Supplementary Data

Supplementary Materials and Methods

We downloaded 90 876 SARS-CoV-2 genomes from GISAID on 10 September 2020 (herein referred as 90K database). The genomes were aligned and cleaned as in Gómez-Carballa et al. (2020b) using the sequence deposited in GenBank under accession number MN908947.3 (GISAID ID #402125) as a reference. Only high-quality (HQ) genomes (>29 kbp in length) having chronological and sampling information in the metadata were retained for follow-up analysis. Indels and ambiguities were eliminated from the analysis, but MNP variations (such as GGG28881AAC, which is diagnostic of A2a4) were retained. The final database consisted of 63 984 SARS-CoV-2 HQ genomes. In the **Supplementary Data**, we discuss several issues related to the data treatment carried out by Zhao et al. (2020). We followed the phylogenetic nomenclature of clades proposed by Gómez-Carballa et al. (2020a, 2020b) and we established the parallelism of nodes with the different signatures defined by Zhao et al. (2020). Thus, the tree skeleton in **Figure 1** was obtained from the large SARS-CoV-2 evolutionary tree in Gómez-Carballa et al. (2020a).

We computed haplotype entropy in an analogous way, as in Zhao et al. (2020), but accounting for haplotypes instead of single positions:

$$HE(h_i) = -\sum_{k \in L} p_k(h_i) \times log_2\big(p_k(h_i)\big)$$

where L represents the unique characters in the SARS-CoV-2 genomes, $p_k(i)$ is the probability of observing combination h of k characters conformed by a growing number of i positions (more than one position would conform a haplotype of i positions). The addition of new i positions would tend to increase $HE(h_i)$ (e.g., excluding phylogenetically redundant variants). We progressively retained the positions yielding the highest $HE(h_{i-1})$ and a number of i positions that empirically allow visualization of a saturation (plateau) in the $HE(h_i)$ function. It is important to note that it is out of the scope of the present study to analyze alternative statistics other than entropy (which measures genome variation), which could potentially yield comparable or better efficiency. Haplogroup frequencies of

selected lineages were represented by geographic maps using SAGA v.7.6.2 (http://www.saga-gis.org/) (Conrad et al., 2015) and the ordinary Kriging method. We only used sampling points with more than 20 viral genomes. Most computations were carried out using R (v.3.5.0.) (R Core Team, 2019).

Comments on sequence alignment and genome database cleaning

Zhao et al. (2020) downloaded >47K genomes from the GISAID database (http://www.gisaid.org) on 17 June 2020. The data were initially filtered by genome length (at least 25K base pairs (bp)) and several genomes were discarded due to a lack of information in the metadata. The sequences were then automatically aligned such that the final alignment length extended to 79 716 nt due to the presence of gaps (approximately 2.6 times the length of the SARS-CoV-2 genome). Entropy was computed for all positions and included variants A, C, G, and T, but also gaps (-) and ambiguities (N); a masked entropy was also computed ignoring gaps and ambiguities. Accidentally, the authors found that "Sites with large number of N's and - should be filtered out because a large number of N's and -'s at a position is typically due to sequencing error or alignment artifact which provides less information about the real nucleotide distribution at this position" (see their legend to Figure S1). It is worth clarifying that filtering out this variation is mandatory because it is probably not real (not because it is less informative). We recently analyzed a similar dataset in size and content (41K genomes) from GISAID (Gómez-Carballa et al., 2020b) (downloaded on 12 June 2020) and our multisequence alignment resulted in a total length of 32K after manually correcting assembly problems generated by individual genomes that were most likely affected by artefactual insertions. For comparison, the alignment that can be directly obtained from GISAID is also 32K in length (downloaded on 17 October 2020, >137K genomes). This suggests that the alignment generated by Zhao et al. (2020) is sub-optimal.

In addition, the genomes in GISAID are labeled as high quality (HQ) and low quality (LQ). It is recommended to use HQ genomes only, especially if the aims of the study require site/variant-specific accuracy (e.g., if selecting ISM). In our recently published study carried out on 41K GISAID genomes (Gómez-Carballa

et al., 2020b), we only used HQ sequences (representing ~64% of the genomes in the database). We additionally removed indels because next-generation sequencing (NGS) read mappers have difficulties in accurately mapping short reads containing complex variation, e.g., indels (Hasan et al., 2019; Mose et al., 2014); moreover, evolutionary patterns of indels are more difficult to reconstruct and can generate conflicts in sequence alignments. Some complex variation may escape from this rule; for example, it seems reasonable to consider (e.g., in a phylogenetic/phylogeographic context) the characteristic MNP GGG28881AAC that identifies one of the most successful SARS-CoV-2 clades worldwide, namely A2a4 (also identified by diagnostic variants: C241T–C3037T–C14408T–A23403G; (Gómez-Carballa et al., 2020a)).

Working with a clean database is the first mandatory step before carrying out genomic analysis. In Zhao et al. (2020), proper filtering would help prevent the generation of entropy values from artefactual variation, and focus attention on the variability with the highest likelihood of being real. Although Zhao et al. (2020) tried to filter out artificial variation *a posteriori* (see legend of their Figure S1.), their 20 compact set still contained two positions with indels, and one of them, namely, position 11083, remained in their 11 ISM set.

Thoughts on future barcode proposals

The paper by Zhao et al. (2020) was published in *PLoS Computational Biology* on 17 September 2020. The authors of the present study attempted to contact their editors by email several times to discuss (in the form of a Formal Letter) the issues observed in this article, but we received no proactive response. We consider that discussions on barcodes are important to prevent future proposals from similar issues observed in these seminal articles. For instance, the article by Guan et al. (2020) emerged almost at the same time as Zhao et al. (2020), and it is highly probable that more articles will also be published.

REFERENCES

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

Gómez-Carballa A, Bello X, Pardo-Seco J, Martinón-Torres F, Salas A. 2020a. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Research*, **30**(10): 1434–1448.

Gómez-Carballa A, Bello X, Pardo-Seco J, Pérez Del Molino ML, Martinón-Torres F, Salas A. 2020b. Phylogeography of SARS-CoV-2 pandemic in Spain: a story of multiple introductions, micro-geographic stratification, founder effects, and super-spreaders. *Zoological Research*, **41**(6): 605–620.

Guan QT, Sadykov M, Mfarrej S, Hala S, Naeem R, Nugmanova R, et al. 2020. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. *International Journal of Infectious Diseases*, **100**: 216–223.

Hasan MS, Wu XW, Zhang LQ. 2019. Uncovering missed indels by leveraging unmapped reads. *Scientific Reports*, **9**(1): 11093.

Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. 2014. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, **30**(19): 2813–2815.

R Core Team. 2019. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Zhao ZQ, Sokhansanj BA, Malhotra C, Zheng K, Rosen GL. 2020. Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. *PLoS Computational Biology*, **16**(9): e1008269.