biomolecules

MDPI

Article

Genome-Wide Analysis of Codon Usage Patterns of SARS-CoV-2 Virus Reveals Global Heterogeneity of COVID-19

Saadullah Khattak [1,2], Mohd Ahmar Rauf [2,3], Qamar Zaman [4], Yasir Ali [5], Shabeen Fatima [6], Pir Muhammad [7], Tao Li [1], Hamza Ali Khan [4], Azhar Abbas Khan [1,8], Ebenezeri Erasto Ngowi [1], Dong-Dong Wu [1,9,]* and Xin-Ying Ji [1,10,]*

[1] Henan International Joint Laboratory for Nuclear Protein Regulation, School of Basic Medical Sciences, Henan University, Kaifeng 475004, Henan, China; saadullah@henu.edu.cn (S.K.); gataoli@yeah.com (T.L.); azharabbaskhan84@gmail.com (A.A.K.); ebenezerngowi92@gmail.com (E.E.N.)
[2] School of Life Sciences, Henan University, Kaifeng 475004, Henan, China; hb7059@wayne.edu or mxr2481@miami.edu
[3] School of Pharmaceutical Sciences, Wayne State University, Detroit, MI 48202, USA
[4] Department of Bioinformatics, Hazara University, Mansehra 21120, Pakistan; qamarktk348@gmail.com (Q.Z.); hamzaalikhan2211@gmail.com (H.A.K.)
[5] National Center for Bioinformatics, Quaid-i-Azam University, Islamabad 45320, Pakistan; yasirkhanqu@gmail.com
[6] Department of Biosciences and Bioinformatics, Capital University of Science and Technology, Islamabad 44000, Pakistan; shabeenfatima53@gmail.com
[7] Henan-Macquarie University Joint Centre for Biomedical Innovation, School of Life Sciences, Henan University, Kaifeng 475004, Henan, China; pir.muhammad786@gmail.com
[8] Department of Biochemistry, Hazara University, Mansehra 21110, Pakistan
[9] School of Stomatology, Henan University, Kaifeng 475004, Henan, China
[10] Kaifeng Key Laboratory of Infection and Biological Safety, School of Basic Medical Sciences, Henan University, Kaifeng 475004, Henan, China
* Correspondence: ddwubiomed2010@163.com (D.-D.W.); 10190096@vip.henu.edu.cn (X.-Y.J.); Tel.: +86-371-23880525 (D.-D.W.); +86-371-23880585 (X.-Y.J.)

check for updates

Citation: Khattak, S.; Rauf, M.A.; Zaman, Q.; Ali, Y.; Fatima, S.; Muhammad, P.; Li, T.; Khan, H.A.; Khan, A.A.; Ngowi, E.E.; et al. Genome-Wide Analysis of Codon Usage Patterns of SARS-CoV-2 Virus Reveals Global Heterogeneity of COVID-19. *Biomolecules* **2021**, *11*, 912. https://doi.org/10.3390/biom11060912

Academic Editor: Vladimir N. Uversky

Received: 20 April 2021
Accepted: 14 June 2021
Published: 18 June 2021

Abstract: The ongoing outbreak of coronavirus disease COVID-19 is significantly implicated by global heterogeneity in the genome organization of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The causative agents of global heterogeneity in the whole genome of SARS-CoV-2 are not well characterized due to the lack of comparative study of a large enough sample size from around the globe to reduce the standard deviation to the acceptable margin of error. To better understand the SARS-CoV-2 genome architecture, we have performed a comprehensive analysis of codon usage bias of sixty (60) strains to get a snapshot of its global heterogeneity. Our study shows a relatively low codon usage bias in the SARS-CoV-2 viral genome globally, with nearly all the over-preferred codons' A.U. ended. We concluded that the SARS-CoV-2 genome is primarily shaped by mutation pressure; however, marginal selection pressure cannot be overlooked. Within the A/U rich virus genomes of SARS-CoV-2, the standard deviation in G.C. (42.91% ± 5.84%) and the GC3 value (30.14% ± 6.93%) points towards global heterogeneity of the virus. Several SARS-CoV-2 viral strains were originated from different viral lineages at the exact geographic location also supports this fact. Taking all together, these findings suggest that the general root ancestry of the global genomes are different with different genome's level adaptation to host. This research may provide new insights into the codon patterns, host adaptation, and global heterogeneity of SARS-CoV-2.

Keywords: Coronavirus; SARS-CoV-2; Codon usage bias; COVID-19; heterogeneity of COVID-19; mutational bias; natural selection

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a non-segmented positive sense, enveloped, single-stranded RNA virus that belongs to genus Coronavirus, order Nidovirales, and family Coronaviridae [1]. It is the seventh member of the Coronaviridae family that infects humans. Among the prior six, NL63, HKU1, 229E and OC43 are known to show mild symptoms, while MERS-CoV, SARS-CoV, and recently emerged SARS-CoV-2 are associated with severe diseases [2]. SARS-CoV-2 propagates a respiratory and gastrointestinal infection referred to as coronavirus disease 2019 (CoVID-19), an ongoing pandemic [3]. The symptoms include dry cough, fever, fatigue, dyspnea, and lymphopenia with occasional human complications like severe acute respiratory syndrome (SARS), pneumonia, and even death [4]. The SARS-CoV-2 genome ranges from 27–32 kb and carries a unique replication strategy [5]. The genome comprises of several open reading frames like ORF1ab, ORF3a, ORF6, ORF7a, ORF7b ORF8, and ORF10 that codes for both structural proteins like spike protein, replicase polyprotein, envelope protein, accessory proteins, nucleocapsid proteins, and other Non-structural proteins (NSP) [6–8].

However, the recent global spread of SARS-COV-2, with surging transmissions in many parts of the world and steep declines in others, has raised fundamental questions about the viral genome's evolution and adaptation with the host genome [9]. The counter causes of global heterogeneity are variation in host immune systems, mutations, deletions, recombination, genetic drift, and the founder's impact [10]. According to some studies, adaptations at nucleotide [1] and amino acid [4] regions, as well as heterogeneity found in structural proteins, significantly spike proteins, are possible causes of global variable transmission patterns [11–13]. The evolution of the viral genome, transmissibility, and virulence can be linked to intra-host viral evolution after infection [13,14]. Virus infectivity is directly affected by deleted variants of accessory proteins and non-structural proteins [15–17]. However, both reports are limited to comparing only a few genomes from a few countries and concentrating primarily on structural proteins. Codon usage research will allow us to understand the global heterogeneity of SARS-CoV-2, the evolution of its genome, and its adaptation to the host genome. In the translation process, different strains use different codons [13,18–20]. The translation process is primarily driven by synonymous codons [21], which code for the same amino acids but at different frequencies [22,23]. Codon usage bias (CUB) is common in eukaryotes, prokaryotes, viruses, and other genes in the same organism [24]. Codon usage patterns are more pronounced in highly expressed genes than in lower expressed genes [25]. The codon usage pattern is influenced by transcriptional factors, translation, G.C. content, secondary motifs, and gene expression level [23]. Natural selection and mutational pressure, on the other hand, are the primary driving forces [26,27]. Several studies have concluded that mutational pressure is the primary driver of codon usage patterns in SARS-CoV-2 [23,28–31].

In contrast to eukaryotes and prokaryotes, viral genomes have unique characteristics such as host dependence for replication, synthesis, and protein transmission. These essential characteristics affect virus development, adaptation, survival, and host immune system disregard [32,33]. Understanding the codon usage pattern can thus provide valuable insights into the evolutionary process, host adaptation, disregard for the host immune system, viral pathogenesis, and global heterogeneity of the SARS-CoV-2 virus [12,14,20,29]. For this purpose, we have characterized the codon usage patterns among sixty SARS-CoV-2 viral strains to unravel the role of codon architecture in viral pathogenesis. The possible causes of this newly emerging coronavirus's global heterogeneity and its adaptation to host and insights into the viral lineages were gained using bioinformatics approaches.

2. Material and Methods:

2.1. Acquisition of Data

The whole-genome sequences of 60 SARS-CoV-19 isolates reported worldwide were retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov (accessed on 16 February 2021)). The sequences were selected based on geographical distribution across five continents, transmission rate, and isolation methods. The CDS sequences were concatenated using the Artemis genome browser [34] and aligned through ClustalW [35]. The codon usage data of the SARS-CoV-2 host, Homo Sapiens, was retrieved from the codon usage database (https://www.kazusa.or.jp/codon/ (accessed on 23 February 2021)).

2.2. Nucleotide Composition of SARS-CoV-2

Nucleotide composition analysis of CDS sequences of 60 SARS-CoV-2 confines was helped out through the Artemis genome program [34], codonW (http://codonw.sourceforge.net//culong.html (accessed on 5 January 2021)) and Emboss pilgrim [36]. The genome-level individual frequencies of the nucleotide (A, T, G and C,) were checked during composition analysis. This was followed by finding the collective frequencies of A.U. and G.C. throughout the genomes and the occurrence of A, U, G and C at 1st, 2nd, and third places of codons. The mean value of AU3, GC12 and GC3 was recorded for all the strains. Herein, AUG and UGG bearing no synonymous codons, while UGA, UAG, and UAA stop codons were neglected in the protocol.

2.3. Codon Preference Characteristics

The overall relative synonymous codon usage (RSCU) is the proportion of codons' observed recurrence compared to the regular recurrence of codons under uniform synonymous codon utilization. An RSCU value equivalent to 1 reflects that codon usage is not biased. RSCU values under 1.0 happen when the observed recurrence is not precisely the regular recurrence [37]. The codon preferences of SARS-CoV-2 were calculated and compared with other hosts like humans, dogs, cats and cattle to show the codon preferences of natural and other hosts.

2.4. Analysis of Codon Usage in SARS-CoV-2

CodonW (available at http://sourceforge.net/projects/codonw (accessed on 5 January 2021)) was employed to perform codon usage bias analysis by calculating relative synonymous codon usage RSCU values [38]. An adequate number of codons ENC plot analysis was also performed to reveal the usage bias pattern in the CDS regions of SARS-CoV-2 [39]. To carry out the correlation studies and demonstrate the adaptation of SARS-CoV-2 to its host, the Codon adaptation index of all the strains was measured against the reference human genome's codons usage pattern [40].

2.5. ENC-Plot Analysis

An ENC plot will shed light on the ENC relationship and the G.C. content at the third codon location (GC3). This method demonstrates gene codon usage bias. It is commonly used to assess the extent of a gene's codon usage bias. To determine the correlation, the predicted ENC values for the corresponding GC3 were calculated using Singh et al. process [41]. There is a solid line that represents the expected curve where if the strains lie close or on the line, represents mutational pressure being the driving force or if lower, shows selection pressure as well in addition to mutational pressure.

2.6. Neutral Evolution Analysis

Neutral evolution analysis or neutrality plot analysis is used to determine the factors that affect codon usage preference [42]. It was used to evaluate the mutation-selection equilibrium in shaping the codon usage bias. Using GC3 as a horizontal coordinate and GC12 as a vertical coordinate, the GC3 and GC12 contents were plotted with a regression line to determine how mutational pressures played a role in forming codon usage bias instead of natural selection [42].

2.7. Codon Adaptation Analysis

Codon usage similarities of host *Homo Sapiens* and SARS-CoV-2 genomes were quantified using codon adaptation index (CAI) [40]. CAI analysis can reveal respective codons for those amino acids that are more efficient for translation and are highly expressed genes. CAI values ranges between 0.0 and 1.0, where higher CAI values depict higher gene expression potential and vice versa. Further, values that are close to one indicate that codons with higher RSCU values are used in the CDS sequences. The host synonymous codon usage bias data was extracted from the codon usage bias database (http://www.kazusa.or.jp/codon/ (accessed on 23 February 2021)), compiled from 93487 CDS sequences of Homo Sapiens. Wilcoxon & Mann Whitney test was utilized to identify statistically significant CAI values [43]. To show that the CAI value's significance is solely due to codon usage pattern preferences, the expected CAI (eCAI) was calculated at a confidence interval of 95% [44].

2.8. Correspondence Analysis (COA)

Correspondence analysis is a multivariate statistical analysis that is used to detect variable and sample relationships. Correspondence analysis (COA) is a broadly used statistical method to analyze multiple factors and their influences on a specific component. COA displays sets of rows and columns in a particular data set [45]. This approach helps to reflect the trend of strain change directly. The codonW program was used in this study to perform COA based on RSCU values.

2.9. Phylogenetic Analysis

Phylogenetic analysis was performed to depict the genetic diversity and evolutionary relationships among SARS-CoV-2 strains retrieved from NCBI. The sequences were aligned using the ClustalW program [46]. The phylogenetic tree was constructed using Mega 7, utilizing the maximum likelihood method [47]. In the present study, we performed comprehensive analyses of codon usage and composition of SARS-CoV-2 strains and checked the possible leading evolutionary element of the biases found.

3. Results and Discussion

3.1. Nucleotide Composition Analysis of SARS-CoV-2

Nucleotide composition being the major force in affecting codon usage pattern was measured to evaluate its impact on the codon usage pattern of CDS sequences of SARS-CoV-2 [48]. The composition frequency and trend of each nucleotide were U (32.19% ± 0.05) > A (29.85% ± 0.02) > G (19.56% ± 0.05) > C (18.39% ± 0.07), which is consistent with the trend in other Coronaviruses like SARS and MERS [29] which manifestly indicate U codon being the more frequent one throughout SARS-CoV-2 genome (Supplementary Table S1, Figure 1A) However, these trends differ from other viruses like RSV which follows A > U > G > C and H1N1 and H3N2 order A > G > U > C [29]. AU's mean values (62.04% ± 0.04) and GC (37.96% ± 0.04) emphasize the CDS genome as AU-rich genome (Supplementary Table S1). The nucleotide composition of codons at third position U3 (42.73% ± 3.93) > A3 (28.80% ± 4.03) > C3 (16.08% ± 2.21) > G3 (14.04% ± 4.90 showed inconsistent trend against SARS-CoV and MERS which follows U3 > A3 > G3 > C3 [29]. Interestingly, this trend is also inconsistent with the overall trend of bases in the genome (Supplementary Table S1, Figure 1B). However, the higher AU content results align with other RNA viruses like SARS, showing high A.U. content and preferred A/U ending codons [49]. The GC12 value of 42.91% ± 5.84 and GC3 value of 30.14% with a standard deviation SD of 6.93 indicates the biasness in the codon usage pattern manifesting global heterogeneity (Supplementary Table S1).

3.2. Relative Synonymous Codon Usage (RSCU) Analysis

The RSCU values for 60 strains were calculated and compared to the host's human, dog, cat, and cattle genomes to get insights into the codon usage bias pattern of the SARS-CoV-2 viral genome. The results implicitly demonstrated that all 18 frequent codons were A/U ended, which shows SARS-CoV-2 genomes, higher bias towards A/U than G/C. These results are consistent with other viral genomes like avian rotaviruses, equine influenza viruses, and Crimean-Congo hemorrhagic fever virus [49]. In contrast, all the ten under preferred codons were G/C ended, which manifests that the SARS-CoV-2 viral genome is mostly under mutational pressure, which helps them avoid the host's innate immunity [50] (Table 1, Figure 2). 14 out of 59 codons were similar in both the host and viral genome, which improves the translation efficiency of the SARS-CoV-2 within humans like BTV viruses which have 9/59 similar codons with its host Bos Taurus [51,52] (Table 1, Figure 2). Analysis of RSCU values of SARS-CoV-2 and its different hosts uncovered the codon preferences of SARS-CoV-2, human, dog, cat, pig, horse, and cattle (Table 1). The average RSCU of SARS-CoV-2 was compared to that of its regular (human) and accidental (dog) hosts along with other animal species, which revealed that the codon preference of SARS-CoV-2 and its hosts (natural, accidental, and other) are not similar (Figure 2) [53].

Table 1. The Relative synonymous codon usage (RSCU) value of 59 codons encoding 18 amino acids of SARS-CoV-2 virus with their hosts Homo sapiens, dog, cat, and cattle's RSCU values [53].

AA	Codon	SARS-CoV-2	Human	Dog	Cat	Cattle	AA	Codon	SARS-CoV-2	Human	Dog	Cat	Cattle
Phe	UUU	1.31	0.97	1.09	0.77	0.85	Ala	GCU	1.78	1.01	1.07	0.69	1.12
	UUC	0.69	1.03	0.91	1.33	1.15		GCC	0.69	1.12	0.27	1.79	0.57
Leu	UUA	1.51	0.50	1.32	0.35	0.38		GCA	1.26	1.18	1.18	0.76	0.94
	UUG	1.25	1.00	0.51	0.76	0.71		GCG	0.27	0.7	0.48	0.50	0.35
	CUU	1.31	0.81	1.22	0.67	0.70	Tyr	UAU	1.06	0.71	1.15	0.78	0.87
	CUC	0.50	1.07	1.09	1.09	1.09		UAC	0.93	1.29	0.85	1.22	1.12
	CUA	0.80	0.46	1.42	0.36	0.36	His	CAU	1.04	0.85	1.2	0.74	0.81
	CUG	0.63	2.33	0.51	2.57	2.59		CAC	0.96	1.15	0.80	1.26	1.18
Iie	AUU	1.44	1.13	1.05	0.95	0.98	Gln	CAA	1.35	0.49	1.25	0.56	0.49
	AUC	0.72	1.37	0.94	1.58	1.57		CAG	0.65	1.51	0.75	1.44	1.50
	AUA	0.84	0.50	1.01	0.47	0.45	Asn	AAU	1.10	0.98	1.18	0.82	0.92
Val	GUU	1.58	0.79	1.12	0.62	0.64		AAC	0.90	1.02	0.82	1.18	1.07
	GUC	0.66	0.90	0.57	1.13	1.01	Lys	AAA	1.36	0.88	1.37	0.86	0.84
	GUA	0.91	0.52	1.67	0.38	0.40		AAG	0.64	1.12	0.63	1.14	1.15
	GUG	0.85	1.79	0.64	1.87	1.95	Asp	GAU	1.13	0.99	1.13	0.84	0.98
Ser	UCU	1.44	1.15	1.35	1.12	1.04		GAC	0.87	1.01	0.87	1.16	1.01
	UCC	0.55	1.17	1.04	1.48	1.37	Glu	GAA	1.33	0.85	1.17	0.86	1.88
	UCA	1.37	0.93	1.27	0.74	0.79		GAG	0.67	1.15	0.83	1.14	1.05
	UCG	0.28	0.36	0.39	0.38	0.39	Cys	UGU	1.21	0.95	0.89	0.87	0.92
	AGU	1.42	0.98	1.91	0.8	0.87		UGC	0.79	1.05	1.11	1.13	1.07
	AGC	0.93	1.42	1.05	1.47	1.53	Arg	CGU	0.68	0.54	1.17	0.41	0.49
Pro	CCU	1.50	1.20	1.41	1.03	1.08		CGC	0.43	1.11	0.92	1.09	0.94
	CCC	0.56	1.22	1.24	1.51	1.39		CGA	0.87	0.76	0.71	0.55	0.74
	CCA	1.58	1.14	0.92	0.97	1.00		CGG	0.35	1.31	0.48	1.19	1.18
	CCG	0.14	0.45	0.43	0.50	0.53		AGA	2.81	1.18	1.29	1.33	1.26
Thr	ACU	1.32	1.03	1.35	0.84	0.89		AGG	1.36	1.1	1.42	1.41	1.36
	ACC	0.82	1.32	1.06	1.59	1.55	Gly	GGU	1.54	0.71	1.02	0.58	0.71
	ACA	1.52	1.19	1.16	0.94	1.01		GGC	0.84	1.35	1.05	1.42	1.36
	ACG	0.34	0.46	0.43	0.63	0.56		GGA	1.10	1.01	1.27	1.01	0.96
								GGG	0.52	0.93	0.66	0.99	0.96

AA, amino acid; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; RSCU. Relative Synonymous codon usage: Underlined codons represent optimal codons of SARS-CoV-2; Blue color represents most favored codons by both host and SARS-CoV-2 having RSCU value greater than 1, pink and dark green colors represent under preferred (RSCU < 0.6) and over preferred (RSCU > 1.6) codons, respectively.

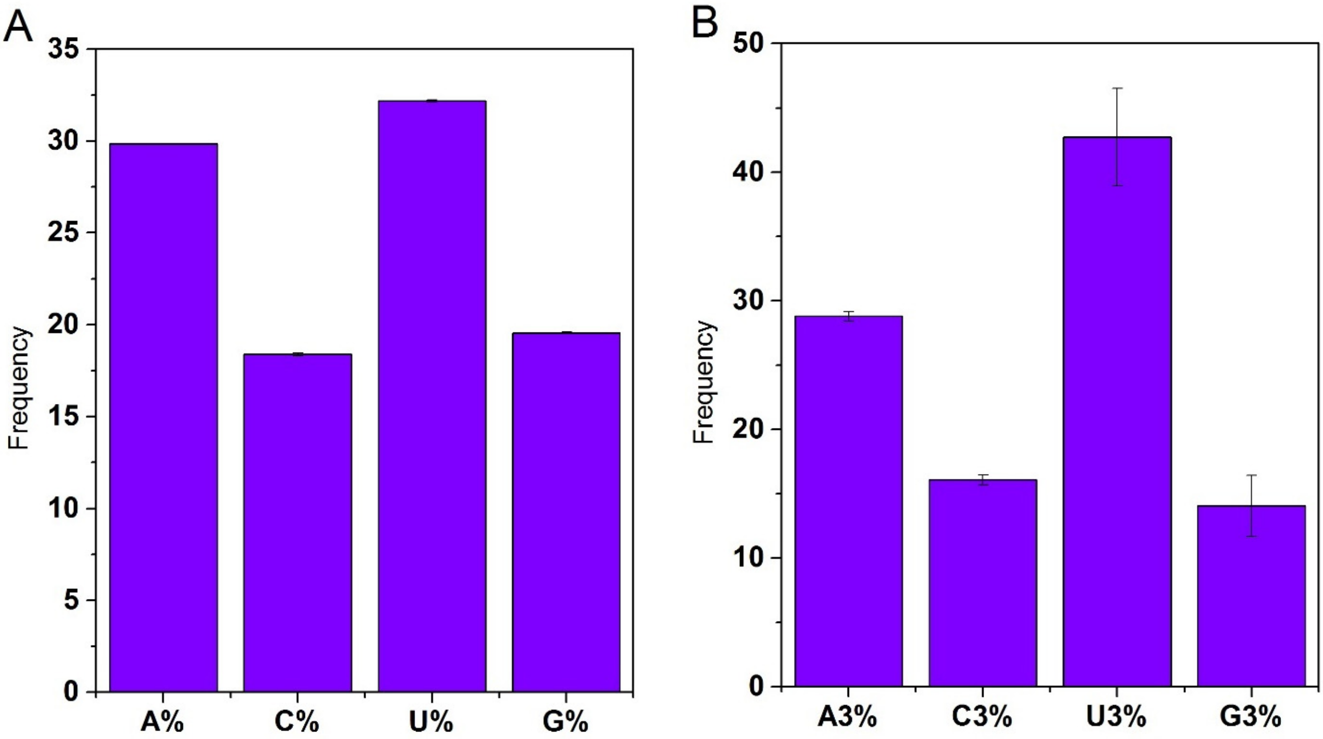

Figure 1. Nucleotide content distribution of CDS sequences of SARS-CoV-2 in percentage. (**A**) The average composition of individual bases A, C, U, G. among 60 strains. (**B**) A, C, U, and G frequency at the third position of the codons.

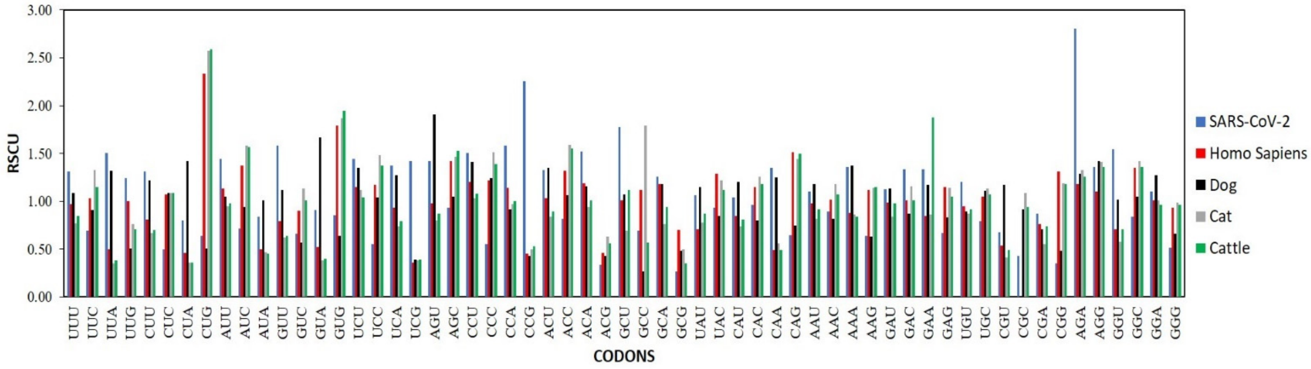

Figure 2. Comparisons of RSCU values of SARS-CoV-2 virus with its host humans, dog, cat, and cattle. The dog, cat, and cattle RSCU is shown as calculated by Dutta et al. [53].

3.3. Codon Usage Bias Analysis of SARS-CoV-2 Genomes

To identify the magnitude of codon usage pattern among CDS sequences of SARS-CoV-2 virus, we have calculated and analyzed all the 60 strains' ENC values. The ENC values ranged from 45.17 of a Chinese strain MT135043.1 to 52.06 of a Spanish strain MT233522.1 with a mean value of 45.80 and a standard deviation of 1.27 (Supplementary Table S2). This value is significantly lower than the mean ENC value of other coronaviruses including BCoV (52.10 ± 2.36), BuCoV HKU11 (51.41 ± 1.85), ECoV (49.31 ± 4.02), FIPV (51.56 ± 1.99) and HCoV-229E (50.29 ± 3.62), among others, which indicates that SARS-CoV-2 uses a relatively lower set of synonymous codons [54,55]. ENC versus GC3 plot was constructed to show that the codon usage pattern of SARS-CoV-2 is mainly under mutational bias (Supplementary Table S3, Figure 3). All the values lie significantly lower to the solid line demonstrate that mutational pressure is not the single factor shaping the codon usage bias. However, other factors such as natural selection are likely to determine the selective constraints on the codon usage bias in 60 strains of SARS-CoV-2 Figure 3. These results are consistent with recently reported SARS-CoV-2 codon usage pattern characterization [20]. However, to show up to which extent both the major driving forces influence the codon usage pattern, the GC12 and GC3 neutrality plot was constructed.

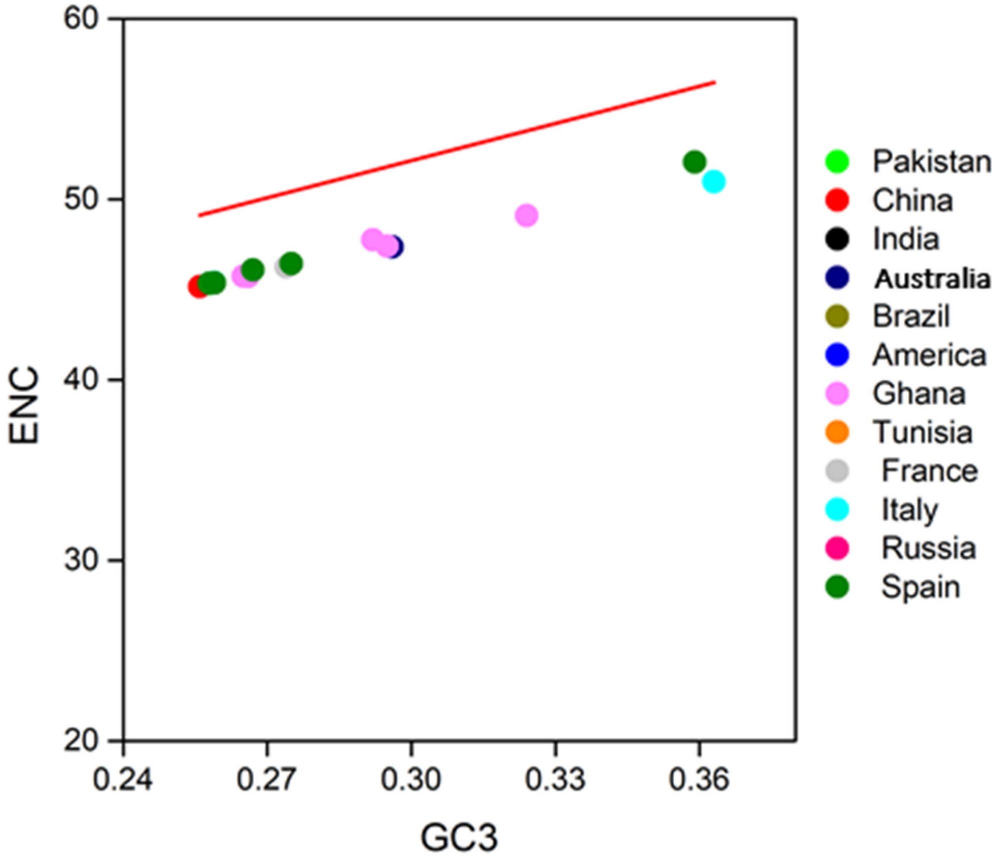

Figure 3. ENC versus GC3 plots of 60 SARS-CoV-2 isolates are represented in different color schemes.

3.4. Neutrality Plot

The degree of Mutation bias and selection pressure was measured through the Neutrality plot between GC12 and GC3 using the Pearson correlation method. The neutrality plot r = −0.31 and $p > 0.01$ indicates that both mutation pressure and natural selection drive the codon usage patterns of SARS-CoV-2. Our analysis also confirmed that most SARS-CoV-2 genomes were present along the unity slope, indicating that the SARS-CoV-2 genome is essentially under mutational pressure. Some of the points were scattered away from the line marking the translational force also marginally but significantly drove the codon usage biases in SARS-CoV-2 (Supplementary Table S4, Figure 4). Our results are consistent with the previously reported results for SARS-CoV-2 [20,29] but inconsistent with some reports [53]. This reconfirms our statement that analysis on a small sample size may result in false positives, so we analyzed a more diverse dataset of 60 genomes.

3.5. Codon Adaptation Analysis

The codon adaptation index analysis was performed to demonstrate the adaptation of the SARS-CoV-2 strains to their host. CAI values are used to determine the level of expression of pathogen proteins in the host and the adaptation of a virus to a host. Sequences with higher CAI values are considered more adapted to a particular host than those with low values. The CAI value of SARS-CoV-2 concerning humans (0.70 ± 0.01) is higher than 0.62 ± 0.01, 0.59 ± 0.00, and 0.61 ± 0.01 concerning the dog, cattle, and cat. These values show the higher adaptation of SARS-CoV-2 to the human environment compared to other hosts. (Supplementary Table S5, Figure S1). The higher tendency of the human CAI value shows that selection pressure from humans can affect the codon patterns of SARS-CoV-2 which have allowed it to use the translation source of humans more efficiently as in line with Marburg virus adaptation to the human host [18] (Supplementary Figure S1). The higher average CAI values of humans compared to dogs, cats, and cattle observed in the present study indicated that dogs and other hosts are less susceptible to COVID 19 than humans. However, the cross-transmission of SARS-CoV-2 between humans and dogs, cats, and cattle, has not been well-understood [53]. Furthermore, to validate the statistical significance, the expected CAI (e-CAI) values were computed for SARS-CoV-2, humans, dogs, cattle, and cat codon usage sets by generating 500 uneven sequences with similar nucleotide contents and amino acid composition as the sequences of interest. The e-CAI values of 0.75, 0.66, 0.61, and 0.65 of humans, dogs, cattle, and cats, respectively, revealed that the generated sequences had a normal distribution.

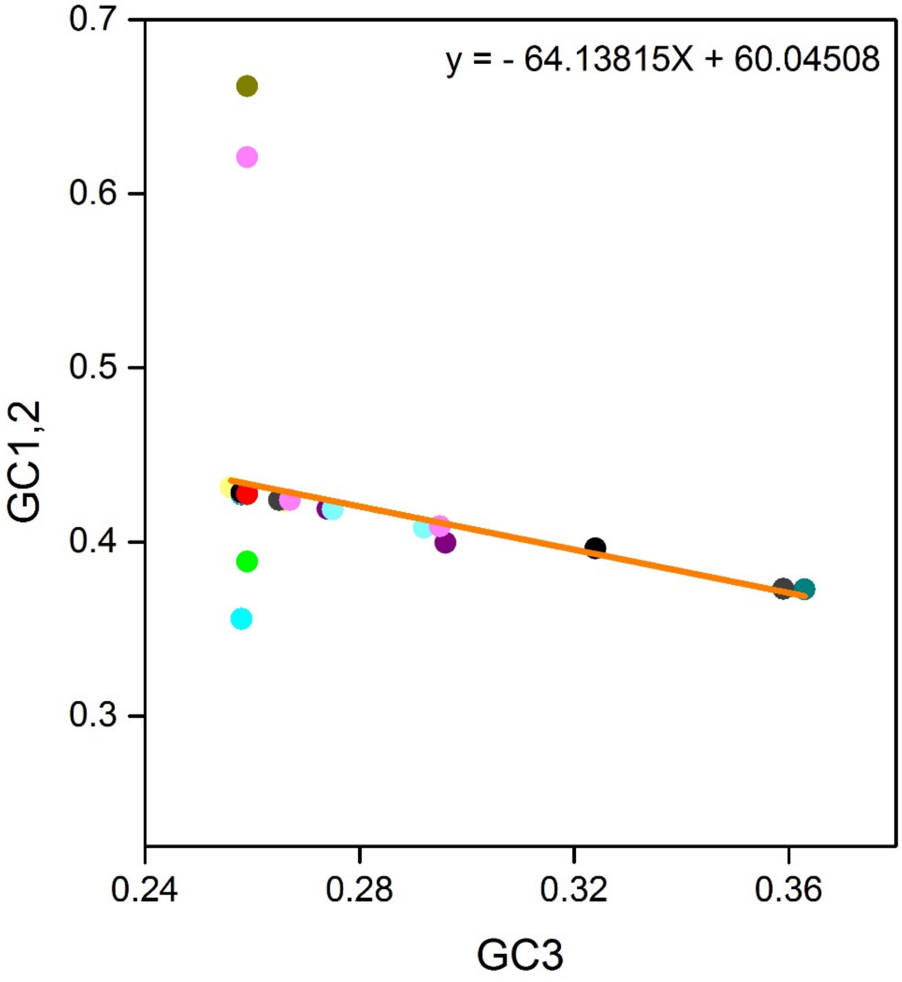

Figure 4. Neutrality plot analysis (GC12 vs. GC3) plots of 60 SARS-CoV-2 isolates are represented in different color schemes. GC12 stands for the average value of G.C. contents at the first and second positions of the codons (GC1 and GC2), while GC3 refers to the G.C. contents at the third codon position.

3.6. COA Analysis

The sixty strains of SARS-CoV-2 were plotted into clusters based on their RSCU values. All the SARS-CoV-2 strains were clustered into two major groups and two minor clusters, while some were found scattered. The first significant cluster has SARS-CoV-2 strains from Australia, Brazil, China, France, Ghana, India, Italy, Pakistan, Spain, and Tunisia. The second considerable cluster has strains from America, Australia, India, Italy, Pakistan, Russia, and Tunisia. The first small cluster has strains from Brazil, India, Italy, and Russia, while the second has strains from France, Spain, and Russia. There were two scattered strains from Ghana, two from Australia, and one from Italy Figure 5. These results suggested that geographical locations play an essential role in the SARS-CoV-2 evolutionary process and a synonymous codon usage pattern. Besides, it is also highlighted that each infected country has emerged from more than one viral genetic lineage, which depicts those geographic locations have a crucial role in shaping codons [56]. Some countries have distinct genomes, probably due to viral transmission through international trade, human traveling, or bird migration, as in Crimean-Congo [56].

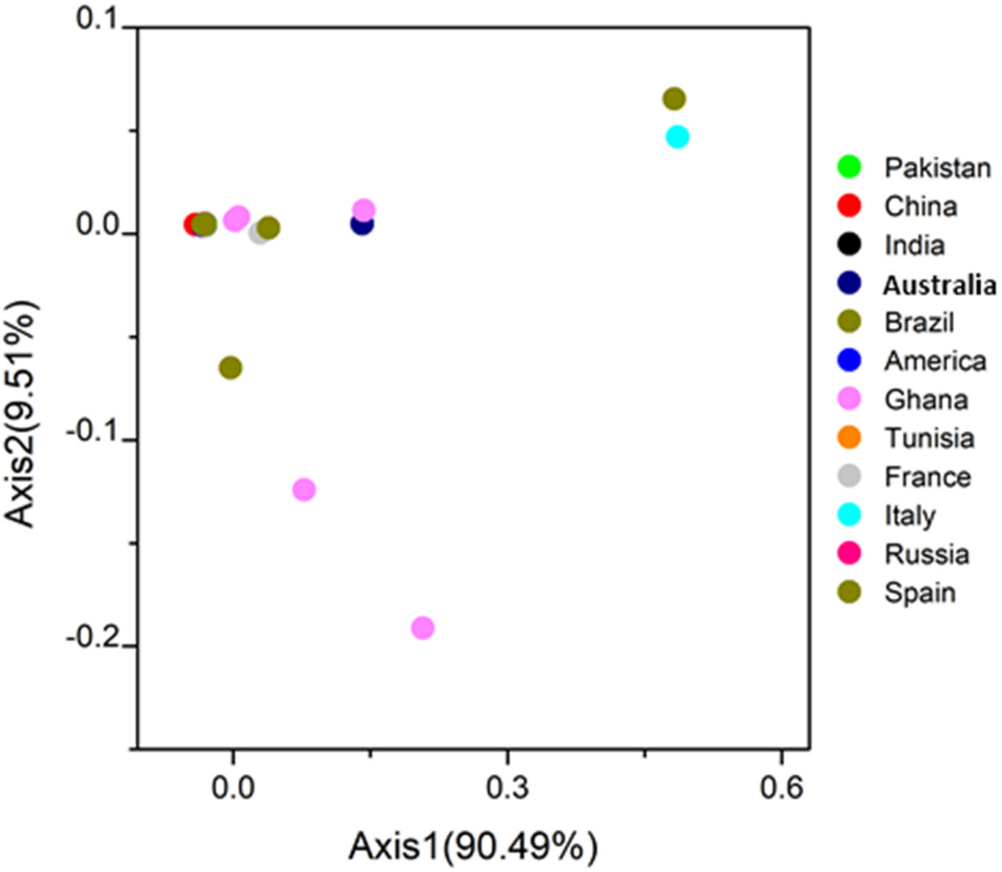

Figure 5. Variation analysis in SARS-CoV-2 genomes: based on the RSCU values, all the strains are plotted in the variance plane. The first and second principal axes represent different geographical origins. Each point represents a strain and shows in different colors.

3.7. Phylogenetic Analysis

Phylogenetic relationships of sixty SARS-CoV-2 CDS genomes were depicted from a Phylogenetic tree. Very high diversity was found in some strains like MT745629.1 and MT745630.1 of Australia, MT89210.1 and MT890211.1 of Ghana, and MT682732.1, MT622321.1 of Italy. Some strains from France (MT709104.1, MT709105.1), Russia (MT-890462.1, MT637143.1), and Spain (MT233522.1) though distinct but were found in the same clade. Apart from these, the remaining strains were divided into two major clades. A few strains of the same country were present in different clades, which depict that SARS-CoV-2 can arise from other viral lineages (Figure 6). Like the diversity found in Ghana's strains, the various evolution in different countries points towards the founder's effect in the SARS-CoV-2 genome. Moreover, it is also noticed that geographical locations may play a role in viral evolution.

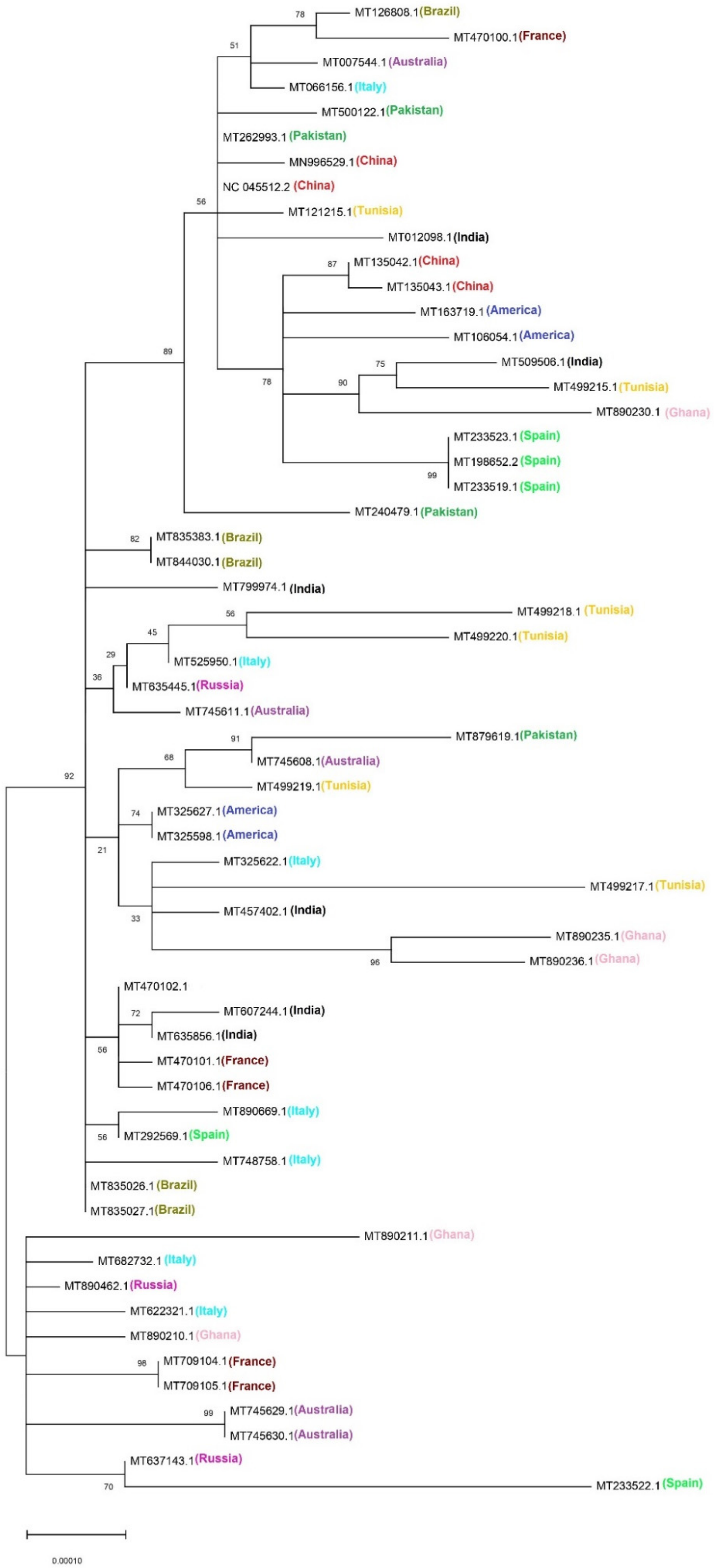

Figure 6. Phylogenetic tree based on the whole genome sequences of 60 SARS-Cov-2 strains. The tree was generated by the maximum likelihood (ML) method using the MUSCLE online tool.

4. Conclusions

Based on our findings, we conclude that the SARS-CoV-2 genome is shaped by a relatively low codon usage bias that is primarily motivated by mutational pressure but is also influenced by translation selection, which cannot be overlooked. Some countries' SARS-CoV-2 strains vary slightly due to different viral lineages. The better codon adaptation with humans can help explain the extensive spreadable nature of the virus. This research makes an essential contribution to the understanding of coronaviruses. Moreover, it is vital to conduct a large-scale comparative analysis of the codon patterns to reduce false positives.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biom11060912/s1. Table S1: ENC Values of sixty SARS-CoV-2 Strains around the globe with mean ENC value, Standard deviation and Maximum and minimum ENC value. Table S2: ENC and GC3 values of all the sixty strains of SARS-CoV-2 viruses from around the globe with their mean and standard deviation. Table S3: Neutrality plot (GC3 and GC12) value comparison of all the sixty strains of SARS-CoV-2. Table S4: CAI values of sixty strains of SARS-CoV-2 with host humans. Figure S1: Mean CAI value and Expected CAI value of SARS-CoV-2 strains against its host Humo sapiens.

Author Contributions: D.-D.W. and X.-Y.J. conceived the concept of the work and supervised the project. S.K., M.A.R., Q.Z., Y.A., S.F., P.M., H.A.K., A.A.K., E.E.N., and D.-D.W. reviewed the literatures, data curation, conceptualization, formal analysis, investigations, resources, software and extracted the data. T.L. assisted in writing the manuscript and performed data curation. S.K., M.A.R. and Y.A. drafted the manuscript. D.-D.W., and X.-Y.J. revised the manuscript and provided intellectual input in the paper. All authors read and approved the final version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 81802718, U1504817), the Foundation of Science & Technology Department of Henan Province, China (Nos. 192102310151, 202102310480), and the Training Program for Young Backbone Teachers of Institutions of Higher Learning in Henan Province, China (No. 2020GGJS038).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available with proper figures, tables, captions, legends in the manuscript, and supplementary files.

Acknowledgments: The authors thank Xin-Ying Ji and Dong-Dong Wu for their guidance in finalizing the manuscript. Besides, we have countless thanks to the all database of websites for their permission to use their data.

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

References

1. Paules, C.I.; Marston, H.D.; Fauci, A.S. Coronavirus Infections-More Than Just the Common Cold. *JAMA* **2020**. [CrossRef] [PubMed]
2. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; Holmes, E.C.; Garry, R.F. The proximal origin of SARS-CoV-2. *Nat. Med.* **2020**, *26*, 450–452. [CrossRef]
3. Pal, M.; Berhanu, G.; Desalegn, C.; Kandi, V. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. *Cureus* **2020**, *12*, e7423. [CrossRef] [PubMed]
4. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
5. Harapan, H.; Itoh, N.; Yufika, A.; Winardi, W.; Keam, S.; Te, H.; Megawati, D.; Hayati, Z.; Wagner, A.L.; Mudatsir, M. Coronavirus disease 2019 (COVID-19): A literature review. *J. Infect. Public Health* **2020**, *13*, 667–673. [CrossRef]
6. Ahmed, S.F.; Quadeer, A.A.; McKay, M.R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* **2020**, *12*, 254. [CrossRef]
7. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *81*, 104260. [CrossRef]
8. Ahmad, S.; Navid, A.; Farid, R.; Abbas, G.; Ahmad, F.; Zaman, N.; Parvaiz, N.; Azam, S.S. Design of a Novel Multi Epitope-Based Vaccine for Pandemic Coronavirus Disease (COVID-19) by Vaccinomics and Probable Prevention Strategy against Avenging Zoonotics. *Eur. J. Pharm. Sci.* **2020**, *151*, 105387. [CrossRef]

9. Rahman, M.S.; Hoque, M.N.; Islam, M.R.; Akter, S.; Rubayet Ul Alam, A.S.M.; Siddique, M.A.; Saha, O.; Rahaman, M.M.; Sultana, M.; Crandall, K.A.; et al. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2, the etiologic agent of COVID-19 pandemic: An in silico approach. *PeerJ* **2020**, *8*, e9572. [CrossRef]
10. Chiara, M.; Horner, D.S.; Gissi, C.; Pesole, G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-CoV-2. *bioRxiv* **2020**. [CrossRef]
11. Armijos-Jaramillo, V.; Yeager, J.; Muslin, C.; Perez-Castillo, Y. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. *bioRxiv* **2020**. [CrossRef]
12. Sardar, R.; Satish, D.; Birla, S.; Gupta, D. Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv* **2020**. [CrossRef]
13. Dilucca, M.; Forcelloni, S.; Georgakilas, A.G.; Giansanti, A.; Pavlopoulou, A. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* **2020**, *12*, 498. [CrossRef]
14. Shen, Z.; Xiao, Y.; Kang, L.; Ma, W.; Shi, L.; Zhang, L.; Zhou, Z.; Yang, J.; Zhong, J.; Yang, D.; et al. Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.* **2020**, *71*, 713–720. [CrossRef] [PubMed]
15. Angeletti, S.; Benvenuto, D.; Bianchi, M.; Giovanetti, M.; Pascarella, S.; Ciccozzi, M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. *J. Med. Virol.* **2020**, *92*, 584–588. [CrossRef]
16. Bal, A.; Destras, G.; Gaymard, A.; Bouscambert-Duchamp, M.; Valette, M.; Escuret, V.; Frobert, E.; Billaud, G.; Trouillet-Assant, S.; Cheynet, V.; et al. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). *Clin. Microbiol. Infect.* **2020**, *26*, 960–962. [CrossRef] [PubMed]
17. Holland, L.A.; Kaelin, E.A.; Maqsood, R.; Estifanos, B.; Wu, L.I.; Varsani, A.; Halden, R.U.; Hogue, B.G.; Scotch, M.; Lim, E.S. An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J. Virol.* **2020**, *94*, e00711–e00720. [CrossRef]
18. Nasrullah, I.; Butt, A.M.; Tahir, S.; Idrees, M.; Tong, Y. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.* **2015**, *15*, 174. [CrossRef]
19. Rahman, S.U.; Yao, X.; Li, X.; Chen, D.; Tao, S. Analysis of codon usage bias of Crimean-Congo hemorrhagic fever virus and its adaptation to hosts. *Infect. Genet. Evol.* **2018**, *58*, 1–16. [CrossRef]
20. Tort, F.L.; Castells, M.; Cristina, J. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. *Virus Res.* **2020**, *283*, 197976. [CrossRef] [PubMed]
21. Chen, H.; Sun, S.; Norenburg, J.L.; Sundberg, P. Mutation and Selection Cause Codon Usage and Bias in Mitochondrial Genomes of Ribbon Worms (Nemertea). *PLoS ONE* **2014**, *9*, e85631. [CrossRef] [PubMed]
22. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **1981**, *146*, 1–21. [CrossRef]
23. Gu, H.; Chu, D.K.W.; Peiris, M.; Poon, L.L.M. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evol.* **2020**, *6*. [CrossRef] [PubMed]
24. Butt, A.M.; Nasrullah, I.; Qamar, R.; Tong, Y. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg. Microbes Infect.* **2016**, *5*, e107. [CrossRef]
25. Jia, W.; Higgs, P.G. Codon Usage in Mitochondrial Genomes: Distinguishing Context-Dependent Mutation from Translational Selection. *Mol. Biol. Evol.* **2007**, *25*, 339–351. [CrossRef]
26. Cristina, J.; Moreno, P.; Moratorio, G.; Musto, H. Genome-wide analysis of codon usage bias in Ebolavirus. *Virus Res.* **2015**, *196*, 87–93. [CrossRef] [PubMed]
27. Sharp, P.M.; Bailes, E.; Grocock, R.J.; Peden, J.F.; Sockett, R.E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **2005**, *33*, 1141–1153. [CrossRef]
28. Sharp, P.M.; Emery, L.R.; Zeng, K. Forces that influence the evolution of codon bias. Sharp P M, Emery L R, Zeng K. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. B* **2010**, *365*, 1203–1212. [CrossRef]
29. Tyagi, N.; Sardar, R.; Gupta, D. Comparative analysis of codon usage patterns in SARS-CoV-2, its mutants and other respiratory viruses. *bioRxiv* **2021**. [CrossRef]
30. Hou, W. Characterization of codon usage pattern in SARS-CoV-2. *Virol. J.* **2020**, *17*, 138. [CrossRef] [PubMed]
31. Alonso, A.M.; Diambra, L. SARS-CoV-2 Codon Usage Bias Downregulates Host Expressed Genes With Similar Codon Usage. *Front. Cell Dev. Biol.* **2020**, *8*. [CrossRef]
32. Moratorio, G.; Iriarte, A.; Moreno, P.; Musto, H.; Cristina, J. A detailed comparative analysis on the overall codon usage patterns in West Nile virus. *Infect. Genet. Evol.* **2013**, *14*, 396–400. [CrossRef]
33. Shackelton, L.A.; Parrish, C.R.; Holmes, E.C. Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses. *J. Mol. Evol.* **2006**, *62*, 551–563. [CrossRef]
34. Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J.; McQuillan, J.A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **2012**, *28*, 464–469. [CrossRef]
35. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680. [CrossRef]

36. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
37. Sharp, P.M.; Li, W.H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **1986**, *24*, 28–38. [CrossRef] [PubMed]
38. Simmonds, P. SSE: A nucleotide and amino acid sequence analysis platform. *BMC Res. Notes* **2012**, *5*, 50. [CrossRef]
39. Xia, X. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **2013**, *30*, 1720–1728. [CrossRef]
40. Sharp, P.M.; Li, W.H. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **1987**, *15*, 1281–1295. [CrossRef]
41. Singh, N.K.; Tyagi, A.; Kaur, R.; Verma, R.; Gupta, P.K. Characterization of codon usage pattern and influencing factors in Japanese encephalitis virus. *Virus Res.* **2016**, *221*, 58–65. [CrossRef]
42. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2653. [CrossRef] [PubMed]
43. Neuhäuser, M. Wilcoxon–Mann–Whitney Test. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1656–1658.
44. Tao, P.; Dai, L.; Luo, M.; Tang, F.; Tien, P.; Pan, Z. Analysis of synonymous codon usage in classical swine fever virus. *Virus Genes* **2009**, *38*, 104–112. [CrossRef] [PubMed]
45. Wong, E.H.M.; Smith, D.K.; Rabadan, R.; Peiris, M.; Poon, L.L.M. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **2010**, *10*, 253. [CrossRef] [PubMed]
46. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef] [PubMed]
47. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef]
48. Jenkins, G.M.; Holmes, E.C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **2003**, *92*, 1–7. [CrossRef]
49. Berkhout, B.; van Hemert, F. On the biased nucleotide composition of the human coronavirus RNA genome. *Virus Res.* **2015**, *202*, 41–47. [CrossRef]
50. Kindler, E.; Thiel, V. To sense or not to sense viral RNA–essentials of coronavirus innate immune evasion. *Curr. Opin. Microbiol.* **2014**, *20*, 69–75. [CrossRef]
51. van Hemert, F.; van der Kuyl, A.C.; Berkhout, B. On the nucleotide composition and structure of retroviral RNA genomes. *Virus Res.* **2014**, *193*, 16–23. [CrossRef]
52. Yao, X.; Fan, Q.; Yao, B.; Lu, P.; Rahman, S.U.; Chen, D.; Tao, S. Codon Usage Bias Analysis of Bluetongue Virus Causing Livestock Infection. *Front. Microbiol.* **2020**, *11*. [CrossRef] [PubMed]
53. Dutta, R.; Buragohain, L.; Borah, P. Analysis of codon usage of severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) and its adaptability in dog. *Virus Res.* **2020**, *288*, 198113. [CrossRef] [PubMed]
54. 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. [CrossRef] [PubMed]
55. Castells, M.; Victoria, M.; Colina, R.; Musto, H.; Cristina, J. Genome-wide analysis of codon usage bias in Bovine Coronavirus. *Virol. J.* **2017**, *14*, 115. [CrossRef]
56. Bente, D.A.; Forrester, N.L.; Watts, D.M.; McAuley, A.J.; Whitehouse, C.A.; Bray, M. Crimean-Congo hemorrhagic fever: History, epidemiology, pathogenesis, clinical syndrome and genetic diversity. *Antivir. Res.* **2013**, *100*, 159–189. [CrossRef]