Insights into The Codon Usage Bias of 13 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Isolates from Different Geo-locations

Ali Mostafa Anwar [*1], Saif M. Khodary [1]

[1] Department of Genetics, Faculty of Agriculture, Cairo University, Giza, 12613, Egypt

*****Correspondence**:** ali.mo.anwar@std.agr.cu.edu.eg

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of Coronavirus disease 2019 (COVID-19) which is an infectious disease that spread throughout the world and was declared as a pandemic by the World Health Organization (WHO). In the present study, we analyzed genome-wide codon usage patterns in 13 SARS-CoV-2 isolates from different geo-locations (countries) by utilizing different CUB measurements. Nucleotide and di-nucleotide compositions displayed bias toward A/U content in all codon positions and CpU-ended codons preference, respectively. Relative Synonymous Codon Usage (RSCU) analysis revealed 8 common putative preferred codons among all the 13 isolates. Interestingly, all of the latter codons are A/U-ended (U-ended: 7, A-ended: 1). Cluster analysis (based on RSCU values) was performed and showed comparable results to the phylogenetic analysis (based on their whole genome sequences) indicating that the CUB pattern may reflect the evolutionary relationship between the tested isolates. To investigate the force (mutation and/or selection) influencing the pattern of CUB in SARS-CoV-2 coding sequences, we employed the following; (i). Effective number of codons (ENc), (ii). ENc-GC3 plot, (iii). Neutrality plot, and (iv) Codon adaptation index (CAI). According to their results, natural selection and/or other factors (not investigated in this study) may be the dominant force driving SARS-CoV-2 CUB. It is also worth mentioning that, by using the most expressed genes in human lung tissues as a reference set, some viral genes such as Nucleocapsid phosphoprotein, ORF7a protein, and surface glycoprotein had high CAI values which may indicate for selection force acting on their codon usage, as they play important roles in viral assembly and may help viruses avoid the host immune system. The outcome of our study may help in understanding the underlying factors involved in the evolution of SARS-CoV-2 viruses, and the interactions with their host. Also, it may aid in vaccine design strategies.

Introduction

Baltimore classified viruses into 6 classes by the means of their genomes. One class is shared between RNA and DNA viruses, while 3 of them are occupied solely by RNA viruses reflecting their great diversity and different replicative mechanisms [1]. Over the past few decades, many human infectious diseases including, Ebola fever, avian influenza, severe acute respiratory syndrome coronavirus (SARS-CoV) resulted from the interspecies transmission of zoonotic RNA viruses [2–4]. Most recently, the new pandemic (COVID-19) caused by 2019-nCoV (SARS-CoV-2) has emerged in Wuhan, China in December 2019 and spread to 199 other countries, areas or territories with 462,684 confirmed cases of infection and 20,834 confirmed deaths globally up today (https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 26 March 2020, 20:33 EET). Preliminary phylogenetic analysis showed that SARS-CoV-2 most closely related viruses were (bat-SL-CoVZC45) and a SARS-like beta-coronavirus of bat origin (bat-SL-CoVZXC21). Many encoded proteins revealed a high sequence identity except for the spike (S) protein and protein 13 (80% and 73% respectively) between SARS-CoV-2 and other bat-derived coronaviruses [5]. Two probable scenarios are suggested to explain the origin of SARS-CoV-2, natural selection in an animal host either before or following a zoonotic transfer [6].

Coronaviruses (CoVs) are a family of enveloped, single-stranded RNA viruses with the largest genomes (~30 kilobases in length) among other RNA viruses. They are known to cause infection in many avian and mammalian hosts, including humans [7]. They contain four structural proteins: namely, the spike (S) protein, membrane (M) protein, envelope (E) protein and the nucleocapsid (N) protein. The (S) protein has two functions; attachment to the receptors of host cells, and activating the fusion of the virion membrane with host cell membranes [8]. The (M) protein is the most abundant glycoprotein in the virion, unlike the (E) protein which is present in minute amounts yet it is essential for coronavirus morphogenesis and envelope formation [9]. Meanwhile, the (N) protein is present inside the virion complexed with the viral RNA to form the helical nucleocapsid structure [10].

During the translation process from mRNA to protein, information is transmitted in the form of nucleotide triplets named codons. Amino acids are degenerate, having more than one codon representing each except for Methionine (Met) and Tryptophan (Trp). Thus, codons encoding the same amino acid are known as synonymous codons. Many studies on different

organisms showed that synonymous codons are not used uniformly within and between genes of one genome. This phenomenon called synonymous codon usage bias (SCUB) or codon usage bias (CUB) [11–14]. Further, the degree of the unequal use of synonymous codons differs between species [15,16]. Hence, each organism has its optimal codons, where an optimal codon is defined as a codon which is more frequently used in highly expressed genes than in the slightly expressed genes [17]. Two main factors shape the codon usage of an organism mutation and selection [18–20]. Other factors also known to influence the CUB of an organism are nucleotide composition [21], synonymous substitution rate [22], tRNA abundance [23], codon hydropathy and DNA replication initiation sites [24], gene length [25], and expression level [26]. Since viruses rely on the tRNA pool of their hosts in the translation process, previous studies suggest that translational selection and/or directional mutational pressure act on the codon usage of the viral genome to optimize or deoptimize it towards the codon usage of their hosts [27,28]. Hence, it is important to examine viral gene structures and compositions at the codon or nucleotide level to reveal the mechanisms of virus-host relationships and virus evolution [29].

Studying the codon usage in RNA viruses in general help in understanding the evolutionary history of viruses and the evolutionary forces that shape the viral genome, which might assist in understanding the characteristics of newly emerging viruses. Moreover, a study on Influenza A virus (IAV) [30] suggested that understanding codon usage and its nucleotide content in viruses may help in creating new vaccines using Synthetic Attenuated Virus Engineering (SAVE). By deoptimizing viral codons it might be possible to attenuate a virus [31]. Another study reported that the replacement of natural codons with synonymous triplets with increased frequencies of CpG gives rise to inactivation of Poliovirus infectivity [32], the same can be applied to IAV [33].

To our knowledge, no attempts have been made to understand the forces (mutation and/or selection) influencing the overall CUB in SARS-CoV-2 coding regions. This study aims to investigate the codon usage bias and factors affecting them in all coding regions (CDS) of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from 13 different countries. The latter may help in better understanding the molecular evolution regarding the codon usage and nucleotides composition of SARS-CoV-2. Also, the codon usage for SARS-CoV-2 was compared to the host codon usage, to examine the virus in relation to its host co-adaptation (*Homo sapiens*).

Materials and Methods

Sequence data collection

All the CDS for the complete genomes were obtained from the NCBI virus portal (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). In this study, 13 isolates of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) were picked from different countries (USA, Pakistan, Spain, Vietnam, Italy, India, Brazil, China, Sweden, Nepal, Taiwan, South Korea, and Australia) according to the most recent collection date and complete genome length. All information about the used isolates can be found in Supplementary file 1. For this study, isolates were named by their geo-location (country).

Codon usage bias measurement:

Nucleotide composition analysis

GC and AU nucleotide content was estimated for each isolate. As well as, GC and AU content in first, second and third codon positions, were used as a parameter for CUB to address the nucleotide composition effect.

Synonymous Dinucleotide Usage (SDU)

A new index named Synonymous Dinucleotide Usage (SDU) has been developed [34] to implement a way to estimate the degree to which a host-driven force acting on the dinucleotide level of viral genomes has skewed the synonymous codon usage of the protein sequence. To examine the occurrences of a given dinucleotide to the null hypothesis that there is equal usage of synonymous codons. The SDU defines three positions for dinucleotide frames, a frame 1 as the first and second nucleotide codon positions, frame 2 as the second and third nucleotide codon positions, and a bridge frame as the third nucleotide codon position and the first nucleotide from a downstream codon on the same coding sequence. As frame 1 and 2 can change dinucleotides in a way without changing the amino acids of a protein sequence (Synonymous dinucleotides). The SDU values for all SARS-CoV-2 CDS tested isolates in this study were calculated using this equation:

$$SDU_{j,h} = \frac{\sum_{i=1}^{k} n_i \frac{o_{i,j,h}}{e_{i,j,h}}}{N} \quad \textbf{(1)}$$

where n_i is the number of occurrences of amino acid i in the sequence, $o_{i,j,h}$ is the synonymous proportion of dinucleotide j in frame position h for amino acid i observed in the sequence,

$e_{i,j,h}$ is the synonymous proportion of dinucleotide *j* in frame position *h* for amino acid or amino acid pair *i* expected under equal synonymous codon usage, and *N* it the total number of amino acids in the sequence. The result from SDU directly indicates the overall synonymous dinucleotide representation in each frame positions for the tested CDS. Where SDU value 1 means a dinucleotide in a given frame position is equal to the expected under equal synonymous codon usage. SDU more than 1 means the dinucleotide is over-represented in a given frame position compared to the excepted under the null hypothesis. Lastly, SDU within 0 and 1 show under-representation in the provided frame position, compared to the representation assumed under the null hypothesis. Using this index, the dinucleotide frequency in frame 2 (dinucleotide at position 2 and 3 in a codon) were measured for the 13 tested isolates in this study.

Relative Synonymous Codon Usage (RSCU)

Using the following equation [35]:

$$RSCU = \frac{O_{ac}}{\frac{1}{k_a}\sum_{c \in C_a} O_{ac}} \quad \textbf{(2)}$$

Where O_{ac} is the count of codon *c* for the amino acid *a*, and k_a is the number of synonymous codons in amino acid *a* family, RSCU values were calculated. An RSCU value of 1 indicates no codon usage bias as the observed frequency is equal to the expected frequency. RSCU values less than 1 indicate negative bias, and values of greater than 1 indicates positive bias. Accordingly, RSCU values are divided into 3 ranges; values ≤ 0.6 indicates under-represented codons, values between 0.6 and 1.6 indicates randomly used or biased codons, and values ≥ 1.6 indicates over-represented or preferred codons [25,35–37].

Effective number of codons (ENc)

The effective number of codons (ENc) measures the bias of using a smaller subset of codons apart from the equal use of synonymous codons. Also, the ENc measures the codon usage imbalance among genes, where an amino acid is encoded by one codon in a gene would be biased and negatively correlated with the ENc value. The ENc value ranges from 20-61, with higher values indicating more codons being used for each amino acid, i.e less bias and vice

versa. This measures codon bias irrespective of gene length and can be an indicator of codon usage regarding mutational bias [38]. ENc was calculated using the following equations [39]:

$$F_{CF} = \sum_{i=1}^{m} \left(\frac{n_i + 1}{n + m}\right)^2 \quad \textbf{(3)}$$

Then, the ENc value is obtained by:

$$N_{c.CF} = \frac{1}{F_{CF}} \quad \textbf{(4)}$$

Where n_i is the count of codon *i* in *m* amino acid family and *m* is the number of codons in an amino acid family.

Mutational pressure mediated codon usage bias:

ENc-GC3 plot

To verify the relationship between ENc and GC3s, ENc-plot was drawn where the expected ENc values from GC3s (denoted by `S') were determined according to the following equation [38,40]:

$$ENc_{expected} = 2 + S\frac{29}{S^2 + (1 - S)^2} \quad \textbf{(5)}$$

Using the above equation, the expected fitting curve of ENc values was drawn then ENc values versus GC3s values for each coding region are plotted. If the distribution of the plotted genes is along/near the curve, then the codon usage bias is assumed to be affected only by mutation. If the distribution of the plotted genes is below the curve, then the codon usage bias is assumed to be affected by selection and other factors [41].

Natural selection mediated codon usage bias:

Neutrality Plot

Neutrality plot is used to estimate the effect of mutation pressure and natural selection on codon usage bias. In this study, the GC contents at the first, second and third codon positions (GC1, GC2, GC3, respectively) of the 13 SARS-CoV-2 isolates were analyzed. Then, GC12 representing the average GC content at the first and second codon positions of each isolate was obtained. Both GC12 and GC3 values were used for neutrality plot analysis. If the correlation between GC12 and GC3 is statistically significant and the slope of the regression

line is close to 1, mutation pressure is assumed to be the main force structuring codon usage bias. Conversely, natural selection would have higher odds leading to a narrow distribution of GC content and a lack of correlation between GC12 and GC3 [42–44].

Codon Adaptation Index (CAI)

Codon adaptation index (CAI) uses a reference set of highly expressed genes (e.g. ribosomal genes), this measure is an indicator of gene expression levels and natural selection; it ranges from 0 to 1 with higher values indicating stronger bias with respect to the reference set, therefore this method is an indicator of selection for a bias toward translational efficiency [35,45]. Codon adaptation index (CAI) was calculated by the equation given by [45,46]:

$$CAI = exp\frac{1}{L}\sum_{k=1}^{L} lnw_{c(k)} \quad \textbf{(6)}$$

Where L is the count of codons in the gene and $w_{c(k)}$ is the relative adaptiveness value for the *k-th* codon in the gene. Twelve genes with the highest level of expression in the lung tissues for human were collected from the human protein atlas project database (Supplementary file 4) (https://www.proteinatlas.org/humanproteome/tissue/lung). Then, CAI values for the 13 isolates were calculated based on those 12 genes as a reference set.

Phylogenetic Analysis

A dendrogram was constructed using 13 complete genomic sequences of 13 SARS-CoV-2 isolates obtained from different geo-locations (Countries). Evolutionary relationships were inferred by using maximum-likelihood statistical method with general time-reversible (GTR) model implemented in MEGA-X software (v10.1) [47]. The bootstrap method with 1000 replicates was used to test the reliability of the phylogenetic tree. For each isolate, the following data are given: Species and country of origin.

Software and Statistical analysis

Spearman's rank correlation and linear regression analyses tests were performed using R Language [48]. Different R packages as vhcub, SeqinR, ggplot2 and stats [49–52] were used to calculate various CUB indices and to draw the graphs in this study. As well as a python package named CAI [46] was used to estimate the CAI for the tested viral isolates. A cluster analysis (Heat map) was performed using CIMminer (https://discover.nci.nih.gov/cimminer/)

based on the RSCU values obtained from the tested isolates. Multiple sequence alignments for the whole genome of the 13 SARS-CoV-2 was done with MAFFT software (v7.450) and the phylogenetic analysis was performed with MEGA-X software (v10.1).

Results and Discussion

Nucleotide and Di-nucleotide Composition Analyses

The analysis of nucleotide content in the 13 tested SARS-CoV-2 isolates showed AT bias in all codon positions. According to **Fig. 1**: (i) The average GC and AT overall content was 38% to 62% respectively, (ii) For GC1, 2, and 3, their average percentages were 45%, 36% and 32% respectively, (iii) The values of AT1, AT2, and AT3 were 55%, 64%, and 68%, respectively. Synonymous Dinucleotide Usage (SDU) was employed to account for the dinucleotide usage in positions 2 and 3 in codons to reveal under- and over-represented dinucleotides. The average SDU values for each dinucleotide from all the examined isolates reported by bar plot (**Fig. 2**) and the red line signifies the cut off (SDU >1) of a dinucleotide to be over-represented. Seven dinucleotides had SDU values more than 1, in descending order (CpU, UpU, ApA, CpA, GpU, GpA, and ApU), all of these dinucleotides either contain A, U or both, reflecting A/U bias. Interestingly, CpG had an SDU value of less than 1 and ranked as the last dinucleotide out of the total 16 (Supplementary file 2).
A recent study [29] on N genes among 13 different coronaviruses (CoVs) reported that a higher AU% over GC% was observed in all the 13 coronaviruses. In this study, the nucleotide composition analysis showed the same pattern of AU%, in which it is over-represented in all codon positions (AU1, AU2, and AU3), as well as the overall AU% (Fig. 1). To examine the effect of dinucleotides (in second and third codon positions only) bias over the CUB of SARS-CoV-2 CDS, a new index named SDU was used to account for dinucleotide frequencies. According to SDU results, CpU was the most over-represented dinucleotide which may indicate that SARS-CoV-2 prefer A/U-ended codons. The under-representation of UpA and CpG in SARS-CoV-2 CDS could be due to the effect of these dinucleotides on the replication rate, where increasing UpA and CpG levels in RNA viruses can lead to a decrease in replication rate and subsequent viral attenuation, also causing a more powerful immune response while decreasing their abundance has the reverse effect [29,53–55]. In HIV-1, the decrease in CpG was explained due to the host-driven force which selects against viruses rich in CpG dinucleotides and drives the observed under-representation [56]. SARS-CoV-2 exhibited the same pattern of nucleotide and dinucleotide compositions as the

aforementioned RNA viruses, which may be explained by the host-driven force that selects against CpG during the immune response against the virus.

Relative Synonymous Codon Usage (RSCU) analysis

In order to determine to what extent A/U ended codons are preferred, and the patterns of synonymous codon usage, RSCU values of each codon was calculated for all 13 SARS-CoV-2 isolates. Synonymous codons with RSCU values ≥ 1.6 were considered over-represented or preferred codons, RSCU values that fell between 1 and 1.6 were considered randomly used or unbiased, and RSCU values ≤ 0.6 were considered under-represented. Among the 18 amino acids used in the analysis, we found 8 over-represented codons for the following amino acids (Arg, Val, Ser, Ala, Thr, Pro, Leu, and Gly) that are common in all 13 isolates with their corresponding average RSCU values (2.46, 2.05, 1.995, 1.815, 1.724, 1.713, 1.687, 1.651, respectively). The amino acid Arginine (Arg) over biased with AGA codon, Valine (Val) over biased with GUU, Serine over biased with UCU, Alanine (Ala) over biased with GCU, Threonine (Thr) over biased with ACU, Proline (Pro) over biased with CCU, Leucine (Leu) over biased with CUU, Glycine (Gly) over biased with GGU. The cluster (Heat map) analysis revealed this pattern as all of the previously mentioned putative preferred codons clustered together **Fig. 3**. It is quite interesting to note that all 8 over-represented codons are A/U-ended (U-ended: 7, A-ended: 1), and none of them was G/C-ended. Meanwhile, most of the under-represented codons are G/C-ended. Furthermore, 4 out of the 8 over-represented codons (RSCU > 1.6) ended with CpU dinucleotide, which was mentioned before as the most over-representative dinucleotide. Thus, it is evident from nucleotide compositional and RSCU analyses that SARS-CoV-2 genomes exhibited higher codon usage bias towards A/U-ended codons compared to G/C ended ones.

Hierarchical clustering (Heat map) and phylogenetic analyses

To observe similar patterns of codon usage bias among the 13 SARS-CoV-2 isolates we performed a hierarchical cluster analysis based on the RSCU values obtained from all coding regions in their genomes. Codons with similar RSCU values were clustered together, more noticeably the 8 over-represented or preferred codons with RSCU values >1.6 (AGA, GUU, UCU, GCU, ACU, CCU, CUU, and GGU) on the right side of **Fig. 3**. The clustering of the 13 isolates on the top side of **Fig. 3** yielded two major branches (A, and B). The isolate from Pakistan separated itself in branch A from the other 12 isolates in branch B indicating its

unique codon usage pattern. Branch B can be further divided into 3 sub-branches: two of them contained South Korea, and the USA being closer to Pakistan isolate than the remaining ones. Meanwhile, the third sub-branch contained two main clusters that are fairly closer to the USA than South Korea. Isolates from India and Spain are grouped in one cluster. The other cluster contained two sub-clusters in which isolates from Australia, Sweden, Brazil and Italy are grouped in one sub-cluster away from China, Nepal, Taiwan and Vietnam isolate in the other sub-cluster. To gain more insight into the evolutionary relationships between the 13 SARS-CoV-2 isolates, phylogenetic analysis was performed based on their whole genome sequences (Fig. 4). It yielded two main branches: one contained the outgroup (Human beta-coronavirus 2c EMC/2012) separated from other SARS-CoV-2 isolates. The branch containing SARS-CoV-2 isolates can be divided into 3 sub-branches: two of them contained the USA, and Pakistan isolates closer to each other than other SARS-CoV-2 isolates, however, the third sub-branch contained two main clusters that are fairly closer to Pakistan isolate than the USA. One cluster containing isolates from Nepal, Taiwan, China, Vietnam, India, and Spain can be further divided into two sub-clusters, where isolates from India and Spain are grouped in one sub-cluster away from others. The other main cluster containing isolates from South Korea, Brazil, Italy, Sweden, and Australia can be also further divided into two sub-clusters, where isolates from Sweden and Australia are grouped in a sub-cluster away from the others.

Despite the differences between hierarchical clustering and phylogenetic analyses, they exhibited similarities that can be summarized in the following 4 points. (i). Both Pakistan and the USA isolates showed clear relatedness in both dendrograms that separated them away from other SARS-CoV-2 isolates. However, Pakistan isolate is further than the USA from the remaining isolates from codon usage bias perspective indicating its unique codon usage bias. That wasn't the case in the phylogenetic tree as Pakistan showed more relatedness to other isolates than the USA from an evolutionary perspective based on their whole-genome sequence similarities. (ii). Isolates from India and Spain grouped themselves in a sub-cluster in the phylogenetic tree and were also present in the same sub-cluster in RSCU clustering analysis indicating their highly similar patterns of codon usage bias and evolutionary relatedness. (iii). Isolates from Sweden, Australia, Brazil, and Italy fell into the same cluster in both analyses where Sweden and Australia being the closest to each other. (iv). Isolates from Nepal, Taiwan, China, and Vietnam also fell into the same cluster in both analyses. The hierarchical clustering and phylogenetic analyses clearly showed that, despite the

evolutionary relatedness of the 13 SARS-CoV-2 isolates based on their whole-genome sequence similarities, their pattern of codon usage bias may differ slightly.

Investigating the forces influencing the pattern of CUB in SARS-CoV-2 coding sequences

Effective number of codons (ENc) and ENc-GC3 plot

In theory, ENc correlates negatively with CUB. For all isolates, the CDS average ENc value was around 48.54 with SD = 0.5 (Supplementary file 3), except for SARS-CoV-2-Pakistan the average ENc was equal to 46.63. As well as, within the genes for each isolate, ENc values ranged from 39.94 to 55.89 (Supplementary files 3), showing a wide range of ENc (CUB). Using equation **(5)**, an expected curve represent CUB dominated by the mutational force was drawn and the ENc-GC3 values plotted for all CDS for each isolate (**Fig. 5)**. The CDS of all isolates were found under the expected values of the standard curve, also, it is worth to mention that as the GC3 increases the CDS gets closer to the curve. Our results suggest that, mutational pressure is not the key factor in structuring the codon usage bias in all 13 SARS-CoV-2 isolates, but other factors are more likely involved, such as natural selection.
A previous study [57] on 50 different human RNA viruses showed a mean of 50.9 for ENc values, ranged from 38.9 to 58.3. Also, many other studies on different RNA viruses for different hosts show ENc values in the same range, as in H1N1pdm IAV (ENc = 52.5) [30], Equine Infectious Anemia Virus (ENc = 43.61) [58], Bovine Viral Diarrhea Virus (ENc = 50.91) [59], and Classical Swine Fever Virus (ENc = 51.7) [60]. The 13 SARS-CoV-2 tested isolates showed the same pattern of ENc values with mean 48.54 and SD = 0.5, for each isolates the average ENc values ranged from 46.64 to 48.56. Together this indicates a relatively stable and conserved genomic composition in these 13 isolates, with slightly low ENc values (high CUB) compared to RNA viruses reported in other studies [61,62] .

Neutrality analysis

The neutrality plot **(Fig. 6**) revealed a narrow range distribution (26.23% to 39.01%) of GC3 values among the 13 SARS-CoV-2 isolates. 12 isolates showed similar weak positive non-significant correlation between GC12 and GC3 values ($r = 0.15$, $p = 0.71$), meanwhile the isolate from Pakistan was slightly different ($r = 0.084$, $p = 0.8$). The slope of the regression line across all isolates showed values ranging from 0.163 to 0.197 with an average of 0.1835 except one isolate from Pakistan had a relatively lower slope value of 0.0559. Therefore, our

results suggest that, the relative neutrality (mutation pressure) had the minor effect (5.59% and 18.35%) on codon usage bias across all 13 isolates, and the main evolutionary force driving codon usage bias is the relative constraints on GC3 (natural selection) (94.41% and 81.65%).

Codon adaptation index (CAI)

In order to determine the degree of adaptation between the human lung tissues (Reservoir of viral host cells) and SARS-CoV-2 codon usage, the CAI values for each isolate was estimated using the most expressed genes in lung tissues as a reference set. Our results showed that, CAI values (**Table 1**) ranged from 0.53 (Pakistan isolate) to 0.54 (Taiwan isolate). As well as, some genes showed a higher CAI with an average of 0.62 (nucleocapsid phosphoprotein gene) and ORF7a protein gene displayed a CAI value of 0.58. The CAI results for SARS-CoV-2 isolates revealed a moderate host-virus co-adaptation, also the CDS for the SARS-CoV-2 can utilize the human cell resources for its translation process efficiently. Moreover, the highest CAI value was recorded for nucleocapsid phosphoprotein gene (0.62) which plays a fundamental role during viral self-assembly [63]. The second highest gene with CAI 0.58 was ORF7a, which may play a role in viral assembly or budding events unique to SARS-CoV [64]. The CAI results for both genes reflect their importance (viral assembly and budding events) for the SARS-CoV-2 virus to complete its life-cycle efficiently in the host cell. Only small amounts of the Envelope (E) protein is sufficient to trigger the formation of virus-like particles [63], which may explain its lowest CAI value (0.48) among other genes.

Conclusion

This study conclusively demonstrates that genome-wide codon usage bias in SARS-CoV-2 coding sequences are similar to what was found in most studied RNA viruses. However, mutation pressure is not the main force acting to shape the CUB in SARS-CoV-2, in contrast to what was observed in other RNA viruses (e.g. Hepatitis A virus and Influenza A virus). Natural selection and/or other factors (not investigated in this study) may be the dominant force driving SARS-CoV-2 CUB. The nucleotide compositions showed an AU% bias in all codon positions (AU1, AU2, AU3, and overall AU). Additionally, the di-nucleotide composition (in frame 2) showed a bias towards CpU-ended codons. Moreover, the 8 common putative preferred codons (over-representative codons) determined in this study had

either A or U end nucleotides with the absence of G and C ones. Furthermore, some genes such as Nucleocapsid phosphoprotein, ORF7a protein, and Surface glycoprotein had slightly high CAI values which may indicate for selection force acting on their codon usage, as they play important roles in viral assembly and may help viruses avoid the host immune system. Overall, there was no significant differences in the patterns of codon usage bias between the tested 13 SARS-CoV-2 isolates from different geo-locations, except for the isolate from Pakistan that might be slightly different, reflecting small evolutionary changes between them. The findings of the present study may help in understanding the underlying factors involved in the evolution of SARS-CoV-2 viruses and the interactions with their host. Also, it may aid in vaccine design strategies.

Table 1: The table shows the average ENc and CAI values for the 13 SARS-CoV-2 isolates.

SARS-CoV-2 (by Country)	Average ENc	Average CAI
Australia	48.547	0.636
Brazil	48.543	0.638
China	48.534	0.641
India	48.501	0.639
Italy	48.541	0.638
Nepal	48.534	0.636
Pakistan	46.64	0.636
South Korea	48.545	0.641
Spain	48.522	0.641
Sweden	48.547	0.641
Taiwan	48.536	0.641
USA	48.555	0.64
Vietnam	48.535	0.641

Table2: RSCU values for all the 13 isolates in each country, and the RSCU > 1.6 are marked.

Codon	AA	SARS-CoV-2-Australia	SARS-CoV-2-Brazil	SARS-CoV-2-China	SARS-CoV-2-India	SARS-CoV-2-Italy	SARS-CoV-2-Nepal
GCA	A	1.401	1.401	1.401	1.402	1.4	1.401
GCC	A	0.453	0.453	0.453	0.454	0.453	0.453
GCG	A	0.321	0.321	0.321	0.321	0.321	0.321
GCU	**A**	**1.825**	**1.825**	**1.825**	**1.822**	**1.826**	**1.825**
UGC	C	0.702	0.702	0.702	0.702	0.702	0.702
UGU	C	0.898	0.898	0.898	0.898	0.898	0.898
GAC	D	0.71	0.71	0.71	0.71	0.71	0.71
GAU	D	1.29	1.29	1.29	1.29	1.29	1.29
GAA	E	1.147	1.147	1.147	1.147	1.147	1.147
GAG	E	0.653	0.653	0.653	0.653	0.653	0.653
UUC	F	0.78	0.78	0.78	0.78	0.78	0.78
UUU	F	1.22	1.22	1.22	1.22	1.22	1.22
GGA	G	0.832	0.832	0.823	0.823	0.832	0.823
GGC	G	1.061	1.061	1.055	1.055	1.061	1.055
GGG	G	0.062	0.062	0.062	0.062	0.062	0.062
GGU	**G**	**1.645**	**1.645**	**1.66**	**1.66**	**1.645**	**1.66**
CAC	H	0.53	0.53	0.53	0.53	0.53	0.53
CAU	H	1.07	1.07	1.07	1.07	1.07	1.07
AUA	I	0.972	0.972	0.972	0.973	0.972	0.972
AUU	I	1.358	1.358	1.358	1.357	1.358	1.358
AAA	K	1.353	1.353	1.353	1.353	1.353	1.353
AAG	K	0.447	0.447	0.447	0.447	0.447	0.447
CUA	L	0.717	0.717	0.717	0.718	0.717	0.717
CUC	L	0.899	0.899	0.899	0.899	0.899	0.899
CUG	L	0.423	0.423	0.423	0.423	0.423	0.423
CUU	**L**	**1.685**	**1.685**	**1.685**	**1.698**	**1.685**	**1.685**
UUA	L	1.217	1.217	1.217	1.19	1.217	1.217
UUG	L	1.06	1.059	1.06	1.073	1.059	1.06
AAC	N	0.73	0.73	0.729	0.73	0.73	0.727
AAU	N	1.27	1.27	1.271	1.27	1.27	1.273
CCA	P	1.422	1.422	1.413	1.421	1.422	1.422
CCC	P	0.211	0.21	0.213	0.21	0.21	0.21
CCG	P	0.647	0.647	0.648	0.647	0.647	0.647
CCU	**P**	**1.72**	**1.721**	**1.726**	**1.722**	**1.721**	**1.721**
CAA	Q	1.244	1.244	1.244	1.244	1.244	1.244
CAG	Q	0.556	0.556	0.556	0.556	0.556	0.556
AGA	**R**	**2.47**	**2.477**	**2.477**	**2.477**	**2.477**	**2.477**
AGG	R	1.079	1.068	1.068	1.068	1.068	1.068
CGA	R	0.371	0.371	0.371	0.371	0.371	0.371
CGC	R	0.365	0.367	0.365	0.365	0.365	0.365
CGG	R	0.086	0.087	0.087	0.087	0.087	0.087
CGU	R	1.629	1.63	1.632	1.632	1.632	1.632
AGC	S	0.453	0.453	0.451	0.452	0.453	0.453
AGU	S	1.205	1.21	1.206	1.198	1.21	1.21
UCA	S	1.494	1.492	1.504	1.532	1.492	1.492
UCC	S	0.603	0.603	0.601	0.596	0.603	0.603
UCG	S	0.233	0.233	0.232	0.226	0.233	0.233
UCU	**S**	**2.011**	**2.009**	**2.005**	**1.996**	**2.009**	**2.009**
ACA	T	1.547	1.547	1.547	1.547	1.547	1.547
ACC	T	0.278	0.278	0.278	0.278	0.278	0.278
ACG	T	0.467	0.467	0.467	0.467	0.467	0.467
ACU	**T**	**1.708**	**1.708**	**1.708**	**1.708**	**1.708**	**1.708**
GUA	V	1.037	1.037	1.041	1.04	1.037	1.041
GUC	V	0.418	0.418	0.42	0.419	0.418	0.42
GUG	V	0.506	0.506	0.507	0.506	0.506	0.507
GUU	**V**	**2.039**	**2.039**	**2.033**	**2.035**	**2.039**	**2.033**
UAC	Y	1.024	1.023	1.024	1.024	1.024	1.024
UAU	Y	0.976	0.977	0.976	0.976	0.976	0.976

Codon	AA	SARS-CoV-2-Pakistan	SARS-CoV-2-South Korea	SARS-CoV-2-Spain	SARS-CoV-2-Sweden	SARS-CoV-2-Taiwan	SARS-CoV-2-USA	SARS-CoV-2-VietNam
GCA	A	1.257	1.401	1.401	1.401	1.401	1.401	1.401
GCC	A	0.762	0.453	0.453	0.453	0.453	0.453	0.453
GCG	A	0.282	0.321	0.321	0.321	0.321	0.321	0.321
GCU	**A**	**1.7**	**1.825**	**1.825**	**1.825**	**1.825**	**1.825**	**1.825**
UGC	C	0.698	0.702	0.702	0.7	0.702	0.702	0.702
UGU	C	0.969	0.898	0.898	0.9	0.898	0.898	0.898
GAC	D	0.737	0.71	0.702	0.71	0.71	0.711	0.71
GAU	D	1.263	1.29	1.298	1.29	1.29	1.289	1.29
GAA	E	1.246	1.147	1.147	1.147	1.147	1.147	1.147
GAG	E	0.588	0.653	0.653	0.653	0.653	0.653	0.653
UUC	F	0.775	0.78	0.78	0.781	0.78	0.779	0.78
UUU	F	1.225	1.22	1.22	1.219	1.22	1.221	1.22
GGA	G	0.742	0.832	0.801	0.832	0.823	0.822	0.823
GGC	G	0.932	1.061	1.061	1.061	1.055	1.054	1.055
GGG	G	0.059	0.062	0.062	0.062	0.062	0.061	0.062
GGU	**G**	**1.601**	**1.645**	**1.676**	**1.645**	**1.66**	**1.663**	**1.66**
CAC	H	0.574	0.53	0.53	0.53	0.53	0.519	0.53
CAU	H	1.093	1.27	1.07	1.07	1.07	1.081	1.07
AUA	I	0.891	0.972	0.972	0.972	0.972	0.972	0.972
AUU	I	1.457	1.358	1.358	1.358	1.358	1.357	1.358
AAA	K	1.232	1.353	1.353	1.353	1.352	1.353	1.353
AAG	K	0.434	0.447	0.447	0.447	0.448	0.447	0.447
CUA	L	0.695	0.723	0.715	0.717	0.717	0.717	0.717
CUC	L	0.794	0.896	0.897	0.899	0.899	0.899	0.899
CUG	L	0.464	0.428	0.42	0.423	0.423	0.423	0.423
CUU	**L**	**1.736**	**1.655**	**1.692**	**1.685**	**1.685**	**1.674**	**1.685**
UUA	L	1.285	1.218	1.21	1.217	1.217	1.239	1.217
UUG	L	1.027	1.08	1.066	1.059	1.06	1.049	1.06
AAC	N	0.659	0.729	0.73	0.73	0.73	0.73	0.73
AAU	N	1.341	1.271	1.27	1.27	1.27	1.27	1.27
CCA	P	1.314	1.422	1.422	1.422	1.422	1.423	1.422
CCC	P	0.194	0.21	0.21	0.21	0.21	0.21	0.21
CCG	P	0.547	0.647	0.647	0.647	0.647	0.647	0.647
CCU	**P**	**1.611**	**1.721**	**1.721**	**1.721**	**1.721**	**1.721**	**1.721**
CAA	Q	1.31	1.244	1.244	1.244	1.244	1.257	1.244
CAG	Q	0.523	0.556	0.556	0.556	0.556	0.543	0.556
AGA	**R**	**2.276**	**2.477**	**2.477**	**2.474**	**2.477**	**2.477**	**2.478**
AGG	R	0.967	1.068	1.068	1.071	1.068	1.068	1.069
CGA	R	0.336	0.371	0.371	0.371	0.371	0.371	0.371
CGC	R	0.35	0.365	0.365	0.365	0.365	0.365	0.365
CGG	R	0.084	0.087	0.087	0.087	0.087	0.087	0.087
CGU	R	1.485	1.632	1.632	1.632	1.632	1.632	1.63
AGC	S	0.398	0.453	0.455	0.453	0.453	0.453	0.453
AGU	S	1.138	1.211	1.202	1.211	1.21	1.227	1.21
UCA	**S**	**1.883**	1.494	1.52	1.492	1.492	1.442	1.492
UCC	S	0.538	0.602	0.597	0.601	0.603	0.611	0.603
UCG	S	0.203	0.227	0.227	0.234	0.233	0.241	0.233
UCU	**S**	1.841	**2.013**	**1.999**	**2.008**	**2.009**	**2.026**	**2.009**
ACA	T	1.427	1.548	1.547	1.547	1.547	1.548	1.547
ACC	T	0.263	0.278	0.278	0.278	0.278	0.277	0.278
ACG	T	0.401	0.466	0.467	0.468	0.467	0.467	0.467
ACU	**T**	**1.909**	**1.708**	**1.708**	**1.708**	**1.708**	**1.708**	**1.708**
GUA	V	0.937	1.037	1.052	1.037	1.041	1.041	1.041
GUC	V	0.388	0.418	0.418	0.418	0.42	0.42	0.42
GUG	V	0.474	0.506	0.506	0.506	0.507	0.507	0.507
GUU	**V**	**2.201**	**2.039**	**2.024**	**2.039**	**2.033**	**2.033**	**2.033**
UAC	Y	0.924	1.024	1.023	1.024	1.023	1.024	1.024
UAU	Y	1.076	0.976	0.977	0.976	0.977	0.976	0.976

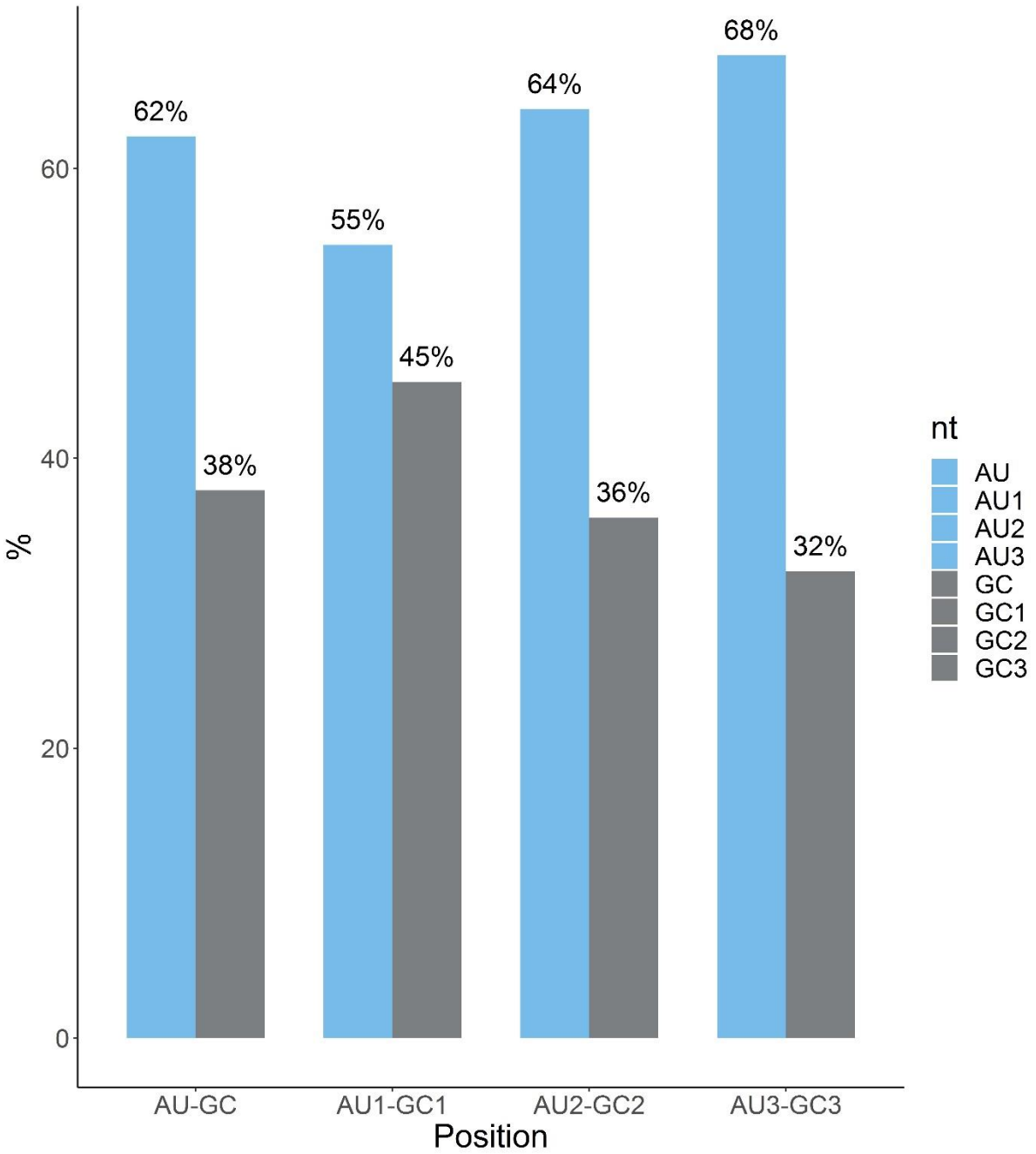


Figure 1. AU% (light blue bars), and GC% (grey bars) nucleotide compositions are plotted together to compare their content in different codon positions, as well as, the overall content. The first two stacked bars represent AT and GC overall content. The rest of the stacked bars represent (AT1 and GC1), (AT2 and GC2) and, (AT3 and GC3) respectively from left to the right.

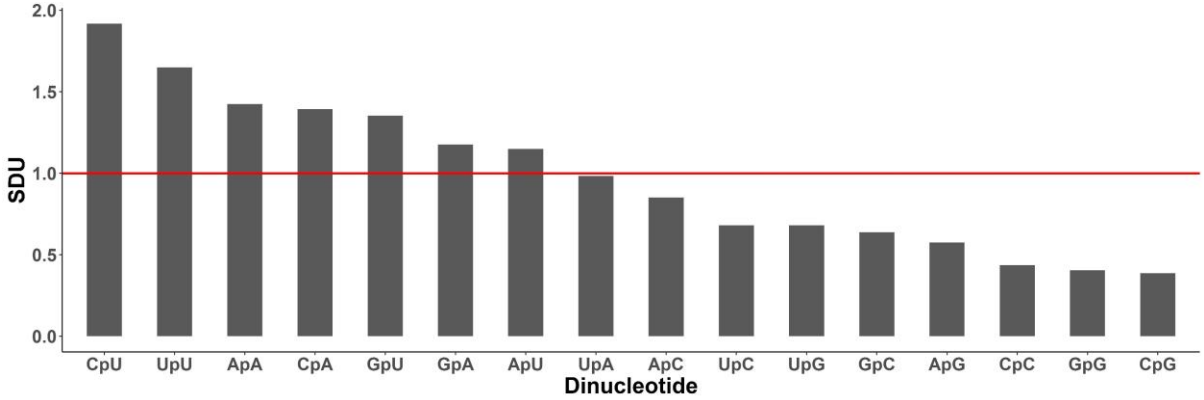


Figure 2. The average synonymous di-nucleotide usage (SDU) values for each di-nucleotide from all examined isolates are estimated and plotted. The x-axis represents each di-nucleotide and the y-axis show the average SDU for each dinucleotide. The red line signifies the cut off (SDU >1) of dinucleotide to be over-represented.

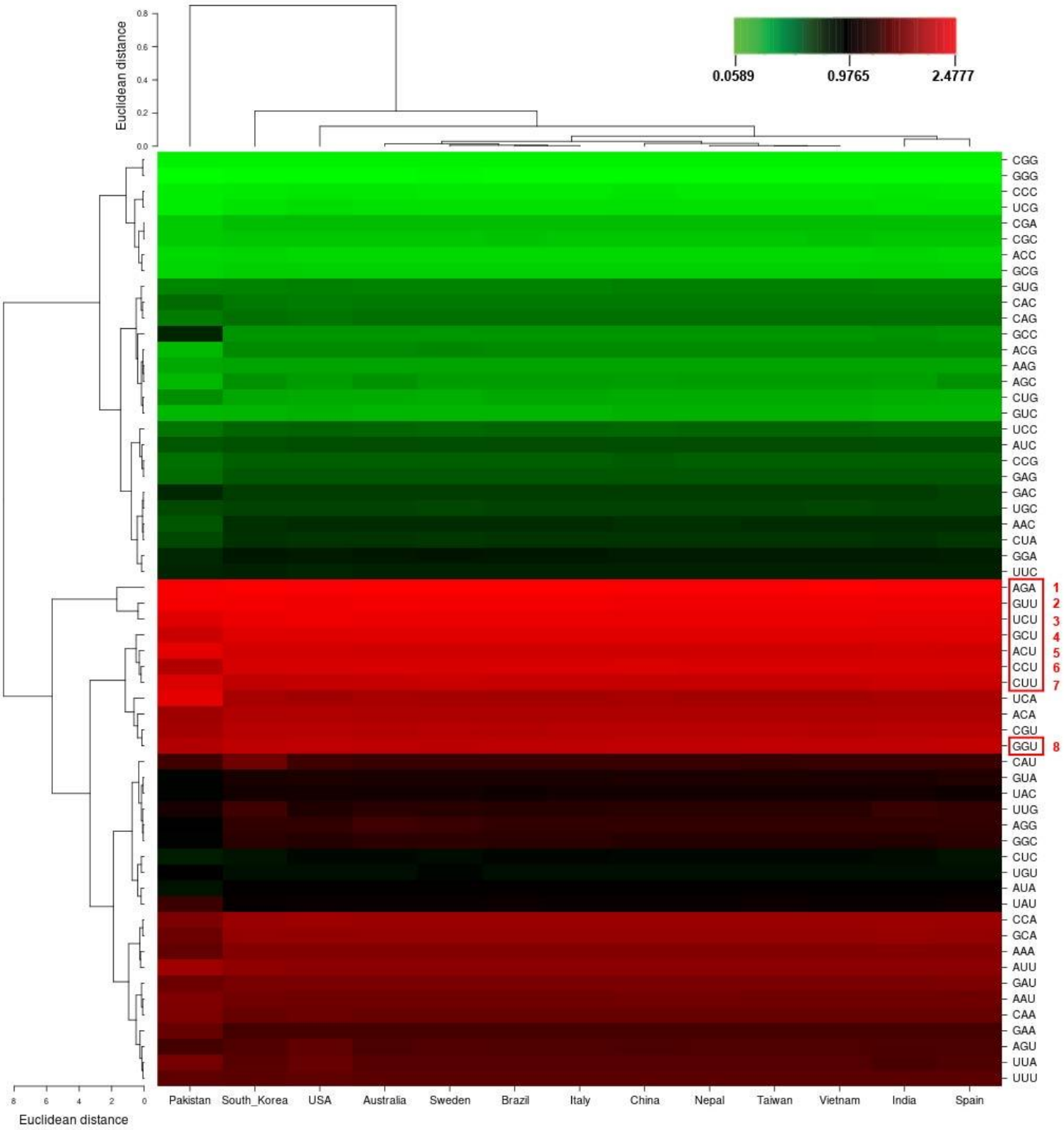


Figure 3. Cluster analysis (Heat map) of RSCU values in 13 SARS-CoV-2 isolates coding regions. The heat map represents the RSCU values divided into 3 ranges; <1 (Green color), 1-1.6 (Dark red), and >1.6 (Distinct red). Euclidean distance and complete-linkage methods were used to produce the clusters.

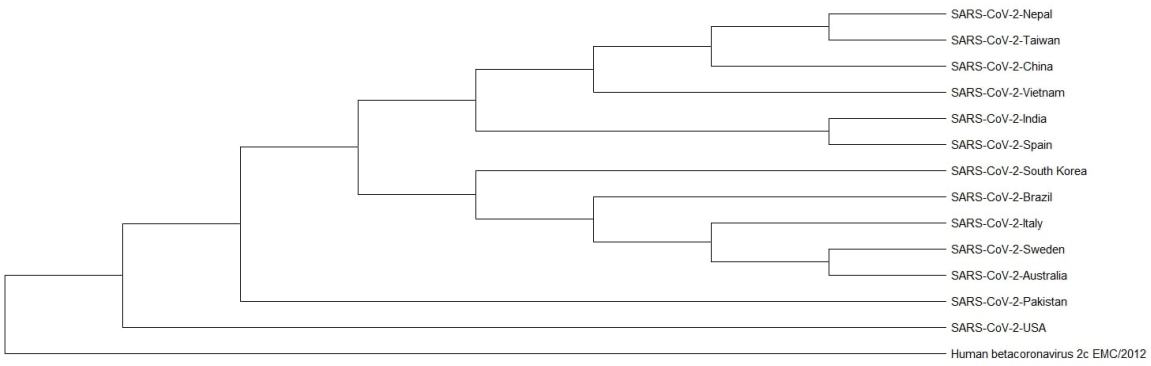


Figure 4. Phylogenetic analysis of 13 SARS-CoV-2 isolates based on their whole genome sequence. The rooted phylogenetic tree was constructed by the maximum-likelihood method with General time-reversible model (GTR), and bootstrap method with 1000 replicates to test the reliability of the phylogenetic tree. The complete genome sequence of (Human beta-coronavirus 2c EMC/2012) was used as an out group.

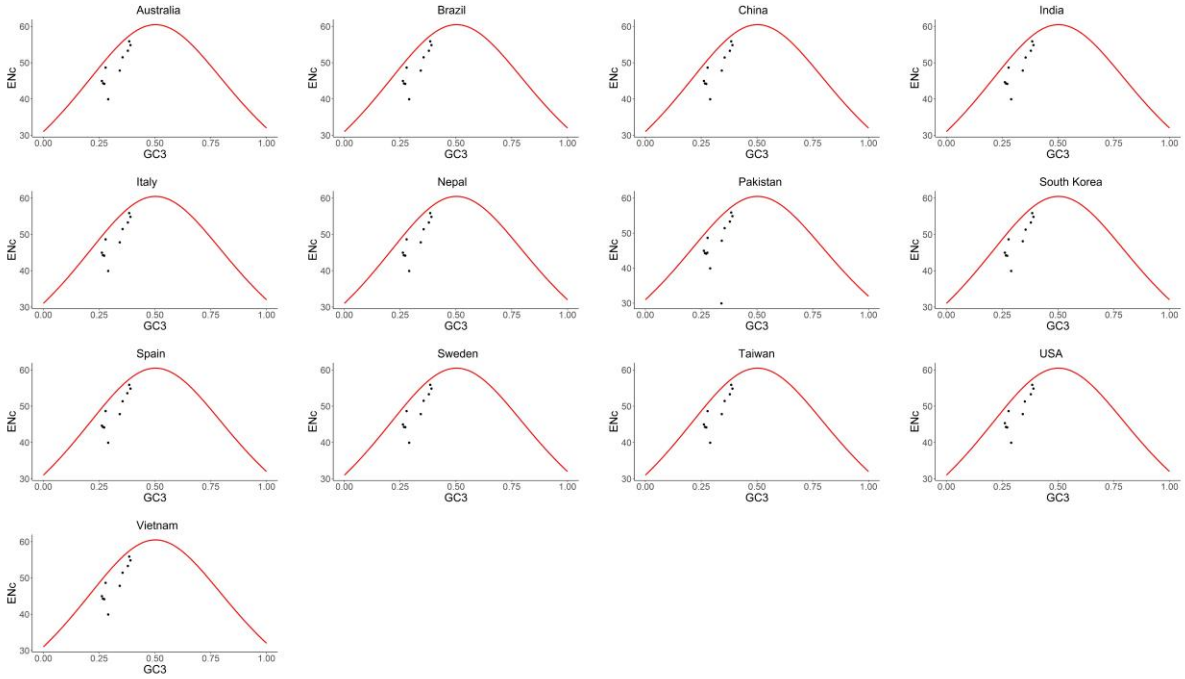


Figure 5. ENc-GC3 plot showing the values of the ENc (y-axis) versus the GC3 content (x-axis) for the 13 SARS-CoV-2 isolates named by their geo-location, the solid red line represents the expected ENc values if the codon bias is affected by GC3s only.

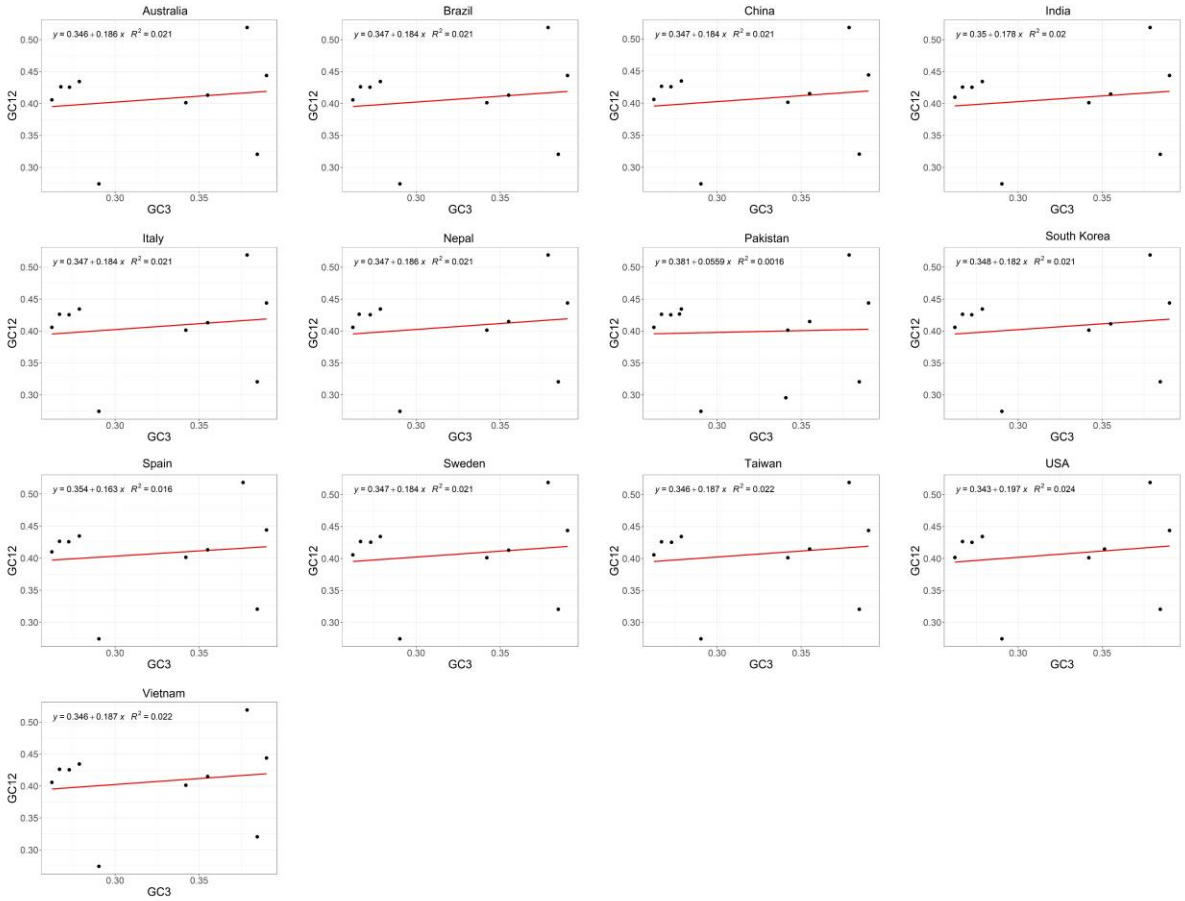


Figure 6. Neutrality plot analysis of 13 SARS-CoV-2 isolates from different countries. GC12 frequencies were plotted against GC3 frequencies. The y-axis (GC12) refers to the average GC frequency at the first and second codon positions. The x-axis (GC3) refers to the GC frequency at the third codon position. The slope value indicates the mutational pressure.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971;35: 235–241. doi:10.1128/mmbr.35.3.235-241.1971

2. Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, et al. Fruit bats as reservoirs of Ebola virus. Nature. 2005;438: 575–576. doi:10.1038/438575a

3. Subbarao K, Klimov A, Katz J, Regnery H, Lim W, Hall H, et al. Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. Science (80-). 1998;279: 393–396. doi:10.1126/science.279.5349.393

4. Epstein JH, McEachern J, Zhang J, Daszak P, Wang H, Field H, et al. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. Science (80-). 2005;310: 676–679. doi:10.1109/NEMS.2006.334722

5. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395: 565–574. doi:10.1016/S0140-6736(20)30251-8

6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;89: 44–48. doi:10.1038/s41591-020-0820-9

7. Weiss SR, Navas-Martin S. Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus. Microbiol Mol Biol Rev. 2005;69: 635–664. doi:10.1128/mmbr.69.4.635-664.2005

8. Cavanagh D. Coronaviridae: a review of coronaviruses and toroviruses. Coronaviruses with Spec Emphas First Insights Concern SARS. 2005; 1–

54. doi:10.1007/3-7643-7339-3_1

9. Narayanan K, Maeda A, Maeda J, Makino S. Characterization of the Coronavirus M Protein and Nucleocapsid Interaction in Infected Cells. J Virol. 2000;74: 8127–8134. doi:10.1128/jvi.74.17.8127-8134.2000

10. Risco C, Antón IM, Enjuanes L, Carrascosa JL. The transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins. J Virol. 1996;70: 4773–4777. doi:10.1128/jvi.70.7.4773-4777.1996

11. Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res. 2004;101: 155–161. doi:10.1016/j.virusres.2004.01.006

12. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evol Biol. 2007;7: 1–17. doi:10.1186/1471-2148-7-226

13. Behura SK, Severson DW. Comparative analysis of Codon usage bias and Codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One. 2012;7. doi:10.1371/journal.pone.0043111

14. Boël G, Letso R, Neely H, Price WN, Su M, Luff J, et al. Codon influence on protein expression in E.coli. 2016;529: 358–363. doi:10.1038/nature16509.Codon

15. Dohra H, Fujishima M, Suzuki H. Analysis of amino acid and codon usage in Paramecium bursaria. FEBS Lett. 2015;589: 3113–3118. doi:10.1016/j.febslet.2015.08.033

16. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species. Genome Biol Evol. 2011;3: 868–880. doi:10.1093/gbe/evr085

17. Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. 2018; 1–17. doi:10.1371/journal.pone.0194372

18. Chen H, Sun S, Norenburg JL, Sundberg P. Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (Nemertea). PLoS One. 2014;9. doi:10.1371/journal.pone.0085631

19. Zalucki YM, Power PM, Jennings MP. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. Nucleic Acids Res. 2007;35: 5748–5754. doi:10.1093/nar/gkm577

20. Prabha R, Singh DP, Sinha S, Ahmad K, Rai A. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. Mar Genomics. 2017;32: 31–39. doi:10.1016/j.margen.2016.10.001

21. Palidwor GA, Perkins TJ, Xia X. A general model of Codon bias due to GC mutational bias. PLoS One. 2010;5. doi:10.1371/journal.pone.0013431

22. Marais G, Mouchiroud D, Duret L. Neutral effect of recombination on base composition in Drosophila. Genet Res. 2003;81: 79–87. doi:10.1017/S0016672302006079

23. Rocha EPC. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 2004; 2279–2286. doi:10.1101/gr.2896904

24. Huang Y, Koonin E V., Lipman DJ, Przytycka TM. Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. Nucleic Acids Res. 2009;37: 6799–6810.

doi:10.1093/nar/gkp712

25. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 1999;96: 4482–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/10200288%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC16358

26. Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes to Cells. 2009;14: 499–509. doi:10.1111/j.1365-2443.2009.01284.x

27. Burns CC, Shaw J, Campagnoli R, Jorba J, Vincent A, Quay J, et al. Modulation of Poliovirus Replicative Fitness in HeLa Cells by Deoptimization of Synonymous Codon Usage in the Capsid Region. J Virol. 2006;80: 3259–3272. doi:10.1128/jvi.80.7.3259-3272.2006

28. Cladel NM, Hu J, Balogh KK, Christensen ND. CRPV genomes with synonymous codon optimizations in the CRPV E7 gene show phenotypic differences in growth and altered immunity upon E7 vaccination. PLoS One. 2008;3: 1–9. doi:10.1371/journal.pone.0002947

29. Sheikh A, Al-taher A, Al-nazawi M, Al-mubarak AI, Kandeel M. Analysis of preferred codon usage in the coronavirus N genes and their implications for genome evolution and vaccine design. J Virol Methods. 2020;277: 113806. doi:10.1016/j.jviromet.2019.113806

30. Goñi N, Iriarte A, Comas V, Soñora M, Moreno P, Moratorio G, et al. Pandemic influenza A virus codon usage revisited: Biases, adaptation and implications for vaccine strain development. Virol J. 2012;9: 1–8. doi:10.1186/1743-422X-9-263

31. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. Science (80-). 2008;320: 1784–1787. doi:10.1126/science.1155761

32. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. J Virol. 2009;83: 9957–9969. doi:10.1128/jvi.00508-09

33. Wimmer E, Paul A V. Synthetic Poliovirus and Other Designer Viruses: What Have We Learned from Them? Annu Rev Microbiol. 2011;65: 583–609. doi:10.1146/annurev-micro-090110-102957

34. Lytras S, Hughes J. Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses. 2020. doi:10.1101/2020.03.02.973438

35. Sharp PM, Li W. Codon Adaptation Index and its potential applications Nucleic Acids Research. 1987;15: 1281–1295.

36. Liu H, He R, Zhang H, Huang Y, Tian M, Zhang J. Analysis of synonymous codon usage in Zea mays. Mol Biol Rep. 2010;37: 677–684. doi:10.1007/s11033-009-9521-7

37. Mandal S De, Mazumder TH, Panda AK, Kumar NS, Jin F. Analysis of synonymous codon usage patterns of HPRT 1 gene across twelve mammalian species. Genomics. 2020;112: 304–311. doi:10.1016/j.ygeno.2019.02.010

38. Wright F. The “effective number of codons” used in a gene. Gene. 1990;87: 23–29.

39. Sun X, Yang Q, Xia X. An improved implementation of effective number

of codons (Nc). Mol Biol Evol. 2013;30: 191–196. doi:10.1093/molbev/mss201

40. Novembre J a. Letter to the Editor Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. Amino Acids. 2000;2: 1390–1394.

41. Liu H, Huang Y, Du X, Chen Z, Zeng X, Chen Y, et al. Patterns of synonymous codon usage bias in the model grass Brachypodium distachyon. Genet Mol Res. 2012;11: 4695–4706. doi:10.4238/2012.October.17.3

42. Song H, Liu J, Song Q, Zhang Q, Tian P, Nan Z. Comprehensive Analysis of Codon Usage Bias in Seven Epichloë Species and Their Peramine-Coding Genes. Front Microbiol. 2017;8: 1–12. doi:10.3389/fmicb.2017.01419

43. Jia J, Xue Q. Codon Usage Biases of Transposable Elements and Host Nuclear Genes in Arabidopsis thaliana and Oryza sativa AT content of TEs and host nuclear. Genomics Proteomics Bioinformatics. 2009;7: 175–184. doi:10.1016/S1672-0229(08)60047-9

44. Wu Y, Zhao D, Tao J. Analysis of Codon Usage Patterns in Herbaceous Peony (Paeonia lactiflora Pall.) Based on Transcriptome Data. 2015;2: 1125–1139. doi:10.3390/genes6041125

45. Ran W, Higgs PG. Contributions of Speed and Accuracy to Translational Selection in Bacteria. PLoS One. 2012;7. doi:10.1371/journal.pone.0051652

46. Lee. Python Implementation of Codon Adaptation Index. J Open Source Softw. 2018;3: 905. doi:10.21105/joss.00905

47. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol. 2011;28: 2731–2739. doi:10.1093/molbev/msr121

48. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016. Available: https://www.r-project.org/

49. Charif D, Lobry JR. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag; 2007. pp. 207–232.

50. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available: http://ggplot2.org

51. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018. Available: https://www.r-project.org/

52. Anwar AM, Soudy M, Mohamed R, Corredor-figueroa AP. vhcub : Virus-host codon usage co-adaptation analysis [version 1 ; peer review : 2 approved]. 2020;2137: 1–10.

53. Witteveldt J, Martin-Gans M, Simmonds P. Enhancement of the Replication of Hepatitis C Virus Replicons of Genotypes 1 to 4 by Manipulation of CpG and UpA Dinucleotide Frequencies and Use of Cell Lines Expressing {SECL}14L2 for Antiviral Resistance Testing. Antimicrob Agents Chemother. 2016;60: 2981–2992. doi:10.1128/aac.02932-15

54. Gaunt E, Wise HM, Zhang H, Lee LN, Atkinson NJ, Nicol MQ, et al. Elevation of CpG frequencies in influenza A genome attenuates

pathogenicity but enhances host response to infection. Elife. 2016;5. doi:10.7554/elife.12735

55. Klitting R, Riziki T, Moureau G, Piorkowski G, Gould EA, de Lamballerie X. Exploratory re-encoding of yellow fever virus genome: new insights for the design of live-attenuated viruses. Virus Evol. 2018;4. doi:10.1093/ve/vey021

56. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. {CG} dinucleotide suppression enables antiviral defence targeting non-self {RNA}. Nature. 2017;550: 124–127. doi:10.1038/nature24039

57. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 2003;92: 1–7. doi:10.1016/s0168-1702(02)00309-x

58. Yin X, Lin Y, Cai W, Wei P, Wang X. Comprehensive analysis of the overall codon usage patterns in equine infectious anemia virus. Virol J. 2013;10: 356. doi:10.1186/1743-422x-10-356

59. Wang M, Zhang J, Zhou J, Chen H, Ma L, Ding Y, et al. Analysis of codon usage in bovine viral diarrhea virus. Arch Virol. 2010;156: 153–160. doi:10.1007/s00705-010-0848-0

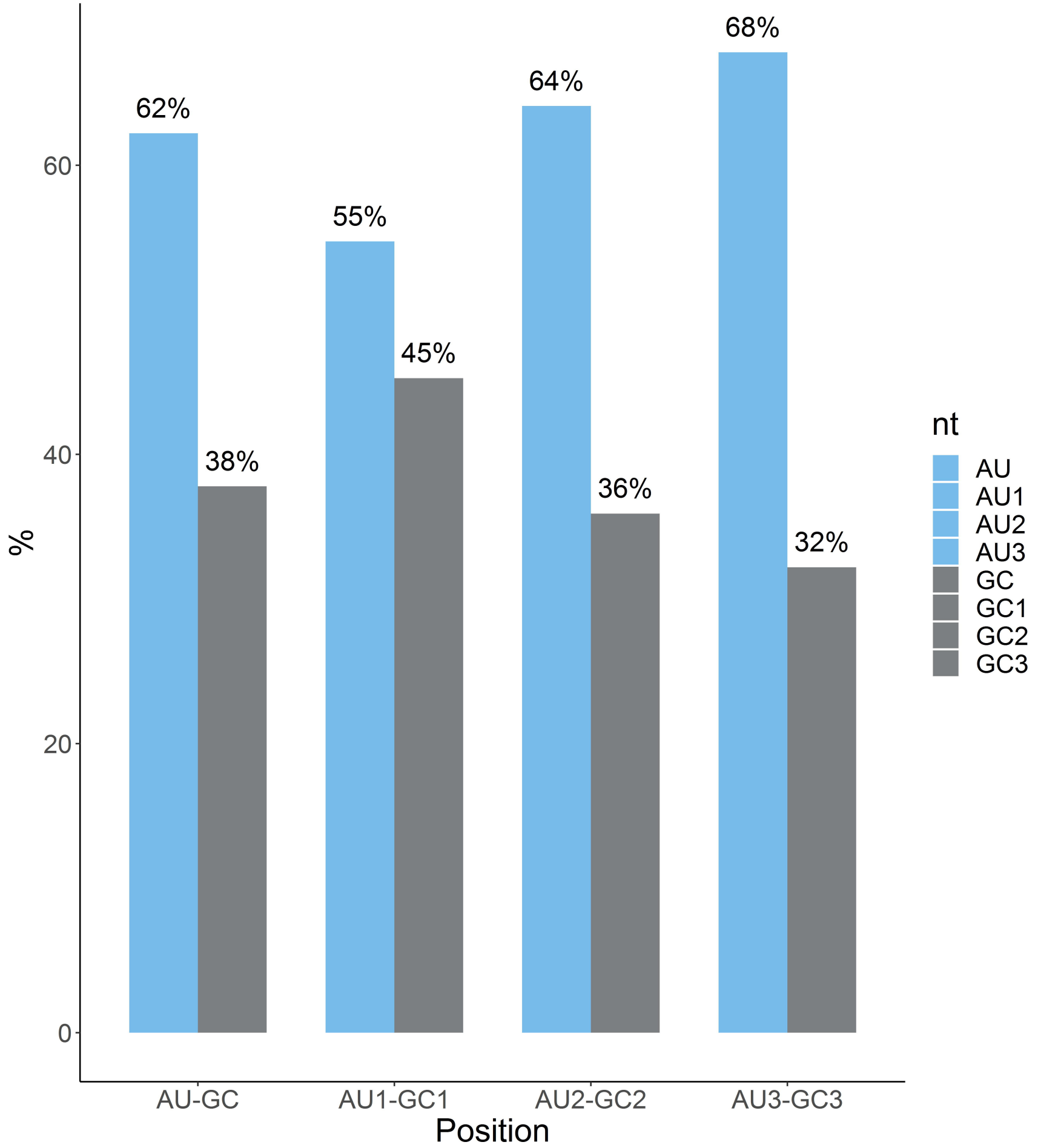
60. Tao P, Dai L, Luo M, Tang F, Tien P, Pan Z. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes. 2008;38: 104–112. doi:10.1007/s11262-008-0296-z

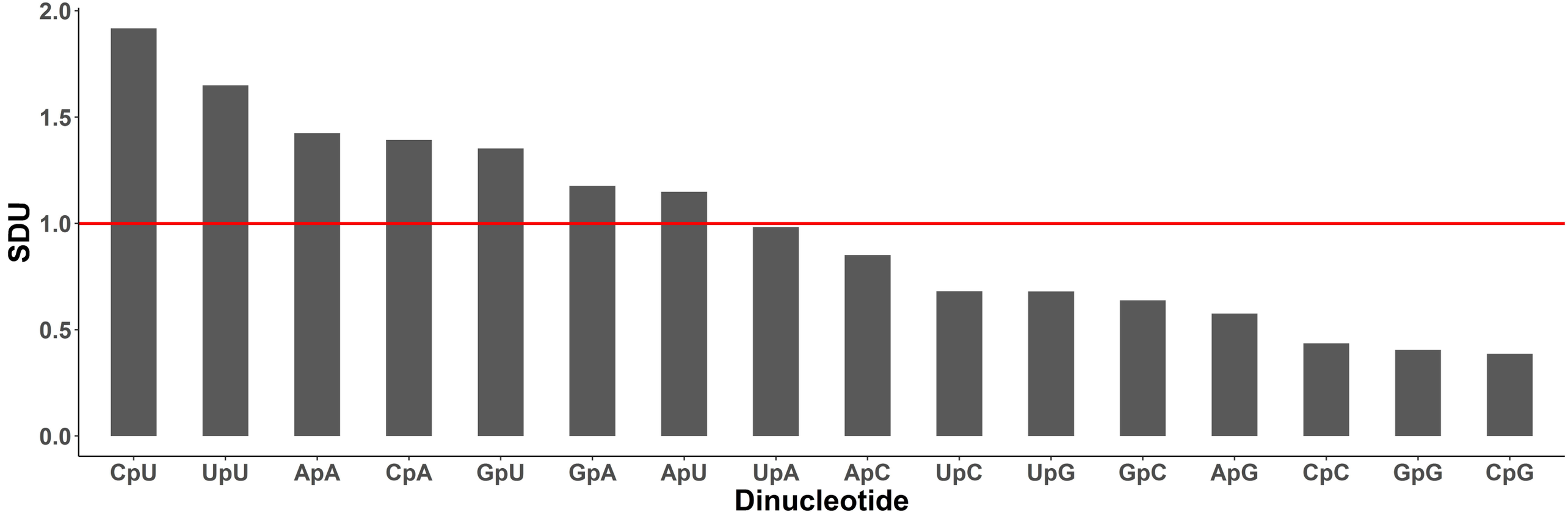
61. Butt AM, Nasrullah I, Tong Y. Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses. Baldanti F, editor. {PLoS} {ONE}. 2014;9: e90905. doi:10.1371/journal.pone.0090905

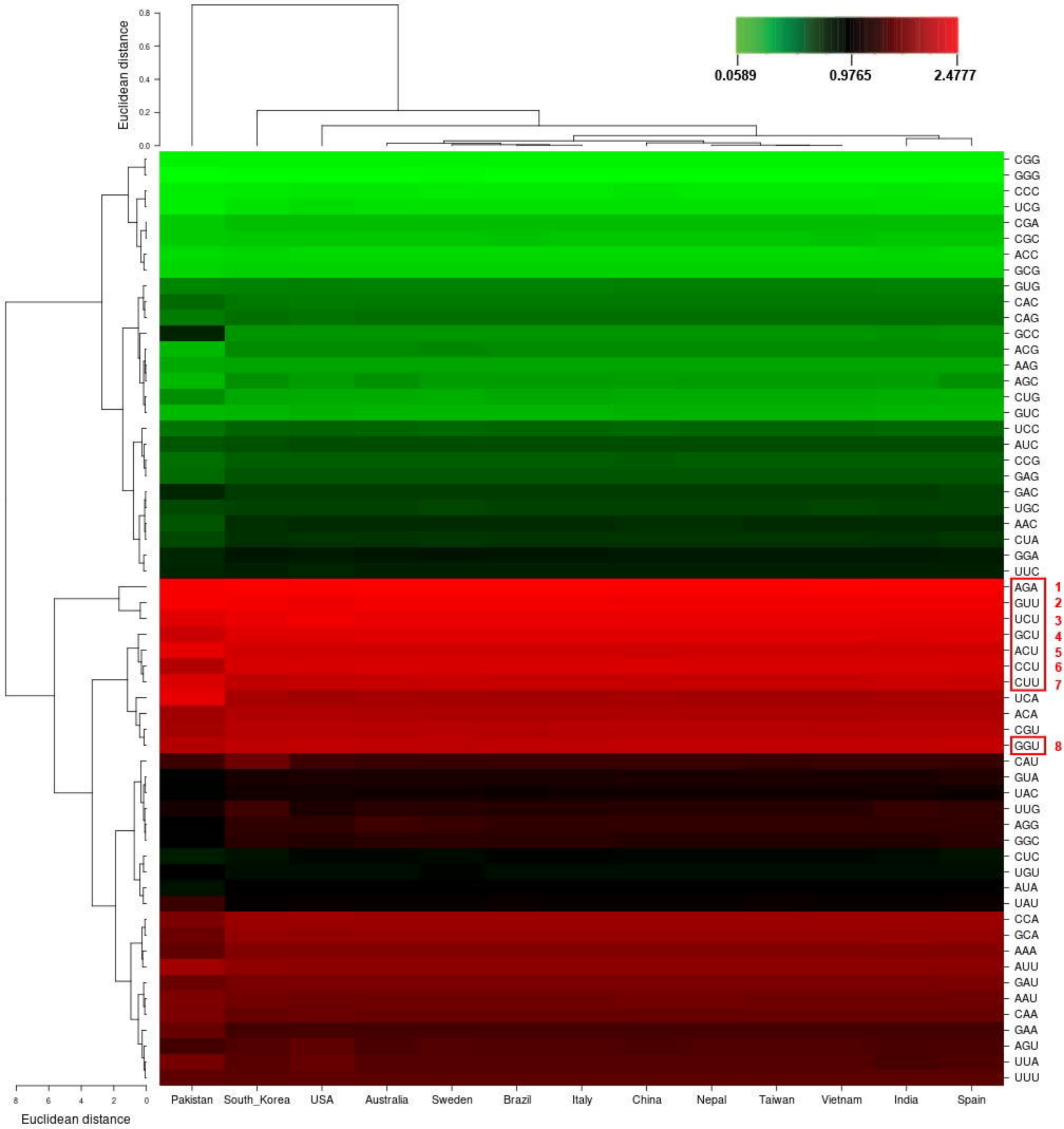
62. Kumar N, Bera BC, Greenbaum BD, Bhatia S, Sood R, Selvaraj P, et al. Revelation of Influencing Factors in Overall Codon Usage Bias of Equine Influenza Viruses. Khudyakov YE, editor. {PLOS} {ONE}. 2016;11: e0154376. doi:10.1371/journal.pone.0154376

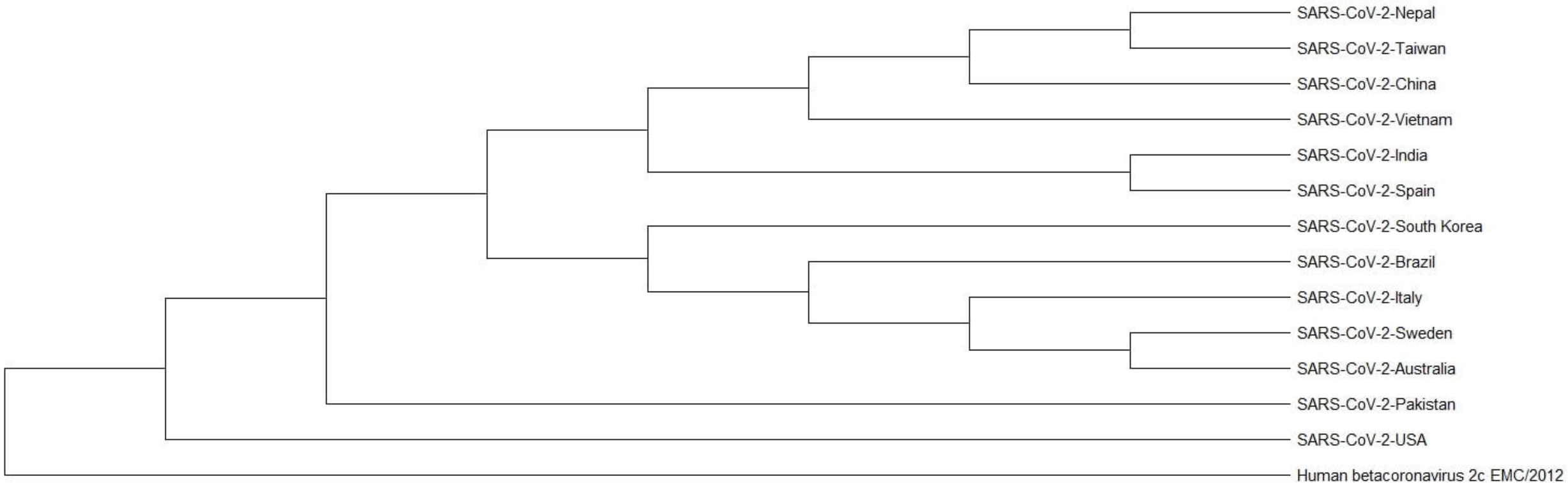
63. Nelson CA, Pekosz A, Lee CA, Diamond MS, Fremont DH. Structure and Intracellular Targeting of the SARS-Corona virus Orf7a Accessory Protein. Structure. 2005;13: 75–85. doi:10.1016/j.str.2004.10.010

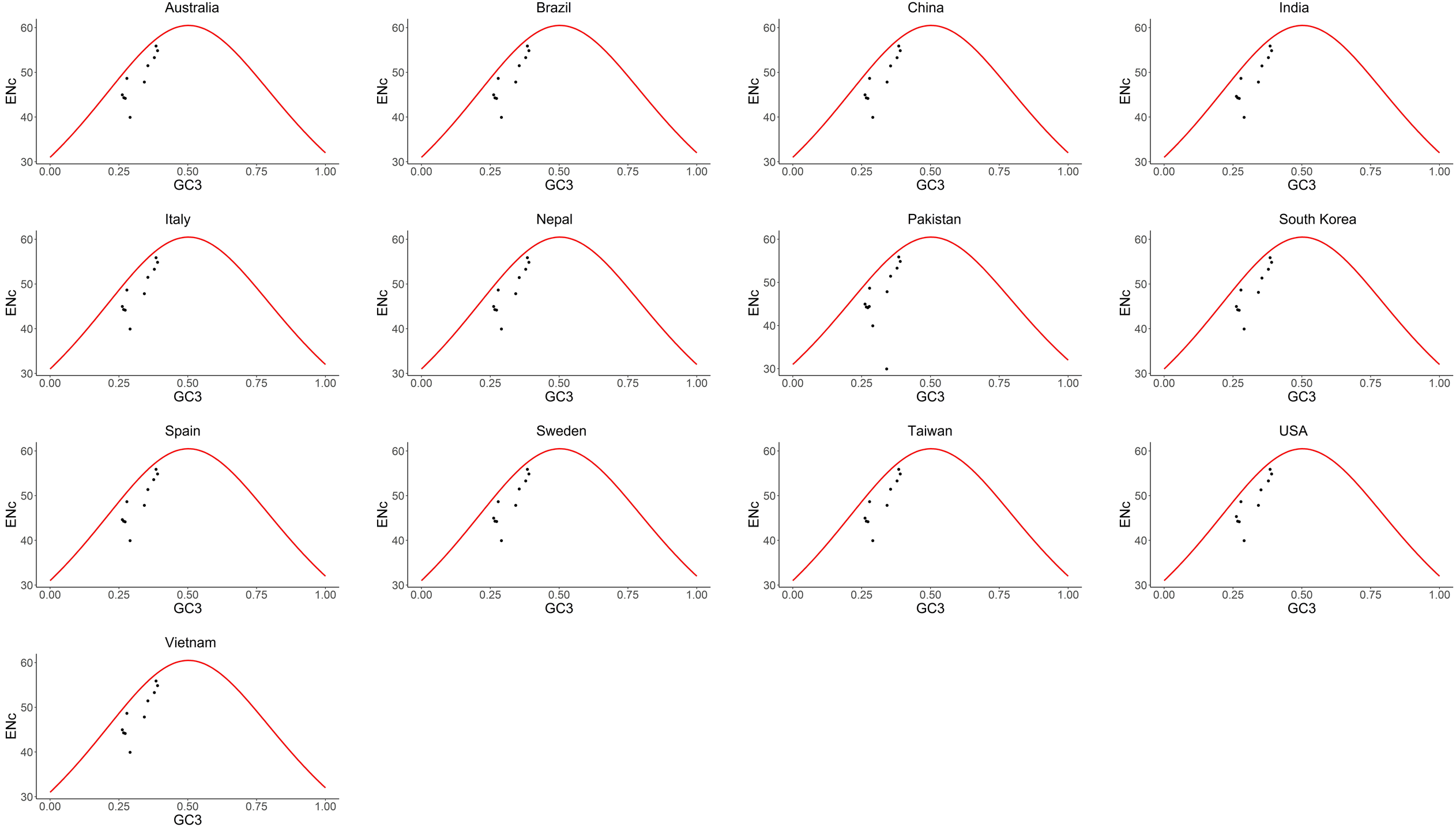
64. Chang C, Hou M-H, Chang C-F, Hsiao C-D, Huang T. The SARS coronavirus nucleocapsid protein--Forms and functions. Antiviral Res. 2014;103: 39–50. doi:10.1016/j.antiviral.2013.12.009

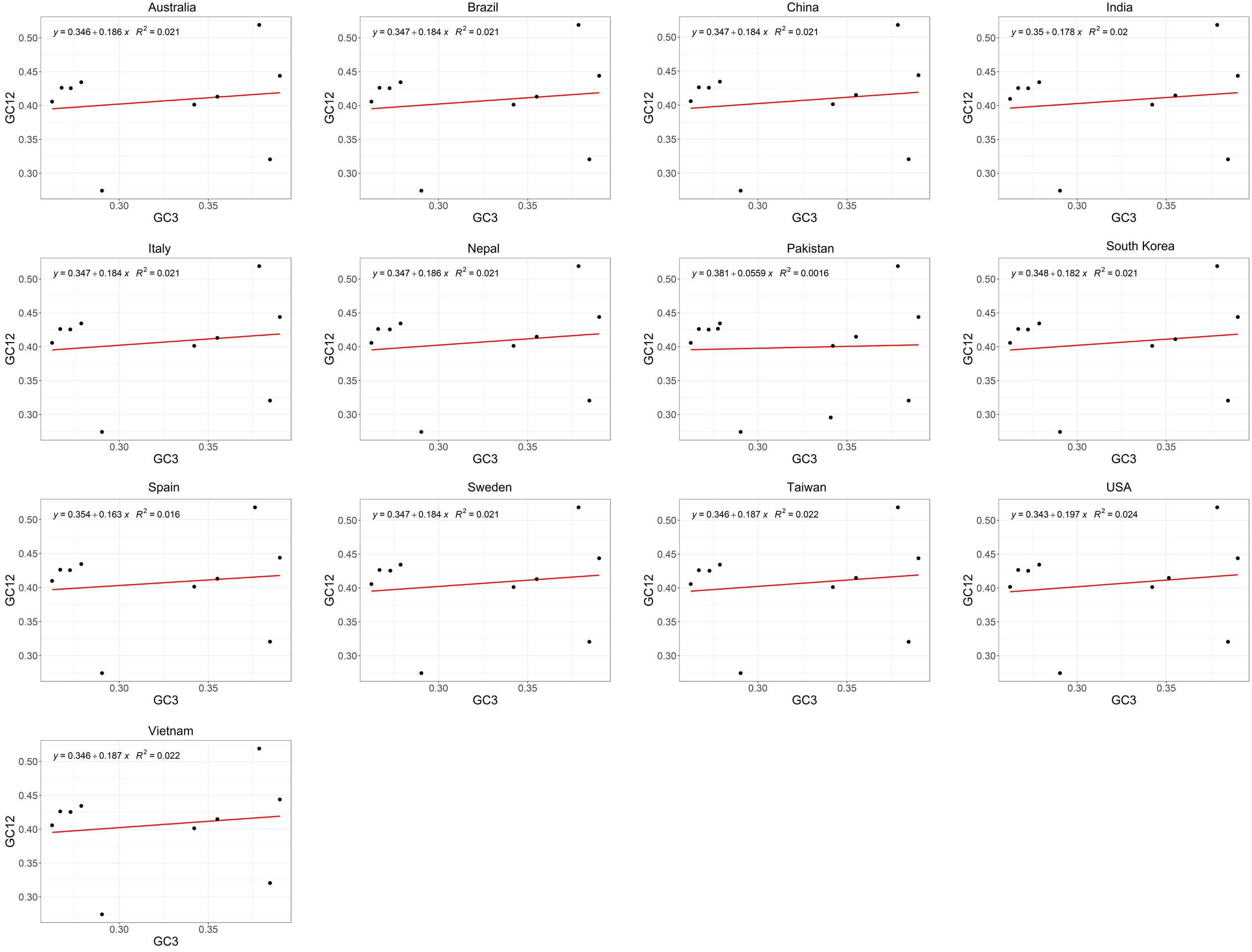

62%
38%
55%
45%
64%
36%
68%
32%
60
40
20
0
%
AU-GC
AU1-GC1
AU2-GC2
AU3-GC3
Position
nt
AU
AU1
AU2
AU3
GC
GC1
GC2
GC3


2.0
1.5
1.0
0.5
0.0
SDU
CpU
UpU
ApA
CpA
GpU
GpA
ApU
UpA
ApC
UpC
UpG
GpC
ApG
CpC
GpG
CpG
Dinucleotide


Euclidean distance
0.8
0.6
0.4
0.2
0.0
0.0589
0.9765
2.4777
CGG
GGG
CCC
UCG
CGA
CGC
ACC
GCG
GUG
CAC
CAG
GCC
ACG
AAG
AGC
CUG
GUC
UCC
AUC
CCG
GAG
GAC
UGC
AAC
CUA
GGA
UUC
AGA
GUU
UCU
GCU
ACU
CCU
CUU
UCA
ACA
CGU
GGU
CAU
GUA
UAC
UUG
AGG
GGC
CUC
UGU
AUA
UAU
CCA
GCA
AAA
AUU
GAU
AAU
CAA
GAA
AGU
UUA
UUU
1
2
3
4
5
6
7
8
8
6
4
2
0
Pakistan
South_Korea
USA
Australia
Sweden
Brazil
Italy
China
Nepal
Taiwan
Vietnam
India
Spain
Euclidean distance


SARS-CoV-2-Nepal
SARS-CoV-2-Taiwan
SARS-CoV-2-China
SARS-CoV-2-Vietnam
SARS-CoV-2-India
SARS-CoV-2-Spain
SARS-CoV-2-South Korea
SARS-CoV-2-Brazil
SARS-CoV-2-Italy
SARS-CoV-2-Sweden
SARS-CoV-2-Australia
SARS-CoV-2-Pakistan
SARS-CoV-2-USA
Human betacoronavirus 2c EMC/2012


Australia
Brazil
China
India
Italy
Nepal
Pakistan
South Korea
Spain
Sweden
Taiwan
USA
Vietnam
ENc
GC3
30
40
50
60
0.00
0.25
0.50
0.75
1.00


Australia
y = 0.346 + 0.186 x R² = 0.021
Brazil
y = 0.347 + 0.184 x R² = 0.021
China
y = 0.347 + 0.184 x R² = 0.021
India
y = 0.35 + 0.178 x R² = 0.02
Italy
y = 0.347 + 0.184 x R² = 0.021
Nepal
y = 0.347 + 0.186 x R² = 0.021
Pakistan
y = 0.381 + 0.0559 x R² = 0.0016
South Korea
y = 0.348 + 0.182 x R² = 0.021
Spain
y = 0.354 + 0.163 x R² = 0.016
Sweden
y = 0.347 + 0.184 x R² = 0.021
Taiwan
y = 0.346 + 0.187 x R² = 0.022
USA
y = 0.343 + 0.197 x R² = 0.024
Vietnam
y = 0.346 + 0.187 x R² = 0.022
GC12
0.50
0.45
0.40
0.35
0.30
GC3
0.30
0.35