frontiers
in Microbiology

ORIGINAL RESEARCH
published: 06 April 2021
doi: 10.3389/fmicb.2021.548275

Base Composition and Host Adaptation of the SARS-CoV-2: Insight From the Codon Usage Perspective

Ayan Roy[1]†, Fucheng Guo[2,3]†, Bhupender Singh[1], Shelly Gupta[1], Karan Paul[4], Xiaoyuan Chen[2], Neeta Raj Sharma[1], Nishika Jaishee[5], David M. Irwin[6,7] and Yongyi Shen[2,3,8]*

[1] Department of Biotechnology, Lovely Professional University, Phagwara, India, [2] College of Veterinary Medicine, South China Agricultural University, Guangzhou, China, [3] Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou, China, [4] Department of Biochemistry, DAV University, Jalandhar, India, [5] Department of Botany, St Joseph's College, Darjeeling, India, [6] Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada, [7] Banting and Best Diabetes Centre, University of Toronto, Toronto, ON, Canada, [8] Key Laboratory of Zoonosis Prevention and Control of Guangdong Province, Guangzhou, China

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been spreading rapidly all over the world and has raised grave concern globally. The present research aims to conduct a robust base compositional analysis of SARS-CoV-2 to reveal adaptive intricacies to the human host. Multivariate statistical analysis revealed a complex interplay of various factors including compositional constraint, natural selection, length of viral coding sequences, hydropathicity, and aromaticity of the viral gene products that are operational to codon usage patterns, with compositional bias being the most crucial determinant. UpG and CpA dinucleotides were found to be highly preferred whereas, CpG dinucleotide was mostly avoided in SARS-CoV-2, a pattern consistent with the human host. Strict avoidance of the CpG dinucleotide might be attributed to a strategy for evading a human immune response. A lower degree of adaptation of SARS-CoV-2 to the human host, compared to Middle East respiratory syndrome (MERS) coronavirus and SARS-CoV, might be indicative of its milder clinical severity and progression contrasted to SARS and MERS. Similar patterns of enhanced adaptation between viral isolates from intermediate and human hosts, contrasted with those isolated from the natural bat reservoir, signifies an indispensable role of the intermediate host in transmission dynamics and spillover events of the virus to human populations. The information regarding avoided codon pairs in SARS-CoV-2, as conferred by the present analysis, promises to be useful for the design of vaccines employing codon pair deoptimization based synthetic attenuated virus engineering.

Keywords: SARS-CoV-2, codon usage, base composition, codon pair usage, codon adaptation index, host adaptation

OPEN ACCESS

Edited by:
Rosa Maria Pintó,
University of Barcelona, Spain

Reviewed by:
Luis Martinez-Sobrido,
Texas Biomedical Research Institute, United States
Francisco Javier Pérez Rodríguez,
Geneva University Hospitals (HUG), Switzerland

****Correspondence:***
Yongyi Shen
shenyy@scau.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *02 April 2020*
Accepted: *12 March 2021*
Published: *06 April 2021*

Citation:
Roy A, Guo F, Singh B, Gupta S, Paul K, Chen X, Sharma NR, Jaishee N, Irwin DM and Shen Y (2021) Base Composition and Host Adaptation of the SARS-CoV-2: Insight From the Codon Usage Perspective. Front. Microbiol. 12:548275. doi: 10.3389/fmicb.2021.548275

INTRODUCTION

The evolution of viruses is a conundrum for mankind. High mutation rates and host-shifts have culminated in multiple world-wide pandemics (Mackay and Arden, 2015; Luk et al., 2019). Recently, since December 2019, a viral outbreak by a novel coronavirus, named the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has emerged in Wuhan, China, and since then has rapidly spread throughout the world (Benvenuto et al., 2020a). In the recent past, two other coronavirus have resulted in major outbreaks, the severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV) in 2002, which resulted in 8,096 cases of infection and 774 deaths, and the Middle East respiratory syndrome (MERS) coronavirus (MERS-CoV) in 2012, which infected 2,494 individuals and claimed 858 lives (Lu et al., 2020). Alarmingly, COVID-19, the disease caused by SARS-CoV-2, was declared to be the first coronavirus-related pandemic by WHO. The SARS-CoV-2 outbreak has already surpassed both the SARS-CoV and MERS-CoV outbreaks in terms of numbers of infected individuals and numbers of deaths (Lu et al., 2020), although the overall case-fatality rate for SARS-CoV-2 appears to be lower than those for both SARS-CoV and MERS-CoV (Wu and McGoogan, 2020).

Severe acute respiratory syndrome coronavirus 2 is a positive-stranded RNA virus that belongs to the genus *Betacoronavirus* within the family Coronaviridae (Benvenuto et al., 2020a). This newly emerged virus appears to be considerably distant from SARS-CoV (with around 79% identity) and MERS-CoV (around 50% identity) (Lu et al., 2020). Though, bats are believed to be the primary reservoirs for these coronaviruses, intermediate hosts are suggested to be involved before their final spillover events into humans (Zhou et al., 2020). Civets and dromedary camels have been reported to be the intermediate hosts for SARS-CoV and MERS-CoV, respectively (Mackay and Arden, 2015; Luk et al., 2019). The Malayan pangolin (*Manis javanica*) has been assumed to be the probable intermediate host for SARS-CoV-2 (Lam et al., 2020; Xiao et al., 2020), although a firm and definite conclusion is yet to be reached (Li et al., 2020).

Host adaptation is an extremely important aspect dictating the survival and reproductive prowess of viral pathogens (Butt et al., 2016; Chen et al., 2017). Given the degeneracy of the genetic code, the preferential usage of specific synonymous codons (codons encoding the same amino acid) leads to a codon usage bias in genes and genomes (Grantham et al., 1980). Codon usage patterns in viral genomes have been reported to be due to the impact of mutational pressure and selection for host translational efficiency (i.e., matching host codon bias) (Butt et al., 2016; Chen et al., 2017). Viruses, owing to their small genomes, largely depend on the cellular machinery of their hosts for processes such as replication and protein synthesis (Butt et al., 2016). Investigations into viral codon usage patterns, relative to their hosts, have proved to be instrumental in elucidating riddles concerning viral adaptation and evasion of host immune responses (Butt et al., 2016). Furthermore, exploring the potential role of intermediate hosts in the transmission route of viral pathogens has unraveled intriguing facets of viral spillover events (Butt et al., 2016; Chen et al., 2017).

Since the emergence of SARS-CoV-2, extensive research has been undertaken to decipher its genomic features and riddles of its transmission, evolutionary dynamics, epidemiology, and mode of infection (Alonso and Diambra, 2020; Benvenuto et al., 2020b; Chen et al., 2020; Dilucca et al., 2020; Gu et al., 2020; Kames et al., 2020; Lam et al., 2020; Lu et al., 2020; Malik et al., 2020; Wang et al., 2020; Wu et al., 2020; Xiao et al., 2020; Zhou et al., 2020). However, the complexities of codon usage by SARS-CoV-2 and its impact on the adaptation and fitness of this virus to human hosts have not yet been addressed. Accordingly, the present research endeavor has targeted base composition analysis and investigation of factors influencing the complex codon usage profile of this newly emergent virus and subsequent adaptation to the human host. Information obtained from this research, in combination with existing knowledge, promises to deepen our understanding of the basic biology, pandemicity, pathogenesis, and host adaptation of SARS-CoV-2.

MATERIALS AND METHODS

Retrieval of Genomic Data

Genome sequences of 99 recently sequenced SARS-CoV-2, isolated from human (*Homo sapiens*) hosts, were retrieved from the GISAID repository[1] (**Supplementary Material 1**) (as per data available on February 24, 2020; time of this work). Full genome sequence-based alignments were generated by employing MAFFT software (version 7.4.2) (Katoh and Standley, 2013) followed by editing with MEGA X software (version 10) (Kumar et al., 2018) according to the reading frames encoded by the reference SARS-CoV-2 genome (GenBank: MN908947). Five closely related Pangolin-CoVs and five Bat-CoVs isolated from the pangolins and bats, respectively, were also retrieved from GISAID and processed accordingly (**Supplementary Material 1**). Annotated coding sequences of the *H. sapiens* genome (GRCh38.p13) were fetched from NCBI GenBank. A total of 78 complete SARS-CoV genomes, isolated from human, civet and bat hosts, and 491 MERS-CoV genomes, representing human, dromedary camel, and bat hosts, were downloaded from NCBI GenBank database (**Supplementary Material 1**). The Genomic tRNA Database (GtRNAdb) (Chan and Lowe, 2009) was used to retrieve information regarding the isoacceptor tRNAs in *H. sapiens*.

Estimation of Codon Usage Indices

Base compositional features of the viral coding sequences and estimates of codon usage, including relative synonymous codon usage (RSCU) and effective number of codons (ENC), were estimated employing CodonW (Ver. 1.4.2) software[2] (Peden, 2000). Correspondence analysis, based on the RSCU data of the viral coding sequences, was also generated using CodonW.

Neutrality Plot and Translational Selection Index

Neutrality plot analysis, an estimate of neutral evolution, was generated by plotting GC3 values (*x*-axis) of the viral genes against the respective GC12 values (*y*-axis) (Nasrullah et al., 2015; Butt et al., 2016).

Translational selection index (P2), an imperative estimate of translational selection, reflects the magnitude of

[1] http://www.GISAID.org

[2] http://www.molbiol.ox.ac.uk/cu

interaction between a codon and its respective anticodon (Gatherer and McEwan, 1997). P2 was calculated as:

$$P2 = \frac{WWC + SSU}{WWY + SSY}$$

where, W denotes the frequency of adenine (A) or uracil (U), S signifies the frequency of cytosine (C) or guanine (G), and Y reflects the frequency of cytosine (C) or uracil (U).

Codon Adaptation Index

Codon adaptation index (CAI) efficiently depicts the probable expression levels of genes of interest with respect to the codon usage pattern of a highly expressed reference gene set (Puigbò et al., 2008; Butt et al., 2016). CAI values lie between 0–1 and higher CAI values of viral genes, with respect to the host codon usage pattern, have been suggested to reflect higher levels of viral adaptation to the host environment (Puigbò et al., 2008). The CAIcal server[3] (Puigbò et al., 2008) was used to estimate CAI of the SARS-CoV-2 genes with respect to the *H. sapiens* host.

Relative Dinucleotide Abundance and Relative Synonymous Codon Pair Usage

Relative abundance (P_{xy}) of dinucleotides in SARS-CoV-2 and *H. sapiens* genes was computed as per the scheme suggested by Karlin and Burge (1995). The relative dinucleotide abundance was estimated as:

$$P_{xy} = \frac{f_{xy}}{f_x f_y}$$

where, f_{xy} and $f_x f_y$ refer to the observed and expected frequencies of the dinucleotide XY, respectively. Dinucleotides with $P_{xy} > 1.25$ were considered to be over-represented, whereas, dinucleotides with $P_{xy} < 0.78$ were inferred as under-represented (Kunec and Osterrieder, 2016).

The ratio of the observed to expected frequencies of a particular codon pair is commonly referred to as the relative synonymous codon pair usage (Kunec and Osterrieder, 2016). Codon pair scores were calculated as the natural logarithm of relative synonymous codon pair usage (Kunec and Osterrieder, 2016).

TABLE 1 | Relative synonymous codon usage (RSCU) patterns of SARS-CoV-2 in comparison with its host *Homo sapiens*.

Codon (Aa)	SARS-CoV-2	*Homo sapiens*	Codon (Aa)	SARS-CoV-2	*Homo sapiens*
UUU(Phe)	***1.40***	0.92	GCU(Ala)	**2.18**	1.08
UUC(Phe)	0.60	1.08	GCC(Ala)	0.58	1.60
UUA(Leu)	***1.63***	0.48	GCA(Ala)	**1.09**	0.92
UUG(Leu)	***1.06***	0.78	GCG(Ala)	0.15	0.44
CUU(Leu)	***1.75***	0.78	UAU(Tyr)	***1.22***	0.88
CUC(Leu)	0.59	1.20	UAC(Tyr)	0.78	1.12
CUA(Leu)	0.66	0.42	CAU(His)	***1.38***	0.84
CUG(Leu)	0.31	2.40	CAC(His)	0.62	1.16
AUU(Ile)	***1.53***	1.08	CAA(Gln)	***1.39***	0.54
AUC(Ile)	0.56	1.41	CAG(Gln)	0.61	1.46
AUA(Ile)	0.92	0.51	AAU(Asn)	***1.35***	0.94
GUU(Val)	***1.96***	0.72	AAC(Asn)	0.65	1.06
GUC(Val)	0.56	0.96	AAA(Lys)	***1.31***	0.86
GUA(Val)	0.91	0.48	AAG(Lys)	0.69	1.14
GUG(Val)	0.57	1.84	GAU(Asp)	***1.28***	0.92
UCU(Ser)	***1.96***	1.14	GAC(Asp)	0.72	1.08
UCC(Ser)	0.46	1.32	GAA(Glu)	***1.44***	0.84
UCA(Ser)	***1.67***	0.90	GAG(Glu)	0.56	1.16
UCG(Ser)	0.11	0.30	UGU(Cys)	***1.54***	0.92
AGU(Ser)	***1.43***	0.90	UGC(Cys)	0.46	1.08
AGC(Ser)	0.36	1.44	CGU(Arg)	**1.46**	0.48
CCU(Pro)	**1.94**	1.16	CGC(Arg)	0.58	1.08
CCC(Pro)	0.30	1.28	CGA(Arg)	0.29	0.66
CCA(Pro)	**1.59**	1.12	CGG(Arg)	0.19	1.20
CCG(Pro)	0.17	0.44	AGA(Arg)†	**2.67**	1.26
ACU(Thr)	***1.78***	1.00	AGG(Arg)	0.81	1.26
ACC(Thr)	0.38	1.44	GGU(Gly)	**2.33**	0.64
ACA(Thr)	***1.64***	1.12	GGC(Gly)	0.72	1.36
ACG(Thr)	0.20	0.44	GGA(Gly)	0.83	1.00
			GGG(Gly)	0.12	1.00

Preferentially employed codons (RSCU > 1.00) in SARS-CoV-2 are marked in bold; AU rich preferentially employed codons in SARS-CoV-2 are marked in italics; Most preferred codons in SARS-CoV-2 and H. sapiens are underlined; Most preferred codons in SARS-CoV-2 displaying antagonism with H. sapiens are marked in red color; Coincident codon of SARS-CoV-2 and H. sapiens are marked with †.
Aa, amino acid.

RESULTS

Base Composition of SARS-CoV-2

Extensive analysis of the nucleotide composition of the viral coding sequences revealed a distinct trend of AU richness among the SARS-CoV-2 genomes. The average AU and GC contents (%) were observed to be 62.56 ± 0.05 and 37.44 ± 0.05, respectively. The mean compositions (%) of the nucleotides A (28.15 ± 0.04) and U (34.41 ± 0.06) were found to be significantly higher than G (17.87 ± 0.02) and C (19.57 ± 0.03) ($P < 0.01$). An analysis of the RSCU revealed that a majority (20 out of 26 codons) of the preferentially employed codons (RSCU > 1) were AU rich (**Table 1**). The average composition (%) of AU3 (70.02 ± 0.05) was found to be significantly higher than GC3 (29.98 ± 0.05) ($P < 0.01$). It was also evident that 25 out of the 26 preferentially employed codons ended with A/U nucleotides (**Table 1**).

Estimates of the Codon Usage in SARS-CoV-2

The average ENC of the viral coding sequences was found to be 46.16 ± 5.93. An analysis of the GC3 versus ENC plot for SARS-CoV-2 revealed that while certain viral genes clustered on the continuous ENC plot curve (or close to it), others fell well below the curve (**Figure 1A**). It has been suggested that the continuous ENC plot curve indicates the expected codon usage of genes if GC compositional constraints alone account for the codon usage bias (Wright, 1990). It was interesting to note that a group of coding sequences in the GC3 versus ENC plot

[3] http://genomes.urv.cat/CAIcal/RCDI/

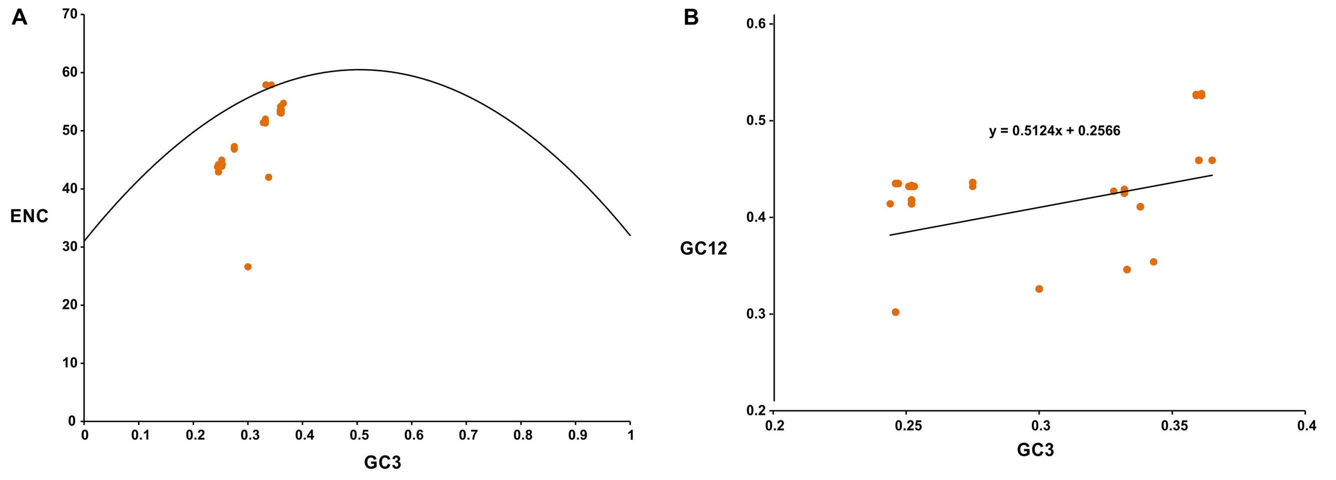

FIGURE 1 | (A) GC3-ENC plot for SARS-CoV-2. Viral genes analyzed are marked as orange colored circles. ENC plot curve is indicated by the bell-shaped solid line. **(B)** Neutrality plot for SARS-CoV-2. Viral genes analyzed are marked as orange colored circles.

for SARS-CoV-2 (**Figure 1A**) presented low ENC values (below 30) and behaved as outliers. After a thorough scrutiny it was observed that the group of coding sequences represented ORF7b of SARS-CoV-2 which was only 129 bp. It has been suggested that coding sequences less than 300 bp might behave as outliers due to small sizes in codon usage analysis (Xiang et al., 2015). A comprehensive analysis of the neutrality plot revealed that the slope of the regression line (**Figure 1B**) was around 0.5124, signifying a 51.24% influence of the compositional constraint on the viral coding sequences. The average translational selection index (P2) value of the SARS-CoV-2 coding sequences was found to be 0.43 ± 0.04, which suggested that apart from mutational bias, natural selection had a role in influencing the SARS-CoV-2 codon usage pattern.

Correspondence analysis, a multivariate statistical method, was performed on the RSCU data of SARS-CoV-2 to identify the determinants of codon usage variation employing CodonW. High significant correlations of GC and GC3 contents with Axes 1 and 2 of the RSCU data, the two major principle axes for the separation of genes, revealed a pronounced impact of compositional constraint on the SARS-CoV-2 genomes (**Table 2**). Strong correlation of the ENC with GC ($r = 0.49$, $P < 0.01$) and GC3 ($r = 0.48$, $P < 0.01$) contents reinforced the governing impact of compositional bias. Axes 1 and 2 of the RSCU data were found to correlate significantly with CAI of the SARS-CoV-2 genomes (**Table 2**), thus, depicting an undeniable influence of natural selection. Significant correlation of CAI with ENC ($r = 0.81$, $P < 0.01$) further signified the impact of natural selection in shaping the codon usage signatures of SARS-CoV-2. It is evident from **Table 2** that the length of the viral coding sequences correlated significantly with Axes 1 and 2 of the RSCU data. Furthermore, factors such as hydropathicity index {GRAVY [positive GRAVY (hydrophobic), negative GRAVY (hydrophilic)]} and aromaticity of the encoded viral gene products also correlated significantly with Axis 1 of the RSCU data (**Table 2**).

Relative Dinucleotide Abundance in SARS-CoV-2

Robust analysis of the relative dinucleotide abundance in SARS-CoV-2 revealed that UpG (1.38 ± 0.02) and CpA dinucleotides (1.27 ± 0.02) were over-represented (**Figure 2A**). Dinucleotide ApC (1.23 ± 0.02) and CpU (1.18 ± 0.03) were found to be marginally preferred in SARS-CoV-2 (**Figure 2A**). An analysis of RSCU revealed that the UpG containing codons, such as UUG and UGU, and CpA containing codons, like UCA, ACA, CCA, GCA, CAU, and CAA, were preferred (RSCU > 1.00) in SARS-CoV-2 (**Table 1**). These observations correlated well with the over-representation of these concerned dinucleotides. The CpG dinucleotide was observed to be highly under-represented (0.39 ± 0.01) in SARS-CoV-2, a pattern consistent with its host

TABLE 2 | Correlation analysis (Spearman's rank correlation) of various codon usage indices of SARS-CoV-2 with the principle axes of separation of the genes to Axes 1 and 2 of the RSCU data.

Codon usage indices	Axis1 (RSCU)	Axis 2 (RSCU)
A	0.14**	−0.41**
G	−0.32**	0.42**
C	−0.44**	0.36**
U	0.21**	0.12**
ENC	−0.59**	−0.33**
GC	−0.49**	0.42**
GC3	−0.58**	0.37**
Length	0.13**	0.11**
Gravy	−0.32**	0.05
Aromo	−0.16**	−0.08
CAI	0.74**	−0.60**

***Statistically significant at $P < 0.01$.*
Aromo, aromaticity of the viral gene product; ENC, effective number of codons; GRAVY, grand average hydropathicity score; Length, length of viral protein coding sequence; CAI, codon adaptation index.

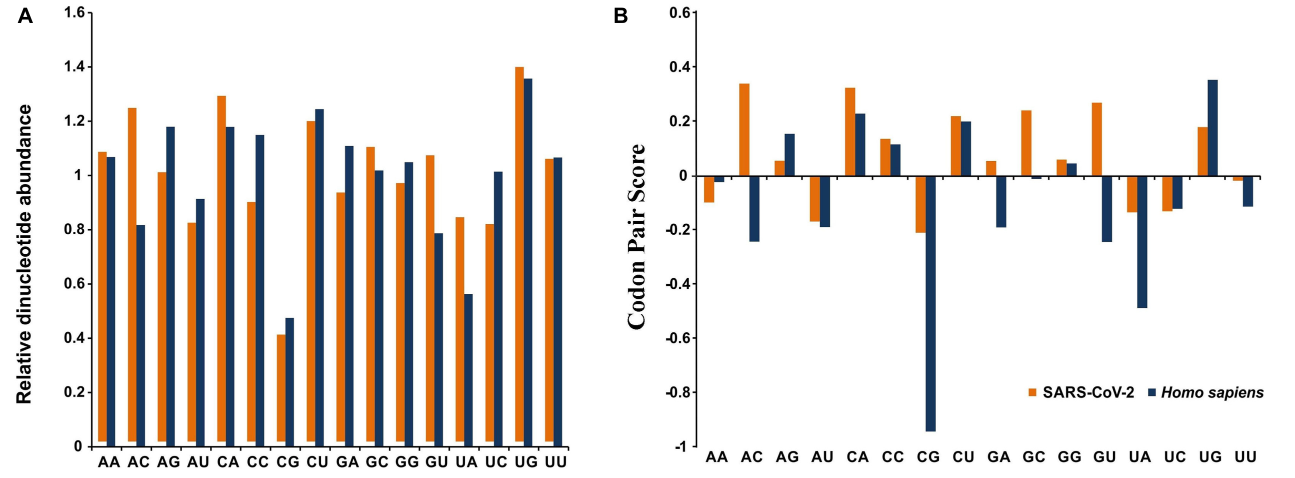

FIGURE 2 | (A) Relative dinucleotide abundance for SARS-CoV-2 in comparison with its host *Homo sapiens*. **(B)** Dinucleotide bias at the codon–codon interface for SARS-CoV-2 in comparison with its host *Homo sapiens*.

H. sapiens (**Figure 2A**). Codons, such as GCG, UCG, CGC, CGA, CGG, CCG, and ACG, containing the CpG dinucleotide were noted to be under-represented (RSCU < 0.60) (**Table 1**). The dinucleotides UpA (0.82 ± 0.03), ApU (0.81 ± 0.02), and UpC (0.82 ± 0.01) were found to be marginally avoided in SARS-CoV-2 (**Figure 2A**).

Patterns of Codon Pair Usage in SARS-CoV-2

An extensive analysis of the relative synonymous codon pair usage, of the 3,721 (61 × 61) codon pairs (excluding stop:stop and stop:sense codon pairs), revealed that 1,530 codon pairs were over-represented (positive codon pair score), whereas, 2,187 were under-represented (negative codon pair score) in SARS-CoV-2 (**Supplementary Material 2**). The codon pair CGG-CGG, coding for the amino acid pair Arg–Arg, was noted to be most over-represented with a codon pair score of 4.27. On the other hand, the codon pair AAU-AUA, encoding the amino acid pair Asn-Ile, displayed the lowest codon pair score of −4.56 and was inferred to be most avoided.

Dinucleotide patterns NNU-GNN, NNC-ANN, NNC-UNN, and NNA-CNN, representing the dinucleotides UpG, CpA, CpU, and ApC, respectively, were predominant at codon-codon junctions in SARS-CoV-2 (**Figure 2B**). The NNU-GNN (representing the UpG dinucleotide at the codon pair junction) was found to be most prevalent (11.24%) among the over-represented codon pairs.

Dinucleotide patterns NNC-GNN, NNU-ANN, and NNA-UNN, representing the dinucleotides CpG, UpA, and ApU, respectively, were noted to be highly avoided at codon pair junctions in SARS-CoV-2 (**Figure 2B**). The NNC-GNN pattern was noted to be most abundant (10.06%) among the under-represented codon pairs, suggesting strong suppression of CpG dinucleotides at the codon-codon interface of the SARS-CoV-2 genomes.

Investigating the Patterns of SARS-CoV-2 Adaptation to Human Hosts

Antagonistic Codon Usage Patterns of SARS-CoV-2 Toward Human Host

Detailed RSCU analysis and profiling the most preferred codons (for each amino acid) in SARS-CoV-2 and *H. sapiens* revealed a distinct trend of antagonism between the viral and human codon usage patterns (**Table 1**). Seventeen out of the eighteen most preferred codons in SARS-CoV-2 were found to exhibit antagonism with *H. sapiens*, whereas the codon AGA (coding for Arg), was the only one to display coincidence (**Table 1**).

Most Preferred Codons in SARS-CoV-2 and Human Isoacceptor tRNAs

Identification of the most preferred codons (for each amino acid) in SARS-CoV-2 and the most abundant isoacceptor tRNAs in human cells revealed that 6 out of the 18 most preferred codons in SARS-CoV-2, namely, GCU, CCU, ACU, UCU, CUU, and AUU (coding for the amino acids Ala, Pro, Thr, Ser, Leu, and Ile, respectively), optimally matched with the respective most abundant isoacceptor tRNAs in human hosts (**Table 3**).

Adaptive Efficacy of SARS-CoV-2 in Human Host in Comparison With SARS-CoV and MERS-CoV

The magnitude of adaptive efficacy and associated fitness of the recently emerged SARS-CoV-2 to the human niche was explored in light of the adaptation exhibited by other relevant and notorious coronaviruses associated with severe pneumonia and past outbreaks, namely SARS-CoV and MERS-CoV. A correspondence analysis based on the RSCU data of the concerned viral genomes (SARS-CoV-2, SARS-CoV, and MERS-CoV) and the human genome led to the formation of four discrete clusters, with the viral clusters falling close to

TABLE 3 | The most preferred codon, for each amino acid, in SARS-CoV-2 and iso-acceptor tRNAs in *Homo sapiens*.

Amino acids	Most preferred codons in SARS-CoV-2	tRNA isotypes in *Homo sapiens*
Ala	GCU	**AGC (22)**, GGC (0), CGC (4), UGC (8)
Gly	GGU	ACC (0), GCC (14), CCC (5), UCC (9)
Pro	CCU	**AGG (9)**, GGG (0), CGG (4), UGG (7)
Thr	ACU	**AGU (9)**, GGU (0), CGU (5), UGU (6)
Val	GUU	AAC (9), GAC (0), CAC (11), UAC (5)
Ser	UCU	**AGA (9)**, GGA (0), CGA (4), UGA (4), ACU (0), GCU (8)
Arg	AGA	ACG (7), GCG (0), CCG (4), UCG (6), CCU (5), UCU (6)
Leu	CUU	**AAG (9)**, GAG (0), CAG (9), UAG (3), CAA (6), UAA (4)
Phe	UUU	AAA (0), GAA (10)
Asn	AAU	AUU (0), GUU (20)
Lys	AAA	CUU (15), UUU (12)
Asp	GAU	AUC (0), GUC (13)
Glu	GAA	CUC (8), UUC (7)
His	CAU	AUG (0), GUG (10)
Gln	CAA	CUG (13), UUG (6)
Ile	AUU	**AAU (14)**, GAU (3), UAU (5)
Tyr	UAU	AUA (0), GUA (13)
Cys	UGU	ACA (0), GCA (29)

Most abundant iso-acceptor tRNAs in Homo sapiens matching the most preferred codons of SARS-CoV-2 are marked in bold.

each other (**Figure 3A**). It was intriguing to note that the cluster formed by MERS-CoV was closest to that formed by the human genome (spatial Mahalanobis distance of 169.41), followed by the SARS-CoV cluster (spatial Mahalanobis distance of 206.71) (**Figure 3A**). Interestingly, the cluster formed by SARS-CoV-2 was observed to be most distant from the human genome cluster (spatial Mahalanobis distance of 1618.40) (**Figure 3A**). Axis 1 of the RSCU data (contributing 79.41% of the variation) was found to display a strong correlation with CAI values of the respective viral genomes (r = 0.87, $P < 0.01$). CAI of the viral genes, calculated with respect to host codon usage patterns, provides insights into the degree of viral adaptation to the host cellular environment. The average CAI value for SARS-CoV-2 with respect to *H. sapiens* was found to be 0.701 ± 0.04, which was significantly lower ($P < 0.01$) than for both MERS-CoV (0.718 ± 0.05) and SARS-CoV (0.715 ± 0.04) (**Figure 3B**). Our observations indicate a relatively lower adaption of the newly emerged SARS-CoV-2, in contrast to SARS-CoV and MERS-CoV, to human cellular systems.

Role of the Intermediate Host in the SARS-CoV-2 Spillover Event to the Human Population

In order to investigate the potential role of the intermediate host in the SARS-CoV-2 transmission event, we compared the CAI values of human isolated SARS-Cov-2, closely related Pangolin-CoVs and Bat-CoVs, employing human codon usage patterns as a reference for calculation. Our results displayed similar patterns of adaptation to the human cellular system for SARS-Cov-2 and closely related Pangolin-CoVs ($P < 0.01$) (**Figure 4**). However, the CAI values of the closely related Bat-CoVs were noted to be significantly lower ($P < 0.01$) (**Figure 4**) in comparison to SARS-Cov-2 and Pangolin-CoVs, when assessed in reference to human codon usage patterns. We further analyzed the patterns of adaptation of SARS-CoV and MERS-CoV genomes isolated from bats (primary reservoir for both of the concerned coronaviruses), intermediate hosts (civets in the case of SARS-CoV and dromedary camels in the case of MERS-CoV) and the terminal hosts human beings. Our results revealed that the CAI values of SARS-CoV and MERS-CoV (estimated employing human codon usage patterns) isolated from bats were significantly lower than the CAI values of the respective viral isolates representing the intermediate and human hosts ($P < 0.01$) (**Figure 4**). Thus, the trend of similar adaptation patterns for SARS-Cov-2, SARS-CoV, and MERS-CoV isolates from intermediate and human hosts was found to be consistent.

DISCUSSION

The recent emergence of the novel SARS-CoV-2 has posed a serious threat to global public health (Benvenuto et al., 2020a). The rapid spread and transmission of this virus demands an investigation into its adaptive fitness and patterns of acclimatization into the newly introduced human host. The present research effort was undertaken to comprehensively explore the complex codon usage profile of SARS-CoV-2, relative to the human host, and confers meaningful knowledge pertaining to viral adaptation to the human niche.

Our analysis revealed that the coding sequences of SARS-CoV-2 had a high average of ENC (ENC = 46.16 ± 5.93), which signified a pattern of low codon usage bias (**Figure 1A**). It has been suggested that a virus with low codon usage bias might be more flexible and able to adapt and maintain its survival cycle in a broad range of hosts with different codon usage signatures (Jenkins and Holmes, 2003; Luo et al., 2020).

Thorough our analysis of a GC3-ENC plot of SARS-CoV-2 we revealed that while certain viral genes fell on or close to the continuous ENC plot curve, others clustered well below the curve (**Figure 1A**). It has been suggested that if codon usage of a gene is governed only by compositional bias, then it would lie on or above the continuous ENC plot curve, whereas, the clustering of genes well below the curve signifies the impact of other factors such as natural selection, in addition to compositional constraint (Wright, 1990). Thus, it was evident that apart from compositional constraint others factors such as translational selection also influenced the codon usage behavior of SARS-CoV-2. A neutrality plot (**Figure 1B**) revealed that the degree of compositional constraint influence on the coding sequences was around 51.24%, which further supports our conclusion (Nasrullah et al., 2015; Butt et al., 2016).

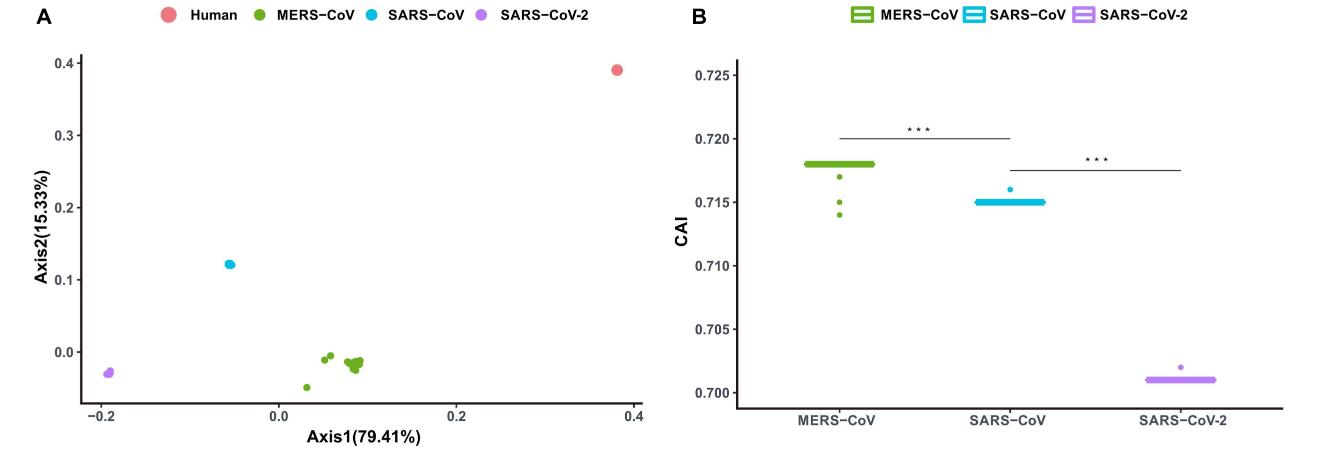

FIGURE 3 | (A) Correspondence analysis depicting Axis 1 and Axis 2 of the RSCU data for MERS-CoV, SARS-CoV, SARS-CoV-2, and their human host. Dots representing MERS-CoV, SARS-CoV, and SARS-CoV-2 are marked in green, blue, and purple, respectively. Dots signifying human coding sequences are marked in red. **(B)** Codon adaptation index (CAI) values for MERS-CoV, SARS-CoV, and SARS-CoV-2. Mann–Whitney U Rank Sum Test was used to compare the average of the CAI values pertaining to the different sets of viruses. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

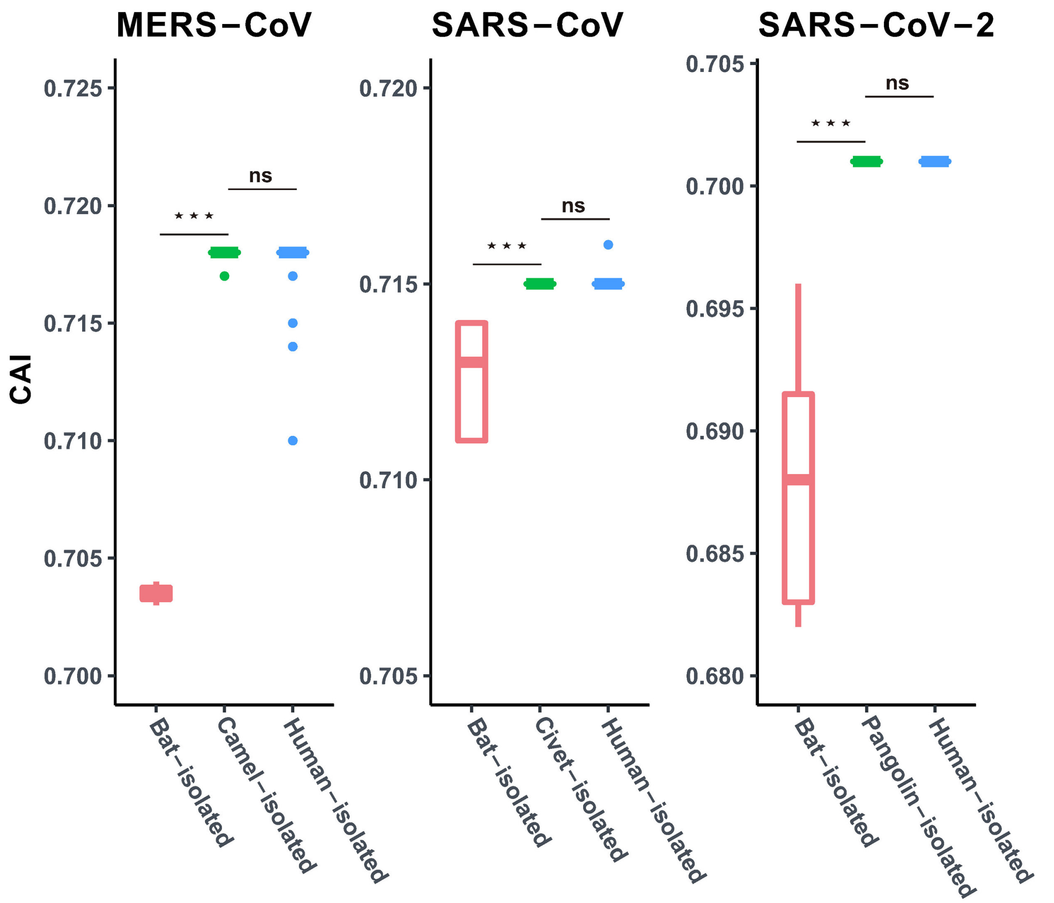

FIGURE 4 | Codon adaptation index (CAI) values for MERS-CoV, SARS-CoV, and SARS-CoV-2, isolated from different hosts, calculated with the human coding sequences as reference. Red, green, and blue boxes represent viral isolates from the bat, corresponding intermediate host (dromedary camel for MERS-CoV, civet for SARS-CoV, and pangolin for SARS-CoV-2) and human host, respectively. Mann–Whitney U Rank Sum Test was employed to compare the mean of the CAI values pertaining to the different sets of viruses. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. ns, not significant.

Strong correlations of GC and GC3 contents of the viral coding sequences with the two major axes of separation of the RSCU data reinforce the pronounced impact of compositional bias (**Table 2**). It has been suggested that the translational selection index (P2) > 0.50 signifies a major role of translational selection acting on the concerned genes (Gatherer and McEwan, 1997). An average P2 value of 0.43 ± 0.04 for the SARS-CoV-2 coding sequences signified that, apart from compositional constraint, translational selection had a considerable influence on the viral codon usage patterns. Apart from compositional bias and translational selection, factors including length of the viral protein coding sequence, hydropathicity index (GRAVY) and aromaticity of the viral protein products also display significant correlations with the RSCU data (**Table 2**). Thus, codon usage signatures of the SARS-CoV-2 appeared to be result of various imperative factors including compositional bias, natural selection, hydropathicity and aromatic character of the viral gene products, and the lengths of the viral coding sequences, with compositional bias displaying the most crucial impact. Similar cases inferring compositional bias as the major contributor of codon usage variations have been previously reported for SARS-CoV (Gu et al., 2004) and MERS-CoV (Chen et al., 2017).

Severe acute respiratory syndrome coronavirus 2 has an AU rich genome with a distinct preference toward the usage of AU rich codons over their GC rich counterparts (**Table 1**). Similar trends have been observed in the closely related SARS-CoV (Gu et al., 2004) and MERS-CoV (Chen et al., 2017) genomes. SARS-CoV-2 was found to strongly abstain from using the dinucleotides CpG and UpA, a pattern consistent with the host *H. sapiens* (**Figure 2A**). Similar patterns of CpG and UpA dinucleotide avoidance was evident at the codon-codon interface of the viral genomes (**Figure 2B**). CpG and UpA under-representation is a trademark of vertebrate genomes (Kunec and Osterrieder, 2016). Avoidance of the CpG dinucleotide is an established feature of RNA viruses (Kunec and Osterrieder, 2016). It has been suggested that the unmethylated CpGs of viral pathogens are recognized by the host intracellular pattern recognition receptor Toll like receptor 9 (TLR9) to stimulate an immune response against the pathogens (Dorn and Kippenberger, 2008; Kunec and Osterrieder, 2016). Thus, suppression of CpG dinucleotides in SARS-CoV-2 appears to be a strategy to evade human immune response. Under-representation of UpA dinucleotides in SARS-CoV-2 might be a reflection of its effort to hone translational efficacy by reducing the risk of nonsense mutations, mRNA degradation, and error-prone translation associated with UpA abundance (Karlin and Burge, 1995). Interestingly, a significant share of the over-represented (52.22%) and under-represented codon pairs (56.29%) in SARS-CoV-2 matched with host *H. sapiens*, which indicates adaptation toward enhanced robustness of SARS-CoV-2 to the human cellular environment.

The rapid spread of COVID-19 necessitates an urgent need of safe and effective vaccines to combat the pandemic. Analysis of codon usage and information regarding under-represented codons and codon pairs in viral genomes offer scopes toward the development of live-attenuated vaccines employing the synthetic attenuated virus engineering approach (Coleman et al., 2008). Synthetic attenuated virus engineering involves recoding and synthesis of a viral genome in a way that preserves the wild-type amino acid sequence but rearranges existing synonymous codons to create a sub-optimal arrangement of codon pairs that are typically under-represented (Coleman et al., 2008). The identification and profiling of under-represented codons and codon pairs containing the CpG and UpA dinucleotides (**Table 1** and **Figure 2**) in SARS-CoV-2 genomes through extensive codon usage analysis, as executed in the present study, promises to be useful in guiding deoptimization of codons and codon pairs for viral attenuation and vaccine development. Such applications have been successfully implemented in the development of live-attenuated vaccines against poliovirus (Coleman et al., 2008), human respiratory syncytial virus (Le Nouën et al., 2014), influenza virus (Mueller et al., 2010), dengue virus (Shen et al., 2015), Lassa virus (Cai et al., 2020), and enterovirus A71 (Tsai et al., 2019).

Severe acute respiratory syndrome coronavirus 2 was found to display antagonistic codon usage patterns with the human host (**Table 1**). Similar trends of antagonism have previously been seen in the Marburg virus (Nasrullah et al., 2015) and the hepatitis A virus (Sanchez et al., 2003) with the human host. In contrast, poliovirus was found to display a complete coincidence (Mueller et al., 2006) and Zika virus (Butt et al., 2016) has been reported to exhibit a mixture of antagonism and coincidence with the codon usage patterns of the human genome. It has been suggested that coincident patterns of codon usage between a virus and its host facilitates translational efficiency, whereas, antagonism facilitates proper folding of the viral proteins, although the efficacy of translation might be reduced (Hu et al., 2011).

Most of the preferred codons in SARS-CoV-2 coding sequences use suboptimal isoacceptor tRNAs from human cells (**Table 3**). A similar pattern of suboptimal tRNA isotype recognition use has been previously reported for the Nipah virus (Khandia et al., 2019). It has been suggested that the usage of suboptimal isoacceptor host tRNAs during the initial phase of an infection might facilitate slow but precise translation, which yields the synthesis of accurate and properly folded viral proteins (Khandia et al., 2019).

The CAI of viral genes, estimated with respect to host codon usage patterns, has been proposed to be an effective index of the degree of viral adaptation to a host's cellular environment (Puigbò et al., 2008). Lower CAI value for SARS-CoV-2, in comparison to MERS-CoV and SARS-CoV (**Figure 3B**), estimated with respect to human codon usage patterns, signifies moderately adapted potential and fitness of this recently emerged pathogen to human cellular system and seems to be in agreement with its long incubation time (between 1 and 14 days), milder infective consequences and relatively lower case-fatality rates of between 3 and 4% (as reported by WHO), in contrast to the enhanced infective manifestations and higher case-fatality rates of 34.40 and 9.56% reported for MERS-CoV and SARS-CoV (Wu and McGoogan, 2020).

Considering the fact that this virus has already spread globally, and that human populations are highly susceptible, there exists a possibility that SARS-CoV-2 might enhance its adaptive finesse to human cells through ongoing processes of adaptive evolution, leading to further risks for transmission and imminent outbreaks.

Viral transmission across different hosts and associated cross-species jumps are puzzling events in the intricate transmission dynamics of these viruses (Smith et al., 2009). Intermediate hosts are believed to play a crucial role in the viral spillover events to human populations. There have been many instances of the involvement of intermediate hosts in viral pandemics and epidemics (Smith et al., 2009; Parrish et al., 2015). It has been suggested that the involvement of an intermediate host, or a series of hosts, facilitates viral transmission from its natural reservoirs and acts as a platform to hone adaptive finesse of the viral pathogens before the final spillover to the human population (Parrish et al., 2015). In this context, it is interesting to note that the viral isolates from the pangolins (Pangolin-CoVs) display highly similar patterns of adaptation with SARS-CoV-2, in the human cellular niche (as evident from similar CAI values) (**Figure 4**). Similar trends were observed for other coronaviruses such as SARS-CoV and MERS-CoV that have infected human populations, where the CAI values of the viral isolates from their respective intermediate hosts matched those isolated from humans, thus signifying similar adaptive patterns (**Figure 4**). In contrast, the respective viral isolates from primary bat reservoirs displayed significantly weaker adaptation (as evident from lower CAI values) to human host systems (**Figure 4**). These results suggest that pangolins have the potential to act as the intermediate host for SARS-CoV-2 and that the Pangolin-CoVs represents a potential future threat to public health (Lam et al., 2020; Xiao et al., 2020). However, more systematic research efforts and exhaustive long-term monitoring of SARS-related coronaviruses in pangolins and other related animals would be necessary to draw a final inference about the mysterious intermediate host of SARS-CoV-2.

Multiple studies have been targeted at the investigation of codon usage patterns of SARS-CoV-2 till date (Alonso and Diambra, 2020; Dilucca et al., 2020; Gu et al., 2020; Kanduc, 2020; Malik et al., 2020; Saha et al., 2021). Recently, Tort et al. (2020) reported a close relationship of SARS-CoV-2 with Bat-CoVs in comparison to SARS-CoV, MERS-CoV, and other coronaviruses isolated from civets and ferrets, based on a comparative codon usage analysis. The study indicates toward the involvement of bats as probable primary reservoirs for SARS-CoV-2 (Tort et al., 2020). However, in the present analysis based on codon usage, apart from bats we have also explored the potential association of pangolins in the transmission route of SARS-CoV-2 to humans. The present research pertaining to the codon usage patterns of SARS-CoV-2, together with its transmission dynamics involving intermediate hosts, and facets of its adaptation to its newly introduced human population promises to significantly contribute toward the elucidation of its infective manifestations and rapid global spread, thus bolstering current efforts on pandemic preparedness. Extensive knowledge pertaining to codon usage and the profiling of preferred and avoided codons and codon pairs in the viral genomes have been effective in synthetic attenuated virus engineering toward the development of live-attenuated vaccines (Burns et al., 2006; Coleman et al., 2008). Deoptimized viral genes and genomes, with under-represented codon pairs, have been reported to exhibit decreased replicative fitness and low protein expression levels without major alterations that evoke host immune responses (Burns et al., 2006; Coleman et al., 2008). The identification and profiling of under-represented codon pairs in SARS-CoV-2, as executed in the present analysis, might prove beneficial toward the rational development of safe attenuated vaccines and combat the COVID-19 pandemic.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YS conceived, designed, and supervised the study. AR, FG, BS, SG, KP, XC, NS, and NJ generated the data. YS, AR, and FG analyzed the data. YS, DI, and AR wrote and prepared the manuscript. All authors have read and agreed to submission of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant number 31822056), Department of Education of Guangdong Province (grant numbers 2019KZDXM004 and 2019KCXTD001), the Guangdong Science and Technology Innovation Leading Talent Program (grant number 2019TX05N098), the 111 Project (grant number D20008), Ministry of Science and Technology of China, Chinese Academy of Engineering, Chinese Academy of Sciences, Department of Science and Technology of Guangdong Province, and Department of Agriculture of Guangdong Province.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.548275/full#supplementary-material

Supplementary Material 1 | Information pertaining to the accession numbers and hosts (isolation source) of the SARS-CoV-2, SARS-CoV, and MERS-CoV genomes used in the present study.

Supplementary Material 2 | Codon pair scores of the 3,721 (61 × 61) codon pairs (excluding stop:stop and stop:sense codon pairs) that were analyzed for SARS-CoV-2.

REFERENCES

Alonso, A. M., and Diambra, L. (2020). SARS-CoV-2 codon usage bias downregulates host expressed genes with similar codon usage. *Front. Cell Dev. Biol.* 8:831. doi: 10.3389/fcell.2020.00831

Benvenuto, D., Giovanetti, M., Ciccozzi, A., Spoto, S., Angeletti, S., and Ciccozzi, M. (2020a). The 2019-new coronavirus epidemic: evidence for virus evolution. *J. Med. Virol.* 92, 455–459. doi: 10.1002/jmv.25688

Benvenuto, D., Giovanetti, M., Salemi, M., Prosperi, M., Flora, C. D., Alcantara, L. C. J., et al. (2020b). The global spread of 2019-nCoV: a molecular evolutionary analysis. *Pathog. Glob. Health* 114, 64–67. doi: 10.1080/20477724.2020.1725339

Burns, C. C., Shaw, J., Campagnoli, R., Jorba, J., Vincent, A., Quay, J., et al. (2006). Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* 80, 3259–3272. doi: 10.1128/JVI.80.7.3259-3272.2006

Butt, A. M., Nasrullah, I., Qamar, R., and Tong, Y. (2016). Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg. Microb. Infect.* 5:e107. doi: 10.1038/emi.2016.106

Cai, Y., Ye, C., Cheng, B., Nogales, A., Iwasaki, M., Yu, S., et al. (2020). A Lassa fever live-attenuated vaccine based on codon deoptimization of the viral glycoprotein gene. *mBio* 11:e0039-20. doi: 10.1128/mBio.00039-20

Chan, P. P., and Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37, D93–D97. doi: 10.1093/nar/gkn787

Chen, Y., Xu, Q., Yuan, X., Li, X., Zhu, T., Ma, Y., et al. (2017). Analysis of the codon usage pattern in middle east respiratory syndrome Coronavirus. *Oncotarget* 8, 110337–110349. doi: 10.18632/oncotarget.22738

Chen, Z., Boon, S. S., Wang, M. H., Chan, R. W., and Chan, P. K. (2020). Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses. *J. Virol. Methods* 289:114032. doi: 10.1016/j.jviromet.2020.114032

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science* 320, 1784–1787. doi: 10.1126/science.1155761

Dilucca, M., Forcelloni, S., Georgakilas, A. G., Giansanti, A., and Pavlopoulou, A. (2020). Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses* 12:498. doi: 10.3390/v12050498

Dorn, A., and Kippenberger, S. (2008). Clinical application of CpG-, non- CpG-, and antisense oligodeoxynucleotides as immunomodulators. *Curr. Opin. Mol. Ther.* 10, 10–20.

Gatherer, D., and McEwan, N. R. (1997). Small regions of preferential codon usage and their effect on overall codon bias–the case of the plp gene. *Biochem. Mol. Biol. Int.* 43, 107–114. doi: 10.1080/15216549700203871

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 8, r49–r62. doi: 10.1093/nar/8.1.197-c

Gu, H., Chu, D. K., Peiris, M., and Poon, L. L. (2020). Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evol.* 6:veaa032. doi: 10.1093/ve/veaa032

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

Hu, J. S., Wang, Q. Q., Zhang, J., Chen, H. T., Xu, Z. W., Zhu, L., et al. (2011). The characteristic of codon usage pattern and its evolution of hepatitis C virus. *Infect. Genet. Evol.* 11, 2098–2102. doi: 10.1016/j.meegid.2011.08.025

Jenkins, G. M., and Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* 92, 1–7. doi: 10.1016/S0168-1702(02)00309-X

Kames, J., Holcomb, D. D., Kimchi, O., DiCuccio, M., Hamasaki-Katagiri, N., Wang, T., et al. (2020). Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *Sci. Rep.* 10:15643. doi: 10.1038/s41598-020-72533-2

Kanduc, D. (2020). Severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2): codon usage and replicative fitness. *Glob. Med. Genet.* 7, 92–94. doi: 10.1055/s-0040-1721080

Karlin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11, 283–290. doi: 10.1016/S0168-9525(00)89076-9

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., et al. (2019). Analysis of nipah virus codon usage and adaptation to hosts. *Front. Microbiol.* 10:886. doi: 10.3389/fmicb.2019.00886

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kunec, D., and Osterrieder, N. (2016). Codon pair bias is a direct consequence of Dinucleotide bias. *Cell Rep.* 14, 55–67. doi: 10.1016/j.celrep.2015.12.011

Lam, T. T., Jia, N., Zhang, Y. W., Shum, M. H., Jiang, J. F., Zhu, H. C., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* 583, 282–285. doi: 10.1038/s41586-020-2169-0

Le Nouën, C., Brock, L. G., Luongo, C., McCarty, T., Yang, L., Mehedi, M., et al. (2014). Attenuation of human respiratory syncytial virus by genome-scale codon-pair deoptimization. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13169–13174. doi: 10.1073/pnas.1411290111

Li, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B. T., et al. (2020). Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J. Med. Virol.* 92, 602–611. doi: 10.1002/jmv.25731

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

Luk, H. K. H., Li, X., Fung, J., Lau, S. K. P., and Woo, P. C. Y. (2019). Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.* 71, 21–30. doi: 10.1016/j.meegid.2019.03.001

Luo, W., Roy, A., Guo, F., Irwin, D. M., Shen, X., Pan, J., et al. (2020). Host adaptation and evolutionary analysis of Zaire ebolavirus: insights from codon usage based investigations. *Front. Microbiol.* 11:570131. doi: 10.3389/fmicb.2020.570131

Mackay, I. M., and Arden, K. E. (2015). MERS coronavirus: diagnostics, epidemiology and transmission. *Virol. J.* 12:222. doi: 10.1186/s12985-015-0439-5

Malik, Y. S., Ansari, M. I., Kattoor, J. J., Kaushik, R., Sircar, S., Subbaiyan, A., et al. (2020). Evolutionary and codon usage preference insights into spike glycoprotein of SARS-CoV-2. *Brief. Bioinform.* 2020:bbaa383. doi: 10.1093/bib/bbaa383

Mueller, S., Coleman, J. R., Papamichail, D., Ward, C. B., Nimnual, A., Futcher, B., et al. (2010). Live attenuated influenza virus vaccines by computer-aided rational design. *Nat. Biotechnol.* 28, 723–726. doi: 10.1038/nbt.1636

Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S., and Wimmer, E. (2006). Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* 80, 9687–9696. doi: 10.1128/jvi.00738-06

Nasrullah, I., Butt, A. M., Tahir, S., Idrees, M., and Tong, Y. (2015). Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.* 15:174. doi: 10.1186/s12862-015-0456-4

Parrish, C. R., Murcia, P. R., and Holmes, E. C. (2015). Influenza virus reservoirs and intermediate hosts: dogs, horses, and new possibilities for influenza virus exposure of humans. *J. Virol.* 89, 2990–2994. doi: 10.1128/JVI.03146-14

Peden, J. F. (2000). *Analysis of Codon Usage*. Doctoral thesis, University of Nottingham, Nottingham.

Puigbò, P., Bravo, I. G., and Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biol. Direct.* 3:38. doi: 10.1186/1745-6150-3-38

Saha, J., Bhattacharjee, S., Pal, M. S., Saha, B. K., Basak, H. K., Adhikary, S., et al. (2021). A comparative genomics-based study of positive strand RNA viruses emphasizing on SARS-CoV-2 utilizing dinucleotide signature, codon usage and codon context analyses. *Gene Rep.* 23:101055. doi: 10.1016/j.genrep.2021.101055

Sanchez, G., Bosch, A., and Pinto, R. M. (2003). Genome variability and capsid structural constraints of hepatitis a virus. *J. Virol.* 77, 452–459. doi: 10.1128/jvi.77.1.452-459.2003

Shen, S. H., Stauft, C. B., Gorbatsevych, O., Song, Y., Ward, C. B., Yurovsky, A., et al. (2015). Large-scale recoding of an arbovirus genome to rebalance its insect versus mammalian preference. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4749–4754. doi: 10.1073/pnas.1502864112

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., et al. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459, 1122–1125. doi: 10.1038/nature08182

Tort, F. L., Castells, M., and Cristina, J. (2020). A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. *Virus Res.* 283:197976. doi: 10.1016/j.virusres.2020.197976

Tsai, Y. H., Huang, S. W., Hsieh, W. S., Cheng, C. K., Chang, C. F., Wang, Y. F., et al. (2019). Enterovirus A71 containing codon-deoptimized VP1 and high-fidelity polymerase as next-generation vaccine candidate. *J. Virol.* 93, e2308–e2318. doi: 10.1128/JVI.02308-18

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., et al. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* 30, 269–271. doi: 10.1038/s41422-020-0282-0

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* 87, 23–29. doi: 10.1016/0378-1119(90)90491-9

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome composition and divergence of the novel Coronavirus (2019-nCoV) originating in China. *Cell Host Microb.* 27, 325–328. doi: 10.1016/j.chom.2020.02.001

Wu, Z., and McGoogan, J. M. (2020). Characteristics of and important lessons from the Coronavirus disease 2019 (COVID-19) outbreak in china: summary of a report of 72314 cases from the Chinese center for disease control and prevention. *JAMA* 323, 1239–1242. doi: 10.1001/jama.2020.2648

Xiang, H., Zhang, R., Butler, R. R. III, Liu, T., Zhang, L., Pombert, J. F., et al. (2015). Comparative analysis of codon usage bias patterns in microsporidian genomes. *PLoS One* 10:e0129223. doi: 10.1371/journal.pone.0129223

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J. J., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* 583, 286–289. doi: 10.1038/s41586-020-2313-x

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.