Multidisciplinary Sciences Reports

Original Article

IN SILICO COMPARATIVE ANALYSIS OF SARS-COV-2 MUTATIONS IN BRAZIL

1 Gabriel Ferrari de Oliveira eng.gabriel.ferrari@gmail.com

2 Sarah de Oliveira Rodrigues sarahrodrigues232@gmail.com

3 Kolawole Banwo kolabanwo@yahoo.com

4 Isabela Bacelar de Assis isabela_bacelar@yahoo.com

5 Celso Iwata Frison celso@pucpcaldas.br

6 Jorge Pamplona Pagnossa jorgepampa@gmail.com

1 Pontifícia Universidade Católica – PUCMINAS, Poços de Caldas, Minas Gerais, Brasil

3 University of Ibadan Faculty of Science: Ibadan, Oyo, Nigeria

Mult. Sci. Rep. 2023; v. 3 n. 3 / ISSN: 2764-0388
DOI: https://doi.org/10.54038/ms.v3i3.46
Subject: 04/06/2023–Accepted: 22/08/23

RESUMO

SARS-CoV-2 é um novo coronavírus que surgiu no fim de 2019 na China. Ele causa Covid-19, uma doença que se tornou pandemia semanas depois do primeiro caso e é responsável por infectar e matar milhões de pessoas ao redor do mundo. Desde o primeiro surto, a comunidade científica tem procurado medidas terapêuticas e profiláticas contra a Covid-19. O objetivo desta pesquisa é trazer discussões que possam contribuir para o entendimento do vírus e o desenvolvimento de tratamentos e prevenções contra a doença, além de validar uma metodologia que possa ajudar no entendimento e controle de outros surtos virais. Para este propósito, 5016 amostras de SARS-CoV-2 coletadas no Brasil foram analisadas através de recursos computacionais. Este trabalho apresenta os resultados da árvore filogenética, entropia da informação do genoma e gráficos e tabelas mostrando informações sobre as mutações do SARS-CoV-2 no Brasil. Com base nesses resultados, evidenciou-se a importância da proteína espícula para a alta transmissibilidade do vírus.

PALAVRAS-CHAVE: Coronavírus. Bioinformática. Pandemia. Nextclade. Proteína espícula.

ABSTRACT

SARS-CoV-2 is a novel coronavirus that emerged in late 2019 in China. It causes Covid-19, a disease that became pandemic weeks after the first case and is responsible for infecting and killing millions of people worldwide. Since the first outbreak, the scientific community has searched for therapeutics and prophylactics measures against Covid-19. This research aims to bring discussions that may contribute to the understanding of the virus and the development of treatments and preventions against the disease, besides validating a methodology that may help understanding and controlling other viral outbreaks. For this purpose, 5016 samples of SARS-CoV-2 collected in Brazil were analysed through computational resources. This work presents the results of the phylogenetic tree, genome information entropy and graphs and tables showing information about SARS-CoV-2 mutations in Brazil. Based on these results, the importance of the spike protein in the high transmissibility was highlighted.

KEYWORDS: Coronavirus. Bioinformatics. Pandemic. Nextclade. Spike protein.

DOI: https://doi.org/10.54038/ms.v3i3.46
Subject: 04/06/2023–Accepted: 22/08/23

INTRODUCTION

In late 2019, in Wuhan City, China, the first case of the disease known as Coronavirus Disease 2019 (or simply Covid-19) emerged. Among the various symptoms caused by Covid-19, fever, cough, myalgia, and fatigue are the most common, in addition to pneumonia present in most infected people. In a few weeks since the first case of Covid-19, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for causing this disease, spread worldwide, infecting thousands, and killing hundreds of people. Until March 11, 2020, the World Health Organization (WHO) declared the Covid-19 pandemic [1].

SARS-CoV-2 is a coronavirus that originated in bats. In taxonomy, the novel CoV belongs to the family *Coronaviridae*, subfamily *Orthocoronavirinae*, genera *Betacoronavirus* and subgenera *Sarbecovirus*. Although SARS-CoV-2 does not cause SARS, it received this name because of its phylogenetic similarity to SARS-CoV, in which approximately 80% of the genomes of the two viruses are the same [2,3].

As with all other coronaviruses, SARS-CoV-2 has a single genomic RNA fragment oriented in a positive direction with a size between 27 and 32 kilobases [2,3]. Its genome has 14 open-reading frames (ORFs). Two of these fourteen ORFs, ORF1a and ORF1b, are responsible for translating polyproteins that produce the virus's 16 non-structural proteins (NSPs). The other ORFs, such as ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF9b, are accessory proteins. Also, there are four structural proteins: spike (S), nucleocapsid (N), membrane (M) and envelope (E) [3].

Every time a virus makes its copies in human cells, it is subject to errors that lead to mutations in the genetic code. Since the novel CoV is an RNA virus, it is highly variable. According to Grudlewska-Buda et al. [4], one mutation every ten thousand nucleotides occurs during the replication process of a CoV and, according to Yeh & Contreras [5], SARS-CoV-2 has an average mutation rate of 7.23 virus mutations. Its high variability generates new variants rapidly, causing mutations as the virus evolves, using the information acquired by the development of immunity to distinguish genetic mutations [5]. Genetic changes in the virus or host cells tend to alter the virus-cell interaction rapidly. The regulatory elements that affect cell tissue tropism and pathology can act on transcription, RNA processing, stability and transport levels, translation, and protein processing. The very high complexity of change compromises cellular functionality in general [6].

Genomic and evolutionary analyses, such as taxonomic, and phylogenetic and evolutionary analyses [7], and molecules structures analyses, such as determination and prediction of 3D structures of proteins and drug design [8], are some examples of the fields that computational resources can aid to understand the Covid-19 pandemic and to develop drugs. So, a bioinformatics approach applied to SARS-CoV-2, as proposed by this work, can be a powerful tool against the virus and the disease. Also, the results of this research may help the development of drugs to combat future outbreaks caused by coronaviruses that do not yet exist, as occurred in the current pandemic: studies about SARS-CoV and middle east respiratory syndrome coronavirus (MERS-CoV) were used to produce vaccines against Covid-19 [9].

The main objective of this research is to bring discussions that contribute to the understanding of the SARS-CoV-2 and the production of therapeutics and prophylactic measures against Covid-19 and maybe other coronaviruses diseases, besides developing a methodology that aids in outbreaks caused by different viruses. The specific objectives are identifying patterns and making statistics on SARS-CoV-2 mutations in Brazil.

METHODOLOGY

To achieve the objectives listed above, a computational methodology similar to Pagnossa et al. was used [10].

The computational resources used were:

- Python, a high-level and object-oriented programming language widely used in science [11].
- Biopython, an open-source library in bioinformatics [11].
- Nextclade, a bioinformatic tool capable of indicating the quality control (QC) metrics of the nucleotide sequences, making the multiple sequence alignment, building the phylogenetic tree, detecting mutations, and presenting the genome information entropy of SARS-CoV-2 about the original strain from Wuhan in 2019 (GenBank accession: MN908947) [12].
- Google Sheets, an online spreadsheet software developed by Google.

The first step was to import the sequences of Brazilian strains of SARS-CoV-2 from the National Center for Biotechnology Information (NCBI) database and organise them into FASTA format. For this purpose, a Python program with the Biopython library was used. Firstly, the algorithm searches in the NCBI database for genomes with the following terms: “SARS-CoV-2”, “Complete Genome”, “Human”, and “BRA”, referring to the country Brazil. Then, the sequences were imported and organised in FASTA format. As of the date of this research, late August 2022, 7016 strains of SARS-CoV-2 found in Brazil with complete genomes were available on NCBI.

The next step was to input the FASTA file in Nextclade. Due to computational resource limitations, 5016 out of 7016 sequences were randomly selected to be processed in the bioinformatics tool. After the processing, the Nextclade gives the following outputs: phylogenetic tree, genome information entropy and a tab-separated values (TSV) file containing information about the mutations of each sample.

A second Python algorithm was created to analyse the information from the TSV file. Firstly, the samples were divided into five groups to improve the analysis. Then, the sequences with a low QC metric (incomplete and with ambiguous bases, for example [12]) were removed. After this, for each group, five statistical analyses were performed: occurrences of nucleotide and amino acid substitutions, average reversion to root per sample, amount of amino acids substitutions per gene as a function of its length, percentage of the presence of amino acid substitutions in each gene and the 25 most frequent amino acid substitutions.

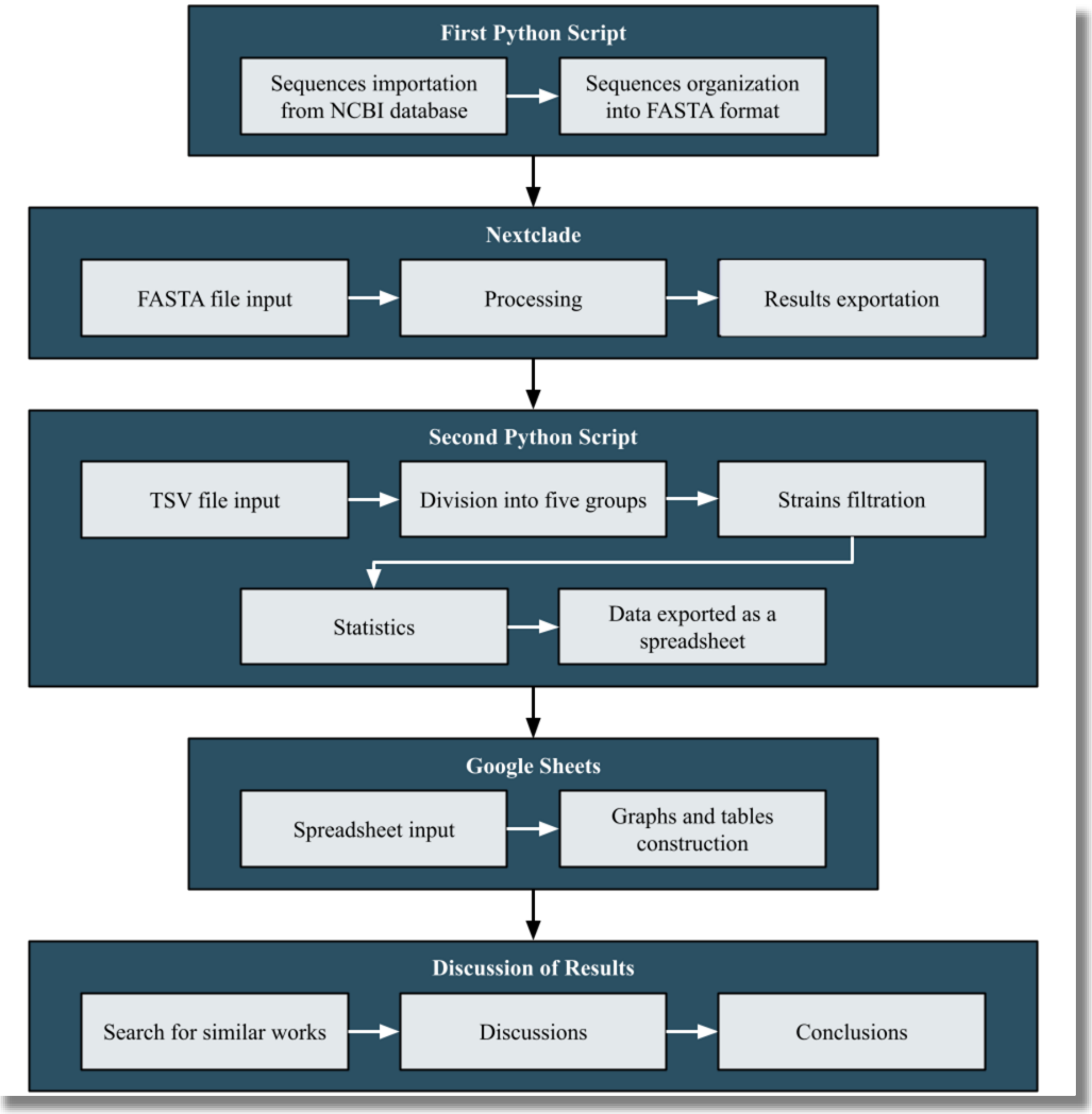
These statistics were exported to a spreadsheet and loaded into Google Sheets, where the graphs and the tables were built, besides the third-degree polynomial regression (the curve that best suited) performed in the amount of amino acids substitutions per gene as a function of its length analysis.

Finally, the results were discussed. In this stage, similar works and works about SARS-CoV-2 mutations were searched on Google Scholar to help discuss the results of this research.

The methodology is summarized in the form of a flowchart in Figure 1.

Figure 1 – The Methodology Flowchart

Source: Own elaboration (2022)



DOI: https://doi.org/10.54038/ms.v3i3.46
Subject: 04/06/2023–Accepted: 22/08/23

RESULTS AND DISCUSSIONS

This section is divided into eight subsections, each one containing analysis/result and its discussion. The results in the order they appear are phylogenetic tree, number of samples, occurrences of nucleotide and amino acid substitutions, average reversion to root per sample, the genome information entropy, amount of amino acid substitutions per gene as a function of its length, percentage of the presence of amino acid substitutions in each gene and the 25 most frequent amino acid substitutions.

The objective of the topics “phylogenetic tree” and “genome information entropy” is to bring a macro and qualitative view of SARS-CoV-2 mutations in Brazil. This overview allows for building some hypotheses and insight into the behaviour of SARS-CoV-2 in Brazil. Later, in the other subsections, these highlights are tested and discussed quantitatively.

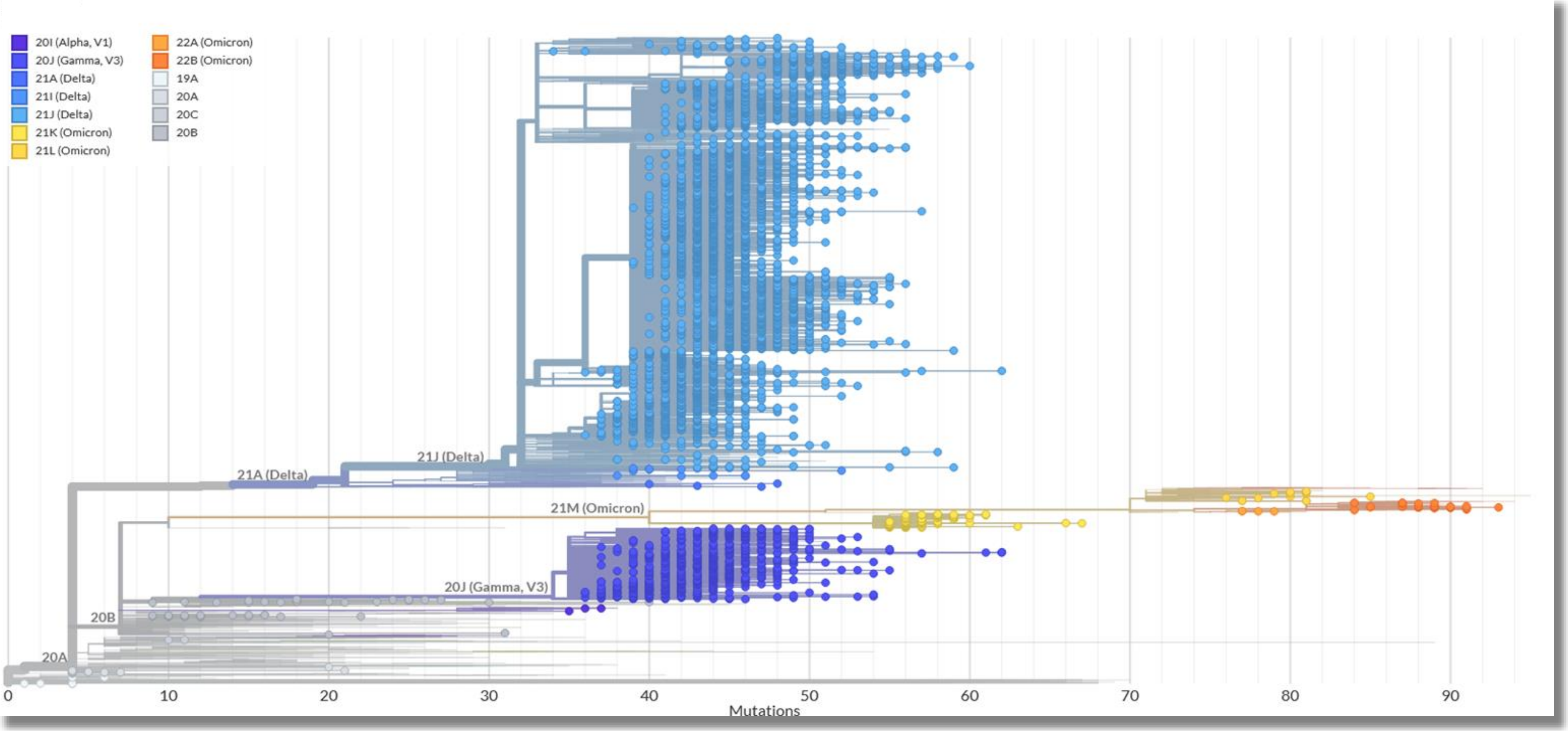
Phylogenetic Tree

A phylogenetic tree is a vital tool to easily visualise how a species evolves, how far it is from its ancestral, the ramifications indicating divergence in the evolutionary process, and the distance between the variants.

In this perspective, it is necessary to analyse the variants in a phylogenetic tree format to emphasise the variants of SARS-CoV-2 virus over time in a didactic way. The phylogenetic tree of the Brazilian strains of SARS-CoV-2 can be seen in Figure 2. The dots indicate the strains imported from the NCBI database and they are plotted over the Nextclade reference tree. Each color represents a clade of the novel coronavirus, and its legend is in the upper left corner. On the bottom is a scale showing the divergence about the original sample from Wuhan. Thus, samples further to the right are samples with more mutations and more distant from the original Wuhan strain. The divergence considers the nucleotide substitutions and reversions to the root. The reversion to the root means the nucleotide mutated back to its original form.

Figure 2 – Phylogenetic Tree

Source: Own elaboration (2022)



The most frequent variants identified in the phylogenetic tree in descending order are the Delta variant, represented by the colour light blue; the Gamma variant, represented by the colour dark blue; and the Omicron variant, represented by the colours yellow and orange. The Alpha variant, represented by the colour purple, and the variants 19A, 20A, 20C, and 20B, represented by the shades of gray, are present in a minimal amount.

The variants with low sampling are older and with less divergence. From the 20A variant, the variants of concern (VOCs) were generated. While the Delta variant originated directly from the 20A variant, the Alpha, Gamma, and Omicron VOCs originated from the 20B intermediate variant. Thus, Alpha, Gamma, and Omicron variants are phylogenetically closer about the Delta variant.

The Delta variant originated in India in October 2020 [13] and rapidly became the dominant VOC of SARS-CoV-2 in the world [14]. It is characterised for being 60% more transmissible than the Alpha variant and having a high viral load in animal models [13]. Through the phylogenetic tree, a divergence between 34 and 62 mutations was identified to this VOC.

The Gamma variant is very similar to the Delta variant. It occurred for the first time one month later in Brazil, with a divergence between 36 and 62 mutations. Regarding the transmissibility, it has a slight increase, 1.7-2.5 times higher than the first variants and having a high viral load in the reinfection case [13].

The Omicron variant originated in South Africa and Botswana. Despite being 3.2 times more transmissible than the Delta variant, thus the most infectious variant of SARS-CoV-2 until now, and having replaced Delta as the dominant variant, the phylogenetic tree shows that it is the less frequent in analysed strains. This may occur because this VOC is the most recent, being reported to the WHO in November 2021, so a few samples were sequenced and uploaded into the database [14]. This VOC has mutations between 55 and 93, becoming the most divergent variant.

These VOCs have essential mutations in the spike glycoprotein protein [13,14], suggesting that this region of the SARS-CoV-2 genome is an essential element in understanding the virus. This protein is located on the surface of the coronaviruses and has a shape that resembles a crown, hence the name “coronavirus”. It is responsible for receptor recognition and binding to the host human cell through the angiotensin-converting enzyme 2 (ACE2) receptor and for membrane fusion [3] and it is one of the main targets of neutralising antibodies produced by the body to block the virus [15].

Number of Samples

Initially, a total of 5016 Brazilian samples of SARS-CoV-2 were analysed. This number of strains (before filtering) is presented in the results “phylogenetic tree” and “genome information entropy”. Subsequently, some samples were removed, remaining 4767 samples (after filtering). This number is found in the results “occurrences of nucleotide and amino acid substitutions”, “average reversion to root per sample”, “amount of amino acids substitutions per gene as a function of its length”, “percentage of occurrence of amino acid substitutions per gene” and “the most frequent amino acid substitutions”.

Based on the previous subsection discussion about the variants' occurrence, the samples were divided into five groups. Three of these groups refer to the most frequent variants: Delta variant, Gamma variant, and Omicron variant; one relates to variants that appear less frequently (Alpha, 19A, 20A, 20C, and 20B); and the last one refers to all variants. The numbers of samples used in this study and their fragmentation into groups, are placed in Table 1. Note that the number of the “all variants” group is the sum of the numbers of the other groups.

Table 1 – Number of Samples Analysed

Source: Own elaboration (2022)

Group	Amount of Samples	
	Before Filtering	After Filtering
All variants	5016	4767
Delta variant	4041	3837
Gamma variant	720	695
Omicron variant	149	149
Other variants (Alpha, 19A, 20A. 20C and 20B)	86	86

Occurrences of Nucleotide and Amino Acid Substitutions

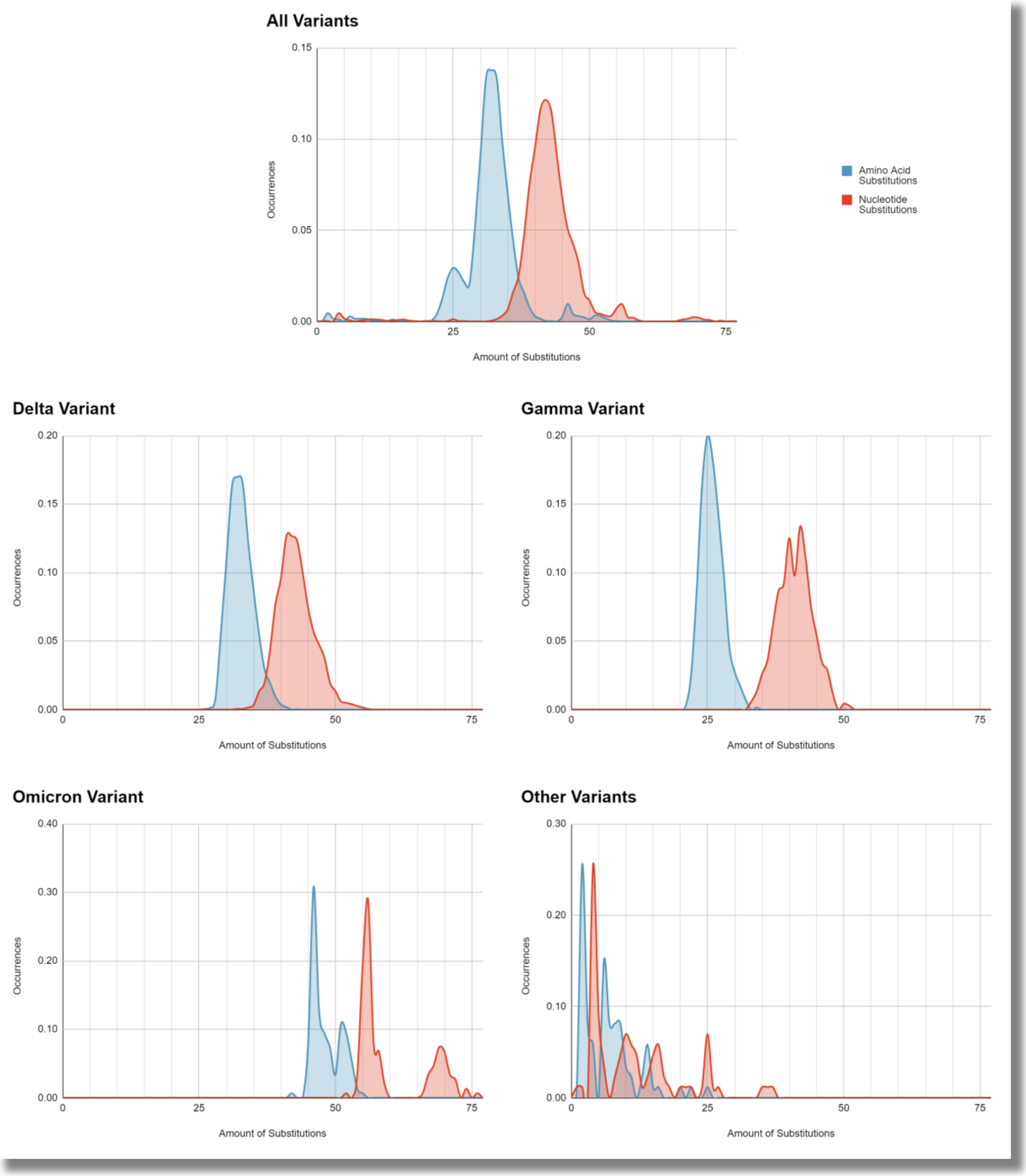
SARS-CoV-2 mutations result from two significant mechanisms: spontaneous nucleotide substitution/deletion and RNA recombination. Nucleotides are definable as molecules composed of three components: a pentose, a nitrogen base, and one or more phosphate groups. RNA and DNA are formed by the esterification reaction between phosphoric acid and nucleotides [16].

Proteins are substances formed by amino acids linked together through peptide bonds. It is understood that amino acid substitutions in protein S in specific areas have aided in their mutations. Amino acid substitution is the change of an amino acid in a protein by a different amino acid due to the point mutation in the DNA sequence. It is caused by a non-synonymous exchanged sense mutation that alters the codon sequence to encode other amino acids instead of the original ones [17].

Given this, it is essential to emphasise the analysis of nucleotide and amino acid substitutions. To obtain the occurrences (i.e., the number of strains that have the respective amount of substitutions), the algorithm counted how many strains have the same amount of amino acid substitutions and how many strains have the same amount of nucleotide substitutions. The results are presented in Figure 3, where the "amount of substitutions" is plotted over the X-axis and the "occurrences" in percentage is plotted over the Y-axis.

Figure 3 – Occurrences of Nucleotide and Amino Acid Substitutions

Source: Own elaboration (2022)



The occurrence of nucleotide and amino acid substitutions follows the phylogenetic tree results about the divergence: the greater the divergence, the greater the number of substitutions. The Omicron variant leads the number of substitutions ranking (far right of Fig. 3). There are Delta and Gamma variants, almost tied in second and third place, respectively (they are in the center of the graph). And finally, with the least substitutions, there is the group of other variants (they are concentrated on the extreme left of the graph).

These results indicate a possible relation between the transmissibility of the variant and its number of substitutions. The Omicron variant has the highest number of substitutions and transmissibility; the Delta and Gamma variants are intermediate in both issues; and the group of other variants has the lowest number of substitutions and transmissibility. Thus, the greater the number of substitutions, the more transmissible the variant tends to be.

Another important aspect is the substantial difference between the number of nucleotide substitutions and the divergence in the subvariants 21L, 22A, and 22B, all belonging to the Omicron.

Average Reversion to Root per Sample

The purpose of this study is to validate, numerically, the idea brought up in the previous paragraph. Particularly at this stage, the Omicron variant group was divided in two: one containing the subvariant 21K and the other containing the subvariants 21L, 22A and 22B. Here, the script counted the number of reversions to root and divided it by the number of samples.

Table 2 – Average Reversion to Root per Sample of each Group

Source: Own elaboration (2022)

Group	Average Reversion to Root per Sample
All variants	0.28
Delta variant	0.32
Gamma variant	0.09
Omicron variant (21K)	0.35
Omicron variant (21L, 22A and 22B)	0.89
Other variants	0.06

Table 2 confirms what was perceived through a qualitative view. The 21L, 22A, and 22B subvariants of Omicron variants have an average reversion to the root of 0.89 per sample, the group with the highest value.

The 21K subvariant, also belonging to the Omicron variant, has an average reversion to root per sample similar to the Delta variant, 0.35 and 0.32, respectively. The group of the other variants has the lowest average reversion to root per sample; its value is 0.06. Perhaps, the justification is that older samples form this group; thus, there was no time for them to mutate and mutate again to the original form found in the Wuhan strain.

The second group with the lowest average reversions to root per sample is the Gamma variant. This VOC is the second most transmissible until now, suggesting that the reversions to root are not related to the virus's transmissibility.

Genome Information Entropy

Originally, information theory was developed to analyse communication systems. However, later on, it was applied to other areas, including the biological sciences. Through this analysis, it provides the concept of information entropy, which measures uncertainty and complexity of the information. In other words, high entropy indicates a high uncertainty and complexity, and vice versa [18].

When applied to the genomic information, high entropy indicates a high level of genetic diversity and, consequently, a high mutation rate. This means that genetic information encompasses a wide range of possible states, suggesting a greater likelihood of genetic mutations occurring [18].

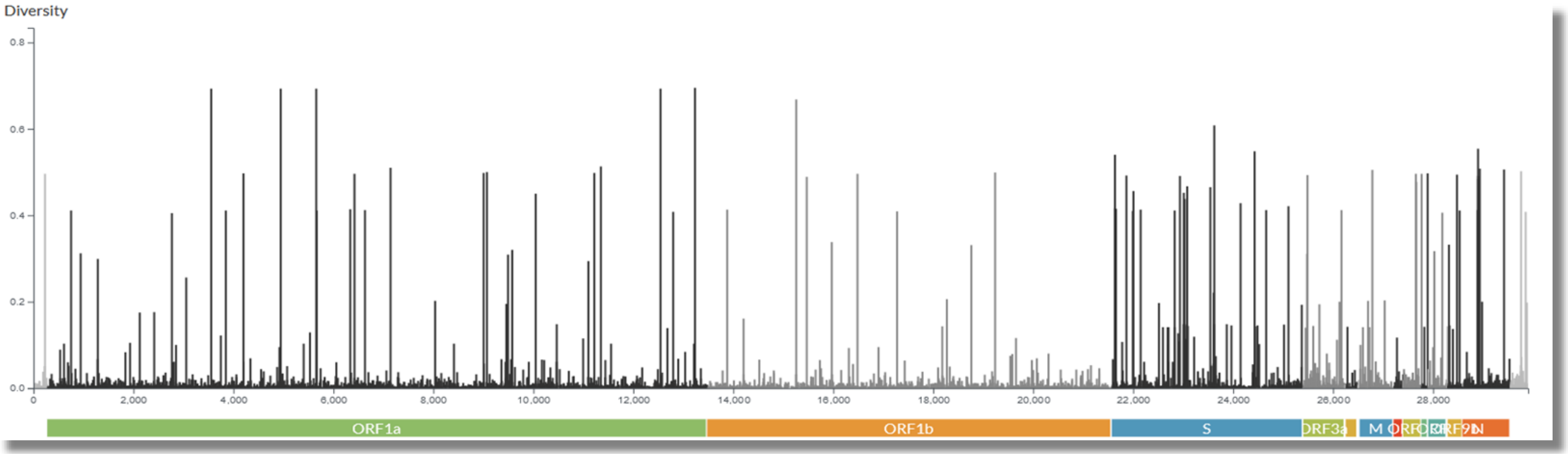
Entropy can be determined by different causes which prevent communication with each other. One of the determining causes for the mutation of just one gene, correlated with a high rate of entropy, is selective pressure, defined as any set of environmental conditions that originate the favouring of specific genes over others in a given population. Cross-pressure correlated with high entropy may result from less intense pressure on a particular gene and may occur when a gene is not essential for virus replication or survival, allowing viability for different emergencies [19].

Another factor is genetic recombination, the process of exchanging genetic material with living or non-living organisms. If the specific gene is involved in a recombination process, it can give rise to high entropy due to incorporating different sequences. In addition, a high mutation rate can promote changes in a viral genome; that is, a particular area of the genome can have a higher mutation rate than others [20].

In the same context of the phylogenetic tree, Figure 4 shows the samples' genome information entropy. The X-axis represents the SARS-CoV-2 genome and indicates the 12 coding regions (in sequence: ORF1a, ORF1b, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, ORF9b, and N) and the 29903 nucleotides. The Y-axis represents the entropy, which means, the longer the bar, the higher quantity of changes the nucleotide has suffered.

Figure 4 – Genome Information Entropy

Source: Own elaboration (2022)



Analysing the genome information entropy, two highlights are noticed. The first concerns the size of the genes. In descending order, the ORF1a, ORF1b and S genes are visibly the largest ones, while the nine genes are tiny and have a similar length. Therefore, it is expected that in a set of strains of SARS-CoV-2, the ORF1a, ORF1b, and S genes have the highest number of mutations. Extending this idea to the other genes, it is expected that the number of mutations in a gene is directly proportional to the gene size.

The second concerns the high density of S protein. The area occupied by the entropy bars in the S protein appears to be larger than the areas in other regions of the virus genome. It suggests that this region has more entropy, thus the one that mutates most.

Amount of Amino Acid Substitutions per Gene as a Function of Its Length

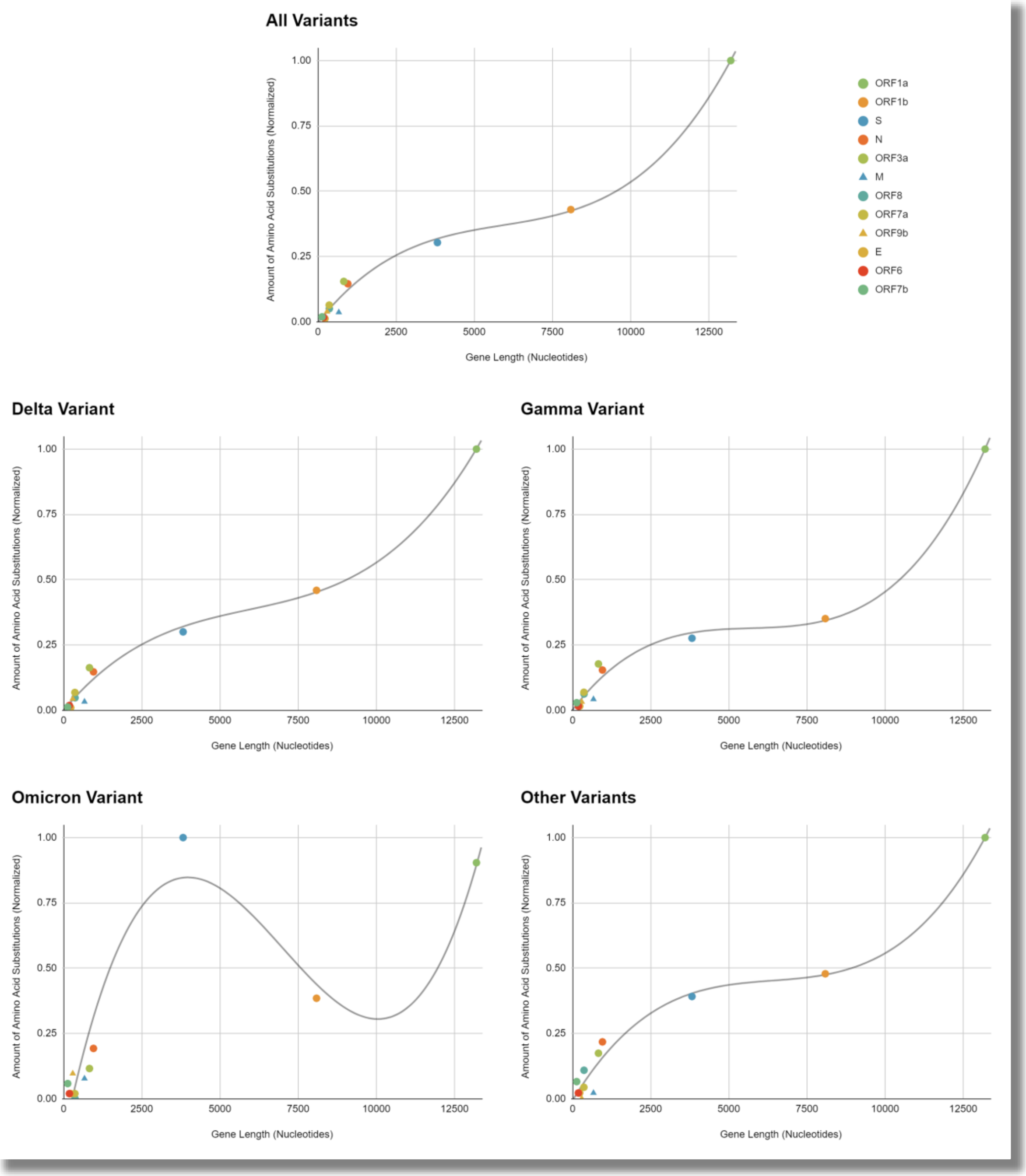
To test the hypothesis that the ORF1a, ORF1b and S genes have the highest number of mutations and the direct proportionality between a number of mutations and gene size, this analysis was proposed. The script verified the samples and summed one in the respective gene for each mutation that occurred at least once — the recurrences were not accounted for. These results are compiled in Table 3, along with their respective percentages and gene lengths. Then, the values of amino acid substitutions were normalised to build the scatter plot "amount of amino acid substitutions per gene as a function of its length" (Figure 5). In this graph, the dots represent the genes, and the line is the third-degree polynomial regression that relates the number of amino acid substitutions to the gene length.

Table 3 – Genes with Their Respective Lengths and Amount of Amino Acid Substitutions. Source: Own elaboration (2022).

Gene	Length	Amount of Amino Acids Substitutions				
		All Variants	Delta Variant	Gamma Variant	Omicron Variant	Other Variants
ORF1a	13202	1908 (44.16%)	1521 (43.53%)	599 (45.07%)	47 (31.33%)	46 (39.32 %)
ORF1b	8088	819 (18.95%)	698 (19.98%)	210 (15.80%)	20 (13.33%)	22 (18.80%)
S	3822	578 (13.38%)	456 (13.05%)	165 (12.42%)	52 (34.67%)	18 (15.38%)
N	956	277 (6.41%)	223 (6.38%)	92 (6.92%)	10 (6.67%)	10 (8.55%)
ORF3a	828	294 (6.80%)	247 (7.07%)	106 (7.98%)	6 (4.00%)	8 (6.84%)
M	669	68 (1.57%)	49 (1.40%)	25 (1.88%)	4 (2.67%)	1 (0.85%)
ORF8	366	95 (2.20%)	72 (2.06%)	37 (2.78%)	0 (0.00%)	5 (4.27%)
ORF7a	362	121 (2.80%)	103 (2.95%)	41 (3.09%)	1 (0.67%)	2 (1.71%)
ORF9b	294	73 (1.69%)	66 (1.89%)	19 (1.43%)	5 (3.33%)	0 (0.00%)
E	228	22 (0.51%)	14 (0.40%)	9 (0.68%)	1 (0.67%)	1 (0.85%)
ORF6	186	31 (0.72%)	27 (0.77%)	9 (0.68%)	1 (0.67%)	1 (0.85%)
ORF7b	132	35 (0.81%)	18 (0.52%)	17 (1.28%)	3 (2.00%)	3 (2.56%)
Total		**4321 (100%)**	**3494 (100%)**	**1329 (100%)**	**150 (100%)**	**117 (100%)**

Figure 5 – Amount of Amino Acid Substitutions per Gene as a Function of its Length

Source: Own elaboration (2022)



Firstly, it is observed that all variants, except the Omicron variant, have similar behaviour. All of them present a third-degree polynomial with no extrema, a rising point of inflection and similarity in numbers. About the coefficient of determination (R2), a high value of R2 was obtained, indicating that the number of mutations can be related to the gene length through a third-degree polynomial. R2 values are 0.99 for the Delta variant, 0.99 for the Gamma variant, and 0.98 for the group of other variants.

The exception in the Omicron variant occurred due to the anomaly in the S protein. The S protein, the third largest gene, has the highest amount of amino acid substitutions, which gives the function a maximum and a minimum point. Due to this discrepancy, the R2 value of the Omicron variant is 0.92. Disregarding this anomaly, the omicron would have a similar behaviour to the other variants and a higher value of R2.

Considering the first part of the hypothesis, the most significant genes by far, ORF1a, ORF1b and S, are the ones that have the highest number of amino acid substitutions in all groups, proving this hypothesis. However, the second part of the hypothesis is validated only in Delta and Gamma variants and other variants, where the number of mutations in a gene is directly proportional to the gene size. In the Omicron variant, as commented before, the third largest gene has the highest number of mutations, and the most significant gene is in second place in the number of mutations.

The anomaly due to the low sampling of the Omicron variant, was discarded for two reasons. The first is because the group of other variants has a lower selection and has a behaviour similar to the Delta and Gamma variants. The second is because when 149 (the number of strains of the Omicron variant) samples of Delta or Gamma variants are randomly selected and analysed, they exhibit the same behaviour as when they are highly sampled.

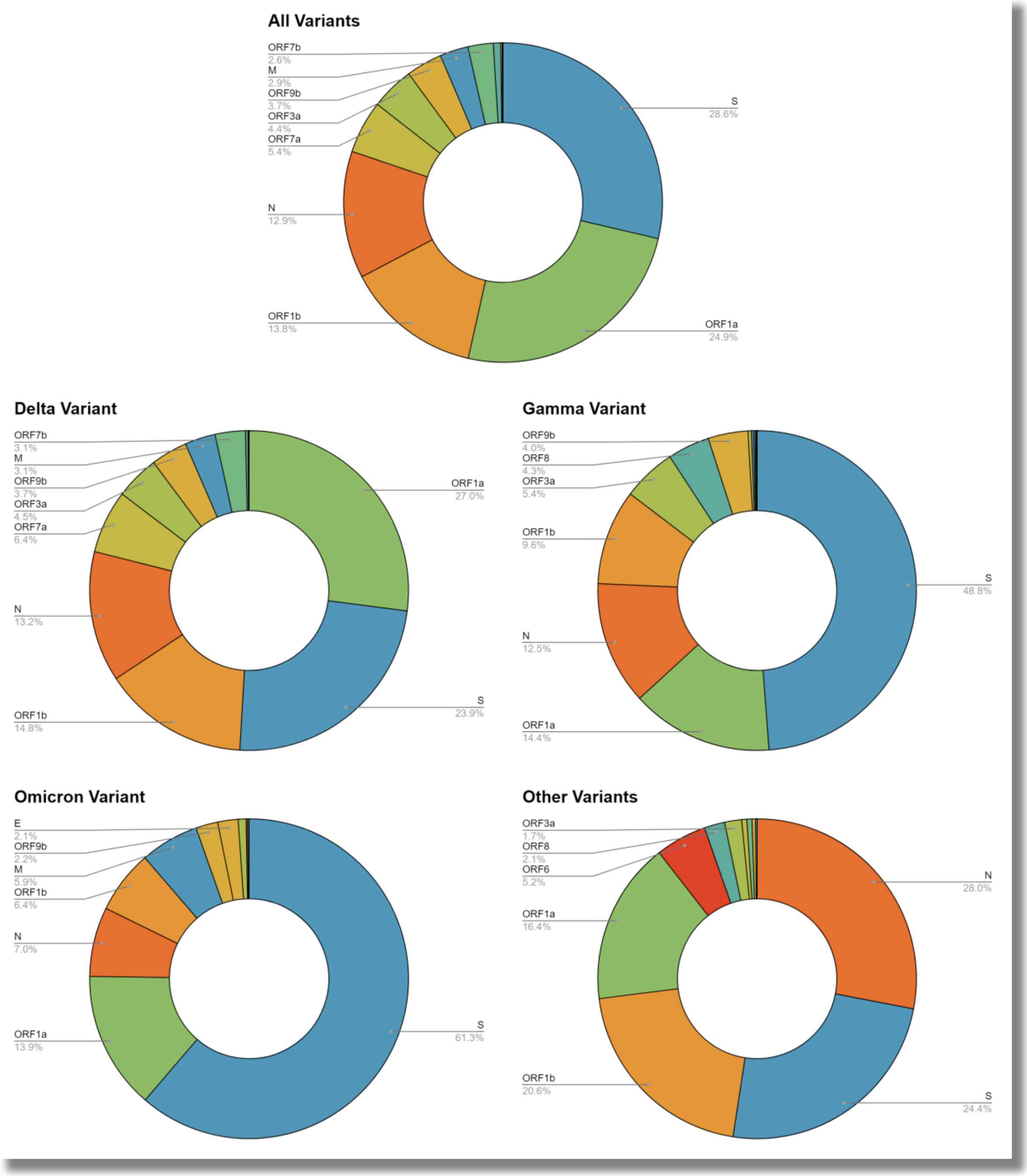
Due to the anomaly occurring only in the Omicron variant, the most transmissible VOC, it can be said that this is a possible explanation why this VOC has the highest transmissibility.

The other genes have a similar size compared to the ORF1a, ORF1b, and S. Consequently, they have identical amounts of mutations. In addition to standard deviation, some of these genes may not follow the idea that the larger the gene, the higher the number of mutations. But in general, when a group of SARS-CoV-2 samples is analysed, the number of mutations in a gene is directly proportional to the gene size and mathematically modeled by a third-degree polynomial.

Percentage of the Presence of Amino Acid Substitutions in Each Gene

This analysis aims to identify the genes with the highest percentage of amino acid substitutions. For this end, an algorithm similar to the previous study was used, but the recurrences were accounted for here. After the counts, the results were divided by the total number of mutations of all genes and multiplied by 100 to get the final results in percentage. The results are presented in Figure 6.

Figure 6 – Percentage of the Presence of Amino Acid Substitutions in Each Gene. Source: Own elaboration (2022)



Common to all groups, the structural proteins S and N and the genes responsible for producing the NSPs, ORF1a and ORF1b, are the coding regions with the highest amino acid substitutions. In the five groups, these four genes together represent more than 75% of the mutations. The other structural proteins, M and E, and the auxiliary proteins have a very low percentage.

Despite ORF1a and ORF1b genes being more significant than the S gene and generally having the highest amount of amino acid substitutions, the S protein has a significant number of mutations. It can be stated in first place in the groups “all variants”, “Gamma variant”, and “Omicron variant”, and in second place in the groups “Delta variant” and “other variants”.

Also, there is a possible connection between these percentages and the transmissibility of the variants. The Omicron variant, the most transmissible VOC, has the highest percentage of S protein (61.3%). The Gamma variant, the second most transmissible VOC, has the second highest percentage of S protein (48.8%). The Delta variant and the group of other variants, the two least transmissible, have similar results for S protein: it is in second place and has 23,9% and 24,4%, respectively. The difference is in the gene that is in the first place, ORF1a for the Delta variant and N for the group of other variants.

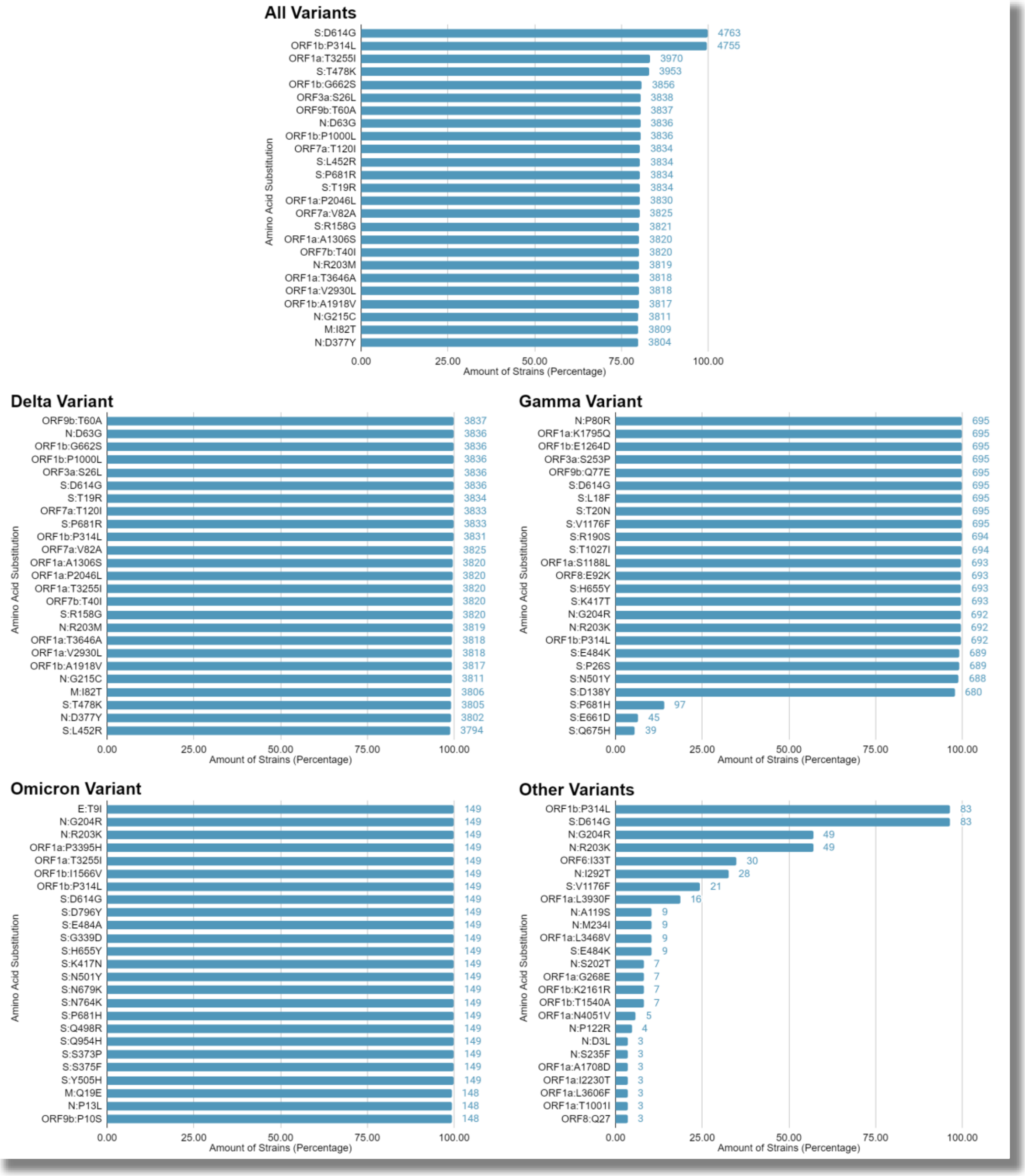
Therefore, these results indicate that ORF1a, ORF1b, N, and in particular S genes are coding regions that contribute to increasing the virus’s transmissibility.

The most Frequent Amino Acid Substitutions

The last analysis (Figure 7) exhibits the 25 most frequent amino acid substitutions. The Y-axis shows the gene followed by its mutation; they are sorted first in descending order of frequency and then in alphabetical order. The X-axis shows the percentage of strains that have the respective mutation. The number of strains that have the respective mutation is in front of each bar.

Figure 7 – 25 Most Frequent Amino Acid Substitutions

Source: Own elaboration (2022)



DOI: https://doi.org/10.54038/ms.v3i3.46
Subject: 04/06/2023–Accepted: 22/08/23

When all variants are considered, two amino acid substitutions draw attention: “S: D614G” and “ORF1b: P314L”, present in almost 100% of the samples and very distant from the third most frequent mutation. These two mutations seem to be linked to each other, which corroborates the research performed by Sharawy et al. [21] and Biswas et tal. [22].

The D614G mutation that occurs in the spike gene is the most frequent in Brazil, and it is observable in 99.92% of the samples analysed (4763 out of 4767 samples). This mutation occurred first in France before spreading to other countries and becoming dominant in the world [23,24]. The gene is located at position 614 of the spike protein, where aspartic acid (D) is replaced by glycine (G). It increases the viral load of SARS-CoV-2, making it more contagious, but it does not interfere with the severity of the disease. In addition to the study conducted by Zhang et al. [24], two other studies achieved the same result regarding the frequency, one analysing Egyptian strains [21] and the different analysing Philippine strains [25].

The second most frequent mutation is present in 99.75% of the samples (4755 out of 4767) and happens in the ORF1b gene, specifically, in NSP12. This mutation consists of the substitution of proline (P) by leucine (L) at position 314 in ORF1b, which corresponds to position 323 in the NSP12 [22]. The NSP12, also known as RNA-dependent RNA polymerase (RdRp), is one of the NSPs responsible for SARS-CoV-2 RNA replication and transcription [3,22]. The RdRp enzyme catalyses the replication of viral RNA. Thus, a mutation in NSP12 may increase the virus infectivity and the severity of the disease [22].

Visualising the “S: D614G” and “ORF1b: P314L” mutations from the evolutionary biology perspective supports the idea that they increase the transmission rate of the virus. The fact that these two mutations are present in approximately 100% of the samples indicates that they brought some survival advantages, probably related to the increase in transmissibility of the virus.

Notably, the Gamma variant has 22 amino acid substitutions that appear in almost all samples. The other mutations occur in less than 14% of samples. This may indicate that these 22 mutations only characterise the Gamma variant.

Regarding the Omicron variant, it is observed that a large part of the amino acid substitutions belongs to the S protein. Of the 22 most frequent mutations in this VOC, 15 are in the S protein and are in 100% of the samples. It may be related to the high transmissibility of this VOC.

CONCLUSIONS

Firstly, the methodology constructed to study SARS-CoV-2 can be easily applied to other viruses to analyse their mutation and help in outbreak situations. However, because it is a new proposed methodology, it may contain topics to be optimised. Therefore, more studies should be done to improve and validate this methodology which has shown to be promising.

Regarding SARS-CoV-2, the spike protein was the most important coding region. Besides being the interface between the virus and the human cell, it is the gene most highlighted in the results. In addition, the results indicate that the high transmissibility of SARS-CoV-2 is mainly linked to the S protein. Thus, producing drugs whose target is S protein has advantages and disadvantages. The advantage is that the drug will attack the viral structure that makes the first contact with the host cell; the disadvantage is that the drug may become obsolete because of the high mutability of this gene.

This research brought information about SARS-CoV-2 mutations in Brazil as of the end of August 2022. This knowledge may aid future studies in developing therapeutic and prophylactic measures against Covid-19 and as well as other diseases caused by coronaviruses because of their genomic similarity. Also, it adds to scientific knowledge, allowing new works based on it to be done. Further studies are required to understand how the results of this work can be used.

ACKNOWLEDGEMENTS

The authors would like to thank for the Pontifical Catholic University for the support in all stages of research.

CONFLICT OF INTEREST

The authors declare there are no conflict of interest.

REFERENCES

1. Carvalho, T., Krammer, F., & Iwasaki, A. (2021). The first 12 months of COVID-19: a timeline of immunological insights. *Nature Reviews Immunology*, *21*(4), 245-256.
2. Khalil, O. A. K., & da Silva Khalil, S. (2020). SARS-CoV-2: taxonomia, origem e constituição. *Revista de Medicina*, *99*(5), 473-479.
3. Arya, R., Kumari, S., Pandey, B., Mistry, H., Bihani, S. C., Das, A., ... & Kumar, M. (2021). Structural insights into SARS-CoV-2 proteins. *Journal of molecular biology*, *433*(2), 166725.
4. Grudlewska-Buda, K., Wiktorczyk-Kapischke, N., Wałecka-Zacharska, E., Kwiecińska-Piróg, J., Buszko, K., Leis, K., ... & Skowron, K. (2021). SARS-CoV-2—morphology, transmission and diagnosis during pandemic, review with element of meta-analysis. *Journal of Clinical Medicine*, *10*(9), 1962.
5. Yeh, T. Y., & Contreras, G. P. (2021). Full vaccination against COVID-19 suppresses SARS-CoV-2 delta variant and spike gene mutation frequencies and generates purifying selection pressure. *Medrxiv*, 2021-08.
6. Sánchez, C. M., Izeta, A., Sánchez-Morgado, J. M., Alonso, S., Sola, I., Balasch, M., ... & Enjuanes, L. (1999). Targeted recombination demonstrates that the spike gene of transmissible gastroenteritis coronavirus is a determinant of its enteric tropism and virulence. *Journal of virology*, *73*(9), 7607-7618.
7. Hu, T., Li, J., Zhou, H., Li, C., Holmes, E. C., & Shi, W. (2021). Bioinformatics resources for SARS-CoV-2 discovery and surveillance. *Briefings in bioinformatics*, *22*(2), 631-641.
8. Waman, V. P., Sen, N., Varadi, M., Daina, A., Wodak, S. J., Zoete, V., ... & Orengo, C. (2021). The impact of structural bioinformatics tools and resources on SARS-CoV-2 research and therapeutic strategies. *Briefings in Bioinformatics*, *22*(2), 742-768.
9. Li, Y. D., Chi, W. Y., Su, J. H., Ferrall, L., Hung, C. F., & Wu, T. C. (2020). Coronavirus vaccine development: from SARS and MERS to COVID-19. *Journal of biomedical science*, *27*(1), 1-23.
10. Pagnossa, J. P., Rodrigues, S. D. O., Oliveira, G. F. D., Adnan, M., Aljaid, M. S., Assis, I. B. D., ... & Batiha, G. E. S. (2023). COVID-19 in a Pre-Omicron Era: A Cross-Sectional Immuno-Epidemical and Genomic Evaluation. *Vaccines*, *11*(2), 272.

11. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423.
12. Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A. (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of open source software*, *6*(67), 3773.
13. Thye, A. Y. K., Law, J. W. F., Pusparajah, P., Letchumanan, V., Chan, K. G., & Lee, L. H. (2021). Emerging SARS-CoV-2 variants of concern (VOCs): an impending global crisis. *Biomedicines*, *9*(10), 1303.
14. Fan, Y., Li, X., Zhang, L., Wan, S., Zhang, L., & Zhou, F. (2022). SARS-CoV-2 Omicron variant: recent progress and future perspectives. *Signal transduction and targeted therapy*, *7*(1), 141.
15. Bai, C., & Warshel, A. (2020). Critical differences between the binding features of the spike proteins of SARS-CoV-2 and SARS-CoV. *The Journal of Physical Chemistry B*, *124*(28), 5907-5912.
16. Na, W., Moon, H., & Song, D. (2021). A comprehensive review of SARS-CoV-2 genetic mutations and lessons from animal coronavirus recombination in one health perspective. *Journal of Microbiology*, *59*, 332-340.
17. Leparc-Goffart, I., Hingley, S. T., Chua, M. M., Jiang, X., Lavi, E., & Weiss, S. R. (1997). Altered pathogenesis of a mutant of the murine coronavirus MHV-A59 is associated with a Q159L amino acid substitution in the spike protein. *Virology*, *239*(1), 1-10.
18. Vinga, S. (2014). Information theory applications for biological sequence analysis. *Briefings in bioinformatics*, *15*(3), 376-389.
19. Ghanchi, N. K., Nasir, A., Masood, K. I., Abidi, S. H., Mahmood, S. F., Kanji, A., ... & Hasan, R. (2021). Higher entropy observed in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. *PloS one*, *16*(8), e0256451.
20. Fan, Z., Yao, B., Ding, Y., Zhao, J., Xie, M., & Zhang, K. (2021). Entropy-driven amplified electrochemiluminescence biosensor for RdRp gene of SARS-CoV-2 detection with self-assembled DNA tetrahedron scaffolds. *Biosensors and Bioelectronics*, *178*, 113015.
21. Sharawy, L., Tantawy, M., Ahmed, Y., Taha, A., Soliman, O., Ibrahim, T. M., & El-Hadidi, M. (2020, October). In-Silico Comparative Analysis of Egyptian SARS CoV-2 with Other Populations: a Phylogeny and Mutation Analysis. In *2020 2nd Novel*

Intelligent and Leading Emerging Sciences Conference (NILES) (pp. 618-622). IEEE.

22. Biswas, S. K., & Mudi, S. R. (2020). Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. *Genomics & informatics*, *18*(4).
23. Li, M., Prasad, N., Hall, D., & Wu, H. (2020, December). Analysis of SARS-CoV-2 sequences reveals transmission path and emergence of S D 614G mutation. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1995-1998). IEEE.
24. Zhang, J., Cai, Y., Xiao, T., Lu, J., Peng, H., Sterling, S. M., ... & Chen, B. (2021). Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science*, *372*(6541), 525-530.
25. Velasco, J. M., Chinnawirotpisan, P., Joonlasak, K., Manasatienkij, W., Huang, A., Valderama, M. T., ... & Klungthong, C. (2020). Coding-complete genome sequences of 23 SARS-CoV-2 samples from the Philippines. *Microbiology resource announcements*, *9*(43), e01031-20.