Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(1): 013
http://doi.org/10.31083/j.fbl2701013

Original Research

Temporal evolution and adaptation of SARS-CoV-2 codon usage

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

[1]Physics Department, Sapienza University of Rome, 00185 Rome, Italy
[2]Liceo Statale Maria Montessori, 00198 Rome, Italy
[3]Izmir Biomedicine and Genome Center (IBG), 35340 Izmir, Turkey
[4]Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, 35340 Balcova, Izmir, Turkey
[5]DNA Damage Laboratory, Physics Department, School of Applied Mathematical and Physical Sciences, National Technical University of Athens (NTUA), 15780 Athens, Greece
[6]INFN Roma1 Unit, 00185 Rome, Italy
*Correspondence: maddalena.dilucca@gmail.com (Maddalena Dilucca)
Academic Editor: Graham Pawelec
Submitted: 31 July 2021 Revised: 20 September 2021 Accepted: 12 November 2021 Published: 11 January 2022

Abstract

Background: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) first occurred in Wuhan (China) in December of 2019. Since the outbreak, it has accumulated mutations on its coding sequences to optimize its adaptation to the human host. The identification of its genetic variants has become crucial in tracking and evaluating their spread across the globe. **Methods**: In this study, we compared 320,338 SARS-CoV-2 genomes isolated from all over the world to the first sequenced genome in Wuhan, China. To this end, we analysed over time the codon usage patterns of SARS-CoV-2 genes encoding for the membrane protein (M), envelope (E), spike surface glycoprotein (S), nucleoprotein (N), RNA-dependent RNA polymerase (RdRp) and ORF1ab. **Results**: We found that genes coding for the proteins N and S diverged more rapidly since the outbreak by accumulating mutations. Interestingly, all genes show a deoptimization of their codon usage with respect to the human host. Our findings suggest a general evolutionary trend of SARS-CoV-2, which evolves towards a sub-optimal codon usage bias to favour the host survival and its spread. Furthermore, we found that S protein and RdRp are more subject to an increasing purifying pressure over time, which implies that these proteins will reach a lower tendency to accept mutations. In contrast, proteins N and M tend to evolve more under the action of mutational bias, thus exploring a large region of their sequence space. **Conclusions**: Overall, our study shed more light on the evolution of SARS-CoV-2 genes and their adaptation to humans, helping to foresee their mutation patterns and the emergence of new variants.

Keywords: SARS-CoV-2; Codon usage bias; Viral adaptation; Forsdyke plot; Similarity index; ENC plot

1. Introduction

The emergence of the 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.

SARS-CoV-2 belongs to the Coronaviridae family, with a genome of approximately 30 kb in length and a structure characteristic of known coronaviruses. The virus genome encodes for structural, non-structural, and accessory proteins [1]. The four structural proteins are the *envelope* protein (E), the *membrane* protein (M), the *nucleocapsid* protein (N), and the *spike* glycoprotein (S). The overlapping open reading frames ORF1a and ORF1ab encode for the two isoforms of the polyprotein complex pp1a and pp1ab, respectively, where the latter is further proteolytically cleaved into 16 different non-structural proteins (NSP) [2–4].

The *nucleocapsid* protein plays an important role in maintaining the RNA conformation stable for replication, transcription, and translation of the viral genome, as well as protecting it. It is highly immunogenic and capable of modulating the metabolism of an infected cell [5]. The *envelope* protein acts as a *viroporin* [6] and plays multiple roles in viral replication and signalling pathways that affect inflammatory and type 1 INF-γ signalling [7].

The *spike* protein S is responsible for receptor recognition and membrane fusion, that leads to viral entry into the host cells [8]. Finally, the membrane protein is associated with the *spike* protein and is responsible for the virus budding process [9].

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. The mutation rate of RNA viruses (10^{-5}–10^{-3} per replication cycle [10]) is markedly higher than that of their mammalian hosts (10^{-9} per year [11]). This high mutation rate is correlated with virulence modulation and evolvability, which are considered beneficial for viral adaptation [12].

Table 1. Length of SARS-CoV-2 genes. For each gene, the length of its encoded protein, in amino acids, and the location inside the reference genome Wuhan-Hu-1 (Accession Number Y P_009724390) are reported. Data from GenBank [16].

Gene	Slope	x-Intercept	R-squared
M	0.2951 ± 0.0017	–0.0027 ± 0.0017	0.086
N	1.9015 ± 0.0012	–0.0657 ± 0.0013	0.904
S	2.1984 ± 0.0023	–0.0071 ± 0.0023	0.784
RdRp	0.5477 ± 0.0022	–0.1692 ± 0.0022	0.161
ORF1a	1.648 ± 0.004	0.012 ± 0.004	0.525
ORF1b	1.0354 ± 0.0025	–0.0265 ± 0.0025	0.369

In this study, we investigate the evolution of SARS-CoV-2 genomes and codon usage patterns over time, by analysing the virus adaptation to the human host. The main hypothesis is that the codon usage should be more optimized with respect to the host's one [13], to maximize the efficiency of translation. This is linked to the exploitation of the host's resources, such as tRNA, which is tightly related to codon usage bias [14,15].

For this purpose, we focused on the four SARS-CoV-2 genes encoding for the structural proteins *membrane* (M), *envelope* (E), *spike* surface glycoprotein (S) and *nucleoprotein* (N), as well as *RNA-dependent RNA polymerase* (RdRp). We also decided to consider the polyprotein complex ORF1ab taking the two segments ORF1a and ORF1b separately as they code for many different proteins. Of note, RdRp is encoded by a segment of the ORF1ab, but it has been isolated due to its key role in viral replication. In Table 1 (Ref. [16]) can be found the lengths of the considered genes, along with their position in the viral genome.

2. Materials and methods

2.1 Sequence data analysed

All available SARS-CoV-2 genomes reported across the world were obtained from GISAID (available at https://www.gisaid.org/epiflu-applications/nexthcov-19-app/), on February 7, 2021. Then, the sequences were classified according to their isolation dates. Only complete genomes (29–30 Kb) were included in the present analysis, with a list of 320,338 SARS-CoV-2 genomes. We used the SARS-CoV-2 coding DNA sequences (CDSs) deposited in January 2020 by Zhu and co-workers [17], formerly called "Wuhan seafood market pneumonia virus", also referred to as "Wuhan-Hu-1" (WSM, NC 045512.2), as reference sequence. 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 genomes under study, by using the Exonerate tool (v 2.2.0, https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) with default parameters [18]. A threshold in length of 95% the total reference gene length has been used to filter out defective sequences.

2.2 Relative synonymous codon usage

Relative Synonymous Codon Usage (RSCU) is the fraction of the observed codon frequency to expected codon frequency, given that all the codons corresponding to a precise amino acid are used equally. RSCU value for each gene is calculated using Eqn. 1 for each codon i, as described by Sharp and Li [19]:

$$RSCU_i = \frac{X_i}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_j} \tag{1}$$

where X_i is the observed number of the i-th codon and n_i the total number of synonymous codons for the corresponding amino acid. Stop codons and non-degenerate codons (methionine (M) and tryptophan (W)) are not considered, so RSCU is a 59-component vector. The values go from 0 to the degeneracy of the codon. Three different types of RSCU values are obtained: (i) codons with RSCU <1 are less frequently used and have a negative usage bias, (ii) codons with RSCU >1 are frequently used and have positive usage bias and (iii) codons with RSCU equal to 1 have no bias.

2.3 Codon adaptation index and similarity index

The two indices Codon Adaptation Index (CAI) and Similarity Index (SiD) are both used to quantify the extent of codon usage adaptation of SARS-CoV-2 to the human coding sequences.

The principle behind CAI [20] is that the codon usage in highly expressed genes can reveal the optimal (i.e., most efficient for translation) codons for the translation of 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 ($RSCU_i$), defined in the previous paragraph, and the relative codon adaptiveness (w_i), as shown in Eqn. 2:

$$w_i = \frac{\mathrm{RSCU}_i}{\sum_{j=1,2\ldots n_i} \max\{RSCU_j\}} \tag{2}$$

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 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, as shown in Eqn. 3.

$$CAI_g = \left(\prod_{i=1}^{l_g} w_i\right)^{1/l_g} \tag{3}$$

where the product runs over the l_g codons belonging to that gene (except for the stop codons and the non-degenerate ones (W and M).

On the other hand, SiD is the cosine of the solid angle between the two RSCU vectors of the virus and the host, without taking into account the expression rate of the genes.

Formally, SiD is defined in Eqn. 4 [21]:

$$SiD = \frac{\sum_{k=1}^{59} a_i \cdot b_i}{\sqrt{\sum_{k=1}^{59} a_i^2 \cdot \sum_{k=1}^{59} b_i^2}} \quad (4)$$

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 corresponding codons of the potential host. Therefore, SiD quantifies the degree of similarity between the virus and the host in terms of their codon usage patterns.

Both indices values range from 0 to 1; the higher the value, the more adapted the codon usage of SARS-CoV-2 to the host. They are fundamental estimators for this research, as they are useful to quantify the codon usage bias viral adaptation to the host's one.

The reference set of highly used genes used for CAI has been retrieved from DAMBE 5.0 [22], while the codon frequencies used to estimate human RSCU, from CUBAP [23].

2.4 Principal component analysis

Principal Component Analysis (PCA) [24] is a multivariate statistical method to transform a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (called principal components), spanning a space of lower dimensionality. The transformation is defined so that the first principal component accounts for the largest possible variance of the data, and each subsequent component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components.

We use this technique on the space of RSCU vectors, so that each gene of SARS-CoV-2 is represented as a 59-dimensional vector with codons as coordinates. Such coordinates are separately normalized to zero mean and unit variance over the whole genome. We then obtain the associated covariance matrix between the dimensions of codon bias and diagonalize it. The eigenvectors of the covariance matrix, ordered according to the magnitude of the corresponding eigenvalues, are the principal components of the original data. PCA has been performed by using Python scikit-learn (v 0.20.3)(https://pypi.org/project/scikit-learn/0.20.3/).

2.5 ENC plot

ENC is used to evaluate the degree of bias of a gene or genome, in respect to codon usage. It quantifies the difference from a random usage of all codons. Its value ranges from 23, when there is a strong bias and only one codon per amino acid family is used, to 61, when all codons are equally used. 6 codon-families have been considered as distinct subfamilies of 2 and 4 codons, as explained in Sun *et al.* (2013) [25].

An ENC plot analysis was performed to estimate the relative contributions of mutational bias and natural selection in shaping the CUB of the SARS-CoV-2 genes. In this plot, the ENC values are plotted against GC3 values. If codon usage is dominated by mutational bias, then a clear relationship is expected between ENC and GC3, given by the formula in Eqn. 5,

$$ENC = 2 + s + \frac{29}{s^2 + (1-s)^2} \quad (5)$$

where s represents the value of GC3 [26]. If the mutational bias is the main force affecting the CUB of the genes, the corresponding points are expected to fall near Wright's theoretical curve. Conversely, if CUB is mainly affected by natural selection, the corresponding points should fall considerably below Wright's theoretical curve [27]. To quantify the relative extent of the natural selection, for each gene, we calculated the Euclidean distance **d** of the corresponding point, from the curve. We then showed the average values of the distance over time with a heatmap.

2.6 Forsdyke plot

To study the mutational rates of genes under study, we performed an analysis by using Forsdyke plots [28]. Each gene of the SARS-CoV-2 sequences under investigation was compared to its orthologous gene in the reference Wuhan-Hu-1 genome. Each pair of orthologous genes is represented by a point in the Forsdyke plot, wherein protein divergence is correlated with RNA divergence. The protein sequences were aligned using Biopython. The RNA sequences were then aligned using the protein alignments as templates. Then, both RNA and protein divergences were assessed, as explained in Methods in [28], by counting the number of mismatches in each pair of aligned sequences. Thus, each point in the Forsdyke plot measures the divergence between pairs of orthologous genes in the two viruses, as projected along with the phenotypic (protein) and nucleotidic (RNA) axis. The first step in each comparison is to compute the regression line between protein versus RNA sequence divergence in the Forsdyke plot for getting values of intercept and slope for each variant of genes.

3. Results

3.1 Dataset distribution

The records of virus genomes according to geographical location and month of isolation are shown in Fig. 1. A great percentage of the annotated SARS-CoV-2 genomes (about 62% and 22% respectively) are distributed in Eu-

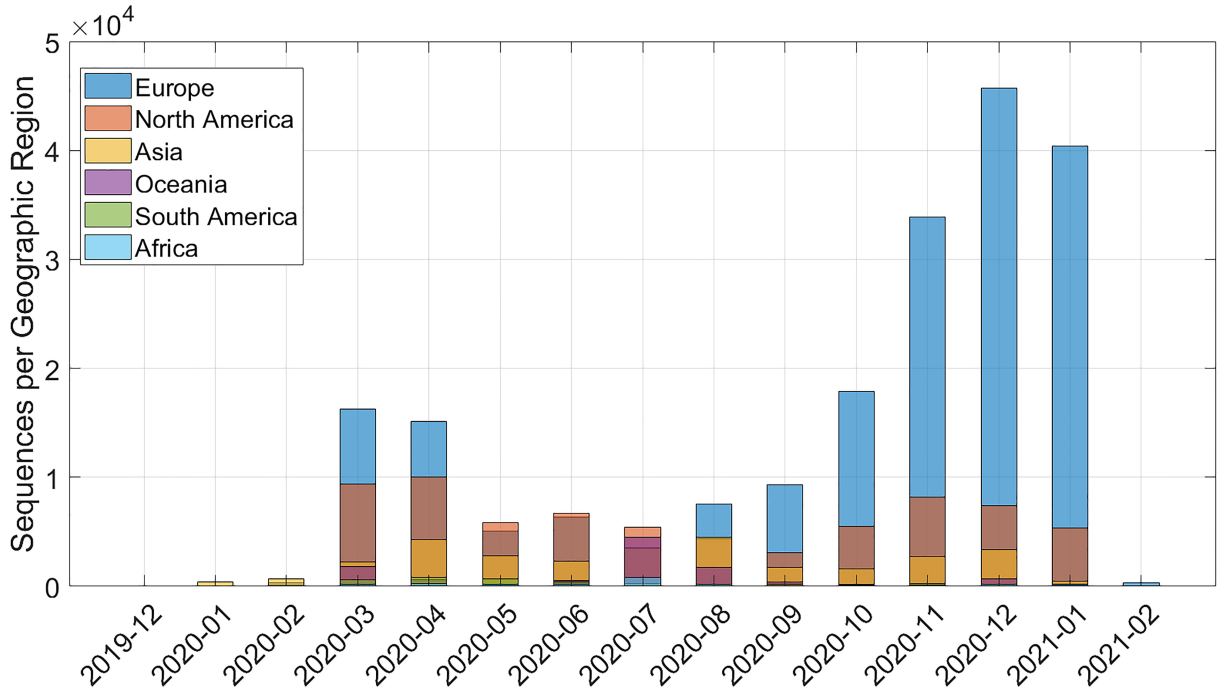

Fig. 1. Distribution of the 320,338 genomes used in this study. Distribution of available genome sequences, subdivided per month and geographical region.

rope and North America. The number of complete viral genomes is non-homogeneous, ranging from 200–300 in the first months, reaching a peak of around 50,000 in December 2020. In the other months, the values fluctuate between 7000 and 40,000.

3.2 Principal component analysis of RSCUs

We performed PCA over the space of RSCU vectors measured for each SARS-CoV-2 gene in all sequences. The two first principal components (PC1 and PC2) appear to represent as much as 78% of the total variance of RSCUs (respectively, 52% for PC1 and 26% for PC2). The projections of the first two principal components on the individual codons show that none of the codons predominantly contributes to the data variability (see Fig. 2). Thus, the placement of a gene on the PC1-PC2 plane depends on a weighted contribution of all codons.

Most codons highly affect PC1, whereas PC2 is mostly affected by triplets that code for phenylalanine (TTT and TTC), leucine (TTA, TTG and CTA), isoleucine (ATT, ATC), lysine (AAG) and valine (GTA).

The PCA plot of all considered genes is shown in Fig. 3. Interestingly, the genes coding for the polyproteins and S protein are very close, in the PCA plane at the bottom of the graph, far apart from the others. Regarding the N protein, the coordinates in PC1 is similar to those of the S-ORF1ab cluster, while is quite distant in PC2. The M protein is closer to N in PC2, denoting a similar usage of codons translating the five amino acids displayed above. Finally, the E protein is the mostly separated and dispersed cluster. The difference in RSCU of the protein E from the others and the dispersion in PCA is due to its shortness (see Table 1) for it does not allow a consistent repetition of amino acids and thus the use of different codons. Taken together, this analysis shows that the longer genes carry a signature RSCU of the virus, suggesting that SARS-CoV-2 genes tend to use the same set of codons. For the two proteins M and N, this set differs in the codons encoding the five aforementioned amino acids.

3.3 Codon adaptation index temporal evolution

To measure SARS-CoV-2 codon usage bias adaptation to the human host, we first evaluated CAI for all genes under investigation. For each gene, we performed a daily mean of the CAI values (see Fig. 4). In line with the PCA analysis, the CAI starting values are very similar for genes *S*, *ORF1a*, *ORF1b* and *RdRp*. The N protein has the highest CAI value, implying a larger adaptation of its codon usage to that of highly expressed genes in humans. Conversely, the M protein exhibits the lowest CAI value, indicating a less adapted codon usage with respect to the host.

The fastest evolving gene appears to be the one coding for the N protein with a globally descending trend. Similarly, the S protein displays the same trend but with a down-peak in the month of July 2020. In the timeline of sequence evolution shown in **Supplementary Fig. 1**, where the sequences are divided by region, that this down-peak is mostly due to Oceania sequences. Other geographical regions display a smoother decrease.

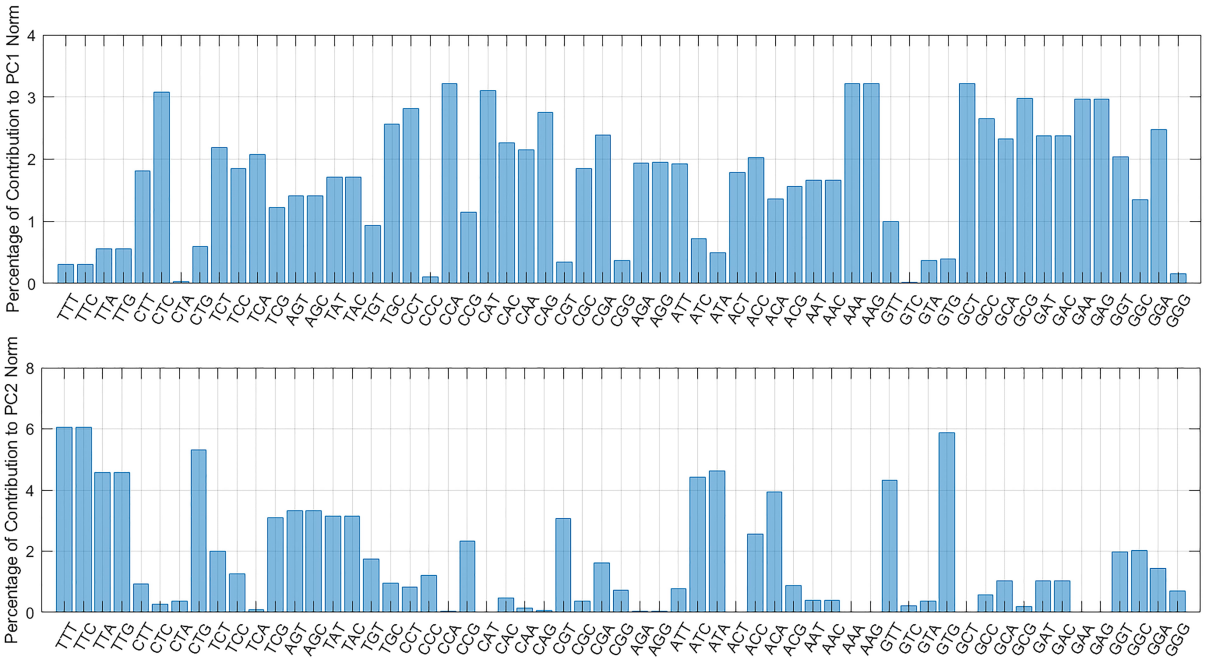

Fig. 2. Projection of the first two PCA components on the RSCU vectors. The distribution of codons in PC1 is more uniform, resulting in a weighted and coherent contribution of all codons, whereas in PC2 the contribution is more focused on specific codons encoding phenylalanine, leucine, isoleucine, lysine and valine.

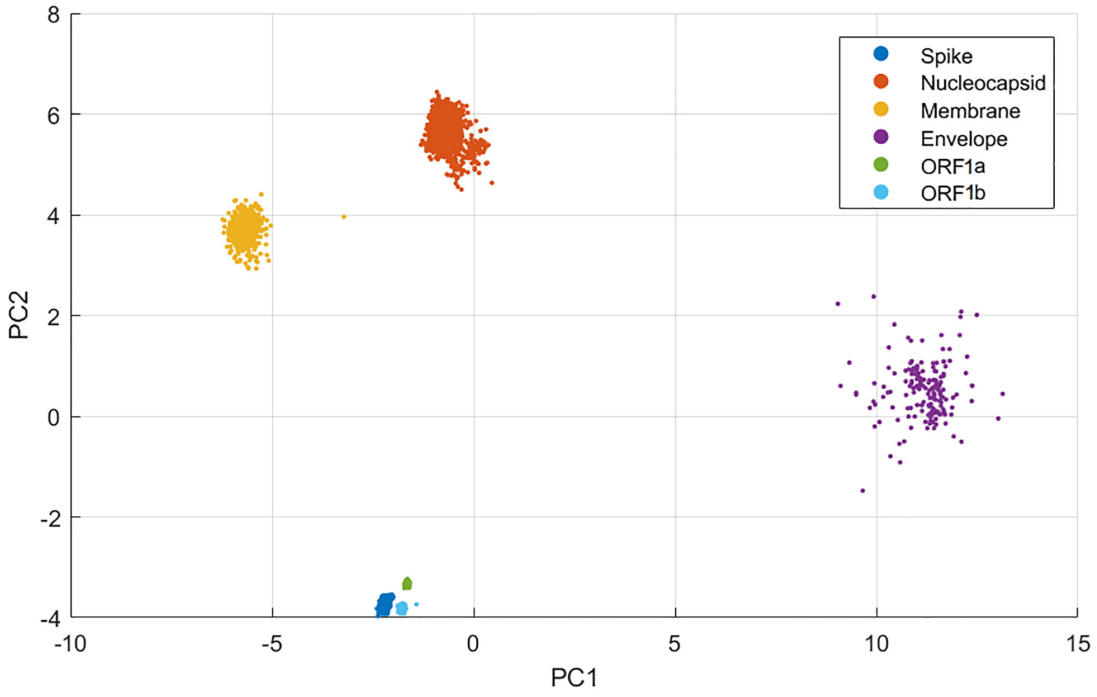

Fig. 3. PCA of SARS-CoV-2 Genes. All available sequences for each gene have been used in this study, by calculating the RSCUs and performing PCA for each of them. Approximately 320,000 sequences were analysed per gene.

The M protein evolves in a constant manner until August, where a steep ascent occurs. On the other hand, RdRp presents a sudden decrease in the first months and a subsequent stabilization on a lower CAI value. This observation suggests an initial adaptation of RdRp to the new host with the emergence of a new variant that rapidly superseded the

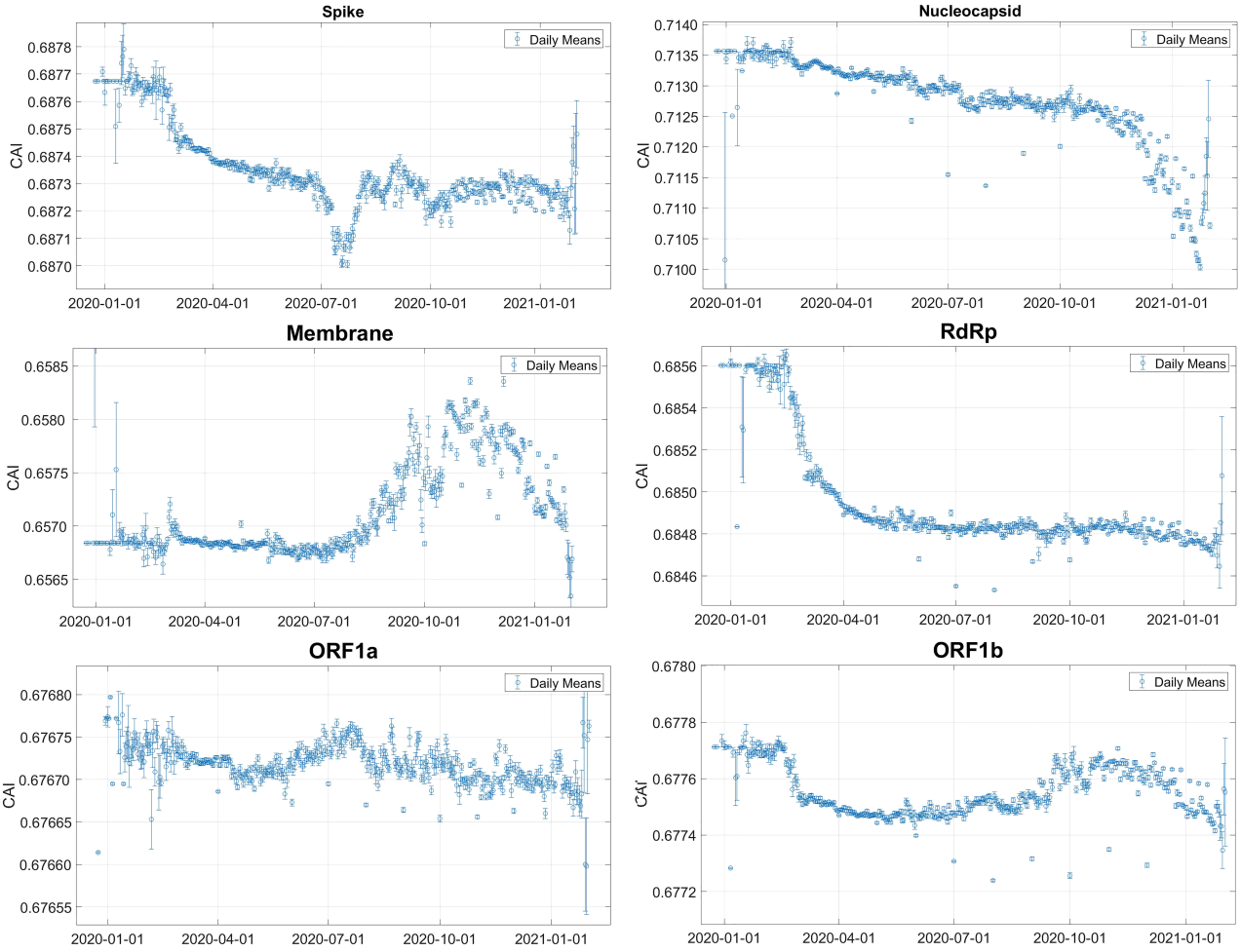

Fig. 4. CAI values daily averages in SARS-CoV-2 genes. Daily averaged CAI values estimated from all the available sequences and the standard errors on their assessments ($\sigma/\sqrt{N}$). CAI values span from 0, for an entirely opposite use of codons, to 1, for identical usage of codons to the host. Each panel refers to the gene indicated in the title.

older, less optimized one. Finally, the two CDSs of gene ORF1ab manifest a lower variability over time, in line with the fundamental role of these proteins in replication.

Overall, the SARS-CoV-2 genes display a descending trend over time, which is in contrast with the hypothesis that a virus evolves to optimize its codon usage to the one of the human host. Our observations suggest a general evolutionary trend of SARS-CoV-2 according to which it has been evolving towards a sub-optimal codon usage bias, which may favour the host survival and, therefore, its spread. Indeed, if the virus replicates efficiently, then it is more likely to be successful at invading a given host, eventually killing it. This finding is in accordance with the virulence trade-off hypothesis [29].

3.4 Similarity index time evolution

In line with our previous study [30], we calculated the SiD of SARS-CoV-2 genes with respect to the human host. 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 [31]. Specifically, we calculated here the SiD values for each gene under study, and performed a daily average of all the available sequences (see Fig. 5).

In accordance with the CAI temporal evolution (see Section 3.3), the N protein corresponds to the most adapted gene; in contrast, the gene encoding the S protein is the least adapted.

Also in this case, the RdRp, ORF1ab, and protein S, show similar SiD values. Moreover, the temporal trends are mostly descending, remarking the previous statements on CAI about a reduced optimization of SARS-CoV-2 codon usage to the human host over time. As observed in Section 3.3, the averaged SiD value for the S protein shows a down peak in July, which is mainly due to the Oceania sequences (**Supplementary Fig. 1**).

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

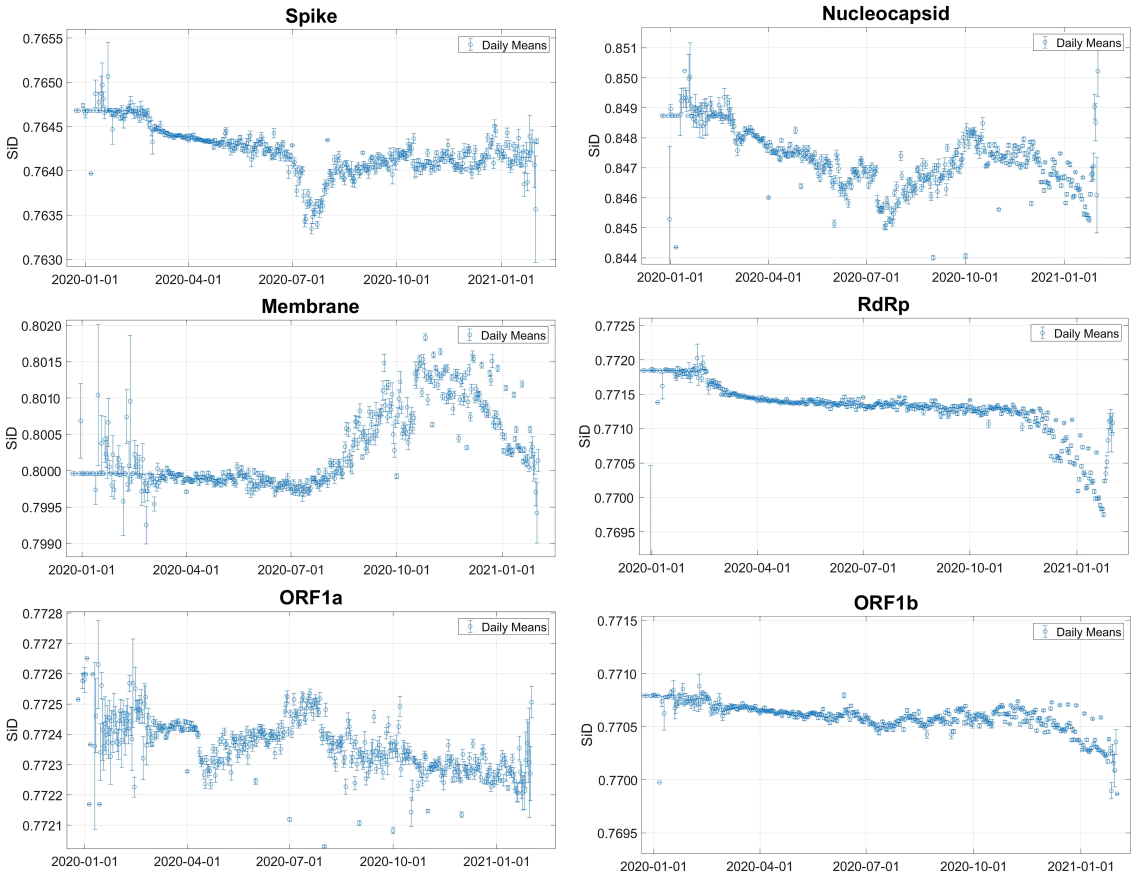

Fig. 5. SiD values daily averages in SARS-CoV-2 genes. Daily averaged SiD values estimated from all the available sequences and the standard errors on their assessments ($\sigma/\sqrt{N}$). SiD values span from 0 for an entirely opposite use of codons, to 1 for identical usage of codons to the host. Each panel refers to the protein indicated in the title.

was conducted separately for each gene here considered. In Fig. 6, we show the ENC-plot obtained for the Wuhan-Hu-1 genes, together with Wright's theoretical curve (see [27] and Section 2.4 for details). This plot was performed to have a reference for the next temporal analysis.

All genes are found below the curve, in accordance with Ayan (2020) [32], an indication that not only mutational bias but also natural selection plays a non-negligible role in their codon choice. The M, E and N proteins have the greatest distance from Wright's curve, pointing out that the codon usage of corresponding genes is the most subject to natural selection.

By using Wuhan-Hu-1 genes as reference, we analysed variations over time in the relative extent to which natural selection and mutational bias affects SARS-CoV-2 genes codon usage by generating ENC plots for each gene under study at different time points. To illustrate more clearly the temporal diversities, 10–day averages of the distances from Wright's theoretical curve have been estimated. In Fig. 7 is shown the heatmap of the differences between these values calculated over time and the reference values of Wuhan-Hu-1. In Fig. 7, a positive difference means that the distance is lower with respect to that of the reference gene (lighter colours), while a negative difference implies that the points are getting farther from the curve (darker colours).

These differences reflect the action of natural selection and mutational bias in shaping codon usage of SARS-CoV-2 genes over time. Interestingly, the average distances from Wright's theoretical curve increase over time for genes *S* and *RdRp*. This indicates that the codon usage of these genes is more affected by natural selection as time goes on. This observation is consistent with the fact that the Spike protein and the RdRp are the key proteins for infecting host's cells, controlling the binding and fusion with the host cell membrane, and the subsequent viral replication.

In contrast, the N and M proteins get closer to Wright's theoretical curve, suggesting that their codon usage is more affected by mutational bias than Wuhan-Hu-1 genes.

These observations, in a nutshell, indicate that S and RdRp have a tendency to explore increasingly more focussed regions of their viable sequence space, whereas M and N are prone to a wider, less contained exploration.

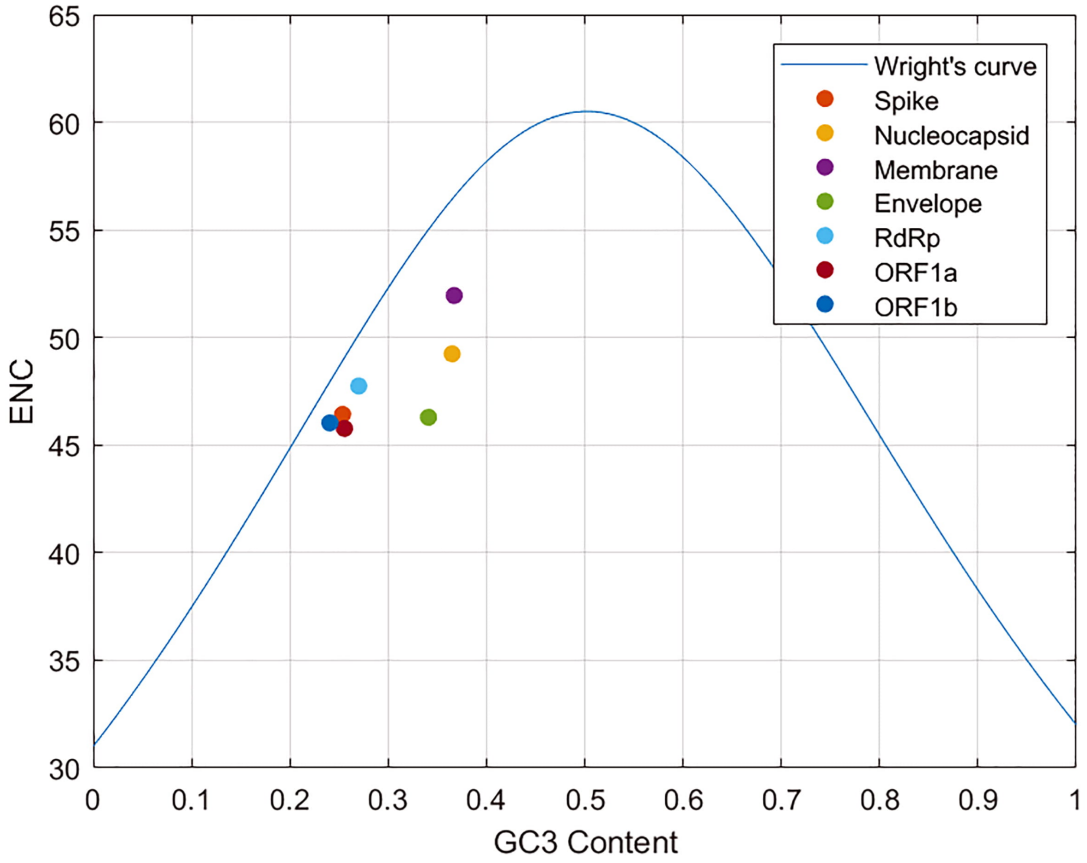

Fig. 6. ENC plot of Wuhan-Hu-1 genes. Representation of Wright's Theoretical curve together with Wuhan-Hu-1 genes ENC Plots (GC3 vs ENC). ENC values span from 23 to 61. GC3 is the fraction of GC content at the third position of the codon, and ranges from 0 if no codon ends in C or G, to 1 if all codons do.

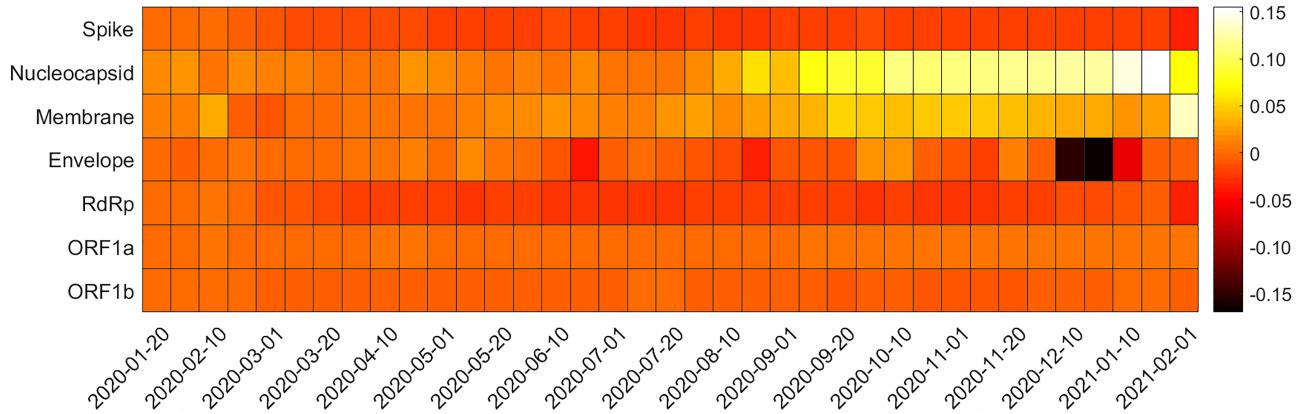

Fig. 7. Heatmap of ENC plots. Each square of the heatmap represents the difference between the distance of Wuhan-Hu-1 gene from Wright's curve and the distance of the examined gene from the same curve; 10-day averages have been calculated. Darker colours represent an increase of the distance from Wright's theoretical curve with respect to the reference genes distance, while lighter colours indicate a closer distance.

3.6 Forsdyke plots

Next, we estimated the evolutionary divergences of the seven genes considered herein. For this purpose, the RNA and protein sequence divergences of these genes were analysed by comparing the nucleotide sequences of the reference Wuhan-Hu-1 genes and their corresponding protein sequences with the other SARS-CoV-2 sequences extrapolated over time (Fig. 8). In this Fig. 8, each panel refers

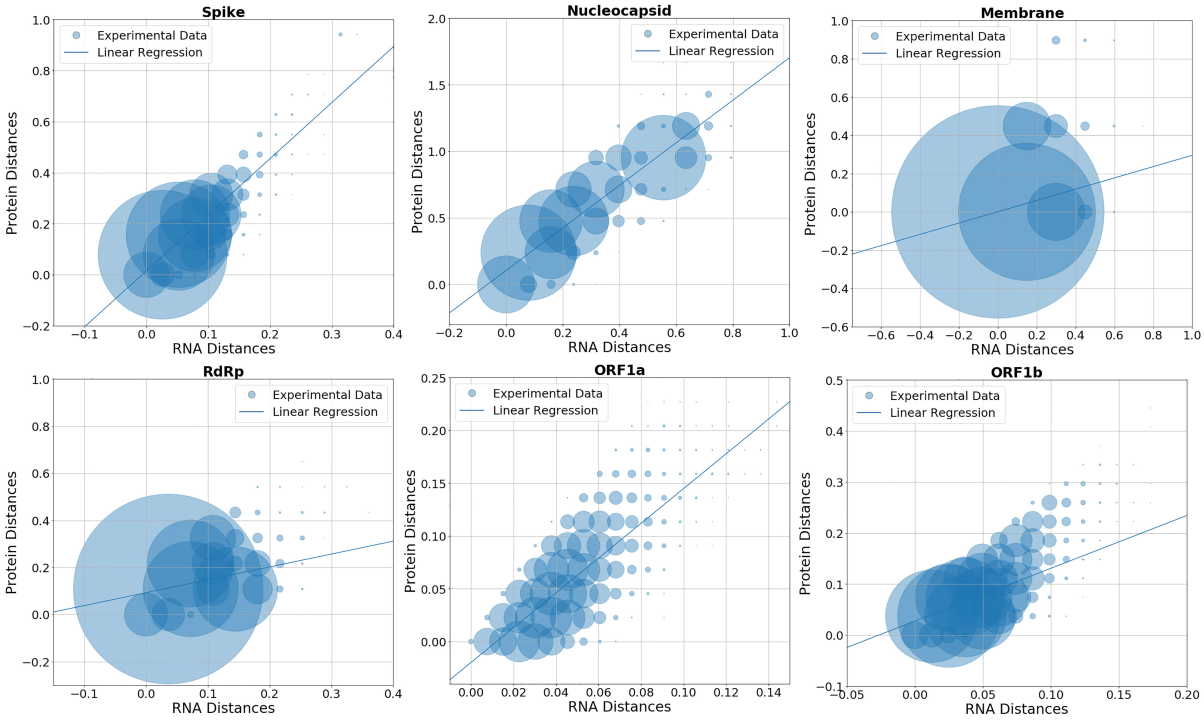

Fig. 8. Forsdyke plots of the considered genes. Forsdyke Plots of structural proteins (panels on the top) and replicase polyprotein complex (panels on the bottom). The radius of each circle is proportional to the number of data points in the same spot. Values span from 0, when the sequence under study is identical to the reference one, to 1, when they are different in every nucleotide/amino acid, for both axes. The inferred data from linear regressions are shown in Table 2.

to a single gene, and each gene sequence is represented by a single point. Because of the redundancy of mutations, many gene sequences characterized by the same pattern of mutations at the nucleotide and amino acid level are often superimposed on this plane. To take into account this redundancy, we report a bubble plot, in which the radius of each circle is proportional to the number of genes with the same coordinates (i.e., same nucleotide and amino acid divergence). Each bubble in these plots measures the divergence between a gene sequenced at different time points and the reference Wuhan-Hu-1 gene, as projected along with the phenotypic (protein) and nucleotide (DNA) axis. Thus, the slope is an estimation of the fraction of RNA mutations that results in amino acid substitutions [27].

Overall, protein and RNA sequence divergences are linearly correlated, and these correlations correspond to slopes and intercepts of the linear regressions in Table 2.

Specifically, gene *M* displays an accumulation of points around the x-axis, suggesting that the corresponding protein tends to evolve slowly by accumulating mutations on its sequence. Conversely, the steeper slopes for genes *N* and *S* suggest that these genes are fast-evolving compared to the other genes. This is in line with our previous study [30], according to which the N and S proteins are the main drivers of the speciation process among Coronaviruses. Considering the results from the ENC plots (section 3.5), the evolution of S protein is characterized by a purifying selection, while the N protein mutates under the action of mutational bias.

Table 2. Forsdyke Plots Linear Regressions. Linear regression parameters of the Forsdyke plots in Fig. 8.

Gene	Length (aa)	Location
ORF1a	4405	266–13 483
ORF1b	2696	13 483–21 555
RdRp	932	13 442 –16 236
S	1273	21 563–25 384
E	75	26 245–26 472
M	222	26 523–27 191
N	419	28 274–29 533

Moreover, ORF1a and ORF1b have a lower slope, pointing out that the phenotypic divergence of those is not as fast as that of the structural proteins. This is in line with the role of ORF1ab in viral replication and, therefore, with its higher sequence conservation. Interestingly, RdRp has a lower slope than ORF1b, pointing out the great importance of this protein in the replication process, that is, it tends to accumulate more synonymous mutations than missense, conserving the protein structure and functionalities.

It is worth noting that all the x-intercepts are close to 0. This observation suggests that *genic* differences (i.e., amino acid changes) have played a significant role in the *gene* di-

vergence.

4. Discussion

In this study, we performed a comprehensive analysis of the evolutionary divergence and codon usage of SARS-CoV-2 genes over time, considering all genomes available in GISAID up to February 7, 2021. After filtering out incomplete genomes, we retained 320,338 complete genomes, with the purpose of investigating their divergences from the first sequenced SARS-CoV-2 genome in Wuhan (NC 045512.2). We focused on seven SARS-CoV-2 genes and their encoded proteins that are crucial for virus structure, synthesis, transmissibility and virulence. Specifically, we considered *Spike*, *Nucleocapsid*, *Membrane*, *Envelope*, *RdRp* and the two segments of ORF1ab separately, *ORF1a* and *ORF1b*. The SARS-CoV-2 genomes have a tendency to diverge constantly from the reference genome, accumulating mutations over time [12]. We assumed herein that the accumulation of these mutations could affect the codon usage bias of the SARS-CoV-2 genes. Thus, we investigated the codon usage preference of SARS-CoV-2 in relation to the human host, using classical measures of codon usage bias such as RSCU, CAI, SiD and ENC.

First, we performed PCA in the space of RSCUs to investigate the major sources of variations in the codon usage of SARS-CoV-2 genes (Fig. 3). The two first principal components (PC1 and PC2) represent as much as 78% of the total variance of codon bias. Focusing on the PC1-PC2 plane, all the genes considered herein form distinct clusters, which indicate a different usage of synonymous codons in their sequence. Specifically, ORF1a, ORF1b and Spike exhibit very similar RSCUs and quite compact clusters. In contrast, Envelope protein corresponds to the most dispersed cluster. As this protein is the shortest one, we suppose that every mutation has a greater impact on the RSCU, resulting in the dispersed cluster that we observed in the PC1-PC2 plane. Finally, the Nucleocapsid protein shows a diffused cluster which is mainly due to the high rate of mutations occurring in its sequence. Notably, except for E, all the genes have similar values of PC1, which is affected by all codons uniformly (see Fig. 2). Conversely, the clusters separated along PC2 reflect a different usage only of codons encoding for phenylalanine (TTT and TTC), leucine (TTA, TTG and CTA), isoleucine (ATT, ATC), lysine (AAG) and valine (GTA).

Next, we analysed the temporal evolution of the codon usage bias by investigating CAI (Fig. 4) and SiD (Fig. 5) values as a function of the isolation dates of genomes across the world up to February 7, 2021. We observed that most SARS-CoV-2 genes show a tendency to use a broader set of synonymous codons over time. This is in contrast with the initial hypothesis of adaptation of the viral codon usage to the human host, which would lead to a greater tendency to use the synonymous set of codons preferred in highly expressed genes. We suppose here that a well-adapted codon usage to the host might be counterproductive, increasing the efficiency of translation, speeding up viral replication, and potentially leading to the death of the host. For this reason, a less-adapted codon usage over time could favour the adaptation of the virus to the human host while ensuring its survival through its spreading across the human population. Interestingly, RdRp shows a sudden decrease in both CAI and SiD values during the first months, to stabilise afterwards. We suggest that the first adaptation can be linked to the use of antiviral drugs, that mainly target this protein.

To understand the evolutionary forces modelling the differential usage of codons over time, we performed an ENC-plot analysis. We compared the position of genes isolated at different time points across the world with the first sequenced genomes in Wuhan, taken here as reference (see Fig. 6 for the reference ENC plot). This analysis revealed that the codon usage of the viral genes are subject to different balances between mutational bias and natural selection as a function of the isolation date. We found that the codon preference of the genes *S* and *RdRp* is mainly determined by natural selection. This may be linked to the crucial role of their corresponding proteins in the infection (S) and replication (RdRp) processes. RdRp is considered less vulnerable to mutations [12] 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 [33].

Conversely, the codon usage of the genes *M* and *N* is strongly affected by mutational bias, leading to a broader exploration of the mutation spectrum.

Performing a Forsdyke plot analysis (Fig. 8), we obtained results confirming those of our previous study [30], showing that the speciation process of Coronaviruses is predominately driven by the genes *N* and *S*. According to Rehman *et al*. (2020) [34], the spike protein, which mediates the virus interaction with the human host cells, is more prone to mutations, although the amino acids implicated in the spike and the angiotensin-converting enzyme 2 (ACE2) interface tend to be more conserved [35], due to the importance of their function [34,36]. In particular, we have found about one missense mutation for each of the 2700 analysed sequences in the Receptor Binding Domain. These divergences though, considering the results from ENC plots, are profoundly different, that is, while S is mutating under the action of a purifying pressure, N evolution is mainly due to mutational bias.

Therefore, the *N* and *S* genes accumulate beneficial mutations that would increase the evolvability and transmissibility of SARS-CoV-2, allowing it to continuously adapt to different populations, like spilling over from bats, or other candidate natural reservoirs, to humans. The greatest rate of mutation of these two proteins may be linked to the high number of epitopes found on them [37–39].

5. Conclusions

In the present study, we have studied the adaptive features of SARS-CoV-2 genomes, taking into account the location and time those genomes were isolated and sequenced. The main question addressed in this study is whether the human codon usage bias, during the pandemic, has affected and optimized the codon bias of SARS-CoV-2.

We focused on the RNA genes coding for M, E, S, N, RdRp and ORF1ab. All these viral genes appear to evolve toward a codon usage pattern that is less efficient than it could be with respect to the human codon usage, at least in the time sequel we have considered. This is in contrast with the hypothesis of a greedy viral adaptation in codon usage bias to the host. On the other hand, it is in agreement with the virulence trade-off hypothesis, according to which the virus would not adapt to the host at the optimal level, to keep it alive and consequently spread the infection, a hypothesis that has been extensively discussed in a recent paper [29].

Moreover, from Forsdyke and ENC plots, we observed that the S and N proteins evolve faster and drive the speciation process. The S protein is subject to positive purifying selection, while N is under mutational bias. ORF1ab, and in particular RdRp, present a lower number of non-synonymous mutations (i.e., they have a low value for the slope in the Forsdyke plots). As a matter of fact, according to the ENC plot in Fig. 7, RdRp mutations are mostly driven by natural selection, consistently with its key role in the replication process.

It is very important to extend the type of investigation presented herein, so as to have a clearer perspective on the adaptation of viral codon bias to the host, a theme of great relevance for the control of epidemics.

Author contributions

EP, MD, SF, AP, AGG and AG designed the research study and performed the research. EP, MD, SF wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

The authors warmly thank the anonymous referees whose constructive comments were of great help in revising and focusing our paper.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflict of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at https://www.imrpress.com/journal/FBL/27/1/10.31083/j.fbl2701013.

References

[1] Yoshimoto FK. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. The Protein Journal. 2020; 39: 198–216.

[2] Meriem L, Tarek A, Souad KMW, Chemao-Elfihri, Mohammed H, Abdelomunim E, *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. PLoS ONE. 2020; 15: e0240345.

[3] Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Reports. 2020; 19: 100682.

[4] Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, *et al*. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochimica Et Biophysica Acta (BBA) - Molecular Basis of Disease. 2020; 1866: 165878.

[5] Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, *et al*. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceutica Sinica B. 2020; 10: 1228–1238.

[6] Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virology Journal. 2019; 16: 69.

[7] DeDiego ML, Nieto-Torres JL, Jimenez-Guardeño JM, Regla-Nava JA, Castaño-Rodriguez C, Fernandez-Delgado R, *et al*. Coronavirus virulence genes with main focus on SARS-CoV envelope gene. Virus Research. 2014; 194: 124–137.

[8] Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual Review of Virology. 2017; 3: 237–261.

[9] Nyayanit DA; Yadav PD; Kharde R; Cherian S. Natural Selection Plays an Important Role in Shaping the Codon Usage of Structural Genes of the Viruses Belonging to the Coronaviridae Family. Viruses. 2021; 13: 3.

[10] Thébaud G, Chadoeuf J, Morelli MJ, McCauley JW, Haydon DT. The relationship between mutation frequency and replication strategy in positive-sense single-stranded RNA viruses. Proceedings. Biological Sciences. 2010; 277: 809–817.

[11] Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99: 803–808.

[12] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, *et al*. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine. 2020; 18: 179.

[13] Nambou K, Anakpa M. Deciphering the co-adaptation of codon usage between respiratory coronaviruses and their human host uncovers candidate therapeutics for COVID-19. Infection, Genetics and Evolution. 2020; 85: 104471.

[14] Postnikova OA, Uppal S, Huang W, Kane MA, Villasmil R, Rogozin IB, *et al*. The Functional Consequences of the Novel Ribosomal Pausing Site in SARS-CoV-2 Spike Glycoprotein RNA. International Journal of Molecular Sciences. 2021; 22: 6490.

[15] Qian W, Yang J, Pearson NM, Maclean C, Zhang J. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genetics. 2012; 8: e1002603.

[16] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Research. 2019; 47: D94–D99.

[17] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al*. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New

England Journal of Medicine. 2020; 382: 727–733.
[18] 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.
[19] Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution. 1987; 24: 28–38.
[20] Sharp PM, Li W. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987; 15: 1281–1295.
[21] Roth A, Anisimova M, Cannarozzi GM. Measuring codon usage bias. Codon Evolution Mechanisms and Models. 2012; 189–217.
[22] Xia X. DAMBE5: a Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. Molecular Biology and Evolution. 2013; 30: 1720–1728.
[23] Hodgman MW, Miller JB, Meurs TE, Kauwe JSK. CUBAP: an interactive web portal for analyzing codon usage biases across populations. Nucleic Acids Research. 2020; 48: 11030–11039.
[24] Jolliffe IT. Principal Component Analysis. Series: Springer Series in Statistics. Springer. 2002; 487: 28.
[25] Sun X, Yang Q, Xia X. An improved implementation of effective number of codons (nc) Molecular Biology and Evolution. 2013; 30: 191–196.
[26] Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87: 23–29.
[27] Forcelloni S, Giansanti A. Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. Journal of Molecular Evolution. 2020; 88: 164–178.
[28] Forcelloni S, Giansanti A. Mutations in disordered proteins as early indicators of nucleic acid changes triggering speciation. Scientific Reports. 2020; 10: 4467.
[29] Rochman N, Wolf YI and Koonin EV. Evolution of human respiratory virus epidemics [version 2; peer review: 2 approved]. F1000Research. 2021; 10: 447.
[30] Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A and Pavlopoulou A, Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. Viruses. 2020; 12: 498.
[31] Gairu L, Huijuan W, Shilei W, Gang X, Cheng Z, Wenyan Z, *et al*. Insights into the genetic and host adaptability of emerging porcine circovirus. Virulence. 2018; 9: 1301–1313.
[32] Ayan R, Fucheng G, Bhupender S, Shelly G, Karan P, Xiaoyuan C, *et al*. Base Composition and Host Adaptation of the SARS-CoV-2: Insight From the Codon Usage Perspective. Frontiers in Microbiology. 2021; 12: 548275.
[33] Pfeiffer JK, Kirkegaard K. A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. Proceedings of the National Academy of Sciences. 2003; 100: 7289–7294.
[34] Rehman SU, Shafique L, Ihsan A, Liu Q. Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. Pathogens. 2020; 9: 240.
[35] Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proceedings of the National Academy of Sciences. 2021; 118: e2104241118.
[36] Forcelloni S, Benedetti A, Dilucca M, Giansanti A. Identification of conserved epitopes in SARS-CoV-2 spike and nucleocapsid protein. BioRxiv. 2020. (in press)
[37] Ahmed SF, Quadeer AA, McKay MR. Preliminary Identification of Potential Vaccine Targets for the COVID19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses. 2020; 12: 254.
[38] 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. Journal of Virological Methods. 2020; 277: 113806.
[39] 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. Journal of Clinical Virology. 2004; 30: 309–312.