bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evidence for ZAP-independent CpG reduction in SARS-CoV-2 genome, and pangolin coronavirus origin of 5'UTR

Ali Afrasiabi≠ [1], Hamid Alinejad-Rokny≠ [1,2,3], Nigel Lovell[4], Zhenming Xu[5], Diako Ebrahimi* [6]

[1]Systems Biology and Health Data Analytics Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, AU.

[2]School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW, 2052, AU.

[3]Health Data Analytics Program, AI-enabled Processes (AIP) Research Centre, Macquarie University, Sydney, 2109, AU.

[4]The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, AU.

[5]Department of Microbiology, Immunology and Molecular Genetics, University of Texas Health Science Center at San Antonio, San Antonio, TX, 78229, USA

[6] Biomedical Informatics Lab, Texas Biomedical Research Institute, San Antonio, TX, 78227, USA

≠ Equal contribution

* Corresponding author: debrahimi@txbiomed.org

Abstract

SARS-CoV-2, the causative agent of COVID-19, has an RNA genome, which is, overall, closely related to the bat coronavirus sequence RaTG13. However, the ACE2-binding domain of this virus is more similar to a coronavirus isolated from pangolin. In addition to this unique feature, the genome of SARS-CoV-2 (and its closely related coronaviruses) has a low CpG content. This has been postulated to be the signature of an evolutionary pressure exerted by the host antiviral protein ZAP. Here, we analyzed the sequences of a wide range of viruses using both alignment-based and alignment free approaches to investigate the origin of SARS-CoV-2 genome. Our analyses revealed a high level of similarity between the 5'UTR of SARS-CoV-2 and that of a Guangdong pangolin coronavirus. These data suggest that not only ACE2, but also the 5'UTR of SARS-CoV-2 likely has a pangolin coronavirus origin. Additionally, we performed a detailed analysis of viral genome compositions as well as expression and RNA binding data of ZAP to show that the low CpG abundance in SARS-CoV-2 is not related to an evolutionary pressure from ZAP.

Key Words: SARS-CoV-2, ZAP, CpG, Dinucleotides abundance, Motif representation, Pangolin, Bat

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19 pandemic, has a single-stranded positive RNA (+ssRNA) genome of ~30kb long, which is one of the largest known viral RNA genomes [1]. 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 these genomic features is essential to improve our current understanding of the source and evolutionary path of this virus.

So far, comparative genomic studies suggest that, overall, SARS-CoV-2 is closely related to RaTG13, which is a coronavirus isolated from a species of bat known as the *intermediate horseshoe bat* (*Rhinolophus affinis*), in the Yunnan Province of China [4-6]. However, the genomic region coding for the ACE2 cellular receptor binding domain of SARS-CoV-2 Spike protein is an exception in that it is highly similar to the sequence of a Sunda pangolin (*Manis javanica*) coronavirus found in the Guangdong Province [7-9]. Both the sequence divergence of SARS-CoV-2 genome and its unique features including a 12-base insert creating a polybasic cleavage site, have sparked extensive discussion on the origin of this new coronavirus [9, 10]. Nonetheless, it has been shown that SARS-CoV-2 was originated in bat, and pangolin was the intermediate species for its leap into human [8, 10-12].

One of the unique genomic features of SARS-CoV-2 is the low abundance of CpG [13-16]. CpG suppression is a well-known phenomenon in viruses particularly those with RNA genomes [17, 18]. It has been reported that the CpG composition of +ssRNA viral genomes often mimics the CpG content of their hosts [19]. This suggests host CpG manipulating mechanisms play a role in shaping +ssRNA viral genomes during cross-species transmission. Nevertheless, these molecular mechanisms are not fully understood. One of the suggested mechanisms is DNA cytosine methylation-induced deamination [19-21]. 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) [14]. ZAP is known to restrict viral replication by binding to the CpG rich regions of viral RNA, and subsequently inducing a viral RNA degradation process [22, 23]. The low CpG content of RNA viruses have been proposed to be linked to this ZAP-dependent depletion of CpG-rich viral RNA sequences [22-24]. Nevertheless, ZAP, despite having a broad antiviral role, does not restrict all viruses [22].

A recent study by Xia suggested that the low CpG content of SARS-CoV-2 might be due to an evolutionary pressure from ZAP [14]. 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. Furthermore, based on a previous report that ZAP identifies CpG-containing RNA [24], Xia linked the low CpG content of SARS-CoV-2 genome to the high expression level of ZAP in the intermediate host tissues [14]. This hypothesis is based on two assumptions, which are not

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

likely to be correct: first, only the frequency of CpG (but not those of other 15 dinucleotides ApA, ApC, …, TpT) is sufficient to make inferences about the origin of viruses, and second, ZAP is the main selection force against viral CpGs. A number of follow up studies have 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 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 [13]. 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 [25]. This finding was also confirmed by viral replication experiments [26]. Digard *et al*. showed that CpG abundance varies significantly across the SARS-CoV-2 genome, with envelop and ORF10 not showing CpG depletion. They show that the CpG levels of SARS and SARS-CoV-2 envelops are even higher than those of envelop from other human coronaviruses. Using a phylogenetic analysis, the authors argue 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 [27].

Here, we investigate the origin of SARS-CoV-2 using a comparative genomic approach. Additionally, we perform a detailed analysis of multiple data sets including the representations of short sequence motifs in viral genomes, ZAP expression, and CpG motif preference of ZAP to investigate the role of ZAP in reducing SARS-CoV-2 CpG level. Our analyses suggest that not only the ACE2 binding domain of SARS-CoV-2 Spike, but also its 5'UTR has a pangolin coronavirus origin. This suggests bat and pangolin coronaviruses have likely recombined at least twice (in the 5'UTR and ACE2 binding domains) to seed the formation of SARS-CoV-2. We show that the representation of almost all dinucleotides, not only CpG is different in SARS-CoV-2 compared to its more distantly related coronaviruses. For example, GpC, TpC, and ApT are all represented at significantly lower levels in the SARS-CoV-2 genome compared to the SARS-CoV genomes. Our analyses indicate that the CpG motifs preferentially targeted by ZAP do not have a lower representation than those not often recognized by ZAP. Altogether, our results provide multiple lines of evidence against the role of ZAP in shaping 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 1 and 2**). **Fig. 1A** shows the PCA of all four virus groups studied here, and **Fig. 1B** shows the results of the same analysis done on coronaviruses only. As indicated in **Fig. 1A**, all of the four virus families HIN1, HBV, HIV-1 and coronaviruses are separated using the first two PCs. PCA of only coronavirus sequences is depicted in **Fig. 1B**. As expected, SARS-COV-2, Bat-RaTG13, and Pangolin-CoV formed a cluster (SARS-CoV-2-like group), which was separated from the rest of coronavirus sequences, Human-SARS-CoV, Paguma-SARS-CoV, Viverridae-SARS-CoV, Paradoxurus-SARS-CoV, Bat-SARS-CoV, Mus-SARS-CoV and Primate-SARS-CoV (SARS-CoV group). The two groups of coronaviruses described above (shown in ovals in **Fig. 1B**) are clearly

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

separated from each other, except for three bat coronaviruses shown by black arrows (accession numbers MG772933, MG772934 and KY352407). In the PC1-PC2 plot of **Fig. 1B** these three sequences are positioned closer to the SARS-CoV-2-like group than to their own SARS-CoV group. Alignment-based comparative genomic studies have shown that SARS-CoV-2 sequences are most closely related to the Bat-RaTG13 sequence [5]. This is also apparent in the PC1-PC2 plot of **Fig. 1B** with Bat-RaTG13 (blue dot) being the closest to the SARS-CoV-2 cluster (red dots). Pangolin-CoV sequences (green dots) also position close to the SARS-CoV-2 sequences (**Fig. 1B**). The Pangolin-CoV sequence with the accession number EPI_ISL_410721, which is labeled in PubMed to be of Guangdong origin, is the second most closely related virus to the SARS-CoV-2 family. The remaining Pangolin-CoV sequences, which are labeled with Guangxi, form a sub-cluster that is located at a farther distance to the SARS-CoV-2 sequences.

5'UTR sequence similarity

To investigate the origin of SAR-CoV-2, we compared, using a moving window analysis, its genome to the sequences of Bat-RaTG13 and two Pangolin-CoV genomes, one from Guangdong (accession number EPI_ISL_410721), and one from Guangxi (accession number EPI_ISL_410539). The results are shown in **Fig. 2**, which depicts the difference between pairwise similarities of SARS-CoV-2/Bat-RaTG13 and SARS-CoV-2/Pangolin-CoV in windows of size 20bp moving along the sequence by 1bp. Windows within which Bat-RaTG13 sequence is more closely similar to SARS-CoV-2 than Pangolin-CoV to SARS-CoV-2 are shown in blue vertical bars, and the opposite is shown in red. As expected, the majority of windows are blue indicating that, overall, Bat-RaTG13 is the most similar virus to SARS-CoV-2. The windows indicated in red are those within which Pangolin-CoV is more similar to SARS-CoV-2. Overall, for both Guangdong and Guangxi sequences there are regions within which these sequences are more similar to SARS-CoV-2 (6% and 9% of windows, red bars in **Fig. 2** for Pangolin Guangxi and Guangdong, respectively). However, there are many more windows within which SARS-CoV-2 is more similar to Bat-RaTG13 (94% and 91% of windows, blue bars in Pangolin Guangxi and Guangdong, respectively panels in **Fig. 2**). The ACE2-binding domain of SARS-CoV-2 Spike is known to have a very high sequence similarity to the same region in the Guangdong pangolin coronavirus genome [7-9]. This is also indicated by two sharp red peaks in **Fig. 2B** and to a lesser extent in **Fig. 2A**. The nucleic acid as well as amino acid sequences of these two regions are shown in boxes above the similarity lines. Most importantly, our window analysis shows a region within 5'UTR that is identical in SARS-CoV-2 and Guangdong pangolin coronavirus sequences but differs in five positions between Bat-RaTG13 and SARS-CoV-2 sequences. These genomic positions are 19, 37, 91, 174 and 190, relative to the SARS-CoV-2 reference sequence (accession number NC_045512). Such a high level of similarity between the Guangdong pangolin coronavirus and SARS-CoV-2 is not observed for the Guangxi pangolin coronavirus.

As indicated in **Fig. 1**, coronaviruses used in this study form two distinct groups (SARS-CoV-2-like and SARS-CoV) based on their motifs representations (D-values). We compared, for each dinucleotide, the median D-values of these two groups (**Fig. 3** and **Supplementary Table 3**). All dinucleotide motifs (not only CpG) except for ApG were significantly different between the two groups. There is an excess of ApA, TpA, GpT, TpT, ApC, CpC and GpG, and a deficit of CpA, GpA, ApT, CpT, TpC, GpC, TpG and CpG in group one (SARS-CoV-2-like) compared to group two

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(SARS-CoV). We then performed an additional representation analysis focusing on the trinucleotide motifs that contain CpG (**Fig. 4** and **Supplementary Table 3**). Our analysis shows that the trinucleotide profiles of the two virus groups are also different. Specifically, there is a higher representation for CGC, CGG, CCG, and ACG and lower representation for CGA, CGT, TCG and GCG in group one compared to group two coronaviruses.

To determine if host ZAP level has played a role in altering viral CpG levels, we investigated the correlation between ZAP mRNA expression level and CpG motif representation in viral genomes (**Fig. 5** and **Supplementary Table 4**). To ensure our analyses are physiologically relevant, we used the ZAP expression data of common cellular/tissue reservoirs for each virus. We observed no significant correlation between CpG motif representation and ZAP mRNA expression (R^2 = 0.04, p = 0.47, **Fig. 5**).

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. 6**). 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 [28]. To further investigate the role of ZAP CpG-binding in reducing SARS-CoV-2 CpG level, we compared the representation of C(n7)G(n)CG in group one (SARS-CoV-2-like) and group two (SARS-CoV) viruses (**Supplementary Table 5**). In spite of CpG reduction in SARS-CoV-2 genome, the ZAP preferred motif C(n7)G(n)CG, is higher in the SARS-CoV-2-like group (**Fig. 7**).

In Xia's study, it is shown that the genome of one of the dog coronaviruses (accession number KC175339) has a significantly higher CpG level than the rest of canine coronaviruses. The author attributed the high CpG level of this virus to its extensive propagation in cell culture before it is sequenced. Xia argues that lack of a selection force against CpG has led to its rebound in this virus [14]. To test this rebounding hypothesis in cell lines, we quantified the CpG motif representations of HIV-1 lab strain IIIB (accession numbers A04321.1), which is commonly cloned and cultured in cell lines, and compared it with the CpG representation of HIV sequences isolated from patients. **Fig. 8** indicates the distribution of CpG representations in 844 HIV-1 isolate sequences. The CpG representation values for A04321.1 falls within the interquartile range, thus is not significantly different from the median of this distribution.

Discussion

Alignment-based comparative genomic studies indicate that the shoehorse bat coronavirus Bat-RaTG13 is the most closely related virus to SARS-CoV-2. However, the ACE2 biding domain sequence of SARS-CoV-2 is more similar to that of a pangolin coronavirus. These results suggest that SARS-CoV-2 originated via a recombination between the bat and pangolin coronaviruses [7, 8]. We have previously shown that an alignment-free based method that uses the representation of short sequence motifs can precisely identify HIV-1 subtypes [29]. Here, we used the same method to investigate the similarities between and within virus families. As

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indicated in **Fig. 1A**, this method successfully classifies different virus families. Additionally, it shows a separate cluster for SARS-CoV-2 with Bat-RaTG13 and pangolin coronaviruses in its close proximity (see SARS-CoV-2-like group in **Fig 1B**). Our method confirmed that SARS-CoV-2 sequences are closely similar to Bat-RaTG13 (MN996532) followed by a Guangdong pangolin coronavirus (EPI_ISL_410721). Additionally, our window analysis revealed that the 5'UTR of SARS-CoV-2 is highly similar to the 5'UTR of Guangdong pangolin coronavirus. Altogether, our analyses indicate that both the 5'UTR and ACE2-binding domain of SARS-CoV-2 genome have a pangolin coronavirus origin. Therefore, our data supports the hypothesis that pangolin was the intermediate species for SARS-CoV-2 transmission into humans [7, 8].

Our analyses confirmed that SARS-CoV-2 and its closely related viruses Bat-RaTG13 and Pangolin-CoV have a lower CpG content compared to other viruses (**Fig. 3**) [13-16]. However, we show that changes in motif representation is not exclusive to CpG. All dinucleotides (except for ApG) have significantly different representations in the viruses of SARS-CoV-2-like group compared to the viruses in the SARS-CoV group **(Fig. 3)**. Not only CpG, but also CpA, GpA, ApT, CpT, and TpC have lower representations in SARS-CoV-2 and its closely related viruses. Additionally, ApA, TpA, TpT, GpT, ApC, CpC, and GpG have higher representations in the viruses of SARS-CoV-2-like group. The tri-nucleotide profiles of these two virus groups are also totally distinct (**Fig. 4**). Altogether, these data suggest that the observed differences in CpG abundance between the two virus groups is part of a global genomic difference that points to a phylogeny difference rather than an exclusive selection force against CpG motifs.

Immune evasion is one of the mechanisms proposed to explain the low CpG abundance in viral genomes including coronavirus sequences [18, 19, 30-33]. Assuming that this hypothesis is correct, 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 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 groups one (SARS-CoV-2-like) and two (SARS-CoV) viruses, respectively. This means, on average, compared to SARS-CoV, SARS-CoV-2 has ~4.6 less CpGs per Kb. It is unlikely that this marginal CpG reduction in the SARS-CoV-2 genome helps the virus to evade our immune system and cause COVID-19. Importantly, a previous study shows no correlation between the pathogenicity and global CpG content of coronaviruses infecting humans [15]. 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 likely not a vital genomic feature in SARS-CoV-2 life cycle and it does not play a role in immune evasion [27]. Altogether, the overall reduction of CpG level in SARS-CoV-2 genome is likely unrelated to the pressure imposed on the virus by the innate immune response.

Among the components of our innate immune system, ZAP has been shown to play a key role in the inhibition of RNA viruses by binding to viral sequences containing CpG, and recruiting a RNA-degradation machinery [24]. A recent study shows that ZAP is capable of inhibiting SARS-CoV-2 *in vitro* [34]. Xia's study postulates that the source of SARS-CoV-2 is a bat coronavirus

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whose genome underwent further CpG reduction by ZAP after it infected an intermediate species with high ZAP expression (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 [14]. We tested this hypothesis using the CpG representation data of multiple viruses and the expression data of ZAP in tissues infected by these viruses. As indicated in **Fig. 5** we observe no correlation between the abundance of CpG in viral genomes and the ZAP mRNA level in viral reservoirs.

Our analyses show little association between the abundance of CpG across viral genomes and the location of ZAP binding peaks (**Fig. 6**). These data suggest that binding of ZAP to viral RNA is likely not governed by the global CpG abundance of viral genomes. In fact, previous studies have shown that ZAP does not inhibit all viruses and its inhibition is independent of the viral CpG content [35]. 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 [36]. Moreover, one of the mechanisms by which ZAP inhibits viruses is through the suppression of viral mRNA translation via blocking eIF4A [22]. This mechanism is independent of ZAP binding to viral RNA.

A previous study has shown that mouse ZAP preferentially binds to C(n7)G(n)CG motifs where n: A, C, G, or U [28]. 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 abundance of C(n7)G(n)CG to be lower in SARS-CoV-2 compared to other coronaviruses. Our analysis does not show a lower abundance for this motif in the SARS-CoV-2 genome (**Fig. 7**). In fact SARS-CoV-2 genome has a slightly higher %C(n7)G(n)CG. This may be considered as yet another line of evidence against the role of ZAP in lowering the CpG content of SARS-CoV-2 genome.

In Xia's study it is postulated that in the absence of ZAP, pressure against CpG is relieved and viruses evolve to replenish their lost CpGs [14]. We tested this hypothesis using the HIV-1 molecular clone IIIB. Our analysis did not show a significant difference between the CpG content of HIV IIIB and those of HIV sequences isolated from patients (**Fig. 8**). This suggests the abundance of CpG motif in viral genomes is likely independent of ZAP-induced immune response in the virus reservoir.

We notice 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 15 out of 16 dinucleotides (i.e. not only CpG) (**Fig. 3**) point to general mechanism(s) with a global impact on the overall composition of SARS-CoV-2 genome. One of these mechanisms might be oxidative stress, although there is currently no data to support this hypothesis. Viral infection is often associated oxidative stress, which results in producing reactive oxygen species (ROS) [37]. It has

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

been shown that coronavirus infection causes a high level of ROS production in host cells [38]. Nucleotides, particularly guanine, are very sensitive to the ROS, which oxidize guanine to 8-oxyguanine in both DNA and RNA [39]. 8-oxyguanine has a similar affinity for binding to adenine and cytosine [40]. It is possible 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 T. Lower representation of CpG, TpG, and GpA accompanied with higher representation of CpT, TpT and TpA in the viruses of SARS-CoV-2-like group (**Fig. 3**) might be due to G>T mutations induced by oxidative stress. Nevertheless, there is currently no data to support this hypothesis.

In summary, we performed several independent analyses to investigate the source of SARS-CoV-2 and determine if ZAP played a role in the emergence of this virus. The high degree of similarity between the 5'UTRs of SARS-CoV-2 and the coronavirus isolated from the Guangdong pangolin supports the hypothesis that pangolin was the intermediate host for SARS-CoV-2. The 5'UTR of viral genomes have important biological functions, including viral transcription and translation [41, 42]. It remains to be determined if the high pathogenicity and transmissibility of SARS-CoV-2 is due, at least in part, to its unique 5'UTR sequence. Our analyses find no evidence to suggest that ZAP exerts an evolutionary pressure on the SARS-CoV-2 genome by targeting its CpG motifs. This, however, does not imply ZAP plays no role in inhibiting SARS-CoV-2. In fact, it has recently been shown that ZAP can inhibit SARS-CoV-2 infection *in vitro* [34].

Methods

Viral sequences acquisition

We analyzed a total of 3,967 full-length genomic sequences from different coronaviruses: 3,621 Human-SARS-CoV, 1 Paguma-SARS-CoV, 3 Viverridae-SARS-CoV, 6 Paradoxurus-SARS-CoV, 45 Bat-SARS-CoV, 41 Mus-SARS-CoV, 1 Primate-SARS-CoV, 242 Human-SARS-CoV-2, 1 Bat-RaTG13, 6 Pangolin-CoV. In addition, full-length sequences of 2,021 HIV-1, 12 HIV-2, 11 Adeno-associated virus, 4 Adeno-associated virus-1, 11 Colorado tick fever virus, 91 Flu H1N1 virus, 7365 H3N2 virus, 141 HBV, 6 HCV, 8 HTL-V2, 6 Human bocavirus, 576 JC polyomavirus, 3 Mouse mammary tumor virus, 9 Murine polyomavirus, 2 Papovavirus Hamster, 2 Parvovirus ADV-G, 4 Poliovirus type 2, and 7 SIV-CPZ were analyzed. GenBank (www.ncbi.nlm.nih.gov) was used to retrieve all 14,246 viral sequences (**Supplementary Table 1**).

Analysis of motif representation

We quantified the representation of short sequence motifs (di- and tri-nucleotides) in all viral genomes using our previously reported D-value method [43, 44]. 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 the total count of the motif in the sequence divided by the total count of all other motifs with the same length. An example of a D-value for a tri-nucleotide sequence is given in Eq. 1.

$$D = \frac{ACG_{obs}}{ACG_{exp}} = \frac{ACG_{obs} \times C_{obs}}{AC_{obs} \times CG_{obs}} \qquad \text{Eq. 1}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To demonstrate the ability of our nonalignment-based D-value method in classifying viral families we performed principal component analysis (PCA). D-values of all dinucleotide and trinucleotide motifs in all sequences form a matrix, which is used as an input for PCA.

We performed two separate principal component analyses. The first analysis was performed on all coronaviruses as well as flu H1N1, HIV, and HBV sequences to determine if analysis of motif representations can distinguish these viruses from coronaviruses. The second PCA was performed on coronavirus sequences to identify viruses whose genome show high similarities to the SARS-CoV-2 genome. Mann–Whitney test was used to determine the difference in the median of D-value between these two groups of coronaviruses.

Analysis of association between CpG abundance and ZAP level

We interrogated the association between viral CpG representation (D-value) and *ZC3HAV1* (ZAP) mRNA expression level using a linear regression model. For viruses with more than one sequence we used an average D-value in our analyses. *ZC3HAV1* mRNA expression level data of common reservoirs for Human-SARS-CoV-2, Human-SARS-CoV, Mus-SARS-CoV, Paguma-SARS-CoV, HIV1, HIV2, H1N1, H3N2, HCV, HBV, Human bocavirus, Murine polyomavirus, Poliovirus type 2 and MMTV viruses were obtained from FANTOM5 [45].

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 [24, 46]. We then calculated the CpG density across the JEV and HIV genomes using a sliding window analysis method. We used 200bp window and 1Bp sliding for JEV and 250bp window and 1Bp sliding for HIV.

The motif C(n7)G(n)CG has recently been demonstrated to be the optimal binding motif for Mouse ZAP [28]. To further investigate the role of ZAP in reducing CpG abundance in shaping the genome of SARS-CoV-2, we compared the relative abundance of C(n7)G(n)CG (#C(n7)G(n)CG / #CG, reported per kilobase) in the two viral groups of SARS-CoV-2-like and SARS-CoV. We used Mann–Whitney test to determine the difference between the two coronavirus groups).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and Figures legend:

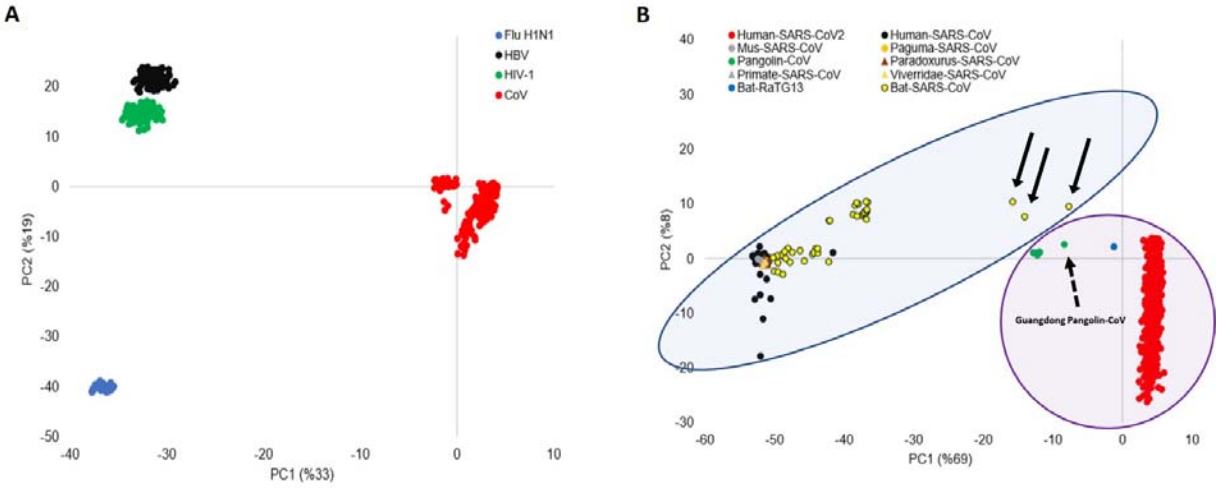


Figure 1: PCA of viral motifs representations. D-values of all dinucleotide and trinucleotide motifs in all viral sequences form a matrix, which is used as an input for PCA. (A) PC1-PC2 plot shows four clusters, one for each virus family: H1N1, SARS-CoV, HBV, and HIV-1. (B) PC1-PC2 plot classifies SARS-CoV viruses into two clusters. SARS-COV-2, Bat-RaTG13, and Pangolin-CoV formed a cluster (SARS-CoV-2-like group), which is separated from the rest of coronavirus sequences Human-SARS-CoV, Paguma-SARS-CoV, Viverridae-SARS-CoV, Paradoxurus-SARS-CoV, Bat-SARS-CoV, Mus-SARS-CoV and Primate-SARS-CoV (SARS-CoV group). SARS-CoV-2-like and SARS-CoV groups are highlighted with purple and blue circles. The three Bat-SARS-CoV sequences are located close to the SARS-CoV-2-like group, which is shown by black arrows. Guangdong Pangolin-CoV is shown by a broken arrow.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

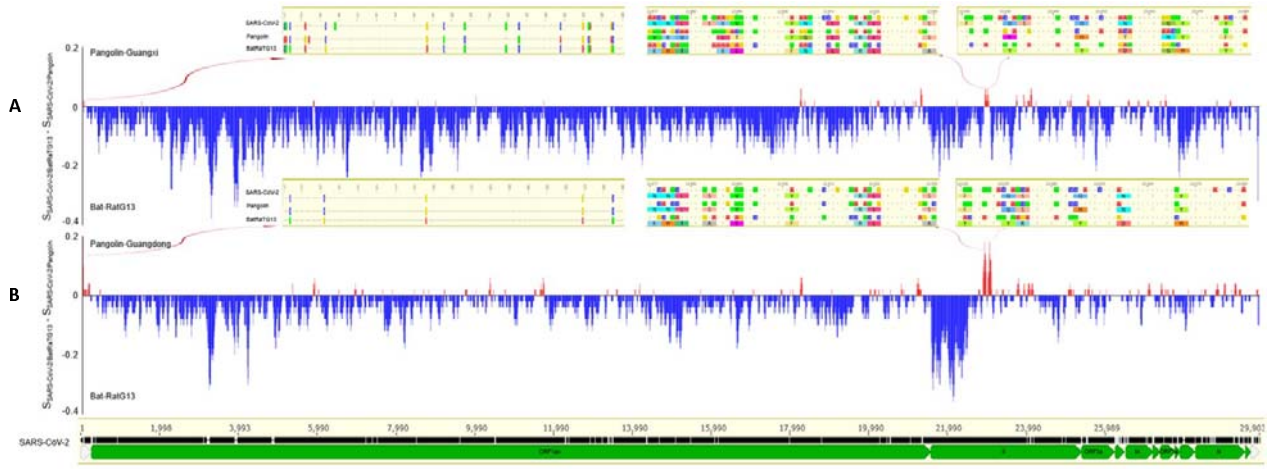


Figure 2: 5'UTR sequence similarity between SARS-CoV-2 and pangolin-CoV. The sequences of Bat-RaTG13 and two Pangolin-CoV genomes, one from Guangdong (accession number EPI_ISL_410721), and one from Guangxi (accession number EPI_ISL_410539) were compared with the SARS-CoV-2 reference sequence. A 20bp sliding window moving by 1bp was used to quantify the difference between two pairwise similarity measures (SARS-CoV-2/Bat-RaTG13 and SARS-CoV-2/Pangolin-CoV). Windows within which Bat-RaTG13 sequence is more closely similar to SARS-CoV-2 than Pangolin-CoV is to SARS-CoV-2 are shown in blue vertical bars, and the opposite is shown in red.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

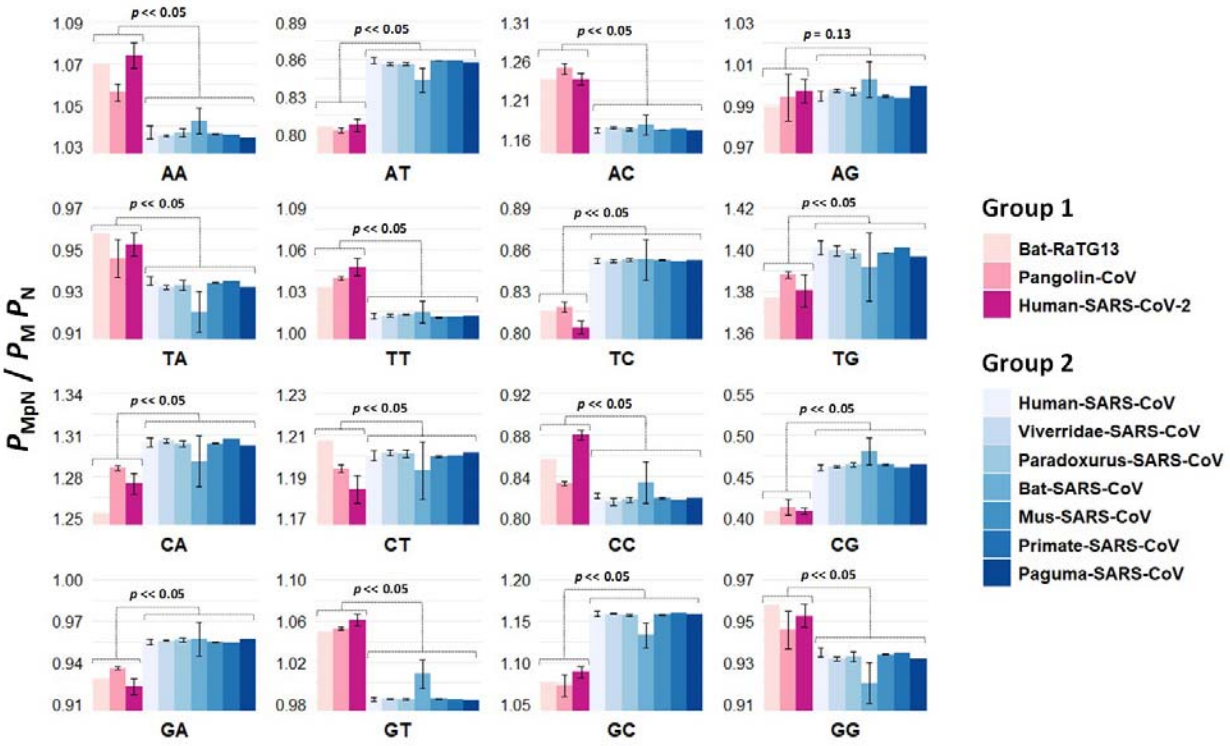


Figure 3: Comparison of dinucleotide motif representations between SRAS-CoV-2-like and SARS-CoV groups. D-values of each dinucleotide were compared between the two virus groups SARS-CoV-2-like (SARS-COV-2, Bat-RaTG13, and Pangolin-CoV) and SARS-CoV (Human-SARS-CoV, Paguma-SARS-CoV, Viverridae-SARS-CoV, Paradoxurus-SARS-CoV, Bat-SARS-CoV, Mus-SARS-CoV and Primate-SARS-CoV). Mann–Whitney test was used to examine the difference in the median of D-values between the two coronavirus groups.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

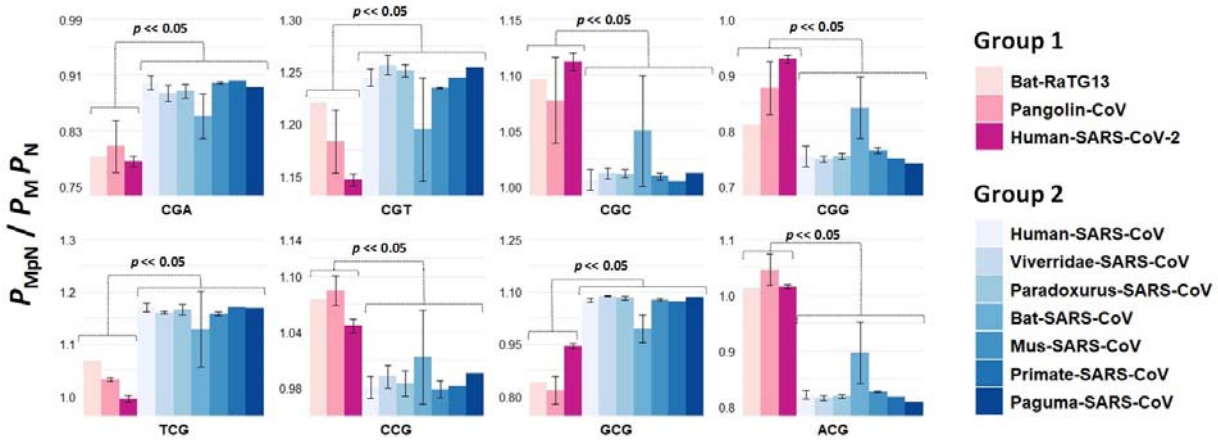


Figure 4: Comparison of CpG trinucleotide motif representations between SRAS-CoV-2-like and SARS-CoV groups. D-values of each trinucleotide were compared between the two virus groups SARS-CoV-2-like (SARS-COV-2, Bat-RaTG13, and Pangolin-CoV) and SARS-CoV (Human-SARS-CoV, Paguma-SARS-CoV, Viverridae-SARS-CoV, Paradoxurus-SARS-CoV, Bat-SARS-CoV, Mus-SARS-CoV and Primate-SARS-CoV). Mann–Whitney test was used to examine the difference in the median of D-value between the two coronavirus groups.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

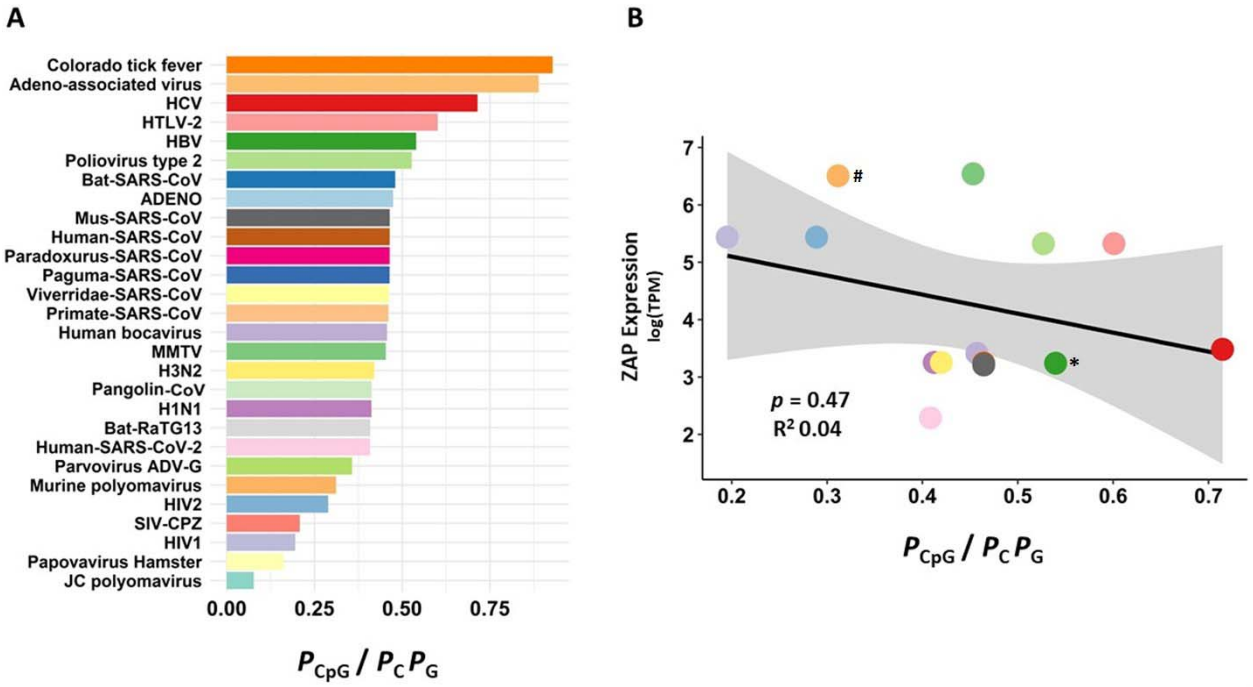


Figure 5: Correlation between ZAP mRNA level and viral CpG motif representation. (A) D-values of CpG motifs in different viruses. (B) Linear regression model was used to test the correlation between the CpG D-values of viral genomes and ZAP mRNA expression levels in common virus reservoirs. Colors match in panels A and B. All viruses in panel B are RNA viruses except HBV (*), which has a partially double stranded circular DNA genome and Murine polyomavirus (#), which is a double stranded DNA virus.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

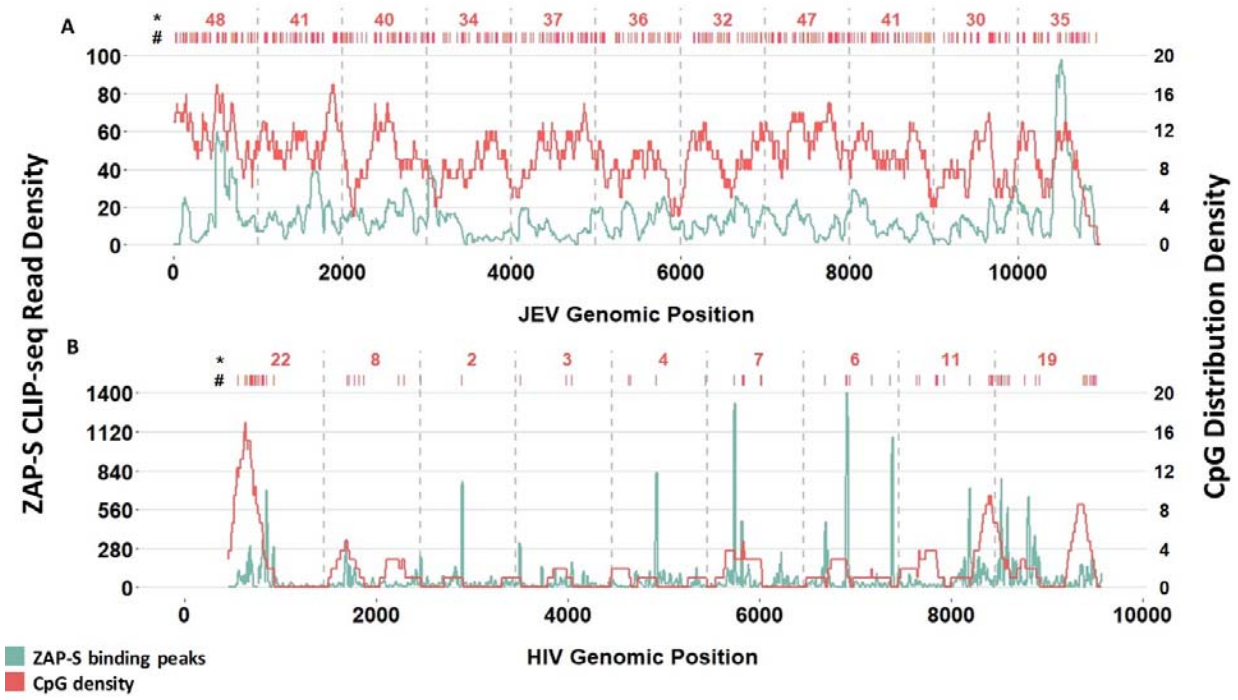


Figure 6: Co-location of ZAP binding regions and CpG motifs. Overlaying of the ZAP binding peaks and CpG densities in (A) JEV genome (Japanese Encephalitis Virus) and (B) HIV-1. The ZAP binding peaks (density of reads aligned to the genome) are estimated using a 250bp sliding window moving by 1bp along the viral genomes. The CpG density was calculated using the same sliding window analysis method, except we used a 200bp window sliding by 1bp in JEV and a 250bp 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 1Kb.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

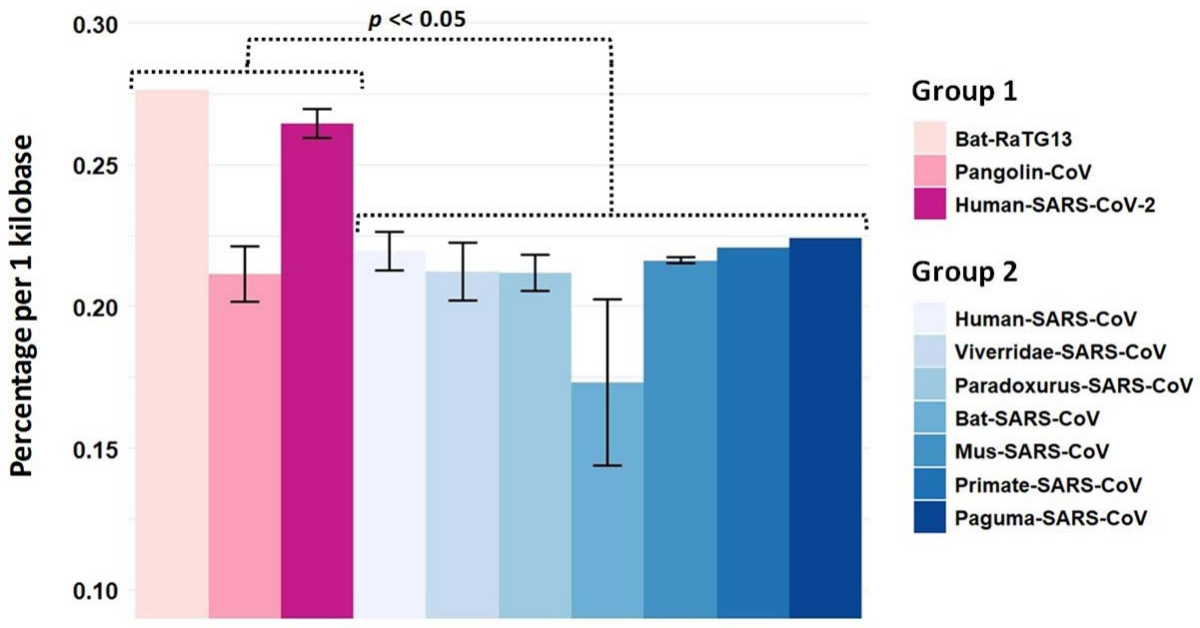


Figure 7: Comparison of the abundance of ZAP optimal binding motif C(n7)G(n)CG in viruses of SARS-CoV-2-like and SARS-CoV groups. The percentages of ZAP optimal binding motif C(n7)G(n)CG per 1kb in the two coronavirus groups we compared. The SARS-CoV-2-like group includes SARS-COV-2, Bat-RaTG13, and Pangolin-CoV. The SARS-CoV group includes Human-SARS-CoV, Paguma-SARS-CoV, Viverridae-SARS-CoV, Paradoxurus-SARS-CoV, Bat-SARS-CoV, Mus-SARS-CoV, and Primate-SARS-CoV. Mann–Whitney test was used to determine the difference in the median of D-values between these two coronavirus groups.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

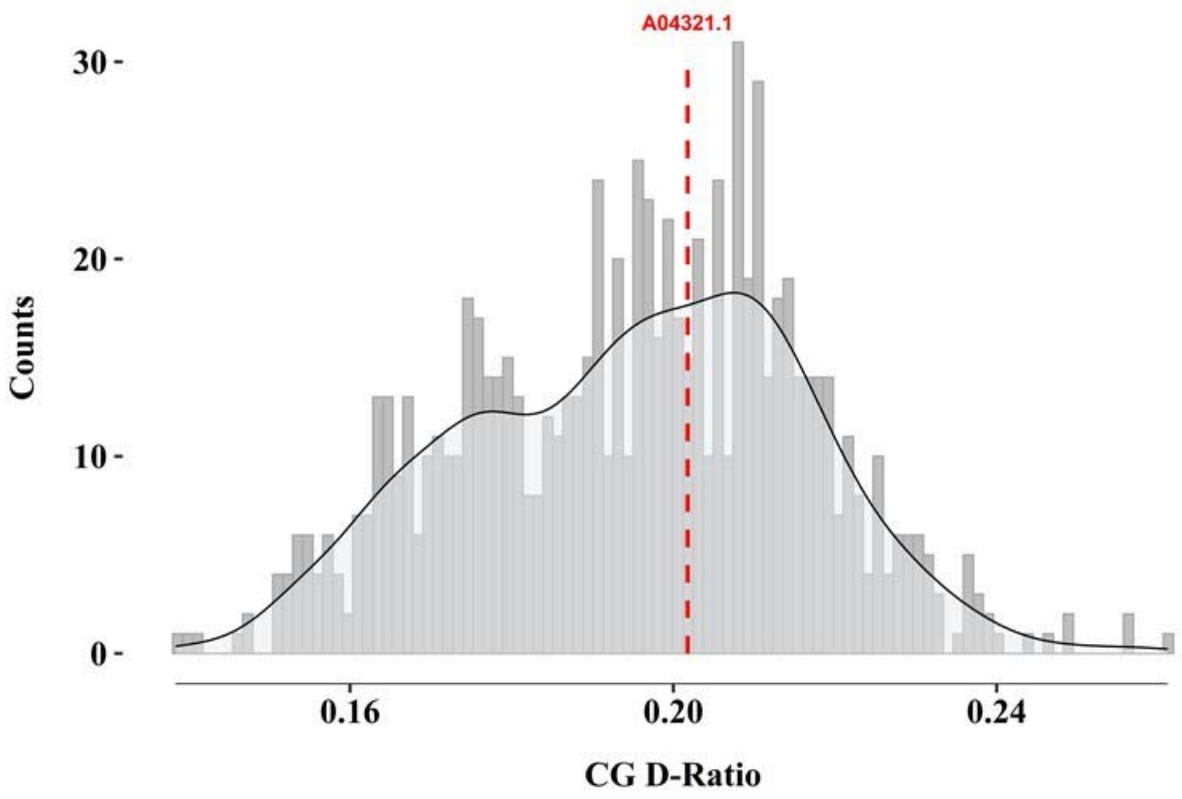


Figure 8: Comparison of CpG abundance in the HIV-1 molecular clone IIIB and HIV-1 isolates obtained from patients. A total of 844 HIV-1 sequences were used to generate a distribution of CpG D-values. The CpG D-value of HIV-1 IIIB (accession number A04321.1) is shown by a vertical dashed line.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Material

Supplementary Table 1: List of accession numbers of all viral genomes used in this study.

Supplementary Table 2: D-values of dinucleotide and trinucleotide motifs used as an input for the first PCA.

Supplementary Table 3: D-values of dinucleotide and CpG trinucleotide motifs of all coronaviruses used in this study.

Supplementary Table 4: Zap (ZC3HAV1) mRNA expression level (TPM) in the common reservoir of selected viruses.

Supplementary Table 5: Abundance of C(n7)G(n)CG motif in coronaviruses used in this study.

Supplementary Table 6: CpG D-values of HIV-1 isolate sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data Availability

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

Conflict of interest

The authors declare no competing financial and non-financial interests.

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. This study is also supported by the UNSW Scientia Fellowship Program to HAR, and a the Australian Government Research Training Program (RTP) Scholarship to AA.

Contributions

AA, DE, and HAR conceived the project. HAR and AA conducted all analysis. HAR generated Figures 1 and 2. AA generated all other figures and tables with the supervision of DE and HAR. All authors contributed to the discussions. AA and DE drafted the manuscript, which was reviewed by all authors. All authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Romano, M., et al., *A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping.* Cells, 2020. **9**(5).

2. Gussow, A.B., et al., *Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses.* Proc Natl Acad Sci U S A, 2020. **117**(26): p. 15193-15199.

3. 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, 2020. **5**(4): p. 536-544.

4. Jaimes, J.A., et al., *Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop.* J Mol Biol, 2020. **432**(10): p. 3309-3325.

5. Zhou, H., et al., *A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein.* Curr Biol, 2020. **30**(11): p. 2196-2203 e3.

6. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.

7. Lam, T.T., et al., *Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins.* Nature, 2020. **583**(7815): p. 282-285.

8. Xiao, K., et al., *Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins.* Nature, 2020. **583**(7815): p. 286-289.

9. Frutos, R., et al., *COVID-19: Time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans.* Infect Genet Evol, 2020. **84**: p. 104493.

10. Boni, M.F., et al., *Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.* Nat Microbiol, 2020.

11. Zhao, X., et al., *Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2.* J Virol, 2020. **94**(18).

12. Zhai, X., et al., *Comparison of Severe Acute Respiratory Syndrome Coronavirus 2 Spike Protein Binding to ACE2 Receptors from Human, Pets, Farm Animals, and Putative Intermediate Hosts.* J Virol, 2020. **94**(15).

13. Pollock, D.D., et al., *Viral CpG Deficiency Provides No Evidence That Dogs Were Intermediate Hosts for SARS-CoV-2.* Mol Biol Evol, 2020. **37**(9): p. 2706-2710.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. Xia, X., *Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense.* Mol Biol Evol, 2020. **37**(9): p. 2699-2705.

15. di Gioacchino, A., et al., *The heterogeneous landscape and early evolution of pathogen-associated CpG and UpA dinucleotides in SARS-CoV-2.* bioRxiv, 2020.

16. Wang, Y., et al., *Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames.* Sci Rep, 2020. **10**(1): p. 12331.

17. Rima, B.K. and N.V. McFerran, *Dinucleotide and stop codon frequencies in single-stranded RNA viruses.* J Gen Virol, 1997. **78 (Pt 11)**: p. 2859-70.

18. Karlin, S., W. Doerfler, and L.R. Cardon, *Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?* J Virol, 1994. **68**(5): p. 2889-97.

19. Cheng, X., et al., *CpG usage in RNA viruses: data and hypotheses.* PLoS One, 2013. **8**(9): p. e74109.

20. Bird, A.P., *DNA methylation and the frequency of CpG in animal DNA.* Nucleic Acids Res, 1980. **8**(7): p. 1499-504.

21. Alinejad-Rokny, H., et al., *Source of CpG Depletion in the HIV-1 Genome.* Mol Biol Evol, 2016. **33**(12): p. 3205-3212.

22. Ghimire, D., M. Rai, and R. Gaur, *Novel host restriction factors implicated in HIV-1 replication.* J Gen Virol, 2018. **99**(4): p. 435-446.

23. Gao, G., X. Guo, and S.P. Goff, *Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein.* Science, 2002. **297**(5587): p. 1703-6.

24. Takata, M.A., et al., *CG dinucleotide suppression enables antiviral defence targeting non-self RNA.* Nature, 2017. **550**(7674): p. 124-127.

25. Damas, J., et al., *Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates.* bioRxiv, 2020.

26. Shi, J., et al., *Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2.* Science, 2020. **368**(6494): p. 1016-1020.

27. Digard, P., et al., *Intra-genome variability in the dinucleotide composition of SARS-CoV-2.* bioRxiv, 2020.

28. Luo, X., et al., *Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein.* Cell Rep, 2020. **30**(1): p. 46-52 e4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Ebrahimi, D., H. Alinejad-Rokny, and M.P. Davenport, *Insights into the motif preference of APOBEC3 enzymes.* PLoS One, 2014. **9**(1): p. e87679.

30. Shackelton, L.A., C.R. Parrish, and E.C. Holmes, *Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses.* J Mol Evol, 2006. **62**(5): p. 551-63.

31. Greenbaum, B.D., et al., *Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses.* Proc Natl Acad Sci U S A, 2014. **111**(13): p. 5054-9.

32. Woo, P.C., et al., *Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses.* Virology, 2007. **369**(2): p. 431-42.

33. Shpaer, E.G. and J.I. Mullins, *Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression.* Nucleic Acids Res, 1990. **18**(19): p. 5793-7.

34. Nchioua, R., et al., *The Zinc Finger Antiviral Protein restricts SARS-CoV-2.* bioRxiv, 2020.

35. Bick, M.J., et al., *Expression of the zinc-finger antiviral protein inhibits alphavirus replication.* J Virol, 2003. **77**(21): p. 11555-62.

36. Ficarelli, M., et al., *CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms.* J Virol, 2020. **94**(6).

37. Schwarz, K.B., *Oxidative stress during viral infection: a review.* Free Radic Biol Med, 1996. **21**(5): p. 641-9.

38. Delgado-Roche, L. and F. Mesta, *Oxidative Stress as Key Player in Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) Infection.* Arch Med Res, 2020. **51**(5): p. 384-387.

39. 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, 1993. **301**(1): p. 91-7.

40. Sekiguchi, M. and T. Tsuzuki, *Oxidative nucleotide damage: consequences and prevention.* Oncogene, 2002. **21**(58): p. 8895-904.

41. Miao, Z., et al., *Secondary structure of the SARS-CoV-2 5'-UTR.* RNA Biol, 2020: p. 1-10.

42. Kim, D., et al., *The Architecture of SARS-CoV-2 Transcriptome.* Cell, 2020. **181**(4): p. 914-921 e10.

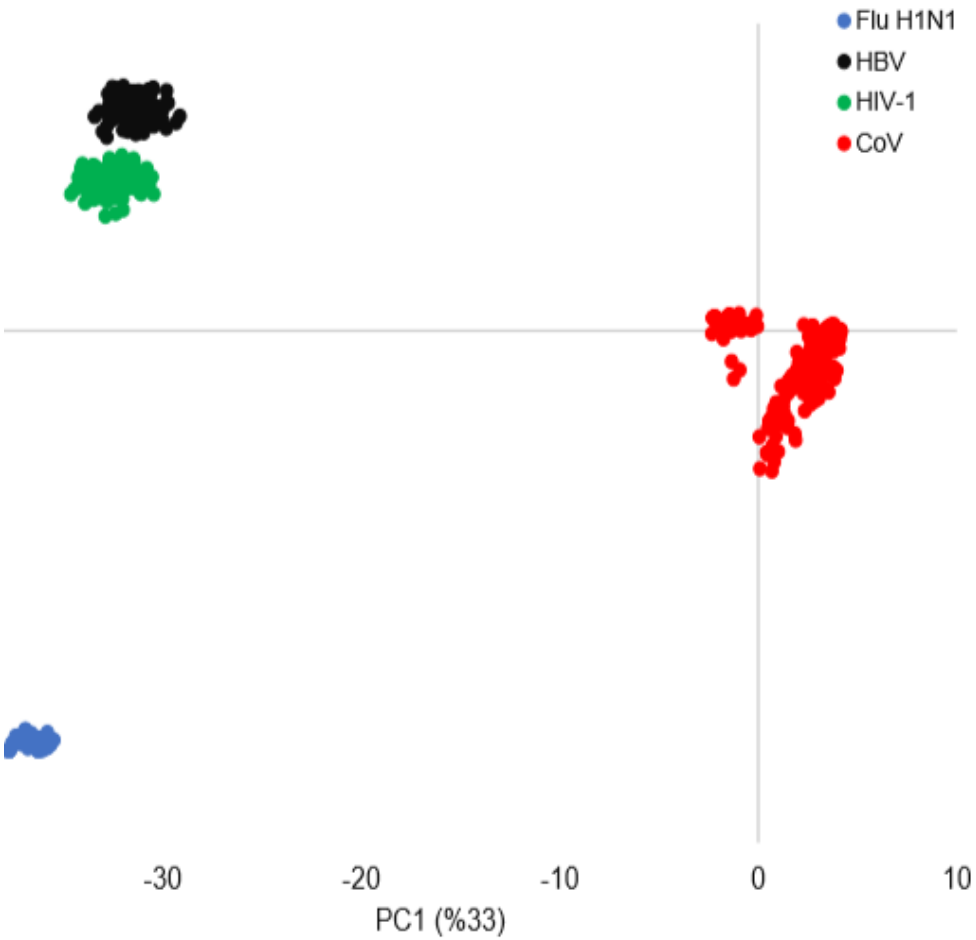
bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

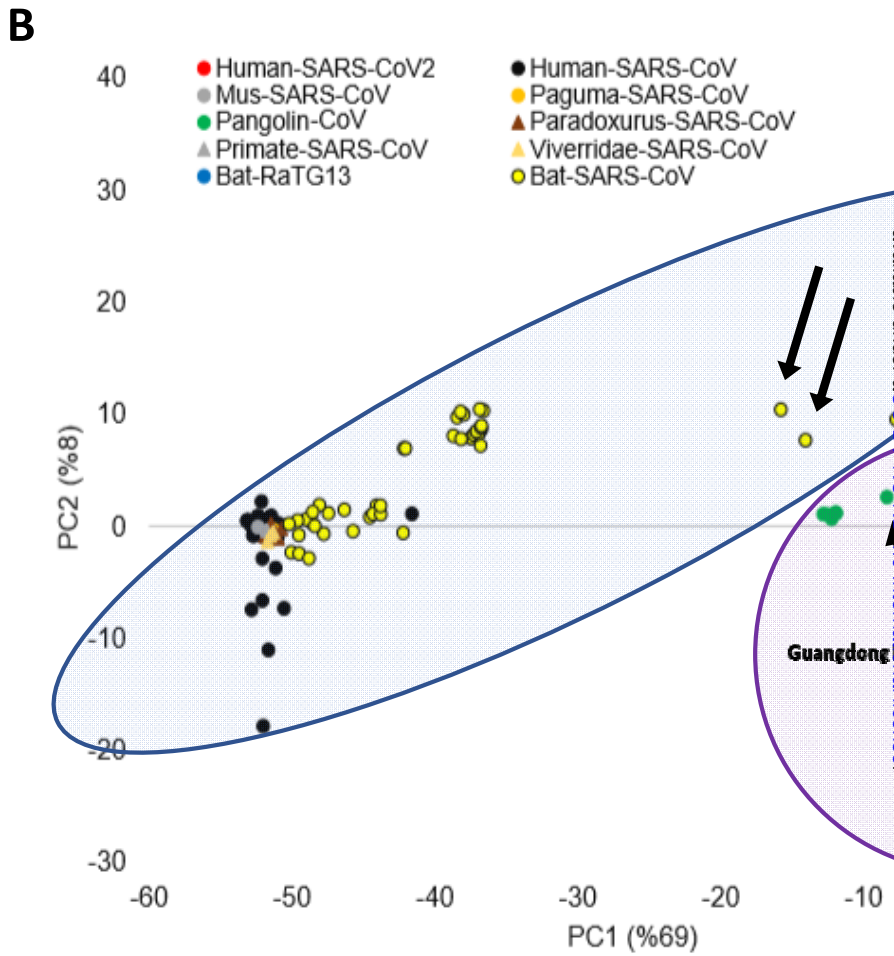
43. Ebrahimi, D., F. Anwar, and M.P. Davenport, *APOBEC3 has not left an evolutionary footprint on the HIV-1 genome.* J Virol, 2011. **85**(17): p. 9139-46.

44. Ebrahimi, D., F. Anwar, and M.P. Davenport, *APOBEC3G and APOBEC3F rarely co-mutate the same HIV genome.* Retrovirology, 2012. **9**: p. 113.

45. Lizio, M., et al., *Gateways to the FANTOM5 promoter level mammalian expression atlas.* Genome Biol, 2015. **16**: p. 22.

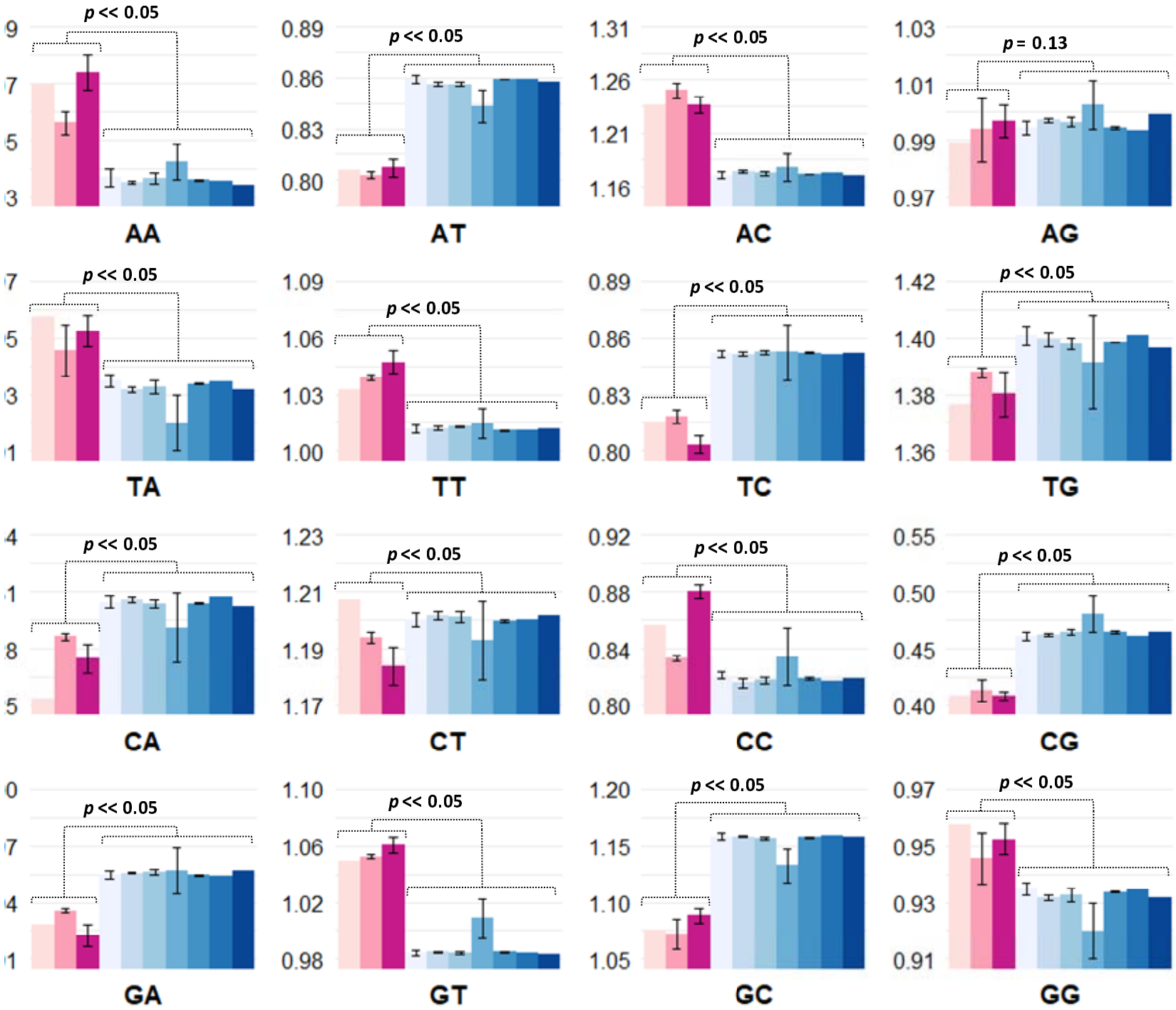
46. Chiu, H.P., et al., *Inhibition of Japanese encephalitis virus infection by the host zinc-finger antiviral protein.* PLoS Pathog, 2018. **14**(7): p. e1007166.



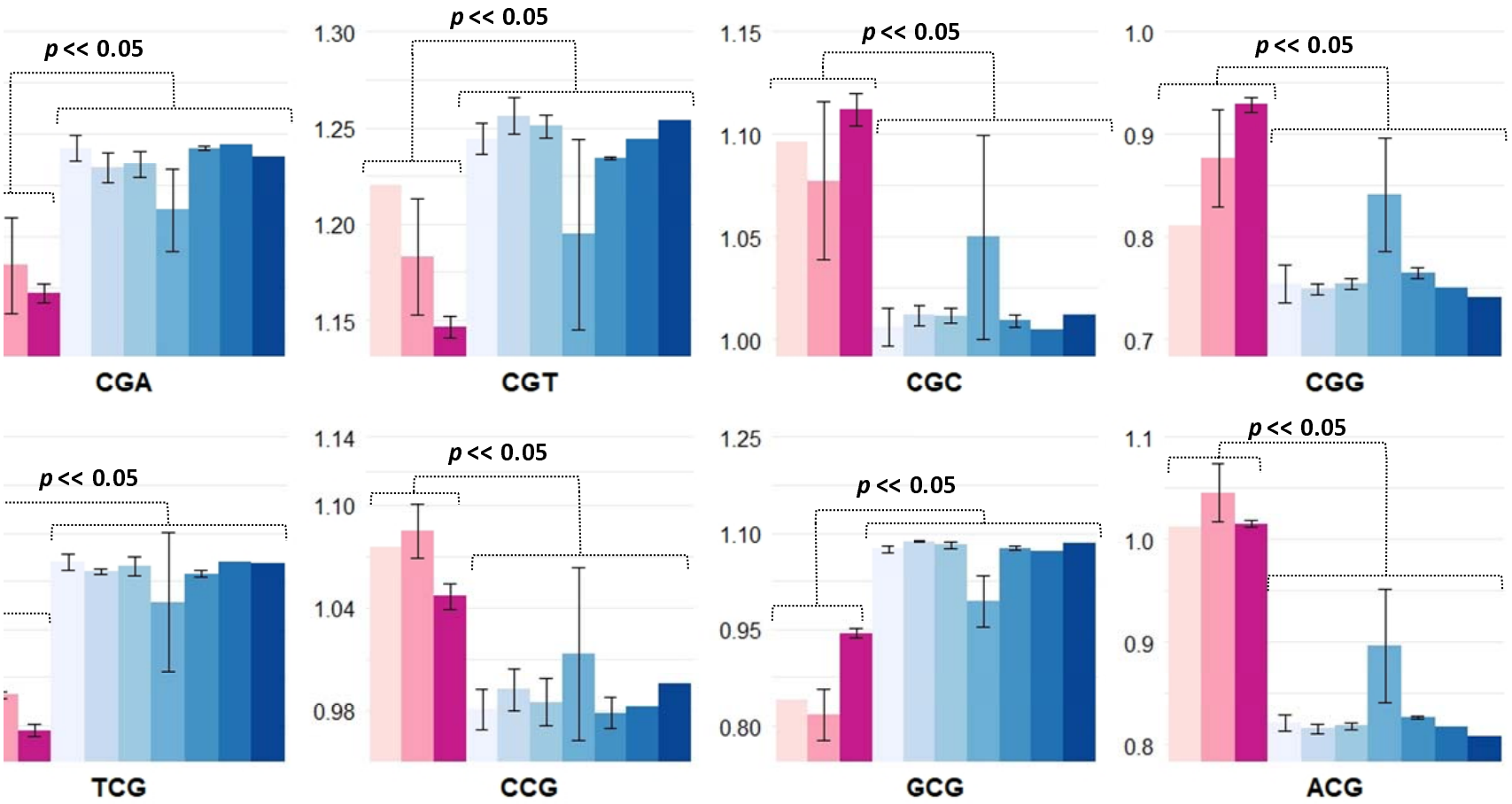


bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

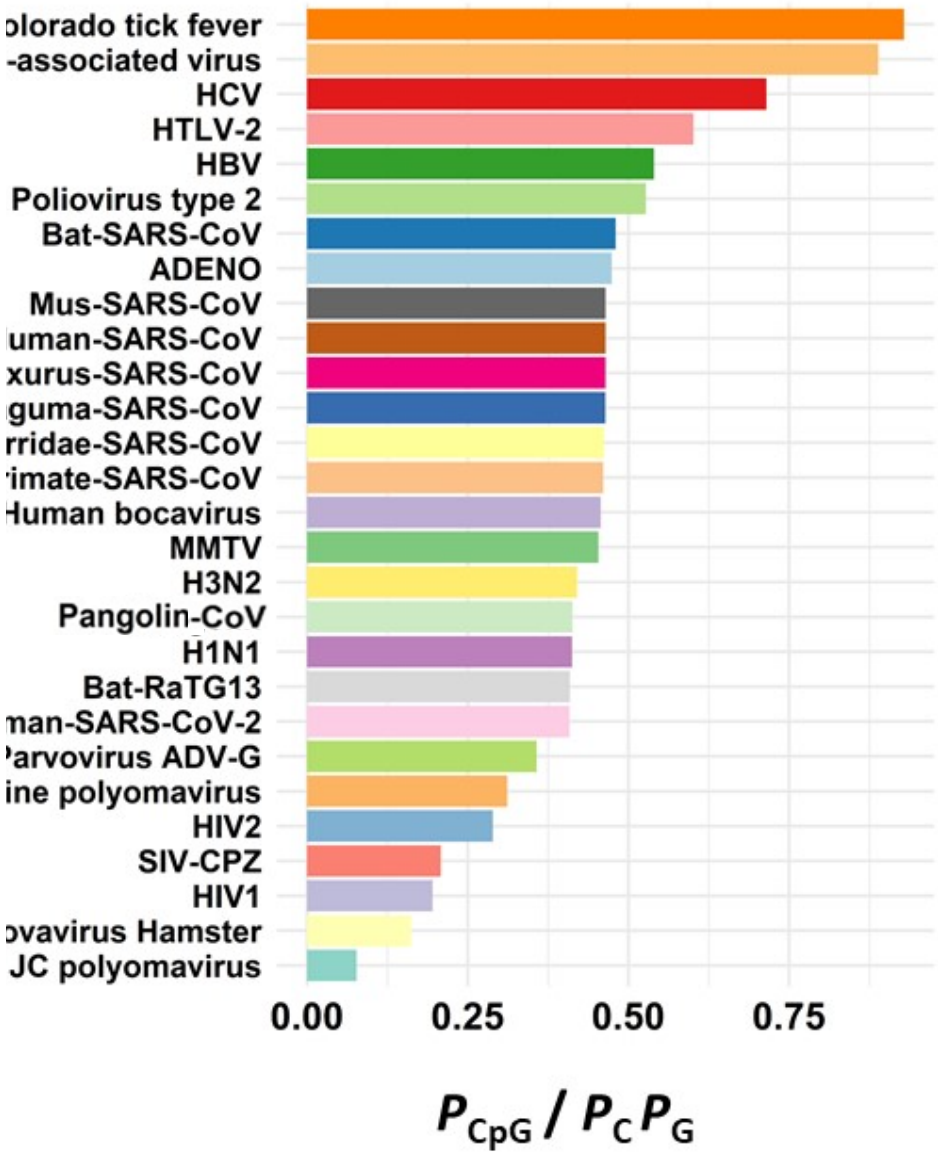
bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

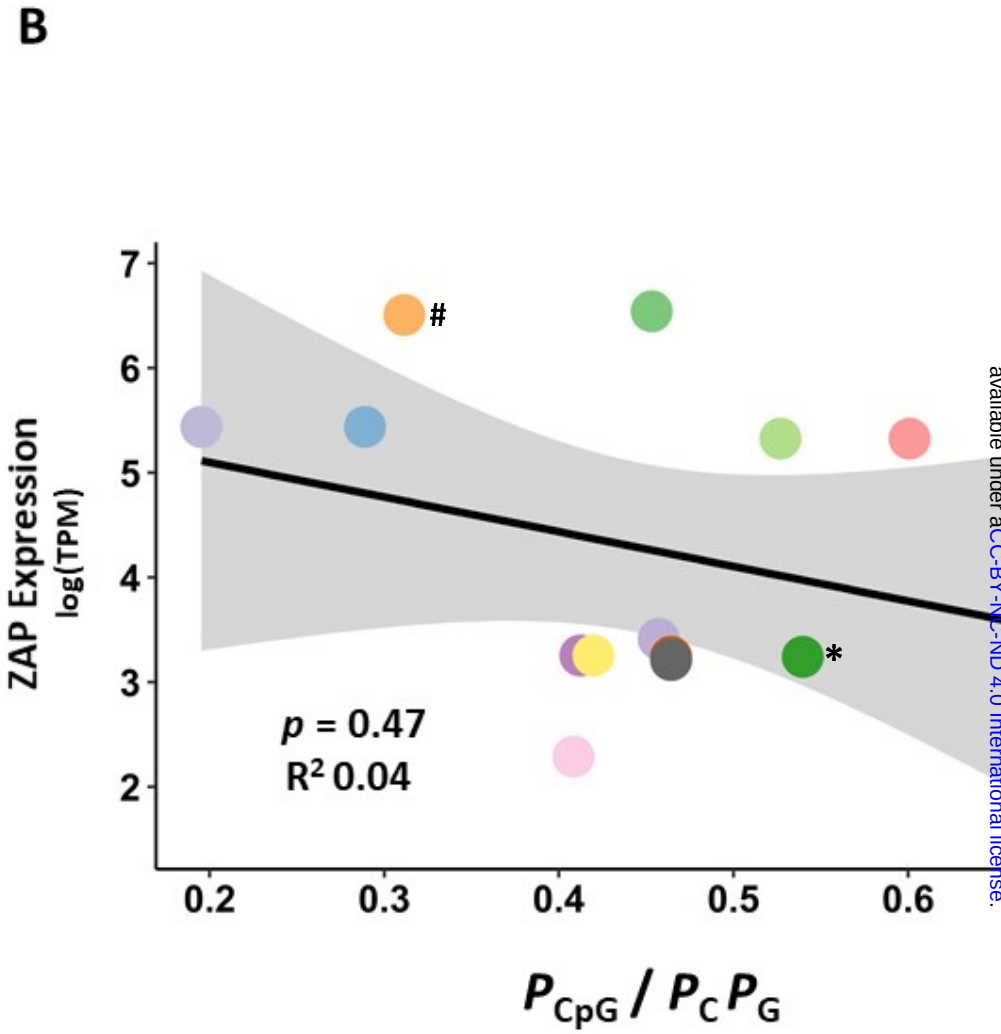


bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

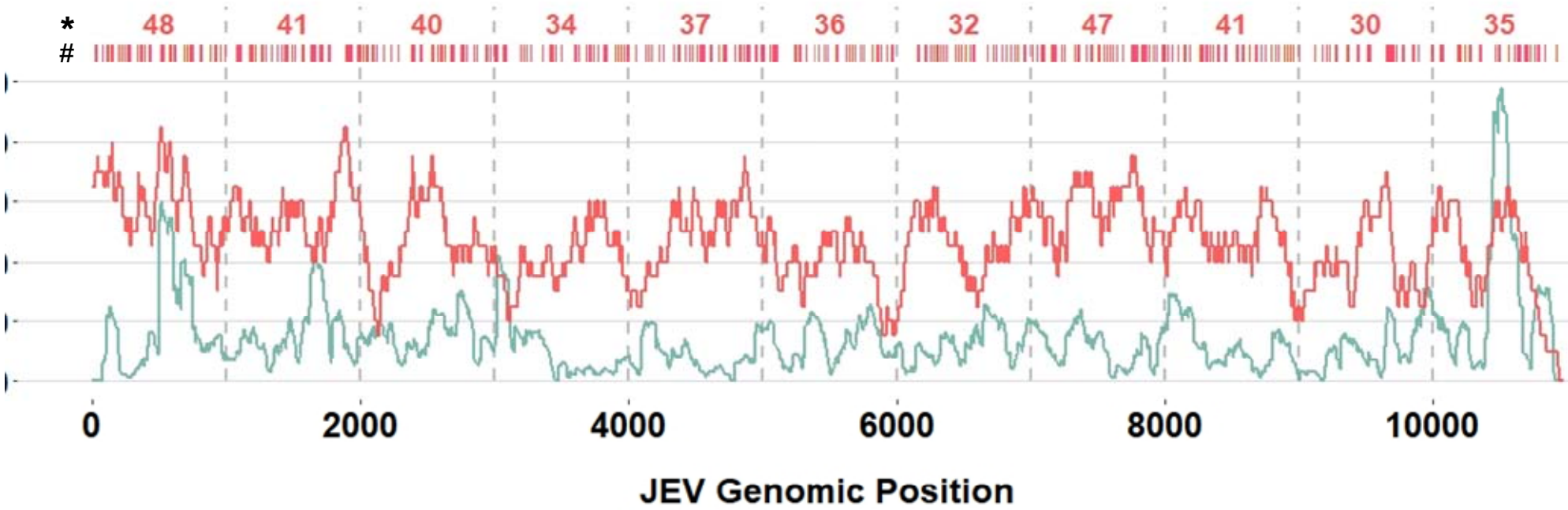


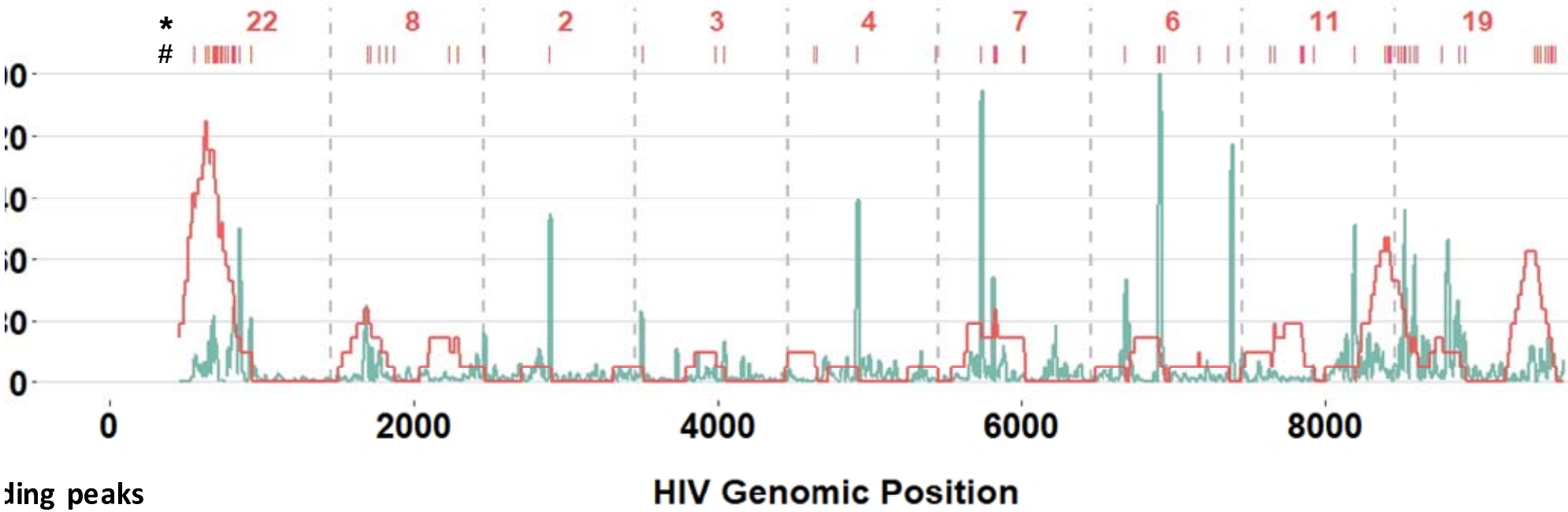
bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





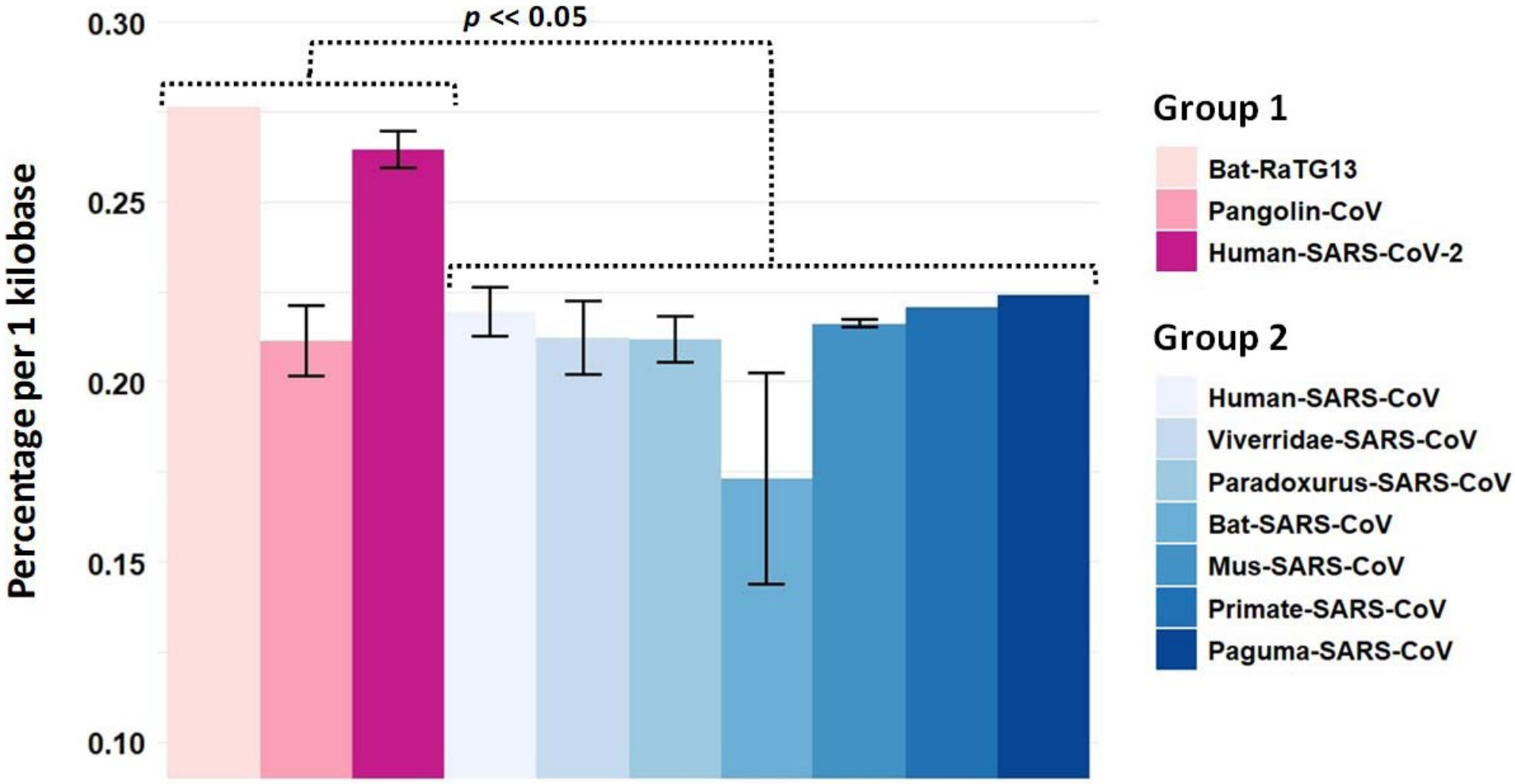
bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



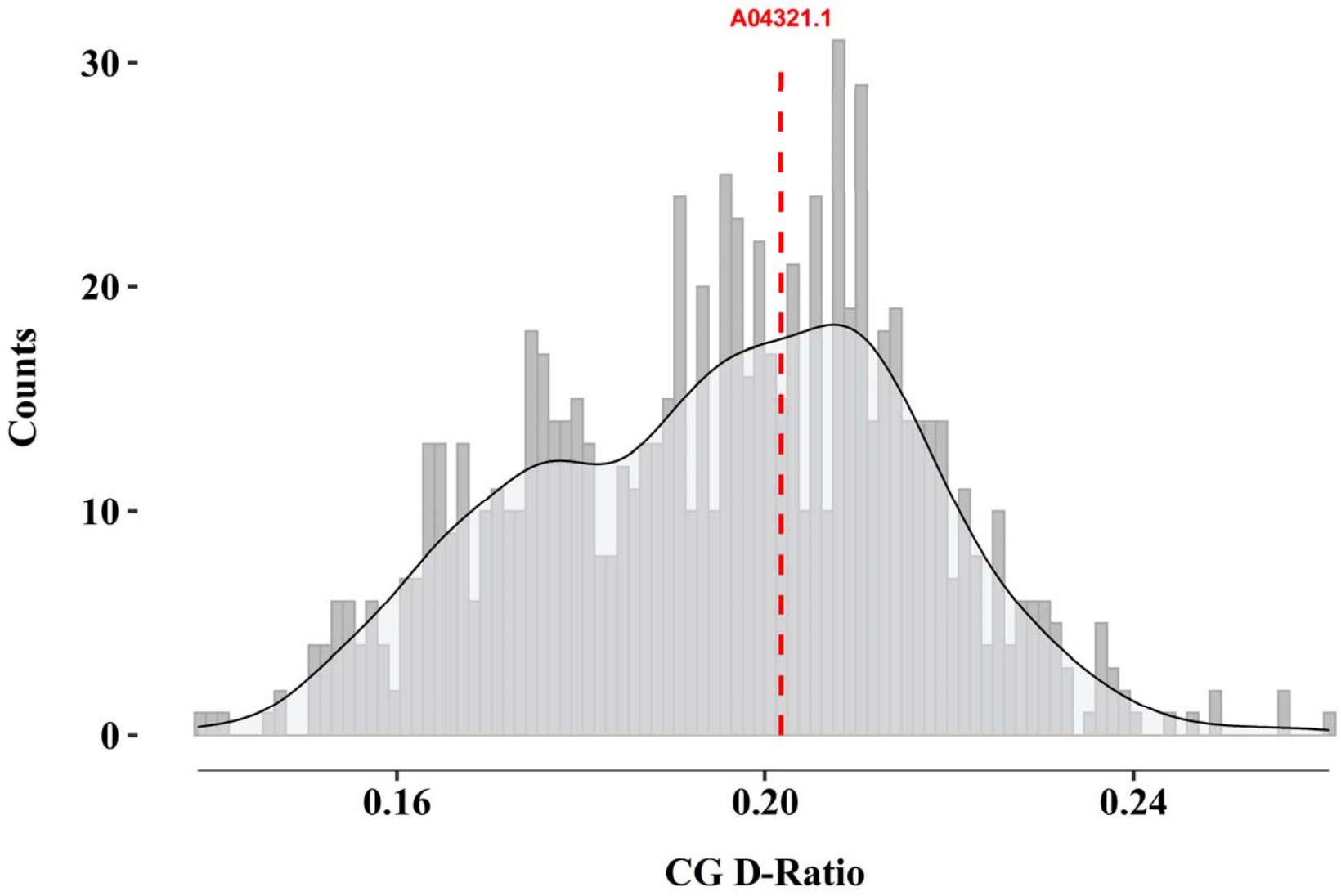


ding peaks
ty

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.351353; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.