The heterogeneous landscape and early evolution of pathogen-associated CpG and UpA 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

May 6, 2020

Abstract

SARS-CoV-2 infection can lead to acute respiratory syndrome in patients, which can be due in part to dysregulated immune signalling. We analyze here the occurrences of CpG and UpA dinucleotides, which are putative pathogen-associated molecular patterns, along the viral sequence. Carrying out a comparative analysis with other ssRNA viruses and within the *Coronaviridae* family, we find the CpG content of SARS-CoV-2, while low compared to other betacoronaviruses, widely fluctuates along its primary sequence. While the CpG relative abundance and its associated CpG force parameter [1] are low for the spike protein (S) and comparable to circulating seasonal coronaviruses such as HKU1, they are much greater and comparable to SARS and MERS for the 3'-end of the viral genome. In particular, the nucleocapsid protein (N), whose transcripts are relatively abundant in the cytoplasm of infected cells and present in the 3'UTRs of all subgenomic RNA, has high CpG content. We speculate this dual nature of CpG content can confer to SARS-CoV-2 high ability to both enter the host and trigger pattern recognition receptors (PRRs) in different contexts. We then investigate the evolution of synonymous mutations since the outbreak of the COVID-19 pandemic. Using a new application of selective forces on dinucleotides to estimate context driven mutational processes, we find that synonymous mutations seem driven both by the viral codon bias and by the high value of the CpG force in the N protein, leading to a loss in CpG content. Sequence motifs preceding these CpG-loss-associated loci match recently identified binding patterns of the Zinc Finger anti-viral Protein (ZAP) 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

The innate immune system and RNA processing machinery recognizes "non-self" patterns and motifs in viruses that are rarely seen in the hosts they infect. These patterns can differ between species [2]. When a virus enters a new host, it can present pathogen-associated molecular patterns (PAMPs) in its new host 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 evolving 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 [3, 2, 4, 1]. 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 amino acid usage of codon bias [5, 1]. 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]. Various treatments to attenuate inflammatory responses have been proposed and are currently under analysis or being clinically tested [16]. 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 [17, 18] to better reflect human-genome features.

In this work we will use the framework developed in [1] 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. In [1] some of us introduced a parameter, which we call the selective force, that characterizes the deviation with respect to a null model in which the number of dinculeotides is the one statistically expected under a set of constraints. For instance, the selective force on CpG dinucledoties is zero if the number of CpG is exactly the one expected, is positive when it is larger than what is expected and negative otherwise. Such a force generalizes the dinucleotide relative abundance introduced in [2], not only when choosing the nucleotide bias as the null model, but also to a null model built on other features such as a fixed amino acid sequence and reference codon bias. 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 [19], to characterize the CpG evolution through synonymous mutations in H1N1 [1], and to characterize local and non-local forces on dinucleotides across RNA viruses [7].

We perform an analysis of the landscape of CpG and UpA motifs and associated selective forces in SARS-CoV-2 in comparison with other genomes in the coronavirus family in order to understand specific PAMPs associated features in the new SARS-CoV-2 strains (Sec. 2). We also focus on the heterogeneity of CpG motif usage along the SARS-CoV-2 genome (Sec. 2 and 3). 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. 4). The latter approach points out at hotspots where new mutations will likely attenuate the virus, while evolving in contact with the human host.

2 Forces on dinucleotides in viral genome

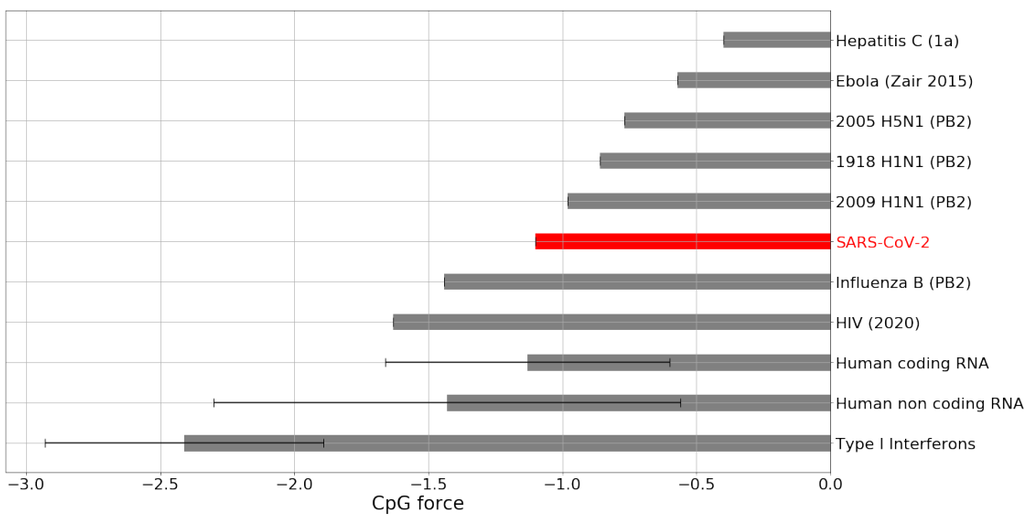


Figure 1: **Comparison of CpG forces for SARS-CoV-2 genome with human coding DNAs and non-coding RNAs and other ssRNA viruses**. The average value and variances are shown for human RNA . Type 1 Interferons coding transcripts are shown separately to highlight their very low CpG force. For human influenza the longest gene (PB2) has been chosen. Data from Gencode [20] and NCBI [21], see SI.1.

We analyze dinucleotide forces across the SARS-CoV-2 genome, in particular those on CpG and UpA motifs. As a first check, in Fig. 1 we compare the total CpG forces of SARS-CoV-2 with other ssRNA viruses (Influenza B, 1918 and 2009 H1N1, H5N1, Ebola, type 1 HIV, type 1a Hepatitis C), and the forces in human transcriptome for coding, non coding region [19] as well as the RNA for type I interferons. Fig. 1 shows that the total SARS-CoV-2 force on CpG motifs is larger than influenza B but smaller than the H1N1 strains, H5N1, Ebola virus, therefore putting SARS-CoV-2 in an intermediate region of CpG forces.

Next we explore more in detail the CpG forces within the *Coronaviridae* family, with a particular emphasis on the genera *Alphacoronavirus* and *Betacoronavirus*, and on those viruses which infect humans [24]. We have first compared the global forces on CpG and UpA dinucleotides, computed across the whole genome without any constraint on amino acid or codon usage. To calculate such forces we need a reference nucleotide bias, to give a null model for nucleotide usage (See Methods Sec. 6.1, 6.2). We will in the following use as nucleotide bias, the one calculated from the human genome [25]. This is the natural choice since we ultimately want to predict regions sensed by PRRs as non-self, which may engage the human innate immune response. We have verified that the specific choice of such bias does not change qualitatively the results in the comparative analysis, as long as the same set of reference frequencies is used to compute the forces across viral genomes. For instance, we obtain similar results in terms of the rank order of viruses, using the average frequencies of nucleotides in coronavirus genomes.

The resulting comparison is summarized in Fig. 2a: in this case 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 a global CpG force comparable or higher than that of SARS-CoV-2, so there is not a straightforward correlation between global CpG force and the pathology of a coronavirus in humans. For the force on UpA dinucleotides, it is clear already from this comparison that it is anti-correlated with the CpG force, as we expect for the complementarity between such motifs[1]. Due to the fact that the genes and the global length of viruses in the *Alphacoronavirus* and *Betacoronavirus* genera are quite similar, the comparison of CpG forces gives the same outcome than the direct comparison of the number of CpG along the sequence, see Supp. Fig. SI.2.

To go beyond the global analysis we have performed a local analysis on CpG motif usage, by looking to the local forces in fixed-length windows along the genome (similar results are also obtained by looking directly at the

[1] transitions have largest probability with respect to transversions and are less likely to result in amino acid substitutions.

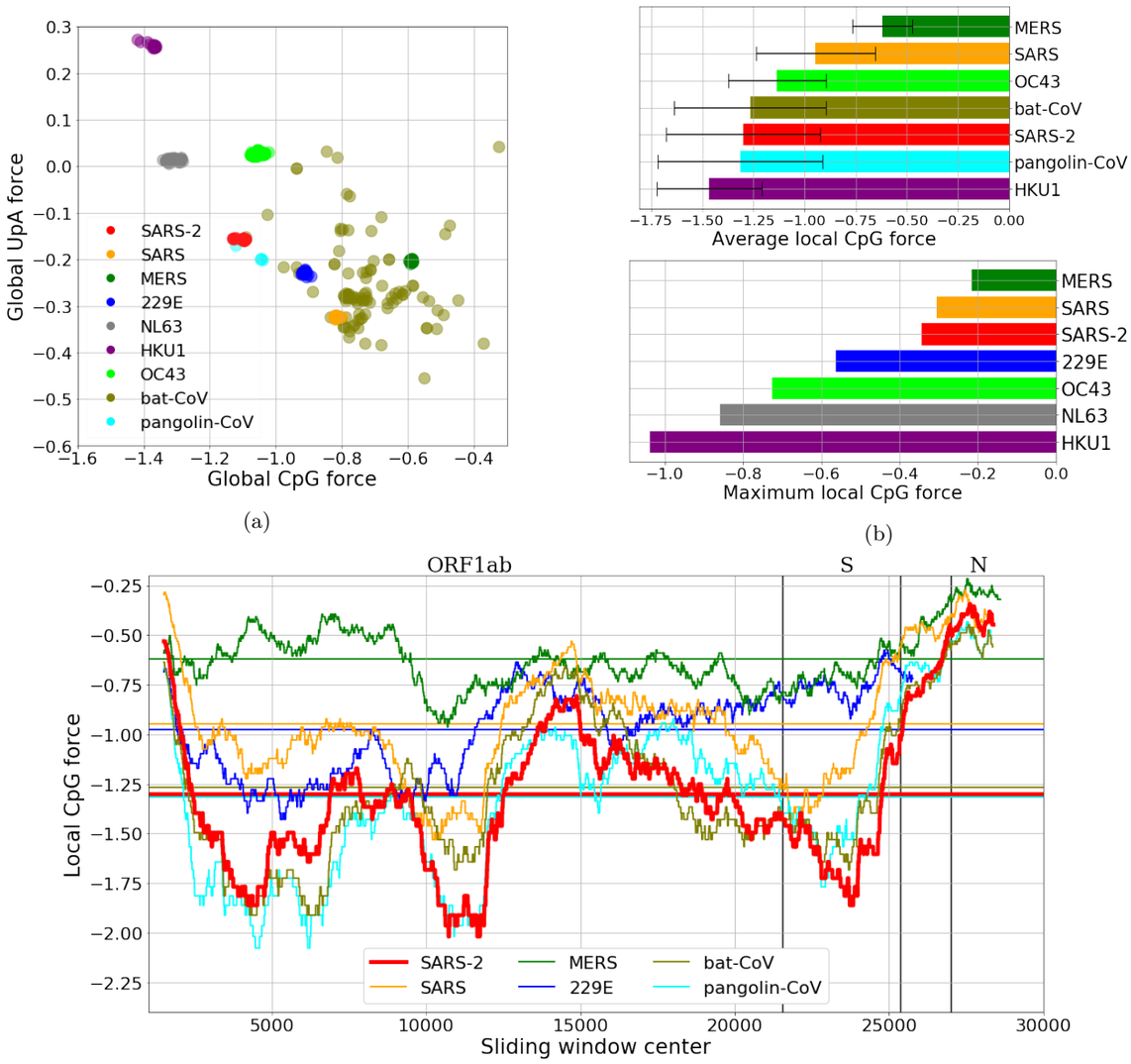


Figure 2: **UpA and CpG forces and their local fluctuations in the coronaviriadae family.** (a): Forces computed on the whole genome. Species are well clustered, due to the large conservation of the sampled sequences, except for the strains originating in bats which is a grouping of several diverse strains. The anticorrelation of the CpG and the UpA force is due to their complementarity. (b, c, d): Local analysis in sliding windows of 3 kb along the genome. Local forces along the genome (b, c): while the average CpG force of SARS-CoV-2 is relatively low, the variance along the genome is high with greater CpG forces in certain regions (such as the coding region for protein N) and lower CpG forces in other (e. g. coding region for protein S). (b)-lower panel: maximum value of the local CpG force showing that values for SARS-CoV-2 moves closer to the most dangerous viruses. The bat sequence analyzed in panels (b) and (c) is RaTG13, while the pangolin sequence has been sequenced at Guangdong, in 2019. Data from VIPR [22] and GISAID [23], see Methods Sec. 6.5 and SI.1.

CpG content in the same windows). This is done in Fig. 2c for SARS, MERS, SARS-CoV-2, hCoV-229E and two representative sequences of bat and pangolin coronaviruses, which are chosen because of their closeness (in Hamming distance) to SARS-CoV-2 (the bat-coronavirus sequence is RaTG13, the pangolin-coronavirus sequence has been sequenced at Guangdong, in 2019).

It is evident that in some regions, especially at the 5' and 3' ends of the sequence, the 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-229E (as well as in the other hCoVs, see Supp. Fig. SI.3). The high CpG content at the extremities can 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 [26, 27]. During the infection many more RNA fragments with these regions are present in the cytoplasm, than the other regions of the viral genome. In this sense, despite the relatively low CpG content of SARS-CoV-2 compared to other coronaviruses, there is can be a high concentrations of CpG rich RNA due to the higher transcription of these regions.

To complement this analysis, in Fig. 2b (upper bar plot) we make a direct comparison among the average local CpG forces and standard deviations of the coronaviruses analyzed. This shows more clearly that, while the SARS-CoV-2 has an intermediate overall CpG level, its variance is very high, comparable with that of SARS and of two examples of bat and pangolin coronaviruses. This suggests a comparison among the coronaviruses based on the part of the sequence with the highest CpG and UpA force, which is presented in Fig. 2b (lower bar plot). In this case, the results are much closer to what may be expected: 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. These results are corroborated by the same qualitative picture which comes from the analysis of CpG motif density, as shown in Suppl. Fig. SI.2c. 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 [28, 11].

3 Forces in coding regions constrained by codon and amino acid usage

We have seen in Fig. 2c that in all the coronavirus genomes there are regions with large deviations in CpG content with respect to the average (whatever we use to measure this content, be it density or force), and that SARS-CoV-2 has a particularly high variance when compared with other coronaviruses circulating in humans. However, most of the viral genome codes for proteins, so the analysis should take this feature into account. The advantage of the method of dinucleotide forces is that its statistical physics framework allows such constraints to be introduced in a straightforward manner (Methods Sec. 6.2). In Fig. 3a we perform the local analysis of forces on dinucleotides with the coding constraints, by using as reference codon usage bias the human one, along all the coding regions of the SARS-CoV-2 genome, and in Figs. 3b, 3c, 3d and 3e, we focus on the genome regions coding for the nucleocapside (N) and spike (S) proteins [29, 30, 17].

A first, relevant observation is that the high variability which we discussed in Fig. 2c is confirmed by this analysis, and in particular the 3' peak of high CpG density and force is still present. Moreover, in Fig 3a some of the small non-coding regions which are excluded from the figure (as they are not used for the computation of the local force) are pointed out with black solid lines. It seems that large changes in local force happen in the vicinity of these lines, thus hinting at a possible role of recombination in the generation of this complex and heterogeneous local CpG-force landscape, though such a hypothesis would have to be confirmed by phylogenetic analysis.

To compare several strains of coronaviruses, we focus on the structural proteins, which are always present and quite similar across the *Coronaviridae* family. Notice that all these proteins are coded for after the ORF1ab coding region in Fig. 2c. We computed the CpG force with coding constraints for each structural protein. The first protein (in the standard 5'-to-3' order) is the S protein, and it corresponds to the low-peak in CpG force and density for SARS-CoV-2 in Fig. 2c. The S protein has to bind with ACE2 human receptor and TMPRSS2 [30]. It is also the longest structural protein. In Fig. 3e the CpG force along the region coding for S (S ORF) computed. SARS-CoV-2 shows the lowest average of CpG force among the human-infecting betacoronaviruses, see Fig. 3b, and also locally it has a force lower than most of the other beta-CoVs. A fascinating (as well as speculative) reason that could explain the low CpG force on the S protein, is again recombination: the S protein coding region may come (at least in part) from other coronaviruses that better bound human entry receptors [31].

The second longest structural protein is the N protein. Fig. 3c shows the local computation of CpG forces under coding constraints restricted to this protein, coded for in a region (N ORF) of about 1.2 kb close to the 3'-end of the viral genomes. In this case it is apparent that the CpG force in the SARS-CoV-2 is much higher, immediately below that of SARS and above that of MERS. This is confirmed by the comparison of the total CpG force computed for N across the human-infecting members of the *Coronaviridae* family, presented in Fig. 3d. This high CpG content, together with the large concentration of N transcripts and similar sequences as a part of all other subgenomic RNAs [27], could confound the "global" prediction that SARS-CoV-2 presents low levels of dinucleotide patterns to receptors that can recognize them, which may be relevant to innate recognition of the virus.

In the comparative analysis of structural proteins in the *Coronaviridae* family, the E protein is the one showing the largest difference in CpG forces between coronaviruses that circulate in humans (see Supp. Fig. SI.4b). However, the fact that the E protein is short (coded for in about 228 nucleotides) makes the computation of the local forces less stable (that is, small variations in CpG number may correspond to large force variations). The M protein is also quite short (coded for in about 669 nucleotides) and has smaller CpG force differences among species (see Supp. Fig. SI.4c). We therefore focus in the next section on the two longest structural proteins, S and N, to check for the ability of our model to explain (and possibly predict) the CpG changes via synonymous mutations in protein coding regions.

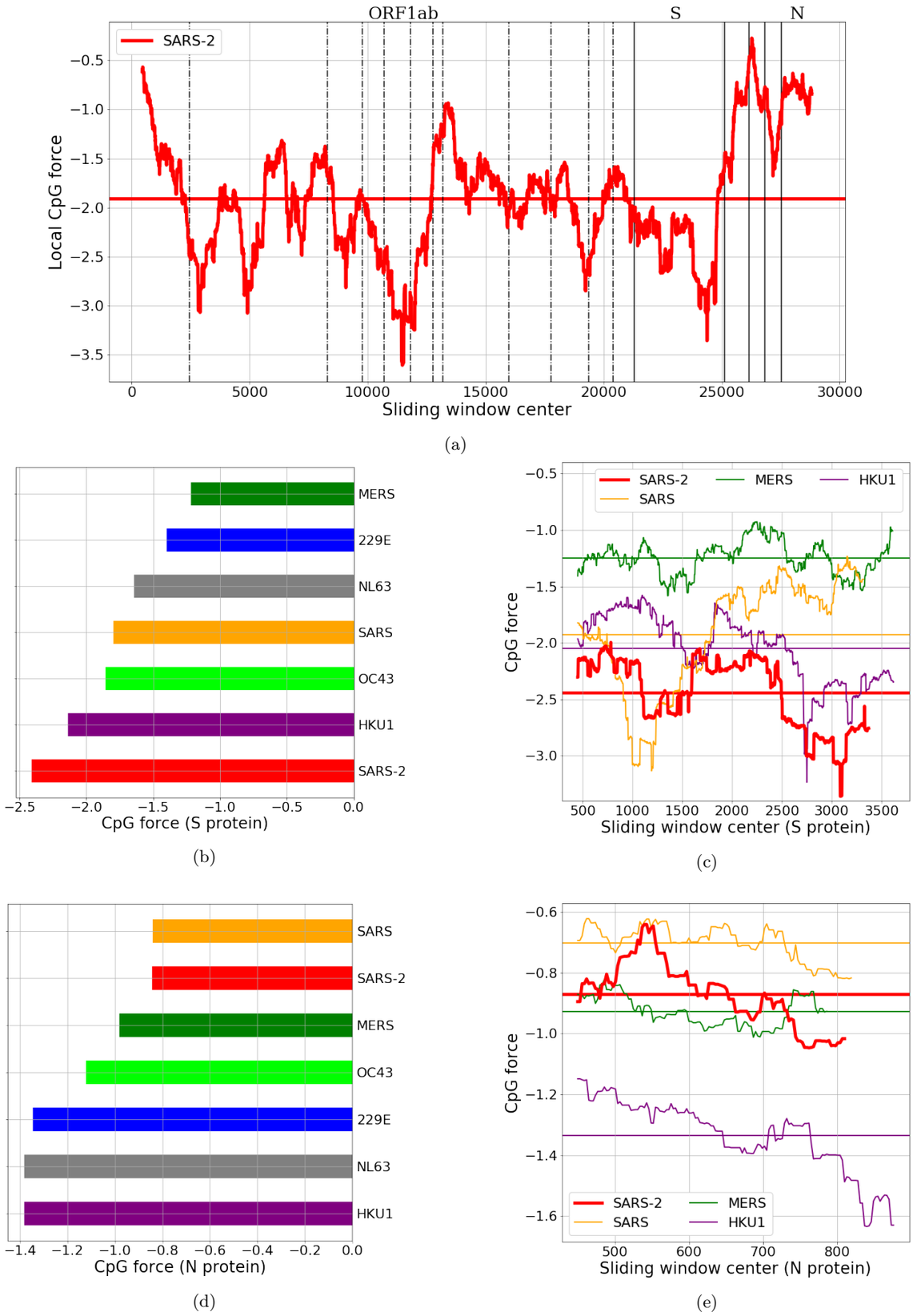


Figure 3: **CpG local Forces on SARS-CoV-2 coding regions constrained by codon and amino acid usage**. (a): Local Forces, on sliding windows over 900 nucleotides, on all the coding regions of SARS-CoV-2 (pre-processed to ensure the correct reading frame, see Methods Sec. 6.5). The horizontal line denotes the average force. Many large fluctuations coincide with boundaries of protein domains, suggesting that recombination could have had a role in creating theses CpG force fluctuations. In panels (b) and (d): bar plots for average local forces for structural proteins in the *Coronaviridae* family. The two largest structural proteins, N and S, have very different forces in SARS-CoV-2. Each bar is obtained by averaging the CpG force over several sequences (from VIPR [22] and GISAID [23], see SI.1), and the standard deviation of the average (not shown) is lower than 0.025. (c) and (e): local forces for N and S compared with the same proteins of other beta-coronaviruses. Horizontal lines denote the average forces.

4 Evolution of CpG motifs in SARS-CoV-2

As seen in the previous sections, SARS-CoV-2 has regions which have higher CpG content (and forces) than strains that have been circulating in humans, while the number of CpGs (and the corresponding force) is much lower in other regions. It is therefore interesting to ask to what extent the CpG force model is able to predict any bias in the synonymous mutations already detectable in the few months of evolution since the first SARS-CoV-2 genome was sequenced (data from GISAID [23], reference sequence Wuhan, 26-12-2019, last updated sequence 2020-04-22, see Methods Sec. 6.5) and, consequently, could allow us to predict putative mutational hotspots in the viral evolution.

Barring confounding effects, we would expect that some regions, such as the N ORF, will be driven by host mimicry towards a lower number of CpG motifs. Other regions, such as the the 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. 4a. We see that the N and S subsequences undergo synonymous mutations that decrease (blue) or increase (red) their numbers of CpG. We also show the local CpG force, constrained by codon and amino acid usage (same as Fig. 3a), along the sequence. Most of the mutations that decrease the number of CpG are located at the 3'-end of the sequence, in correspondence with the high peak in CpG force 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.

We first focus on the N protein. The locations of synonymous mutations and multiplicities (the number of sequences in which they are found to occur) are indicated by star symbols of corresponding sizes in Fig. 4b. We observed a total number of $M_s = 262$ synonymous variants (with 55 unique mutations). Out of these M_s variants 101 and 9, respectively, lower and increase CpG, while the remaining 152 leave CpG content unchanged. It is remarkable than more than 10 in 11 mutations that affect the CpG count decrease it.

When restricting the analysis on the 101 mutated sequences in which the CpG count decreases, the losses take place in at 9 different loci. The nucleotide motifs preceding these loci are listed in the top 9 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. Six out of 9 of these motifs, which represent 68 out of the 101 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 6 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 3 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.

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	60
CTGGCAAT**G**G**CG**	7	642	2
CACAAGCTTT**CG**	-	822	22
GTCACACCTT**CG**	-	980	6
CAAGCCTTAC**CG**	-	1148	5
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 1: Analysis of nucleotidic motifs preceding CpG in the N protein. The top 6 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 some of the mutated sequences. 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 3 lines show the other 3 CpG lost through mutations and their 10 preceding nucleotides, which do not correspond to motifs tested in [9]. The last 6 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.

For the S protein (see Fig. 4d and star sizes) we observed $M_s = 393$ synonymous variants (with 123 unique mutations). Among these unique mutations, 20 and 58, respectively, lower and increase the CpG content. Therefore, only 1 in 5 mutations that affect the CpG count decreases it.

These results seem to support the existence of early selection pressure to lower CpG occurrence in N ORF, but not in S ORF. Moreover, according to our model, mutations are not only driven by the local CpG force, but also by viral codon bias computed on the reference sequence[2]. It is therefore theoretically possible (and in full accord with our model) that, even in low CpG force regions, a mutation decreasing the CpG number is favorable due to the biases introduced by the distribution of synonymous codons. We developed a computational framework to estimate the odds of transition to a mutated sequence from SARS-CoV-2 (Wuhan, 26-12-2019) based on the CpG-force model (Methods Sec. 6.3). For a synonymous mutation that does not change CpG, our synonymous mutation score (SMS), defined in Methods Sec. 6.3, will only depend on the change in codon biases due to the mutation. Additionally, mutations affecting CpG content bring contributions to the transition probabilities that depend on the local CpG

[2]We consider the virus codon bias rather than the human codon bias, as SARS-CoV-2 is likely not in equilibrium with his host.

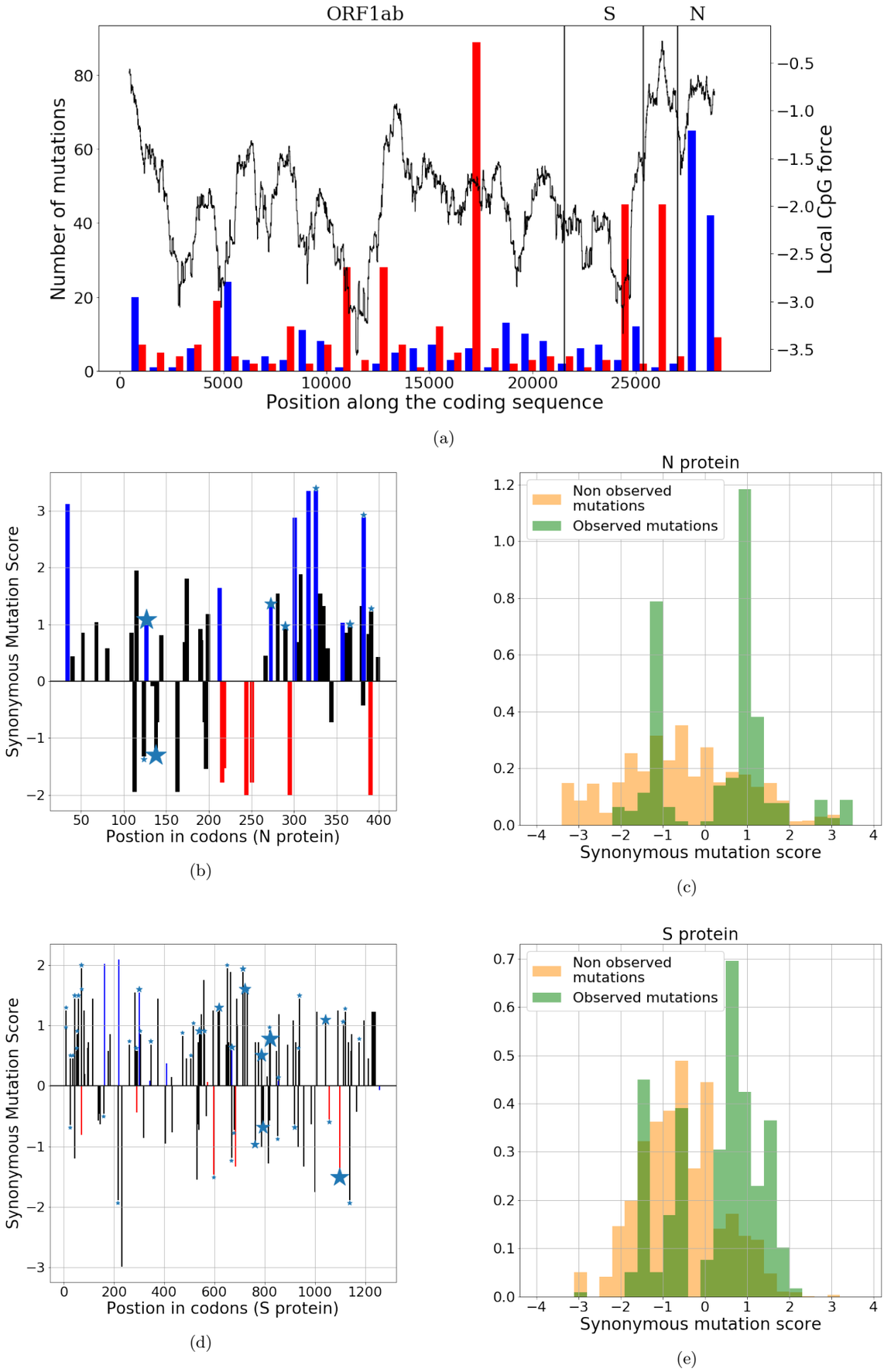


Figure 4: **Analysis of the synonymous mutations from one among the first sequenced SARS-CoV-2 genomes (Wuhan, 26-12-2019)**. (a): CpG-changing mutations. Red and blue bars show the numbers of mutations that, respectively, increased and decreased the mean CpG content over 900 b windows (left scale). The black curve shows the CpG force along the sequence (right scale). (b),(d) 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. (c),(e): histograms of SMS for observed mutations (in green) and for non-observed putative mutations (in yellow) in the N and S proteins. Data from GISAID [23], see Methods Sec. 6.5.

force. In Fig. 4 we show our predictions for synonymous mutations in the N (4b, 4c) and the S (4d, 4e) proteins. Figs. 4b and 4d show SMS along, respectively, the N and S sequences and the mutations, respectively, lowering (blue), increasing (red), or leaving unchanged (black) the CpG content. We observe that the majority of mutations (taking into account multiplicity, labeled by stars in the figure) 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. 4c and 4e we show the histograms of the SMS corresponding to observed synonymous mutations (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 mutations 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: 94 for N and 247 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). A simpler model for the SMS using codon bias only, without force (Suppl. Fig. SI.5) is also able to discriminate between observed and non-observed mutations, albeit with a smaller score for protein N (ANOVA F-test=35) and a comparable result for S (ANOVA F-test=239), further demonstrating that S ORF synonymous mutations are likely due to a neutral reversion to the mean, while N ORF mutations may contain an additional pressure.

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), [32] (trs-trv bias) (Methods Sec. 6.4.), 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.6. 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.

	F-test (S ORF)	F-test (N ORF)	p-value (S ORF)	p-value (N ORF)
Uniform bias	1	1	-	-
Uniform bias + CpG force	32	189	$2 \cdot 10^{-8}$	$3 \cdot 10^{-40}$
trt-trv bias	0.5	$2 \cdot 10^{-4}$	> 0.05	> 0.05
trt-trv bias + CpG force	22	124	$3 \cdot 10^{-6}$	$2 \cdot 10^{-27}$
Virus codon bias	239	35	$< 10^{-50}$	$4 \cdot 10^{-9}$
Virus codon bias + CpG force	247	94	$< 10^{-50}$	$2 \cdot 10^{-21}$

Table 2: Anova F-test 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.

Consequently, mutation data seem to confirm the hypothesis that the CpG force drives the evolution of synonymous mutations to lower number of CpG motifs in the N ORF. For the S ORF, which already has quite low CpG number, the situation is less clear: the observed tendency to CpG increase through synonymous mutations is compatible with the presence of a force pushing up this number, but is equally well explained by the viral codon bias model. This uncertainty is expected from the relatively small number of mutations with a CpG change in S protein region in the sequences collected so far.

5 Conclusions

The present work contains three main directions in the early analysis of SARS-CoV-2 genomes and dinucleotide motifs, particularly CpG, usage. First, a comparative analysis with other genomes in the *Coronaviridae* family, which has stressed that a peculiarity of SARS-CoV-2 with respect to the other coronavirus genomes is a globally lower CpG content (though not particularly low compared to other RNA viruses) accompanied by large fluctuations along the genome. Notably the segment coding for the S protein has a much lower CpG content and force. 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 the 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 [27] to be the most abundant transcript in the cytoplasm. The high concentration of this feature 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, need further experimental investigation. The sharp heterogeneity of the CpG abundance along the SARS-CoV-2 genome is also compatible with viral recombination, in agreement with the hypothesis stated in [31]. The degree to which this heterogeneity in any way reflects zoonotic origins should be further carried out using phylogeny. A first analysis of the evolution of synonymous mutations since the outbreak of COVID-19 shows that mutations increasing CpG content have occurred in the middle and low CpG content regions, in particular the S protein region. Conversely, mutations lowering the number of CpG have taken place in regions with higher CpG content, in particular, the N protein 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 [34, 23, 22], and on a longer

evolutionary time are needed to confirm these hypothesis. SARS-CoV-2 has a low mutation rate for an RNA virus due to the presence of a proofreading mechanism in its replication [35]. An integrative study of the transmissivity of the SARS-CoV-2 and its mutation rate could be performed to predict the time scale at which such natural evolution driven by host mimicry would bring the virus to an equilibrium with its host [5, 1].

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.

Acknowledgment

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 GISAID's EpiCoV(TM) Database on which this research is based. This work was partially supported by the ANR-19 Decrypted CE30-0021-01 grants. 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.

6 Methods

6.1 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.2 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 Sec. 6.2 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) [2], see Eq. 4 and SI Sec. SI.2. 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 Fig. 1.
- As shown in Sec. 6.2, 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 always used here the human codon bias as reference 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 in Sec. 3 and it was crucial to investigate the dynamics of the synonymous mutations in viral evolution, Sec. 4.

6.2 Force-based model

The model at the core of many of the analyses made here is taken from [1]. 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 the i-th nucleotide is 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, Sec. SI.2, the force computed as above turns out to be the logarithm of the relative abundance index, that is

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

In Suppl. Fig. SI.1 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 [1]).

6.3 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 $T(n_{CpG} \to n'_{CpG})$ 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 [1] for the evolution of the CpG number of a sequence under a initial force f through a equilibrium force f_{eq}:

$$\tau \frac{dn}{dt} = (f_{eq} - f). \tag{6}$$

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 [1] 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

$$T(n_{CpG} \to n'_{CpG}) \propto e^{(f_{eq}-f)\Delta n}, \tag{7}$$

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

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

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,

$$\text{SMS} = (f_{eq} - f)\Delta n + \log\left(\frac{p(c_i')}{p(c_i)}\right) \ . \tag{9}$$

6.4 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 pyrimidine).
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, which is not very different from the uniform one, to perform ANOVA F-tests together with other codon usage biases, see Table 2 and Suppl. Fig. SI.6.

6.5 Data Analysis

SARS-CoV-2 sequences are taken from GISAID [23]. We collected each sequence present in the database, which was obtained by 2020-04-22. Before any of our analyses, we discarded all the sequences where one or more nucleotides were wrongly read. This left us with 3329 SARS-CoV-2 sequences. To obtain Fig. 2 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 [22]. 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. 2a and for Fig. 2b (lower panel) we used the largest number possible of sequences, up to a maximum of 100. For Fig. 2b (upper panel) and Fig. 2c 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. 3 and Fig. 4, we considered as reference SARS-CoV-2 sequence the one which has been collected on 26-12-2019 (ID: EPI_ISL_406798). Sequences has been processed to ensure the correct reading frame. This means, for instance, that 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. To produce the bar plots in Figs. 3b

and 3d 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 presented in Figs. 3b and 3d is smaller than 0.025 (and, for most of the viruses, much smaller). More detailed information about the genomes used in this work are given in Supplementary SI.1

References

[1] 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, 9, 10, 11

[2] 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, 10, 14

[3] 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

[4] 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

[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, 9

[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, 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, 4

[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, Chao-Lin 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] Chuang Guo, Bin Li, Huan Ma, Xiaofang Wang, Pengfei Cai, Qiaoni Yu, Lin Zhu, Liying Jin, Chen Jiang, Jingwen Fang, Qian Liu, Dandan Zong, Wen Zhang, Yichen Lu, Kun Li, Xuyuan Gao, Binqing Fu, Lianxin Liu, Xiaoling Ma, Jianping Weng, Haiming Wei, Tengchuan Jin, Jun Lin, and Kun Qu. Tocilizumab treatment in severe covid-19 patients attenuates the inflammatory storm incited by monocyte centric immune interactions revealed by single-cell analysis. *bioRxiv*, 2020. 1

[17] Fatima Amanat and Florian Krammer. Sars-cov-2 vaccines: status report. *Immunity*, 2020. 1, 4

[18] 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

[19] 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

[20] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012. 2

[21] 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. 2, 14

[22] Brett E Pickett, Eva L Sadat, Yun Zhang, Jyothi M Noronha, R Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, et al. Vipr: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research*, 40(D1):D593–D598, 2012. 3, 5, 8, 11

[23] 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, 5, 6, 7, 8, 11

[24] 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

[25] 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

[26] Michael MC Lai and David Cavanagh. The molecular biology of coronaviruses. In *Advances in virus research*, volume 48, pages 1–100. Elsevier, 1997. 4

[27] 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. 4, 8

[28] 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. 4

[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. 4

[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. 4

[31] 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. 4, 8, 14

[32] 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. 8

[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] 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. 9

The heterogeneous landscape and early evolution of pathogen-associated CpG and UpA 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. Sequences shown in Fig. 1: Human cDNA and ncRNA as annotated in HG38 assembly, Type I interferon's cDNA as annotated in HG38. Viral ssRNAs were obtained from NCBI [21] Virus database (strains used: H5N1: A/Anhui/1/2005, H1N1: A/Aalborg/INS132/2009 and A/Brevig Mission/1/1918, MERS: MERS-CoV_England-KSA/1/2018, SARS: CUHK-AG01, Ebola: COD/1995/Kikwit-9510623, Influenza B: B/Massachusetts/07/2020, HIV: HK_JIDLNBL_S003, HCV: NC_004102).
The SARS-CoV-2 sequence used in Figs. 2c and 2b (upper panel) has GISAID accession ID: EPI_ISL_420793. The GenBank accession numbers for the specific genomes used in Figs. 2c and 2b (upper panel) are: AY427439 (SARS), NC_038294 (MERS), MF542265 (hCoV-229E), JX524171 (hCoV-NL63), KT779555 (hCoV-HKU1) and KF923918 (hCoV-OC43). For these figures, choose the bat and pangolin sequences closest to the SARS-CoV-2 points in Fig. 2a (these two sequences are also known to be very similar to the SARS-CoV-2 genome from other works [31]). 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. 3a and 4a. For Figs. 3c and 3e 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 From CpG force to CpG relative abundance

We want to show in which limit that the CpG force (without codon constraints) is equivalent to the relative dinucleotide abundance [2], 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

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

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.1 shows the correlation between the CpG force with the nucleotide bias and the CpG relative abundance .

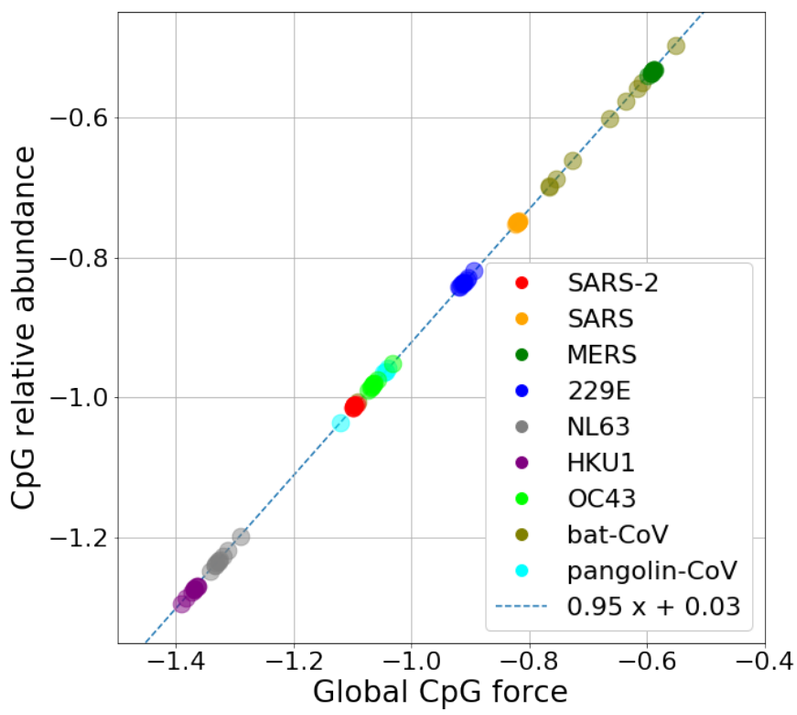


Figure SI.1: Comparison between the CpG force and the CpG relative abundance index. As discussed in Sec. SI.2, these two quantities are almost identical when the genome is long. To show that, here 10 different genomes for several coronavirus species are used to compute these two quantities, and the dashed blue line is a linear fit of the resulting points.

SI.3 Supplementary Figures

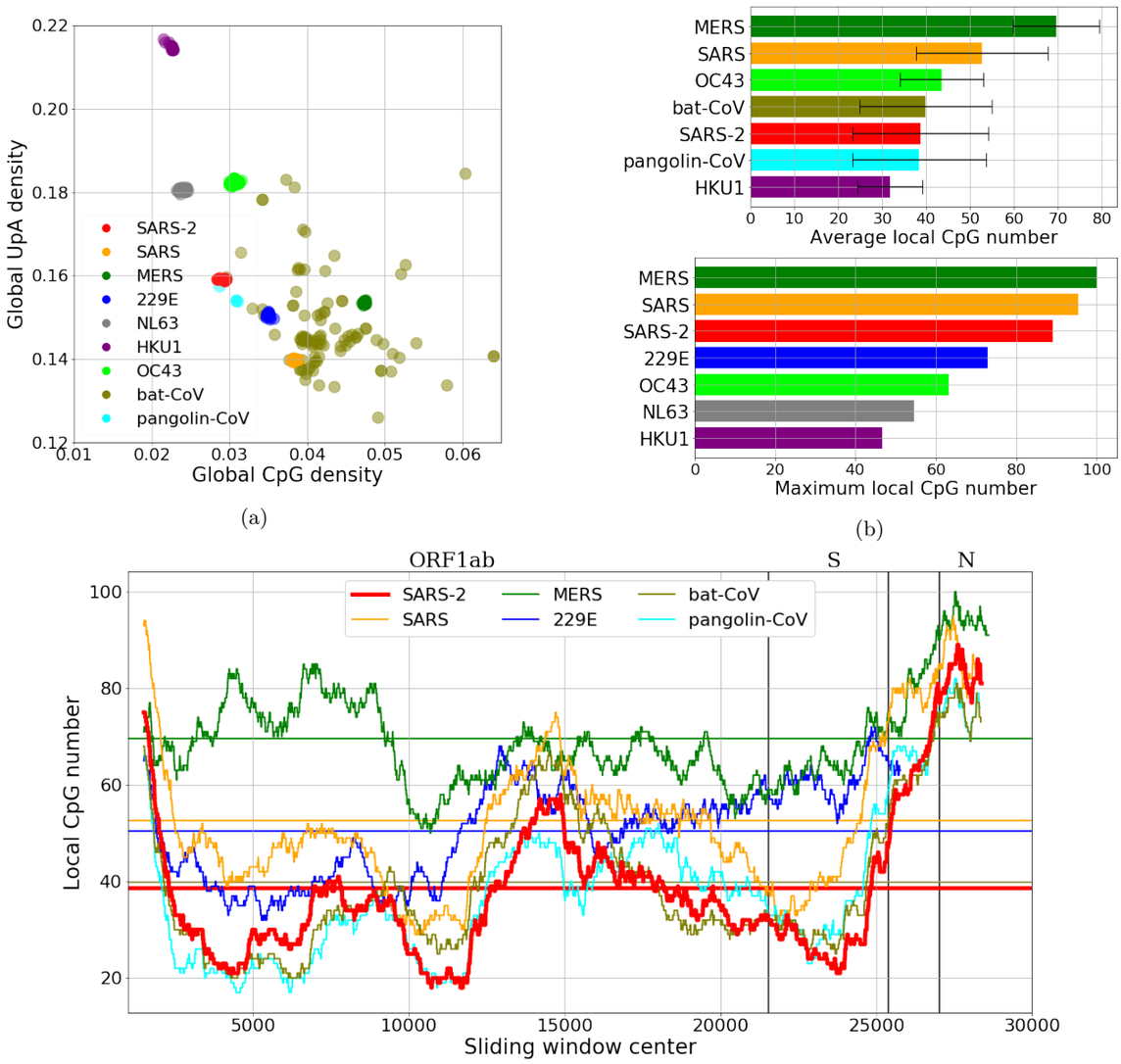
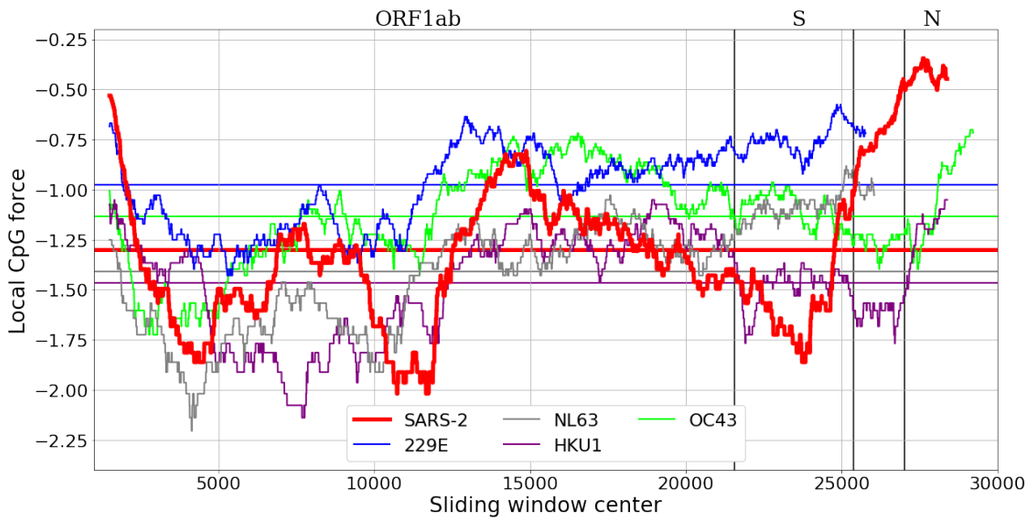
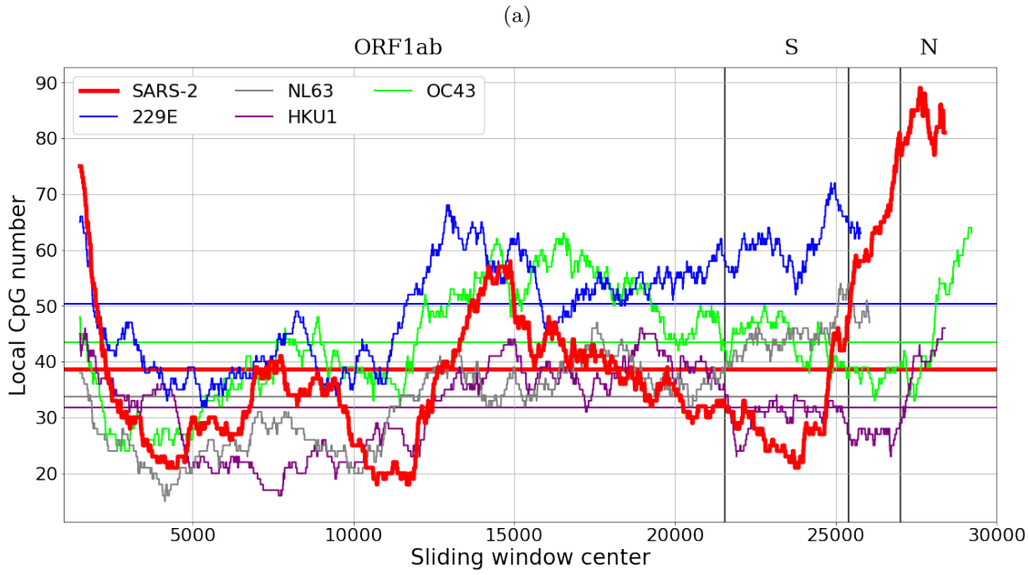


Figure SI.2: The same analysis performed in Fig. 2, but here we used CpG densities instead of CpG forces. The results obtained are qualitatively the same.

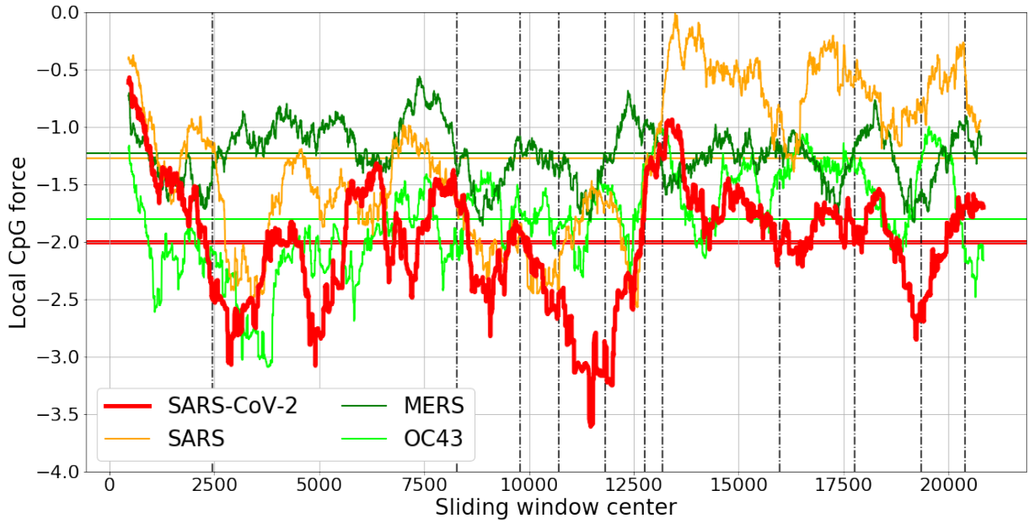


(a)

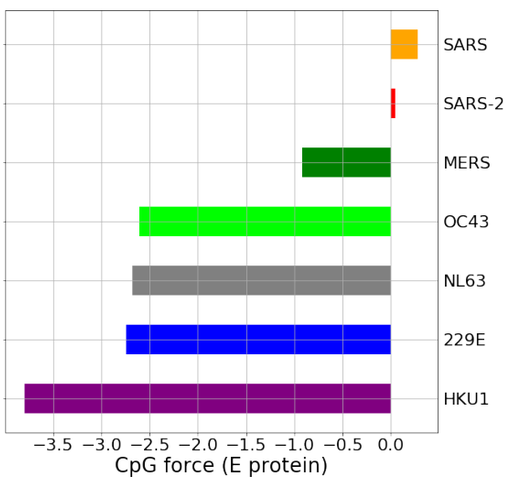


(b)

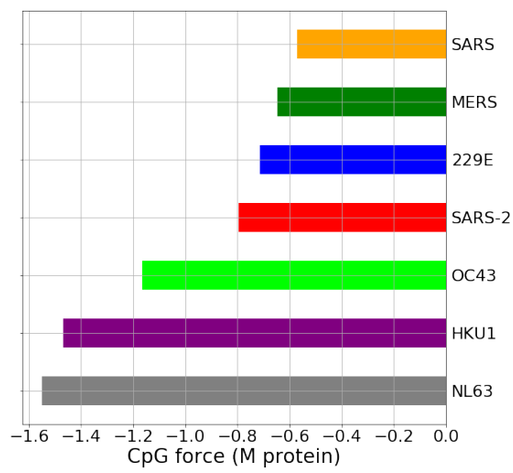
Figure SI.3: Supplement to Fig. 2c and Fig. SI.2, where all the coronaviruses associated with circulating human strains are compared with SARS-CoV-2 in terms of CpG force (panel (a)) or density (panel (b)). Again, even though the final regions of the hCoVs has relatively high CpG force with respect to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



(a)

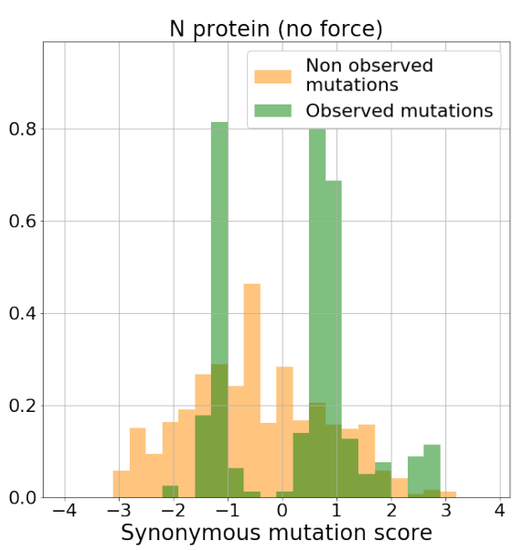


(b)

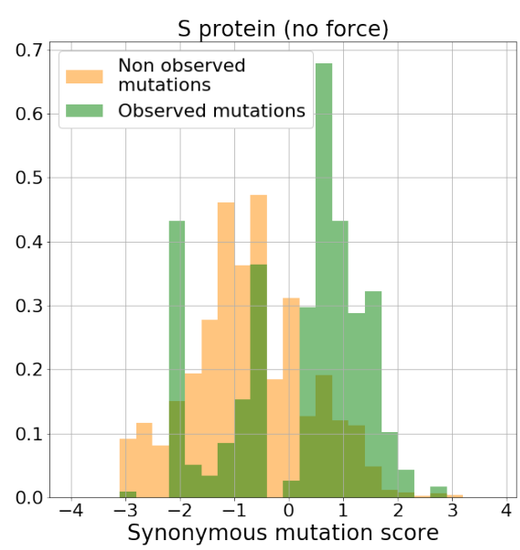


(c)

Figure SI.4: Extension of the comparison performed in Fig. 3. 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.

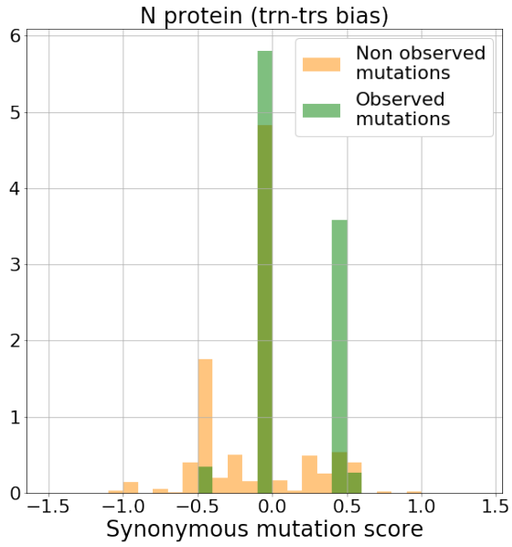


(a)

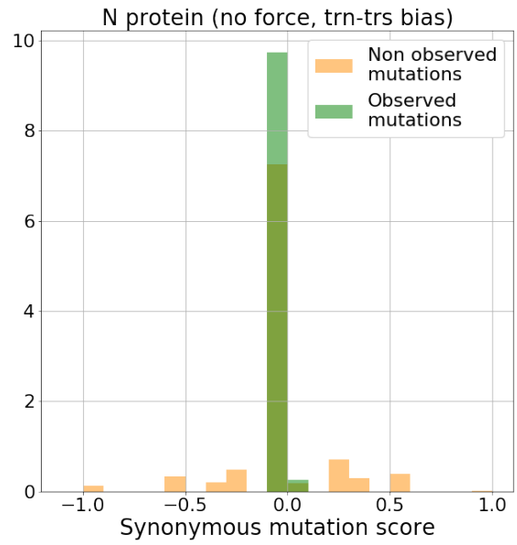


(b)

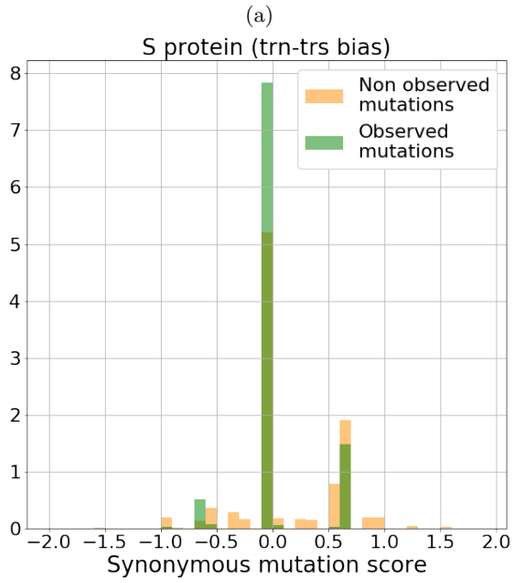
Figure SI.5: The same analysis discussed in Figs. 4c and 4e have been performed here computing the SMSs without force.



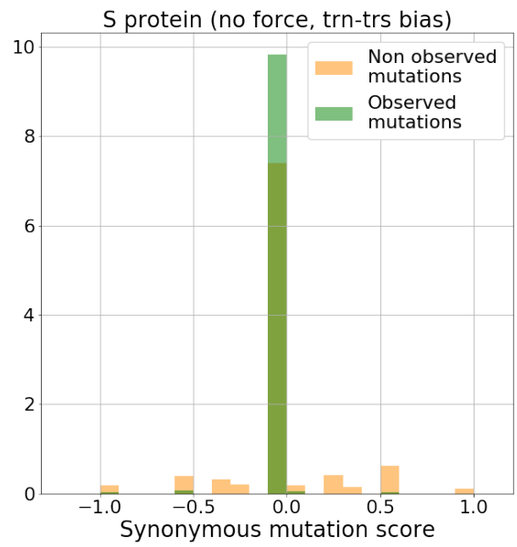
(a)



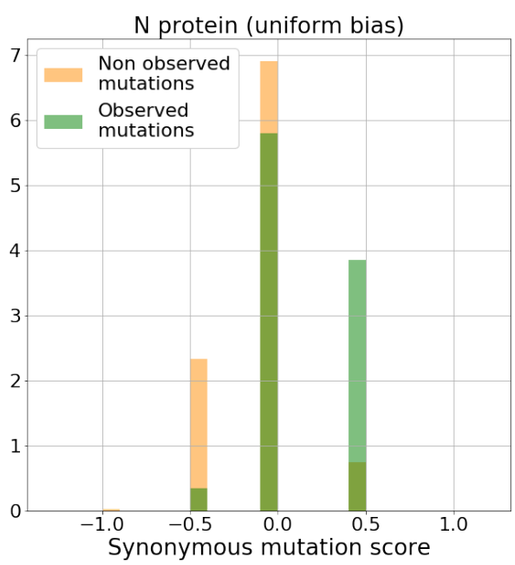
(b)



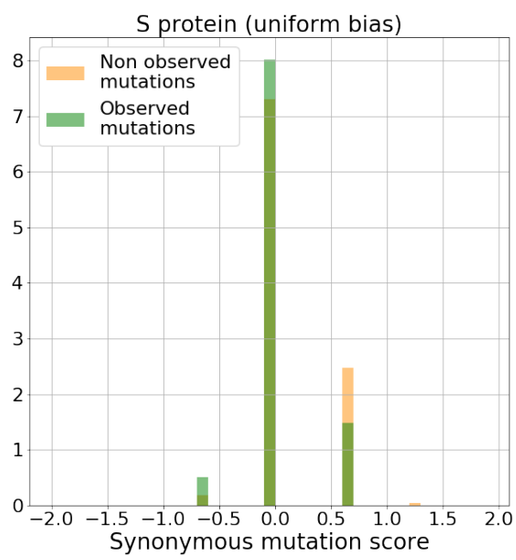
(c)



(d)



(e)



(f)

Figure SI.6: The same analysis discussed in Figs. 4c and 4e 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 used the included also a penalty for transversions with respect to translations.