The heterogeneous landscape and early evolution of pathogen-associated CpG dinucleotides in SARS-CoV-2

Andrea Di Gioacchino[1], Petr Šulc[2], Anastassia V. Komarova[3], Benjamin D. Greenbaum[4], Rémi Monasson[1], and Simona Cocco[1]

[1]Laboratoire de Physique de l'Ecole Normale Supérieure, PSL & CNRS UMR8063, Sorbonne Université, Université de Paris, F-75005 Paris, France
[2] School of Molecular Sciences and Center for Molecular Design and Biomimetics, The Biodesign Institute, Arizona State University, 1001 South McAllister Avenue, Tempe, Arizona 85281, USA
[3]Molecular Genetics of RNA viruses, Department of Virology, Institut Pasteur, CNRS UMR-3569, 75015 Paris, France.
[4]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 1275 York Avenue New York, NY 10065

August 25, 2020

Abstract

COVID-19 can lead to acute respiratory syndrome in patients, which can be due to dysregulated immune signaling. We analyze the distribution of CpG dinucleotides, a pathogen-associated molecular pattern, in the SARS-CoV-2 genome. We find that CpG relative abundance, which we characterize by an adequate force parameter taking into account statistical constraints acting on the genome at the nucleotidic and amino-acid levels is, on the overall, low compared to other pathogenic betacoronaviruses. However, the CpG force widely fluctuates along the genome, with particularly low value, comparable to the circulating seasonal HKU1, in the Spike protein (S) coding region and high value, comparable to SARS and MERS, in the highly expressed nucleocapside (N) coding regions, whose transcripts are relatively abundant in the cytoplasm of infected cells and present in the 3'UTRs of all subgenomic RNA. This dual nature of CpG content could confer to SARS-CoV-2 the ability to avoid triggering pattern recognition receptors upon entry, while eliciting a stronger response during replication. We then investigate the evolution of synonymous mutations since the outbreak of the COVID-19 pandemic. Using a model of the viral gene evolution under human host pressure, we find that synonymous mutations seem driven, in the N protein coding region, both by the viral codon bias and by the high value of the CpG content, leading to a loss in CpG. Sequence motifs preceding these CpG-loss-associated loci match recently identified binding patterns of the Zinc finger Anti-viral Protein.

Keywords— ssRNA viruses, SARS-CoV-2, pathogen-associated molecular patterns (PAMPs), pattern recognition receptors (PRRs), viral host mimicry, CpG motifs, evolution of synonymous mutations

1 Introduction

When a virus enters a new host, it can present pathogen-associated molecular patterns (PAMPs) that are rarely seen in circulating strains that have adapted to that host's immune environment over evolutionary timescales. The emergence of SARS-CoV-2, therefore, provides a rare window into innate immune signaling that may be relevant for understanding immune-mediated pathologies of SARS-CoV-2, anti-viral treatment strategies, and the evolutionary dynamics of the virus, where evidence for selective pressures on viral features can reflect what defines "self" in its new host. As a case in point, the 1918 influenza pandemic was likely caused by a strain that originated in water fowl and entered the human population after possible evolution in an intermediate host. That viral genome presented CpG dinucleotides within a context and level of density rarely found in the human genome where they are severely underespressed, particularly in a set of genes coding for the proteins associated with antiviral innate immunity [1, 2, 3, 4]. Over the past century the 1918 H1N1 lineage evolved in a directed manner to lower these motifs and gain UpA motifs, in a way that could not be explained by its usage of amino-acid codon bias [4, 5]. It has since been found that these motifs can engage the pattern recognition receptors (PRRs) of the innate immune system [6, 7], and directly bind the Zinc finger Anti-viral Protein (ZAP), both in a CpG-dependent manner [8, 9, 10]. Hence, the interrogation of emergent viruses from this perspective can predict novel host virus interactions.

COVID-19 presents, thus far, a different pathology than that associated with the 1918 H1N1, which was disproportionately fatal in healthy young adults. It has been characterized by a large heterogeneity in the immune response to the virus [11, 12, 13] and likely dysregulated type I interferon signaling [14, 15, 16, 17]. Various treatments to attenuate inflammatory responses have been proposed and are currently under analysis or being clinically tested [18]. It is therefore essential to quantify pathogen-associated patterns in the SARS-CoV-2 genome for multiple reasons. The first is to better understand the pathways engaged by innate immune agonism and the specific agonists to help build better antiviral therapies. Another is to better predict the evolution of motif content in synonymous mutations in SARS-CoV-2, as it will help understand the process and timescales of attenuation in humans. Third is to offer a principled approach for optimizing vaccine strategy for designed strains [19, 20] to better reflect human-genome features.

In this work we will use the computational framework developed in [4] to carry out a study of non-self associated dinucleotide usage in SARS-CoV-2 genomes. The statistical physics framework is based on the idea of identifying the abundance or scarcity of dinucleotides given their expected usage based on host features. It generalizes the

standard dinucleotide relative abundance introduced in [1], as it can easily incorporates constraints in coding regions coming from amino-acid content and codon usage. The outcome of the approach are forces [4] that characterize the deviations with respect to null models in which the number of dinculeotides is the one statistically expected under a set of various constraints. CpG forces could be related to the evolutionary constraint to lower or increase CpG number, under the pressure of host PRRs that recognize a pathogen. Such formalism has further been applied to identify non-coding RNA from repetitive elements in the human genome expressed in cancer that can also engage PRRs [21], to characterize the CpG evolution through synonymous mutations in H1N1 [4], and to characterize local and non-local forces on dinucleotides across RNA viruses [7].

We perform an analysis of the landscape of CpG motifs and associated selective forces in SARS-CoV-2 in comparison with other genomes in the coronavirus family in order to understand specific PAMP associated features in the new SARS-CoV-2 strains (Sec. 2.1). We also focus on the heterogeneity of CpG motif usage along the SARS-CoV-2 genome (Sec. 2.1 and Sec. 2.2). Finally we use a model of the viral gene evolution under human host pressure, characterized by the CpG force, to study synonymous mutations, and in particular those which change CpG content, observed since the SARS-CoV-2 entered the human population (Sec. 2.3). The latter approach points out at hotspots where new mutations will likely attenuate the virus, while evolving in contact with the human host.

2 Results

2.1 The landscape of CpG forces in SARS-CoV-2 is strongly heterogeneous

We first compute the global force on CpG dinucleotides for SARS-Cov-2 and a variety of other viruses from the *Coronaviridae* family affecting humans or other mammals (bat, pangolin), see Fig. 1a, using as null model the nucleotide usage calculated from human genome [22] (see Methods Sec. 4.2)[1]. The value $\simeq -1.1$ of the global force for SARS-CoV-2 is lower than for SARS and MERS, and other strongly pathogenic viruses in humans, such as H1N1, H5N1, Ebola (Suppl. Fig. SI.1). MERS shows the highest CpG force among the human coronaviruses, followed by SARS, while some bat coronaviruses have stronger CpG force. It is worth noticing that SARS-CoV-2 is among the viruses with smallest global CpG force and some hCoV that circulate in humans with less pathogenicity have global CpG forces comparable or higher than that of SARS-CoV-2. The absence of a straightforward correlation between global CpG force and the pathology of a coronavirus in humans calls for a finer, local analysis of CpG forces we report below.

Figure 1b compares the forces acting on CpG and UpA motifs within the *Coronaviridae* family, with a particular emphasis on the genera *Alphacoronavirus* and *Betacoronavirus*, and on those viruses which infect humans [23]; for other dinucleotides, see Suppl. Fig. SI.2[2]. We observe an anti-correlation between UpA and CpG forces. UpA is the CpG complementary motif corresponding to the two nucleotidic substitutions more likely to occur in terms of mutations; Transitions have larger probability with respect to transversions and are less likely to result in amino-acid substitutions. Such anti-correlations are not observed with motifs that are one mutation away from CpG (Suppl. Fig. SI.4).

To go beyond the global analysis we study the local forces acting on CpG in fixed-length windows along the genome. Results for SARS, MERS, SARS-CoV-2, hCoV-HKU1 and two representative sequences of bat and pangolin coronaviruses, chosen for their closeness to SARS-CoV-2 are reported in Fig. 1c. In some genomic regions, especially at the 5' and 3' extremities, SARS-CoV-2, SARS and MERS (together with the bat and pangolin viruses) have a peak in CpG forces, which is absent in the hCoV-HKU1 (as well as in the other hCoVs, see Suppl. Fig. SI.5). The high CpG forces at the extremities could have an important effect on the activation of the immune response via sensing, as the life cycle of the virus is such that the initial and final part of the genome are those involved in the subgenomic transcription needed for viral replication [24, 25]. During the infection many more RNA fragments from these regions are present in the cytoplasm than from the other parts of the viral genome. Consequently, despite the relatively low CpG content of SARS-CoV-2 compared to other coronaviruses, there can be high concentrations of CpG-rich RNA due to the higher transcription of these regions.

The similarity between the high values of the maximum local forces of SARS-CoV-2 and those of SARS, bat and pangolin coronaviruses shown in Fig. 1a confirm this pattern: MERS and SARS, viruses that are likely less well adapted to a human host, have the highest local peaks in CpG content, followed by SARS-CoV-2 and then by seasonal strains that circulate in humans. It is interesting to notice that high and very high levels of proinflammatory cytokines/chemokines (such as IL-6 and TNF-α) have been observed in, respectively, SARS and MERS and, at times, SARS-CoV-2 infection [11, 17, 26]. These results are qualitatively corroborated by the simpler analysis of CpG motif density (Suppl. Fig. SI.3).

2.2 Forces acting on coding regions widely vary across structural proteins

We now restrict our analysis to the coding regions of SARS-CoV-2 and, in particular, on two structural proteins, N (nucleocapside) and S (spike) [19, 29, 30]. To account for the extra constraints on the amino-acid content and codon usage acting over these coding regions we modify our local force calculation, see Methods Sec. 4.1 and Sec. 4.2.

The landscape of forces when taking into account these coding constraints is shown, restricted to the coding regions of SARS-CoV-2 genome (see Methods Sec. 4.5), in Fig. 2a, together with the forces computed without the coding constraints with the same human nucleotide frequency used in Fig. 1 (dashed lines).The global shift of the forces with or without coding constraints is due to the different background used. Indeed, when the human nucleotide frequency is modified into that computed through only the human coding RNAs [22], the result without coding constraints becomes much similar to the one with coding constraints (dotted lines in Fig. 2a). Apart from this global shift, the qualitative result does not substantially differ from the findings of Fig. 1c. In particular the peak of high CpG density and force is still present at the 5' and the 3' ends of the genome, including the N-ORF, the envelope E-ORF and membrane glycoprotein M-ORF regions. On the contrary in the S-ORF region the CpG forces under coding constraint are small.

Detailed results for the S (S ORF) and N (N ORF) proteins are shown in, respectively, Figs. 2b&2c. These structural proteins are present and quite similar across the *Coronaviridae* family, and allow us to compare several strains of coronaviruses. In the S ORF, SARS-CoV-2 shows the lowest global CpG force among the human-infecting

[1] This is a natural choice in order to predict regions sensed by PRRs as non-self, which may engage the human innate immune response. Similar results in terms of the rank order of viruses are obtained using for null model the average nucleotide usage in coronavirus genomes.

[2] Comparison based on dinucleotide number rather than force give similar results, see Suppl. Fig. SI.3

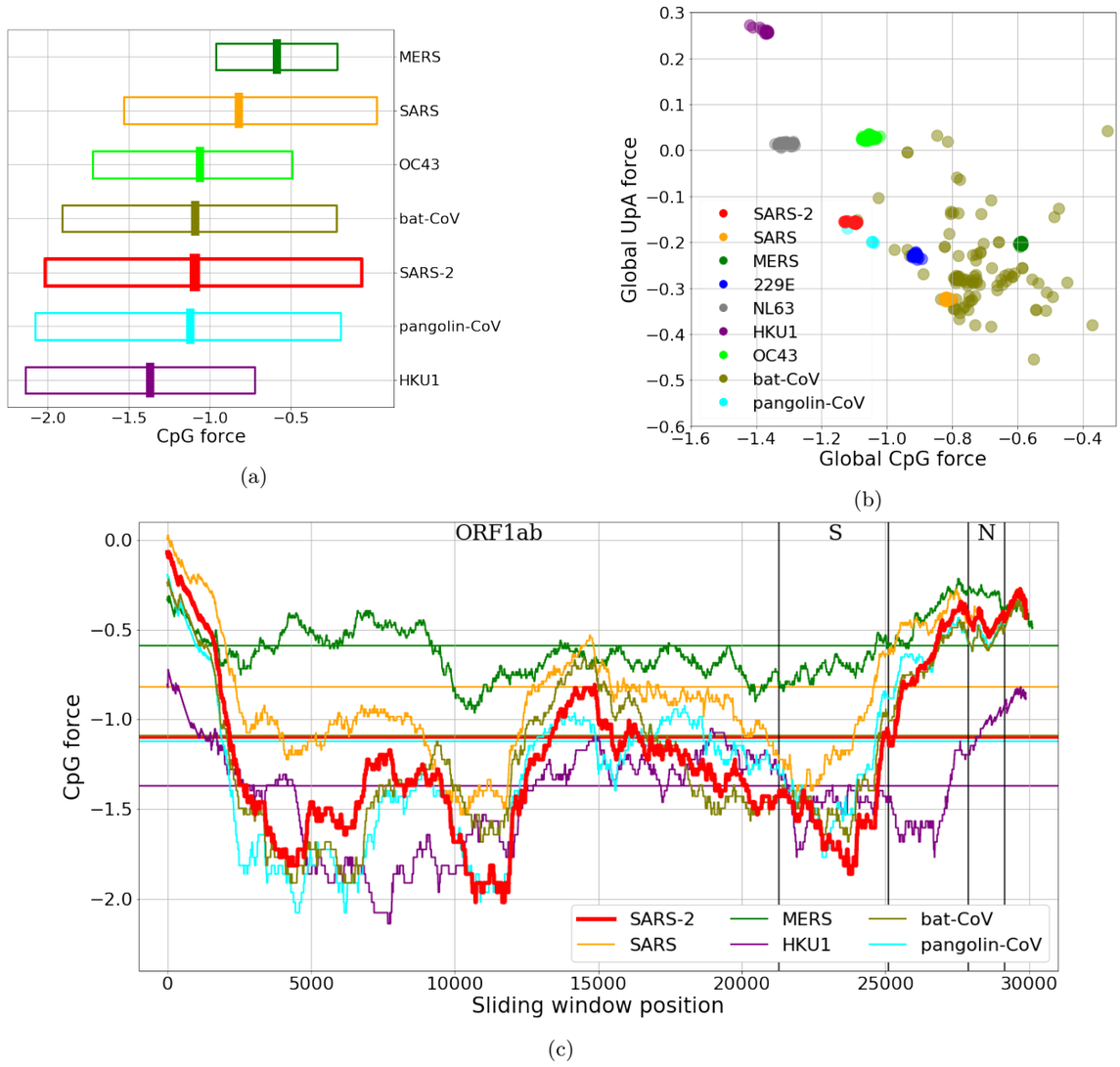


Figure 1: **CpG and UpA forces at fixed nucleotidic usage, and local fluctuations in the coronaviriadae family.** (a): Forces on CpG dinucleotides for SARS-Cov-2 and other coronaviruses. The central thick bars show global forces over the whole genomes, while lateral bars locate minimal and maximal values of the local forces over sliding windows from 1.5kb (at the boundary) up to 3kb (in the bulk), see text. (b): UpA vs. CpG global forces. Species are well clustered, due to the large conservation of the sampled sequences, except for viruses circulating in bats composed of several diverse strains. Notice the anticorrelation between CpG and UpA forces ($r^2 = 0.48$). (c): Local CpG force analysis in sliding windows of 3 kb along the genome for Coronaviruses; Horizontal lines correspond to the global forces shown in panel (b). The force is highly variable along the genome, with much larger values in certain regions (such as the coding region for protein N) than in others (e.g. coding region for protein S). The maximum value of the local CpG force hints at the similarity of SARS-CoV-2 with the most pathogenic viruses. Bat sequence analyzed in panels (a) and (c): RaTG13; Pangolin sequence obtained in Guangdong in 2019. Data from VIPR [27] and GISAID [28], see Methods Sec. 4.5 and Suppl. Sec. SI.1.

betacoronaviruses, see Fig. 2c. The CpG force is much higher for protein N in SARS-CoV-2, immediately below the level of SARS and above that of MERS, see comparison with human-infecting members of the *Coronaviridae* family presented in Fig. 2b. The comparative analysis of forces in the E-ORF (Suppl. Fig. SI.6b) gives results similar to the N-ORF, while smaller differences in CpG force among coronaviruses that circulate in humans are observed for the M-ORF (Suppl. Fig. SI.6c).

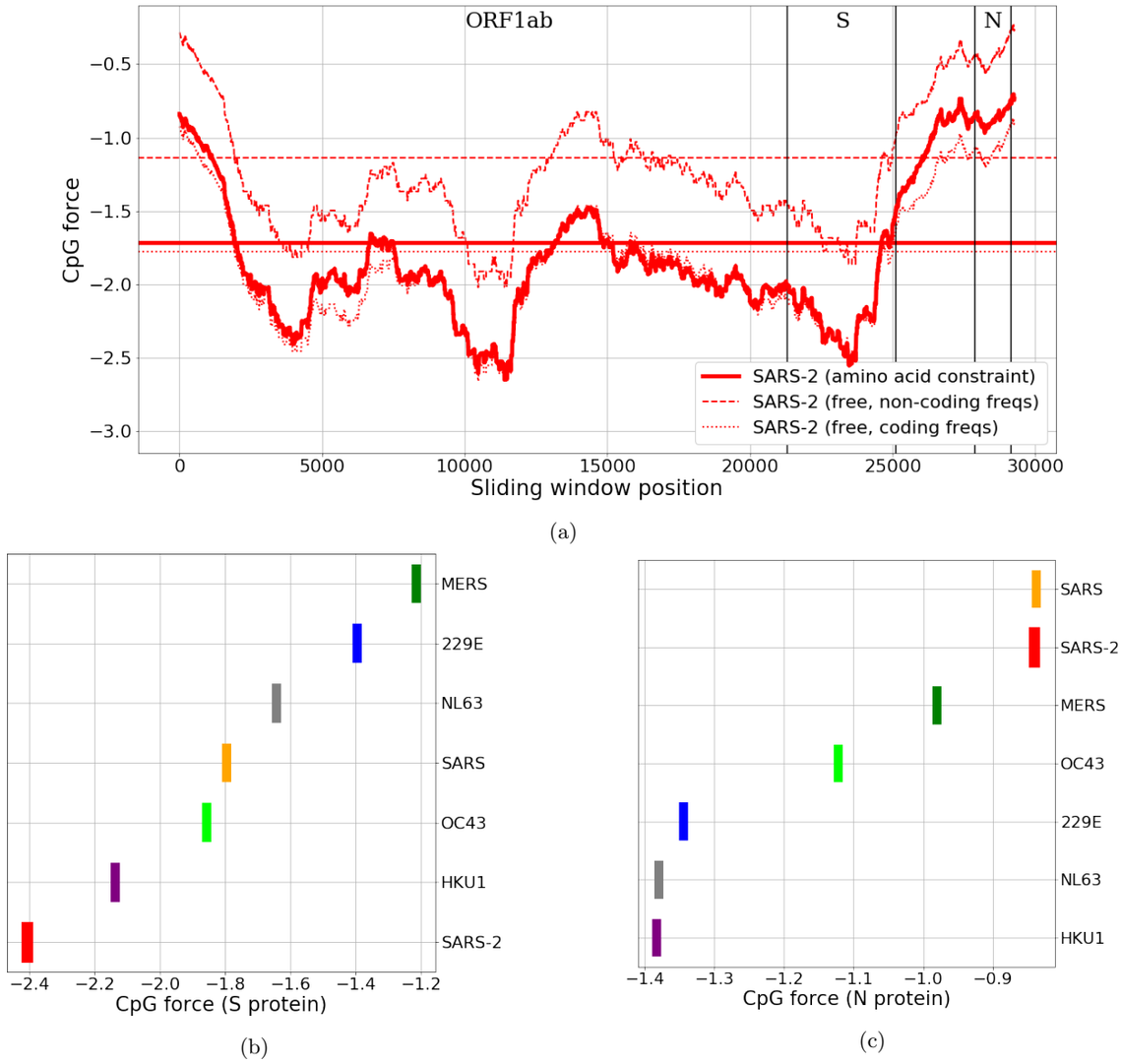


Figure 2: **CpG local Forces on SARS-CoV-2 coding regions, at fixed amino acids and codon usage**. (a): Local forces, over sliding windows ranging from 1.5kb (at edges) up to 3kb (in the bulk), in the the coding regions of SARS-CoV-2 (pre-processed to ensure the correct reading frame, see Methods Sec. 4.5). Full, thick red curve shows forces computed at fixed amino-acid content. The dashed and dotted line shows forces computed with respect to a null model with no constraint on amino acids and nucleotide frequencies computed from, respectively, all human RNAs and coding human RNAs. Horizontal lines locate global forces. Vertical lines show boundaries of protein domains. (b) and (c): global forces for structural proteins (S and N) in the *Coronaviridae* family. This values were averaged on 4 to 20 sequences from VIPR [27] and GISAID [28], see Methods Sec. 4.5 and Suppl. Sec. SI.1.

2.3 Force-based model accounts for early evolution of synonymous CpG-related mutations in SARS-CoV-2

We now assess the ability of our CpG force model to predict biases in the synonymous mutations already detectable across the few months of evolution following the first sequencing of SARS-CoV-2 (data from GISAID [28], reference sequence Wuhan, 26-12-2019, last updated sequence 2020-07-20, see Methods Sec. 4.5). Barring confounding effects, we expect that high-force regions, such as N ORF, will be driven by host mimicry towards a lower number of CpG motifs. Other regions, such as S ORF, have already low CpG content and would feel no pressure to keep the CpG content at that level, so random mutations would likely increase their CpG numbers.

These predictions are in good agreement with the observed mutations in current SARS-CoV-2 data, as shown in Fig. 3a. Most of the mutations that decrease the number of CpG are located at the 5' and 3'-end of the sequence, in correspondence with the high peak in CpG force, notably in the N ORF region. Conversely, mutations that increase the number of CpG are found in ORF1ab, in the low-CpG-force regions and in the S ORF region.

2.3.1 Analysis of N ORF suggests implication of ZAP in progressive loss of CpG

We now focus on the N protein. Figure 3b shows the locations of synonymous mutations, indicated by bars, and the number of variants in which they are found to occur, indicated by star symbols of corresponding sizes (only if observed more than 5 times).

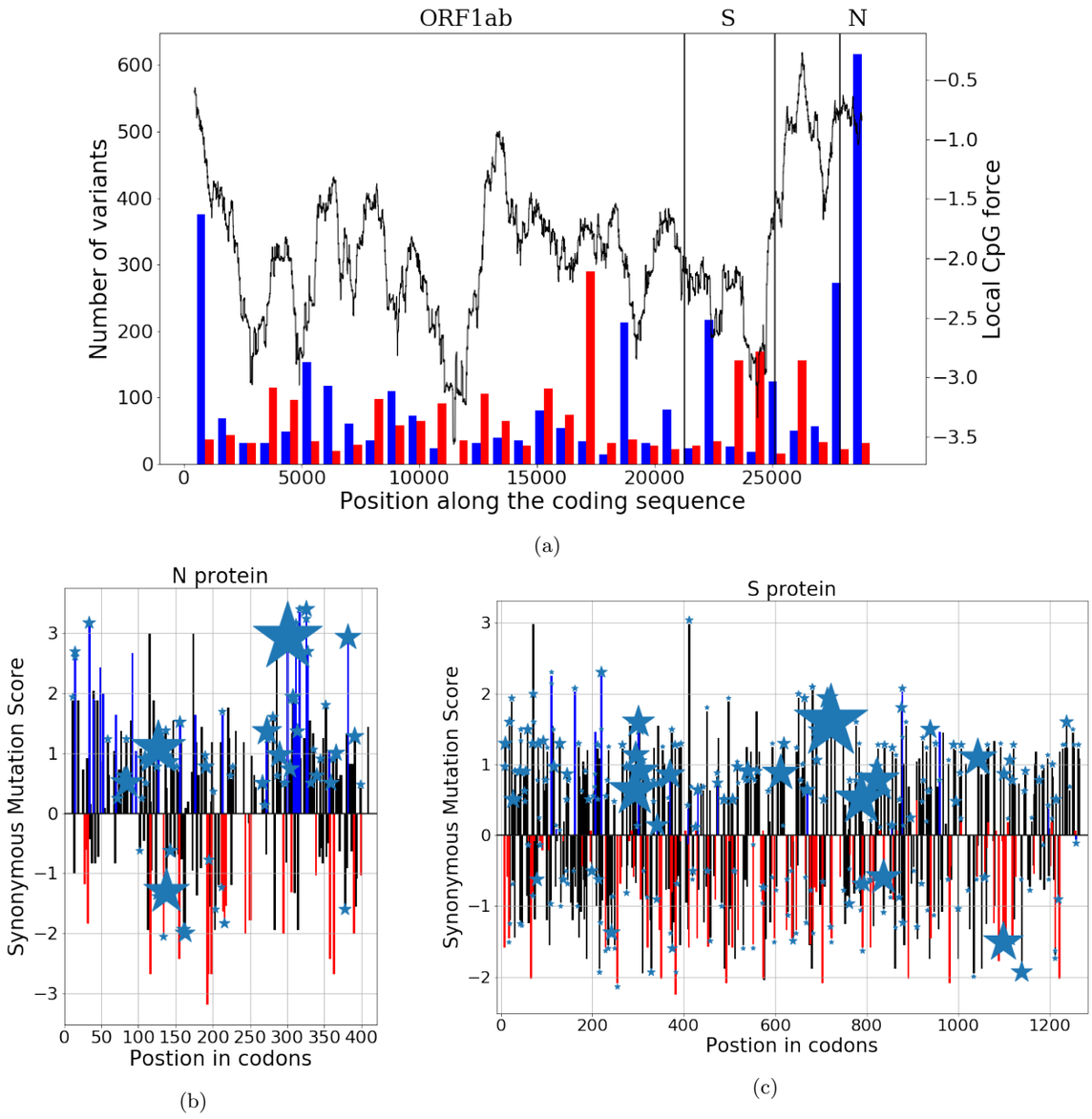
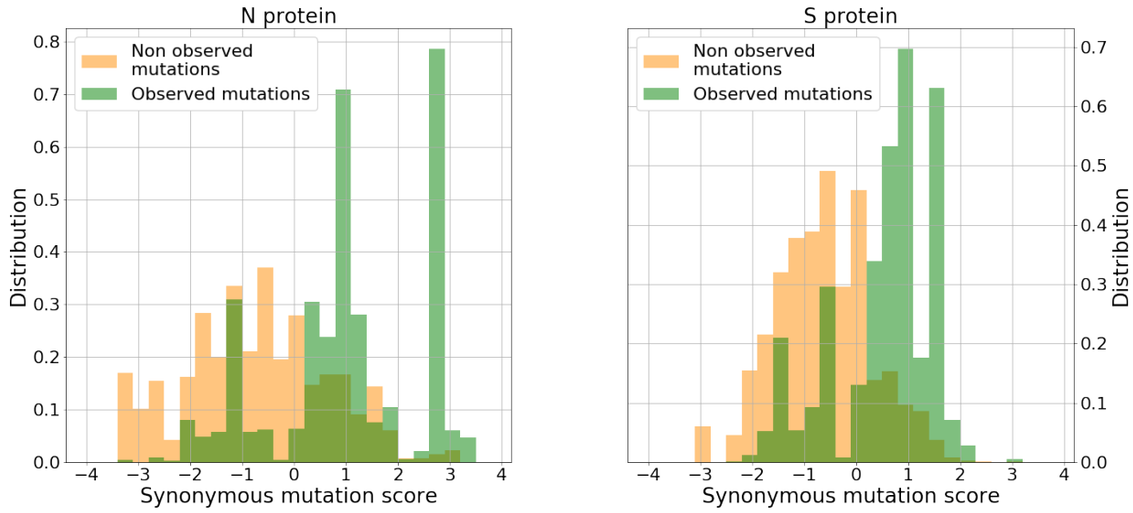


Figure 3: **Analysis of synonymous mutations in the early evolution of SARS-CoV-2 genome.** Reference genome is Wuhan, 26-12-2019. (a): CpG-changing mutations. Red and blue bars show the numbers of variants that, respectively, increased and decreased the mean CpG content over 900 b windows (left scale). The black curve shows the CpG force along the sequence (right scale). (b), (c) for N and S protein: bars show the synonymous mutation score (SMS) associated to mutations along the sequence. Color code for mutations: black–no change in CpG; blue–increase in CpG, red–decrease in CpG. Stars show mutations observed at least 5 times in sequence data; the size of the star is proportional to the number of observations. Data from GISAID [28], see Methods Sec. 4.5.



(a) (b)

Figure 4: **Model can distinguish observed from non-observed mutations.** Reference genome is Wuhan, 26-12-2019. (a), (b): histograms of SMS distribution for observed variants (in green) and for non-observed putative mutations (in yellow) in the N and S proteins. Data from GISAID [28], see Methods Sec. 4.5.

We observed a total number of $M_s = 2111$ variants with synonymous mutations, out of which 208 are unique. Out of these M_s variants 965 and 55, respectively, lower and increase CpG, while the remaining 1091 leave CpG content unchanged. It is remarkable than more than 94% of the variants differing in CpG count actually have it decreased.

When restricting the analysis on the 965 variants in which the CpG count decreases, the losses take place in at 22 different loci. The nucleotide motifs preceding these loci are listed in the top 22 lines of Table 1, together with their positions along SARS-CoV-2 (Wuhan, 26/12/2019) and their number of occurrences in the sequence data. 10 out of 22 of these motifs, which represent 709 out of the 965 observed CpG losses, are of the type CnxGxCG, where nx is a spacer of n nucleotides and were identified as ZAP binding patterns in [9]. The binding affinity of ZAP to the motifs strongly depends on the spacer length, n, with top affinity for n=7 [9]. Notice that 3 out of the 10 CpG-suppression related motifs in SARS-CoV-2 correspond to n=7. Other motifs of the type CnxGcCG are also present in SARS-CoV-2, but their CpG is not lost in sequence data, see last 2 lines of Table 1; the dissociation constants associated to their spacer lengths are on average larger than the ones of the motifs showing CpG loss.

For the S protein (see Fig. 3c and star sizes), we observe $M_s = 4161$ synonymous variants (with 536 unique mutations). Among these variants, 350 and 440, respectively, lower and increase the CpG content. Therefore, only about 44% of the variants that affect the CpG count decrease it. These results support the existence of early selection pressure to lower CpG occurrence in N ORF, but not in S ORF.

2.3.2 Model is able to discriminate between observed and non-observed mutations

Our model can be further used to predict the odds of synonymous mutations from the original SARS-CoV-2 (Wuhan, 26-12-2019) sequence (see Sec. 4.5). For this purpose, we introduce a synonymous mutation score (SMS), defined in Methods Sec. 4.3, whose value expresses how likely the mutation is to appear under the joint actions of the force and of the reference codon usage[3], see Eq. 1. The effect of codon usage is important, in particular for synonymous mutations that do not change CpG and for which it is the only driving factor in our model.

In Fig. 3 we show our predictions for synonymous mutations in the N (3b, 4a) and the S (3c, 4b) proteins. Figs. 3b and 3c show SMS along, respectively, the N and S sequences and the unique mutations, respectively, lowering (blue), increasing (red), or leaving unchanged (black) the CpG content. The majority of mutations, both unique and taking into account multiplicities (which are described by the stars in Figs. 3b and 3c), in SARS-CoV-2 correspond to high SMS, in agreement with our model.

To make our arguments more quantitative, we tested the ability of our model to discriminate between observed and non-observed mutations. In Figs. 4a and 4b we show the histograms of the SMS corresponding to observed synonymous variants (in green) and to putative mutations that would leave amino-acid content unchanged but have not been observed so far (in yellow). The distribution of SMS for observed variants is shifted to higher values compared to their counterparts for non-observed mutations, both for the proteins N and S. Hence, our model is able to statistically discriminate between non-observed and observed synonymous mutations (ANOVA F-test: 314 for N and 1083 for S). Note that for the null mutational models in which synonymous mutational rate are uniform, the score distribution for observed and unobserved mutations is equally peaked at zero (ANOVA F-test=1). We have checked that our model is able to discriminate between non-observed and observed synonymous mutations even when we restrict only to unique mutations, dropping any information about multiplicity, (ANOVA F-test=100 (N protein), 193 (S protein) see Suppl. Fig. SI.8). We have further performed comparative tests of our model, in which mutations are driven by codon bias and CpG forces, with simpler models using: i) only the transition versus transversion rate (with ratio 4:1), [31] (trs-trv bias) (Methods Sec. 4.4.), ii) the transition versus transversion rate and CpG force, iii) a uniform rate (null model described above), and iv) a uniform rate and CpG force. The results of these additional tests are shown in Suppl. Fig. SI.9. The ANOVA F-test and p-values are shown in Table 2 and confirm that while the uniform rate and the transition versus transversion bias are not enough to separate the score distributions between observed and unobserved mutations, for the N ORF adding a CpG force gives a very clear separation, in the two cases, while for the S ORF we observe a still present but less marked separation. Finally we checked the consistency of our results at different times since our first analysis (dated 2020-04-22, see Suppl. Sec. SI.3).

[3]We consider here the virus codon bias, calculated on the Wuhan sequence, rather than the human codon bias, as SARS-CoV-2 is likely not to be in equilibrium with its host yet.

Motif	n	Position of CpG	Nb. of variants with CpG loss
CATTG**G**C**CG**	4	905	411
CTACCA**G**A**CG**	5	277	3
CGGAAT**G**T**CG**	5	953	9
CATATT**G**A**CG**	5	1074	1
CGCAGTGG**G**G**CG**	7	104	17
CTAACAAA**G**A**CG**	7	384	255
CTGGCAAT**G**G**CG**	7	642	5
CAAAACAAC**G**T**CG**	8	121	1
CGAGGACAA**G**G**CG**	8	213	3
CGTGGTGGT**G**A**CG**	8	294	4
CCCCGCATTA**CG**	-	47	16
AATAATACTG**CG**	-	149	3
CTTGGTTCAC**CG**	-	162	3
ATGCTGCAAT**CG**	-	471	14
AGAAGGGA**G**G**CG**	-	534	4
CACAAGCTTT**CG**	-	822	74
TTGCCCCCAG**CG**	-	930	7
AGCGCTTCAG**CG**	-	938	10
CAGCGTTCTT**CG**	-	945	20
GTCACACCTT**CG**	-	980	31
CCTTCGGGAA**CG**	-	986	13
CAAGCCTTAC**CG**	-	1148	61
CGGCA**G**A**CG**	4	829	0
CACGTA**G**T**CG**	5	571	0

Table 1: **Analysis of nucleotidic motifs preceding CpG in the N protein.** The top 10 lines show subsequences of SARS-CoV-2 (Wuhan, 26/12/2019) of the type CnxGxCG, where the spacer nx includes n = 4, 5, 7 or 8 nucleotides, for which the CpG dinucleotide was lost in one or more of the synonymous variants. These motifs were shown to be binding patterns for the ZAP protein in [9]; the dissociation constants were measured for repeated A spacers, with values (in μM) $K_d(4) = 0.33 \pm 0.05$, $K_d(5) = 0.49 \pm 0.10$, $K_d(7) = 0.12 \pm 0.04$, $K_d(8) = 0.64 \pm 0.14$, [9]. The next 12 lines show the other CpG lost through mutations and their 10 preceding nucleotides, which do not correspond to motifs tested in [9]. The last 2 lines show other subsequences in the N protein, known as binding motifs of ZAP from [9], but for which no loss of CpG is observed in the sequence data.

	F-test (S ORF)	F-test (N ORF)	p-value (S ORF)	p-value (N ORF)
Uniform bias	1	1	-	-
Uniform bias + CpG force	398	821	$< 10^{-50}$	$< 10^{-50}$
trt-trv bias	0.6	1	> 0.05	> 0.05
trt-trv bias + CpG force	295	627	$< 10^{-50}$	$< 10^{-50}$
Virus codon bias	2366	874	$< 10^{-50}$	$< 10^{-50}$
Virus codon bias + CpG force	2473	906	$< 10^{-50}$	$< 10^{-50}$

Table 2: **Anova F-test results.** Test results with the corresponding p-values for distributions of synonymous mutation scores (SMS) computed from observed and non-observed mutations in S ORF and N ORF and for different models of synonymous mutations (uniform bias, transition versus transversion bias (trt-trv), virus codon bias, with, and without CpG force). In the uniform case, we did not specify the p-value since all mutations have the same SMS.

3 Discussion

The present work reports analysis of dinucleotide motif usage, particularly CpG, in the early evolution of SARS-CoV-2 genomes. First, a comparative analysis with other genomes shows that the overall CpG force, and the associated CpG content are not as large as for highly pathogenic viruses in humans (such as H1N1, H5N1, Ebola and SARS and MERS in the Coronaviridae family). However, the CpG force, when computed locally, displays large fluctuations along the genome. This strong heterogeneity is compatible with viral recombination, in agreement with the hypothesis stated in [32]. The degree to which this heterogeneity in any way reflects zoonotic origins should be further worked out using phylogenetic analysis. In particular, the segment coding for the Spike protein has a much lower CpG force. The S protein has to bind ACE2 human receptors and TMPRSS2 [17, 30]. A fascinating reason that could explain the low CpG force on this coding region, is that it may come (at least in part) from other coronaviruses that better bound human entry receptors [32]. Other regions, in particular the region after the slippage site in ORF1ab and the initial and final part of the genome including the N ORF, are characterized by a larger density of CpG motifs (and corresponding CpG force), which are comparable to what is found in SARS and MERS viruses in the *Betacoronavirus* genus. Interestingly the initial and final part of the genome are implied in the full-genome and subgenomic viral replication. In particular, the coding region of the N protein and its RNA sequence, present in the 3' untranslated region (UTRs) of all SARS-CoV-2 subgenomic RNAs, has been shown in [25] to be the most abundant transcript in the cytoplasm. The high concentration of N transcripts in the cytoplasm could contribute to

a dysregulated innate immune response. SARS-CoV-2, due to its complex replication machinery, does not express its RNA at uniform concentration. A mechanism generating different densities of PAMPs being presented to the immune system at different points in the viral life cycle can affect immune recognition and regulation. The precise way this can contribute to immuno-pathologies associated with COVID-19 and how this is related to the cytokine signaling dysfunction associated with severe cases [16], need further experimental investigation.

Second, a first analysis of the evolution of synonymous mutations since the outbreak of COVID-19 shows that mutations lowering the number of CpG have taken place in regions with higher CpG content, at the 5' and 3' ends of the sequence, and in particular in the N protein coding region. The sequence motifs preceding the loci of the CpG removed by mutations match some of the strongly binding patterns of the ZAP protein [9]. Natural sequence evolution seems to be compatible for protein N with our model, in which synonymous mutations are driven by the virus codon bias and the CpG forces leading to a progressive loss in CpG. These losses are expected to lower the CpG forces, until they reach the equilibrium values in human host, as is seen in coronaviruses commonly circulating in human population [33]. More data, collected at an unprecedented pace [27, 28, 34], and on a longer evolutionary time are needed to confirm these hypothesis. Since the data collected are likely affected by relevant sampling biases, a more precise analysis of synonymous mutations could be carried out using the available phylogenetic reconstruction of viral evolution [35]. Nevertheless our results seem robust, since they are consistent both considering unique mutations and all collected synonymous variants. They coherently point to the presence of putative mutational hotspots in the viral evolution.

While the results presented here are preliminary due to the early genomics of this emerging virus, they point to interesting future directions to identifying the drivers of SARS-CoV-2 evolution and building better antiviral therapies. It would interesting to further model transmission and mutations (in the presence of a proofreading mechanism [36]) processes in SARS-CoV-2 to predict the time scale at which natural evolution driven by host mimicry would bring the virus to an equilibrium with its host [4, 5].

After our work was posted on the bioarxiv, R. Nchioua and colleagues have shown the importance of ZAP in controlling the response against SARS-CoV-2, see [37], by demonstrating that a knock-out of this protein increases SARS-CoV-2 replication. This finding supports our prediction that recognition of SARS-CoV-2 by ZAP imposes a significant fitness cost on the virus, as demonstrated by its early evolution to remove ZAP recognition motifs. Two other recent theoretical works [38, 39], corroborate our results showing that at the single nucleotide level there is a net prevalence of C→U synonymous mutations (the most common nucleotide mutation which may cause a CpG loss) in the early evolution of SARS-CoV-2. Moreover a recent analysis of the immune profile of patients with moderate and severe disease revealed an association between early, elevated cytokines and worse disease outcomes identifying a maladapted immune response profile associated with severe COVID-19 outcome [40].

Acknowledgments

We thank Nicolas Vabret for discussions and reading of the manuscript, Eddie Holmes and Marta Łuksza for helpful exchanges. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAIDs EpiCoV(TM) Database on which this research is based. This work was partially supported by the ANR-19 Decrypted CE30-0021-01 grant and by the Fondation de la Recherche Médicale (FRM) (ANR-Flash Covid, project SARS-Cov-2immunRNAs). B.G. was supported by National Institutes of Health grants 7R01AI081848-04, 1R01CA240924-01, a Stand Up to Cancer Lustgarten Foundation Convergence Dream Team Grant, and The Pershing Square Sohn Prize Mark Foundation Fellow supported by funding from The Mark Foundation for Cancer Research.

4 Material and Methods

4.1 CpG density versus local and global forces

The aim of this section is to give an overview of the methods used throughout this work, and to explain why for some analyses we used one method rather than the others.

We want to characterize the CpG content of a given genome. The different methods that we used to achieve this result, discussed with their usage cases and their limits, are the following:

- A first possibility is to simply count the number of dinucleotide motifs (or to compute their density), along the whole genome. This simple count can be useful to see if there is an evolution of the motif number over time, or to study local fluctuations along a sequence to identify regions in which a motif is abundant or scarce, but it is not suitable to make comparisons among viruses of different families, mainly because of the different length and usage biases of viral genomes. However, since we focused mostly on the *Coronaviridae* family, these differences are not so important, and indeed we can see in Suppl. Fig. SI.3 that some of our results are also apparent from the motif density analysis.
- The force defines the abundance or scarcity of a motif given its expected usage based on the nucleotide bias. It can be computed on the whole or part of the genome. In this work we always use, to calculate the force, the human nucleotide bias as reference bias. In the following we detail the force calculation. An important remark is that the force is directly related to the relative abundance $\frac{f(ab)}{f(a)f(b)}$ of a motif of nucleotides a and b with respect to their bias ($f(ab)$ is the frequency of the motif and $f(a)$, $f(b)$ the reference biases of nucleotides a and b) [1], see Eq. 4 and SI Sec. SI.4. This method gives results qualitatively similar to the CpG-density methods for the local analysis. However it allows, in principle, also comparisons among more different genomes, as was done in Suppl. Fig. SI.1.
- As shown below, the calculation of the force can be extended to constrain variability of nucleotidic sequences at fixed codons, and using as reference bias the codon bias. This way of computing forces takes into account the fact that the virus has to code for certain specific proteins in its genome. We used here the human codon bias and the SARS-CoV-2 codon bias (the latter only for the computation of SMS since the virus is not in equilibrium with the human host) as references to compute this force with codon constraints. Calculating forces at fixed codon usage allows us to confirm also in this framework the identification of high- and low-CpG force regions, and it was crucial to investigate the dynamics of the synonymous mutations in viral evolution.

4.2 Force-based model

The model at the core of many of the analyses made here is taken from [4]. Here we briefly review the model, together with its simplified version which does not take into account the codon constraint. Let us start from the latter. Given a motif m and a sequence $s_0 = \{s_1, \ldots, s_N\}$ of length N, we consider the ensemble of all sequences with length N, which we denote with $\mathcal{S}$, and we suppose the probability of observing s out of this ensemble to be

$$p(s) = \frac{1}{Z}\left(\prod_{i=1}^{N} f(s_i)\right) e^{xN_m(s)}. \tag{1}$$

Here, $f(s_i)$ is the nucleotide bias, that is the probability of the i-th nucleotide being s_i (for example, we always used in this work the human frequency of nucleotides as $f(s_i)$), x is the force we want to compute, and N_m is the number of times the motif m appears in the sequence. Z is the normalization constant, that is

$$Z = \sum_{s\in\mathcal{S}}\left(\prod_{i=1}^{N} f(s_i)\right) e^{xN_m(s)}. \tag{2}$$

Therefore the force x is a parameter which quantifies the lack (if negative) or abundance (if positive) of occurrences of m with respect to the number of occurrences due to the local probabilities $f(s_i)$. We can fix x by requiring that the number of motifs in the observed sequence, $N_m(s_0) = n_0$, is equal to the average number of motifs computed through the model, $\langle n\rangle$, that is

$$\langle n\rangle = \sum_{s\in\mathcal{S}}\left(\prod_{i=1}^{N} f(s_i)\right) N_m(s)\, e^{xN_m(s)}. \tag{3}$$

Notice that this is exactly equivalent to the request that x maximizes the probability of having observed s_0.

Let us focus now on the specific case of a dinucleotide motif, that is our motif m consists of the pair ab, where a and b are two consecutive nucleotides (for example, $a = C$ and $b = G$ for the CpG motif). In this case, within an approximation discussed in the SI, Suppl. Sec. SI.4, the force computed as above turns out to be the logarithm of the relative abundance index, that is

$$x \simeq \log\left(\frac{f(ab)}{f(a)f(b)}\right), \tag{4}$$

where $f(ab)$ is the number of motifs ab divided by the total length of the sequence N. In Suppl. Fig. SI.11 we tested the accuracy of this approximation in our specific case.

Our model can be improved to take into account the fact that not all the possible sequences with length N can be observed if the genome is coding for one (or more) protein(s). If we restrict the ensemble of sequences to those coding for the same protein, we obtain the model with the codon constraints that we used for several of our analyses here. In this case, we write each sequence s as a series of codons, and its probability is defined as

$$p(s) = \frac{1}{Z}\left(\prod_{i=1}^{N/3} f(c_i)\right) e^{xN_m(s)}, \tag{5}$$

where now the bias takes the form of a codon usage bias, and the normalization constant Z changes accordingly into a sum over all possible synonymous sequences. The force x can be computed, analogously with the procedure for the simpler case, by requiring that the number of motifs observed in s_0 is equal to the model average (although this creates some technical difficulties, which have been overcome in [4]).

4.3 Definition of synonymous mutation score

We use the model introduced above in Eq. 1 to assign a score, which we call synonymous mutation score (SMS), to each possible synonymous mutation of a reference sequence. We consider a system evolving for a small time scale, and a mutation which changes the i-th codon c_i into a synonymous c_i'. The transition probability, that is the probability of observing the mutation, for such evolution can be decomposed in the product of two evolution operators: The first $T(N_{CG} \to N_{CG}')$ representing the change in the number of CpG motifs in the mutated sequence, and the second $T(c_i \to c_i')$ representing the gain in mutating the particular codon in position i.

The first operator can be computed from the dynamical equation introduced in [4] for the evolution of the CpG number N_{CG} of a sequence under a initial force x through a equilibrium force x_{eq}:

$$\tau \frac{dN_{CG}}{dt} = (x_{eq} - x). \tag{6}$$

The equilibrium force is the force computed on a viral strain which is supposed to be to the equilibrium with the human innate immune system, because it has evolved in the human host since a long time. Eq. 6 was used in [4] to describe the evolution of the CpG numpber in H1N1, taking as the equilibrium force the one of the Influenza B strain. In analogy with this approach we take here as f_{eq} the average force calculated for the given segment or coding region on the seasonal hCoVs (that is hCoV-229E, hCoV-NL63, hCoV-HKU1, hCoV-OC43). τ is a parameter determining the characteristic time scale for synonymous mutations. Based on Eq. (6) we define the transition operator for CpG number as

$$T(N_{CG} \to N_{CG}') \propto e^{(x_{eq}-x)\Delta N_{CG}}, \tag{7}$$

where $\Delta N_{CG} = N_{CG}' - N_{CG}$. Notice that for all the synonymous mutations leaving unchanged the CpG number the above operator is one. The codon mutational operator is

$$T(c_i \to c_i') \propto \left(\frac{f(c_i')}{f(c_i)}\right), \tag{8}$$

where $f(c_i)$ is the frequency of codon c_i from the chosen codon usage bias. Putting together these two terms allows us to estimate how likely a specific synonymous mutation is to happen. The synonymous mutation score (SMS) accompanying a mutation is defined as the logarithm of this quantity,

$$\text{SMS} = (x_{eq} - x)\Delta N_{CG} + \log\left(\frac{f(c_i')}{f(c_i)}\right) \ . \tag{9}$$

4.4 Uniform codon usage bias with transversion penalties

It is well known that transversions (i. e. mutations of purines in pyrimidines and vice-versa) are suppressed with respect to transitions (i. e. mutations of purines in purines or pyrimidines in pyrimidines).
We introduce here a simple way to account for penalties for transversions in the uniform codon bias. We suppose that a mutation with n transversions happens 4 times less than a mutation with $n-1$ transversions. Therefore, starting from a uniform probability of mutating a codon c into a synonymous codon c', we insert the transvertion penalty and obtain that this probability becomes

$$p(c \to c') = \frac{1}{\mathcal{N}} \left(\frac{1}{4}\right)^{t(c,c')}, \tag{10}$$

where $t(c, c')$ is the number of transvertions needed to mutate c into c'. Here $\mathcal{N}$ is a normalization constant, such that

$$\sum_{c'} p(c \to c') = 1, \tag{11}$$

the sum runs over the synonymous of c (without including c). Then we can define, for a fixed set of synonym codons, a transition matrix

$$T_{c,c'} = p(c \to c'). \tag{12}$$

The codon usage bias, for synonyms mutations with transversion penalties, is the stationary probability distribution of the Markov chain having the matrix defined in Eq. 12 as transition operator. This stationary distribution is therefore given by the unique vector of probabilities $b(c)$ such that

$$b(c') = \sum_c b(c)\, T_{c,c'}. \tag{13}$$

By solving this set of equations, together with the requests that $\sum_c b(c) = 1$, for each set of synonymous codons, we obtain our codon usage bias. We have repeated this calculation for all amino acids.

We used this modified bias to perform ANOVA F-tests together with other codon usage biases, see Table 2 and Suppl. Fig. SI.9.

4.5 Data Analysis

SARS-CoV-2 sequences are taken from GISAID [28]. We collected each sequence present in the database on 2020-08-23 (the most recent sequence was collected on 2020-07-20). Before any of our analyses, we discarded all the sequences where one or more nucleotides were wrongly read. This left us with 23762 SARS-CoV-2 sequences. To obtain Fig. 1 we considered, in addition to the SARS-CoV-2 sequences are taken from GISAID, other *Alphacoronavirus* and *Betacoronavirus* sequences (whole genomes and genes) which have been obtained from VIPR [27]. The pre-processing consisted again of discarding all the sequences with one error or more. After this process we collected 341 SARS, 48 MERS, 20 hCoV-229E, 48 hCoV-NL63, 14 hCoV-HKU1, 124 hCoV-NL63, 166 bat-CoVs and 5 pangolin-CoVs whole genomes. For Fig. 1b we used the largest number possible of sequences, up to a maximum of 100. For Fig. 1a (viral sequences) and Fig. 1c we chose a single sequence for each species. However, we checked that the result is qualitatively the same if we use other sequences from the same species for human coronaviruses.
To obtain the plots in Fig. 2 and Fig. 3, we considered as reference SARS-CoV-2 sequence the one which has been collected on 26-12-2019 (ID: EPI_ISL_406798). In Fig. 2a the SARS-CoV-2 sequence has been processed to ensure the correct reading frame. Therefore the ORF1ab gene is read in the standard frame up to the ribosomal shifting site,

and it is read in the shifted frame from that site up to the end of the polyprotein. Moreover, the small non-coding parts between successive proteins have been cut, resulting in a loss of 634 nucleotides. To produce the bar plots in Figs. 2c and 2b we collected genes data on VIPR. Then we discarded as usual all the sequences with one or more errors, and we computed for each gene an average of up to 20 different sequences (coming from the same species). For some structural proteins we did not find 20 different genes but in any case the standard deviation of the averages of Figs. 2c and 2b is smaller than 0.02 (and, for most of the viruses, much smaller). In particular, we used 20 sequences of SARS-CoV-2, MERS, hCoV-NL63, hCoV-OC43 proteins, 14 sequences for hCoV-229E, 13 for hCoV-HKU1 and 4 for SARS. More detailed information about the genomes used in this work are given in Suppl. Sec. SI.1.

The code used to compute forces, both in the coding and non-coding cases, is publicly available at https://github.com/adigioacchino/dinucleotide_forces.

References

[1] Samuel Karlin and Jan Mrázek. Compositional differences within and between eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 94(19):10227–10232, 1997. 1, 2, 9, 21

[2] S Karlin, W Doerfler, and LR Cardon. Why is cpg suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *Journal of virology*, 68(5):2889–2897, 1994. 1

[3] Xiaofei Cheng, Nasar Virk, Wei Chen, Shuqin Ji, Shuxian Ji, Yuqiang Sun, and Xiaoyun Wu. Cpg usage in rna viruses: data and hypotheses. *PloS one*, 8(9), 2013. 1

[4] Benjamin D Greenbaum, Simona Cocco, Arnold J Levine, and Rémi Monasson. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proceedings of the National Academy of Sciences*, 111(13):5054–5059, 2014. 1, 2, 8, 9, 10

[5] Benjamin D Greenbaum, Arnold J Levine, Gyan Bhanot, and Raul Rabadan. Patterns of evolution and host gene mimicry in influenza and other rna viruses. *PLoS pathogens*, 4(6), 2008. 1, 8

[6] Sonia Jimenez-Baranda, Benjamin Greenbaum, Olivier Manches, Jesse Handler, Raúl Rabadán, Arnold Levine, and Nina Bhardwaj. Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. *Journal of virology*, 85(8):3893–3904, 2011. 1

[7] Nicolas Vabret, Nina Bhardwaj, and Benjamin D Greenbaum. Sequence-specific sensing of nucleic acids. *Trends in immunology*, 38(1):53–65, 2017. 1, 2

[8] Matthew A Takata, Daniel Gonçalves-Carneiro, Trinity M Zang, Steven J Soll, Ashley York, Daniel Blanco-Melo, and Paul D Bieniasz. Cg dinucleotide suppression enables antiviral defence targeting non-self rna. *Nature*, 550(7674):124–127, 2017. 1

[9] Xiu Luo, Xinlu Wang, Yina Gao, Jingpeng Zhu, Songqing Liu, Guangxia Gao, and Pu Gao. Molecular mechanism of rna recognition by zinc-finger antiviral protein. *Cell Reports*, 30(1):46–52, 2020. 1, 6, 7, 8

[10] Jennifer L Meagher, Matthew Takata, Daniel Gonçalves-Carneiro, Sarah C Keane, Antoine Rebendenne, Heley Ong, Victoria K Orr, Margaret R MacDonald, Jeanne A Stuckey, Paul D Bieniasz, et al. Structure of the zinc-finger antiviral protein in complex with rna reveals a mechanism for selective targeting of cg-rich viral sequences. *Proceedings of the National Academy of Sciences*, 116(48):24303–24309, 2019. 1

[11] Toshio Hirano and Murakami Masaaki. Covid-19: a new virus, but an old cytokine release syndrome. *Immunity*, DOI: 10.1016/j.immuni.2020.04.003, 2020. 1, 2

[12] Puja Mehta, Daniel F McAuley, Michael Brown, Emilie Sanchez, Rachel S Tattersall, and Jessica J Manson. Covid-19: consider cytokine storm syndromes and immunosuppression. *The Lancet*, 2020. 1

[13] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, 2020. 1

[14] Eveline Kindler and Volker Thiel. Sars-cov and ifn: too little, too late. *Cell host & microbe*, 19(2):139–141, 2016. 1

[15] Jerome Hadjadj, Nader Yatim, Laura Barnabei, Aurelien Corneau, Jeremy Boussier, Helene Pere, Bruno Charbit, Vincent Bondet, Camille Chenevier-Gobeaux, Paul Breillat, et al. Impaired type i interferon activity and exacerbated inflammatory responses in severe covid-19 patients. *medRxiv*, 2020. 1

[16] Dina Ragab, Haitham Salah Eldin, Mohamed Taeimah, Rasha Khattab, and Ramy Salem. The covid-19 cytokine storm; what we know so far. *Frontiers in Immunology*, 11:1446, 2020. 1, 8

[17] Nicolas Vabret, Graham J. Britton, Conor Gruber, Samarth Hegde, Joel Kim, Maria Kuksin, Rachel Levantovsky, Louise Malle, Alvaro Moreira, Matthew D. Park, et al. Immunology of covid-19: Current state of the science. *Immunity*, 52(6):910 – 941, 2020. 1, 2, 7

[18] Xiaoling Xu, Mingfeng Han, Tiantian Li, Wei Sun, Dongsheng Wang, Binqing Fu, Yonggang Zhou, Xiaohu Zheng, Yun Yang, Xiuyong Li, et al. Effective treatment of severe covid-19 patients with tocilizumab. *Proceedings of the National Academy of Sciences*, 117(20):10970–10975, 2020. 1

[19] Fatima Amanat and Florian Krammer. Sars-cov-2 vaccines: status report. *Immunity*, 2020. 1, 2

[20] Jacob Kames, David Dillon Holcomb, Ofer Kimchi, Michael DiCuccio, Nobuko Hamasaki-Katagiri, Tony Wang, Anton A Komar, Aikaterini Alexaki, and Chava Kimchi-Sarfaty. Sequence analysis of sars-cov-2 genome reveals features important for vaccine design. *BioRxiv*, 2020. 1

[21] Antoine Tanne, Luciana R. Muniz, Anna Puzio-Kuter, Katerina I. Leonova, Andrei V. Gudkov, David T. Ting, Rémi Monasson, Simona Cocco, Arnold J. Levine, Nina Bhardwaj, et al. Distinguishing the immunostimulatory properties of noncoding rnas expressed in cancer cells. *Proceedings of the National Academy of Sciences*, 112(49):15154–15159, 2015. 2

[22] Elizabeth D Howe and Jun S Song. Categorical spectral analysis of periodicity in human and viral genomes. *Nucleic acids research*, 41(3):1395–1405, 2013. 2

[23] Zhiqi Song, Yanfeng Xu, Linlin Bao, Ling Zhang, Pin Yu, Yajin Qu, Hua Zhu, Wenjie Zhao, Yunlin Han, and Chuan Qin. From sars to mers, thrusting coronaviruses into the spotlight. *Viruses*, 11(1):59, 2019. 2

[24] Michael MC Lai and David Cavanagh. The molecular biology of coronaviruses. In *Advances in virus research*, volume 48, pages 1–100. Elsevier, 1997. 2

[25] Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. The architecture of sars-cov-2 transcriptome. *Bioinformatics*, 29:15–21, 2020. 2, 7

[26] Jie Zhou, Hin Chu, Cun Li, Bosco Ho-Yin Wong, Zhong-Shan Cheng, Vincent Kwok-Man Poon, Tianhao Sun, Candy Choi-Yi Lau, Kenneth Kak-Yuen Wong, Jimmy Yu-Wai Chan, et al. Active replication of middle east respiratory syndrome coronavirus and aberrant induction of inflammatory cytokines and chemokines in human macrophages: implications for pathogenesis. *The Journal of infectious diseases*, 209(9):1331–1342, 2014. 2

[27] Brett E. Pickett, Eva L. Sadat, Yun Zhang, Jyothi M. Noronha, R. Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, Liwei Zhou, et al. Vipr: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research*, 40(D1):D593–D598, 2012. 3, 4, 8, 10

[28] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017. 3, 4, 5, 6, 8, 10

[29] Xintian Xu, Ping Chen, Jingfang Wang, Jiannan Feng, Hui Zhou, Xuan Li, Wu Zhong, and Pei Hao. Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life Sciences*, 63(3):457–460, 2020. 2

[30] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020. 2, 7

[31] David W Collins and Thomas H Jukes. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, 20(3):386–396, 1994. 6

[32] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. The proximal origin of sars-cov-2. *Nature medicine*, 26(4):450–452, 2020. 7, 13

[33] Sahar Abdul-Rasool and Burtram C Fielding. Understanding human coronavirus hcov-nl63. *The open virology journal*, 4:76, 2010. 8

[34] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018. 8

[35] Ana S. Gonzalez-Reiche, Matthew M. Hernandez, Mitchell J. Sullivan, Brianne Ciferri, Hala Alshammary, Ajay Obla, Shelcie Fabre, Giulio Kleiner, Jose Polanco, Zenab Khan, et al. Introductions and early spread of sars-cov-2 in the new york city area. *Science*, 369(6501):297–301, 2020. 8

[36] Mark R. Denison, Rachel L. Graham, Eric F. Donaldson, Lance D. Eckerle, and Ralph S. Baric. Coronaviruses. *RNA Biology*, 8(2):270–279, 2011. PMID: 21593585. 8

[37] Rayhane Nchioua, Dorota Kmiec, Janis Mueller, Carina Conzelmann, Ruediger Gross, Chad M Swanson, Stuart Neil, Steffen Stenger, Daniel Sauter, Jan Muench, et al. The zinc finger antiviral protein restricts sars-cov-2. *bioRxiv*, 2020. 8

[38] Yulong Wei, Jordan R. Silke, Parisa Aris, and Xuhua Xia. Coronavirus genomes carry the signatures of their habitats. *bioRxiv*, 2020. 8

[39] Mukhtar Sadykov, Tobias Mourier, Qingtian Guan, and Arnab Pain. Cytosine deamination in sars-cov-2 leads to progressive cpg depletion. *bioRxiv*, 2020. 8

[40] Carolina Lucas, Patrick Wong, Jon Klein, Tiago BR Castro, Julio Silva, Maria Sundaram, Mallory K Ellingson, Tianyang Mao, Ji Eun Oh, Benjamin Israelow, et al. Longitudinal analyses reveal immunological misfiring in severe covid-19. *Nature*, 2020. 8

[41] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21, 2007. 13

The heterogeneous landscape and early evolution of pathogen-associated CpG dinucleotides in SARS-CoV-2

Supplementary Information

Andrea Di Gioacchino, Petr Šulc, Anastassia V. Komarova, Benjamin D. Greenbaum, Rémi Monasson, Simona Cocco

SI.1 Genomes analyzed

Here we report some additional information about the genomes used in this work. The SARS-CoV-2 sequence used in Figs. 1a and 1c has GISAID accession ID: EPI_ISL_420793. The other GenBank accession numbers for the specific genomes used in Figs. 1a and 1c are: AY427439 (SARS), NC_038294 (MERS), MF542265 (hCoV-229E), JX524171 (hCoV-NL63), KT779555 (hCoV-HKU1) and KF923918 (hCoV-OC43). For these figures, we choose the bat and pangolin sequences closest to the SARS-CoV-2 points in Fig. 1b (these two sequences are also known to be very similar to the SARS-CoV-2 genome from previous works [32]). These sequences have GISAID accession IDs EPI_ISL_402131 (bat coronavirus sequence known with the name RaTG13) and EPI_ISL_410721 (pangolin coronavirus sequence collected in 2019 in Guangdong).
The SARS-CoV-2 reference sequence which has been collected on 26-12-2019 has GISAID accession ID: EPI_ISL_406798. This sequence has been used in Figs. 2a and 3a. For Figs. 2b and 2c we used specific sequences, with the following GenBank accession numbers: MT300186:28249-29508 (SARS-CoV-2), AY291315:28120-29388 (SARS), NC_038294:28565-29800 (MERS) and KT779555:28281-29606 (hCoV-HKU1) for the N protein; MT300186:21538-25359 (SARS-CoV-2), AY291315:21492-25259 (SARS), NC_038294:21455-25516 (MERS) and KT779555:22903-26973 (hCoV-HKU1) for the S protein.

SI.2 Supplementary Figures

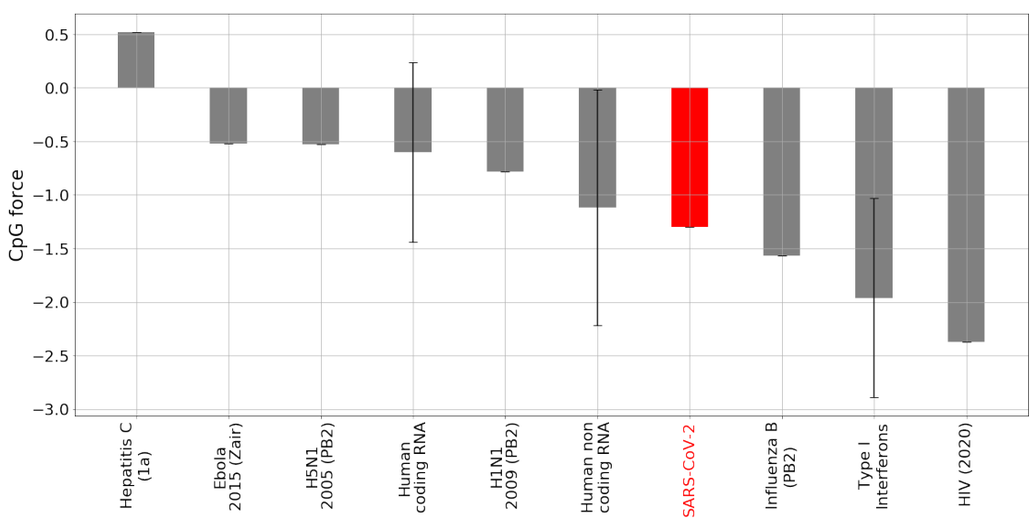


Figure SI.1: Comparison of global (whole-genome) CpG forces for SARS-CoV-2 genome with human coding DNAs and non-coding RNAs and other ssRNA viruses. The average value and standard deviation are shown for human RNAs. Type 1 Interferons transcripts are shown separately to highlight their very low CpG force. Data used: human cDNA and ncRNA as annotated in HG38 assembly, Type I interferon's cDNA as annotated in HG38. Viral ssRNAs were obtained from NCBI [41] Virus database (strains used: H5N1: A/Anhui/1/2005, H1N1: A/Aalborg/INS132/2009 and A/Brevig Mission/1/1918, Ebola: COD/1995/Kikwit-9510623, Influenza B: B/Massachusetts/07/2020, HIV: HK_JIDLNBL_S003, HCV: NC_004102).

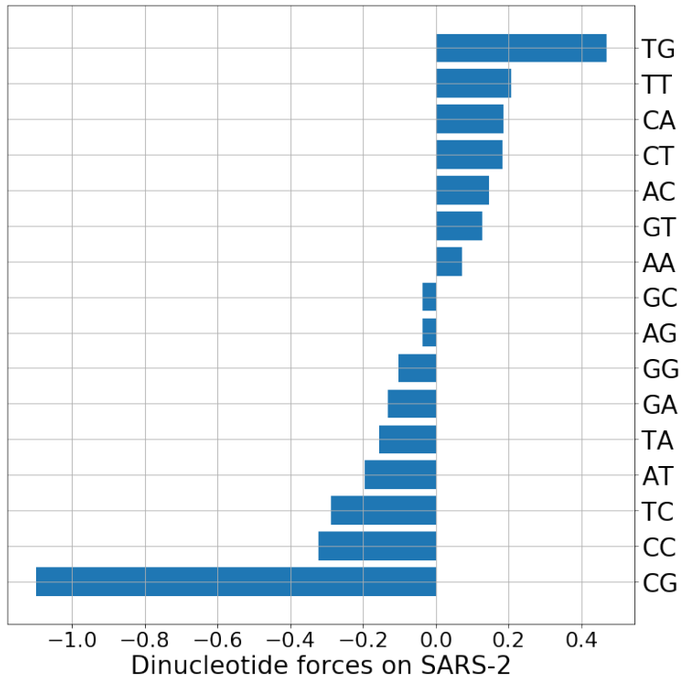


Figure SI.2: All dinucleotide forces computed on the whole SARS-CoV-2 genome, without any codon constraint. The CpG motif is the one with the largest force in absolute value, and the second one is TpG, which is one transition away from CpG.

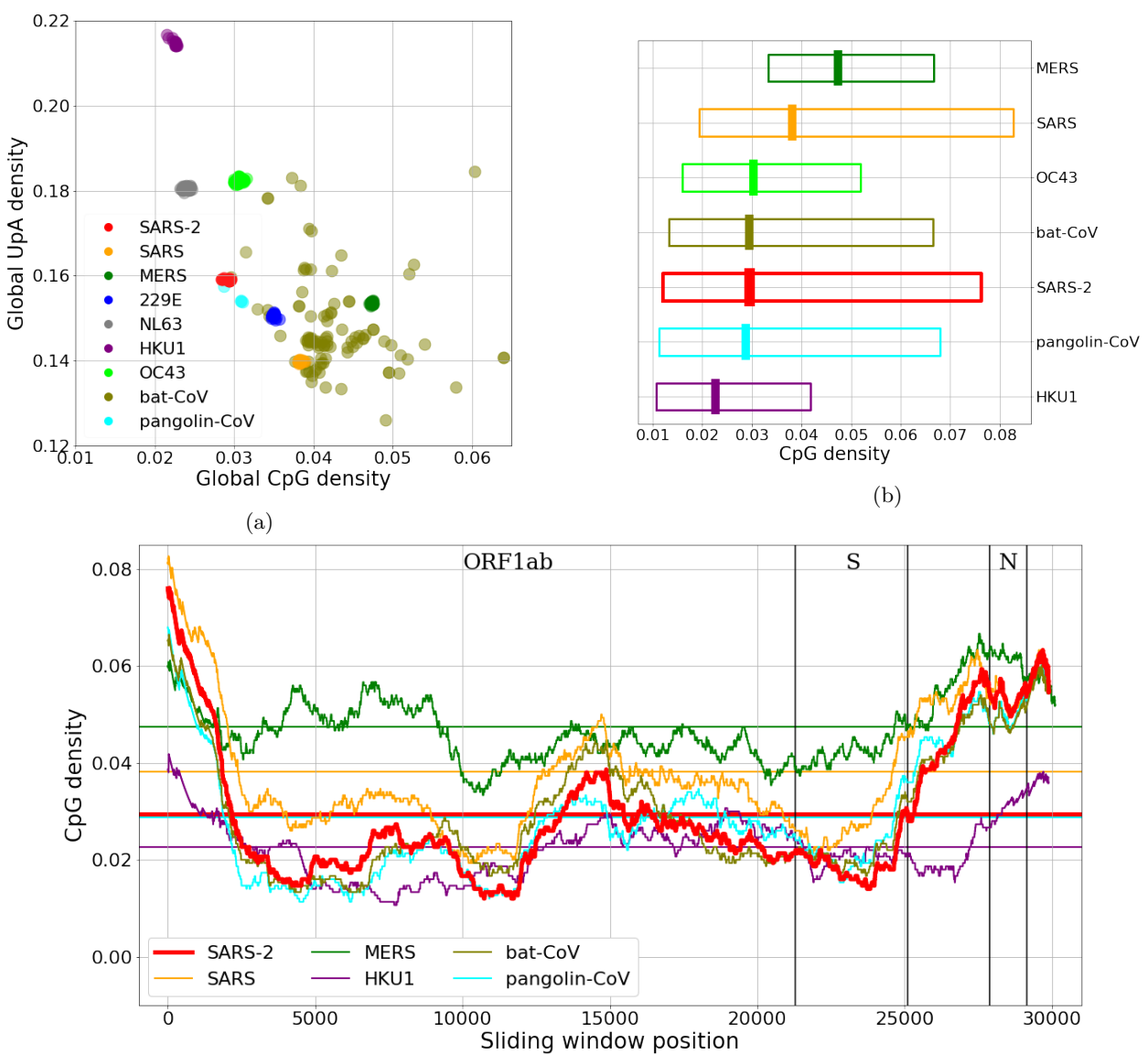
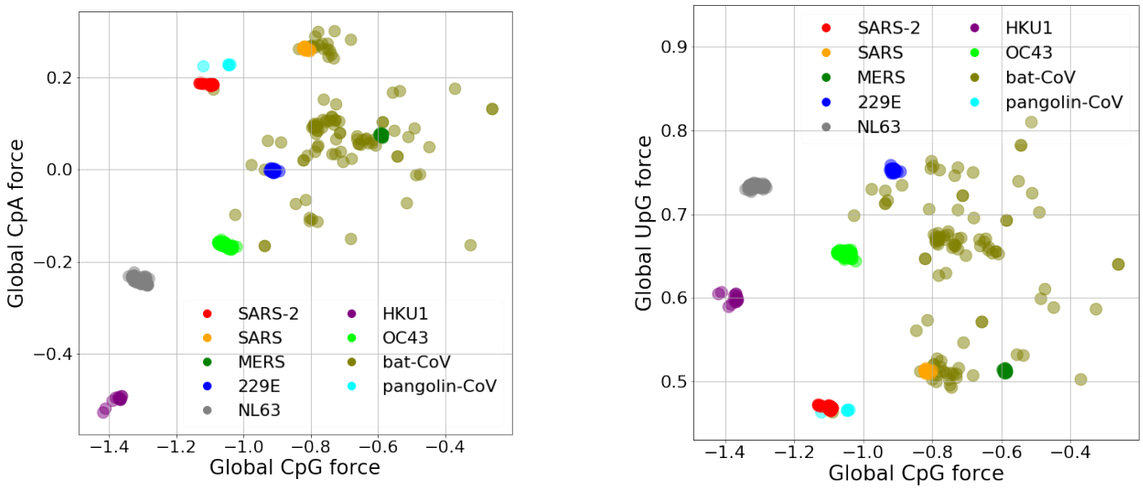
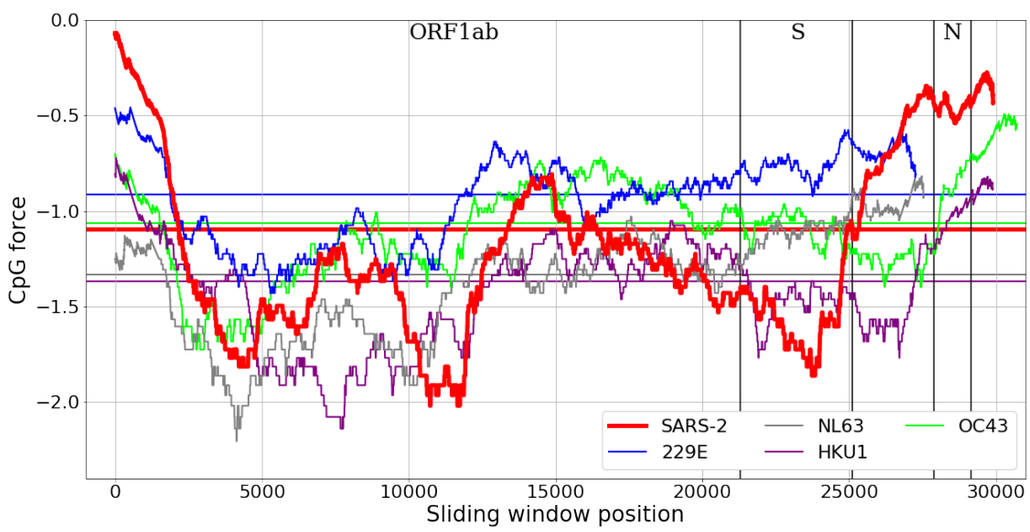


Figure SI.3: The same analysis performed in Fig. 1, but here we used CpG densities (CpG dinucleotides divided by the total number of dinucleotides), instead of CpG forces. The results obtained are qualitatively similar.

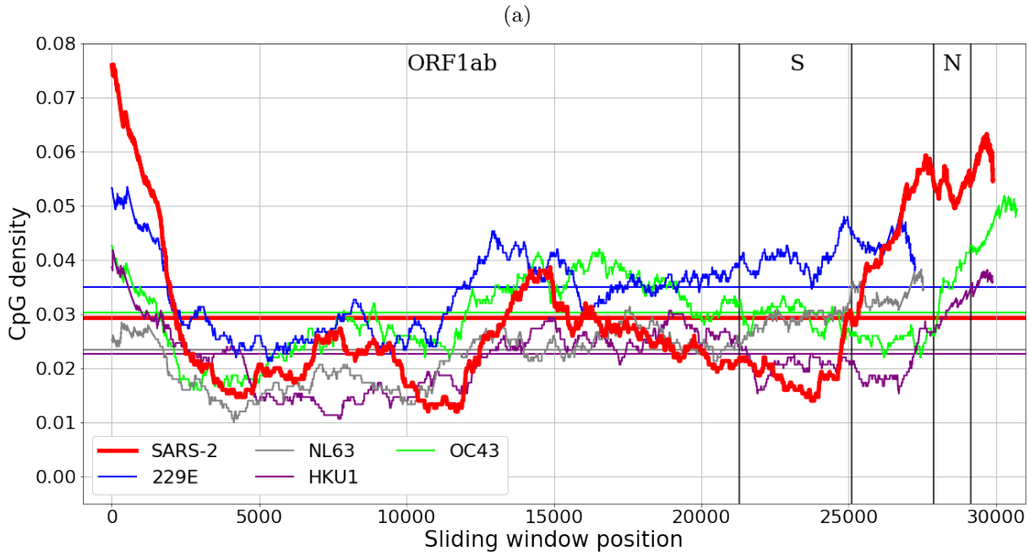


(a) (b)

Figure SI.4: CpG versus CpA (left) and UpG (right) forces. Differently from the CpG versus UpA case (Fig. 1b), no clear correlations (r^2 equal to 0.2 and 0.01, respectively, for the left and right plots) are found.

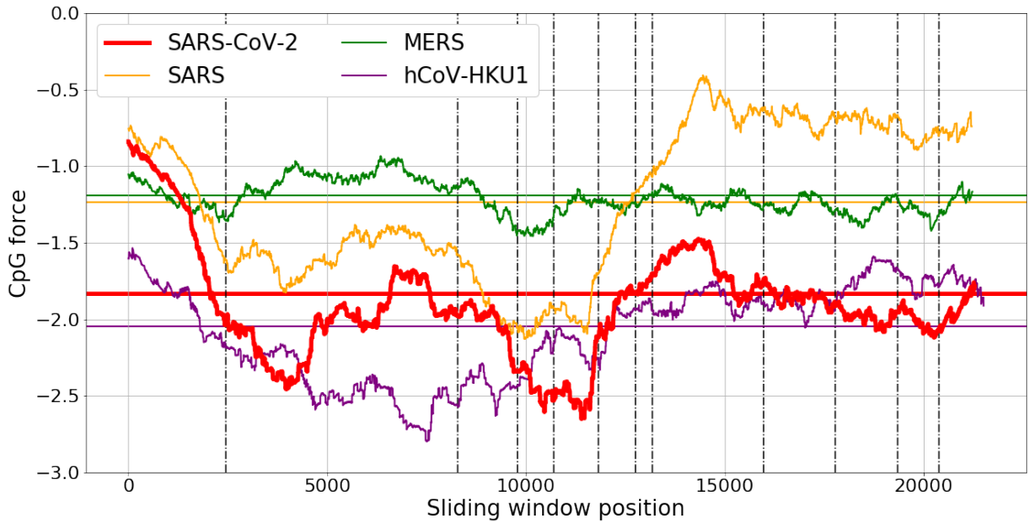


(a)

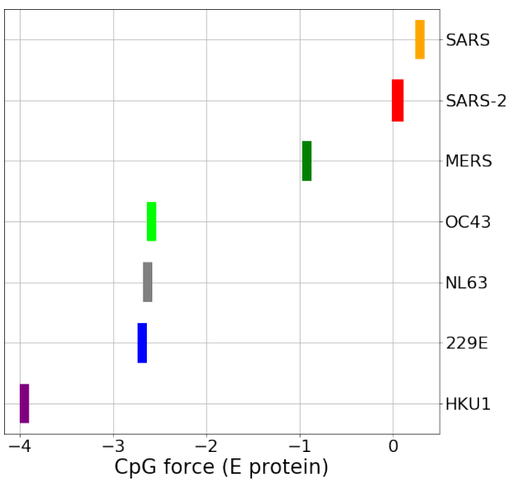


(b)

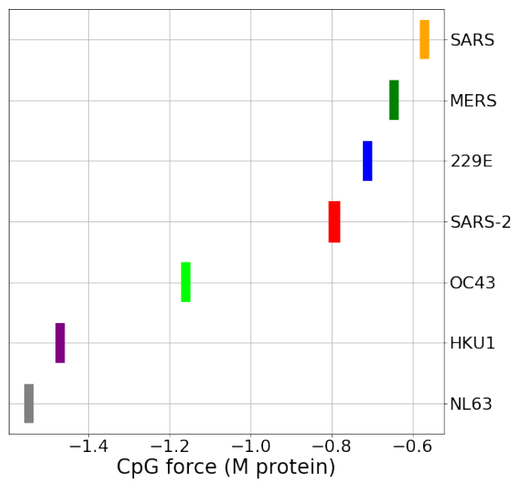
Figure SI.5: Supplement to Fig. 1 and Suppl. Fig. SI.3, where all the coronaviruses associated with circulating human strains are compared with SARS-CoV-2 in terms of CpG force (panel (a)) or number in fixed-length windows (panel (b)). Again, even though the final regions of the hCoVs has relatively high CpG force with respect to the other parts of their sequences, SARS-CoV-2 has a 3' CpG force peak well above the final region of hCoV virus.



(a)

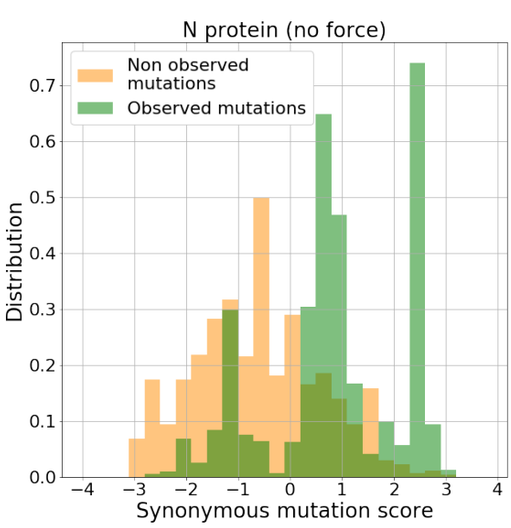


(b)

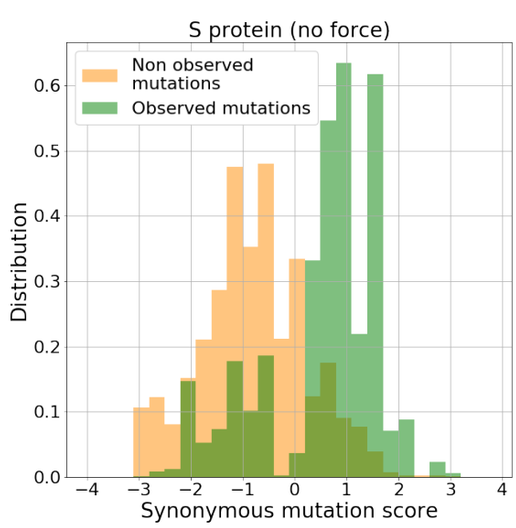


(c)

Figure SI.6: Extension of the comparison performed in Fig. 2. In panel (a) the genome coding for polyprotein ORF1ab is compared among several coronaviruses and in panels (b) and (c) the structural proteins E (envelope) and M (membrane) are considered. Notice that, due to the small size of proteins E and M, only one window is used so the boxes collapse in one line corresponding to the global CpG force on the protein, computed as an average over 4 to 20 viral genomes.

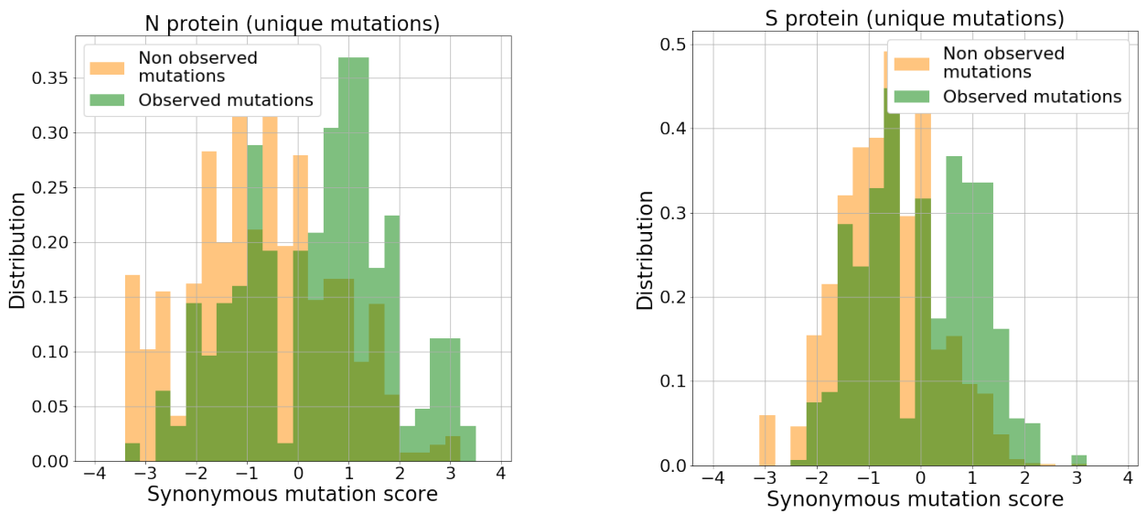


(a)



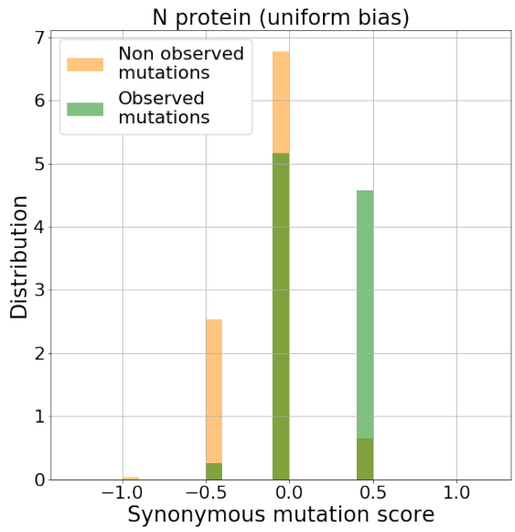
(b)

Figure SI.7: The same analysis discussed in Fig. 4 have been performed here computing the SMSs without force.

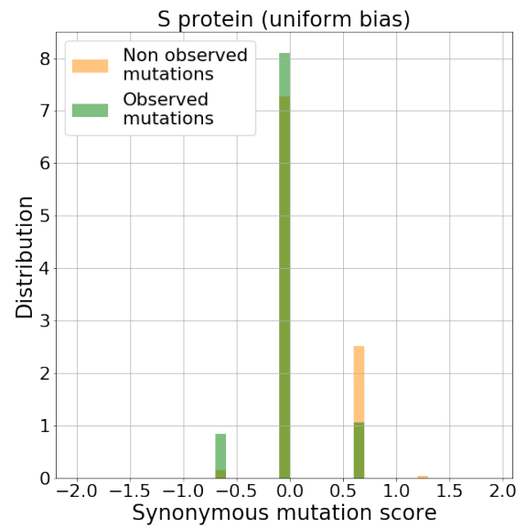


(a) (b)

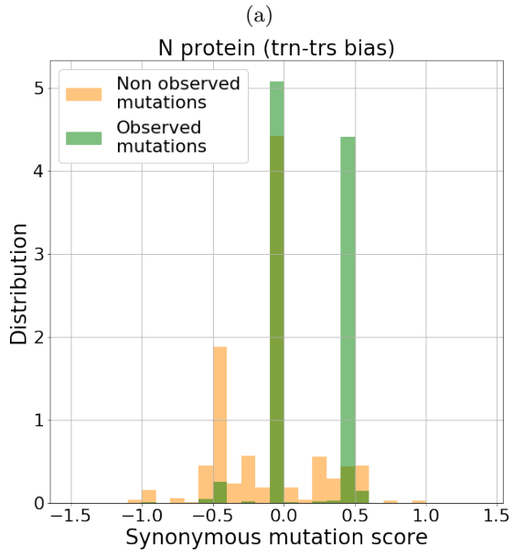
Figure SI.8: The same analysis discussed in Fig. 4 have been performed here computing the SMSs without taking into account the mutation multiplicities. We run the ANOVA F-test, obtaining $F = 100$ for the N ORF, and $F = 193$ for the S ORF, which in turn respectively correspond to p-values of $1 \cdot 10^{-22}$ and $1 \cdot 10^{-42}$, thus showing the robustness of our results.



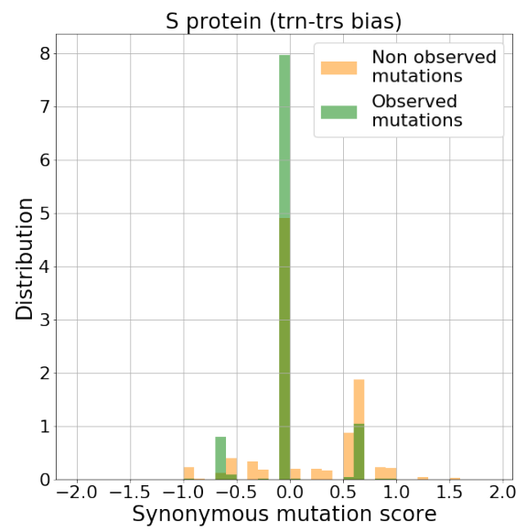
(a)



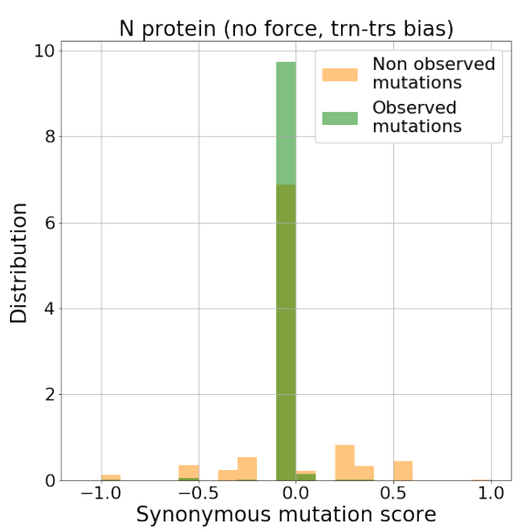
(b)



(c)



(d)



(e)

(f)

Figure SI.9: The same analysis discussed in Fig. 4 have been performed here computing the SMSs with different biases: in panels (a), (b), we used the uniform bias, while in panels (c), (d), (e) and (f) we also included a penalty for transversions with respect to translations.

SI.3 Results with previously collected sequences

SARS-CoV-2 genomes are collected at an unprecedented pace, with tens of genomes added daily. We updated our analysis several times since the first version of this manuscript has been posted on bioRxiv, and our results are consistently verified by mutations observed in recent genomes. To compare the analyses performed with the most recent genomes and those that we obtained previously, we collected all the SARS-CoV-2 sequences submitted to GISAID up to 2020-04-30 (last updated sequence collected on 2020-04-21), and we report here the analogous of Tables 1 and 2, and of Fig. 3 computed with those sequences, respectively as Tables 3, 4 and Fig. SI.10. We also show in Table 5 the comparison of the number of synonymous variants (total and restricted to CpG increasing or decreasing) observed in S and N ORFs with early and present-day genomes.

Motif	n	Position of CpG	Nb. of sequences with CpG loss
CATTG**G**C**CG**	4	905	2
CGGAAT**G**T**CG**	5	953	2
CATATT**G**A**CG**	5	1074	1
CGCAGTGG**G**G**CG**	7	104	1
CTAACAAA**G**A**CG**	7	384	59
CTGGCAAT**G**G**CG**	7	642	2
ATGCTGCAAT**CG**	-	471	1
CACAAGCTTT**CG**	-	822	22
GTCACACCTT**CG**	-	980	7
CAAGCCTTAC**CG**	-	1148	6
CGGCA**G**A**CG**	4	829	0
CTACCA**G**A**CG**	5	277	0
CACGTA**G**T**CG**	5	571	0
CAAAACAAC**G**T**CG**	8	121	0
CGAGGACAA**G**G**CG**	8	213	0
CGTGGTGGT**G**A**CG**	8	294	0

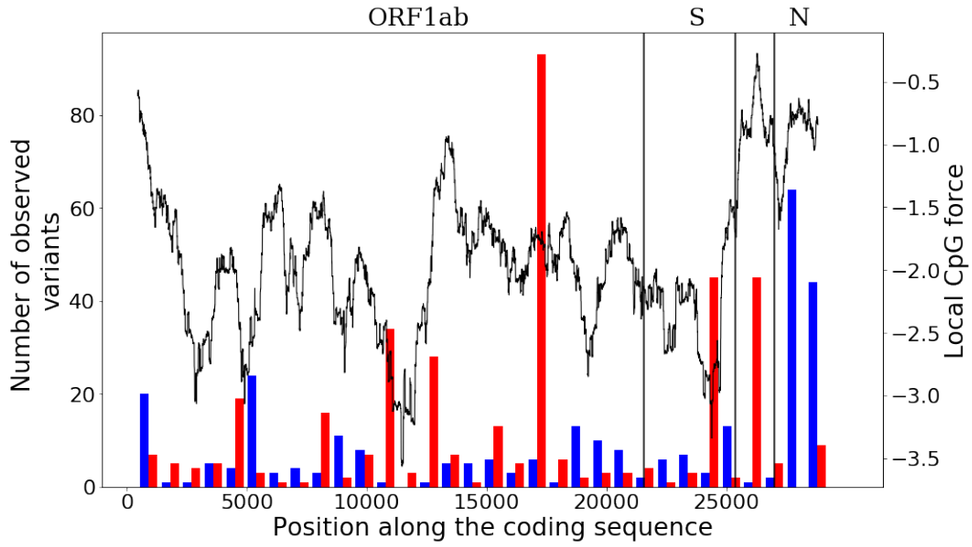
Table 3: Analogous of Table 1, computed with the sequences collected up to 2020-04-22. It is apparent that many of the ZAP biding motifs of the form CnxGxCG already mutated in some of the sequences obtained at the very beginning of the data collection.

	F-test (S ORF)	F-test (N ORF)	p-value (S ORF)	p-value (N ORF)
Uniform bias	1	1	-	-
Uniform bias + CpG force	34	193	$6 \cdot 10^{-9}$	$5 \cdot 10^{-41}$
trt-trv bias	0.3	$7 \cdot 10^{-4}$	> 0.05	> 0.05
trt-trv bias + CpG force	23	127	$2 \cdot 10^{-6}$	$4 \cdot 10^{-28}$
Virus codon bias	243	32	$< 10^{-50}$	$2 \cdot 10^{-8}$
Virus codon bias + CpG force	250	86	$< 10^{-50}$	$7 \cdot 10^{-20}$

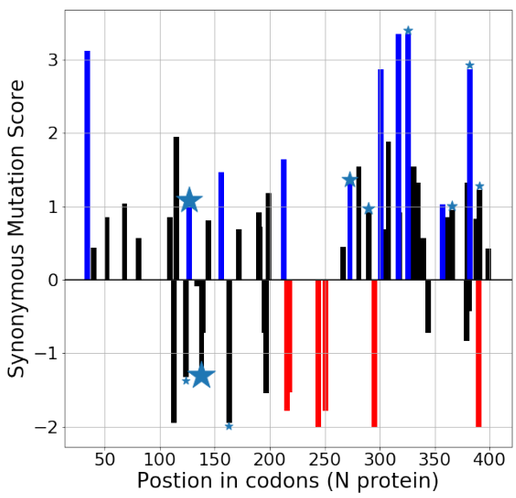
Table 4: Analogous of Table 2, computed with the sequences submitted to GISAID by 2020-04-30. Although the availability of fewer data lowers the F-test results most of the times (and therefore gives a higher p-value) with respect to Table 2, the qualitative results are very similar. For instance, it remains true that the score given through the transition-transversion bias alone cannot distinguish between the observed and non-observed mutations, while these two cases become distinguishable if the CpG force is added, especially for N ORF.

ORF (date)	# of syn. variants	# of CpG-decreasing variants	# of CpG-increasing variants
N (2020-04-30)	271	103	9
S (2020-04-30)	391	22	58
N (2020-06-02)	619	245	17
S (2020-06-02)	872	67	106

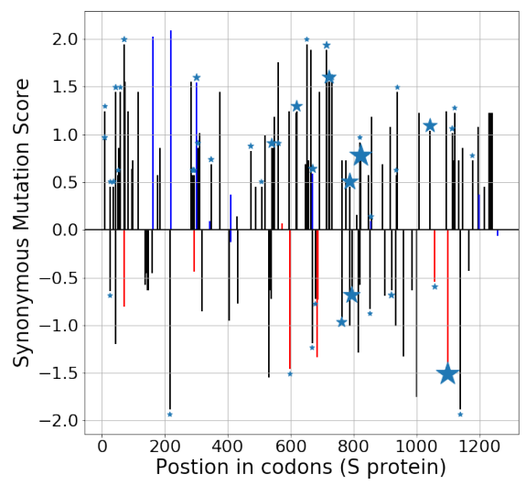
Table 5: Total number of synonymous variants and CpG decreasing/increasing variants observed for ORF N and S at two stages of the early evolution of the SARS-CoV-2 in contact with the human host. The dates refer to the upper limit for submission to GISAID of the sequences used.



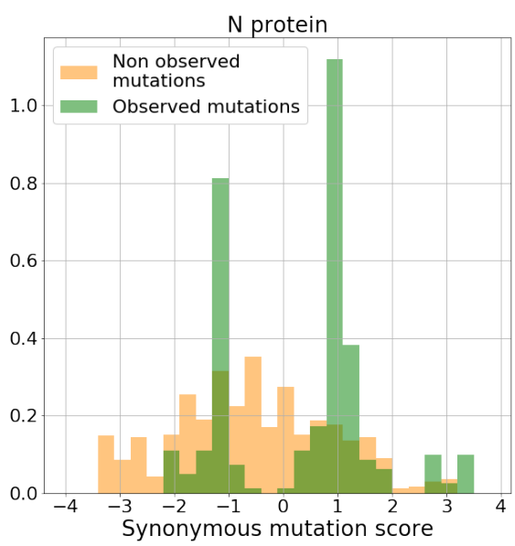
(a)



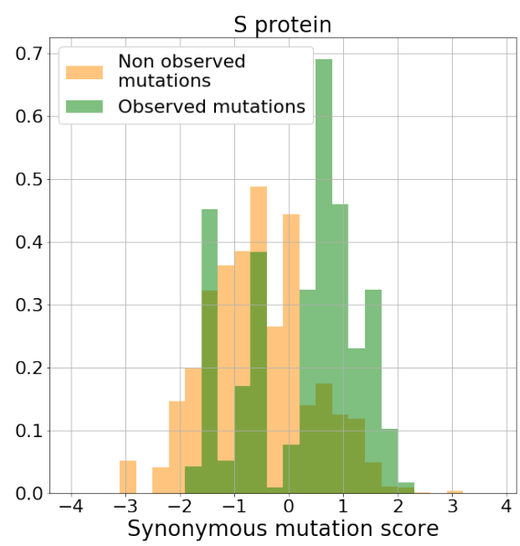
(b)



(c)



(d)



(e)

Figure SI.10: Analogous of Figs. 3 and 4, computed with the sequences submitted to GISAID by 2020-04-30. We observe less mutations than those in Figs. 3 and 4 due to the smaller database of sequences, but the results are qualitatively very similar.

SI.4 From CpG force to CpG relative abundance

We want to show in which limit the CpG force (without codon constraints) is equivalent to the relative dinucleotide abundance [1], Eq. 4. We start from the partition function:

$$Z = \sum_{s_1,\dots,s_N} \left(\prod_{i=1}^{N} f(s_i) \right) \prod_{i=1}^{N-1} e^{x\delta(s_i,a)\delta(s_{i+1},b)}, \tag{SI.1}$$

where δ denotes the Kroneker delta function. In the spirit of a cluster expansion, we write

$$e^{x\delta(s_i,a)\delta(s_{i+1},b)} = 1 + g_{i,i+1}, \tag{SI.2}$$

where

$$g_{i,i+1} = (e^x - 1)\,\delta(s_i,a)\,\delta(s_{i+1},b). \tag{SI.3}$$

Inserting back this into Eq. (SI.1), we obtain

$$\begin{aligned} Z &= \sum_{s_1,\dots,s_N} \left(\prod_{i=1}^{N} f(s_i) \right) \prod_{i=1}^{N-1} (1 + g_{i,i+1}) \\ &= \sum_{s_1,\dots,s_N} \left(\prod_{i=1}^{N} f(s_i) \right) \left[1 + \sum_i g_{i,i+1} + \sum_{i<j} g_{i,i+1}\, g_{j,j+1} + \dots \right]. \end{aligned} \tag{SI.4}$$

Now we can compute each term in the cluster expansion, and we get for the k-th term (for $a \neq b$, as in the CpG case)

$$\sum_{s_1,\dots,s_N} \left(\prod_{i=1}^{N} f(s_i) \right) \sum_{i_1<\dots<i_k} g_{i_1,i_1+1} \dots g_{i_k,i_k+1} = \binom{N-k}{k} \left((e^x - 1)\, f(a)\, f(b) \right)^k = \binom{N-k}{k} g^k. \tag{SI.5}$$

where we defined $g = (e^x - 1)\, f(a)\, f(b)$. Now we suppose $N = 2m$, that is N is even (however, we will consider soon the large-N limit, where this request is not necessary anymore). Therefore, we have

$$Z = \sum_{k=0}^{m} \binom{2m-k}{k} g^k = \frac{(1 + 2g - \sqrt{1+4g})^m (\sqrt{1+4g} - 1) + (1 + 2g + \sqrt{1+4g})^m (\sqrt{1+4g} + 1)}{2^{m+1}\sqrt{1+4g}}. \tag{SI.6}$$

To proceed further, we can consider the case where $g \ll 1$. This is a good approximation when $x \simeq 0$, and it is also fairly good as long as x is lower than 0, as in all the cases studied here. Under this hypothesis, we have

$$Z = (1+g)\, e^{(m-1)2g} \simeq e^{N(e^x-1)\, f(a)\, f(b)}, \tag{SI.7}$$

where in the last step we used also that $N \gg 1$. From this, by using that $\langle n \rangle = \partial_x \log Z$ and requesting $\langle n \rangle = n_0 = N f(ab)$, we obtain Eq. 4. Fig. SI.11 shows the correlation between the CpG force with the nucleotide bias and the CpG relative abundance.

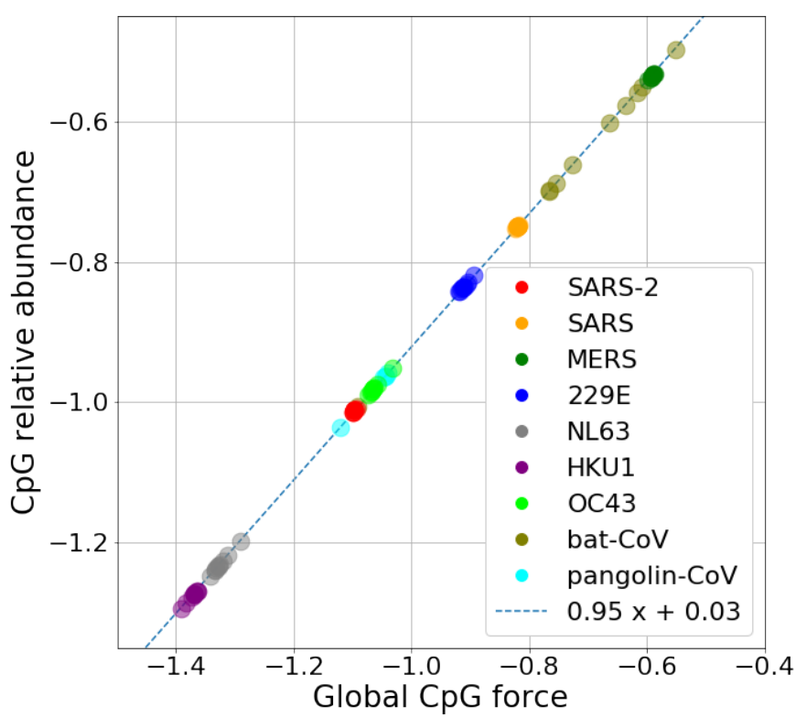


Figure SI.11: Comparison between the CpG force and the CpG relative abundance index. As discussed in Sec. SI.4, these two quantities are almost identical when the genome is long. To show that, here 10 different genomes for several coronavirus species are used to compute these two quantities, and the dashed blue line is a linear fit of the resulting points.