Title: Evaluation of mutagenesis from Spike protein, ORFab, ORF3a genes from SARS-CoV2 genome using Python

Nathaly Dongo, Allison Aldoradin, María de Fátima Salazar, Yomali Ferreyra, Alberto Donayre

Abstract

The SARS-CoV-2 virus, infectious agent for Coronavirus disease 2019 (COVID-19), represents a significant impact on health and daily life around the world. Recently, the virus mutations showed an increase in infections, resulting in long-term health sequelae. In this work, we conduct the analysis of SARS-CoV-2 genome using Python algorithms, demonstrating our scripting could be used to process sequencing information providing organized and real time information. We studied data from 1104 viral sequences from the United States focusing on the mutation incidence in 2020. We identified mutations as well as microsatellites in both structural and nonstructural proteins. Unlike mutations, the microsatellite analysis used different SARS-CoV-2 genomes from 21 countries available at the worldwide database NCBI.

Python algorithms showed in the ORFab gene sequence corresponding to T256I, L3352F, M2606I, and T6668W positions. In addition, a mutation from ORFab located in G172V showed high incidence during November-December of 2020 (*). seems to stabilize the β-barrel, which means hydrophobic interaction is relatively better.

Our program evidenced high rate mutations from the nonstructural ORF Gene 3a, corresponding to Q57H, G172V, T151I, and V112F. Noticeably, the mutation D614G has been detected in spike protein with 99% of incidence and has been related to the transmission increase during the last two months of 2020.

Finally, 14 of the microsatellites found using Python scripting in the SARS-Cov 2 genome are conserved in the 21 variants. Therefore, they can be used as genetic markers since they are repetitive regions in the strains, except for the poly-A tail.

Objectives

- Implement a toolbox for a quick analysis of sequencing data from SARS-CoV-2 genomes using python.
- Graphical representation of genomic variations from SARS-CoV-2 using public databases.
- Provide a tool for a robust analysis of SARS-CoV-2 mutations, microsatellite sequences, and analyzed protein sequences from structural and nonstructural viral proteins.

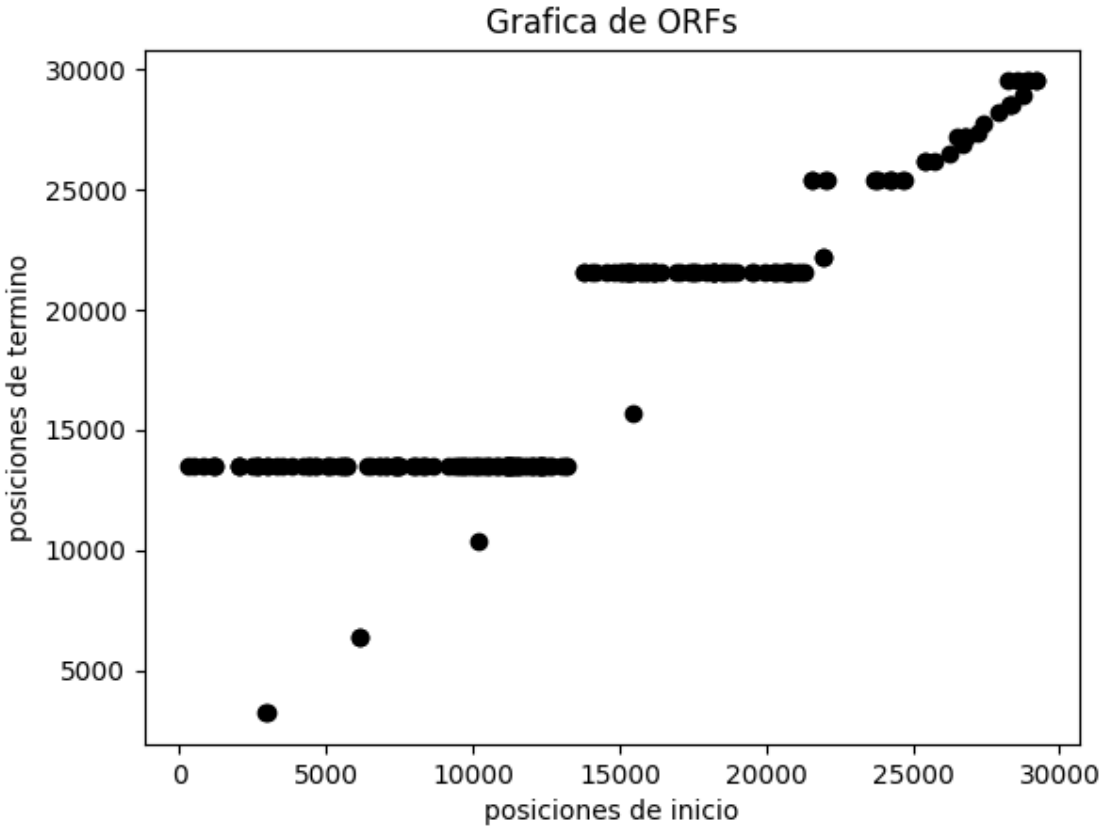

Figure 1. Proteome representation from SARS-CoV-2 using Python. The identification of the ORFs allowed us to evaluate their function in the genome. For example, finding reading frames with a regulatory function or in-frame internal ORFs within existing ORFs generating truncations at the N-terminus, forming a new polypeptide [11]. Dots in the x-axis represent AUG within the full viral genome. Dots in the y axis represent stop codons variants (UAG, UAA, UGA) from the SARS-CoV-2 genome. The figure shows the genomic location of putative sequences coding for theoretical open reading frames (ORFs). A pattern of horizontal dots produced a line. Dotted lines represent a single start codon and multiple stop codons producing viral polypeptides such as ORF1ab. Lines indicate that the SARS-CoV-2 viral genome could produce frameshifts.

A Python program selects DNA fragments that start with an AUG codon and end at one of the three known stop codons (TAA, TAG, TGA). Hypothetical gene sequences with a minimum size of 210 nucleotides are shown. The abscissa axis shows the ATG positions in the viral genome while the ordinate axis shows the positions of stop codons. The colorful dots highlight the proteins reported for SARS-CoV-2. A total of 205 hypothetical sequences were obtained, of which 180 coincide with the sequences reported for the genes that code for the 12 main SARS-CoV-2 proteins.

This part of the code specifically analyzed microsatellites. For this task, the code used 21 isolated sequences of SARS-CoV-2 that were obtained from NCBI viral databases (available at https://www.ncbi.nlm.nih.gov/sars-cov-2/). This database was accessed on 06 January 2021. The selection criteria for these genomes was to ensure to have sequences from Africa, Asia, Europe, Oceania North America and South America apart from Wuhan, which was taken as the reference genome.

Simple sequence repeats (SSRs), also known as microsatellites, refer to sequence units that are repeated in tandem in a genome ranging in length from one to six base pairs. Those short motifs of DNA are distributed ubiquitously in the genome of eukaryotes [1].

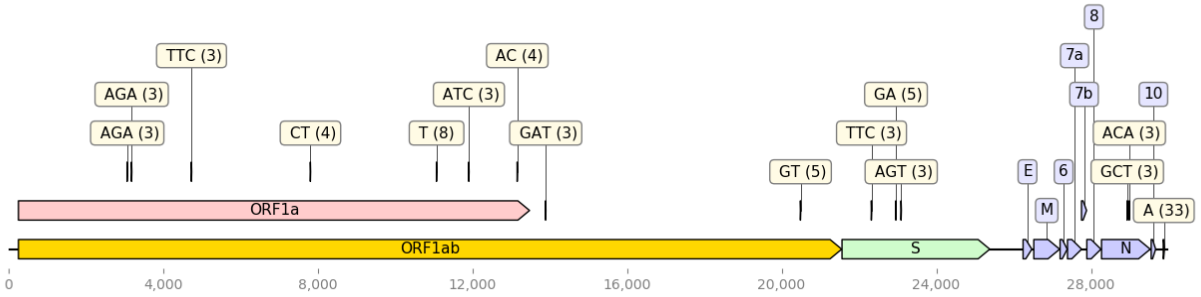

Figure 2. Microsatellites in SARS Cov 2. This graph represents 14 microsatellite markers developed for SARS-Cov 2 that were obtained using a Python algorithm. The algorithm forms DNA fragments according to the length of the atomic nucleus specified by the user and subsequently, searches if this fragment is repeated consecutively. The minimum number of repetitions that is set is 3 except for atomic nucleus lengths of 1 and 2 which are set to 7 and 4, respectively. Mononucleotide, dinucleotide and trinucleotide microsatellites were searched.

The microsatellites represented in the graph are found in the 21 species of SARS-CoV 2 so that they can be used as genetic markers since they are repetitive regions in the strains, except for the poly-A tail. Nine microsatellites are found in the coding region, which means that they have limited usefulness as a marker since they usually occur due to adaptive variations. Likewise, 5 microsatellites were identified which are more appropriate markers since they occur by evolutionary or neutral polymorphisms [2].

The next part of the code specifically analyzed mutations. In this part the code used 1104 isolated sequences of SARS-CoV-2 reported in the USA, they were obtained from NCBI viral databases on 06 January 2021 (available at https://www.ncbi.nlm.nih.gov/sars-cov-2/).

The sequences were introduced as sixth inputs in txt files with FASTA format according to their temporal distributions, separated every two months throughout 2020. As a result, the code then extracted the non-structural protein genes ORF1ab, ORF3a and structural proteín Spike for each input, as can be seen in Figures 3, 4 and 5. For each input, an output was obtained with its respective synonymous, non-synonymous, silent, nonsense mutations, deletions, and insertions were collected in dictionaries. The reference sequence taken into account was NC_045512.2 from Wuhan, with which the nomenclature of the mutations was generated.

Finally, the code developed in Python compiles the number of incidents of the mutation with higher incidence throughout the 2020 period, divides it by the total sequences corresponding to the sample (one for every two months), which vary between 150 and 200 sequences, and gives the incidence rate through a graphical code.

The reason why mutations were analyzed was that viral virulence and infectivity are some of the most important SARS-CoV-2 characteristics and are determined by its molecular structure and function, a change in the molecular structure could determine a change of the characteristics aforementioned [3].

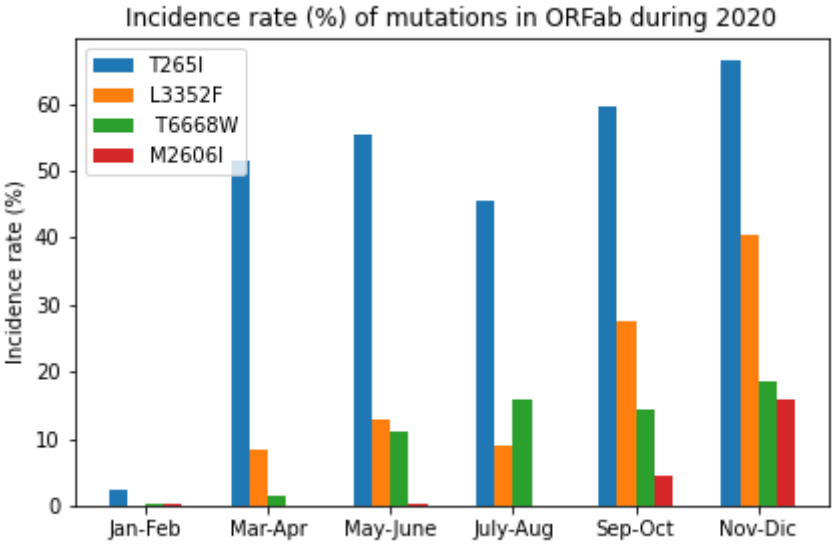

Figure 3. Mutations with greater incidence in the non-structural ORFab gene. The upper graph shows the mutations with the highest incidence in ORFab, corresponding respectively to T256I, L3352F, M2606I, T6668W.

According to the graph, in the first months of the year, there were no incidents in the rate of mutations. As of March, incidences were observed in three mutations (T256I, L3352F and T6668W); of which, the T256I obtains the highest percentage with 51.24%. This mutation preserves logarithmic growth, except for the months of July-August, the maximum incidence rate reached is 66.17% between November-December.

On the other hand, the L3352F mutation does not present relatively high levels (greater than 50%) until the end of the year (40.30%). L3352F creates a mutation from leucine (L) to phenylalanine (F) at residue 89 which could improve hydrophobic packaging and improve protein stability [4]. M2606I and T6668W mutations reached proportions of 15% and 18% in the last November-December two-month period. Also, it was found that M2606I mutation affected the nsp3 protein within what is predicted to be the 3Ecto or C-terminal domain, involved in the anchoring of the replication-transcription complex (RTC) to the endoplasmic reticulum membrane [4]. Finally, no information was found regarding the T256I and T6668W mutations in the non-structural ORFab gene, although the T256I mutation obtained the highest incidence throughout all the months.

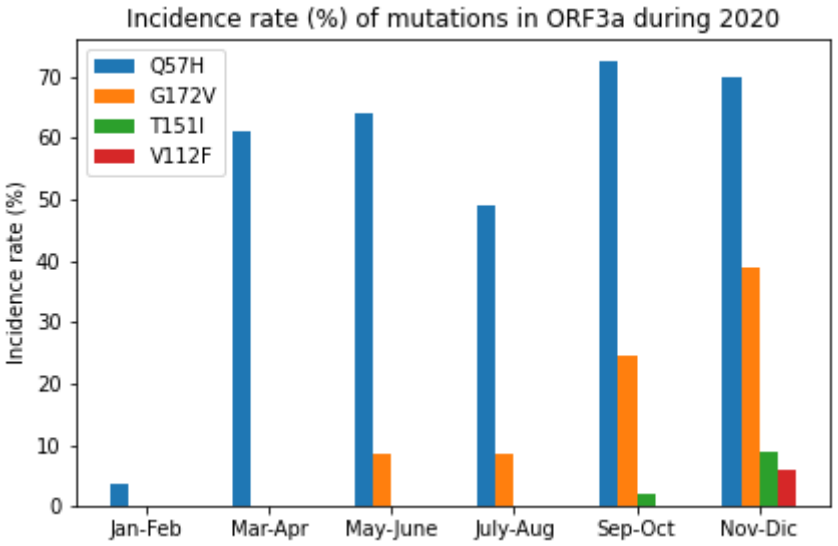

Figure 4. Mutations with greater incidence in the non-structural ORF3a gene. The upper graph shows the mutations with the highest incidence in ORF3a, corresponding respectively to Q57H, G172V, T151I, and V112F.

As can be seen in the graph, the Q57H mutation reaches extremely high levels during the first two months, starting with percentages from 3.33% in January-February to 81.33% in March-April. This mutation preserves logarithmic growth, except for the months of July-August, the maximum incidence rate reached is 96.67% between September-October, ending the 2020 period with 93.33%. The Q57H mutation changes the amino acid glutamine (Q) with a non-charged polar side chain to the positively charged polar side chain of amino acid histidine (H). This tendency could indicate that mutation Q57H becomes popular in the viral patients of the United States, which may make the SARS-CoV-2 more infectious [5].

On the other hand, the G172V mutation does not show relatively high levels (greater than 50%) until the end of the year (52%). Finally, the T151I and V112F mutations reached ratios of 12% and 8% in the last November-December two-month period. So far, G172V may be speculated to stabilize the β-barrel by adding hydrophobic interaction. The b-barrel is a domain linked to virulence, infectivity, and virus release [6]. On the other hand, the T151F and V112F mutations have been reported in the literature, but no evidence was found to demonstrate their effect on the stability of the protein.

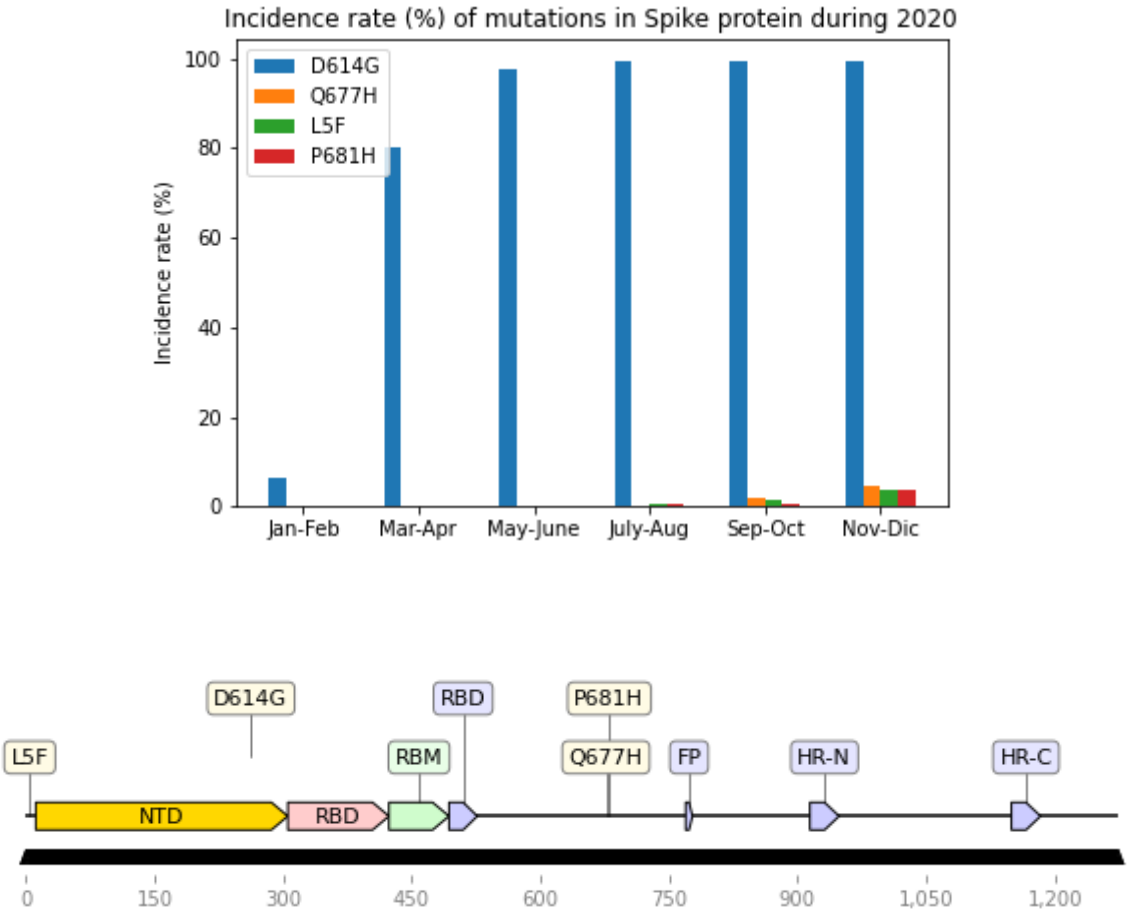

Figure 5. (a) Number of mutations per bimester for the pandemic in the United States. The upper graph shows the mutations with the highest incidence in Spike protein, corresponding respectively to D614G, Q677H, L5F, and P681H.

As can be seen in figure 5.a, the D614G mutation shows a large increase in its incidence since March-April. Reaching values of 99% from May-June, much higher than the rest of the Q677H, L5F, and P681H mutations, which barely reached notable values from the month of November-December. D614G rose from 8.7% in the first two months to 99%, the effects of D614G mutation were studied and it was demonstrated that it "enhances viral replication in human lung epithelial cells and primary human airway tissues by increasing the infectivity and stability of virions" [7], which means the mutation enhances infective agent masses within the higher tract of COVID-19 patients which will increase transmission.

(b) Mutations in the spike protein (protein S) of SARS-CoV-2 from the United States. Due to the high levels of incidence of D614G, a code was created to detect mutations and know in which part of the spike protein occurs. Thus, L5F and D614G mutations were found in NTD sites in the N Terminal Domain (NTD), which function is to contribute to the conformational changes in the protein during the interaction with a host cell; meanwhile, P681H and Q677H were discovered between RBD and FP sites. Regarding the first mutation, it has been found that the L5F mutation is independent of the D614G mutation [8]. Besides, the P681H non-synonymous mutation has been observed in global data outside and likely represents independent, various studies indicating that it may contribute to the

transmission of the virus [9]. On the other hand, information on the Q677H mutation has not been recorded. Finally, the D614G mutation is one of the most studied since this mutation is associated with higher viral loads and younger patients; yet, greater severity of the infection has not been demonstrated [10]. This implies that there is a need to know what effects this mutation causes as a function of structural change and molecular interaction.

References

[1] Sahu, B.P., Majee, P., Singh, R.R. et al. Comparative analysis, distribution, and characterization of microsatellites in Orf virus genome. Sci Rep 10, 13852 (2020). https://doi.org/10.1038/s41598-020-70634-6

[2] Davis, C. L. et al. (1999). Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. Journal of virology, https://doi.org/10.1128/JVI.73.8.6265-6270.1999

[3] Wang, R., Chen, J., Gao, K. et al. Analysis of SARS-CoV-2 mutations in the United States suggests the presence of four substrains and novel variants. Commun Biol 4, 228 (2021). https://doi.org/10.1038/s42003-021-01754-6.

[4] Pater, AA et al. (2021). Appearance and evolution of a new prevalent variant of SARS-CoV-2 in the United States. doi: https://doi.org/10.1101/2021.01.11.426287.

[5] R. Wang, J. Chen, K. Gao, Y. Hozumi, C. Yin, and G. Wei, “Characterizing SARS-CoV-2 mutations in the United States,” 2020.

[6] M. Bianchi, A. Borsetti, M. Ciccozzi, and S. Pascarella, “SARS-Cov-2 ORF3a: Mutability and function,” International Journal of Biological Macromolecules, vol. 170, pp. 820–826, 2021.

[7] D. C. Groves, S. L. Rowland-Jones, and A. Angyal, “The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design,” Biochemical and Biophysical Research Communications, 2020.

[8] R. Wang, J. Chen, K. Gao, Y. Hozumi, C. Yin, and G. Wei, “Characterizing SARS-CoV-2 mutations in the United States,” 2020.

[9] Pauloluniyi, “Detection of SARS-CoV-2 P681H Spike Protein Variant in Nigeria,” Virological, 23-Dec-2020. [Online]. Available: https://virological.org/t/detection-of-sars-cov-2-p681h-spike-protein-variant-in-nigeria/567.

[10] R. Garry, “Mutations arising in SARS-CoV-2 spike on sustained human-to-human transmission and human-to-animal passage,” Virological, Jan. 02, 2021. https://virological.org/t/mutations-arising-in-sars-cov-2-spike-on-sustained-human-to-human-transmission-and-human-to-animal-passage/578 (accessed Feb. 19, 2021).

[11] Finkel, Y., Mizrahi, O., Nachshon, A. *et al.* The coding capacity of SARS-CoV-2. *Nature* 589, 125–130 (2021). https://doi.org/10.1038/s41586-020-2739-1