to sequence SARS-CoV-2 genomes, and corresponding differences between them are evident when analyzing the processed GISAID files (**Figure S14**). We examined several quality variables in the genomes sequenced using different NGS technologies, by taking 60 random genomes from GISAID, respectively 30 LQ and 30 HQ (information on quality cannot be datamined from GISAID but has to be extracted manually). The number of N's islands and the length of NNN's stretches are significantly higher in the LQ dataset, while the length of the genomes is higher in the HQ genomes after eliminating the ambiguities; the latter is partly due to the much higher number of ambiguities existing in the 5' and 3' ends in the LQ datasets. When using the LQ filter, Illumina yields a significant higher number of N's compared to Nanopore, but these differences disappear when looking at the HQ set (**Figure S14**). Indels are more common in the HQ dataset, and Nanopore seems to capture more than Illumina.

In order to minimize the effects of potential sequencing errors, only the HQ genomes were used for the subsequent analysis. In addition to LQ sequences, we also eliminated 53 of the HQ sequences lacking sampling date or having ambiguous haplogroup adscription (see below). Thereby, the final number of HQ genomes used for all the analyses was 3,393 (1,758 unique sequences; 51.81%). These HQ genomes represent six different continental regions: Africa (*n* = 58; 5 countries from North and Sub-Saharan Africa), Asia (*n* = 520; 14 countries from the Asian continent in a broad sense, including e.g. Middle East, South, East, Center), Europe (*n* = 1575; 29 countries), North America (*n* = 859; USA and Canada), Oceania (*n* = 359; Australia and New Zealand), and South America (*n* = 22; 4 countries). Five sequences could not be classified into either haplogroup A or B, and 19 have ambiguous sub-haplogroup classification; moreover, 48 did not have sampling date in the meta-data file; therefore, totals in tables and text do not always match.

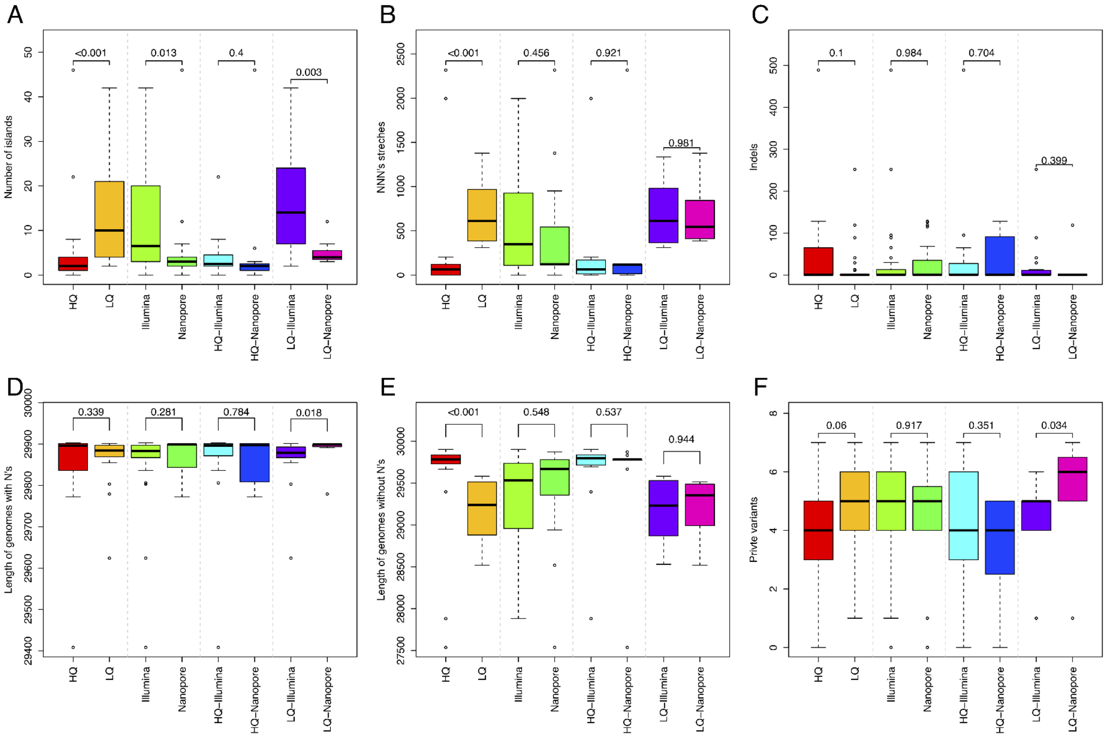


Figure S14. A total of 30 LQ and 30 HQ sequences were downloaded from GISAID, with information on their corresponding NGS technology retrieved manually from the database. Comparisons between technologies were carried out on the FASTA files already processed by GISAID.

Limitations of the present study

Several limitations of the present study are related to the inherent constraints of the meta-data available in GISAID, particularly those related to sampling records of the genomic sequences, which are nonetheless common to analysis and inferences generally made by researchers using these data.

First, some inferences rely on the sampling chronology of the genome sequences. There is no way of checking for deviations between the data recorded by GISAID and the real case dating; we assume however that these deviations should not seriously affect the inferences that were carried out at a global scale. Furthermore, we acknowledge that, as stated by Villabona-Arenas et al. (2020), the same phylogeny might be consistent with multiple transmission histories and thus, ideally, the reconstruction of transmissions should be supported by epidemiological and environmental factors, and human air-travel data; the phylogeographic patterns described in the present study are exclusively based on SARS-CoV-2 genome variation and associated meta-data, and therefore transmission routes of the virus worldwide should be corroborated in future studies with other sources of complementary data.

Second, in some instances, samples could have been obtained not at random from the affected population; it is however not possible to deduce the impact of a non-random sampling; the global scale of our study should account for deviations from randomness in sampling.

Third, sampling was limited at the beginning of the pandemic and thus it is possible that some key lineages holding key information on the origin of SARS-CoV-2 may have been missed; we anticipate that analysis of stored samples from patients treated at the beginning of the pandemic might help to address this limitation.

Fourth, we took the pragmatic decision of allocating the root of the tree in the most parsimonious SARS-CoV-2 tree to haplogroup A (**Figure 2** and **3**) based on a hypothesis that holds reasonable uncertainty but requires further investigations; this decision was necessary in order to allow a clade nomenclature that is coherent with the most basic cladistic rules (e.g. B is an ancestral clade of B1); should future research reveal a different root, modification of the present nomenclature would not require big adjustments due to the mutational proximity of the best candidates to allocate the root.

Fifth, we detected a number of super-spreader candidates in different continental locations; with the information available in the SARS-CoV-2 genomes it is not possible to differentiate which genomes in the basal node of the phylogenies describing these candidates (and those emerging from them) are the result of a super-spreader event with several carriers (horizontal transmission) from those that derived from basically a single super-spreader individual; however, their star-like phylogenies (and the statistical indices describing these topologies) coupled with the short time period all these genomes emerge in the database suggest an important role of super-spreader individuals in the COVID-19 pandemic, a proposition that is gaining more and more support from epidemiological observations and the high heterogeneity transmission rate associated to the disease.

Finally, we used the FASTA sequence information available in GISAID, which corresponds to the consensus sequence of the pathogen genome detected in a patient; however, there is a chance that the same patient carries different strains; this possibility does not seem to be the rule, and should not contribute to significant noise in the analysis carried out in the present study. In the same vein, recombination, which could occur when two different strains coincide in the same host, seems to be low as indirectly observed in our database, where most of the genomes could be corrected classified into haplogroups.

References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat Med* **26**: 450-452.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.

Ebersberger I, Metzler D, Schwarz C, Paabo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70**: 1490-1497.

Rosenfeld JA, Malhotra AK, Lencz T. 2010. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. *Nucleic Acids Res* **38**: 6102-6111.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30**: E386-E394.

Villabona-Arenas CJ, Hanage WP, Tully DC. 2020. Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* **5**: 876-877.

Wan Y, Shang J, Graham R, Baric RS, Li F. 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *Journal of virology* **94**.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32**: 246-251.