SARS-CoV-2 Genotype Network

Iker Atienza-Diez[1,2], Luís F Seoane[1,2], and Susanna Manrubia[1,2],
[1] Systems Biology Department, Centro Nacional de Biotecnología (CSIC), Madrid, Spain
[2] Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain

Genotype networks are powerful representations of great aid in the interpretation of evolutionary processes, especially for highly heterogeneous molecular populations [1]. These networks can be constructed at different scales, for instance by deep-sequencing evolving in-vitro populations [2] or through geographically extended data of a circulating pathogen [3]. In this contribution, we present the SARS-CoV-2 (SARS2) genotype network (GN) reconstructed from genomic data spanning from December 2019 to March 2023, with Wuhan-Hu-1 strain (GenBank: MN908947.3) as wild-type (WT) reference sequence [4].

SARS2 genome is about 30,000 base-pairs long. The reconstruction of the network using whole genomes is computationally unfeasible, so we have selected the Receptor Binding Domain (RBD) section of the viral Spike protein. The RBD contains 223 amino acids (S:319-541) involved in the recognition of the human ACE2 receptor, and thus in cell entry of the virus [5]. Its location in SARS2 genome is illustrated in Figure 1.

A haplotype is any sequence that differs from the WT. After curating the original dataset, we analyse $5,799,310$ complete genomes to extract the set of different haplotypes in the RBD section of interest; we identify $28,686$ unique haplotypes with an abundance ranging from 1 to $1,915,492$ sequences. Each identified haplotype is a node in our GN, and two nodes are connected through an edge if their sequences differ by a single mutation: SARS2 GN has $27,634$ nodes and $56,122$ edges. Fig. 1 shows the obtained GN with nodes colored according to the number of mutations accumulated with respect to the WT, as specified in the legend. Our analysis of the topological properties of this GN reveals that it is weakly disassortative and has an average degree $\langle k\rangle \simeq 4$.

Since genomes in the dataset are labelled according to the variant they belong to, an analysis using the subset of haplotypes in each variant is possible. Since our study is limited to the RBD, there is some degeneracy in this classification, with 916 (3.31%) multi-variant haplotypes, that is, sequences that can be classified in two or more variants. All variants of concern, except Omicron, are relatively close to the WT (< 5 mutations). Omicron-labelled haplotypes are more diverse in terms of mutations, suggesting that this variant has explored a larger region of genotype space. Our analysis supports as well that the fitness landscape around this variant is flatter, since its associated subnetwork has a significantly larger number of nodes with high degree, consistently leading to a less disassortative pattern than that of previous variants. Interestingly, the SARS2 GN contains a large number of cycles, pointing at a non-uniqueness of evolutionary pathways linking different haplotypes within and between variants.

We have also explored the temporal appearance of different haplotypes and found, first, a burst of haplotype diversity (12/2021-01/2022) associated to the emergence of Omicron and, second, a waxing and waning pattern in haplotype abundance caused by the sequential emergence of new successful variants. We observe that some early-explored, but not fixed, haplotypes re-emerge when Omicron arises, possibly due to other accompanying mutations out of the RBD region.

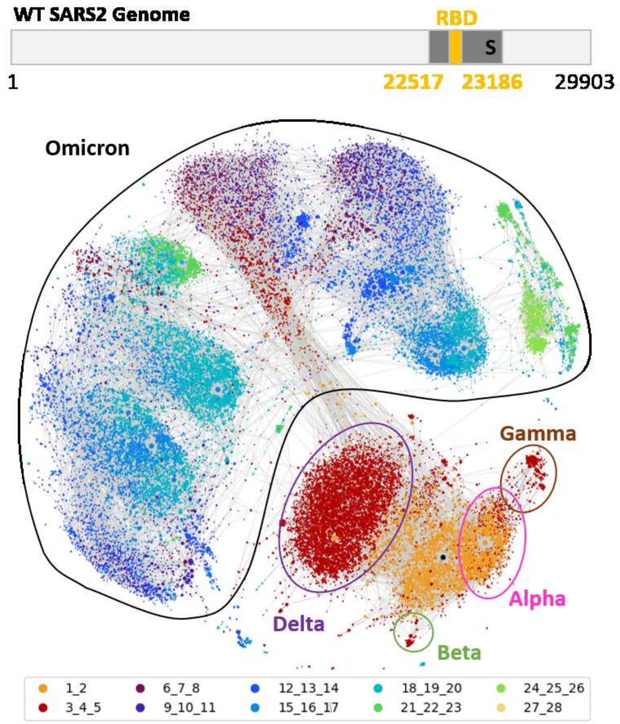

Fig. 1. SARS-CoV-2 genome and RBD Genotype Network. Above: schematic of SARS2 genome highlighting the position of the spike protein and its RBD motif. Below: SARS2 GN. Node size is proportional to (the logarithm of) haplotype abundance and node color indicates the number of mutations accumulated by each particular haplotype. Colors stand for the number of mutations, as indicated in the legend, ranging from 1-2 mutations (orange) to 27-28 (yellow) when compared to the WT RBD sequence (black node).

[1] J. Aguirre, P. Catalán, JA. Cuesta, and S.Manrubia. Open Biol. **81** (2018).

[2] A. Villanueva, H. Secaira-Morocho, LF Seoane, E. Lzaro and S. Manrubia. Biophysica **2(4)**, 381-399 (2022).

[3] A. Wagner. Proc. R. Soc. B. **281** (2014).

[4] `https://docs.nextstrain.org/en/latest/`

[5] Y. Huang, C. Yang, Xf. Xu, W. Xu and SW. Liu. Acta. Pharmacol. Sin. **41**, 11411149 (2020).

[6] RA. Neher. Virus Evolution, **8(2)** (2022).