scientific reports

Check for updates

OPEN

The low abundance of CpG in the SARS-CoV-2 genome is not an evolutionarily signature of ZAP

Ali Afrasiabi[1,9], Hamid Alinejad-Rokny[1,2,3,9], Azad Khosh[4], Mostafa Rahnama[5], Nigel Lovell[6], Zhenming Xu[7] & Diako Ebrahimi[8]✉

The zinc finger antiviral protein (ZAP) is known to restrict viral replication by binding to the CpG rich regions of viral RNA, and subsequently inducing viral RNA degradation. This enzyme has recently been shown to be capable of restricting SARS-CoV-2. These data have led to the hypothesis that the low abundance of CpG in the SARS-CoV-2 genome is due to an evolutionary pressure exerted by the host ZAP. To investigate this hypothesis, we performed a detailed analysis of many coronavirus sequences and ZAP RNA binding preference data. Our analyses showed neither evidence for an evolutionary pressure acting specifically on CpG dinucleotides, nor a link between the activity of ZAP and the low CpG abundance of the SARS-CoV-2 genome.

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19) pandemic, has a ~ 30 kb single-stranded positive RNA (+ssRNA) genome, which is one of the largest known viral RNA genomes[1]. The SARS-CoV-2 RNA has unique genomic features with likely roles in the high pathogenicity and cross-species transmission of this virus[2,3]. An expansive view of the SARS-CoV-2 genomic features is an essential step to improve our current understanding of the evolutionary path of this virus. One of these unique features is the low abundance of CpG in the SARS-CoV-2 genome[4–8]. CpG depletion is a well-known phenomenon in viruses particularly those with RNA genomes[9,10]. It has been reported that the CpG composition of + ssRNA viral genomes often mimics the CpG content of their hosts, however the underlying molecular mechanisms are not well understood[11]. One of the suggested mechanisms is DNA cytosine methylation-induced deamination[11–13]. However, SARS-CoV-2 does not have a DNA stage, thus this mechanism is not likely to be relevant. Another suggested mechanism is recognition of CpG sites within viral RNA by the host RNA-binding protein ZAP (CCCH-type zinc finger antiviral protein)[8]. ZAP is known to restrict the replication of a broad range of viruses including SARS-CoV-2 by binding to the CpG rich regions of viral RNA, and subsequently inducing viral RNA degradation[14–17]. The low CpG content of RNA viruses have been proposed to be linked to this mechanism[14–16].

Xia[8] and Wei et al.[18] have proposed that the low CpG content of SARS-CoV-2 might be due to an evolutionary pressure from ZAP[8]. Among the coronaviruses studied in Xia's study, those isolated from canines were shown to have the most CpG depleted genomes. Therefore, it was postulated that dogs may have been the intermediate species for the emergence of SARS-CoV-2. This hypothesis is based on two assumptions, which are not likely to be correct: first, only the frequency of CpG (but not those of other 15 dinucleotides ApA, ApC, …, UpU) is sufficient to make inferences about the origin of viruses, and second, ZAP is the main source of low CpG abundance in the SARS-CoV-2 genome. A number of follow up studies challenged Xia's methodology and conclusion. For instance, Pollock et al. repeated Xia's analysis using a larger number of SARS-CoV-2 related viruses and

[1]BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia. [2]UNSW Data Science Hub, The University of New South Wales, Sydney, NSW 2052, Australia. [3]Health Data Analytics Program, AI-Enabled Processes (AIP) Research Centre, Macquarie University, Sydney 2109, Australia. [4]Department of Biology, Yazd University, Yazd 8915818411, Iran. [5]Department of Plant Pathology, University of Kentucky, Lexington, KY 40546, USA. [6]The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia. [7]Department of Microbiology, Immunology and Molecular Genetics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. [8]Texas Biomedical Research Institute, San Antonio, TX 78227, USA. [9]These authors contributed equally: Ali Afrasiabi and Hamid Alinejad-Rokny. ✉email: debrahimi@txbiomed.org

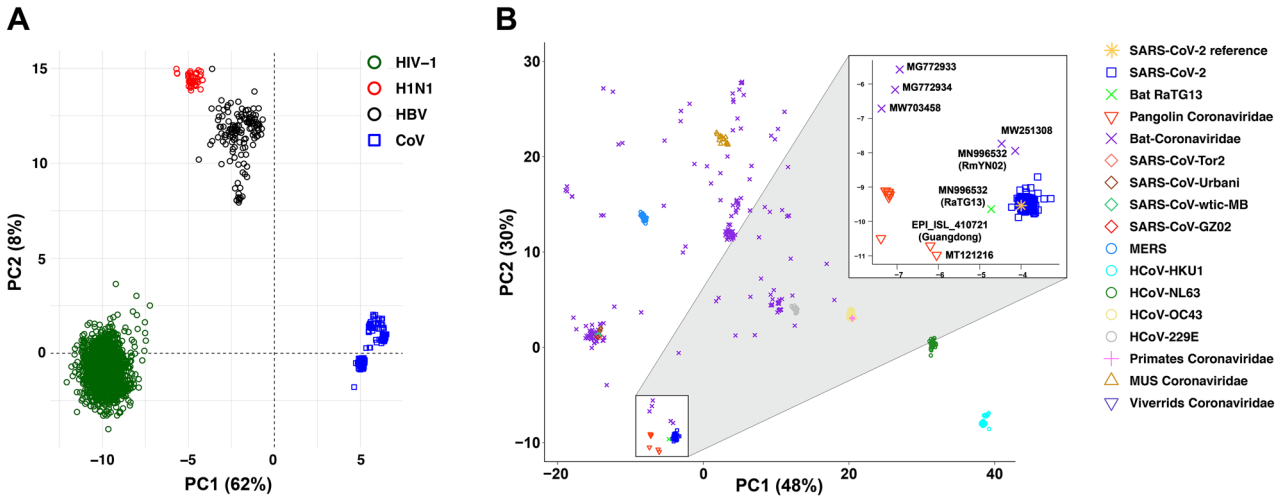

Figure 1. PCA of viral motifs representations. D-values of all dinucleotide, trinucleotide and tetranucleotide motifs in all viral sequences form a matrix, which is used as an input for PCA. (**A**) PC1-PC2 plot shows four clusters, one for each virus family: H1N1, coronaviruses (CoV), HBV, and HIV-1. (**B**) PC1-PC2 plot classifies coronaviruses into two clusters. All 664 SARS-CoV-2 (including reference sequence), Bat-RaTG13, RmYN02, 4 Bat-Coronaviridae viruses (MW703458, MW251308, MG772933, MG772934), and 10 Pangolin-Coronaviridae (EPI_ISL_410721, EPI_ISL_410539, EPI_ISL_410542, EPI_ISL_410543, EPI_ISL_410538, EPI_ISL_410541, EPI_ISL_410540, MT040336.1, MT040333.1, MT121216.1) formed a cluster (SARS-CoV-2-like group), which is separated from the rest of coronavirus sequences including Human coronavirus 229E, Bat-Coronaviridae, Human coronavirus HKU1, Murine coronavirus, MERS coronavirus, Human coronavirus NL63, Human coronavirus OC43, Primates-Coronaviridae, SARS coronavirus Tor2, SARS coronavirus Ubani, Viverrids-Coronaviridae, SARS coronavirus wtic_MB, SARS coronavirus GZ02. SARS-CoV-2-like are highlighted with a square.

found that CpG deficiency is not specific to dog coronaviruses or SASR-CoV-2, and it is observed in pangolin coronaviruses and to a greater extent in pangolin pestiviruses[5]. Moreover, modeling of the binding affinity of ACE2 and SARS-CoV-2 Spike protein in 410 vertebrates showed a low score of susceptibility to SARS-CoV-2 infection for dogs[19]. This finding was also confirmed by viral replication experiments[20]. Digard et al. showed that CpG abundance varies significantly across the SARS-CoV-2 genome, with envelope and ORF10 not showing CpG depletion[21]. They showed that the CpG levels of SARS-CoV and SARS-CoV-2 envelope sequences are even higher than those of envelope from other human coronaviruses. Using a phylogenetic analysis, the authors argued that these genomic composition changes are more likely to be an ancestrally-driven traits related to the origin of these viruses in bats, not due to a post-zoonotic transfer selection force. These data suggest that the overall CpG content alone is not a reliable index for inferring the host origin of viruses[21]. Furthermore, Xia et al.[8] and Wei et al.[18] based their argument on the baseline levels of ZAP in the SARS-CoV-2 intermediate host tissues. It is reasonable to argue that ZAP expressions in healthy/uninfected cells/tissues do not reflect the ZAP expression levels during viral infection[22–24].

Here, we perform a detailed analysis of multiple data sets including the representations of short sequence motifs in viral genomes and patterns of ZAP binding to viral RNA to investigate the role of ZAP in reducing the SARS-CoV-2 CpG level. Our analyses show that the representations of almost all dinucleotides, not only CpG, are different in the SARS-CoV-2 genome compared to the genomes of other coronaviruses. For example, UpC, and ApU are all represented at significantly lower levels in the SARS-CoV-2 genome compared to the SARS-CoV genomes. Our analyses indicate that not only the CpG motifs preferentially targeted by ZAP but also those not often recognized by ZAP have lower representation in SARS-CoV-2 compared to SARS-CoV. Altogether, our results provide multiple lines of evidence against the role of ZAP in the evolution of the SARS-CoV-2 genome.

Methods

Viral sequences. For the analysis presented in Fig. 1A, we used full-length sequences of 3967 coronaviruses, 2021 HIV-1, 91 Flu H1N1, and 141 HBV, totaling 6220 sequences from GenBank. For the analysis presented in Fig. 1B, we obtained a total of 1546 full-length coronavirus genomic sequences (323 human SARS-CoV, 10 viverridis coronaviruses, 190 bat coronaviruses, 41 mus SARS-CoV, 5 primates coronaviruses, 256 MERS coronaviruses, 88 murine coronaviruses, 10 pangolin coronaviruses, and 664 SARS-CoV-2). To ensure that our analysis is not affected by variations accumulated in the SARS-CoV-2 genome over time, we only used 664 SARS-CoV-2 sequences reported before January 31st, 2020 (Supplementary Table 1).

Analysis of motif representation. We quantified the representation of short sequence motifs (di-, tri- and tetra-nucleotides) in all viral genomes using our previously reported Markov-based representation (D-value) method[25–27]. Briefly, motif representation (D-value) is defined as the ratio of the observed frequency (P_{obs}) of a

motif over its expected frequency (P_{exp}). P_{obs} is simply the observed relative frequency of the motif. P_{exp} is quantified using the frequency of the motif in the sequence and the frequencies of the smaller constituting motifs[12]. An example of a D-value for the tri-nucleotide motif ACG is given in Eq. (1).

$$D = \frac{ACG_{obs}}{ACG_{exp}} = \frac{ACG_{obs} \times C_{obs}}{AC_{obs} \times CG_{obs}} \quad (1)$$

For each analysis, we arranged the D-values of all dinucleotide, trinucleotide, and tetranucleotide motifs of all sequences in a data matrix. The matrices were then analyzed by principal component analysis (PCA). We performed two separate PCAs. The first analysis was done on 3967 coronaviruses 2021 HIV-1, 91 Flu H1N1, and 141 HBV sequences to demonstrate that motif representation data can be used to separate different virus families. The second PCA was performed on 1546 coronavirus sequences to identify viruses whose genome show high similarities to the SARS-CoV-2 genome. To quantify similarities (i.e. distances from the SARS-CoV-2 cluster) in the principal component space, we used Mahalanobis distance[28]. Further, Mann–Whitney test was used to determine the difference in the median of sequence motif representation (D-value) between the SARS-CoV-2-like group (SARS-CoV-2 and closely related coronaviruses) and viruses of the SARS-CoV group.

Analysis of association between CpG abundance and ZAP. To investigate the association of ZAP binding regions in the viral genomes with the number of CpGs co-located with these binding regions, we used the publicly available datasets of cross-linking immunoprecipitation (CLIP-seq) reported for ZAP. We obtained the processed ZAP CLIP-seq data (density of reads aligned to the genome) for wild type JEV (Japanese Encephalitis Virus) and wild type HIV[16,29]. We then calculated the CpG density across the JEV and HIV genomes using a sliding window analysis method. We used 200 bp window and 1Bp sliding for JEV and 250 bp window and 1Bp sliding for HIV.

The motif C(n7)G(n)CG has been demonstrated to be the optimal binding motif for Mouse ZAP[30]. To further investigate the role of ZAP in reducing CpG abundance in shaping the genome of SARS-CoV-2, we analyzed the abundance of C(n7)G(n)CG in SARS-CoV-2-like viruses. We used the abundance of motif C(n7)C(n)CG as a negative control in our analysis. Mann–Whitney test was used to determine the significance of difference in the median of these two motifs in the SARS-CoV-2 genome.

Results

Analysis of motif representations using PCA. We applied principal component analysis (PCA) on motif representation (D-value) matrices to interrogate similarities between and within virus families (Supplementary Table 2 and 3). Figure 1A shows the PCA of all four virus groups studied here, and Fig. 1B shows the results of PCA on coronaviruses only (see "Methods" section for details). As indicated in Fig. 1A, all of the four virus families HIN1, HBV, HIV-1 and coronaviruses (CoV) are separated using the first two PCs. PCA of only coronavirus sequences is depicted in Fig. 1B. All groups of coronaviruses are clearly separated from each other except Bat-Coronaviridae viruses which are diverse and form multiple clusters. All 664 SARS-CoV-2 (including reference sequence) and several bat and pangolin coronaviruses formed a cluster (SARS-CoV-2-like group shown in a square in Fig. 1B; see Supplementary Table 4), which was separated from the rest of coronavirus sequences including human coronaviruses 229E, the rest of bat coronaviruses, human coronaviruses HKU1, MERS, murine coronaviruses, human coronaviruses NL63, human coronaviruses OC43, primate coronaviruses, and SARS coronaviruses (Tor2, Urbani, Viverrids, Wtic-MB and GZ02, which we refer to as SARS-CoV here).

Representation of dinucleotides in the genome of SARS-CoV-2-like and SARS-CoV viruses. As indicated in Fig. 1B, coronaviruses used in this study form a distinct group (SARS-CoV-2-like) based on their motif representations (D-values). We compared, for each dinucleotide, the median D-values of SARS-CoV-2-like (group 1) with SARS-CoV (group 2) (Fig. 2 and Supplementary Table 3). All dinucleotide motifs (not only CpG) except for ApG and CpU were significantly different between the two groups. There is an excess of ApA, ApC, CpC, GpG, UpA and UpU, and a deficit of ApU, CpA, GpA, GpC, GpU, UpC, UpG and CpG in group one (SARS-CoV-2-like) compared to group two (SARS-CoV).

Analysis of ZAP binding to CpG sites. To determine the CpG motif specificity of ZAP, we overlaid the ZAP CLIP-seq data and the CpG distribution of two viruses, HIV and JEV (Japanese Encephalitis Virus) (Fig. 3). We found no clear and consistent pattern of co-localization between ZAP binding regions and CpG sites in HIV and JEV genomes.

It has been shown that ZAP binds preferentially to C(n7)G(n)CG motifs[30]. To further investigate the role of ZAP binding in reducing the SARS-CoV-2 CpG level, we compared the relative frequency of C(n7)G(n)CG and C(n7)C(n)CG in SARS-CoV-2-like viruses (Supplementary Table 5). The ZAP preferred motif C(n7)G(n)CG and the motif C(n7)C(n)CG which is not a preferred ZAP binding motif have similar abundances (p value $> > 0.05$, Fig. 4).

Discussion

It is critical to understand the evolutionary trajectory of SARS-CoV-2 and determine the molecular mechanisms that have contributed to its pathogenicity and adaptation to human cells. This information can help prevent and/or better control future pandemics caused by coronaviruses. Previous studies have shown that CpG is depleted in the genome of SARS-CoV-2 and its related bat and pangolin coronaviruses. To better understand the source of CpG depletion in these viruses, the first step was to identify the sequences that are closely related to SARS-CoV-2.

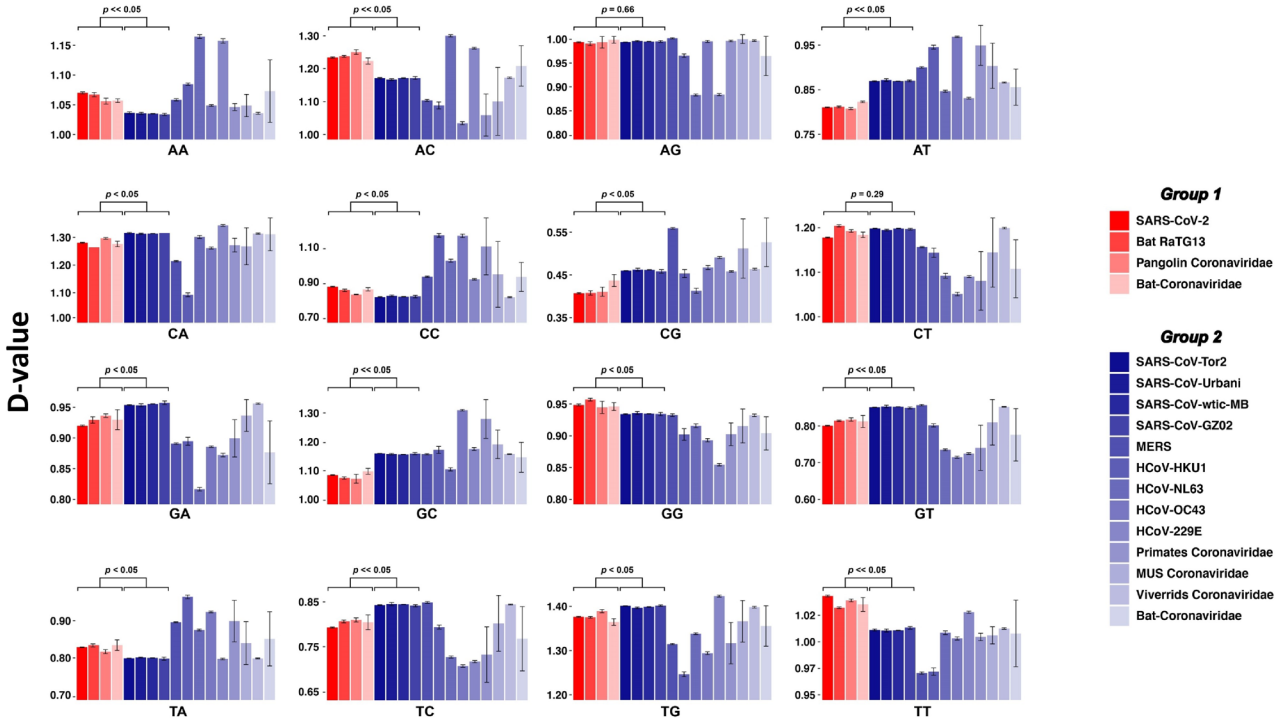

Figure 2. Comparison of dinucleotide motif representations between SRAS-CoV-2-like and SARS-CoV groups. D-values of each dinucleotide were compared between the two viral groups SARS-CoV-2-like and SARS-CoV. Mann–Whitney test was used to examine the difference in the median of D-values between the two coronavirus groups. D-value (motif representation) is defined as the ratio of the observed frequency (Pobs) of a motif over its expected frequency (Pexp). Pobs is simply the observed relative frequency of the motif. Pexp is quantified using the frequency of the motif in the sequence and the frequencies of the smaller constituting motifs.

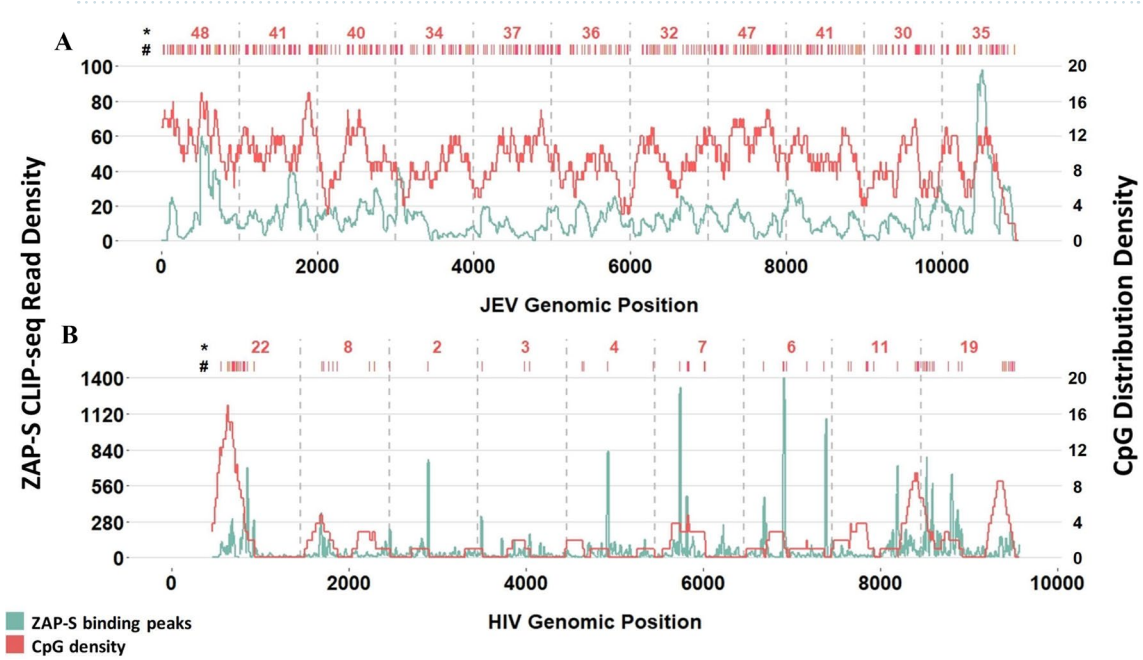

Figure 3. Co-location of ZAP binding regions and CpG motifs. Overlaying of the ZAP binding peaks and CpG densities in (**A**) JEV genome (Japanese Encephalitis Virus) and (**B**) HIV-1. The ZAP binding peaks (density of reads aligned to the genome) are estimated using a 250 bp sliding window moving by 1 bp along the viral genomes. The CpG density was calculated using the same sliding window analysis method, except we used a 200 bp window sliding by 1 bp in JEV and a 250 bp window sliding by 1pb in HIV-1. ZAP binding peaks and CpG densities are shown in green and red, respectively. # Location of CpGs. * Number of CpGs per 1 Kb.

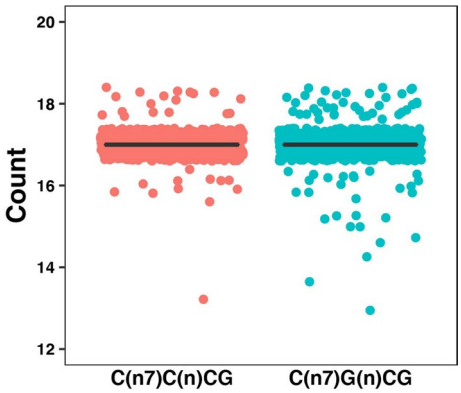

Figure 4. Comparison of the abundance of ZAP optimal binding motif C(n7)G(n)CG with the control motif C(n7)C(n)CG in viruses of SARS-CoV-2-like group. The abundance of ZAP optimal binding motif C(n7)G(n)CG was compared to C(n7)C(n)CG in the SARS-CoV-2-like group. The motif C(n7)C(n)CG was used here as a control. Mann–Whitney test was used to determine the difference in the median of abundance between these two motifs.

We have previously shown that an alignment-free method that uses the representation of short sequence motifs can precisely identify HIV-1 subtypes[31]. Here, we used the same method to investigate the similarities between and within virus families. This method successfully classified different virus families. As expected, it showed a separate cluster for SARS-CoV-2 with several bat and pangolin coronaviruses. Our analyses confirmed that SARS-CoV-2 and its closely related coronaviruses have a lower CpG content compared to other coronaviruses. However, we found that reduction in the representation of CpG was comparable to the reduction in GpC and ApT dinucleotides. More importantly, changes in motif representation are not exclusive to CpG. Most of the dinucleotides have significantly different representations in the viruses of the SARS-CoV-2-like group compared to the viruses in the SARS-CoV group. Altogether, our data suggest that the low abundance of CpG is not exclusive to the SARS-CoV-2 genome and is a general feature of several bat and pangolin coronaviruses. Most importantly, changes in the abundance of dinucleotides are not specific to CpG. Therefore, CpG reduction is likely part of a global genomic difference rather than being a signature of an exclusive selection force against CpG motifs. These data are in line with a study by Di Giallonardo et al. showing that dinucleotide composition of RNA viruses is shaped by the virus family not their hosts[32].

Immune evasion is one of the mechanisms proposed to explain the low CpG abundance in viral genomes including coronavirus sequences[9,11,33–36]. Assuming that this hypothesis is true, one would expect to observe a significant CpG depletion in the genome of SARS-CoV-2 to justify its high transmission rate and pathogenicity. By contrast, there is a little difference between the SARS-CoV-2-like and SARS-CoV viruses in terms of CpG abundance. The average CpG counts per kilobase (Kb) is 14.7 and 19.3 in group one (SARS-CoV-2-like) and group two (SARS-CoV) viruses, respectively. This means, on average, compared to SARS-CoV, SARS-CoV-2 has ~ 4.6 less CpGs per 1 Kb. It is unlikely that such a marginal CpG difference plays a critical role in SARS-CoV-2 immune evasion. Importantly, a previous study showed no correlation between the pathogenicity and global CpG content of coronaviruses infecting humans[4]. Furthermore, it has been shown that the CpG content of SARS-CoV-2 sequences is highly variable across the viral genome. In some regions of the SARS-CoV-2 genome such as envelop, not only CpG is not depleted, but it is more abundant compared to some of the other coronaviruses. This suggests the global CpG content is unlikely to be a vital genomic feature of the SARS-CoV-2 genome with a role in immune evasion[21]. Altogether, the overall CpG reduction in the SARS-CoV-2 genome is likely unrelated to the pressure imposed on the virus by the innate immune system.

Among the components of the human innate immune system, ZAP has been shown to play a key role in the inhibition of RNA viruses by binding to CpG containing sequences and recruiting a RNA-degradation machinery[16]. Xia's study postulates that the source of SARS-CoV-2 is a bat coronavirus whose genome underwent further CpG reduction by ZAP after the virus infected an intermediate species with a high ZAP expression level (possibly a canine tissue). It was suggested that ZAP-induced CpG depletion of viral RNA in this intermediate species led to the generation of SARS-CoV-2, which was able to infect human cells[8]. Our analyses show a poor association between the abundance of CpG across viral genomes and the location of ZAP binding peaks. These data suggest that the binding of ZAP to viral RNA is likely not governed by the global CpG abundance of viral genomes. In support of our results, a study has shown that ZAP inhibition is independent of the viral CpG content[37]. Additionally, a recent study shows that the location and sequence context of CpGs but not the overall CpG abundance of viral genome play a role in inducing ZAP antiviral activity[17]. Moreover, one of the mechanisms by which ZAP inhibits viruses is through the suppression of viral mRNA translation via blocking eIF4A[15], which is independent of ZAP binding to viral RNA. Although it has been shown that ZAP is capable of inhibiting SARS-CoV-2 in vitro[38], there is no evidence to support a role for ZAP in reducing the CpG level of the SARS-CoV-2 genome and shaping the genome of this virus.

A previous study has shown that mouse ZAP preferentially binds to C(n7)G(n)CG motifs where n: A, C, G, or U[30]. Assuming that human ZAP has the same motif preference and that it has induced an evolutionary

pressure on the SARS-CoV-2 genome, one would expect the relative abundance of C(n7)G(n)CG to be lower than a non-ZAP binding motif (e.g. C(n7)C(n)CG) in the SARS-CoV-2 genome. Our analysis does not show a significant difference between the abundance of these motifs in the SARS-CoV-2 genome. This may provide yet another line of evidence against the role of ZAP in lowering the CpG content of SARS-CoV-2 genome.

We note that some of the studies of the evolutionary footprint of host immune mechanisms on viral genomes, focus merely on specific motifs, and ignore the overall composition of viral genomes. In many cases, this can lead to gross misinterpretation of data. For example, a phylogeny effect can be misinterpreted as an evolutionary signature. To better understand the role of ZAP and other restriction factors in the inhibition and/or evolution of viruses, a global analysis of viral genomic composition is needed. Differences observed in 14 out of 16 dinucleotides (i.e. not only CpG) point to general mechanism(s) with a global impact on the overall composition of SARS-CoV-2 genome. One of the mechanisms could be oxidative stress, although there is currently no data to support this hypothesis. Viral infection is often associated with oxidative stress, which results in producing reactive oxygen species (ROS)[39]. It has been shown that coronavirus infection causes a high level of ROS production in host cells[40]. Nucleotides, particularly guanine, are more prone to the oxidative damage caused by ROS, which oxidize guanine to 8-oxyguanine in both DNA and RNA[41]. It is known that 8-oxyguanine has a similar affinity for binding to adenine and cytosine[42]. There is a possibility that during the SARS-CoV-2 replication process[1], which includes synthesizing a negative strand from the genome followed by making a positive strand using the newly synthesized negative strand, G is substituted with U. The lower representations of UpG and GpA accompanied with higher representations of UpU and UpA in the viruses of SARS-CoV-2-like group might be due to G > U mutations induced by oxidative stress. Nevertheless, there is currently no data to support this hypothesis.

In summary, we performed several independent analyses to determine if ZAP played a role in the emergence of SARS-CoV-2. Our analyses found no evidence to suggest that ZAP exerts an evolutionary pressure on the SARS-CoV-2 genome by targeting its CpG motifs.

Data availability

The data underlying this article are available at the online supplementary material.

Received: 17 September 2021; Accepted: 28 December 2021
Published online: 14 February 2022

References

1. Romano, M., Ruggiero, A., Squeglia, F., Maga, G. & Berisio, R. A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. *Cells* https://doi.org/10.3390/cells9051267 (2020).
2. Coronaviridae Study Group of the International Committee on Taxonomy of, V. The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544. https://doi.org/10.1038/s41564-020-0695-z (2020).
3. Gussow, A. B. *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci. U S A* **117**, 15193–15199. https://doi.org/10.1073/pnas.2008176117 (2020).
4. Di Gioacchino, A. *et al.* The heterogeneous landscape and early evolution of pathogen-associated CpG dinucleotides in SARS-CoV-2. *Mol. Biol. Evol.* **38**, 2428–2445. https://doi.org/10.1093/molbev/msab036 (2021).
5. Pollock, D. D. *et al.* Viral CpG deficiency provides no evidence that dogs were intermediate hosts for SARS-CoV-2. *Mol. Biol. Evol.* **37**, 2706–2710. https://doi.org/10.1093/molbev/msaa178 (2020).
6. Subramanian, S. The long-term evolutionary history of gradual reduction of CpG dinucleotides in the SARS-CoV-2 lineage. *Biology* https://doi.org/10.3390/biology10010052 (2021).
7. Wang, Y. *et al.* Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. *Sci. Rep.* **10**, 12331. https://doi.org/10.1038/s41598-020-69342-y (2020).
8. Xia, X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol. Biol. Evol.* **37**, 2699–2705. https://doi.org/10.1093/molbev/msaa094 (2020).
9. Karlin, S., Doerfler, W. & Cardon, L. R. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?. *J. Virol.* **68**, 2889–2897. https://doi.org/10.1128/JVI.68.5.2889-2897.1994 (1994).
10. Rima, B. K. & McFerran, N. V. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J. Gen. Virol.* **78**(Pt 11), 2859–2870. https://doi.org/10.1099/0022-1317-78-11-2859 (1997).
11. Cheng, X. *et al.* CpG usage in RNA viruses: Data and hypotheses. *PLoS ONE* **8**, e74109. https://doi.org/10.1371/journal.pone.0074109 (2013).
12. Alinejad-Rokny, H., Anwar, F., Waters, S. A., Davenport, M. P. & Ebrahimi, D. Source of CpG Depletion in the HIV-1 Genome. *Mol. Biol. Evol.* **33**, 3205–3212. https://doi.org/10.1093/molbev/msw205 (2016).
13. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504. https://doi.org/10.1093/nar/8.7.1499 (1980).
14. Gao, G., Guo, X. & Goff, S. P. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. *Science* **297**, 1703–1706. https://doi.org/10.1126/science.1074276 (2002).
15. Ghimire, D., Rai, M. & Gaur, R. Novel host restriction factors implicated in HIV-1 replication. *J. Gen. Virol.* **99**, 435–446. https://doi.org/10.1099/jgv.0.001026 (2018).
16. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124–127. https://doi.org/10.1038/nature24039 (2017).
17. Ficarelli, M. *et al.* CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. *J. Virol.* https://doi.org/10.1128/JVI.01337-19 (2020).
18. Wei, Y., Silke, J. R., Aris, P. & Xia, X. Coronavirus genomes carry the signatures of their habitats. *PLoS ONE* **15**, e0244025. https://doi.org/10.1371/journal.pone.0244025 (2020).
19. Damas, J. *et al.* Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U S A* **117**, 22311–22322. https://doi.org/10.1073/pnas.2010146117 (2020).
20. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* **368**, 1016–1020. https://doi.org/10.1126/science.abb7015 (2020).
21. Digard, P., Lee, H. M., Sharp, C., Grey, F. & Gaunt, E. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. *Virus Evol.* **6**, veaa057. https://doi.org/10.1093/ve/veaa057 (2020).

22. Afrasiabi, A. *et al.* Evidence from genome wide association studies implicates reduced control of Epstein-Barr virus infection in multiple sclerosis susceptibility. *Genome Med.* **11**, 26. https://doi.org/10.1186/s13073-019-0640-z (2019).
23. Li, M. M., MacDonald, M. R. & Rice, C. M. To translate, or not to translate: Viral and host mRNA regulation by interferon-stimulated genes. *Trends Cell Biol.* **25**, 320–329. https://doi.org/10.1016/j.tcb.2015.02.001 (2015).
24. Afrasiabi, A. *et al.* The interaction of human and epstein-barr virus miRNAs with multiple sclerosis risk loci. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms22062927 (2021).
25. Ebrahimi, D., Anwar, F. & Davenport, M. P. APOBEC3 has not left an evolutionary footprint on the HIV-1 genome. *J. Virol.* **85**, 9139–9146. https://doi.org/10.1128/JVI.00658-11 (2011).
26. Ebrahimi, D., Anwar, F. & Davenport, M. P. APOBEC3G and APOBEC3F rarely co-mutate the same HIV genome. *Retrovirology* **9**, 113. https://doi.org/10.1186/1742-4690-9-113 (2012).
27. Rajaei, P. *et al.* VIRMOTIF: A user-friendly tool for viral sequence analysis. *Genes* https://doi.org/10.3390/genes12020186 (2021).
28. Li, S. Z. & Jain, A. (eds) *Encyclopedia of Biometrics* 953–953 (Springer, 2009).
29. Chiu, H. P. *et al.* Inhibition of Japanese encephalitis virus infection by the host zinc-finger antiviral protein. *PLoS Pathog.* **14**, e1007166. https://doi.org/10.1371/journal.ppat.1007166 (2018).
30. Luo, X. *et al.* Molecular mechanism of RNA recognition by zinc-finger antiviral protein. *Cell Rep.* **30**, 46-52 e44. https://doi.org/10.1016/j.celrep.2019.11.116 (2020).
31. Ebrahimi, D., Alinejad-Rokny, H. & Davenport, M. P. Insights into the motif preference of APOBEC3 enzymes. *PLoS ONE* **9**, e87679. https://doi.org/10.1371/journal.pone.0087679 (2014).
32. Di Giallonardo, F., Schlub, T. E., Shi, M. & Holmes, E. C. Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *J. Virol.* https://doi.org/10.1128/JVI.02381-16 (2017).
33. Greenbaum, B. D., Cocco, S., Levine, A. J. & Monasson, R. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proc. Natl. Acad. Sci. U S A* **111**, 5054–5059. https://doi.org/10.1073/pnas.1402285111 (2014).
34. Shackelton, L. A., Parrish, C. R. & Holmes, E. C. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.* **62**, 551–563. https://doi.org/10.1007/s00239-005-0221-1 (2006).
35. Shpaer, E. G. & Mullins, J. I. Selection against CpG dinucleotides in lentiviral genes: A possible role of methylation in regulation of viral expression. *Nucleic Acids Res.* **18**, 5793–5797. https://doi.org/10.1093/nar/18.19.5793 (1990).
36. Woo, P. C., Wong, B. H., Huang, Y., Lau, S. K. & Yuen, K. Y. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. *Virology* **369**, 431–442. https://doi.org/10.1016/j.virol.2007.08.010 (2007).
37. Bick, M. J. *et al.* Expression of the zinc-finger antiviral protein inhibits alphavirus replication. *J. Virol.* **77**, 11555–11562. https://doi.org/10.1128/jvi.77.21.11555-11562.2003 (2003).
38. Nchioua, R. *et al.* SARS-CoV-2 is restricted by zinc finger antiviral protein despite preadaptation to the low-CpG environment in humans. *MBio* https://doi.org/10.1128/mBio.01930-20 (2020).
39. Schwarz, K. B. Oxidative stress during viral infection: A review. *Free Radic. Biol. Med.* **21**, 641–649. https://doi.org/10.1016/0891-5849(96)00131-1 (1996).
40. Delgado-Roche, L. & Mesta, F. Oxidative stress as key player in severe acute respiratory syndrome coronavirus (SARS-CoV) infection. *Arch. Med. Res.* **51**, 384–387. https://doi.org/10.1016/j.arcmed.2020.04.019 (2020).
41. Schneider, J. E. Jr. *et al.* Methylene blue and rose bengal photoinactivation of RNA bacteriophages: Comparative studies of 8-oxoguanine formation in isolated RNA. *Arch. Biochem. Biophys.* **301**, 91–97. https://doi.org/10.1006/abbi.1993.1119 (1993).
42. Sekiguchi, M. & Tsuzuki, T. Oxidative nucleotide damage: Consequences and prevention. *Oncogene* **21**, 8895–8904. https://doi.org/10.1038/sj.onc.1206023 (2002).

Author contributions

A.A., D.E., and H.A.R. conceived the project. H.A.R., A.A. and A.K. conducted all analysis. M.R. generated all other figures. A.A. generated all supplementary tables. All figures and supplementary tables were generated with the supervision of D.E. All authors contributed to the discussions. A.A. and D.E. drafted the manuscript with the assistance of N.L. and Z.X., which was reviewed by all authors. All authors read and approved the final manuscript.

Funding

This study is supported by funding from the UT Health San Antonio COVID-19 Rapid Response Pilot Program and San Antonio Partnership for Precision Medicine (SAPPT) to DE and ZX (Grant No. g341947). HAR is funded by the UNSW Scientia Program Fellowship, and AA is supported by an Australian Government Research Training Program (RTP) Scholarship.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-06046-5.

Correspondence and requests for materials should be addressed to D.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.