Journal of Virological Methods 289 (2021) 114032

Contents lists available at ScienceDirect

Journal of Virological Methods

journal homepage: www.elsevier.com/locate/jviromet

# Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses

Zigui Chen [a,*], Siaw S. Boon [a], Maggie H. Wang [b], Renee W.Y. Chan [c], Paul K.S. Chan [a,d]

[a] Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[b] Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[c] Department of Paediatrics, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[d] Stanley Ho Centre for Emerging Infectious Diseases, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

*Keywords:*
COVID-19
SARS-CoV-2
SARS-CoV
MERS-CoV
Phylogeny
Codon usage
Dinucleotide suppression

ABSTRACT

Three highly pathogenic human coronaviruses can cause severe acute respiratory syndrome (SARS-CoV, SARS-CoV-2 and MERS-CoV). Although phylogenetic analyses have indicated ancient origin of human coronaviruses from animal relatives, their evolutionary history remains to be established. Using phylogenetics and “high order genomic structures” including trimer spectrums, codon usage and dinucleotide suppression, we observed distinct clustering of all human coronaviruses that formed phylogenetic clades with their closest animal relatives, indicating they have encompassed long evolutionary histories within specific ecological niches before jumping species barrier to infect humans. The close relationships between SARS-CoV and SARS-CoV-2 imply similar evolutionary origin. However, a lower Effective Codon Number (ENC) pattern and CpG dinucleotide suppression in SARS-CoV-2 genomes compared to SARS-CoV and MERS-CoV may imply a better host fitness, and thus their success in sustaining a pandemic. Characterization of coronavirus heterogeneity via complementary approaches enriches our understanding on the evolution and virus-host interaction of these emerging human pathogens while the underlying mechanistic basis in pathogenicity warrants further investigation.

## 1. Introduction

The novel coronavirus disease (COVID-19) pandemic emerged in December 2019 is dramatically threatening global public health and economy, resulting in over 7.6 million confirmed cases and >420,000 deaths in at least 216 countries/cities as of June 15, 2020 (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). This disease is caused by a novel coronavirus, SARS-CoV-2, belonging to the genus *Betacoronavirus* in the family *Coronaviridae*, with major clinical symptoms of cough and fever but also acute respiratory distress syndrome or multiorgan failure (Huang et al., 2020; Lu et al., 2020).

The family *Coronaviridae* consists of four genera: *Alpha-*, *Beta-*, *Gamma-* and *Deltacoronavirus*, with the former two genera infecting mammals only (Cui et al., 2019). Before COVID-19, six human coronaviruses (HCoVs) have been identified, including two (SARS-CoV, MERS-CoV) highly pathogenic clusters that cause severe respiratory pneumonia, and the other four (HCoV-OC43, HCoV-229E, HCoV-NL63 and HCoV-HKU1) that mainly cause common cold. In late 2002, the first severe acute respiratory syndrome coronavirus (SARS-CoV) emerged in Guangdong, China and spread to more than 30 countries. The overall fatality was ca. 10 % with more than 8000 infections (www.who.int) (Zhong et al., 2003). SARS-CoV was not found in humans since 2004. In 2012, another highly pathogenic coronavirus, Middle East respiratory syndrome coronavirus (MERS-CoV) emerged in Middle Eastern countries (Zaki et al., 2012), causing severe pneumonia and renal failure in humans, with a mortality of ~30 %. Phylogenetic analyses indicate that all human coronaviruses have an ancient origin from animals: SARS-CoV, MERS-CoV, HCoV-NL63, HCoV-229E, and perhaps SARS-CoV-2 are related to coronaviruses detected in bats; whereas HCoV-OC43 and HCoV-HKU1 are closely related to coronaviruses detected in rodents (Cui et al., 2019; Forni et al., 2017; Zhou et al., 2020). Evolution and transmission of coronaviruses from their natural reservoirs to humans likely involve intermediate hosts (e.g., palm civets for SARS-CoV, dromedary camels for MERS-CoV) (Reusken et al., 2013; Wang et al., 2005). Phylogenetically, SARS-CoV and SARS-CoV-2 share a most recent common ancestor within the subgenus *Sarbecovirus*, and are

* Corresponding author at: Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, 1/F, Prince of Wales Hospital, Shatin, Hong Kong Special Administrative Region.

*E-mail address:* zigui.chen@cuhk.edu.hk (Z. Chen).

https://doi.org/10.1016/j.jviromet.2020.114032
Received 27 August 2020; Received in revised form 15 November 2020; Accepted 2 December 2020
Available online 5 December 2020

relatively distant to MERS-CoV (belonging to the subgenus *Merbecovirus*) in the genus *Betacoronavirus* (Fig. 1). This genus also includes two other HCoV clusters, HCoV-OC43 and HCoV-HKU1, whereas HCoV-NL63 and HCoV-229E are classified within the genus *Alphacoronavirus*.

Coronaviruses form enveloped and spherical particles of 100–160 nm in diameter. It contains a single-stranded, positive (+)-sense RNA (+ssRNA) genome ranging from 27 to 32 kilobases in length (Cui et al., 2019). The 5′-terminal two-thirds of the genome encodes two large nonstructural polyproteins 1a and 1b which are involved in genome transcription and replication. The 3′ terminus encodes structural proteins, including envelope glycoproteins spike (S), envelope (E), membrane (M) and nucleocapsid (N). The S protein plays a critical role in viral attachment, fusion, entry and transmission. The N-terminal S1 subunit is responsible for virus-host receptor binding and the C-terminal S2 subunit is for virus–cell membrane fusion. SARS-CoV-1 and SARS-CoV-2 use host angiotensin-converting enzyme 2 (ACE2) as a receptor, whereas MERS-CoV binds dipeptidyl peptidase 4 (DPP4) (Du et al., 2009, 2017; Wrapp et al., 2020). During infection, HCoV first binds the host cell through interaction between its S1 receptor-binding domain (S1-RBD) and the cell membrane receptor, followed by cleavage by furin proteases leading to conformational changes in the S2 subunit for virus fusion and entry. Currently, there is no vaccine or therapeutics for effective prevention or treatment of HCoV infection while the research and development of neutralizing antibodies, mainly targeting S1 and/or S2 regions have been vigorously undertaken (Jiang et al., 2020).

With advances in DNA sequencing and bioinformatics analysis, multiple sequence alignments have been used to interrogate genomic diversity and evolution. The highly divergent homologous sequences among HCoV genomes, however, may lead to ambiguous alignments that degrade resolution and bias phylogenetic inference. Alternatively, nonparametric agnostic approaches based on the distribution of exact sub-sequences (*k*-mer spectrum, the DNA 'word' with defined length) and additional genetic metrics (for example, codon usage preference, dimer nucleotide composition) provide complementary information on the complexity and relationship of homologous sequences (Chan and Ragan, 2013; Chor et al., 2009; Shackelton et al., 2006; Vinga and Almeida, 2003). These agnostic methods also avoid complex computation or model selection while capturing signals otherwise lost to indel, recombination or shuffling (Chan and Ragan, 2013). In an effort to interrogate the evolutionary process driving the divergence of coronaviruses, we have primarily focused on human coronaviruses because of the association with severe respiratory pneumonia in human beings. Multiple parametric and nonparametric algorithms were applied. Using the rich resource of a large number of genomes, we sought to uncover hidden biological signals not easily discovered with homology-based methods alone.

## 2. Results

### 2.1. SARS-CoV-2 phylogeny and genomic diversity

To better understand the evolutionary position of SARS-CoV-2, we first examined the phylogenetic relationship of the subgenus *Sarbecoviruses* within the genus *Betacoronaviruses* using the concatenated nucleotide sequences of 12 ORFs/genes (ORF1a, ORF1b, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N and ORF10) (Fig. 2A, Table S1). In consistent with previous reports, SARS-CoV-2 shares sequence similarities of 96.13 ± 0.06 % with a bat coronavirus isolate (RaTG13, NCBI accession number MN996532, GISAID accession number

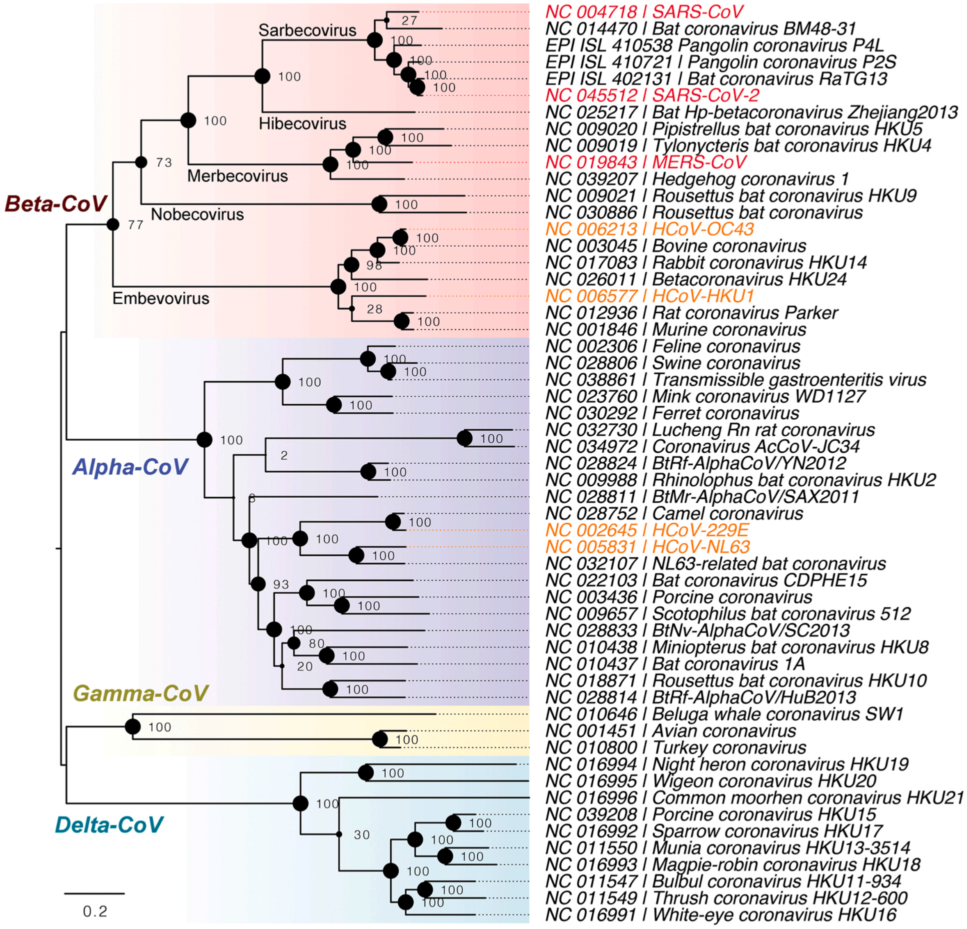
Beta-CoV
Sarbecovirus
Hibecovirus
Merbecovirus
Nobecovirus
Embevovirus
Alpha-CoV
Gamma-CoV
Delta-CoV
27
100
100
100
100
100
100
73
100
100
100
77
100
100
100
98
100
28
100
100
100
100
100
100
2
100
6
100
100
100
100
100
93
100
100
100
80
100
20
100
100
100
100
100
30
100
100
100
100
100
100
0.2
NC 004718 | SARS-CoV
NC 014470 | Bat coronavirus BM48-31
EPI ISL 410538 Pangolin coronavirus P4L
EPI ISL 410721 | Pangolin coronavirus P2S
EPI ISL 402131 | Bat coronavirus RaTG13
NC 045512 | SARS-CoV-2
NC 025217 | Bat Hp-betacoronavirus Zhejiang2013
NC 009020 | Pipistrellus bat coronavirus HKU5
NC 009019 | Tylonycteris bat coronavirus HKU4
NC 019843 | MERS-CoV
NC 039207 | Hedgehog coronavirus 1
NC 009021 | Rousettus bat coronavirus HKU9
NC 030886 | Rousettus bat coronavirus
NC 006213 | HCoV-OC43
NC 003045 | Bovine coronavirus
NC 017083 | Rabbit coronavirus HKU14
NC 026011 | Betacoronavirus HKU24
NC 006577 | HCoV-HKU1
NC 012936 | Rat coronavirus Parker
NC 001846 | Murine coronavirus
NC 002306 | Feline coronavirus
NC 028806 | Swine coronavirus
NC 038861 | Transmissible gastroenteritis virus
NC 023760 | Mink coronavirus WD1127
NC 030292 | Ferret coronavirus
NC 032730 | Lucheng Rn rat coronavirus
NC 034972 | Coronavirus AcCoV-JC34
NC 028824 | BtRf-AlphaCoV/YN2012
NC 009988 | Rhinolophus bat coronavirus HKU2
NC 028811 | BtMr-AlphaCoV/SAX2011
NC 028752 | Camel coronavirus
NC 002645 | HCoV-229E
NC 005831 | HCoV-NL63
NC 032107 | NL63-related bat coronavirus
NC 022103 | Bat coronavirus CDPHE15
NC 003436 | Porcine coronavirus
NC 009657 | Scotophilus bat coronavirus 512
NC 028833 | BtNv-AlphaCoV/SC2013
NC 010438 | Miniopterus bat coronavirus HKU8
NC 010437 | Bat coronavirus 1A
NC 018871 | Rousettus bat coronavirus HKU10
NC 028814 | BtRf-AlphaCoV/HuB2013
NC 010646 | Beluga whale coronavirus SW1
NC 001451 | Avian coronavirus
NC 010800 | Turkey coronavirus
NC 016994 | Night heron coronavirus HKU19
NC 016995 | Wigeon coronavirus HKU20
NC 016996 | Common moorhen coronavirus HKU21
NC 039208 | Porcine coronavirus HKU15
NC 016992 | Sparrow coronavirus HKU17
NC 011550 | Munia coronavirus HKU13-3514
NC 016993 | Magpie-robin coronavirus HKU18
NC 011547 | Bulbul coronavirus HKU11-934
NC 011549 | Thrush coronavirus HKU12-600
NC 016991 | White-eye coronavirus HKU16

**Fig. 1.** Phylogeney of the family *Coronaviridae*. A maximum likelihood (ML) tree was contructed using RAxML MPI v8.2.12 inferred from the concatenated nucleotide sequence alignments of 6 open reading frames (1a-1b-S-E-M-N) of 55 reference genomes. The dot size on the nodes is proportional to the bootstrap support values. The HCoV clusters associated with severe acute respiratory syndrome (SARS-CoV, SARS-CoV-2 and MERS-CoV) and common cold (HCoV-OC43, HCoV-HKU1, HCoV-229E and HCoV-NL63) were highlighted in red and orange, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

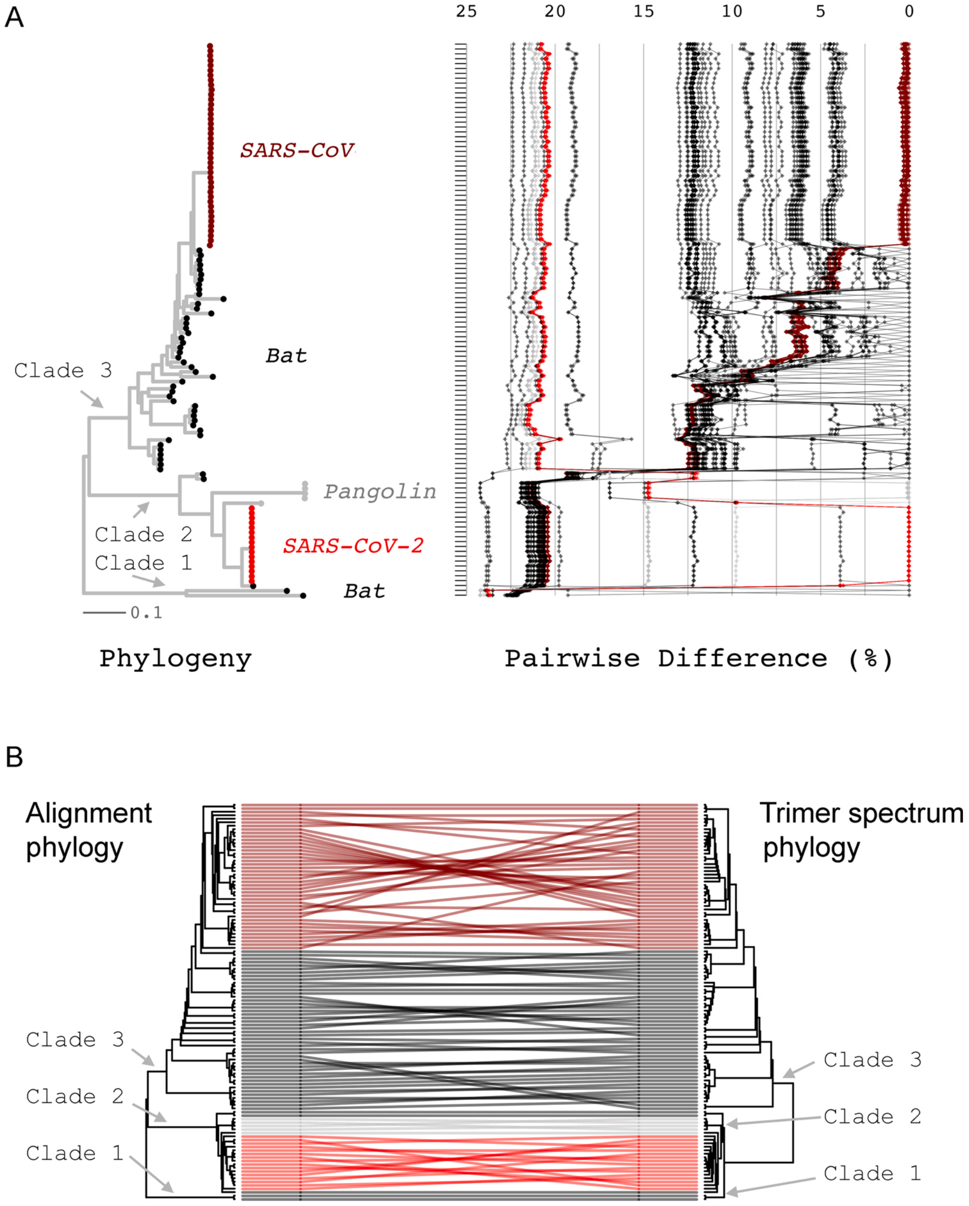

**Fig. 2.** Phylogeny of the subgenus *Sarbecovirus* in the genus *Betacoronavirus*. (A) A maximum likelihood (ML) tree was constructed using RAxML MPI v8.2.12 inferred from the concatenated nucleotide sequence alignments of 12 open reading frames (1a-1b-S-3a-E-M-6-7a-7b-8-N-10) of 114 genomes. The percent nucleotide differences are shown in the panel to the right of the phylogeny. Values for each comparison of a given isolate are connected by lines and the comparison to self is indicated by the 0.0 % difference point. Coloured lines are used to distinguish SARS-CoV-1 and SARS-CoV-2 clusters. **(B)** Tanglegram of tree topologies between the hierarchical clustering. Trimer spectrum and maximum likelihood of 114 *Sarbecovirus* genomes inferred from the concatenated nucleotide sequences of 12 ORF/genes. The bar to the side of each panel indicates the subgenus assignment as coloured according to the key in the figure.

EPI_ISL_402131), similar to the genomic differences between SARS-CoV and bat SARS-like CoVs (e.g., KY417150, KY417146) (similarities of 96.06 ± 0.25 %), implying potential zoonotic transmission of SARS-CoV-2 from bats as their natural reservoirs to humans through unknown intermediate hosts. SARS-CoV-2 also shares close genomic similarities with coronavirus isolates from pangolin animals (EPI_ISL_410984, 90.22 ± 0.03 %; EPI_ISL_410538 - EPI_ISL_410542, 85.27 ± 0.03 %), whereas the complete genome sequence similarities between SARS-CoV and SARS-CoV-2 are 79.4 ± 0.17 %. The subgenus *Sarbecovirus* has average sequence similarities of ≤ 48.4 % with other coronaviruses (data not shown).

### 2.2. *K-mer spectrum clustering of SARS-CoV-2*

The *k*-mer and other nonparametric approaches, such as codon usage and dinucleotide composition, constitute a “higher order genomic structure” that may reflect the influence of evolutionary processes extending beyond common measures of Darwinian selection. In order to explore an agnostic evolutionary model, we used trimer ($k$ = 3) frequency distribution to construct the phylogeny of the subgenus *Sarbecovirus* (Figs. 2B). Both parametric (alignment phylogeny) and nonparametric (trimer spectrum phylogeny) approaches showed distinct separation between SARS-CoV and SARS-CoV-2, with their animal relatives closely clustering to each other.

### 2.3. *Codon usage bias of human betacoronaviruses*

The trimer spectrums between human *bCoV* clusters may imply differential codon usages between viruses and host cells. We then measured the ENC values of coronaviruses across 6 genes (ORF1a, ORF1b, S, E, M, and N) shared by all members, with a main focus on the difference between seven human coronavirus clusters (Table S2). The ENC statistic is a way of analyzing how biased a gene is in terms of its codon usage, with values ranging from 20 when a gene is effectively using only a single codon for each amino acid (strongest bias) to 61 when a gene tends to use all codons with equal frequency (no bias). The ENC values of the surveyed coronavirus genomes ranged between 35.6 and 54.1 (a mean value of 45.4 ± 4.7), with a significant correlation with GC contents (adjusted $R^2$ = 0.953, p < 0.001) (Figure S1). Strong codon usage bias was observed in HCoV-HKU1 (35.7 ± 0.09) and HCoV-NL63 (37.3 ±

0.08) genomes when compared to other HCoV clusters (45.9 ± 4.2, $p < 0.001$) (Fig. 3A). Interestingly, SARS-CoV-2, HCoV-OC43 and HCoV-229E genomes shared similar ENC values, but demonstrated stronger codon usage bias than SARS-CoV and MERS-CoV ($p = 0.008$).

In order to obtain a better understanding on the relationship between codon usage bias and gene composition, a plot of ENC values against GC contents at the synonymous third codon position (GC3s) was constructed (Fig. 3B). This method was used to estimate the factors shaping the codon usage pattern. If codon usage pattern was affected by GC3s alone, the observed ENC values should be on or just below the expected ENC* curve indicating the synonymous codon usage bias may be subject to GC-biased mutational pressure. As shown, all surveyed HCoV genomes lay slightly under the expected ENC* curve (the red curve in Fig. 3B), implying an important role of uneven base composition and hence, of mutational pressure that affected the formation of codon usage bias. However, codon usage bias tended to be less dependent on variation of GC3s when GC contents decreased, with measures of (ENC* - ENC) / ENC* suggesting that other factors, such as translational or natural selection, may act as forces affecting stronger codon usage bias and higher genomic diversity of HCoV-HKU1 and HCoV-NL63 (Fig. 3C).

Differential codon usage patterns within distinct ORFs were observed between HCoV clusters (Figure S2). SARS-CoV and MERS-CoV, for example, had the least codon usage biases across ORFs while HCoV-HKU1 and HCoV-NL63 were the strongest ones. The S and N genes, however, showed no significant difference in ENC values between SARS-CoV-2 and SARS-CoV. The M gene may encompass more diversified selection forces amongst HCoV clusters, as lower values of (ENC* - ENC) / ENC* were observed, probably due to the relatively small protein encoded. It is noted that the E gene was not included for comparison because of the small protein size (< 89 aa), for which interpretation may not be applied.

### 2.4. Synonymous codon usage pattern in human betacoronaviruses

ENC value measures the codon usage bias of an entire genome/gene but not of individual codon. We further calculated the Relative Synonymous Codon Usage (RSCU) values across 6 ORFs of HCoV genomes to better estimate the differential usage of each synonymous codon. Among 59 codons encoding for 18 amino acids (except for Met, Trp, and stop codons), 6 and 9 were defined as preferred (RSCU values > 1.6) and suppressed codons (< 0.6) within all HCoV genomes, respectively (Fig. 3D, Table 1). For example, HCoV genomes prefer to GGT (RSCU ≥ 1.91) rather than GGG (≤ 0.26) to encode Glycine. Interestingly, the majority of human-preferred codons were less commonly found in HCoV genomes, such as CTG encoding for Leucine (Human, 2.37; HCoV, ≤ 0.58) and GCC encoding for Alanine (Human, 1.60; HCoV, ≤ 0.59); in contrast, CGT, the optimized codon encoding for Arginine in HCoV genomes (mean of 2.03) was rarely found in human genomes (0.48). The animal CoV genomes share similar RSCU patterns as the HCoV ones.

In contrast to HCoV genomes that intended to use A/T-ending codons, as observed in 96 % (25/26) of codons with average values of RSCU less than 1.0, human host tends to use G/C-ending codons (74 %, 20/27). However, both HCoV and human displayed a strong tendency to avoid using CG-ending codons (HCoV ≤ 0.28, human ≤ 0.46),

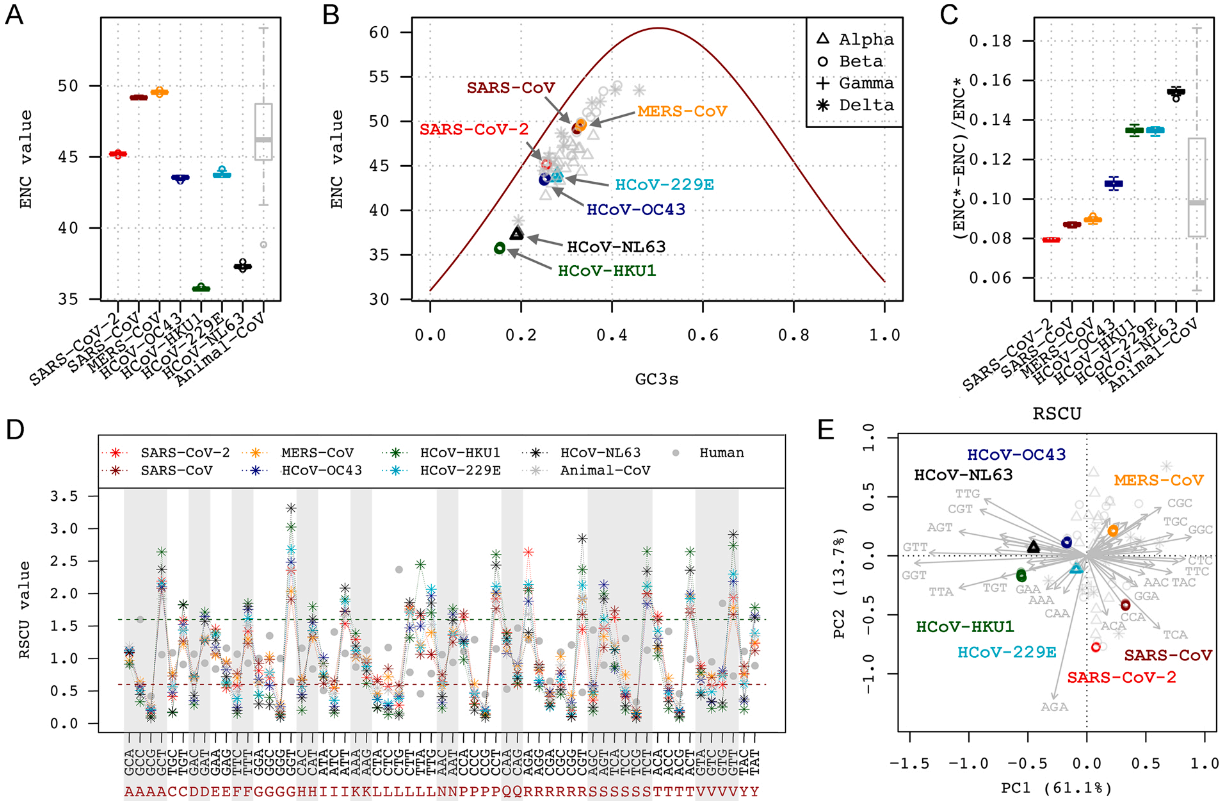

**Fig. 3.** Synonymous codon usage of coronavirus genomes based on concatenated nucleotide sequences of 6 ORFs (ORF1a-1b-S-E-M-N). (A) Boxplot of Effective Number of Codon (ENC) between HCoV clusters. The ENC values range from 20 when a gene is effectively using only a single codon for each amino acid (strongest bias) to 61 when a gene trends to use all codons with equal frequency (no bias). **(B)** Plot of ENC and the synonymous third codon position (GC3s) content. The red curve indicates the expected ENC* if codon usage pattern is only affected by GC3s. **(C)** Boxplot of differences between the observed and expected ENC values among HCoV clusters. **(D)** Mean values of Relative Synonymous Codon Usage (RSCU) for 59 codons (except for Met, Trp, and stop codons) amongst HCoV clusters. The preferred and suppressed codon usages were defined as RSCU values > 1.6 or < 0.6, respectively. **(E)** Scatter biplot of RSCU of HCoV clusters. The clustering was performed using redundancy analysis (RDA), with colours assigned to different human betacoronavirus clusters. The x-axis and the y-axis represent the first two principal coordinate component (PCoA) axes. For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

**Table 1**
Relative synonymous codon usage (RSCU) patterns of the surveyed human betacoronaviruses inferred from the concatenated 6 ORFs (ORF1a, 1b, S, E, M, N). (For interpretation of the references to colour in this Table legend, the reader is referred to the web version of this article).

| Amino acid | Codon | SARS-CoV-2 | SARS-CoV | MERS-CoV | HCoV-OC43 | HCoV-HKU1 | HCoV-229E | HCoV-NL63 | Animal-CoV | | Human |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala (A) | GCA | 1.08 | 1.09 | 0.97 | 1.13 | 0.91 | 1.12 | 1.10 | 1.18 | | 0.91 |
| | GCC | 0.57 | 0.59 | 0.65 | 0.57 | 0.33 | 0.51 | 0.45 | 0.62 | | 1.60 |
| | GCG | 0.16 | 0.24 | 0.28 | 0.20 | 0.11 | 0.22 | 0.08 | 0.28 | | 0.42 |
| | GCT | 2.19 | 2.08 | 2.10 | 2.10 | 2.64 | 2.16 | 2.37 | 1.92 | | 1.06 |
| Cys (C) | TGC | 0.42 | 0.73 | 0.78 | 0.48 | 0.17 | 0.56 | 0.18 | 0.55 | | 1.09 |
| | TGT | 1.58 | 1.27 | 1.22 | 1.52 | 1.83 | 1.44 | 1.82 | 1.45 | | 0.91 |
| Asp (D) | GAC | 0.71 | 0.73 | 0.70 | 0.35 | 0.29 | 0.72 | 0.43 | 0.67 | | 1.07 |
| | GAT | 1.29 | 1.27 | 1.30 | 1.65 | 1.71 | 1.28 | 1.57 | 1.33 | | 0.93 |
| Glu (E) | GAA | 1.45 | 1.07 | 1.05 | 1.17 | 1.39 | 1.36 | 1.28 | 1.14 | | 0.84 |
| | GAG | 0.55 | 0.93 | 0.95 | 0.83 | 0.61 | 0.64 | 0.72 | 0.86 | | 1.16 |
| Phe (F) | TTC | 0.58 | 0.77 | 0.70 | 0.24 | 0.15 | 0.38 | 0.20 | 0.49 | | 1.07 |
| | TTT | 1.42 | 1.23 | 1.30 | 1.76 | 1.85 | 1.62 | 1.80 | 1.51 | | 0.93 |
| Gly (G) | GGA | 0.81 | 0.92 | 0.64 | 0.68 | 0.43 | 0.44 | 0.30 | 0.62 | | 1.00 |
| | GGC | 0.70 | 0.99 | 1.03 | 0.58 | 0.40 | 0.78 | 0.29 | 0.79 | | 1.35 |
| | GGG | 0.13 | 0.18 | 0.30 | 0.26 | 0.14 | 0.10 | 0.10 | 0.22 | | 1.00 |
| | GGT | 2.36 | 1.91 | 2.04 | 2.48 | 3.02 | 2.68 | 3.31 | 2.37 | | 0.65 |
| His (H) | CAC | 0.57 | 0.67 | 0.68 | 0.44 | 0.20 | 0.57 | 0.39 | 0.61 | | 1.16 |
| | CAT | 1.43 | 1.33 | 1.32 | 1.56 | 1.80 | 1.43 | 1.61 | 1.39 | | 0.84 |
| Ile (I) | ATA | 0.92 | 0.62 | 0.73 | 1.01 | 0.92 | 0.76 | 0.71 | 0.89 | | 0.51 |
| | ATC | 0.54 | 0.65 | 0.55 | 0.26 | 0.16 | 0.36 | 0.20 | 0.46 | | 1.41 |
| | ATT | 1.53 | 1.73 | 1.72 | 1.72 | 1.91 | 1.88 | 2.08 | 1.65 | | 1.08 |
| Lys (K) | AAA | 1.30 | 1.07 | 1.01 | 1.02 | 1.39 | 1.12 | 1.19 | 1.02 | | 0.87 |
| | AAG | 0.70 | 0.93 | 0.99 | 0.98 | 0.61 | 0.88 | 0.81 | 0.98 | | 1.13 |
| Leu (L) | CTA | 0.68 | 0.64 | 0.46 | 0.37 | 0.24 | 0.37 | 0.24 | 0.55 | | 0.43 |
| | CTC | 0.57 | 0.84 | 0.72 | 0.27 | 0.14 | 0.30 | 0.22 | 0.51 | | 1.17 |
| | CTG | 0.28 | 0.58 | 0.45 | 0.42 | 0.16 | 0.39 | 0.13 | 0.48 | | 2.37 |
| | CTT | 1.75 | 1.81 | 1.77 | 1.47 | 1.32 | 1.71 | 1.86 | 1.72 | | 0.79 |
| | TTA | 1.66 | 1.07 | 1.21 | 1.50 | 2.45 | 1.16 | 1.69 | 1.22 | | 0.46 |
| | TTG | 1.06 | 1.07 | 1.40 | 1.97 | 1.70 | 2.07 | 1.85 | 1.52 | | 0.77 |
| Asn (N) | AAC | 0.65 | 0.75 | 0.59 | 0.32 | 0.24 | 0.63 | 0.41 | 0.63 | | 1.06 |
| | AAT | 1.35 | 1.25 | 1.41 | 1.68 | 1.76 | 1.37 | 1.59 | 1.37 | | 0.94 |
| Pro (P) | CCA | 1.64 | 1.69 | 1.23 | 1.29 | 0.98 | 1.23 | 1.27 | 1.37 | | 1.11 |
| | CCC | 0.30 | 0.40 | 0.61 | 0.48 | 0.29 | 0.46 | 0.20 | 0.48 | | 1.29 |
| | CCG | 0.14 | 0.14 | 0.16 | 0.21 | 0.13 | 0.20 | 0.09 | 0.27 | | 0.45 |
| | CCT | 1.92 | 1.77 | 2.00 | 2.02 | 2.60 | 2.11 | 2.44 | 1.88 | | 1.15 |
| Gln (Q) | CAA | 1.40 | 1.22 | 1.11 | 1.08 | 1.38 | 1.28 | 1.31 | 1.07 | | 0.53 |
| | CAG | 0.60 | 0.78 | 0.89 | 0.92 | 0.62 | 0.72 | 0.69 | 0.93 | | 1.47 |
| Arg (R) | AGA | 2.64 | 2.08 | 1.33 | 1.88 | 2.05 | 2.14 | 1.40 | 1.68 | | 1.29 |
| | AGG | 0.81 | 0.91 | 0.83 | 0.63 | 0.57 | 0.66 | 0.86 | 0.88 | | 1.27 |
| | CGA | 0.31 | 0.42 | 0.47 | 0.49 | 0.32 | 0.26 | 0.27 | 0.35 | | 0.65 |
| | CGC | 0.60 | 0.81 | 1.03 | 0.76 | 0.45 | 0.67 | 0.52 | 0.90 | | 1.10 |
| | CGG | 0.20 | 0.11 | 0.43 | 0.32 | 0.25 | 0.14 | 0.10 | 0.26 | | 1.21 |
| | CGT | 1.45 | 1.68 | 1.91 | 1.93 | 2.37 | 2.13 | 2.85 | 1.93 | | 0.48 |
| Ser (S) | AGC | 0.36 | 0.50 | 0.44 | 0.58 | 0.19 | 0.48 | 0.25 | 0.55 | | 1.44 |
| | AGT | 1.46 | 1.17 | 1.34 | 2.14 | 2.00 | 1.61 | 1.92 | 1.48 | | 0.90 |
| | TCA | 1.63 | 1.73 | 1.20 | 0.84 | 0.87 | 1.14 | 1.13 | 1.23 | | 0.90 |
| | TCC | 0.44 | 0.41 | 0.75 | 0.45 | 0.21 | 0.50 | 0.27 | 0.56 | | 1.31 |
| | TCG | 0.11 | 0.20 | 0.17 | 0.15 | 0.08 | 0.16 | 0.09 | 0.26 | | 0.33 |
| | TCT | 2.00 | 1.99 | 2.11 | 1.84 | 2.65 | 2.11 | 2.34 | 1.92 | | 1.13 |
| Thr (T) | ACA | 1.65 | 1.59 | 1.15 | 1.31 | 1.04 | 1.42 | 1.20 | 1.33 | | 1.14 |
| | ACC | 0.39 | 0.54 | 0.70 | 0.51 | 0.23 | 0.46 | 0.34 | 0.60 | | 1.42 |
| | ACG | 0.19 | 0.17 | 0.19 | 0.19 | 0.09 | 0.20 | 0.11 | 0.31 | | 0.46 |
| | ACT | 1.77 | 1.70 | 1.96 | 1.99 | 2.64 | 1.92 | 2.36 | 1.77 | | 0.99 |
| Val (V) | GTA | 0.89 | 0.85 | 0.75 | 0.67 | 0.78 | 0.48 | 0.44 | 0.69 | | 0.47 |
| | GTC | 0.58 | 0.71 | 0.73 | 0.33 | 0.23 | 0.49 | 0.34 | 0.57 | | 0.95 |
| | GTG | 0.60 | 0.77 | 0.74 | 0.81 | 0.25 | 0.74 | 0.32 | 0.75 | | 1.85 |
| | GTT | 1.93 | 1.68 | 1.78 | 2.19 | 2.74 | 2.30 | 2.91 | 1.99 | | 0.73 |
| Tyr (Y) | TAC | 0.77 | 0.88 | 0.75 | 0.34 | 0.22 | 0.61 | 0.37 | 0.65 | | 1.11 |
| | TAT | 1.23 | 1.12 | 1.25 | 1.66 | 1.78 | 1.39 | 1.63 | 1.35 | | 0.89 |
| | | | | | | | | | | | |
| | Preferred | 12 | 12 | 9 | 15 | 18 | 13 | 16 | 10 | | 3 |
| | Suppresse | 23 | 13 | 12 | 24 | 28 | 22 | 28 | 18 | | 10 |

ˆ the preferred codons (RSCU > 1.6) and the suppressed codons (RSCU < 0.6) are highlighted in red and green, respectively.

suggesting a potential role of CpG dinucleotide depletion in shaping the codon usages.

To further investigate the variation of synonymous codon usage bias in modulating HCoV genomic diversity, correspondence analysis based on the RSCU patterns was performed (Fig. 3E). Scatter plots of the first two axes well supported the distinct separation of HCoV clusters, with SARS-CoV and SARS-CoV-2 sharing relatively similar RSCU patterns, such as preferred codon usage of ACA (Threonine), CCA (Proline) and TCA (Serine) when compared to other HCoV clusters, but relatively lower values of TTG (Leucine) and CGT (Arginine) (Table 1). The four HCoV clusters associated with common cold (HCoV-OC43, HCoV-HKU1, HCoV-229E and HCoV-NL63) may form a separate group based on the codon usage patterns, with higher RSCU values of GT-ending codons when compared to the SARS-related HCoV genomes (SARS-CoV, SARS-CoV-2 and MERS-CoV). When individual ORFs were accessed, we observed similar codon usage patterns amongst HCoV clusters when compared to the concatenated 6 ORFs/genes (Fig. 4). Interestingly, the N gene of SARS-CoV and SARS-CoV-2 shared nearly identical codon usage patterns, implying similar biological property of nucleoprotein between these two viral clusters in packaging viral particles. Within individual HCoV cluster, ORF1a, ORF1b and S gene usually shared a more similar codon usage patterns that were different to that of M and N genes (Figure S3).

### 2.5. Dinucleotide suppression in human betacoronaviruses

Since CG-ending codons may be less frequent in coronavirus genomes (Table 1), we then measured the relative abundance of dinucleotide across 6 ORFs/genes to determine the influence of dinucleotide suppression on codon usage bias. As expected, CpG dinucleotide was mostly depleted (mean value of observed/expected ratio of 0.48 ± 0.06) (Fig. 5A, Table 2), with significant association with overall GC content (pearson correlation of 0.709, $p < 0.001$) and GC content at the third position (GC3) (cor. 0.736, $p < 0.001$) (Figure S4A). TpC dinucleotides represented the second most suppressed dinucleotide (0.78 ± 0.06), consistent with the strong scarcity of nTC codon usage in the surveyed coronavirus genomes (Tables 1 and 2). Since TC dinucleotide is one of the preferred target sequences of host restriction factor APOBEC3 proteins, the suppression of TC dinucleotide could be a result of evolutionary selection allowing viruses to evade restriction from host immune

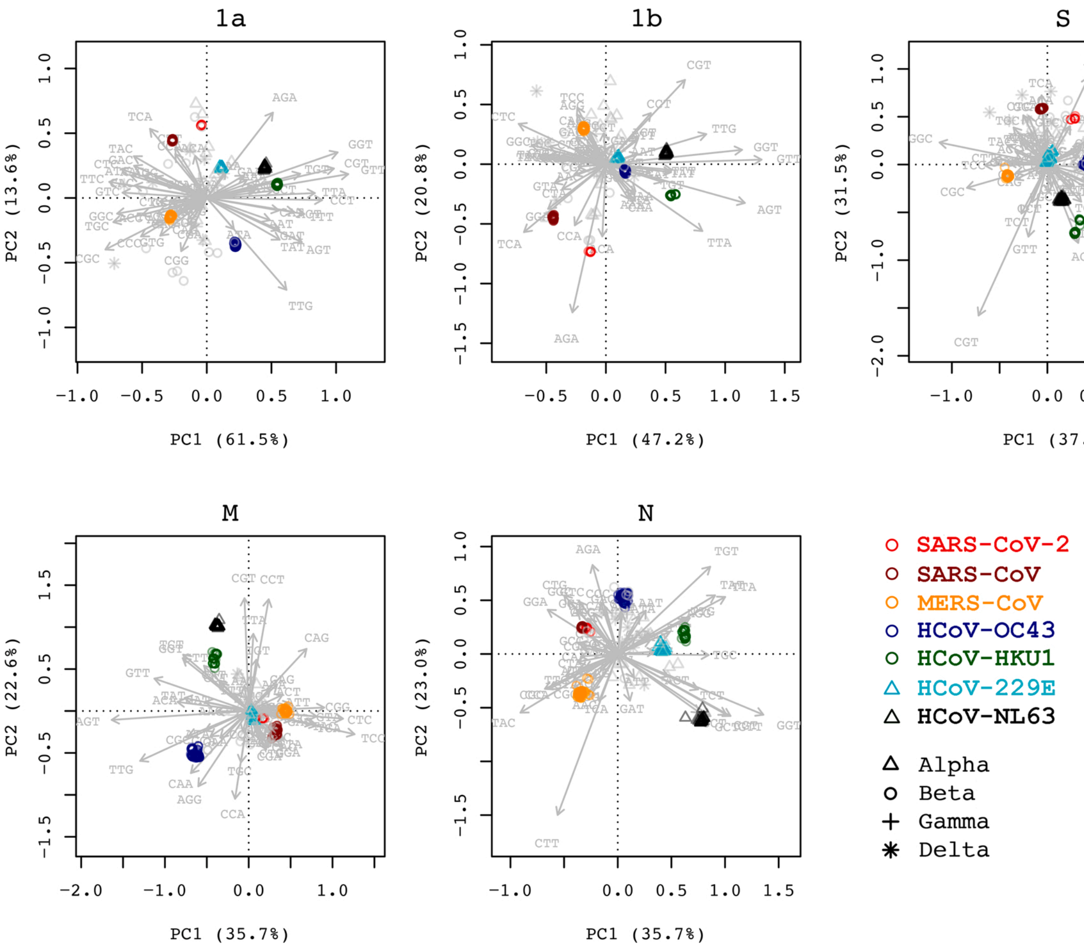

**Fig. 4.** Scatter biplot of Relative Synonymous Codon Usage (RSCU) of HCoV clusters inferred from distinct ORF/gene. The clustering was performed based on RSCU patterns for individual gene using redundancy analysis (RDA), with colours assigned to different coronavirus clusters. The x-axis and the y-axis represent the first two principal coordinate component (PCoA) axes.

protection. TpA dinucleotides were also stringently excluded (0.88 ± 0.05), probably due to in part usage of universal stop (TAA, TAG) and the increased susceptibility of TpA to ribonuclease digestion (Beutler et al., 1989). Interestingly, the loss of TpA seems to be significantly associated with the gain of TpG (Pearson correlation of 0.940, $p <$ 0.001) (Figure S4B, Table S3). Deamination of 5-methylcytosine (5mC) within the CpG island may lead to a cytosine (C) to thymine (T) transition, potentially resulting in a loss of CpG and a gain of TpG. However, the correlation between the loss of CpG and the gain of TpG was not strong (cor=0.138, $p = 0.006$) in the surveyed coronavirus genomes. The gain of TpG was also associated with the loss of ApT (cor=0.730, $p <$ 0.001), but with gain of CpA (cor=0.772, $p < 0.001$) and ApC (cor=0.605, $p < 0.001$).

Discriminative dinucleotide suppression patterns were observed between HCoV clusters. For example, two *Alphacoronavirus* HCoV cluters (HCoV-229E, NCoV-NL63) had a significant gain of ApA but a loss of ApG when compared to the *Betacoronavirus* HCoV clusters ($p < 0.001$) (Figs. 5B and 5C). We also observed higher O/E ratio of CpT, TpC and GpG within MERS-CoV, SARS-CoV and SARS-CoV-2 when compared to the common cold-related HCoV clusters. Interestingly, MERS-CoV genomes had the least loss of CpG dinucleotide, followed by SARS-CoV while SARS-CoV-2 had the most, which might imply differential viral gene expression between these three HCoV clusters. Different levels of dinucleotide suppression were found amongst ORFs/genes (Figure S5). For example, the S gene of SARS-CoV-2 had a significant loss of CpG dinucleotide compared to other genes or other HCoV clusters; the N gene had an overall gain of ApG but a loss of ApC. These differences probably imply an evolutionary apomorphy between HCoV genes that warrants further investigation on their biological properties and clinical relevancies.

## 3. Discussion

Analyses of the origin and evolutionary tree of life have long been based on sequence-aligned dissimilarity to present the homology of organisms in association with genotype, phenotype and phylotype. However, the genetic heterogeneity, such as recombination, duplication, insertion/deletion, genetic fusion and shuffling, and potential sequencing errors, challenges the accuracy of sequence alignment and comparison that strongly relies on heuristic solution and data quality. The relevance of alignment scores to homology may be deducted for coronavirus genomes because of complex histories evolving from animal hosts; hence, complementary approaches containing as much homological signals as possible will provide an opportunity to explore the underlying evolution (Chor et al., 2009). In this study, we applied multiple parametric and nonparametric algorithms (e.g., evolutionary phylogeny, trimer spectrums, codon usage bias and dinucleotide suppression) to compare the differential genomic characterisation between the main coronavirus clusters that infect humans (SARS-CoV, SARS-CoV-2, MERS-CoV, HCoV-OC43, HCoV-HUK1, HCoV-229E and HCoV-NL63). Using the features of a large number of viral genomes we provide evidences with different evolutionary constraints driving the heterogeneity of HCoV genomes. The sharp patterns of codon usage and dinucleotide suppression amongst HCOV clusters (e.g., stronger codon

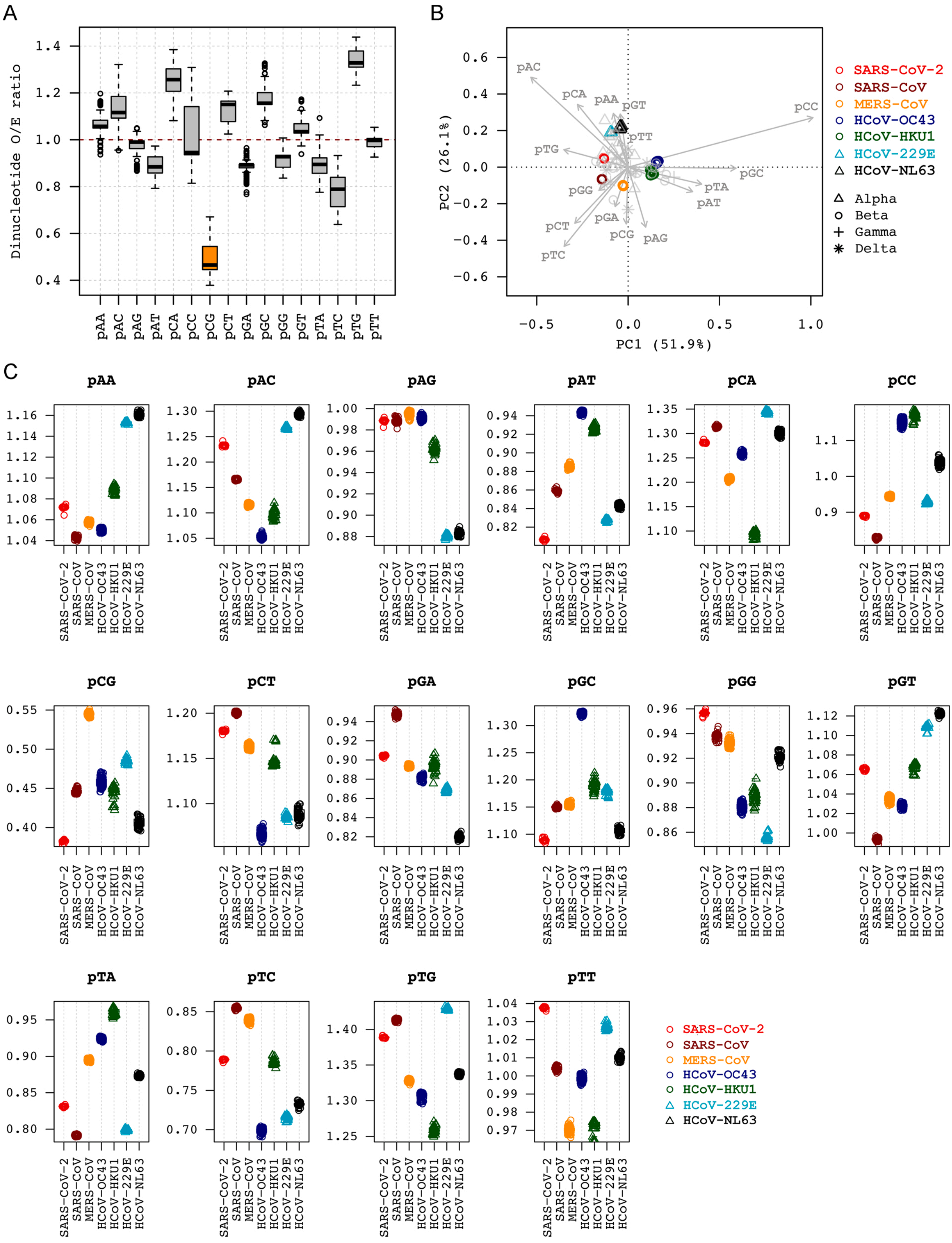

**Fig. 5.** Dinucleotide suppression of HCoV genomes inferred from the concatenated nucleotide sequences of 6 ORFs (ORF1a-1b-S-E-M-N). (A) Boxplot of dinucleotide observed/expected (O/E) ratio. The *ρXY* dinucleotide exhibits suppression if the O/E ratio is less than 1. **(B)** Scatter biplot of relative abundance of dinucleotides of HCoV genomes. The clustering was performed using redundancy analysis (RDA), with colours assigned to different clusters. The x-axis and the y-axis represent the first two principal coordinate component (PCoA) axes. **(C)** Boxplot of the O/E ratios of each dinucleotide amongst HCoV clusters.

**Table 2**
Dinucleatide depletion of the surveyed human betacoronaviruses inferred from the concatenated 6 ORFs (ORF1a, 1b, S, E, M, N).

| Amino acid | SARS-CoV-2 | SARS-CoV | MERS-CoV | HCoV-OC43 | HCoV-HKU1 | HCoV-229E | HCoV-NL63 | Animal-CoV |
|---|---|---|---|---|---|---|---|---|
| pAA | 1.07 | 1.04 | 1.06 | 1.05 | 1.09 | 1.15 | 1.16 | 1.05 |
| pAC | 1.23 | 1.17 | 1.12 | 1.05 | 1.10 | 1.27 | 1.30 | 1.17 |
| pAG | 0.99 | 0.99 | 1.00 | 0.99 | 0.96 | 0.88 | 0.88 | 0.97 |
| pAT | 0.81 | 0.86 | 0.88 | 0.94 | 0.93 | 0.83 | 0.84 | 0.88 |
| pCA | 1.28 | 1.31 | 1.21 | 1.26 | 1.09 | 1.34 | 1.30 | 1.28 |
| pCC | 0.89 | 0.83 | 0.94 | 1.15 | 1.17 | 0.93 | 1.04 | 0.97 |
| pCG | **0.38** | **0.45** | **0.54** | **0.46** | **0.44** | **0.48** | **0.41** | **0.52** |
| pCT | 1.18 | 1.20 | 1.16 | 1.07 | 1.15 | 1.09 | 1.09 | 1.10 |
| pGA | 0.90 | 0.95 | 0.89 | 0.88 | 0.89 | 0.87 | 0.82 | 0.87 |
| pGC | 1.09 | 1.15 | 1.16 | 1.32 | 1.19 | 1.18 | 1.11 | 1.18 |
| pGG | 0.96 | 0.94 | 0.93 | 0.88 | 0.89 | 0.86 | 0.92 | 0.91 |
| pGT | 1.07 | 0.99 | 1.03 | 1.03 | 1.07 | 1.11 | 1.12 | 1.07 |
| pTA | 0.83 | 0.79 | 0.89 | 0.92 | 0.96 | 0.80 | 1.12 | 0.88 |
| pTC | 0.79 | 0.85 | 0.84 | 0.70 | 0.79 | 0.71 | 1.12 | 0.76 |
| pTG | 1.39 | 1.41 | 1.33 | 1.31 | 1.26 | 1.43 | 1.12 | 1.35 |
| pTT | 1.04 | 1.00 | 0.97 | 1.00 | 0.97 | 1.03 | 1.12 | 1.00 |

usage bias of HCoV-HKU1/HCoV-NL63 probably associated with low viral expression) and genes (e.g., distinct patterns between ORF1a/1b/S, M/N and E respective of structure and non-structure proteins) may explain, in part, the different strategies employed in the viral life cycle to evade/manoeuvre host responses, such as deregulated expression in persistent infection, capacity for cell invasion and damage leading to pneumonia, and immune response to host barriers. Both parametric and nonparametric algorithms support distinct separation of HCoV clusters, suggesting different evolutionary histories that the viruses may have encompassed within certain ecological niches or host animals before transmission to humans, while the origin of SARS-CoV and SARS-CoV-2 sharing similar codon usage bias and dinucleotide suppression may converge.

It has been reported that codon usage preferences among viruses and bacteria reflect a balance between mutational biases, genetic drift and natural selection for translational optimization (Bulmer, 1987; Hershberg and Petrov, 2008). Among seven identified human coronavirus clusters, SARS-CoV, SARS-CoV-2 and MERS-CoV are highly pathogenic and have been linked to the development of acute respiratory distress syndrome while increasing evidences indicate that they were also different in transmission, mortality, susceptible population, and early clinical manifestations. Since synonymous mutation is often thought to be selectively neutral, the observed variation in codon usage between different HCoV clusters and their genes may suggest the presence of mutational bias and/or selective pressure that may impact translational efficiency. In RNA viruses, constraints on RNA structures necessary for replication and packaging have also been invoked as another selection pressure for codon bias (Goodfellow et al., 2000). Lastly, the difference in codon usage bias between coronavirus genes could also be explained in part by mutational pressure and selection on genes with different lengths, since selection may be acting to maximize translational efficiency of energetically costly longer genes but reduce the size of highly expressed proteins (Moriyama and Powell, 1998; Zhao and Chen, 2011). These observations raise the complexity of the mechanistic basis that may contribute to niche adaptation, immune exposure, expression and evasion, virulence potential, and probably pathogenicity of human coronaviruses that warrants further investigation.

Viruses rely on host cellular machinery for transcription and replication. However, the virus may have significantly different codon usage to the host genomes for replication and duplication, an evolutionary adaptation to strong host defences and replicative/repair mechanism. Firstly, deoptimized codon usage in coronavirus genomes with respect to that of their hosts may facilitate viral fitness by limiting viral gene expression and eliciting host immune response (Lauring et al., 2012). Suppression by means of codon usage maladaptation may allow viruses to better escape immune surveillance for persistent infection. On the other hand, overexpression of critical viral genes using host cellular machinery may leave the virus more vulnerable; for example, avian-derived influenza A virus M2 overexpression in mammalian model systems was associated with intracellular accumulation of autophagosomes, which is a critical aspect of influenza A virus host adaptation (Calderon et al., 2019). Secondly, codon usage in HCoV genomes might be subject to host innate immune pressure, such as from ABOPEC3, a family of cellular cytidine deaminases that introduce directional C > T substitutions. It has been reported that APOBEC3-mediated cytidine deaminase activity could inhibit replication of HCoV-NL63 (Milewska et al., 2018). Although hypermutation in progeny viruses was not observed, the dramatically underrepresented TpC dinucleotide and TC-ending codons, the preferred dinucleotide target site of many APOBEC3 members, may be responsible for the long-term accumulation of genomic changes that affect the success of niche adaptation or function fitness. Additionally, in HIV, editing cDNA by APOBEC3 could introduce additional genomic composition bias (Liddament et al., 2004). Various phagocytic cells, cytokines, interferons (IFNs), and IFN-stimulated genes have also been reported to play critical roles as defence against DNA/RNA virus infection, but the codon usage optimization of human coronaviruses, particularly for the highly pathogenic clusters in facilitating the immune evasion warrants further investigation. Thirdly, codon usage and dinucleotide composition in HCoV genomes could be due in part to host driven CpG elimination pressure. In mammalian host genomes the CpG dinucleotide is underrepresented, because in this context C is methylated and then deaminated, producing C-T transition (Lister et al., 2009). This creates an obvious mutational pressure on codon usage and results in underrepresentation of nCG codons. DNA methylation functions as a host defence mechanism by regulating gene expression (Shackelton et al., 2006; Upadhyay et al., 2013). When CpG residues of foreign DNA were methylated, pathogen activity can be repressed due to alterations in pathogen transcriptional profiles. CpG repression in RNA viruses might be associated with viral base composition, synonymous codon usage and host selection (Upadhyay et al., 2013). Similar to a number of + ssRNA viruses, HCoV genomes mimic host's CpG usage, which could be tied to evolutionary selection in regulating gene expression, assisting immune evasion, and avoiding C to T mutation elicited by host CpG elimination pressure. However, the extent and dynamics of CpG methylation in HCoV genome and the role in vegetative viral replication and progression to disease remains uncertain.

Codon usage bias of viruses may facilitate persistent infections and/or reinfection in hosts. In contrast, optimization to hosts' codon usage dramatically increases the expression levels of pathogenic genes, may provide an important consideration in developing effective vaccines. Vaccination with codon-optimization for specific viral genes, such as HPV16 E6/E7 (Lin et al., 2006; Steinberg et al., 2005), avian influenza virus H5N1 HA (Stachyra et al., 2016), HIV-1 reverse transcriptase

(Latanova et al., 2018), may promot host immune response in vitro that may inform vaccine development. Although SARS-CoV-2-specific neutralizing monoclonal antibodies are currently not available, the developed anti-SARS-CoV neutralizing antibodies may have potential cross-reactivity against SARS-CoV-2 infection since they share a most recent common ancestor, and both use S1-RBD-ACE2 as a binding-receptor pathway for attachment (Lu et al., 2020; Wrapp et al., 2020). For example, a SARS-CoV specific human monoclonal antibody, CR3022, has been reported to be able to bind potently with SARS-CoV-2 receptor-binding domain, providing an alternative candidate for SARS-CoV-2 prevention (Tian et al., 2020). Hence, codon optimization may provide a promising strategy for the development of more efficient vaccine against human coronaviruses.

## 4. Conclusions

In summary, we applied “high order genomic structures” including trimer spectrums, codon usage and dinucleotide suppression to present a comprehensive analysis of coronaviruses by comparing the diversity and genetic features amongst seven HCoV clusters. Distinct codon usage patterns were observed, which is mainly consistent with the phylogenetic relationships revealed by parametric algorithms. We also observed sharp codon usage patterns amongst genes, mainly between long and short genes, and between structural and non-structural genes. The close relationships between SARS-CoV and SARS-CoV-2 imply similar evolutionary origin, while the lower ENC values in SARS-CoV-2 genome may indicate stronger codon usage bias demonstrating its lower gene expression and/or better host adaptation. The greater variability in synonymous codon usage within coronavirus genomes indicates more complex evolutionary histories in which ancient viral divergence coupled to niche and/or host adaptation has fixed a number of conserved properties, whereas virus divergence and other evolutionary considerations have led to variability within certain limits. The observations raise the complexity of the mechanistic basis that may contribute to niche adaptation, immune exposure, expression and evasion, virulence potential, and probably pathogenicity of human coronaviruses. The forces affecting synonymous variation in coronaviruses cannot be definitively identified by computational means alone; given the short period of SARS-CoV-2 pandemic and potential mutation of RNA viruses through transmission. However, characterization of coronavirus heterogeneity via complementary approaches provides an opportunity to further explore viral genome evolution, regulation and pathogenesis, and the fundamental mechanism of virus-host interaction.

## 5. Materials and methods

### 5.1. Data availability

A total of 3557 complete genome sequences assigned to the genera *Alphacoronavirus* and *Betacoronavirus* available on the Global Initiative on Sharing All Influenza Data (GISAID) and USA National Center for Biotechnology Information (NCBI) were clustered to 2522 genomes using a similarity threshold of 0.1 %. These sequences were globally aligned to check potential errors, and 5′- and 3′-UTR regions were trimmed. Among them, all sequences assigned to the subgenus *Sarbecovirus* (n = 114) and belonging to HCoV clusters (n = 355) were retained for further analysis. In addition, fifty-two reference sequences representing each coronavirus species within the family *Coronaviridae* were downloaded from NCBI public domain (Tables S1 and S2).

### 5.2. Phylogenetic analysis

The nucleotide sequences of each ORF were aligned using translation algorithm based on the aligned amino acid sequence matrix using MAFFT within Geneious Primer package. Maximal likelihood (ML) trees were constructed using RAxML MPI v8.2.12 (Stamatakis, 2006) with optimized parameters via CIPRES Science Gateway (Miller et al., 2010). Phylogenetic trees were constructed using MAFFT v7.402 (Katoh and Toh, 2010) inferred from the concatenated nucleotide sequence alignments of 6 open reading frames (ORFs) shared by all coronaviruses.

### 5.3. K-mer spectrum clustering

We chose $k = 3$ (a total of $4^3 = 64$ trimers) in consideration of the relative short size of viral genome. The trimer frequencies in percentage were normalized and a Kullback-Leibler (KL) distance matrix (Kullback, 1987) was calculated, based on which a hierarchical cluster analysis was performed for phylogenetic topology. We used *count* function in R’s package ‘seqinr’ (Charif and Lobry, 2007) to count the number of each trimer; and *hclust* function in R’s package ‘stats’ (Team, 2014) to construct the hierarchical tree.

### 5.4. Codon usage bias

The effective number of codons (ENC) statistic is a way of analysing how biased a gene is in terms of its codon usage (Wright, 1990). The ENC values range from 20 when a gene is effectively using only a single codon for each amino acid (strongest bias) to 61 when a gene trends to use all codons with equal frequency (no bias). The codonW package (http://codonw.sourceforge.net/) was used to calculate the ENC values. In order to examine the influence of GC content on codon usage, we plotted the relationship between ENC and GC3s (GC content at the synonymous third codon position). This was compared to the expected ENC* if GC content were solely responsible for the codon biases, calculated as (Shackelton et al., 2006; Wright, 1990):

$$ENC^{*} = 2 + GC3s + \frac{29}{GC3s^2 + (1 - GC3s)^2}$$

### 5.5. Relative synonymous codon usage (RSCU)

Relative Synonymous Codon Usage (RSCU) values for 59 codons (except for Met, Trp, and stop codons) were calculated by using the ratio of the observed frequency of codons relative to the expected frequency in the absence of usage bias to measure the extent of non-random usage of synonymous codons (Sharp et al., 1986). Given that all the synonymous codons for the same amino acids are used equally, the RSCU value would be 1. The value would be much less than 1 if a codon were used less frequently than expected and vice versa. The website tool CAIcal (http://genomes.urv.cat/CAIcal/) was used to calculate the RSCU values. The RSCU values in the host genomes served as the references, as retrieved from the Kazusa codon usage database (http://www.kazusa.or.jp/codon/) (Nakamura et al., 2000).

### 5.6. Measuring dinucleotide suppression

It is a well-known phenomenon that CpG dinucleotide is uncommon in most mammalian (including human) genomes, termed CpG suppression. To further understand the potential interplay between viruses and their hosts, we measured the genetic metrics including GC content, dinucleotide suppression ($\rho XY$) and CpG methylation of betacoronaviruses surveyed in this work. The GC content of a string of DNA/RNA is simply the fraction of the letters that are C plus those that are G. A measure of dinucleotide suppression was calculated as the observed frequency of the dinucleotide relative to the product of the frequencies of the individual nucleotides (Karlin and Mrazek, 1997). For example, the ratio

$$\rho CG = \frac{f_{CG}}{f_C * f_G}$$

where $f_{CG}$ represents the frequency of a dimer $CG$, $f_C$ and $f_G$ denote the probabilities of its constituent monomers, respectively. The dinucleotide

O/E ratio would be 1 if the occurrences of its individual nucleotide were independent, and the genome exhibits suppression if it has $\rho XY$ much less than 1.

### 5.7. Correspondence analysis

We used correspondence analysis to study the correlation between codon usage and coronavirus genomic composition. This analysis positions each coronavirus genome (row) and relative synonymous codon usage or dinucleotide suppression (column) to create a series of orthogonal axes to identify variation affecting codon usage bias, with each subsequent axis explaining a decreasing amount of the variation. Multidimensional scaling of the redundancy analysis (RDA), or optionally principal coordinate analysis (PCoA) was applied using *rda* function in R's package 'vegan' to generate two-dimensional representations for matrix 1 and 2, and visualized using the *biplot*. By definition, the axes are ordered according to the amount of variance in the data, with samples sharing similar variations clustering together. Differences in codon usage bias were assessed with permutational multivariate analysis of variance (PERMANOVA) using the *adonis2* function in R's package 'vegan'.

### 5.8. Statistical analysis

The significance of differences in the genetic metric measures was tested using phylogenetic generalized linear models (*pgls* function in R's package 'caper'), non-parameter Wilcoxon and Mann-Whitney *U* test test (implemented in R's package 'stats'), or a two-way analysis of variance (ANOVA). The Pearson's correlation coefficient was used to test the association between paired observations (*cor.test* in R's package 'stats'). All plotting and statistical comparisons were performed in R v3.0.2 (Team, 2014) using scripts developed in-house (available upon request).

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data and materials

The complete genome sequences of coronaviruses analysed in this study were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) and USA National Center for Biotechnology Information (NCBI) (see list in Table S1).

## Authors' contributions

ZC, data curation, formal analysis, writing-original draft and editing; SSB, formal analysis, writing-review and editing; MHW, formal analysis, writing-review and editing; RWYC, formal analysis, writing-review and editing; PKSC, supervision, writing-review and editing.

## Funding

The work was partially supported by funding from Health and Medical Research Fund (Project number COVID190107) (P.K.S.C.), Food and Health Bureau of the Hong Kong Special Administrative Region, China, and the Project Impact Enhancement Fund(Project number PIEF/Ph2/COVID/11) (Z.C.) from Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China.

## Declaration of Competing Interest

The authors declare that they have no competing interests. PC is not involved in the review of this manuscript.

## Acknowledgements

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database (https://www.epicov.org/) on which this research is based.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jviromet.2020.114032.

## References

Beutler, E., Gelbart, T., Han, J.H., Koziol, J.A., Beutler, B., 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc Natl Acad Sci U S A 86, 192–196.

Bulmer, M., 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325, 728–730.

Calderon, B.M., Danzy, S., Delima, G.K., Jacobs, N.T., Ganti, K., Hockman, M.R., Conn, G. L., Lowen, A.C., Steel, J., 2019. Dysregulation of M segment gene expression contributes to influenza A virus host restriction. PLoS Pathog. 15, e1007892.

Chan, C.X., Ragan, M.A., 2013. Next-generation phylogenomics. Biol. Direct 8, 3.

Charif, D., Lobry, J.R., 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. Structural Approaches to Sequence Evolution. Springer, pp. 207–232.

Chor, B., Horn, D., Goldman, N., Levy, Y., Massingham, T., 2009. Genomic DNA k-mer spectra: models and modalities. Genome Biol. 10, R108.

Cui, J., Li, F., Shi, Z.L., 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181–192.

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.J., Jiang, S., 2009. The spike protein of SARS-CoV–a target for vaccine and therapeutic development. Nat. Rev. Microbiol. 7, 226–236.

Du, L., Yang, Y., Zhou, Y., Lu, L., Li, F., Jiang, S., 2017. MERS-CoV spike protein: a key target for antivirals. Expert Opin. Ther. Targets 21, 131–143.

Forni, D., Cagliani, R., Clerici, M., Sironi, M., 2017. Molecular evolution of human coronavirus genomes. Trends Microbiol. 25, 35–48.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J.W., Barclay, W., Evans, D.J., 2000. Identification of a cis-acting replication element within the poliovirus coding region. J. Virol. 74, 4590–4600.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., Cao, B., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506.

Jiang, S., Hillyer, C., Du, L., 2020. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. Trends Immunol.

Karlin, S., Mrazek, J., 1997. Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci U S A 94, 10227–10232.

Katoh, K., Toh, H., 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26, 1899–1900.

Kullback, S., 1987. The kullback-leibler distance. Am. Stat. 41, 340-340.

Latanova, A.A., Petkov, S., Kilpelainen, A., Jansons, J., Latyshev, O.E., Kuzmenko, Y.V., Hinkula, J., Abakumov, M.A., Valuev-Elliston, V.T., Gomelsky, M., Karpov, V.L., Chiodi, F., Wahren, B., Logunov, D.Y., Starodubova, E.S., Isaguliants, M.G., 2018. Codon optimization and improved delivery/immunization regimen enhance the immune response against wild-type and drug-resistant HIV-1 reverse transcriptase, preserving its Th2-polarity. Sci. Rep. 8, 8078.

Lauring, A.S., Acevedo, A., Cooper, S.B., Andino, R., 2012. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe 12, 623–632.

Liddament, M.T., Brown, W.L., Schumacher, A.J., Harris, R.S., 2004. APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. Curr. Biol. 14, 1385–1391.

Lin, C.T., Tsai, Y.C., He, L., Calizo, R., Chou, H.H., Chang, T.C., Soong, Y.K., Hung, C.F., Lai, C.H., 2006. A DNA vaccine encoding a codon-optimized human papillomavirus type 16 E6 gene enhances CTL response and anti-tumor activity. J. Biomed. Sci.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., Ecker, J.R., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574.

Milewska, A., Kindler, E., Vkovski, P., Zeglen, S., Ochman, M., Thiel, V., Rajfur, Z., Pyrc, K., 2018. APOBEC3-mediated restriction of RNA virus replication. Sci. Rep. 8, 5960.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE) 2010, 1–8.

Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. 26, 3188–3193.

Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292.

Reusken, C.B., Haagmans, B.L., Muller, M.A., Gutierrez, C., Godeke, G.J., Meyer, B., Muth, D., Raj, V.S., Smits-De Vries, L., Corman, V.M., Drexler, J.F., Smits, S.L., El Tahir, Y.E., De Sousa, R., van Beek, J., Nowotny, N., van Maanen, K., Hidalgo-Hermoso, E., Bosch, B.J., Rottier, P., Osterhaus, A., Gortazar-Schmidt, C., Drosten, C., Koopmans, M.P., 2013. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. Lancet Infect. Dis. 13, 859–866.

Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–563.

Sharp, P.M., Tuohy, T.M., Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Stachyra, A., Redkiewicz, P., Kosson, P., Protasiuk, A., Gora-Sochacka, A., Kudla, G., Sirko, A., 2016. Codon optimization of antigen coding sequences improves the immune potential of DNA vaccines against avian influenza virus H5N1 in mice and chickens. Virol. J. 13, 143.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Steinberg, T., Ohlschlager, P., Sehr, P., Osen, W., Gissmann, L., 2005. Modification of HPV 16 E7 genes: correlation between the level of protein expression and CTL response after immunization of C57BL/6 mice. Vaccine 23, 1149–1157.

Team, R.C., 2014. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0.

Tian, X., Li, C., Huang, A., Xia, S., Lu, S., Shi, Z., Lu, L., Jiang, S., Yang, Z., Wu, Y., Ying, T., 2020. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. Emerg. Microbes Infect. 9, 382–385.

Upadhyay, M., Samal, J., Kandpal, M., Vasaikar, S., Biswas, B., Gomes, J., Vivekanandan, P., 2013. CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J. Virol. 87, 13816–13824.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison-a review. Bioinformatics 19, 513–523.

Wang, M., Yan, M., Xu, H., Liang, W., Kan, B., Zheng, B., Chen, H., Zheng, H., Xu, Y., Zhang, E., Wang, H., Ye, J., Li, G., Li, M., Cui, Z., Liu, Y.F., Guo, R.T., Liu, X.N., Zhan, L.H., Zhou, D.H., Zhao, A., Hai, R., Yu, D., Guan, Y., Xu, J., 2005. SARS-CoV infection in a restaurant from palm civet. Emerg Infect Dis 11, 1860–1865.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B. S., McLellan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science.

Wright, F., 1990. The' effective number of codons' used in a gene. Gene 87, 23–29.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D., Fouchier, R.A., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367, 1814–1820.

Zhao, K.N., Chen, J., 2011. Codon usage roles in human papillomavirus. Rev. Med. Virol. 21, 397–411.

Zhong, N.S., Zheng, B.J., Li, Y.M., Poon, XieZ.H., Chan, K.H., Li, P.H., Tan, S.Y., Chang, Q., Xie, J.P., Liu, X.Q., Xu, J., Li, D.X., Yuen, K.Y.Peiris, Guan, Y., 2003. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet 362, 1353–1358.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.