Mega-phylogeny sheds light on SARS-CoV-2 spatial phylogenetic structure

Leandro R. Jones [a,b,*] & Julieta M. Manrique [a,b]

[a] Laboratorio de Virología y Genética Molecular, Facultad de Ciencias Naturales y Ciencias de la Salud, Universidad Nacional de la Patagonia San Juan Bosco, 9 de Julio y Belgrano s/n, (9100), Trelew, Chubut, Argentina.

[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

[*] Corresponding author. E-mail address: LRJ lrj000@gmail.com.

Keywords: SARS-CoV-2, COVID19, Phylogeny, Phylogenetic Structure

Relevance. A cluster of pneumonia cases of unknown etiology was reported in December 2019 in Wuhan, Hubei province, China. Since then, hundreds of thousands of people have died all around the planet. Quickly after the pandemic onset, metagenomic studies showed the causative agent, now named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), belonged to the Betacoronavirus genus, responsible for spillover events in 2002 and 2012 (severe acute respiratory syndrome and Middle East respiratory syndrome, respectively). The zoonotic origins of these viruses (possibly bats, camelids, pangolins and/or palm civets) have received much attention. However, other evolutionary aspects, such as spatial variation, have received comparatively little attention. This study shows that SARS-CoV-2 variants, which we call virotypes, are heterogeneously distributed on Earth and demonstrates that the virus phylogeny is geographically structured. We explain how this may be due to founder effects combined with high mutation rates.

Abstract. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an emergent RNA virus that spread around the planet in about 4 months. The consequences of this rapid spread on the virus evolution are under investigation. In this work, we analyzed *ca*. 9,000 SARS-CoV-2 genomes from Africa, America, Asia, Europe, and Oceania. We show that the virus is a complex of slightly different genetic variants that are unevenly distributed on Earth. Furthermore, we demonstrate that SARS-CoV-2 phylogeny is spatially structured. We hypothesize this could be the result of founder effects occurring as a consequence of, and local evolution occurring after, long-distance dispersal. In light of our results, we discuss how dispersal may constitute an opportunity for the virus to fix otherwise rare, and/or develop new, mutations. Based on previous studies, the possibility that this could significantly affect the virus phenotype is not remote.

Introduction

A cluster of acute atypical pneumonia syndrome cases of unknown etiology was reported late December 2019 in China. Rapidly, metagenomic RNA sequencing of bronchoalveolar lavage fluid showed that the etiological agent was a new Betacoronavirus, now named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1; 2]. Human coronaviruses are well known to produce mild to moderate upper-respiratory tract illnesses. However, severe acute respiratory syndrome coronavirus (SARS-CoV) responsible for a 2002-2003 epidemic, Middle East respiratory syndrome coronavirus (MERS-CoV) responsible for a 2012 outbreak and the novel SARS-CoV-2, cause severe syndromes with high fatality rates (10%, 36% and 6,8 %, respectively) [3; 4]. SARS-Co*V* and SARS-CoV-2, both belonging to the Sarbecovirus sub-genus, are ~79% identical at the genetic level and share important features as the same cell receptor and immunopathogenic mechanism.

SARS-CoV-2 has dispersed to about 200 countries, causing hundreds of thousands of infections. The consequences of this rapid spread on the virus evolution are still unclear. Recent studies

suggested that SARS-CoV-2 genotypes are heterogeneously distributed [5; 6]. However, quantitative assessment is still lacking. When relatedness between spatially coexisting lineages is greater than expected, their distribution is said to be phylogenetically structured [7]. One of the consequences of structuring is spatially close sequences being more similar to each other than expected by chance (Figure S1). The phenomenon can be assessed by comparing measured phylogenetic distances with those expected under no structuring, which can be accomplished by Monte Carlo simulations. Here, we describe an analysis of SARS-CoV-2 genomes collected from the beginning of the pandemic to April 25, 2020. We observed an uneven distribution of genetic variants, and a statistically significant spatial phylogenetic structure. As discussed below, this suggests that long-distance dispersal can favor the establishment of otherwise rare, and/or the emergence of new, mutations.

Materials and Methods

Virotypes delimitation and Datasets. Structure analyses requires to determine the abundance patterns of a set of operational taxonomic units (OTUs) [7]. To define viral OTUs, which we call 'virotypes', we initially studied a small dataset of 831 genomes. The smallness of this sample allowed us to identify unique sequences from all the possible pair-wise Needleman-Wunsch alignments in the dataset. By the other side, we compiled a much larger dataset (n=8613 genomes), aiming to gather a wider genomes sample to assess the virus spatial phylogenetic structure. The procedure used to delimit the small dataset OTUs has a time complexity of $O(n^2L)$ - $O(n^3L)$, where n and L are the sequences' number and length, respectively. Thus, the approach could not be used with the large dataset, which was processed by a fast *k*-mer distance method. The small dataset studied here consisted in all the genomes described from the beginning of the pandemic to March 14, 2020 (Table S1). These sequences were first inspected for the presence of ambiguous and undetermined positions by the *base.freq()* function from the *Ape* package [8]. The selected data

(genomes with no non-DNA characters) were aligned with *Mafft* [9] (details below) and the obtained alignment was visually inspected with *Jalview* [10]. This revealed the rare presence of sequences with internal and terminal gaps, which were dismissed (Table S1). After that, we generated Needleman-Wunsch pairwise alignments and, for each one, counted the number of mismatches and indels. End gaps were excluded from the counts since these correspond to sequencing differences. Pairwise alignments and hamming distances were obtained with the *nwhamming()* function of the *Dada2* package [11]. Shortly, we made an R script that takes a sequence (the *query*) from a dataset, aligns it against the rest of sequences and calculates the corresponding hamming distances:

```
pDists <- foreach(i=1:length(rest_of_sequences),
                  .packages = c("ape", "dada2")) %dopar%
                 {
                 nwhamming(query_sequence, rest_of_sequences[[i]])
                 }
```

Based on the obtained distances, the script determines which sequences are identical to the query sequence, put these together in an identity cluster, and starts a new iteration. This was repeated until all the sequences in the dataset were processed. We ensured that the sequences representative of each virotype used in posterior analyses were the longest of each identity cluster. The large dataset studied here consisted in all the sequences in GISAID's EpiCoV™ Database sampled between December 2019 and April 25, 2020 presenting less than 1 % non-DNA characters (*high coverage only* option) [12; 13]. The corresponding sequences and metadata are described in Table S2. Pairwise *K*-mer (k=21) distances (n=37,078,966) were obtained with *Mash* software [14], input in R and processed similarly to what was done with the small dataset.

Virotype abundances. If a virus is evenly distributed, the probability of detecting a given virotype i, in some area A_j, $P(V_i|A_j)$, should be the same for all j: $P(V_i|A_1)=P(V_i|A_2)=...=P(V_i|A_N)=P(V_i)$, where N is the number of sampled areas. We modeled the probability of observing a given virotype,

$P(V_i)$, by its frequency in the data. Then, the expected number of virotype *i* sequences in region A_j can be obtained as $P(i) \times M_j$, where M_j is the number of sequences from A_j. Standardized deviations from expectation can be obtained by dividing the squared deviations $(O_{ij} - E_{ij})^2$ by E_{ij}, where O_{ij} and E_{ij} are the observed and expected numbers of virotype *i* sequences in region *j*, respectively. This statistic is not affected by the order of magnitude of the data, which allows to straightforwardly compare very abundant and rare virotypes, and shallowly and much sampled regions.

Phylogenetic analysis. Sequence alignments were obtained with *Mafft*'s *FFT-NS-2* algorithm. Under this strategy, the program first generates a rough alignment (namely *FFT-NS-1* alignment) from a guide tree created from a distance matrix made by computing the number of hexamers shared between all sequence pairs. Then the *FFT-NS-1* alignment is used to generate a new tree, which is used to guide a second progressive alignment. Once aligned, the small dataset presented 512 variable sites of which 109 were parsimony informative. The large dataset alignment presented 8,750 polymorphic sites of which 1,873 were parsimony informative. The small dataset was phylogenetically analyzed with *PhyML* version 3.3 [15] under a generalized time-reversible model with a proportion of invariable sites and 4 rate categories. The nucleotide frequencies, proportion of invariable sites and value of the gamma shape parameter were maximum-likelihood estimated. Tree searches consisted in 100 parsimony starting trees that were refined by *NNI* and *SPR* branch swaps (option *-s BEST*). Branch supports were estimated with *PhyML*'s bootstrap routine. The large dataset analyses were performed with the double-precision version of *FastTree* [16] under the fast Jukes-Cantor + CAT rate heterogeneity algorithm. We started the searches by making a BIONJ tree, which was refined by the default tree-optimization strategy. This consist in performing up to *4Xlog2(N)* rounds of minimum-evolution *NNI* swaps, where *N* is the number of sequences in the data, followed by 2 minimum-evolution *SPR* rounds, and up to *2Xlog(N)* maximum-likelihood *NNI* rounds. Branch supports were estimated by the Shimodaira-Hasegawa test. Trees were inspected and plotted by *Dendroscope [17]*, *Ape* [8] and *ggtree* [18].

Spatial phylogenetic structure. Structure analyses were performed by the *picante* package [19]. Null distributions were inferred by shuffling virotypes across tree tips 10,000 times. Using the obtained permutations and the actual data, the software calculates the following weighted metric for each region:

$$SES_{MPD} = (MPD_{OBS} - \text{mean}(MPD_{H0}))/\text{sd}(MPD_{H0}),$$

where MPD_{OBS} is the mean phylogenetic distance (*MPD*) between all sequences from the region, $mean(MPD_{H0})$ is the average of the mean phylogenetic distance (*MPD*) between the sequences from the region in the randomized data and $sd(MPD_{H0})$ is the corresponding standard deviation. Negative and low SES_{MPD} values support structured distributions. Significance levels are calculated as the MPD_{OBS} rank divided by the number of permutations minus one.

Results

OTU analyses. Out of 62,481 pairwise comparisons performed from the small dataset, 61,759 (98.84%) revealed 1 or more mutations and 36,771 (58.8%) showed between 5 (Q1) and 11 (Q3) changes (Fig. 1). The analysis of the large dataset also revealed a very large diversity, as depicted in Figure 1. The number of detected OTUs relative to the total number of sequences analyzed was very similar for both datasets and methods. For the small dataset we observed one virotype every 1.55 sequences, while for the large one the ratio was of 1.61. The small dataset presented 228 unique sequences or virotypes. The most abundant virotype was represented by 30 sequences. Twenty-five were represented by 2 sequences each, 9 by 3 sequences, 3 by 4, 4 by 5 and the remaining three by 7, 9 and 14 sequences. The rest of virotypes were singletons. These results are summarized in Table S3. The large dataset presented 5,305 operational taxonomic units or virotypes. The most abundant virotype was represented by 223 sequences. Five virotypes presented

more than 100 sequences each, 12 presented at least 50 sequences and 695 presented between 2 and 50 sequences. These results are detailed in Table S4.

Virotypes distribution. In the small dataset, the 24 most abundant virotypes in Europe (n=78 genomes; virotypes 2, 6, 7, 8, 10, 11, 16, 17, 18, 19, 20, 24, 29, 30, 34, 37, 38, 39, 40, 41, 42, 43, 44 and 45; Table S3) were absent from other continents and virotype 1 was very prevalent in Asia (22 genomes) in comparison to North America (7 genomes). This can be easily appreciated from the corresponding deviations from expectation (Table S3). Virotypes 2, 6, 7, 8, 10, 11, 16, 17, 18, 19, 20, 24, 29, 30, 34, 37, 38, 39, 40, 41, 42, 43, 44 and 45 had a mean departure from expectation of 2.4 in Europe, whereas, for the same region, the rest of virotypes had a mean deviation of 0.68. Likewise, virotype 1 deviations from expectation in Asia and Europe were 8.37 and 11.08, respectively, whereas in North America, Oceania, and South America these were of only 2.26, 0.93 and 0.085, respectively. This reflects a virotype's over-abundance in Asia and rarity in Europe, and frequencies in the rest of continents relatively well predicted by the global virotype's abundance (Materials and Methods). The counts and distributions of the 5,305 virotypes identified from the large dataset are detailed in Table S4. As observed for the small dataset, these virotypes distribution was very uneven. Here we highlight the most relevant cases. Of the virotypes represented by at least 10 sequences, only virotypes 6, 10, 36, 41, 56 and 57 were distributed more or less homogeneously. Virotype 1 abundance was greater than expected in North America and smaller than expected in Europe and Asia. Virotypes 4, 8, 11, 14, 15, 19, 20, 25, 29, 31, 34, 42, 43, 45 and 58 were over-represented in North America while virotypes 2, 3, 7, 9, 12, 16, 17, 18, 21, 22, 23, 24, 27, 30 and 32 were excessively abundant in Europe. Virotypes 5, 13, 26, 38 and 40, and 28, 33 and 55 were over-abundant in Asia and Oceania, respectively, and virotype 59 was too abundant in both Asia and Oceania. Africa shared 4 virotypes with Europe (virotypes 186, 237, 300 and 561), 2 with Europe and South America (virotypes 24 and 61), 1 with Europe and Oceania (virotype 18) and 1 with North America (virotype 308). The remaining 44 virotypes present in Africa were endemic (that is,

were present only there). In contrast, many of the virotypes in South America displayed wide distributions. Four of them (virotypes 3, 6, 12 and 35) were also detected in Asia, Europe, North America and Oceania, three (virotypes 11, 130 and 133) were shared with Europe and North America and two (virotypes 24 and 61) with Africa and Europe. Virotype 19 was shared with Asia, North America, and Oceania, virotype 22 with Asia, Europe, and Oceania and virotype 46 with Europe and Oceania. Finally, virotypes 212 and 658 were also present in North America and Europe, respectively.

Phylogenetic structure analyses. The small dataset phylogeny is shown in Figure 2. A heat map describing the virotypes distribution is shown alongside the phylogeny. Tree terminals are aligned with the heatmap cells, so that the Figure depicts the tree lineages' abundances at each continent (heatmap columns). Clades *A*, *B*, H, *I* and *J* were distributed in Europe. Likewise, the majority of sequences located at clades *A*-*D*, and the sequences grouped in clades *F* and *G* were common in Europe. The exceptions were virotypes 226, 46 and 112, which correspond to samples from North America (virotypes 226 and 112) or from North America and Europe (virotype 46). Most sequences in the branch containing clades *M* and *N* are from North America. Clades *E* and *K* sequences are all from Asia. As expected from the visual inspection of the virotypes phylogeny and distribution, structuring was highly significant for North America and Europe (Table 1). The phylogeny and distributions of the large dataset virotypes are detailed in Figure S2. A thorough inspection of these results revealed six major sections of the tree in which the virotypes were clumped according to their origins, as shown in a simplified manner in Figure 3. For practical reasons, we refer to these six tree sections as *A*, *B*, *C*, *D*, *E*, and *F*. The North American virotypes were more abundant in sections *A* and *D* while the European ones were prevalent in sections *B*, *C*, *E* and *F*, and moderately represented in section *A*. The Asian virotypes clustered preferentially in sections *D*, *E* and *F* but in proximal positions relative to the positions occupied by the North American (region *D*) and European (regions *E* and *F*) virotypes. Many of the virotypes in Africa were clustered in a relatively

well supported branch (Fig. S2). Conversely, the virotypes present in South America and Oceania were scattered all across the tree (Figs. 3 d and S2). Limitation in space was significant for Africa, Europe, and North America (Table 2).

Discussion

SARS-CoV-2 virotypes. Using an exact algorithm, we identified 228 unique sequences (here called 'virotypes') among 354 SARS-CoV-2 genomes sampled from mid-December 2019 to March 14, 2020. In a similar way, analyses of 8,612 sequences sampled between December 2019 and April 25, 2020 using a fast, approximate method, revealed 5,305 genetic variants or virotypes. Less than half the here studied sequences were identical to at least one of the rest of sequences and, out of 62,481 genome pairs analyzed with the exact algorithm, only 722 showed no mutations. A raw extrapolation of these results suggests that, only among the ~20,000 SARS-CoV-2 genomes sequenced so far, there could be some 12,000 virotypes. These data are overwhelming. In addition to the threat this may pose to human health (discussed below), we anticipate significant methodological challenges, especially if the number of SARS-CoV-2 sequences continues to increase at present rates. As explained in the Materials and Methods, the approach used here to analyze the small dataset is $O(N^2L)$ to $O(N^3L)$ complex. This means that, depending on the algorithm and software used, it would be very expensive to replicate the analyses for more than several hundred or some few thousand sequences. Likely impossible for more than 10,000 sequences [20]. An alternative may be to base sequence clustering on a multiple sequence alignment (MSA), thus avoiding the costly pairwise alignment step. However, MSA is NP-complete, which is to say that generating optimal MSAs is computationally intractable for more than 3-4 sequences [21; 22]. Also, MSAs usually present misaligned ends attributable to sequencing differences (data not shown). Misaligned ends can be difficult to handle, especially in large datasets, since there is no standardized procedure to deal with the problem. The issue (which is not

exclusive of SARS-CoV-2) is frequently addressed by trimming off the problematic alignment regions. However, this involves data dismissal, which requires reproducible procedures. Standard procedures are neither available for dealing with incomplete sequences, which are sometimes padded with *N*s. Although in some cases it can be handy to simply discard such sequences, some incomplete sequences may contain useful information that is absent from other sequences. The fast distance method used here to analyze the large dataset produced results congruent with those obtained with the exact algorithm used to analyze the small dataset. Fast approaches similar to the one used here have already proven useful in previous virologic studies and high-throughput applications [23-25]. Further research is needed to determine the degree to which fast distance comparisons reflect actual similarities, especially if detailed mutational analyses are to be performed. However, considering that SARS-CoV-2 genomic data are increasing rampantly, the here-implemented approach is potentially applicable to the development of real-time diversity-analysis tools.

Spatial Phylogenetic Structure. Biogeographical patterns have been previously observed in other Betacoronavirus [26] and very different viruses as phages and retroviruses [27; 28]. That SARS-CoV-2 has developed a biogeography despite its high propagation rate may seem contradictory. However, spatial diversification depends not only on dispersion constraints but also on evolutionary rates. In particular, spatially structured phylogenies can be the consequence of speciation rates being very high relative to dispersion rates [7]. As far as we know, travels between remote places constitute the only dispersal mechanism of the virus. This implies that founder viruses usually carry very small fractions of the total genetic variation of the source populations. Therefore, each time the virus spreads, substantial losses of diversity can occur, possibly combined with rare mutations settlement, due to founder effects. In addition, it stands to reason that, after dispersal, mutations accumulate quickly as the newly established population enlarges, due to the high error rates of viral polymerases [29; 30]. Some of these mutations can lead to novel virotypes, as strongly suggested by

the results in Figure 3. Based on these considerations, it is reasonable to hypothesize that long-distance dispersal constitutes an opportunity for the virus to fix otherwise rare, and/or develop new, mutations.

It has been shown that slight mutations can produce significant phenotypic effects in MERS-CoV and other coronaviruses [26; 31; 32]. On the other hand, the 2002-2003 SARS-CoV epidemic was subdivided into three genetically different phases, suggesting that sarbecoviruses can mutate recurrently along relatively short periods of time [33]. Furthermore, there is evidence that SARS-CoV-2 could be tolerant to non-synonymous mutations at the *nsp1* and accessory ORFs, which in other coronaviruses are related to immune modulation and viral evasion [3; 34]. Thus, the here presented results must be taken as a call for attention. The virus evolution should continue to be monitored and relationships should be sought between viral diversity and pathogenesis.

Acknowledgements

We acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiCoV™ Database used in this study. This work was partially supported by PIP 11220130100255CO (CONICET), PICT 2016-2795 (ANPCyT) and Asociación Civil Argentina Genetics. LRJ and JMM conceived the study; LRJ compiled the data; LRJ and JMM analyzed the data and interpreted the results; LRJ and JMM wrote the manuscript.

References

[1] **Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; Yuan, M.-L.; Zhang, Y.-L.; Dai, F.-H.; Liu, Y.; Wang, Q.-M.; Zheng, J.-J.; Xu, L.; Holmes, E. C. and Zhang, Y.-Z.** (**2020**). *A new coronavirus associated with human respiratory disease in China*, Nature 579 : 265-269.

[2] **Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M.-Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S.; Zhao, K.; Chen, Q.-J.; Deng, F.; Liu, L.-**

L.; Yan, B.; Zhan, F.-X.; Wang, Y.-Y.; Xiao, G.-F. and Shi, Z.-L. (**2020**). *A pneumonia outbreak associated with a new coronavirus of probable bat origin*, Nature 579 : 270-273.

[3] **Cui, J.; Li, F. and Shi, Z.-L.** (**2019**). *Origin and evolution of pathogenic coronaviruses*, Nature Reviews Microbiology 17 : 181-192.

[4] WHO (**2020**). *Emergencies*, https://www.who.int/topics/emergencies/en/. 2020.

[5] **Forster, P.; Forster, L.; Renfrew, C. and Forster, M.** (**2020**). *Phylogenetic network analysis of SARS-CoV-2 genomes*, Proceedings of the National Academy of Sciences 117 : 9241-9243.

[6] **Rambaut, A.; Holmes, E. C.; Hill, V.; O'Toole, Á.; McCrone, J.; Ruis, C.; du Plessis, L. and Pybus, O. G.** (**2020**). *A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology*, bioRxiv doi:10.1101/2020.04.17.046086.

[7] **Webb, C. O.; Ackerly, D. D.; McPeek, M. A. and Donoghue, M. J.** (**2002**). *Phylogenies and Community Ecology*, Annual Review of Ecology and Systematics 33 : 475-505.

[8] **Paradis, E. and Schliep, K.** (**2018**). *Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R*, Bioinformatics 35 : 526-528.

[9] **Katoh, K. and Toh, H.** (**2008**). *Recent developments in the MAFFT multiple sequence alignment program*, Briefings in Bioinformatics 9 : 286-298.

[10] **Waterhouse, A. M.; Procter, J. B.; Martin, D. M. A.; Clamp, M. and Barton, G. J.** (**2009**). *Jalview Version 2--a multiple sequence alignment editor and analysis workbench*, Bioinformatics 25 : 1189-1191.

[11] **Callahan, B. J.; McMurdie, P. J.; Rosen, M. J.; Han, A. W.; Johnson, A. J. A. and Holmes, S. P.** (**2016**). *DADA2: High-resolution sample inference from Illumina amplicon data*, Nature Methods 13 : 581-583.

[12] **Elbe, S. and Buckland-Merrett, G.** (**2017**). *Data, disease and diplomacy: GISAID's innovative contribution to global health*, Global Challenges 1 : 33-46.

[13] **Shu, Y. and McCauley, J.** (**2017**). *GISAID: Global initiative on sharing all influenza data – from vision to reality*, Eurosurveillance 22.

[14] **Ondov, B. D.; Treangen, T. J.; Melsted, P.; Mallonee, A. B.; Bergman, N. H.; Koren, S. and Phillippy, A. M.** (**2016**). *Mash: fast genome and metagenome distance estimation using MinHash*, Genome Biology 17.

[15] **Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W. and Gascuel, O.** (**2010**). *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0*, Systematic Biology 59 : 307-321.

[16] **Price, M. N.; Dehal, P. S. and Arkin, A. P.** (**2010**). *FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments*, PLoS ONE 5 : e9490.

[17] **Huson, D. H. and Scornavacca, C.** (**2012**). *Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks*, Systematic Biology 61 : 1061-1067.

[18] **Yu, G.** (**2020**). *Using ggtree to Visualize Data on Tree-Like Structures*, Current Protocols in Bioinformatics 69.

[19] **Kembel, S. W.; Cowan, P. D.; Helmus, M. R.; Cornwell, W. K.; Morlon, H.; Ackerly, D. D.; Blomberg, S. P. and Webb, C. O.** (**2010**). *Picante: R tools for integrating phylogenies and ecology*, Bioinformatics 26 : 1463-1464.

[20] **Blackshields, G.; Sievers, F.; Shi, W.; Wilm, A. and Higgins, D. G.** (**2010**). *Sequence embedding for fast construction of guide trees for multiple sequence alignment*, Algorithms for Molecular Biology 5 : 21.

[21] **Wang, L. and Jiang, T.** (**1994**). *On the Complexity of Multiple Sequence Alignment*, Journal of Computational Biology 1 : 337-348.

[22] **Nakamura, T.; Yamada, K. D.; Tomii, K. and Katoh, K.** (**2018**). *Parallelization of MAFFT for large-scale multiple sequence alignments*, Bioinformatics 34 : 2490-2492.

[23] **Liu, Z. H. and Sun, X.** (**2008**). *Coronavirus phylogeny based on Base-Base Correlation*, International Journal of Bioinformatics Research and Applications 4 : 211.

[24] **Huang, H.-H.** (**2016**). *An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses*, Journal of Theoretical Biology 398 : 136-144.

[25] **Goussarov, G.; Cleenwerck, I.; Mysara, M.; Leys, N.; Monsieurs, P.; Tahon, G.; Carlier, A.; Vandamme, P. and Houdt, R. V.** (**2020**). *PaSiT: a novel approach based on short-oligonucleotide frequencies for efficient bacterial identification and typing*, Bioinformatics 36 : 2337-2344.

[26] **Chu, D. K. W.; Hui, K. P. Y.; Perera, R. A. P. M.; Miguel, E.; Niemeyer, D.; Zhao, J.; Channappanavar, R.; Dudas, G.; Oladipo, J. O.; Traoré, A.; Fassi-Fihri, O.; Ali, A.; Demissié, G. F.; Muth, D.; Chan, M. C. W.; Nicholls, J. M.; Meyerholz, D. K.; Kuranga, S. A.; Mamo, G.; Zhou, Z.; So, R. T. Y.; Hemida, M. G.; Webby, R. J.; Roger, F.; Rambaut, A.; Poon, L. L. M.; Perlman, S.; Drosten, C.; Chevalier, V. and Peiris, M.** (**2018**). *MERS coronaviruses from camels in Africa exhibit region-dependent genetic diversity*, Proceedings of the National Academy of Sciences 115 : 3144-3149.

[27] **Díaz-Muñoz, S. L.; Tenaillon, O.; Goldhill, D.; Brao, K.; Turner, P. E. and Chao, L.** (**2013**). *Electrophoretic mobility confirms reassortment bias among geographic isolates of segmented RNA phages*, BMC Evolutionary Biology 13 : 206.

[28] **Rodriguez, S. M.; Golemba, M. D.; Campos, R. H.; Trono, K. and Jones, L. R.** (**2009**). *Bovine leukemia virus can be classified into seven genotypes: evidence for the existence of two novel clades*, Journal of General Virology 90 : 2788-2797.

[29] **Gago, S.; Elena, S. F.; Flores, R. and Sanjuan, R.** (**2009**). *Extremely High Mutation Rate of a Hammerhead Viroid*, Science 323 : 1308-1308.

[30] **Moya, A.; Holmes, E. C. and González-Candelas, F.** (**2004**). *The population genetics and evolutionary epidemiology of RNA viruses*, Nature Reviews Microbiology 2 : 279-288.

[31] **Rasschaert, D.; Duarte, M. and Laude, H.** (**1990**). *Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions*, Journal of General Virology 71 : 2599-2607.

[32] **Zhang, X.; Hasoksuz, M.; Spiro, D.; Halpin, R.; Wang, S.; Vlasova, A.; Janies, D.; Jones, L. R.; Ghedin, E. and Saif, L. J.** (**2007**). *Quasispecies of bovine enteric and respiratory coronaviruses based on complete genome sequences and genetic changes after tissue culture adaptation*, Virology 363 : 1-10.

[33] **The Chinese SARS Molecular Epidemiology Consortium** (**2004**). *Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China*, Science 303 : 1666-1669.

[34] Cagliani, R.; Forni, D.; Clerici, M. and Sironi, M. (**2020**). *Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2*, Journal of Virology in press.

Figure Legends

Figure 1. Pairwise genetic distances obtained between the genomes in the small (n=62,481 pair-wise hamming distances) and large (n=37,078,966 pair-wise *k*-mer distances) datasets studied here.

Figure 2. Phylogeny and distribution of 228 SARS-CoV-2 virotypes corresponding to 354 genomes from the small dataset studied here. Dot sizes indicate the number of genomes represented by each virotype, as indicated in the legend to the right of the Figure. Numbers in parenthesis indicate branch supports of selected tree branches, provisionally named clades *A* to *O* for practical reasons. Tree terminals are aligned with the heatmap cells, which depicts the virotypes' distributions and abundances, as indicated by the scale to the right. *ASI Asia*; *EUR Europ*e; *NAM North America*; *OCE Oceania*; *SAM South Americ*a.

Figure 3. Phylogeny and distribution of 5,305 SARS-CoV-2 virotypes representative of 8,612 genomes from around the world. The virotypes present in in North America (**a**), Europe (**b**), Asia (**c**) and Oceania (**d**) are indicated by colored dots. Six sections of the tree are indicated (*A*, *B*, *C*, *D*, *E*, and *F*) to facilitate results interpretation (please see main text). Dots diameters are proportional to the number of genomes accrued in each virotype, as indicated in each panel separately.

Table 1. Small dataset structure analysis.

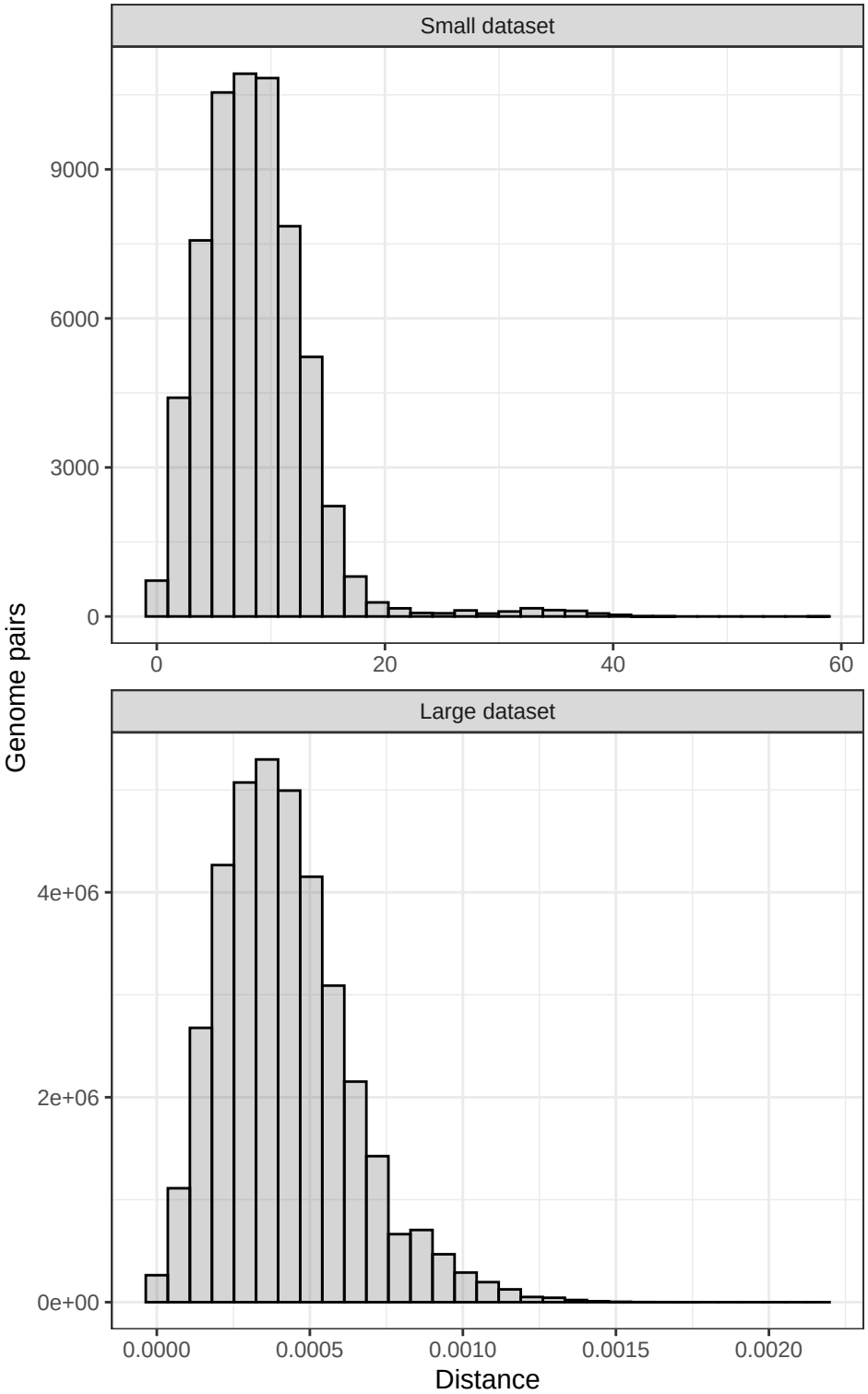
	ntaxa[1]	obs[2]	rand.mean[3]	rand.sd[4]	obs.rank[5]	SES_{MPD}[6]	*p*-value[7]
Asia	89	0.00048	0.00044	0.00008	8198	0.55351	0.81972
Europe	99	0.00031	0.00045	0.00006	1	-2.28880	0.00010
North America	37	0.00027	0.00043	0.00009	19	-1.80454	0.00190
Oceania	10	0.00043	0.00041	0.00014	6616	0.15601	0.66153

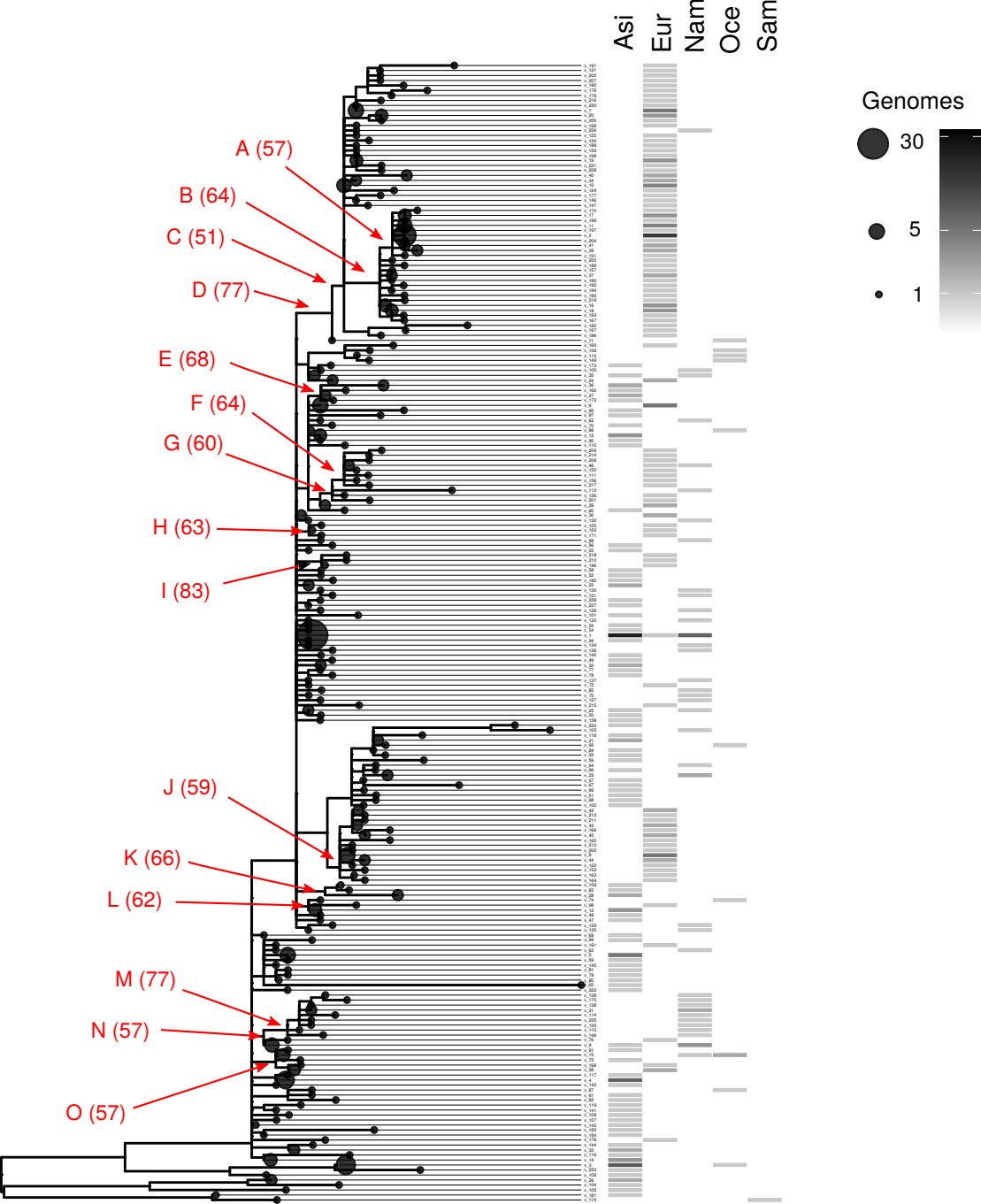
[1] Number of virotypes from each region (rows).
[2] Observed Mean Phylogenetic Distances (MPD).
[3] Average MPD in Monte Carlo (M-C) randomizations (dispersed model).
[4] Standard deviation of M-C MPDs.
[5] Observed MPDs ranks.
[6] Standardized effect of geographical structure.
[7] H_0: evenly distributed virotypes.

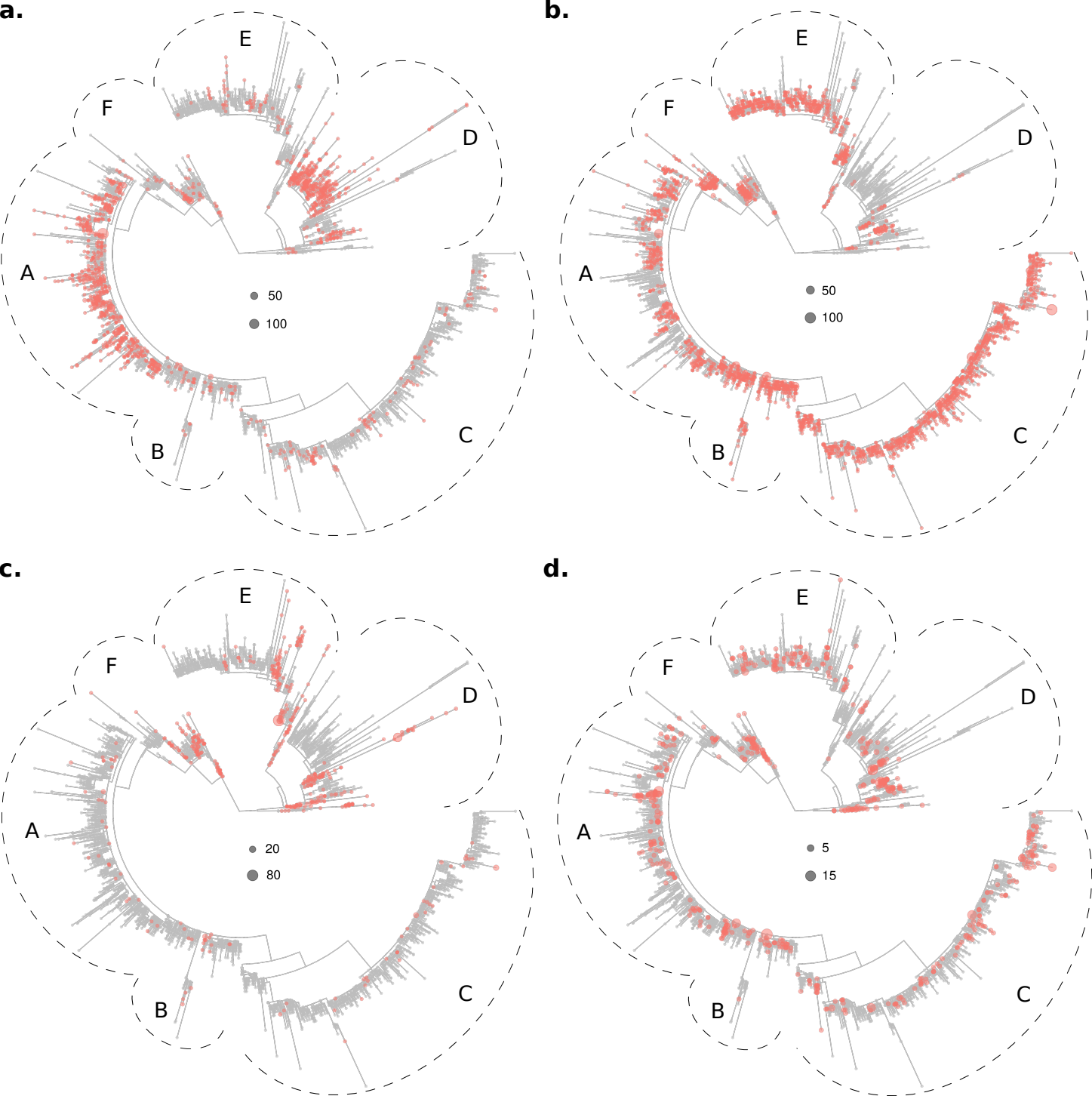
Table 2. Large dataset structure analysis.

	ntaxa[1]	obs[2]	rand.mean[3]	rand.sd[4]	obs.rank[5]	SES_{MPD}[6]	p-value[7]
Africa	53	0.00476	0.00744	0.00053	1	-5.07282	0.0001
Asia	494	0.0074	0.00747	0.00046	4507	-0.14996	0.45065
Europe	2747	0.00671	0.00757	0.00019	1	-4.57045	0.0001
North America	1479	0.00623	0.00753	0.00032	1	-3.99469	0.0001
Oceania	657	0.00769	0.00757	0.00021	6973	0.53542	0.69723
South America	39	0.00683	0.00731	0.00072	2336	-0.67678	0.23358

[1] Number of virotypes from each region (rows).
[2] Observed Mean Phylogenetic Distances (MPD).
[3] Average MPD in Monte Carlo (M-C) randomizations (dispersed model).
[4] Standard deviation of M-C MPDs.
[5] Observed MPDs ranks.
[6] Standardized effect of geographical structure.
[7] H_0: evenly distributed virotypes.


Small dataset
9000
6000
3000
0
0
20
40
60
Genome pairs
Large dataset
4e+06
2e+06
0e+00
0.0000
0.0005
0.0010
0.0015
0.0020
Distance


Asi
Eur
Nam
Oce
Sam
Genomes
30
5
1
A (57)
B (64)
C (51)
D (77)
E (68)
F (64)
G (60)
H (63)
I (83)
J (59)
K (66)
L (62)
M (77)
N (57)
O (57)


a.
E
F
D
A
50
100
B
C
b.
E
F
D
A
50
100
B
C
c.
E
F
D
A
20
80
B
C
d.
E
F
D
A
5
15
B
C