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. 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 .

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280

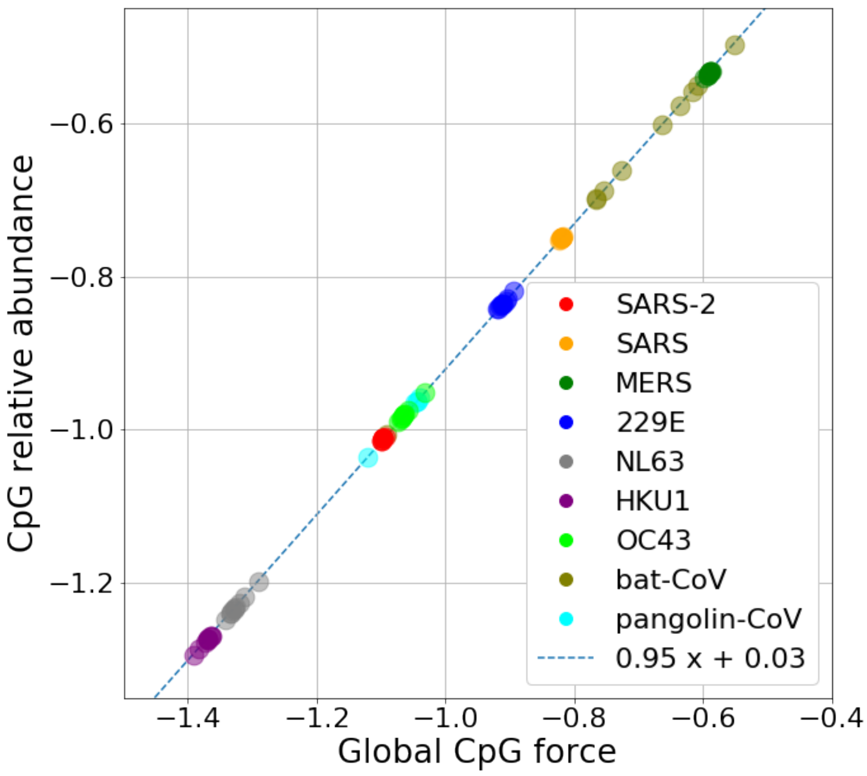


Figure SI.1: Comparison between the CpG force and the CpG relative abundance index. As discussed in Sec. SI.2, these two quantities are almost identical when the genome is long. To show that, here 10 different genomes for several coronavirus species are used to compute these two quantities, and the dashed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

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280

SI.3 Supplementary Figures

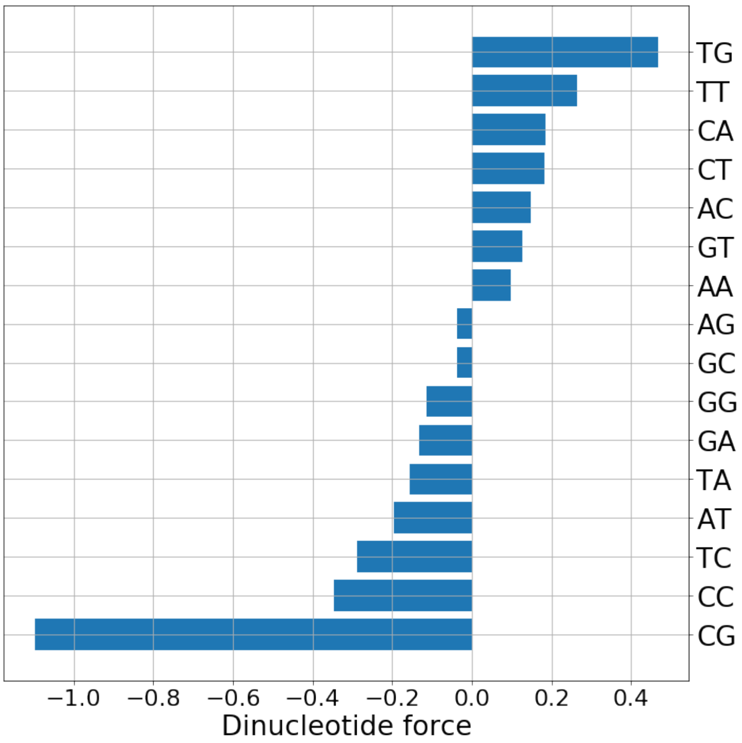
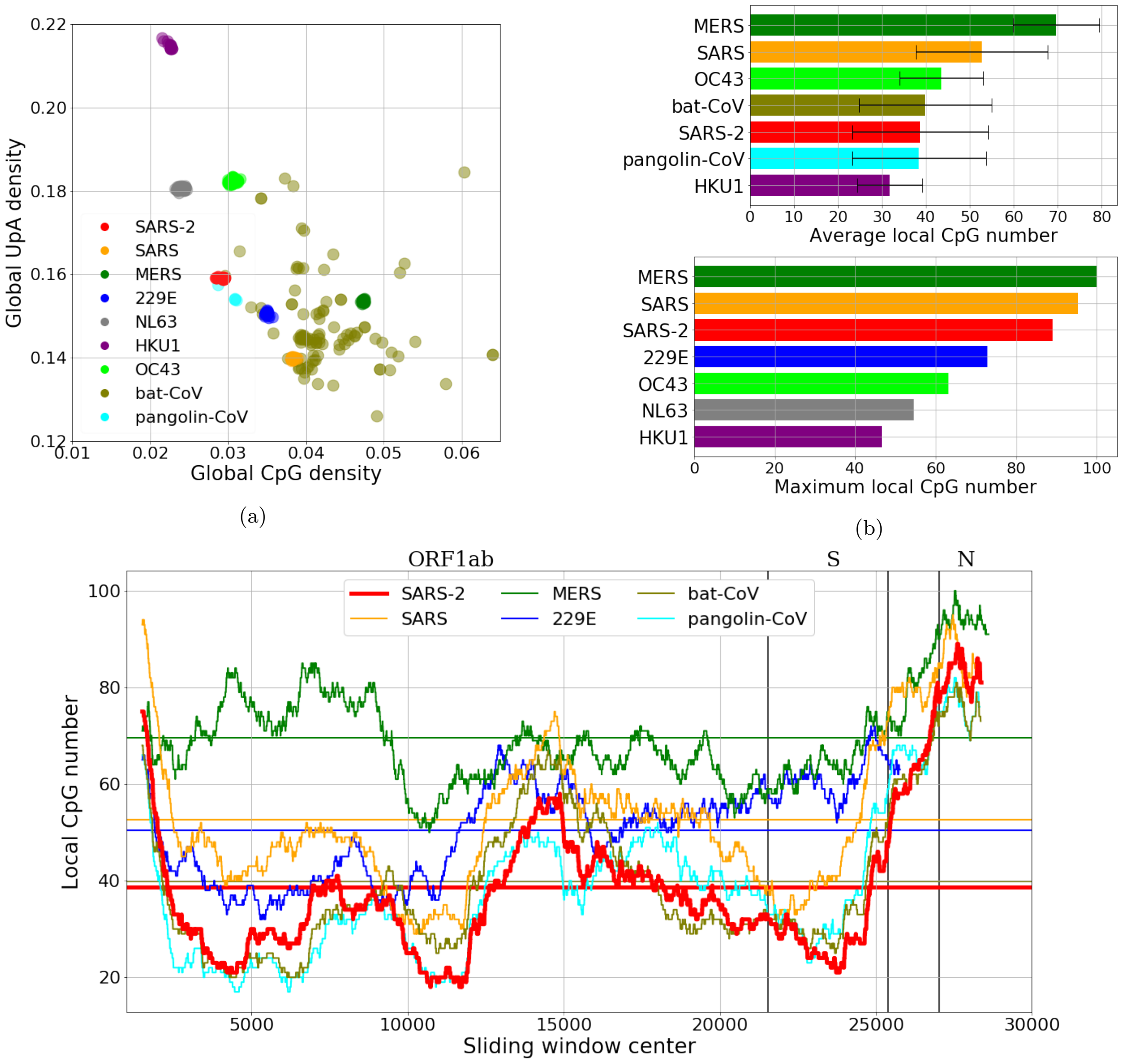


Figure SI.2: All dinucleotide forces computed on the whole SARS-CoV-2 genome, without any codon constraint. The CpG motif is the one with the largest force in absolute value, and the second one is TpG which is one transition away from CpG.

	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 3: Analogous of Table 2, computed with the sequences collected up to 2020-04-22. Although the availability of fewer data lowers the F-test results most of the times (and therefore gives a higher p-value), 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 obersved and non-observed mutations, while these two cases become distinguishable if the CpG force is added, especially for the N ORF.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280



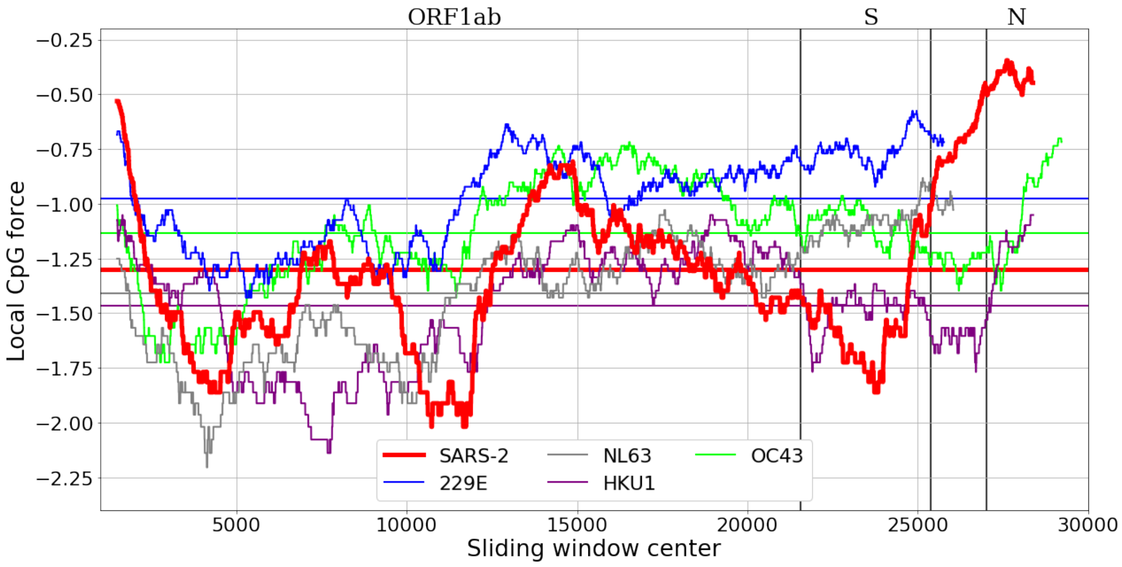
(a)

(b)

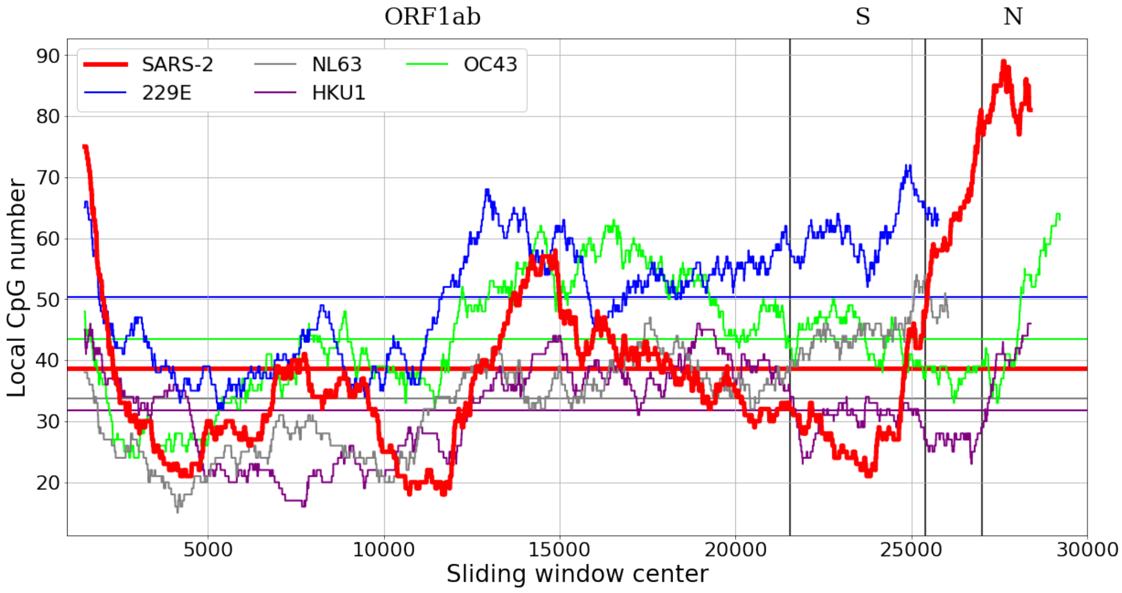
(c)

Figure SI.3: The same analysis performed in Fig. 2, but here we used CpG densities instead of CpG forces. The results obtained are qualitatively the same.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280



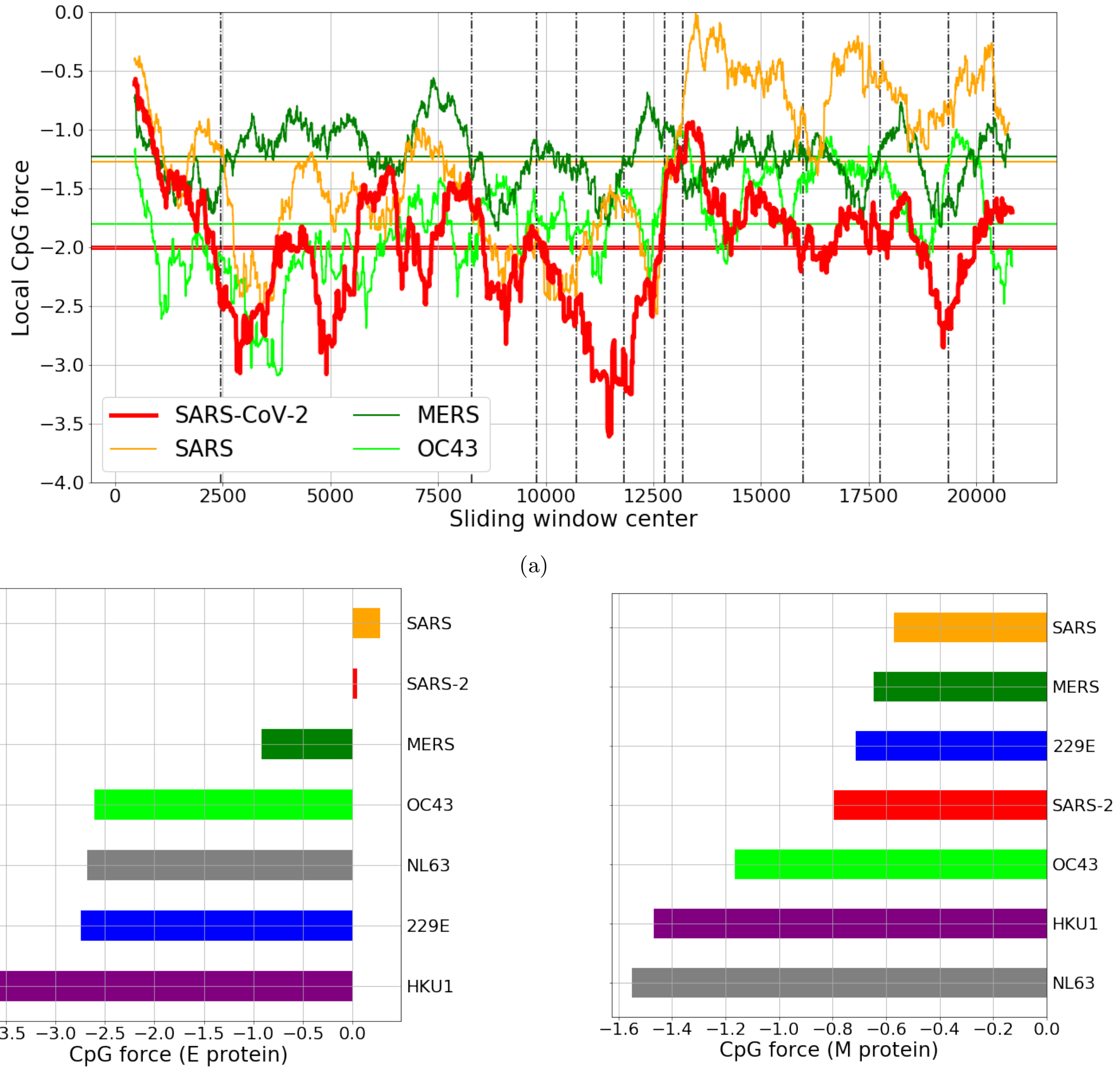
(a)



(b)

Figure SI.4: Supplement to Fig. 2c and 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 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.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280

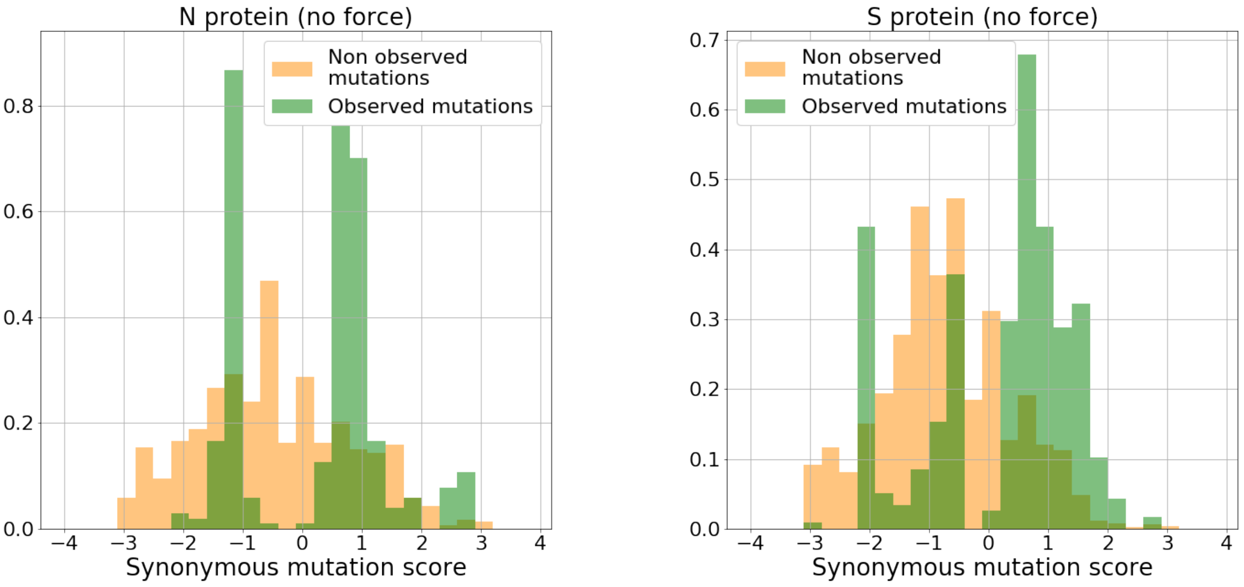


(a)

(b)

(c)

Figure SI.5: 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.6: The same analysis discussed in Figs. 4c and 4e have been performed here computing the SMSs without force.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280

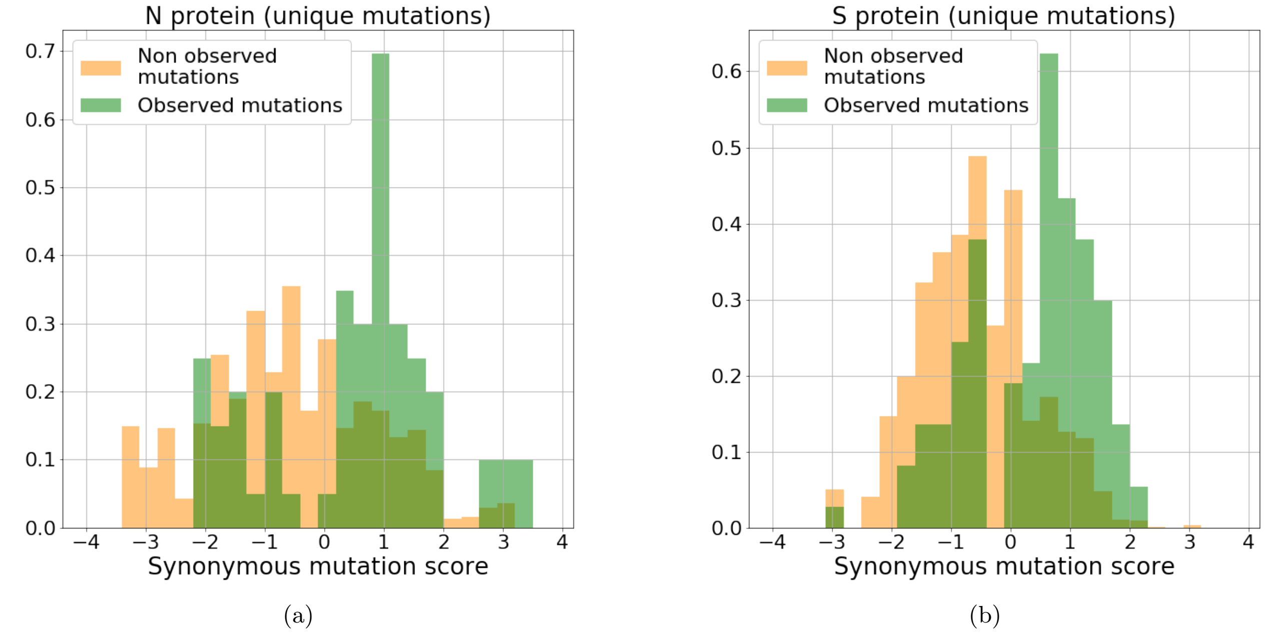
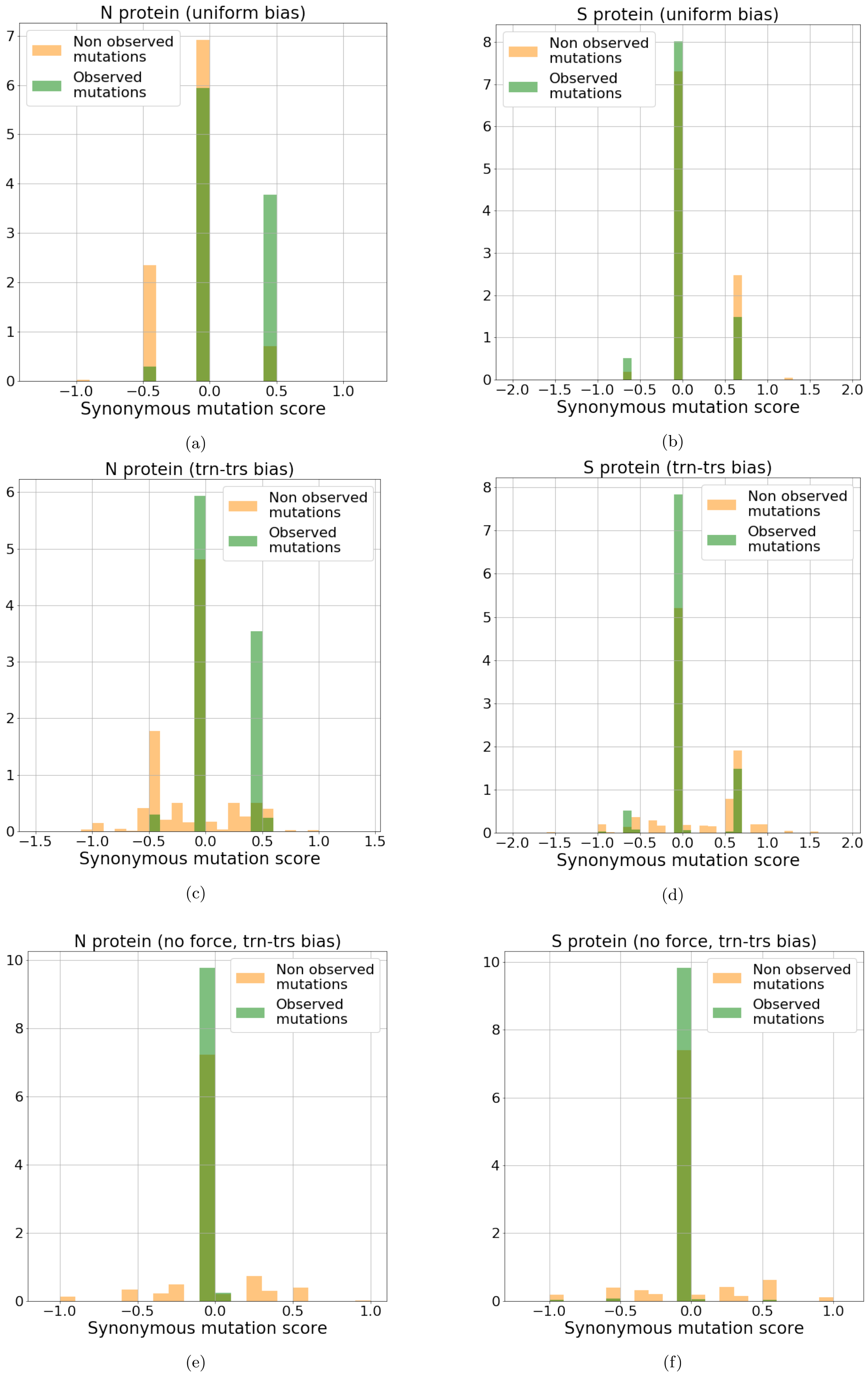


Figure SI.7: The same analysis discussed in Figs. 4c and 4e have been performed here computing the SMSs without taking into account the variant multiplicities. We run the ANOVA F-test, obtaining $F = 37$ for the N ORF, and $F = 145$ for the S ORF, which in turn respectively correspond to p-values of $1.6 \cdot 10^{-9}$ and $1.1 \cdot 10^{-32}$, thus showing the robustness of our results.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280



(a)

(b)

(c)

(d)

(e)

(f)

Figure SI.8: 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.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=3611280