Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Comparative analysis of human coronaviruses focusing on nucleotide variability and synonymous codon usage patterns

Jayanta Kumar Das a,*, Swarup Roy b,*

a Department of Pediatrics, Johns Hopkins University School of Medicine, MD, USA
b Network Reconstruction & Analysis (NetRA) Lab, Department of Computer Applications, Sikkim University, Gangtok, India

ARTICLE INFO

Keywords:
Nucleotide
Codon
RSCU
Amino acid
Phylogeny
Coronaviruses

ABSTRACT

The prevailing COVID-19 pandemic has drawn the attention of the scientific community to study the evolutionary origin of Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2). This study is a comprehensive quantitative analysis of the protein-coding sequences of seven human coronaviruses (HCoVs) to decipher the nucleotide sequence variability and codon usage patterns. It is essential to understand the survival ability of the viruses, their adaptation to hosts, and their evolution.

The current analysis revealed a high abundance of the relative dinucleotide (odds ratio), GC and CT pairs in the first and last two codon positions, respectively, as well as a low abundance of the CG pair in the last two positions of the codon, which might be related to the evolution of the viruses. A remarkable level of variability of GC content in the third position of the codon among the seven coronaviruses was observed. Codons with high RSCU values are primarily from the aliphatic and hydroxyl amino acid groups, and codons with low RSCU values belong to the aliphatic, cyclic, positively charged, and sulfur-containing amino acid groups. In order to elucidate the evolutionary processes of the seven coronaviruses, a phylogenetic tree (dendrogram) was constructed based on the RSCU scores of the codons. The severe and mild categories CoVs were positioned in different clades. A comparative phylogenetic study with other coronaviruses depicted that SARS-CoV-2 is close to the CoV isolated from pangolins (Manis javanica, Pangolin-CoV) and cats (Felis catus, SARS(r)-CoV). Further analysis of the effective number of codon (ENC) usage bias showed a relatively higher bias for SARS-CoV and MERS-CoV compared to SARS-CoV-2. The ENC plot against GC3 suggested that the mutational bias might have a role in determining the codon usage variation among candidate viruses.

A codon adaptability study on a few human host parasites (from different kingdoms), including CoVs, showed a diverse adaptability pattern. SARS-CoV-2 and SARS-CoV exhibit relatively lower but similar codon adaptability compared to MERS-CoV.

1. Introduction

Coronavirus (CoV) is a large, enveloped virus (family-Coronaviridae, subfamily-Coronavirinae) with non-segmented, single-stranded and positive-sense RNA genomes [1]. Seven coronaviruses have been known to infect a human host and cause respiratory diseases. The severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle-East respiratory syndrome coronavirus (MERS-CoV) are the two most lethal coronaviruses. SARS-CoV was first reported in China in 2002 [2,3] and caused about 2000 deaths worldwide. MERS-CoV was reported in Saudi Arabia and South Korea in 2012 and 2015, respectively [4,5]. SARS-CoV-2 is the most recently reported novel CoV (2019), which provoked a large-scale COVID-19 epidemic. SARS-CoV-2 was originated from Wuhan, the largest metropolitan area in the Hubei province in China. SARS-CoV-2 is highly infectious due to its high dissemination rate worldwide. According to the World Health Organization (WHO) report,1 more than 600,000 people were deceased as of July 15, 2020, due to COVID-19. All the three coronaviruses are highly pathogenic, resulting in global outbreaks. The other four human coronaviruses (HCoVs), such as OC43 (HCoV-OC43), HKU1 (HCoV-HKU1), 229E (HCoV-229E), and NL63 (HCoV-NL63), are considered to be a mild category due to their low infection and mortality rate.

* Corresponding authors.
E-mail addresses: dasjayantakumar89@gmail.com (J.K. Das), sroy01@cus.ac.in (S. Roy).

https://doi.org/10.1016/j.ygeno.2021.05.008
Received 19 July 2020; Received in revised form 9 May 2021; Accepted 14 May 2021
Available online 19 May 2021
0888-7543/© 2021 Elsevier Inc. This article is made available under the Elsevier license (http://www.elsevier.com/open-access/userlicense/1.0/).
The complete genome length of SARS-CoV, SARS-CoV-2, and MERS-CoV is approximately 27–30 kbp. The genome of SARS-CoV-2 shows a high sequence similarity (79%) with SARS-CoV and relatively low similarity (50%) with MERS-CoV [6]. The several putative coding regions available in SARS-CoV-2, encode essential genes that include nonstructural proteins such as orf1ab, structural proteins namely spike glycoprotein (S), envelope (E), membrane (M), and nucleocapsid (N), and several accessory protein chains [2, 7]. The two-third of the genome is at the 5’-end of the sequence, encoding the nonstructural proteins, and one-third is at the 3’-end, encoding four structural proteins [7]. The CoV proteins exert diverse functional roles, while nonstructural proteins block the host’s innate immune response [10], the four structural proteins show various functionalities. For example, the envelope protein promotes viral assembly and release [11], spike protein composes the spikes on the viral surface and helps in binding with host receptors [11], nucleocapsid protein self-associates through a C-terminal and activates the expression of cyclooxygenase-2 [12,13], and membrane protein promotes the membrane fusion, regulates viral replication, and packs genomic RNA into viral particles [14,15]. On the other hand, accessory proteins that play a significant role during CoV infection contain several overlapping regions that have been explored slightly; these usually play significant roles during coronavirus infection [16]. However, the accessory proteins may not be functional [17]. Several sequence variability features of structural and nonstructural proteins are yet to be investigated thoroughly.

A total of 61 codons genetically code 20 standard amino acids, and the remaining three codons are for the translation of the termination signal [18]. Therefore, a single amino acid can be coded by multiple codons, which are termed as synonymous codons. The number of synonymous codons for different amino acids varies between 1 and 6. The virus genome differs from each other due to frequent mutations that can prevent the PCR from binding to target sequences [19]. Similar to other RNA viruses, SARS-CoV-2 mutates [20] and creates diverse functionality [21]. Mutation plays a key role that triggers a zoonotic virus to jump from animal to human host [22]. Due to several other biological factors, the mechanism of pathogenicity might differ in highly pathogenic strains that diversify the virus target hosts, even in closely related strains [23,24]. Genome-wide codon usage signature can predict evolutionary forces [25]. The inter- and intra-species codon usage patterns may vary significantly in different organisms [26]. Therefore, it is genetically important to study the nucleotide base composition in all three positions of codon as it could influence the codon usage and mutational bias [27–29]. The frequency of the dinucleotide features is also critical as it might affect the usage of codons [27,30]. Thus, GC (G + C) content may be a good indicator towards understanding the expression of viral genes while interacting with the host proteins [31,32].

Previous studies have demonstrated that the usage of a synonymous codon is a non-random procedure [33,34]. The relative synonymous codon usage (RSCU) standardizes the codon usage of the amino acids encoded by multiple codons. The RSCU value is independent of the amino acid composition and has been used widely to estimate the codon usage bias. Several studies have been performed on CoVs, primarily focusing on the independent genome [27,35] and different strains within the same genome [36]. The recent studies on SARS-CoV-2 have focused on elucidating the proximal origin of SARS-CoV-2 and the host-specific adaption mechanism [37–39]. A study on the codon usage pattern provided an insight into the evolution of viruses and their adaption to hosts [40]. However, several crucial roles are yet to be unveiled from all CoVs, including SARS-CoV-2, to fight against COVID-19 related diseases. Therefore, a genomic level comparative study could help in understanding the molecular and structural resemblance among all HCoVs.

The present study emphasized a comprehensive and integrated study on seven HCoVs. Several bias indexing measures, such as nucleotide composition, dinucleotide odds ratio [41], relative synonymous codon usage pattern [42], effective number of codon usage [43,44], and codon adaptation [45,46] are used to quantify the variability of the candidate HCoVs.

2. Material and methods

2.1. Data retrieval and filtering

The seven candidate species of HCoVs, used in this study, are known to infect the human host. For this research, the nucleotide sequences each of length ≈28kb are collected for candidate viruses from NCBI database2 during April 2020 (Supplementary material 1). All the partial, incomplete, and duplicate genome sequences were removed. Then, for each complete sequence, the single coding sequences are obtained by concatenating the coding regions of all the genes, Orf1ab, S, E, M, N, Orf3a, Orf3b, Orf6, Orf7a, Orf7b, Orf8(a/b), and Orf10 (length 150 bp). The sequence general information for our study are summarized in Table 1.

| Human coronaviruses | Number of unique sequences | Number of protein coding genes |
|---------------------|---------------------------|-------------------------------|
| SARS-CoV            | 134                       | 1766                          |
| SARS-CoV-2          | 401                       | 4525                          |
| MERS-CoV            | 233                       | 2509                          |
| HCoV-OC43           | 157                       | 1257                          |
| HCoV-HKHU1          | 34                        | 277                           |
| HCoV-229E           | 29                        | 225                           |
| HCoV-NL63           | 56                        | 389                           |

2.2. Quantitative measuring indices of nucleotide composition

We reported some of the popular quantitative measuring indices to quantify the composition variability (or similarity) among seven CoVs.

2.2.1. Quantifying nucleotide composition

The quantity of nucleotide base composition (A, T, C, and G) in three different codon positions can be calculated based on the frequency. The base composition and GC-content at the first (GC1), second (GC2), and third positions (GC3) of synonymously variable sense codons vary from 0 to 1. The nucleotide composition of any nucleotide X at position P can be calculated as follows:

\[
N_X(P) = \frac{f_X}{f_A + f_T + f_C + f_G}
\]

where, \( X \in \{A, T, C, G\} \), \( p \in 1, 2, 3 \) and \( f_A, f_T, f_C, f_G \) are the nucleotide frequencies at particular position P for A, T, C and G respectively.

Similarly, we calculated the pair nucleotide composition in a particular position P as follows:

\[
N_{XY}(P) = \frac{f_X + f_Y}{f_A + f_T + f_C + f_G}
\]

where, \( X, Y \in \{A, T, C, G\} \).

2.2.2. Relative dinucleotide abundance

Dinucleotide (DN) composition and variability are essential as they represent the possible bonding and abundances of two consecutive nucleotides over the sequences. In RNA viruses, the relative abundance of dinucleotide has been shown to affect codon usage [28]. The dinucleotide frequency is often used to determine the favorable or unfavorable nucleotide pairs. The patterns of dinucleotide frequency indicate both

---

2 https://www.ncbi.nlm.nih.gov/
selection and mutational pressures [27, 47]. The total possible dinucleotide combinations are 16. The relative dinucleotide abundance frequency can be calculated as follows:

$$D_{xy} = \frac{f_{xy}}{f_x f_y},$$

(3)

where, $f_x$ and $f_y$ represent the individual frequency of nucleotides $x$ and $y$ respectively, and $f_{xy}$ is the frequency of dinucleotide $xy$ in the same sequence. The ratio of observed to expected dinucleotide frequency is known as the odds ratio. The odds ratio $\leq 0.78$ indicated that dinucleotide is underrepresented, whereas a value of $\geq 1.25$ indicates over-representation [41].

2.2.3. Relative synonymous codon usage (RSCU)-pattern

RSCU is the ratio between the observed number of codons and the expected uniform synonymous codon usage [42] (Eq. (4)). The RSCU is used to standardize the codon usage of these amino acids encoded by multiple codons. The RSCU value is independent of the amino acid composition and has been used widely to estimate the codon usage bias. The RSCU value $\geq 1.0$ is considered a positive codon usage bias, and the value $\leq 1.0$ is considered a negative codon usage bias. Thus, a high RSCU value for a codon indicates frequent usage of that codon.

$$RSCU_{ij} = \frac{X_{ij}}{\sum_{i=1}^{n} X_{ij}},$$

(4)

where, $X_{ij}$ is the number of occurrences of the $j$th codon for the $i$th amino acid, which is encoded by $n_i$ synonymous codons.

2.2.4. Effective number of codon (ENC) usage

The ENC (or $N_c$) usage can be obtained by Eq. (5) [43, 44].

$$N_c = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6},$$

(5)

where $F_i$ denotes the average homozygosity for the class with $i$ synonymous codons. The ENC value ranges from 20 to 61. An ENC of 20 represents extreme bias as only one codon is used for each amino acid, and a value of 61 suggests no bias. In contrast to the RSCU value, a high ENC value correlates to a weak codon usage bias. An alternate approach for calculating the ENC is based on the GC3-content, shown in Eq. (6).

$$N_c = 2 + s + \left\{ \frac{29}{s^2 + (1 - s)^2} \right\}$$

(6)

2.2.5. Neutrality plot

A neutrality plot is an analytical method for assessing codon usage to account for mutation-selection equilibrium. A dot in the plot represents each independent sequence. In this plot, the mean GC-content at the third codon position (X-axis), represented by GC3, is compared to the mean GC-content at the first and second codon positions, represented by GC12 (Y-axis). The slope of a regression line represents the effect of mutation pressure on the biased usage of codons. A regression line close to 1 implies mutation bias as a central force for influencing the codon usage [48].

Typically, the correlation, $r$ (or $R$) indicates the strength of the linear association between two variables (x and y). In current study, we considered x and y as the GC-content in different codon positions (GC3, GC12, GC1, and GC2). The $R^2$ value indicated the amount of variability in y explained by the predictor (regressor) x. The $R^2$ value always ranges between 0 and 1.
2.2.6. Codon adaptation index (CAI)

The CAI is a measure of the synonymous codon usage bias for a DNA or RNA sequence. It quantifies the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set [45,46]. The CAI value ranges from 0 to 1; 1 indicates if a gene always uses the most frequently used synonymous codons in the codon usage of a gene and the synonymous codon frequency of a reference set. Therefore, it can be used to understand the codon adaptability between the host and parasite [49,50].

3. Results and discussion

This section highlights the comparative distribution of nucleotide and dinucleotide features in all the HCoVs. Then, codon usage bias (RSCU) was estimated among candidate gene sequences. The RSCU features are used for establishing a phylogenetic relation among the candidate species. Finally, the correlation score was calculated between GC3 and GC content at various codon positions.

3.1. Variability of nucleotide composition

The variability of nucleotide composition is shown using a box-plot in Fig. 1. Notably, the virus nucleotide composition differs significantly with highly diverse base composition of the MERS-CoV genome. Next, we focused on the high and low abundance of four nucleotide bases in seven candidate strains. In addition, a high abundance of base A in SARS-CoV-2 and low in MERS-CoV and HCoV-NL63 was observed. Base T was highly available in HCoV-HKU1 and HCoV-NL63, whereas low in SARS-CoV, SARS-CoV-2 and MERS-CoV contained a high content of base C, while a low availability of base C was detected in HCoV-HKU1. Similarly, the G content was high in HCoV-OC43 and HCoV-229E and low in HCoV-HKU1. General trends of CDSs of all CoV strains were found to be rich in A and T (58–67%) compared to G and C nucleotides. This characteristic was similar to that of the Nipah virus [27], although, we observed a higher abundance of T than A in their genome.

3.2. Dinucleotide (DN) variability and usage pattern

We focused on both the first two codon positions (CP12) and the last two codon positions (CP23) to quantify the consecutive nucleotide pairs (dinucleotide). The observed frequency distribution of nucleotide pair is shown in Fig. 2. We observed a high abundance of nucleotide pairs (AA, GA, GT, and TT) in CP12, when compared to high abundance nucleotide pairs (AT, CT, GT, and TT) in CP23. It is worth mentioning here that dinucleotides, GA and CT, belong to purine and pyrimidine groups, respectively. We also observed less variability in the nucleotide pairs in the CP23 position than CP12 (Fig. 2) position across the seven viruses. Overall, this indicates a distinct pattern of nucleotide pairs in two consecutive codon positions.

Furthermore, we reported the relative abundance of dinucleotide content (odds ratio) for the two highest and lowest usage nucleotide pairs in Table 2. Quantitatively, we observed a similar pattern for the highest and lowest usage nucleotide pairs (GC and CG, respectively) in CP12 position across the seven strains. A similar observation was reported for MERS-CoV [36]. However, we found different structures for the second-highest and lowest usage of nucleotide pairs. We also observed a distinct arrangement of high usage (CT) and low usage (CG) nucleotide pairs in CP23 in the majority of viruses and is also different from the CP12 position as stated above. These two high-usagedinucleotides, GC (in CP12) and CT (in CP23), belong to strong and pyrimidine groups.

3.3. GC content usage pattern

GC content at three different codon positions was calculated for each sequence. We obtained mean (or average) GC content and standard deviation for each strain of CoV as shown in Table 3. It was calculated using Eq. (2). Interestingly, we observed that GC3 shows a high variability among seven viruses, although the GC content in GC1 and GC2 was much greater than that of GC3. Pairwise, the mean GC content for SARS-CoV, MERS-CoV, SARS-CoV-2, and HCoV-229E appeared to be identical. The position-wise GC-content distribution was examined

![Fig. 2. Distribution of mean (average) dinucleotide (DN) content. The mean DN is calculated for both the first two codon positions (CP12) and the last two codon positions (CP23) for seven HCoVs.](image)
We observed that GC content in the three codon positions is well-balanced for HCoV-OC43 and SARS-CoV-2 (mean difference \( \approx 0.9 \)), whereas HCoV-HKU1 and HCoV-NL63 showed similar GC-content between the first and second positions of the codon (mean difference < 0.9), and SARS-CoV, MERS-CoV, and HCoV-229E showed a similar GC content trend in the second and third positions of the codon (mean difference < 0.9).

### 3.4. Synonymous codon usage pattern and phylogenetic clustering

We calculated the RSCU values of 59-non trivial codons (Eq. (4)). The RSCU value for each amino acid and synonymous codons are shown in Table 4, and the distribution pattern is shown in Fig. 4. Next, a high RSCU score (> 1.5) and a low RSCU score (< 0.5) were obtained for all the seven CoVs. Overall, the comparison of SARS-CoV-2’s RSCU values to those of the SARS-CoV and MERS-CoV revealed a similar pattern for most of the codons [36,51].

**High RSCU score codons:** The number of codons with high RSCU value in the seven viruses were as follows: SARS-CoV (13), SARS-CoV-2 (14), MERS-CoV (9), HCoV-OC43 (18), HCoV-HKU1 (18), HCoV-NL63 (11), and HCoV-OC43 (18). Interestingly, a maximum of 18 codons was detected in the mild category of CoVs (HCoV-OC43, HCoV-HKU1, and HCoV-NL63) with high RSCU values, while a minimum of 9 codons was detected in one severe category, MERS-CoV with high RSCU values. In the severe category coronaviruses, high RSCU value codons were mapped to 10 amino acids (A, C, G, I, L, P, R, S, T, and V). However, in the case of mild category CoVs, codons were mapped to the same set of amino acids as in severe category, with an additional five amino acids (D, F, H N, and Y). Together, we observed only 7 common codons among the seven HCoVs: ATT, ACT, TCT, CCT, GTT, GCT, and GGT. Among these 7 codons, 4 were from the aliphatic amino acid group (I-ATT, V-GTT, A-GCT, G-GGT), 2 were from the sulfur-containing amino acid group (S-TCT and T-ACT), and 1 is from the cyclic amino acid group (IP-CCT) (Table 4) according to the 8 chemical groups of amino acid categorization [52,53]. On the other hand, significant differences were obtain in the frequencies of two glutamine codons (GAG and GAA) among SARS-CoV-2 and SARS and MERS-CoVs (Table 4).

**Low RSCU score codons:** The number of codons with low RSCU values in 7 viruses were as follows: SARS-CoV (9), SARS-CoV-2 (13), MERS-CoV (10), HCoV-OC43 (19), HCoV-HKU1 (27), HCoV-229E (19), and HCoV-NL63 (18).
and HCoV-NL63 (28). Thus, it can be stated that high-score and low-score codons are detected in mild category CoVs (except HCoV-229E). We also observed 7 common codons, GCG, GGG, CCG, CGA, CGG, TCG, and ACG, among the 7 CoVs with low RSCU scores. Of these 7 codons, 2 were from the aliphatic group (A-GCG and G-GGG), 1 from the cyclic group (P-CCG), 2 from the basic group (positively-charged) (R-CGA and R-CGG), and 2 from the sulfur-containing groups (S-TCG and T-ACG) (Table 4). Overall, significant differences were observed in the frequencies of 2 glutamine codons (GAG and GAA) in the CoVs in the severe category. In the case of GAG, this codon was lowly expressed in SARS-CoV-2 (0.55) compared to SARS-CoV and MERS-CoVs (0.95 and 0.94, respectively), whereas for GAA, this codon is highly expressed in SARS-CoV-2 (1.44) compared to lowly expressed in SARS-CoV and MERS-CoVs (1.04 and 1.05, respectively).

To understand the evolutionary structure of the CoVs, we obtained an average RSCU value for each codon within the multiple sequences of the same virus (or strain). Next, we created a virus-based 59-dimensional codon feature vector, i.e., RSCU scores of 59 codons. Then, a phylogenetic tree (dendrogram) was constructed using the unweighted pair group method with arithmetic mean (UPGMA), a hierarchical clustering method, based on average linkage [54] algorithm (Fig. 5). Subsequently, the severe- and mild category CoVs were clustered in different clades. SARS-CoV-2 was distantly clustered from the other two CoVs in severe category, and two of the mild coronaviruses, HCoV-OC43 and HCoV-NL63 (28). Thus, it can be stated that high-score and low-score codons are detected in mild category CoVs (except HCoV-229E). We also observed 7 common codons, GCG, GGG, CCG, CGG, TCG, and ACG, among the 7 CoVs with low RSCU scores. Of these 7 codons, 2 were from the aliphatic group (A-GCG and G-GGG), 1 from the cyclic group (P-CCG), 2 from the basic group (positively-charged) (R-CGA and R-CGG), and 2 from the sulfur-containing groups (S-TCG and T-ACG) (Table 4). Overall, significant differences were observed in the frequencies of 2 glutamine codons (GAG and GAA) in the CoVs in the severe category. In the case of GAG, this codon was lowly expressed in SARS-CoV-2 (0.55) compared to SARS-CoV and MERS-CoVs (0.95 and 0.94, respectively), whereas for GAA, this codon is highly expressed in SARS-CoV-2 (1.44) compared to lowly expressed in SARS-CoV and MERS-CoVs (1.04 and 1.05, respectively).

To understand the evolutionary structure of the CoVs, we obtained an average RSCU value for each codon within the multiple sequences of the same virus (or strain). Next, we created a virus-based 59-dimensional codon feature vector, i.e., RSCU scores of 59 codons. Then, a phylogenetic tree (dendrogram) was constructed using the unweighted pair group method with arithmetic mean (UPGMA), a hierarchical clustering method, based on average linkage [54] algorithm (Fig. 5). Subsequently, the severe- and mild category CoVs were clustered in different clades. SARS-CoV-2 was distantly clustered from the other two CoVs in severe category, and two of the mild coronaviruses, HCoV-OC43
and HCoV-229E, were proximal to severe category CoVs. This finding also indicated a closer codon usage pattern with severe category CoVs (Fig. 4 and Table 4). The principal component analysis of RSCU data also supported the phylogenetic correlations (Fig. 5) of seven CoVs (Supplementary Fig. S1).

Furthermore, we extended our study across 28 different host-virus pairs (Supplementary material 1) that included 18 unique CoVs species targeting 16 different hosts to infer the evolutionary process using codon usage pattern. Next, we constructed the phylogenetic tree of the virus species by applying the same hierarchical clustering method (UPGMA, average linkage) based on the RSCU score. As observed in Figs. 6, 7 candidate HCoVs were distributed into four different clades. The majority of the similar coronaviruses, such as Alpha-CoV 1, Beta-CoV 1, SARS-CoV, and MERS-CoV were collected from different hosts are clustered together. Due to the high sequence similarity among the intragroup species, a similar codon usage pattern was observed, although targeted to different hosts. We also observed one of the evolutionarily close species of SARS-CoV-2, Pangolin-CoV (*Manis javanica*) (same has been confirmed in a previous study [55]). Interestingly, the current analysis highlighted that a possible evolutionary closeness of SARS-CoV-2 with SARS(r)-CoV isolated from the cat host (*Felis catus*).
To assess the potential influence of low or high codon usage bias (intragenic codon bias), we calculated the ENC value for all the 7 viruses. Next, we found that the mean ENC value was ranges from 36.40 (for HCoV-HKU1) to 49.8 (for SARS-CoV). An ENC value >45 was considered as a lower codon usage bias. We observed that the mean ENC values for MERS-CoV and SARS-CoV were relatively higher than those of SARS-CoV-2 and other CoVs. Although SARS-CoV-2 showed a high codon usage bias, it could infect other animals, such as cats and ferrets [56]. However, the ENC values for two CoVs (HCoV-OC43 (ENC:43.794) and HCoV-229E (ENC:43.1)) from mild category were higher than the ENC values for the other two mild CoVs (HCoV-229E (ENC:36.4) and HCoV-NL63 (ENC:37.32)). The previous studies on SARS-CoV and MERS-CoV confirmed similar findings [36,51]. Several other studies on different viruses, such as influenza A [57], and classical swine fever virus [58], reported a low codon usage bias. On the other hand, high codon bias was observed in hepatitis A virus [59].

Next, we analyzed the correlation between the ENC value and the GC content in the third site of codons (GC3) in all the 7 CoVs and plotted (Fig. 7). The plot indicated a possible impact of mutational or selection pressure on codon usage [27]. In addition, the ENC curve showed the expected codon usage. We found all the points lying near the solid line on the left region, i.e., observed value was smaller than that of the expected value. These findings suggested that mutational bias might have a role in determining the codon usage variation in candidate viruses. Furthermore, the codon usage bias for severe category CoVs is low. The present study also confirmed the additional role of dinucleotide abundance for the evolution of severe category coronaviruses.

### 3.5. Association between ENC and GC3

To assess the potential influence of low or high codon usage bias (intragenic codon bias), we calculated the ENC value for all the 7 viruses. Next, we found that the mean ENC value was ranges from 36.40 (for HCoV-HKU1) to 49.8 (for SARS-CoV). An ENC value >45 was considered as a lower codon usage bias. We observed that the mean ENC values for MERS-CoV and SARS-CoV were relatively higher than those of SARS-CoV-2 and other CoVs. Although SARS-CoV-2 showed a high codon usage bias, it could infect other animals, such as cats and ferrets [56]. However, the ENC values for two CoVs (HCoV-OC43 (ENC:43.794) and HCoV-229E (ENC:43.1)) from mild category were higher than the ENC values for the other two mild CoVs (HCoV-229E (ENC:36.4) and HCoV-NL63 (ENC:37.32)). The previous studies on SARS-CoV and MERS-CoV confirmed similar findings [36,51]. Several other studies on different viruses, such as influenza A [57], and classical swine fever virus [58], reported a low codon usage bias. On the other hand, high codon bias was observed in hepatitis A virus [59].

Next, we analyzed the correlation between the ENC value and the GC content in the third site of codons (GC3) in all the 7 CoVs and plotted (Fig. 7). The plot indicated a possible impact of mutational or selection pressure on codon usage [27]. In addition, the ENC curve showed the expected codon usage. We found all the points lying near the solid line on the left region, i.e., observed value was smaller than that of the expected value. These findings suggested that mutational bias might have a role in determining the codon usage variation in candidate viruses. Furthermore, the codon usage bias for severe category CoVs is low. The present study also confirmed the additional role of dinucleotide abundance for the evolution of severe category coronaviruses.

### 3.6. Neutrality plot and regression analysis

As discussed earlier, a neutrality plot indicates the degree of mutational pressure on codon usage. In synonymous codons, only the last nucleotide differed (except two codons each from arginine, leucine, and serine amino acid). Thus, the nucleotide change at the third position of a codon implies the possible role of mutational force [27,60], rendering that it is an indicator of the extent or the degree of biasness towards base composition [61]. The correlation between GC12 and GC3 could be attributed to mutational forces [62]. Next, we constructed a neutrality plot and a linear regression analysis between GC3 and GC12, between GC3 and GC1, and between GC3 and GC2. First, we showed the distribution of average GC content for the seven strains of HCoVs (Fig. 8(a)), where we observed various overlapping regions between SARS-CoV-2 and HCoV-229E, SARS-CoV and MERS-CoV, with similar GC content usage pattern. It has also been observed that position-wise distribution might vary among viruses (Fig. 8(b)). The neutrality plot analysis of GC3 against GC1, GC2, and GC12 for all seven CoVs is shown in Fig. 9. The solid line in the figure represents the regression line. The details of the regression line with significant statistical p-value and coefficient of determination ($R^2$) value are shown in Table 5. The slope of the regression line suggests the relative neutrality (mutation pressure) for GC1/GC2/GC12 and the relative constraint on the GC3 (natural selection). Intriguingly, a strong correlation was established between GC12 and GC3 for SARS-CoV-2 ($R^2 = 0.965, p < 0.001$) and HCoV-229E ($R^2 = 0.931, p < 0.001$) (Table 5), a strong correlation was established between GC12 and GC3 for SARS-CoV-2 [62]. A positive correlation was established between GC3 and GC2 for HCoV-NL63 ($R^2 = 0.81, p < 0.001$) and a negative correlation between GC3 and GC1 for HCoV-229E ($R^2 = 0.70, p < 0.001$) explicates the role of mutational pressure towards 2nd and 3rd codon position, respectively. Further, we observed a comparatively low correlation between GC3 and GC1 for HCoV-NL63 ($R^2 = 0.48, p < 0.001$) and SARS-CoV ($R^2 = 0.37, p < 0.001$).

### 3.7. Comparative host adaptability of different parasites

The host-parasite relation is a complex process and influenced by multiple interacting factors [63]. The viruses are pure parasites and co-evolved with their host. A specific position in the parasite genome may be involved in host-specific adaptations [64] that can exploit the host codon usage [49,65]. The highly expressed viral proteins typically show similar codon usage bias to target host proteins [49,50,66,67]. Similar to viruses, various harmful parasites also cause diseases in humans by the process of host co-adaptation. Therefore, it may be interesting to see how different human parasites including HCoVs, adapt to their hosts by exploiting the host codon usage.

To compare the codon adaptability of different human host parasitic species, we collected a total of 8 species of parasites, 8 species from each of the protozoa (Plasmodium falciparum, Giardia intestinalis), bacterium (Mycobacterium tuberculosis, Bordetella pertussis), fungus (Trichophyton rubrum, Trichophyton mentagrophytes), and virus (Varicella zoster, Rhinovirus A/B/C) groups. Next, we compared the codon adaptability score with the above human parasites and the 7 candidate HCoV species. The CAI was calculated using Homo sapiens as the reference species and CAIcal, a web-server tool. The CAI scores are presented using a box plot for different genes (Fig. 10) and data are reported in supplementary file (Supplementary material 2). We found that CAI scores showed a diverse adaptability pattern among species from the four kingdoms. The CAI in two species of the same group varies except for the fungus group that shows a low CAI score. The CAI scores for the bacteria are relatively higher than other parasites, indicating that bacteria are efficient in adapting to human host. The CAI for viruses showed moderate (0.68–0.72) variations. Among the HCoVs, codon adaptability indicated a diverse adaptability pattern. The CAI value of SARS-CoV-2 is low and closer to SARS-CoV, but both the CAI values were lower than those of MERS-CoV. The low CAI value of SARS-CoV-2 suggested that the gene expression of SARS-CoV-2 was less efficient than that of SARS-CoV, MERS-CoV, and other viruses [36,68,69]. On the other hand, within the mild category, CAI scores show high HCoV-OC43 and HCoV-229E (close to MERS-CoV from severe category) and low HCoV-HKU1 and HCoV-NL63 (close SARS-CoV-2 and SARS-CoV from severe category). This phenomenon indicates that HCoV-OC43 and HCoV-229E have a higher adaptation to the human host and a higher rate of acute respiratory tract infections than other CoVs [70].

\[^3\] \url{https://ppuiqbo.me/programs/CAIcal/}
4. Conclusions

We performed an extensive quantitative study on the genome sequence of 7 HCoVs. The critical outcomes of this comparative study on various CoVs have been presented. The percentage of high AT content (58 – 67%) for all CoVs indicates the existence of compositional bias. The current analysis described a high usage of GC and GA dinucleotides (first two positions of codons) and CT dinucleotide (last two positions of codons) that belong to strong hydrogen and pyrimidine groups, respectively. In contrast, CG dinucleotide is low in both cases for all seven CoVs. Next, we observed that the GC content in the third codon position (GC3) in HCoVs. In terms of synonymous codon usage pattern, we observed a high degree of similarity within mild category HCoVs, while in the severe category, the SARS-CoV-2 has the highest codon preference. The common synonymous codon usage for all 7 CoVs was from aliphatic and hydroxyl amino acid groups with high-RSCU values. Simultaneously, low-usage codons are from aliphatic, cyclic, positively-charged, and sulfur-containing groups. A phylogenetic study based on

Fig. 8. (a) The distribution of overall GC-content composition taking all three codon positions for seven human coronaviruses; (b) The position wise distribution of GC-content (GC-content at the first position, Position = GC1; GC-content at the second position, Position = GC2; GC-content at the third position, Position = GC3).
RSCU can differentiate between severe- and mild category CoVs. The phylogenetic study with other coronaviruses revealed that the two CoV species isolated from pangolins (Manis javanica, pangolin-CoV) and cats (Felis catus, SARS(r)-CoV) were in proximity with SARS-CoV-2. It has also been observed that the same CoV species from different hosts are clustered together in the phylogenetic tree that depicts a similar synonymous codon usage pattern. The lowest ENC and CAI values (very close to mild category CoVs) for SARS-CoV-2 clearly indicated a poor adaptation to human codon usage. The overall analysis utilizing different bias indices suggested a potential role of mutation pressure on codon usage, and these findings provide cues for understanding the mechanism of mutations among HCoVs. The analysis of host codon adaptability depicted a lower CAI score for the fungi group and a higher for the bacteria group. Furthermore, CAI scores indicated relatively closer codon adaptability for three coronaviruses, SARS-CoV-2, SARS-CoV, and HCoV-HKU1. Although SARS-CoV-2 exhibits a codon adaptability similar to SARS-CoV, the RSCU-based phylogenetic tree showed proximity among SARS-CoV and MERS-CoV. The current analysis might explain the unique aspects of the virus concerning their resistance to innate immunity and future drug discovery experiments.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.05.008.

Author statement

JKD and SR conceived and designed the study. JKD collected the data and performed the computational study. JKD and SR wrote the original manuscript. Both the authors edited and approved for final submission.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Table 5

Linear regression computed by comparing GC3 to GC1, GC2 and GC12 for all seven CoVs. The coefficient of determination (R2), regression line with a slope, intercept, and p-value are calculated.

| Strains                      | GC3 Vs. | Regression line | R²     | p-value |
|-----------------------------|---------|----------------|--------|---------|
| SARS-CoV                    | GC1     | y = 0.33x + 0.38 | 0.3768 | 0.0     |
|                             | GC2     | y = 0.03x + 0.38 | 0.0141 | 0.1713  |
|                             | GC12    | y = 0.18x + 0.38 | 0.2521 | 0.0     |
| SARS-CoV-2                  | GC1     | y = 0.00x + 0.47 | 0.0005 | 0.6602  |
|                             | GC2     | y = 0.65x + 0.20 | 0.9659 | 0.0     |
|                             | GC12    | y = 0.36x + 0.33 | 0.9695 | 0.0     |
| MERS-CoV                    | GC1     | y = -0.12x + 0.44 | 0.0534 | 0.0004  |
|                             | GC2     | y = -0.19x + 0.46 | 0.022  | 0.0236  |
|                             | GC12    | y = -0.02x + 0.45 | 0.0005 | 0.7275  |
| HCoV-OC43                   | GC1     | y = -0.08x + 0.48 | 0.0098 | 0.2173  |
|                             | GC2     | y = 0.30x + 0.28  | 0.09   | 0.0001  |
|                             | GC12    | y = -0.10x + 0.39 | 0.0502 | 0.0048  |
| HCoV-HKU1                   | GC1     | y = -0.82x + 0.58 | 0.2321 | 0.0039  |
|                             | GC2     | y = 0.72x + 0.22  | 0.3552 | 0.0002  |
|                             | GC12    | y = -0.02x + 0.39 | 0.0022 | 0.7941  |
| HCoV-229E                   | GC1     | y = -0.14x + 0.51 | 0.7023 | 0.0     |
|                             | GC2     | y = 0.66x + 0.18  | 0.9799 | 0.0     |
|                             | GC12    | y = 0.26x + 0.34  | 0.9311 | 0.0     |
| HCoV-NL63                   | GC1     | y = -0.25x + 0.50 | 0.4835 | 0.0     |
|                             | GC2     | y = 0.43x + 0.28  | 0.8114 | 0.0     |
|                             | GC12    | y = 0.11x + 0.39  | 0.2866 | 0.0     |

Fig. 9. The neutrality plot analysis of GC3 against GC1, GC2, and GC12. The solid line represents the regression line.

Comparison

- GC3-GC1
- GC3-GC2
- GC3-GC12

Table 5

Linear regression computed by comparing GC3 to GC1, GC2 and GC12 for all seven CoVs. The coefficient of determination (R²), regression line with a slope, intercept, and p-value are calculated.
Fig. 10. A box-plot shows the range of codon adaptability indices of 15 different human host parasites (from four kingdoms: virus, fungus, protozoa, bacteria), including seven HCoVs. The CAI range for all protein-coding genes (<20000 bp) of the respective parasite species.

Acknowledgments

The authors thank Dr. Aditi Mitra, Dr. Amit Chakraborty, and Ms. Papia Basu Thakur for their constructive suggestions on improving this manuscript.

References

[1] P.C. Woo, S.K. Lau, C.S. Lam, C.C. Lau, A.K. Tsang, J.H. Lau, R. Bai, J.L. Teng, C. C. Tsang, M. Wang, et al., Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus, J. Virol. 86 (2012) 3995–4008.

[2] S. Perlman, J. Netland, Coronaviruses post-sars: update on replication and pathogenesis, Nat. Rev. Microbiol. 7 (2009) 439–450.

[3] J. Pépin, S. Lai, L. Yam, W. Lim, J. Nicholls, W. Yee, W. Yan, M. Cheung, et al., Coronavirus as the possible cause of severe acute respiratory syndrome, Lancet 363 (2004) 1319–1325.

[4] R.J. de Groot, S.C. Baker, R.S. Baric, C.S. Brown, C. Drosten, L. Enjuanes, R. Grieshaber, et al., Commentary: middle east respiratory syndrome coronavirus (mers-cov): announcement of the coronavirus study group, J. Virol. 87 (2013) 7790–7792.

[5] A. Chafekar, B.C. Fielding, Mers-cov: understanding the latest human coronavirus threat, Viruses 10 (2018) 93.

[6] L.E. Gralinski, V.D. Menachery, Return of the coronavirus: 2019-ncov, Viruses 12 (2020) 135.

[7] Y. Ruan, C.L. Wei, A.E. Ling, V.B. Vega, H. Thoreau, S.Y.S. Thoe, J.-M. Chia, P. Ng, I.S. Belalov, A.N. Lukashov, Causes and implications of codon usage bias in rna genomes and mutations in sars-cov-2 proteins from indian isolates, Gene Rep. (2021) 101044.

[8] C. Yin, Genotyping coronavirus sars-cov-2: methods and implications, Genomics 112 (5) (2020) 3588–3596.

[9] V.D. Menachery, K. Devibhui, Coronavirus non-structural protein 16: evasion, attenuation, and possible treatments, Virus Res. 194 (2014) 191–199.

[10] T. Goto, The pathogenesis of influenza in humans, Rev. Med. Virol. 11 (2001) 227–241.

[11] P.J. Walker, J.A. Cowley, Viral genetic variation: implications for disease diagnosis and detection of shrimp pathogens, FAO fisheries, Tech. Paper (2000) 54–59.

[12] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1) (1980) 197.

[13] R.J. Grocock, P.M. Sharp, Synonymous codon usage in cryptoparvovirus parvum: identification of two distinct trends among genes, Int. J. Parasitol. 31 (2001) 402–412.

[14] R.K. Singh, A. Singh, A. Tiwari, K. Dhamma, J. Das, A. Munjal, R.K. Nair, Analysis of nipa virus codon usage and adaptation to hosts, Front. Microbiol. 10 (2019) 886.

[15] S.P. Belakov, A.N. Lukashov, Causes and implications of codon usage bias in rna viruses, PLoS One 8 (2013) e56642.

[16] J.K. Das, A. Sengupta, P.P. Choudhury, S. Roy, Characterizing genomic variants and mutations in sars-cov-2 proteins from indian isolates, Gene Rep. (2021) 101044.

[17] C. Yin, Dinucleotide repeats in coronavirus sars-cov-2 genome: evolutionary implications, arXiv Preprint (2020) https://arxiv.org/abs/2006.00280 (arXiv:2006.00280).

[18] P. Ausewakhali, Composition bias and genome polarity of rna viruses, Virus Res. 109 (2005) 33–37.

[19] R. Klinting, E.A. Gould, X. De Lamblaire, G+C content differs in conserved and variable amino acid residues of flaviviruses and other evolutionary groups, Infect. Genet. Evol. 45 (2016) 332–340.

[20] J.B. Plotkin, G. Kula, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42.
