GENOME AND GENETIC VARIATIONS OF CORONAVIRUSES

António A. N. de Alcochete

Department of Biology, Faculty of Sciences, Agostinho Neto University;

ABSTRACT

The knowledge of genetics, using innovative technologies and methodologies that allow the analysis and comparison of multiple genomes of viruses, vectors and hosts, as a function of time and space is very informative because it allows the understanding of the origin, evolution and geographical dispersion of the disease.

Key-words: SARS-CoV-2, CORONAVÍRUS, DNA

INTRODUCTION

The term genome, introduced in 1920 by Hans Winkler, was used to refer to all the hereditary information of an organism that is encoded in its DNA or RNA, or simply, to the complete DNA sequence of a set of chromosomes. Today, the concept of genome comprises the information needed to build, maintain and know the evolutionary history of an organism (Cristescu, 2019).

CORONAVIRUS GENOME

The genome of coronaviruses is a single-stranded, positive-sense RNA molecule whose size ranges from 27 to 32 kpb and contains at least six "Open Reading Frames" (ORFs). The first ORF (ORF1a/b), located at the 5' end, occupies about two-thirds of the genome and encodes polyprotein 1a,b (pp1a, pp1b). The remaining ORFs are located at the 3'end and encode at least four structural proteins: the capsid/envelope spiculated gliprotein (S), responsible for recognition of host cell

receptors; membrane proteins (M), responsible for the shape of the capsid/envelope; capsid/envelope proteins (E), responsible for virus assembly and release; nucleocapsid proteins (N) involved in genome packaging that play a role in pathogenicity as an interferon (IFN) inhibitor. There are also species-specific structural and accessory proteins, such as HE, 3a/b and 4a/b proteins (Alanagreh et al., 2020).

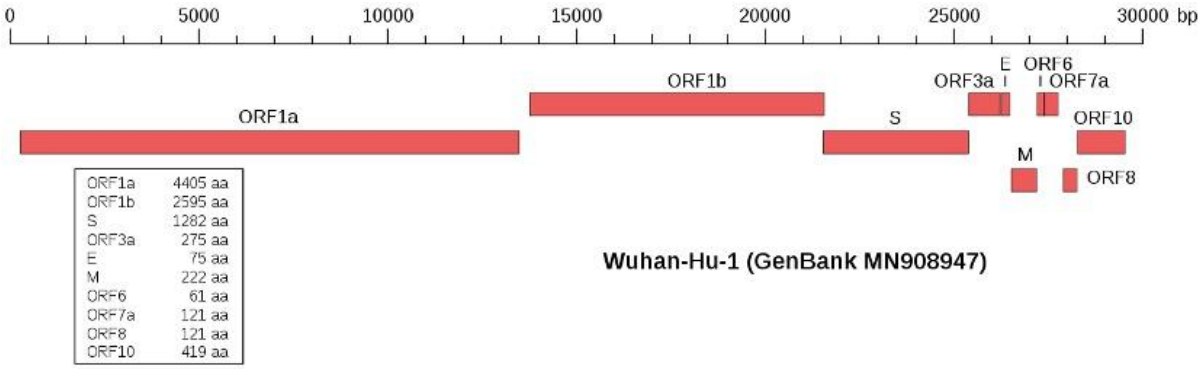

Fig 1 - Genome of SARS-Cov-2 (Source: Wu et al., 2020a).

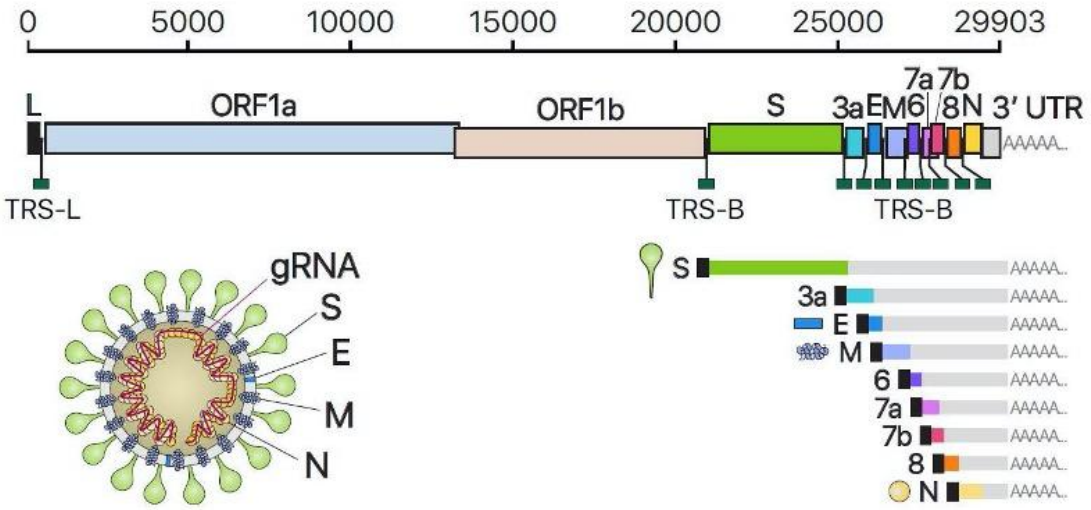

Fig 2 - SARS-Cov-2 genome showing the genomic and subgenomic RNAs ((S, E, M, N, 3a, 6, 7a, 7b, 8) (Source: Fernandez-Rua, 2020)

GENETIC VARIATIONS

Despite the small size of the coronavirus genome, the adaptive evolutionary process in different host environments and the wide geographic dispersion allows recording structural and functional genetic changes. Comparative genetic analysis studies have shown that coronavirus genomes retain between 50% and 95% similarity. The genome of SARS-CoV-2, which infects humans, appears to contain up to 15 genes very similar to SARS-CoV found in Manis javanica (Pangolin) and in bats, especially with Beta-CoV found in bats, at 96.2%, and with Bat-CoV-RatTG13, at 79.2% with SARS-CoV (Zhu et al., 2020).

Although the similarity of the genomes is remarkable, the S protein subunit of Pangolin virus showed greater similarity with SARS-CoV-2 than with SARS-CoV and Bat-RaTG13.

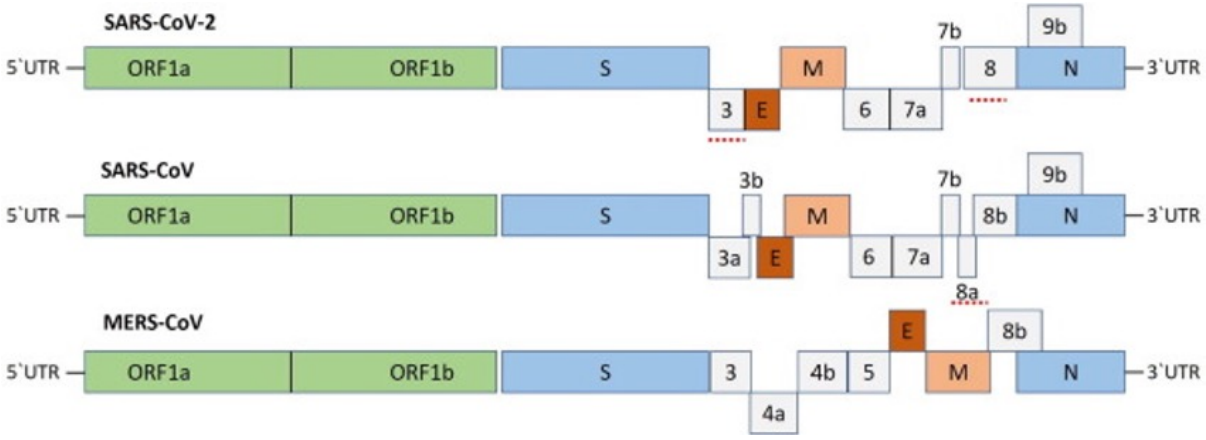

Fig 3 - Genetic similarity and variations of the genomes of SARS-Cov-1, SARS-Cov and MERS-Cov (Source: Shereen et al., 2020)

Genetic changes are also reflected by the high mutation rate shown by coronaviruses. Foster et al. (2020) applied a mathematical algorithm to a phylogenetic analysis of 160 human patient genomes and identified three variants of SARS-Cov-2 (A, B and C), with variants A and C being found in Europe and America while variant B is found in Asia. However, it should be emphasized that variant A, which is the original genome of the virus identified in Wuhan, was not the predominant variant in the city.

Xiaolu et al. (2020) analyzed 103 genomes of SARS-Cov-2 and identified two strains of this virus, designated L and S, differentiated by two single nucleotide polymorphisms or single nocleotide polymorphisms (SNPs). The L strain was identified as being more prevalent in the sample than the S strain. The lack of clarity about the implication of these evolutionary changes in the etiology of the disease suggests the need for further studies of the virus. Analysis of the viral S gene that interacts as a host cell receptor indicated the occurrence of genetic recombination (Wu et al., 2020b).

Zhang et al. (2020) suggested the occurrence of mutations as a result of classifying 27 genomes into six gene groups, despite high similarity, in a study of 27 patients from three cities in China (Wuhan, Zejiang and Guangdong) and Thailand, all with contacts from Wuhan. These experts find that the most basal gene pool was from Guangdong and that the new coronavirus pool had 380 amino acid substitutions.

Laarmarti et al. (2020) analyzed 3067 SARS-CoV-2 genomes from 59 countries using comparative genomic analysis by means of mutation profiling and frequency comparison, as well as monitoring their geography. This group of scientists identified 716 mutations, of which 457 were nonsynonymous effect mutations, 39 recurrent nonsynonymous effect mutations, including 10 hot-spot mutations with prevalence greater than 0.10 distributed in six SARS-Cov-2 genes. The study showed site-specific genotypes and the simultaneous occurrence of mutations due to the presence of several haplotypes, suggesting the action of a co-accumulation mechanism of mutations and a grouping of the viruses into 3 subgroups.

Coppe et al. (2020) reported 2334 non-synonymous mutations after analysis of SARS-CoV-2 sequences obtained from CoV_GLUE5 37 (http://cov-glue.cvr.gla.ac.uk/: 9,028 available sequences ('low coverage' excluded), including 4973 sequences from European patients. The two major mutations (S-D614G & nsp12-P323L) diverging from SARS-CoV-2 virus, NCBI Reference (NC_045512), are seen in all continents, with only 3 cases in Asia. D614G mutations were identified in the S protein (found in 2342 samples), which determines Group G, and co-evolves with the P323L mutation in the nsp12 protein (found in 2318 samples); the ORF8-L84S mutation (third most frequent mutation), which determines Group S; and the L84S amino acid substitution mutation, which co-evolves with three other mutations: 55 nsp4-F308Y, ORF3a-G196V, and N-S197L. These last three mutations, together with S197L mutations and P13L substitution in the N protein are infrequent; the fourth most frequent mutation is ORF3a-Q57H found in 734

sequences; this is followed by N-R203K and N-G204R mutations) found in Europe; and finally, we mention nsp6-L37F and ORF3a-G251V mutations, corresponding to Group V.

Dorp and Balloux (2020) concluded that the genomic sequences studied share a common ancestor that corresponds to the period when SARS-CoV-2 first infected man. These authors identified regions of the genome that did not vary and 198 recurrent mutations, with approximately 80% producing nonsynonymous changes in the S protein; over 15% recurrent mutations in the Nsp6, Nsp11 and Nsp13 regions of ORF1ab and the S protein, indicating convergent evolution of particular interest in the adaptive process of SARS-Cov-2 to man.

Eskier et al., (2020) in studying the effects of RNA-dependent RNA polymerase (RdRp) mutations, in particular the 4408C> mutation, on the mutation rate and dispersal of the virus, concluded that the 14408C>T mutation increases the mutation rate, while the 15324C>T mutation of RdRp, has the opposite effect, suggesting that the 14408C>T mutation may have contributed to the dominance of its co-mutations in different regions. Vankadari (2020), in an analysis of complete viral genomic sequences from 12 different countries, identified 47 SNPs with impact on virulence and response against antivirals, with the Nsp1, RdRp "pico" gliprotein proteins and the ORF8 region mutated within 3 months of human transmission.

Yang et all., (2020) identified six phylogenetic clusters with geographic preference in the analysis of 1932 complete genome sequences of SARS-CoV-2. These authors believe that single nucleotide variations (SNVs) in genomes underlie the results, as

well as contribute to detection, clinical treatment, drug design and vaccine development against the virus.

CONCLUSIONS

Although the coronavirus genome is small and comprises a single-stranded RNA molecule, the positive direction of the molecule allows it to translate rapidly immediately after infection of host cells, thus producing the proteins necessary for its replication. On the other hand, the high infection rate, the high specificity of the "spike" protein (S) for the host cell receptor, and the wide geographic dispersion are qualities inherent in the viruses' ability to mutate structurally and functionally. The query shows the large number of mutations recorded in virus genomes, with varying degrees of importance, and have been classified as recurrent (frequent in space and time), non-synonymous (point mutations, i.e. involving single nucleotide substitution), and "hot spot" regions (regions where mutations occur at a higher frequency) have been identified.

REFERENCES

Cristescu, Melania E. The concept of genome after one century of usage. Genome 62: iii-v (2019) dx.doi.org/10.1139/gen-2019-0129

Doğa Eskier, Gökhan Karakülah, Aslı Suner, Yavuz Oktay (2020) RdRp mutations are associated with SARS-CoV-2 genome evolution. bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.104885

Fernandez-Rua, José M (2020) Nuevo mapa genético del SARS-Cov-2. Biotech magazine & News, https://biotechmagazineandnews.com/nuevo-mapa-genetico-del-sars-cov-2/.

Hsin-Chou Yang, Chun-houh Chen, Jen-Hung Wang, Hsiao-Chi Liao, Chih-Ting Yang, Chia-Wei Chen, Yin-Chun Lin, Chiun-How Kao, and James C. Liao (2020) Genomic,

geographic and temporal distributions of SARS-CoV-2 mutations. bioRxiv preprint doi: https://doi.org/10.1101/2020.04.22.055863

Lucyvan Dorp and François Balloux (2020) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution, Volume 83, September 2020, 104351. https://doi.org/10.1016/j.meegid.2020.104351

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N. and Siddique, R. (2020) COVID-19 infection: Origin, transmission, and characteristics of human Coronaviruses. Journal of Advanced Research 24(2020):91-98.

Sen S, Anand KB, Karade S, Gupta RM. Coronaviruses: origin and evolution. Medical Journal Armed Forces India. 2020 Apr 27;76(2):136-41. doi: 10.1016/j.mjafi.2020.04.008. Epub ahead of print. PMID: 32341622; PMCID: PMC7183968.

Vankadari, N. (2020) Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. Gene. 2020 May 20;752:144792. doi: 10.1016/j.gene.2020.144792. Epub ahead of print. PMID: 32445924; PMCID: PMC7239005.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., and Song, Z. G., et al . (2020a) A new coronavirus associated with human respiratory disease in China. Nature [Epub ahead of print]. https://doi.org/10.1038/s41586-020-2008-3

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020b) Composição do genoma e divergência do novo coronavírus (2019-nCoV) originário da China. Hospedeiro celular e micróbio [Epub antes da impressão]: 1931– 3128.

Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, Jie Cui, Jian Lu (2020) Sobre a origem e evolução contínua da SRA-CoV-2 . National Science Review. https://doi.org/10.1093/nsr/nwaa036

Zhang, L., Shen, F. M., Chen, F., e Lin, Z. (2020a) Origem e evolução do novo coronavírus de 2019. Doenças Infecciosas Clínicas [Epub ahead of print]. pii: ciaa112.

Zhu N, Zhang D, Wang W, et al. (fevereiro de 2020). "A Novel Coronavirus from Patients with Pneumonia in China, 2019". The New England Journal of Medicine. 382 (8): 727–733. PMID 31978945. doi:10.1056/NEJMoa2001017