Temporal evolution and adaptation of SARS-COV 2 codon usage.

Maddalena Dilucca[1,2,*], Sergio Forcelloni[1], Alexandros G. Georgakilas[3], Andrea Giansanti[1,4] Athanasia Pavlopoulou[5,6]

1 Physics Department, Sapienza University of Rome, Rome, Italy
2 Liceo Scientifico Statale Augusto Righi, Rome, Italy
3 DNA Damage Laboratory, Physics Department, School of Applied Mathematical and Physical Sciences, National Technical University of Athens (NTUA), Athens, Greece
4 INFN Roma1 unit, Rome, Italy
5 Izmir Biomedicine and Genome Center (IBG), 35340 Balcova, Izmir, Turkey
6 Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, 35340 Balcova, Izmir, Turkey

*** maddalena.dilucca@gmail.com**

Abstract

The outbreak of severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) has caused an unprecedented pandemic. Since the first sequenced whole-genome of SARS-CoV-2 on January 2020, the identification of its genetic variants has become crucial in tracking and evaluating their spread across the globe.

In this study, we compared 15,259 SARS-CoV-2 genomes isolated from 60 countries since the outbreak of this novel coronavirus with the first sequenced genome in Wuhan to quantify the evolutionary divergence of SARS-CoV-2. Thus, we compared the codon usage patterns, every two weeks, of 13 of SARS-CoV-2 genes encoding for the membrane protein (M), envelope (E), spike surface glycoprotein (S), nucleoprotein (N), non-structural 3C-like proteinase (3CLpro), ssRNA-binding protein (RBP), 2-O-ribose methyltransferase (OMT), endoRNase (RNase), helicase, RNA-dependent RNA polymerase (RdRp), Nsp7, Nsp8, and exonuclease ExoN.

As a general rule, we find that SARS-CoV-2 genome tends to diverge over time by accumulating mutations on its genome and, specifically, on the coding sequences for proteins N and S. Interestingly, different patterns of codon usage were observed among these genes. Genes S, $Nsp7$, $NSp8$, tend to use a norrower set of synonymous codons that are better optimized to the human host. Conversely, genes E and M consistently use a broader set of synonymous codons, which does not vary with respect to the reference genome. We identified key SARS-CoV-2 genes (S, N, $ExoN$, $RNase$, $RdRp$, $Nsp7$ and $Nsp8$) suggested to be causally implicated in the virus adaptation to the human host.

1 Introduction

The recent emergence of the novel, human pathogen Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in China and its rapid spread poses a global health emergency. On March 11 2020, WHO publicly declared the SARS-CoV-2 outbreak as a pandemic. As of May 27, 2020, the COVID-19

pandemic had affected more than 190 countries and territories, with more than 5,731,956 confirmed cases and 353,854 deaths ($https://www.worldometers.info/coronavirus/$).

The size of the SARS-CoV2 genome is approximately 30 kb and its genomic structure is characteristic of that of known coronaviruses. SARS-CoV-2 genome is translated into two large overlapping polyproteins, ORF1a and ORF1ab. These polyproteins are cleaved into five structural structural proteins, including spike protein (S), membrane protein (M), envelope protein (E), nucleocapsid protein (N) and 26 non-structural proteins. There are also nine putative ORFs (ORF3a, ORF3b, ORF4, ORF5, ORF6, ORF7a, ORF7b, ORF8 and ORF10) predicted as hypothetical proteins [1].

Characterization of viral mutations can provide valuable information for assessing the mechanisms linked to pathogenesis, immune evasion and viral drug resistance. In addition, viral mutation studies can be crucial for the design of new vaccines, antiviral drugs and diagnostic tests. Mutation rate of RNA viruses is dramatically higher than their hosts. This high mutation rate is correlated with virulence modulation and evolvability, which are considered beneficial for viral adaptation [2]. Tang and coworkers have recently characterized 13 variation sites in SARS-CoV-2: ORF1ab, S, ORF3a, ORF8 and N regions, among which positions 28144 in ORF8 and 8782 in ORF1a showed a mutation rate of 30.53% and 29.47%, respectively. A previous study based on an analysis of 103 genomes of SARS-CoV-2 indicates that this virus has evolved into two main types: type L is derived from the more ancestral type S and is more widespread compared to type S [3].

According to data from the public database Global Initiative on Sharing All Influenza Data (GISAID), three major clades of SARS-CoV-2 can be identified and are referred to as clade G (variant of the spike protein S-D614G), clade V (variant of the ORF3a coding protein NS3-G251), and clade S (variant ORF8-L84S) [4]. In particular, Giorgi et. al showed that clade G, prevalent in Europe, carries a D614G mutation in the Spike protein, which is responsible for the initial interaction of the virus with the host human cells [5] .

In the present study, we investigate the evolution of SARS-CoV-2 genomes and codon usage patterns over time, as well as virus adaptation to the human host. For this purpose, we focused on 13 SARS-CoV-2 genes encoding the structural proteins membrane (M), envelope (E), spike surface glycoprotein (S) and nucleoprotein (N); the non-structural 3C-like proteinase (3CLpro) which is the main protease responsible for cleaving the viral polyprotein into individual proteins, ssRNA-binding protein (RBP), 2'-O-ribose methyltransferase (OMT), endoRNase (RNase), helicase, RNA-dependent RNA polymerase (RdRp), ExoN which plays a crucial role in proofreading by correcting any errors made by RdRp, as well as Nsp7and Nsp8 which form a supercomplex with RdRp.

2 Materials and Methods

2.1 Sequence data analyzed

A total of 16,352 SARS-CoV-2 genomes reported across the world were obtained from GISAID (available at https://www.gisaid.org/epiflu-applications/nexthcov-19-app/), on 09 May 2020. Then, the sequences were classified according to their isolation dates. Only complete genomes (28-30 Kb) were included in the present analysis. Thus, a list of 15,259 SARS-CoV-2 genomes was generated; representing 94% of total number in GISAID. We used the SARS-CoV-2 coding DNA sequences (CDSs) deposited in January 2020 by Zhu and coworkers [6], formerly called "Wuhan seafood market pneumonia virus" (WSM, NC_045512.2). We retrieved these sequences from NCBI public database at $https://www.ncbi.nlm.nih.gov/$. The CDSs of the reference SARS CoV-2 genome (NC_045512.2) were used to retrieve the homologous

protein-coding sequences from the 15,259 genomes under study, by using Exonerate with default parameters [7].

2.2 Effective Number of Codons Analysis

We calculated the effective number of codons (ENC) to estimate the extent of the codon usage bias of SARS-CoV-2 genes. The values of ENC range from 20 (when just one codon is used for each amino acid) to 61 (when all synonymous codons are equally used for each amino acid) [8]. For each sequence, the computation of ENC starts from F_α, a quantity defined for each family α of synonymous codons:

$$F_\alpha = \left(\frac{n_{k\alpha}}{n_\alpha}\right)^2 \tag{1}$$

where m_α is the number of different codons in α (each one appearing $n_{1_\alpha}, n_{2_\alpha}, ..., n_{m_\alpha}$ times in the sequence) and $n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha}$.

Finally, the gene-specific ENC is defined as:

$$ENC = N_s + \frac{K_2 \sum_{\alpha=1}^{K_2} n_\alpha}{\sum_{\alpha=1}^{K_2}(n_\alpha F_\alpha)} + \frac{K_3 \sum_{\alpha=1}^{K_3} n_\alpha}{\sum_{\alpha=1}^{K_3}(n_\alpha F_\alpha)} + \frac{K_4 \sum_{\alpha=1}^{K_4} n_\alpha}{\sum_{\alpha=1}^{K_4}(n_\alpha F_\alpha)} \tag{2}$$

where N_S is the number of families with one codon only and K_m is the number of families with degeneracy m (the set of 6 synonymous codons for Leu can be split into one family with degeneracy 2, similar to that of phenylalanine (Phe), and one family with degeneracy 4, similar to that, for example, of proline (Pro).

ENC was evaluated by using DAMBE 5.0 [9].

2.3 Codon Adaptation Index

The codon adaptation index (CAI) [10, 11] was used to quantify the extent of codon usage adaptation of SARS-CoV-2 to the human coding sequences. The principle behind CAI is that the codon usage in highly expressed genes can reveal the optimal (i.e., most efficient for translation) codons for each amino acid. Hence, CAI is calculated based on a reference set of highly expressed genes to assess, for each codon i, the relative synonymous codon usages ($RSCU_i$) and the relative codon adaptiveness (w_i):

$$RSCU_i = \frac{X_i}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_j}; \qquad w_i = \frac{RSCU_i}{\max_{j=1,...,n_i}\{RSCU_j\}}; \tag{3}$$

In the $RSCU_i$, X_i is the number of occurrences of codon i in the genome, and the sum in the denominator runs over the n_i synonymous codons. $RSCU$ measures codon usage bias within a family of synonymous codons. w_i is defined as the usage frequency of codon i compared to that of the optimal codon for the same amino acid encoded by i (i.e., the the most used one in a reference set of highly expressed genes). Finally, the CAI value for a given gene g is calculated as the geometric mean of the usage frequencies of codons in that gene, normalized to the maximum CAI value possible for a gene with the same amino acid composition:

$$CAI_g = \left(\prod_{i=1}^{l_g} w_i\right)^{1/l_g}, \tag{4}$$

where the product runs over the l_g codons belonging to that gene (except the stop codon).

This index ranges from 0 to 1, where the score 1 represents a greater tendency of the gene to use optimal codons in the host organism. The CAI analysis was performed using $DAMBE$ 5.0 [9]. The synonymous codon usage data of human host were retrieved from the codon usage database (`http://www.kazusa.or.jp/codon/`).

2.4 Similarity Index

The similarity index (SiD) was used to provide a measure of similarity in codon usage between SARS-CoV-2 and various potential host genomes. Formally, SiD is defined as follows:

$$R(a,b) = \frac{\sum_{k=1}^{59} a_i \cdot b_i}{\sqrt{\sum_{k=1}^{59} a_i^2 \cdot \sum_{k=1}^{59} b_i^2}} \tag{5}$$

$$SiD = \frac{1 - R(a,b)}{2} \tag{6}$$

where a_i is the RSCU value of 59 synonymous codons of the SARS-CoV-2 coding sequences; b_i is the RSCU value of the identical codons of the potential host. R(a,b) is defined as the cosine value of the angle included between A and B spatial vectors, and therefore, quantifies the degree of similarity between the virus and the host in terms of their codon usage patterns. In our analysis, we considered as hosts the species shown in Figure 3 by Dilucca et al. [12]. SiD values range from 0 to 1; the higher the value of SiD, the more adapted the codon usage of SARS-CoV-2 to the host [13].

ENC plot

An ENC plot analysis was performed to estimate the relative contributions of mutational bias and natural selection in shaping CUB of 13 genes encoding proteins that are crucial for SARS-CoV-2. In this plot, the ENC values are plotted against GC_3 values. If codon usage is dominated by the mutational bias, then a clear relationship is expected between ENC and GC_3:

$$ENC = 2 + s + \frac{29}{s^2 + (1-s)^2} \tag{7}$$

s represents the value of GC_3 [8]. If the mutational bias is the main force affecting CUB of the genes, the corresponding points will fall near the the Wright's theoretical curve. Conversely, if CUB is mainly affected by natural selection, the corresponding points will fall considerably below the Wright's theoretical curve.

To quantify the relative extent of the natural selection, for each gene, we calculated the Euclidean distance d of its point from the theoretical curve. We then show the average values of the distance over time with heatmap, drawn with the CIMminer software [14], which uses Euclidean distances and the average linkage algorithm.

2.5 Estimation of Evolutionary Divergence

MUSCLE program was used to conduct multiple sequence alignments [15, 16]. Estimates of Evolutionary Divergence (i.e., the number of base substitutions per site) from the reference sequence (WSM, NC_045512.2) was calculated by using the Maximum Composite Likelihood model, implemented in the software package MEGA version 10.1 [17], based on sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option).

Distributions of divergence were assessed for each group of viruses divided in different period. The statistical analysis was performed with the R software to calculate average values and their standard deviations. Mann-Whitney tests on distributions were used to test for statistical significance. All p-values were calculated from 2-sided tests using 0.05 as cut-off.

2.6 Protein-Protein Network Analysis

In this study, we used the 332 high-confidence SARS-CoV-2-human protein-protein interactions (PPI) collected by Gordon et al. [19] who identified

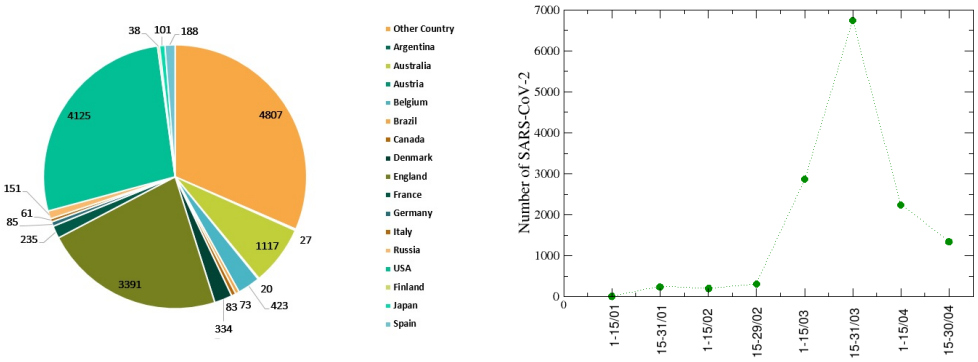

Figure 1. Distribution of the 15,259 genomes used in this study by country and date of isolation. On the left, the pie chart represents the number of genomes used in this study according to their geographic origins. The colors indicate different countries. On the right, the number of genomes of complete pathogens, distributed over a period of 4 months from the beginning of January to the end of April.

the viral proteins that physically associate with human proteins using affinity-purification mass spectrometry (AP-MS). We downloaded these PPI from NDEx (https://*public.ndexbio.org/network*/43803262 − 6*d*69 − 11*ea* − *bfdc* − 0*ac*135*e*8*bacf*).

To detect communities of PPI, we used the app Molecular Complex Detection (MCODE) [20] in Cytoscape. In a nutshell, MCODE iteratively groups together neighboring nodes with similar values of the core-clustering coefficient, which for each node is defined as the density of the highest k-core of its immediate neighborhood times k.[1] MCODE detects the densest regions of the network and assigns to each detected community a score that is its internal link density times the number of nodes belonging to it.

We also characterized the first ten communities c with the mean value $\bar{x}_c$ and standard deviation σ_c of codon bias values within the community, and use them to compute a Z-score as $Z_c = (\bar{x}_c - \bar{x}_n)/\sqrt{\sigma_c^2 + \sigma_n^2}$ (where $\bar{x}_n$ and σ_n are, respectively, the mean value and standard deviation of codon bias values computed for all proteins). In this way, a value of $Z_c > 1$ ($Z_c < -1$) indicates that community c features significantly higher (lower) codon bias than the population mean. Cytoscape was used to detect the degree k of a protein.

3 Results

3.1 SARS-CoV-2 genome diversity analyses

The records of viruses according to geographical location and date of isolation are shown in Figure 1. A great percentage of the 15,259 annotated SARS-CoV-2 genomes (about 50%) are distributed in England and USA. The number of complete virus genomes is quite constant except for the period from 1st of March to 1st of April, when the number rapidly increases.

[1]The density of a graph G with n nodes and l links is the ratio between l and the maximum number of possible links, namely $n(n-1)/2$, whereas, a k-core is a graph G of minimal degree k, meaning that each node belonging to G has degree greater or equal than k.

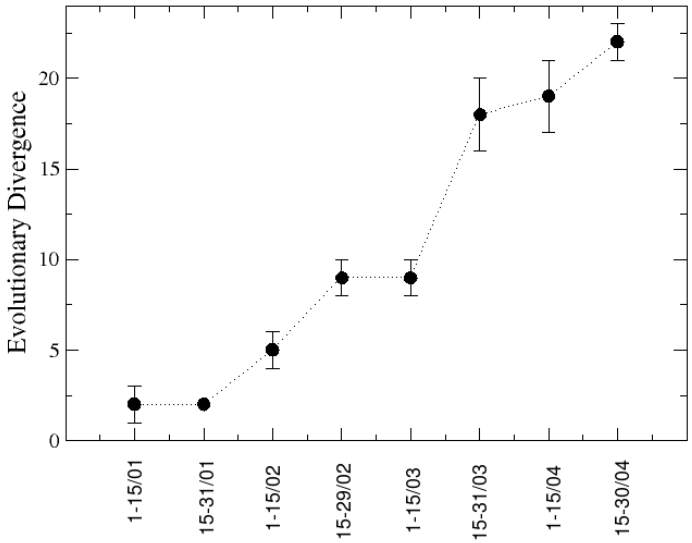

Figure 2. Evolutionary divergence of the SARS-CoV-2 genomes from the reference SARS-CoV-2 sequence.

3.2 SARS-CoV-2 Genomes divergence

The evolutionary divergence from the reference sequence was calculated for each of the 13 genes in all genomes under study. As shown in Figure 2 evolutionary divergence, in terms of number of substitutions per site, increases over time. Regarding the evolutionary trajectory of each gene (Figure 3), there is a statistically significant variation of evolutionary divergence is only for gene N and S. For the other genes most of 85% of viruses present identical sequence with the reference one. This observation is in line with our previous observation that genes encoding nucleocapsid (N) and spike proteins (S) tend to evolve faster in comparison to the two genes encoding the integral membrane proteins M and E [12].

3.3 Correlation between codon usage bias and time

To measure the codon usage bias in the SARS-CoV-2 genomes, we used the effective number of codons (ENC) and the Competition adaptation index (CAI). For the funtionally important genes in each genome, we calculated the average values of CAI and ENC over time, as compared to the reference SARS-CoV-2 sequence (WSM). To visually illustrate the differences among different time periods, the average ENC and CAI values of the coronavirus were depicted using a heatmap (see Figures 4,5). The 13 different genes of the coronavirus show different patterns of codon usage. All the genes have ENC and CAI values that differ significantly from the corresponding values of reference sequences (|Z-score| >2). Specifically, the ENC values associated with E, M and N are significantly higher than the corresponding one in the reference sequences, indicating that these genes use a broader set of synonymous codons in their coding sequences. This observation implies that these genes maintained high genetic variability in the context of synonymous codon usage that might render some advantages for those genes to express under diverse cellular and eviromental conditions. S, Nsp7, Nsp8, ExoN, RBP, OMT and 3CLpro have higher values of ENC, compared to the

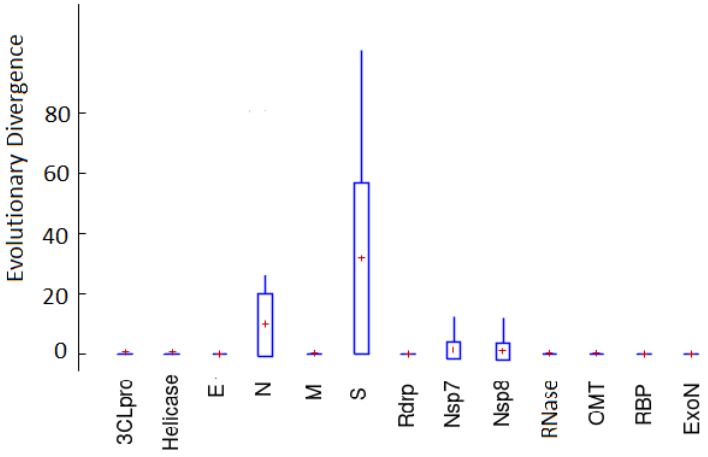

Figure 3. Box plots of evolutionary divegence for each gene under study from the reference sequence.

reference sequence, that decrease significantly over time. It suggests a tendency for these genes to otpimize the choice of codons over time to the respective host. Helicase has an intermediate value of ENC. The ENC value of protein RdRP increases only in the last two weeks, whereas RNase exhibits constantly low ENC values. Noteworthy, the CAI of all genes is markedly higher than the reference sequence 5, underscoring that these genes use codons that are better adapted to the human hosts. In particular, proteins N, ExoN and S have the highest values of CAI, suggesting that these genes accumulate preferential mutations to adapt better to the host.

3.4 SiD

In line with our previous study [12], we calculated the similarity index (SiD) of SARS-CoV-2 genomes with respect human and other hosts species shown in Table 1 by Woo et al. [4]. To understand the rationale behind these results, the higher the value of SiD, the more adapted the codon usage of SARS-CoV-2 to the host under study [13]. In our precedent analysis, SARS-CoV-2 exhibited high SiD values for human (0.78), snakes (0.75), as well as pangolins (SiD = 0.76), bats (SiD =0.70), and rats (SiD = 0.71), which have been suggested to be possible hosts for SARS-CoV-2 [21]. In this analysis, we found that the average value of SiD is 0.81 ± 0.2 (Figure 6). This value in human was increased compared to our previous analysis. In contrast, the values of SiD for other hosts are not statistically significant. Based on the SiD combined with the CAI results (Figure 5), we suggest that SARS-CoV-2, over time, has preferentially accumulated mutations in its genome which correspond to codons that adapt better to the human host.

3.5 ENC Plot Analysis of SARS-CoV-2 genes

To further investigate the evolutionary forces that affect the SARS-CoV-2 codon usage, an ENC-plot analysis was conducted separately for each of the 13 genes considered herein. We then investigated variations over time by performing ENC plots for sets of genes binned by the time. In these plots, each point represents a single gene retrieved from each genome. To show clearer the temporal difference, we calculated the average values of the distance d from the Wright theoretical

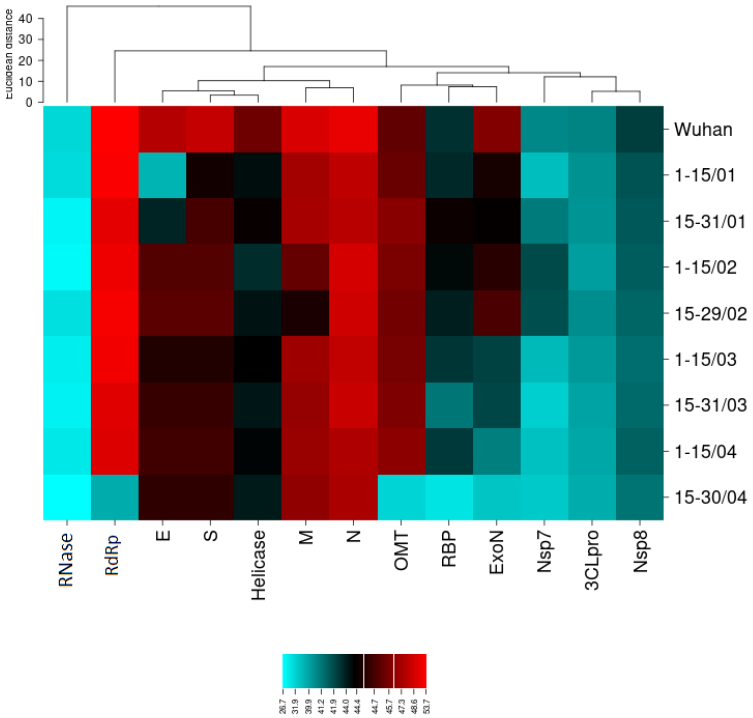

Figure 4. Heatmap of ENC values in SARS-CoV-2 genes. Heatmap was generated with the CIMminer software [14], using Euclidean distances and the average linkage algorithm.

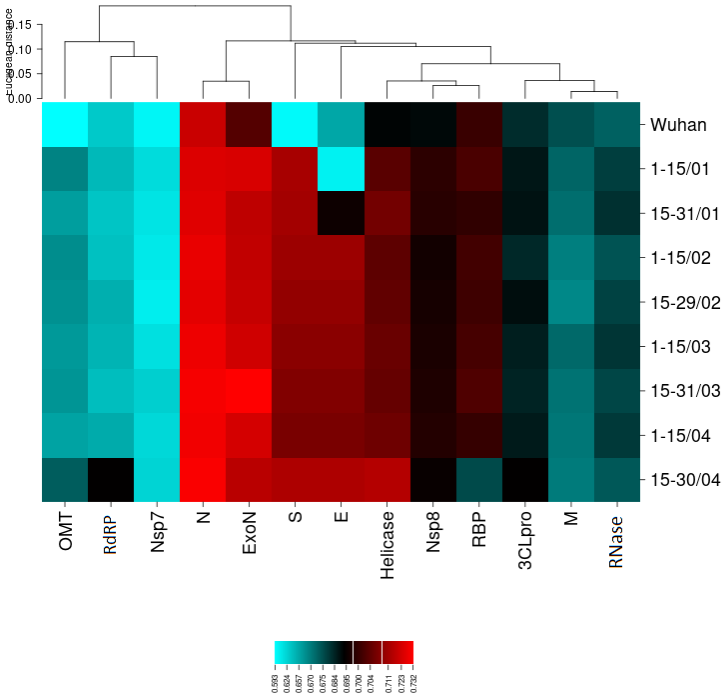

Figure 5. Heatmap of CAI values in SARS-CoV-2 genes. The conventions are the same as in Figure 4.

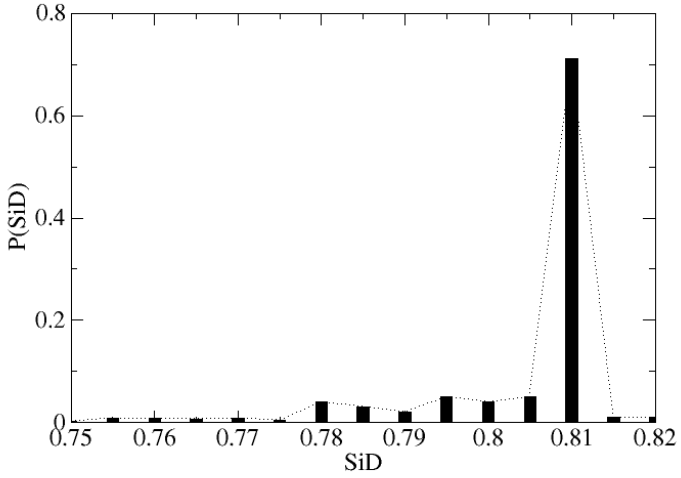

Figure 6. Distribution of frequency of SiD for human. Average value of SiD is 0.81 ± 0.02.

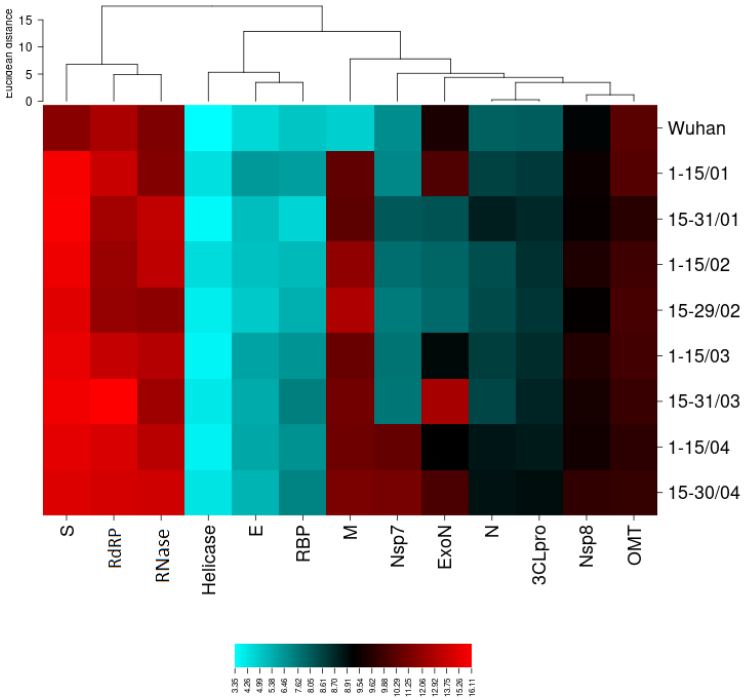

Figure 7. Heatmap of distances from Wright's theoretical curve over time. The conventions are the same as in Figure 4.

curve, and represented them graphically through a heatmap in Figure 7. As a genaral rule, the average distances from Wright theoretical curve increases over time for all the genes, except for $ExoN$, $Nsp8$, and OMT, which appear to be stable. This means that the vast majority of the genes tend to be under a stronger action of natural selection over time. Moreover, we note that genes encoding for Helicase, E, and RBP have a shorter distance from the Wright theoretical curve as a funtion of the time of isolation, meaning that the codon usage of these genes tends to be ruled by a mutational bias. Looking at the numerical values associated with genes ecnding for S, $RdRP$, and $RNase$, we observed that these genes are more scattered, on average, below the theoretical curve, indicating that the codon usage of these genes is most one affected by natural selection. Conversely, the rest of the genes revealed larger deviations from the Wright theoretical curve, thus showing a strict control of natural selection on their codon usage.

3.6 Codon Bias and the Connectivity Patterns of SARS-CoV-2 Protein Interaction Network

As far as the network of interacting proteins in SARS-CoV-2 is concerned, we first investigated codon usage bias in relation with the connectivity patterns of the network. The degree distribution of the network suggests that it is scale-free (Figure. 8), meaning that the network contains a large number of poorly connected proteins and a relatively small number of highly connected proteins or 'hubs'. The corresponding genes of these hub proteins have consistently higher values of codon usage bias when this is measured by CAI and lower when this is measured by ENC. Of note, the two codon bias indices are anticorrelated 9.

Moreover, we examined codon bias in relation with the community structure of the PPI, where a community is a group of proteins that are more densely connected within each other than with the rest of the network. Table 1 shows the

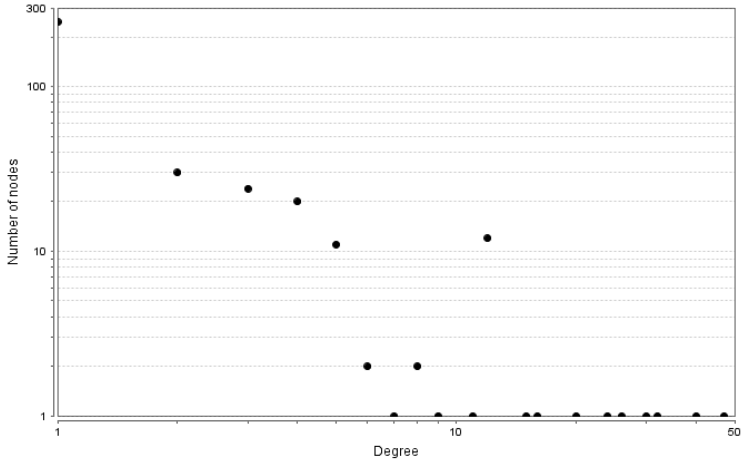

Figure 8. Degree distribution of proteins P(k). The degree distribution of the network follows a power law, indicating that the network is scale-free.

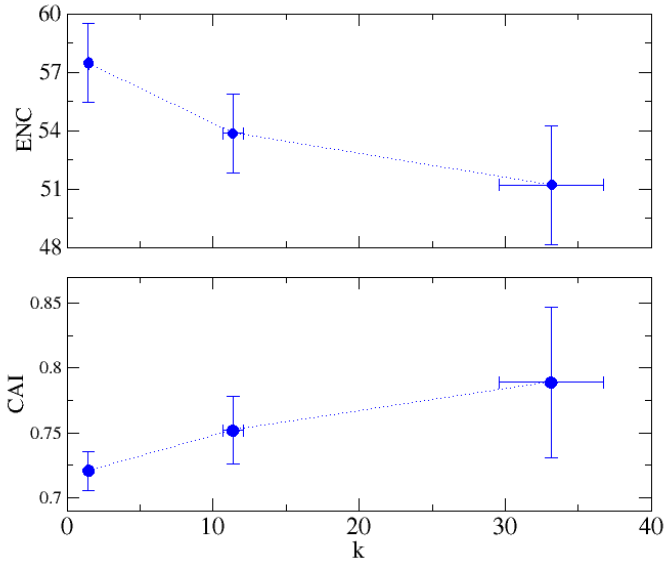

Figure 9. Correlation between codon bias indices of genes and the degree k of the corresponding proteins in the PPI. CAI of a gene consistently increases with the connectivity of the corresponding protein in the PPI, whereas ENC decreases.

features of the first ten communities together with their average degree and their average of the two codon bias indices (CAI and ENC). We also calculate the internal average value $\bar{x}_c$ and the Z-scores, comparing the distribution of bias inside the community with all the proteins. We note that all ten communities superate the Z-score test (Z>2). Regarding the first community, which includes only 13 proteins (3.9% of the whole network) that basically overlaps with the main core of the PPI (*i.e.*, the k-core with the highest possible degree). Notably, proteins belonging to this community have on average a codon bias index (as measured by CAI and ENC) that is significantly higher than the average of the rest of the network (the Z-score > 1).

Table 1. Features of ten top-scoring communities. Number of nodes (n), community score (n times the internal density), mean degree, average codon bias indices (CAI and ENC).

ID	n	score	k	CAI	ENC
1	13	13,00	14,15	0,77	51,99
2	6	6,00	5,67	0,69	55,88
3	6	6,00	6,83	0,71	53,24
4	5	5,00	4,80	0,70	53,70
5	5	5,00	4,40	0,70	53,05
6	4	4,00	4,00	0,71	53,11
7	4	4,00	4,00	0,70	53,12
8	4	4,00	3,25	0,72	53,12
9	4	4,00	6,25	0,70	53,10
10	4	3,33	3,50	0,72	53,13

4 Discussion

In this study, we performed a comprehensive analysis of the evolutionary divergence and codon usage of SARS-CoV-2 over time, considering all genomes available in GISAID on May 09, 2020. After filtering out incomplete genomes, we retained a total of 15,259 complete genomes, with the purpose of investigating the divergence of these viral genomes from the first sequenced SARS-CoV-2 genome (NC_045512.2). We focused on 13 SARS-CoV-2 genes/proteins that are crucial for its structure, synthesis, transmissibility and virulence.

The SARS-CoV-2 genomes have a tendency to diverge constantly from the reference genome. This is in accordance with a recent study by Pachetti and colleagues (2020) where they have demonstrated that the number of SARS-CoV-2 mutations change over time [2]. This trend is more pronounced in the last two weeks of March, where the genome divergence is more rapid. Of note, the sequences of the genes encoding the structural nucleocapsid (N) and spike (S) proteins, vary significantly from the reference sequences, as well as the non-structural proteins Nsp7 and Nsp8. In support of that, the ENC analysis revealed that the codon usage patterns of all 13 genes under study differentiate over time from the reference sequences. This means that a percentage mutations on the SARS-CoV-2 genome are synonymous and, therefore, alter the patterns of codon usage. The CAI values of all genes, especially S, N and $ExoN$, have increased significantly as compared to the reference. The value of the SiD estimated from the average codon usage of the SARS-CoV-2 genomes against the codon usage of the human host was higher (0.81 ± 0.2) as compared to the reference one (0.78). It is suggested that the coronaviral genome has undergone genetic recombinations and beneficial nucleotide changes, which have likely contributed to its enhanced adaptability to the human host. In other words, SARS-CoV-2 might use the human translational machinery most effectively than

that of other animals. This could explain, at least partially, the global distribution and increasing prevalence of SARS-CoV-2 [22, 23]. The ENC-plot analysis revealed that the codon usage of the viral genes here considered are subject to different balances between mutational bias and natural selection. For instance, the codon preference of the genes S, $RdRp$, $RNase$, N, and $Nsp8$, are mainly determined by natural selection, as opposed to the genes M and E, the codon usage of which is rather affected by mutation bias. These results are similar to ones of our previous study [12], showing that the codon usage of the genes N, S and $RdRp$ was found to be under stronger selection than genes encoding for proteins M and E. On the basis of our findings, the SARS-CoV-2 genes S and N consistently displayed higher genetic diversity, increased host adaptation, and more proneness to natural selection. According to Rehman et al. (2020), the spike protein, which mediates the virus interaction with the human host cells, is more prone to mutations and particularly those occurring in the amino acids implicated in the spike-angiotensin-converting enzyme 2 (ACE2) interface [22]. The N protein, responsible for virus assembly and RNA transcription [24], is considered the most conserved and stable coronaviral structural protein [25]. It is tempting to speculate that N gene accumulates mutations that are do not affect its structure and function, but rather enable it to evade the host's immune responses and enhance SARS-CoV-2's pathogenicity. Therefore, these key genes, N and S, accumulate beneficial mutations that would increase the evolvability and transmissibility of SARS-CoV-2, and enable it to continuously adapt to different populations, like spilling over from bats, or other candidate natural reservoirs, to human.

RdRp, which catalyzes the transcription and replication of the coronaviral genome, binds to Nsp7 and Nsp8 to form the core RNA synthesis machinery of SARS-CoV-2 [26]. Accordingly, the genes coding for RdRp and its co-factor Nsp8 appear to be subject to natural selection, suggesting that they have an increased ability to adapt into novel hosts. Despite the fact that RdRp is considered less vulnerable to mutations [2] due to its vital role in maintaining viral genome fidelity, the mutations that occur in RdRp likely promote the virus adaptive flexibility and enhance its resistance to antiviral drugs [27]. The exoribonuclease ExoN, which ensures viral replication fidelity by correcting nucleotides incorrectly incorporated by RdRP [28], has a relatively high CAI value, indicating greater adaptability to the human host.

Furthermore, we found that in the human-SARS-CoV-2 network, the most highly connected or hub proteins have consistently higher codon usage bias relatively to the less connected proteins. This observation leads to the suggestion that evolutionary pressure is exerted upon the genes encoding those proteins,most probably because of their great biological significance for the virus. In other words, if these nodes are removed the entire network will eventually collapse [29].

References

1. Laamarti M et al. Large scale genomic analysis 1 of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations, https://doi.org/10.1101/2020.05.03.074567
2. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18(1):179. Published 2020 Apr 22. doi:10.1186/s12967-020-02344-6
3. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020. DOI: 10.1093/nsr/nwaa036
4. Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster, Phylogenetic network analysis of SARS-CoV-2 genomes, PNAS, https://doi.org/10.1073/pnas.2004999117

5. Mercatelli, Giorgi. Geographic and Genomic Distribution of SARS-CoV-2 mutations, doi:10.20944/preprints202004.0529.v1

6. N. Zhu, D. Zhang, W. Wang, et al.A novel coronavirus from patients with pneumonia in China, 2019, N Engl J Med, 382 (2020), pp. 727-733

7. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. Genome Research 1998, 8: 967–974.

8. Wright, F. The 'effective number of codons' used in a gene. *Gene* **1990**, *87*, 23–29

9. Xia, X. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **2013**, *30*, 1720–1728, doi:10.1093/molbev/mst064.

10. Sharp, P.M.; Wen-Hsiung, L. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **1987**, *15*, 3.

11. Sharp, P.M.; Wen-Hsiung, L. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **1986**, *24*, 28–38.

12. Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A and Pavlopoulou A, Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes, MDPI

13. Lia, G.; Wang, H.; Wanga, S.; Xinga, G.; Zhanga, C.; Zhanga, W. Insights into the genetic and host adaptability of emerging porcine circovirus. *Virulence* **2018**, *9*, 1301–1313, doi:10.1080/21505594.2018.1492863.

14. Weinstein, J.N.; Myers, T.G.; O'Connor, P.M.; Friend, S.H.; Fornace, A.J., Jr.; Kohn, K.W.; Fojo, T,; Bates, S.E.; Rubinstein, L.V.; Anderson, N.L.; et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343–349.

15. Edgar RC, MUSCLE: Multiple Sequence Alignment With High Accuracy and High Throughput, Nucleic Acids Res. 2004 Mar 19;32(5):1792-7. doi: 10.1093/nar/gkh340. Print 2004.

16. Edgar RC, MUSCLE: A Multiple Sequence Alignment Method With Reduced Time and Space Complexity, 2004 Aug 19;5:113. doi: 10.1186/1471-2105-5-113.

17. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-1549. doi:10.1093/molbev/msy096

18. Jia Y, Shen G, Zhang Y, Huang KS, Ying Ho H, Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of 2 RBD mutant with lower ACE2 binding affinity, https://doi.org/10.1101/2020.04.09.034942

19. Gordon, D. E. et al.A SARS-CoV-2 protein interaction mapreveals targets for drug repurposing. Nature https://doi.org/10.1038/s41586-020-2286-9 (2020).

20. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2. doi:10.1186/1471-2105-4-2

21. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; Holmes, E.; Garry, R.F. The proximal origin of SARS-CoV-2. *Nat. Med.* **2020**, doi:10.1038/s41591-020-0820-9.

22. Rehman SU, Shafique L, Ihsan A, Liu Q. Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. Pathogens. 2020;9(3):240. Published 2020 Mar 23. doi:10.3390/pathogens9030240

23. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181-192. doi:10.1038/s41579-018-0118-9

24. Timani KA, Ye L, Ye L, Zhu Y, Wu Z, Gong Z. Cloning, sequencing, expression, and purification of SARS-associated coronavirus nucleocapsid protein for serodiagnosis of SARS. J Clin Virol. 2004;30(4):309-312. doi:10.1016/j.jcv.2004.01.001

25. Sheikh A, Al-Taher A, Al-Nazawi M, Al-Mubarak AI, Kandeel M. Analysis of preferred codon usage in the coronavirus N genes and their implications for genome evolution and vaccine design. J Virol Methods. 2020;277:113806. doi:10.1016/j.jviromet.2019.113806

26. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat Commun. 2019;10(1):2342. Published 2019 May 28. doi:10.1038/s41467-019-10280-3

27. Pfeiffer JK, Kirkegaard K. A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. Proc Natl Acad Sci U S A. 2003;100(12):7289-7294. doi:10.1073/pnas.1232294100

28. Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. The Curious Case of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity. Front Microbiol. 2019;10:1813. Published 2019 Aug 7. doi:10.3389/fmicb.2019.01813

29. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101-113. doi:10.1038/nrg1272