Briefings in Bioinformatics, 00(00), 2020, 1–17

doi: 10.1093/bib/bbaa383
Problem Solving Protocol

Evolutionary and codon usage preference insights into spike glycoprotein of SARS-CoV-2

Yashpal Singh Malik, Mohd Ikram Ansari, Jobin Jose Kattoor, Rahul Kaushik, Shubhankar Sircar, Anbazhagan Subbaiyan, Ruchi Tiwari, Kuldeep Dhama, Souvik Ghosh, Shailly Tomar and Kam Y. J. Zhang

Corresponding authors: Yashpal Singh Malik, Division of Biological Standardization, ICAR-Indian Veterinary Research Institute, Izatnagar 243122, Uttar Pradesh, India. Tel.: +91-5812302777; Fax: +91-5812301757. E-mail: malikyps@gmail.com

Abstract

Interaction of SARS-CoV-2 spike glycoprotein with the ACE2 cell receptor is very crucial for virus attachment to human cells. Selected mutations in SARS-CoV-2 S-protein are reported to strengthen its binding affinity to mammalian ACE2. The N501T mutation in SARS-CoV-2-CTD furnishes better support to hotspot 353 in comparison with SARS-CoV and shows higher affinity for receptor binding. Recombination analysis exhibited higher recombination events in SARS-CoV-2 strains, irrespective of their geographical origin or hosts. Investigation further supports a common origin among SARS-CoV-2 and its

Yashpal Singh Malik, M.V.Sc., Ph.D., serving as Dean, College of Animal Biotechnology, GADVASU, Ludhiana and formerly was ICAR-National Fellow and Professor. He is an expert on enteric viral infections, zoonosis and emerging viral diseases of animals and humans. He did postdoc from the University of Minnesota, USA and advanced training in molecular virology from University of Ottawa, Canada and Wuhan Institute of Virology, China. He is the Secretary General of Indian Virological Society and Secretary-E for the World Society for Virology (USA).
Mohd Ikram Ansari, M.Sc., PhD microbiology from Aligarh Muslim University India. He is presently working as a research associate in an ICAR-National Fellow Scheme at the ICAR Indian Veterinary Research Institute, India. He has done his postdoc from KAUST, Saudi Arabia and worked on molecular aspects of gene and plasmid in Technical University Berlin, Germany.
Jobin Jose Kattoor, M.V.Sc., Ph.D. from ICAR-Indian Veterinary Research Institute, India. He is currently working as a postdoc in Purdue University, USA.
Rahul Kaushik is a research scientist at Laboratory for Structural Bioinformatics, RIKEN Center for Biosystems Dynamics Research, Japan. He works in the area of complex biological functions of proteins, development of methods for protein structure prediction and applying design principles to create proteins with novel architectures, new biological functions or effective therapeutics.
Shubhankar Sircar PhD scholar received his Master's degree from Integral University, India and is presently serving as a senior research Fellow in an ICAR–National Fellow Scheme at the ICAR-Indian Veterinary Research Institute, India. He has received a few awards and recognitions (Best poster and Young Scientist).
Anbazhagan Subbaiyan is a PhD scholar in the division of microbiology at ICAR-Indian Veterinary Research Institute, Izatnagar, Uttar Pradesh, India.
Ruchi Tiwari is currently working as an assistant professor in the Department of Vet erinary Microbiology, DUVASU, Mathura, India. She is currently pursuing her PhD (Hons) degree from DUVASU. She has been honored with the Young Scientist Award, Best Paper Awards (10) and Outstanding Women Faculty Award (2019). She is serving as an Editor and Member, Editorial Board & Reviewer of 15 International Journals.
Kuldeep Dhama, M.V.Sc., Ph.D. (Gold Medalist), is working as a principal scientist in Division of Pathology, ICAR-Indian Veterinary Research Institute, Izatnagar, Uttar Pradesh, India. He has to his credit 600 publications, 06 books and 65 book chapters. Dr. Dhama has been recognized as an extremely productive researcher in the 'Nature' journal publication. He is National Academy of Agricultural Science (NAAS, India) associate, worked as a Nodal Officer, WTO and Member of Wildlife Health Specialist Group (IUCN).
Souvik Ghosh, M.VSc. and PhD, is an associate professor of infectious disease and director of One Health Center for Zoonoses and Tropical Veterinary Medicine, Basseterre, St. Kitts, West Indies. Dr. Ghosh has an interdisciplinary teaching and research experience, encompassing the fields of Veterinary and Medical Virology, Molecular Epidemiology, Viral genomics and Zoonosis/One-Health.
Shailly Tomar is a professor in the Department of Biotechnology, IIT, Roorkee, India. He is involved in the molecular and structural virology, antiviral research, discovery of structure-based antivirals against RNA arboviruses (Chikungunya), structural studies X-ray and Cryo Electron microscopy (CryoEM) of pathogenic virus/viral proteins.
Kam Y. J. Zhang is a team leader at Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, Japan. He works in the area of complex biological functions of proteins, development of methods for protein structure prediction and applying design principles to create proteins with novel architectures, new biological functions or effective therapeutics.
Submitted: 28 September 2020; **Received (in revised form):** 31 October 2020

predecessors, SARS-CoV and bat-SARS-like-CoV. The recombination events suggest a constant exchange of genetic material among the co-infecting viruses in possible reservoirs and human hosts before SARS-CoV-2 emerged. Furthermore, a comprehensive analysis of codon usage bias (CUB) in SARS-CoV-2 revealed significant CUB among the S-genes of different beta-coronaviruses governed majorly by natural selection and mutation pressure. Various indices of codon usage of S-genes helped in quantifying its adaptability in other animal hosts. These findings might help in identifying potential experimental animal models for investigating pathogenicity for drugs and vaccine development experiments.

Key words: COVID-19; SARS-CoV-2; S-protein; ACE2 receptor; recombination; codon usage analysis

Background

The first quarter of the 21st century has witnessed the outbreak of the major pathogenic human coronaviruses (CoV) that are believed to have crossed the species barriers and spill over in humans causing fatal pneumonia. Among them, severe acute respiratory syndrome (SARS)-CoV-2 has been very contagious exhibiting a high infection rate and mortality. The interaction of the virus spike glycoprotein (S-protein) anchored onto the CoV envelope with the host cell receptor, angiotensin-converting enzyme 2 (ACE2) consequences to the viral entry into human host. Considering the very crucial role of S-protein in SARS-CoV-2 infection to humans, we have performed a comprehensive study on S-protein to furnish evolutionary and codon usage insights. The current study comprehends the insights into the architecture of CoV genome, the configuration of S-protein, the potential recombination events among different strains of SARS-CoV-2, S-protein and hACE2 interactions mediated viral entry to human host, codon usage analysis of S-gene of SARS-CoV-2 and its benchmarking with nine different beta-coronaviruses, adaptability of S-gene to different potential hosts through codon and tRNA adaptation indices. The viruses use a controlled expression of viral proteins to achieve the replicative suitability. The means of attaining replicative suitability varies among viruses as some prefer optimized codon usage while others escape host immune system. The study may help in the identification of potential experimental animal model for investigating pathogenicity for drug and vaccine development experiments.

Introduction

The three major pathogenic human coronaviruses (CoVs) are the SARS-CoV, Middle East respiratory syndrome (MERS)-CoV and SARS-CoV-2 [1]. These CoVs have crossed the species barrier to cause fatal pneumonia in humans since the beginning of the 21st century: SARS-CoV in 2002 [2, 3], MERS-CoV in 2012 [4] and SARS-CoV-2 in late 2019 [5, 6]. SARS-CoV appeared in the Guangdong province of China and spread to five different continents through air travel routes, contaminating 8098 persons and causing 774 deaths. After a decade, MERS-CoV appeared in the Arabian Peninsula as a significant public health concern and spread to 27 countries, infecting 2494 persons and causing 858 deaths. A novel CoV, SARS-CoV-2, appeared in Wuhan, Hubei province of China, in December 2019, and it was sequenced and isolated in January 2020 [5, 7].

SARS-CoV-2 is related to progressive atypical pneumonia (COVID-19) that has infected more than 24 million individuals and caused nearly 0.85 million deaths in more than 215 nations as of now. The World Health Organization declared SARS-CoV-2 infection as a pandemic and public health emergency of international concern on 11 March 2020. The closely related viruses, SARS-CoV and SARS-CoV-2, probably originated from bats, and most likely, bats act as the reservoir hosts for them [7–9]. Though palm civets and raccoon dogs have been documented as intermediate hosts for zoonotic SARS-CoV transmission among bats and people [10, 11], the intermediate host for SARS-CoV-2 remains unknown. It was proposed that MERS-CoV also originated from bats, but dromedary camels were considered the reservoir hosts, fueling the spillover of the virus to humans [12, 13].

The CoV infection in humans mostly occurs due to the interaction of the virus spike glycoprotein (S-protein), anchored onto the CoV envelope with the host cell receptor, angiotensin-converting enzyme 2 (ACE2) receptor. The S-protein consists of two subunits, S1, as the receptor-binding domain (RBD), and S2, involved in the fusion of the cell and virus membrane [14, 15]. S-protein is cleaved from the host at the S2 site, located upstream of the fusion protein, by proteases [16, 17]. This cleavage activates membrane fusion protein through extensive irreversible conformational changes [17–19]. Thus, CoV entry into the host cell is a complex process, which requires receptor-binding and proteolytic processing of the S-protein [14]. The affinity of SARS-CoV-2 S-RBD to ACE2 is approximately ten times higher than that of SARS-CoV RBD, suggesting that ACE2, on the host cell, is the specific receptor that binds with the virus [20]. In this article, we explained the genome structure of CoV, with special attention to the S-protein and its role in CoV entry into the host cell, its codon usage bias analysis, and the current trends in antibody production for COVID-19 treatment.

Methods and results

Structure of the CoV genome

The CoV genome is a positive-sense single-stranded RNA of ∼30 kb, and it is 5′ capped and 3′ polyadenylated. The size of CoVs ranges from 100 to 160 μm in diameter, and exceptionally enormous (20 nm in size), intensely glycosylated homotrimeric spikes (S) of ∼200 kDa form the virus envelope, representing a crown. The viral RNA genome is embodied in a helical nucleocapsid phosphoprotein (N), also known as ribonucleoprotein, and wrapped into a virus particle with a membrane (M) and envelope (E) glycoproteins (as depicted in Figure 1). Another protein, the hemagglutinin-acetyl esterase (HE) glycoprotein, is found in some of the CoVs (beta-CoV 2a) and toroviruses. It plays a role in the attachment and release of virus progenies [21, 22]. The poly (A) tail permits CoVs to translate their gene products directly after infection without requiring an intermediate translation stage. Transcription initiation in CoVs is controlled by different types of consensus transcription regulating sequences (TRS), TRS1-L, TRS2-L, 5'-CUAAAC-3′ and 5'-ACGAAC-3′, and these are merged into TRS3-L and 5'-CUAAACGAAC-3′. These different TRS offer sites for subgenomic polycistronic mRNAs to encode accessory, structural and non-structural proteins. In MERS-CoV, transcription mainly begins in TRS2-L, whereas in other CoVs, transcription occurs indistinguishably in all TRS [23, 24].

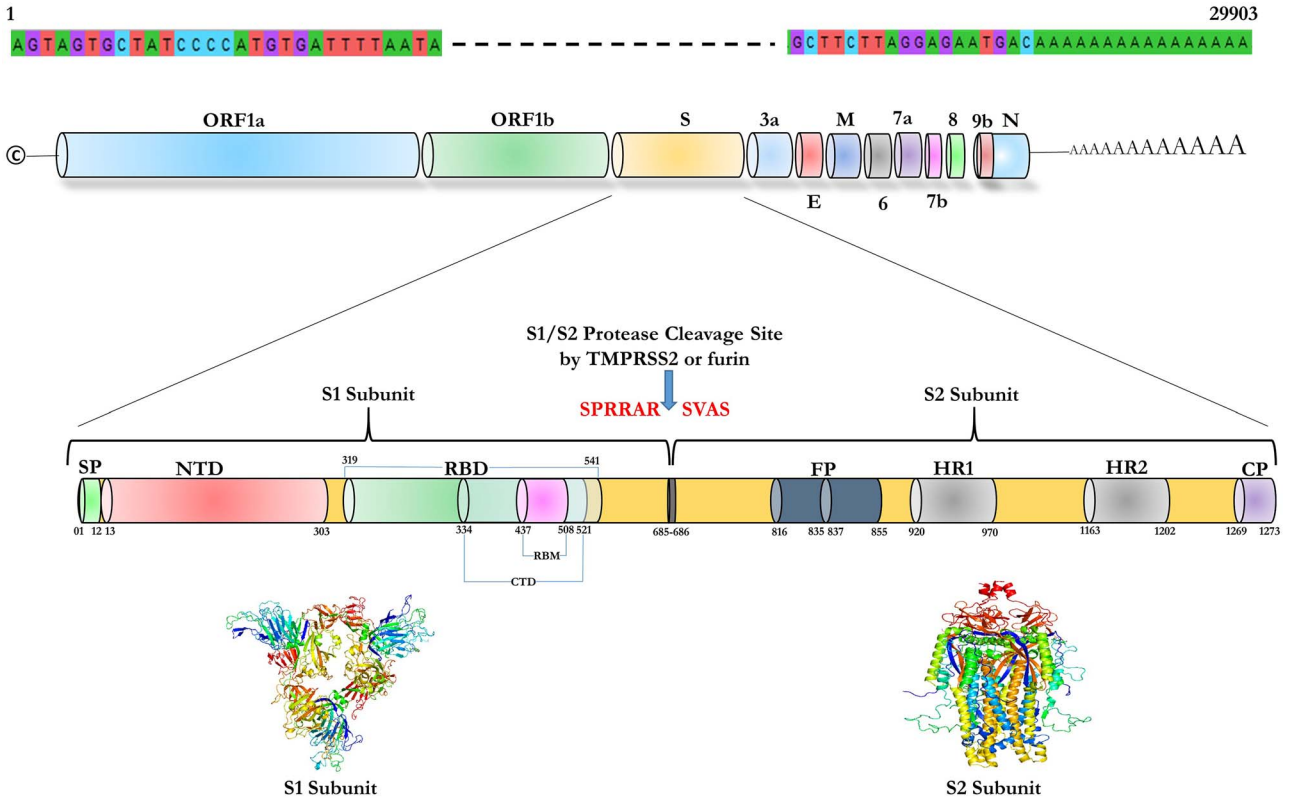


Figure 1. Genome organization and functional domains of SARS-CoV-2 based on the reference genome of strain Wuhan-Hu-1 (Accession no. NC_045512). The genome of SARS-CoV-2 consists of two large genes encoding ORF1a and ORF1b. Apart from these two, there are structural genes spike (S), envelope (E), membrane (M) and nucleocapsid (N), which encodes structural proteins. There are few accessory genes along with the structural genes. The structure of S-protein has been expanded below the genome organization from protein ID: YP_009724390.1 of the reference strain. The S-protein consists of S1 and S2 subunits. The numbers depicted are residues that represent different functional regions in the S-protein. A furin cleavage site between position numbers 685–686 is also described. Below are the ribbon diagrams for S-protein subunits S1 & S2. SP: Signal peptide; NTD: N terminal domain; RBD: receptor binding domain; RBM: receptor binding motif; CTD: C terminal domain; FP: fusion; HR: heptad repeats; CP: cytoplasmic domain.

S-protein configuration

The functional surface S-protein is a homotrimer, and it possesses two subunits, S1 and S2, which are associated with host receptor binding and virus-cell membrane fusion, respectively [14, 15]. The S1 subunit of the S-protein possesses the C-terminal domain (CTD), N-terminal domain (NTD), and two subdomains (SD1 and SD2), and the ACE2 receptor-binding domain (RBD) lies within the CTD [25] (as depicted in Figure 1). Reports have suggested that RBD experiences a conformational change from a stable closed state to a progressively less favorable partly open state [15, 26, 27]. In the closed state, the determinants are submerged and unreachable to the receptors. In contrast, in the partially open state, they are uncovered and essential for the interaction with host cells [28]. In SARS-CoV and SARS-CoV-2, the S-glycoprotein is characteristically found in both the closed and open states, and this property has been reported to exist in the most pathogenic CoVs [14, 29]. Although the partially open RBD plays a significant role in the infection of human cells [30], few studies have reported on this protein conformation, which occurs at the residue level and can play a vital role in the prevention and treatment of the infection.

The S1 subunit is composed of the NTD and three CTDs (CTD1, CTD2 and CTD3). The CTD1 (320–516 residues for SARS-CoV and 333–529 residues for SARS-CoV-2) functions as the RBD, which allows SARS-CoV-2 to bind to the peptidase domain of ACE2 [25, 31]. The recognition step is a unique procedure, during which the protein surfaces of both accomplices adjust to each another [27]. Most of the residues that separate human S-glycoproteins from bat or pangolin S-glycoproteins are present in the S1 subunit (as depicted in Figure 1). All other residues (including residue 797 and cysteine residues C822–C833) are located outside the S2 subunit crown, and consequently, interact with the receptor on the host cell membrane. Remarkably, canonical ACE2 binding residues, conserved in the pangolin and human lineage b strains of SARS-CoV-2, do not seem to be under severe selection pressure [32]. These mutations may allow host adaption, consequently leaving the question of how and when these residues may have developed unanswered.

S-protein-based entry of SARS-CoV-2 through ACE2

With the advent of next-generation sequencing, many whole-genome sequences of SARS-CoV-2 are being added to the public database daily [33]. A phylogenetic analysis of SARS-CoV-2 revealed its close similarity to SARS-CoV [33, 34]. Xu [35] used computer-based homology modeling to report that the RBDs of the SARS-CoV and SARS-CoV-2 S-proteins have nearly the same 3D structures that are maintained by van der Waals forces. Biochemical interaction studies and crystal structure analysis showed that the SARS-CoV S-protein binds to human ACE2 (hACE2) with a strong affinity [36]. SARS-CoV and SARS-CoV-2 S-proteins show significant homology with 76.5% amino acid

sequence identity [35]. There are specific crucial amino acid residues, which play a critical role in the more efficient binding of the SARS-CoV-2 S-protein to the hACE2 receptor. The Lys31 on the hACE2 receptor is typically identified by a glutamine residue, which is present at the 493rd position in SARS-CoV-2 and the 479th position in SARS-CoV [34, 37]. Atomic details at the binding interface exhibit that key residue substitutions (N501T mutation) in the SARS-CoV-2 CTD, compared to the SARS-CoV RBD, reinforce this interaction and lead to receptor binding with higher affinity [38]. The asparagine (Asn) residue at the N501 position provides more support to the hot spot 353 in SARS-CoV-2 than to the hot spots S487 or T487 in SARS-CoV [34]. The other three residues of SARS-CoV-2 RBD are Leu455, Phe486 and Ser494. The first two residues support the hot spot Lys31, whereas the third residue supports the hot spot Lys353. *In silico* analysis of the 3D) structure of SARS-CoV-2 S-glycoprotein by homology modeling displayed that residue mutation Q483V [32] in the CTD1 domain allowed RBD to bind to the ACE2 receptor with more affinity. In this way, the SARS-CoV-2 S-protein is considered to bind with hACE2 with strong affinity.

For predicting the host tropism, we analyzed the major ACE2 residues (K31, E35, D38, M82 and K353) involved in the recognition of S-protein of SARS-CoV-2, as shown in Figure 2. Apart from hACE2, orthologues of other ACE2 in different animal species (species) were analyzed by homology modeling software (*SWISS-MODEL*) [39]. The distance between contact residues measured with PyMoL 2.3.4 software [40], and also protein–protein docking, was performed with HawkDock server [41]. SARS-CoV-2 virus has been reported recently in dogs, cats, tiger and mink. We selected a few representative species like dog, cat, bovine, pigs, chicken, rat, mouse (which co-exist near/in human habitats), African green monkey, orangutan, tiger, mink, ferret, pangolin, palm civet and horseshoe bat. The findings showed that apart from humans, animal species like African green monkey, orangutan, dog, cat, tiger, cattle and pig exhibit the key residues, which are responsible in the interaction of S-protein with ACE2 receptors (Table 1), thus making these species more susceptible host for SARS-CoV-2 virus attachment. A score (+/−) for the binding capacity of S-protein with ACE2 of human and different animal species is provided in Table 1. Overall, the observation indicates that S-protein of SARS-CoV-2 can bind to ACE2 from some wild (tiger, mink, ferret, pangolin, palm civet) as well as companion species like dogs and cats. These species further needs to be investigated to ascertain whether they can serve as an intermediate host for SARS-CoV-2 infection.

Recombination analysis of the SARS-CoV-2 S-protein

Viruses are well known for their diversity, generating recombination mechanisms. RNA viruses can adapt to their hosts because they mutate faster than their counterpart DNA viruses. Homologous recombination was first identified by Hirst [42] in polioviruses. It was later identified in other families of viruses, including *Coronaviridae* [43]. Favorable mutations accumulate during errors in genome replication, allowing viruses to adapt to different environmental selection pressures [44]. Most of the recombination programs are required to define two non-recombinant reference strains to identify a possible new recombined strain. However, RDP4 utilizes a fast and powerful heuristic approach, which sequentially tests different combinations of three different sequences for possible recombination events [45].

In the current study, we analyzed the possibility of recombination events in the newly emerged SARS-CoV-2 using RDP4 because similar recombination events occurred in SARS-CoV-1 during its emergence. The sequences used in the recombination analysis were retrieved from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/). Representative whole-genome sequences of SARS CoV-2, available in different countries, were acquired along with the sequences of the most similar bat CoV RaTG13 strain (96.2% similarity at the nucleotide level) and a bat SARS-CoV-like virus (bat-SL-CoVZXC21) and the reference sequences of MERS-CoV and SARS-CoV-1 (Supplementary Table S1). The whole-genome sequences were trimmed down to a length that contained only the CDS of the S-gene. The trimmed S-gene region sequences were aligned using the Clustal W program in MEGA software (version 6). Aligned sequences were checked for all possible recombination events, parental strains, and recombination breakpoints using default settings in the recombination detection program (RDP) software (version 4.100) and by all available methods, including RDP, GENECONV, MaxChi, BootScan and SiScan. To obtain a statistically reliable result, RDP was performed at a *P*-value of 0.05 in more than two methods. Recombination detection was executed with none of the sequences set as non-recombinant references to achieve an unbiased result.

In the UPGMA tree, obtained from the recombinant region that contained the segments taken from both the major and minor parents, all the sequences of SARS-CoV-2 and the RaTG13 strain of the bat CoV fell into the same clade, irrespective of the country of isolation (Figure 3). At the *P*-value of 0.05, all the representative SARS-CoV-2 S-genes, which acquired the gene segments from a major parent [bat SARS-like CoV (MG773924)] and a minor parent [SARS-CoV-1 strain from Canada (NC_004718)], showed similar recombination. The bat CoV (RaTG13 strain, MN996532) with the highest nucleotide level similarity to SARS-CoV-2 also showed recombination in the S-gene alignment nearly identical to that in SARS-CoV-2. The major parent (MG772934, bat SARS-like-CoV bat-SL-CoVZXC21) was, in turn, found to be a recombinant of the minor parent (NC_004718, SARS-CoV-1, Canada), and no possible recombination events were detected in MERS-CoV (Figure 3).

In the breakpoint analysis, a single recombinant tract, spanning from region 2051 to 2334 in the aligned bases, was observed in the SARS-CoV-2 S-gene sequences at a confidence interval of 99%. In contrast, the area involved in recombination, extending from residue 1769 to 2321, was observed in the RaTG13 strain (as shown in Supplementary Figure S1).

Codon usage analysis of the beta-coronaviruses S-gene

Codon Usage Bias (CUB) is generally observed among different organisms and mainly governed through specific selection pressures such as natural selection, mutation pressure and nucleotide compositional constraints [46, 47]. The codon usage bias may be analyzed in two different ways, viz. protein level where preferential amino acid usage is considered, or synonymous codon usage where preferences of codon usage for the same amino acid are considered [48, 49]. The selective pressure from the host cells is believed to be a critical driving force for synonymous codon usage bias in viruses. Various studies demonstrated the significant impact of selection pressure in the evolution of preferential viral codon usage [48–51]. It is observed that the preferred codons in an organism are translated more efficiently than the non-preferred codons [46, 47]. As the viruses are obligate intracellular parasites and utilize host cell mechanisms for their gene expression, a comprehensive codon usage analysis (CUA) of viral genes may

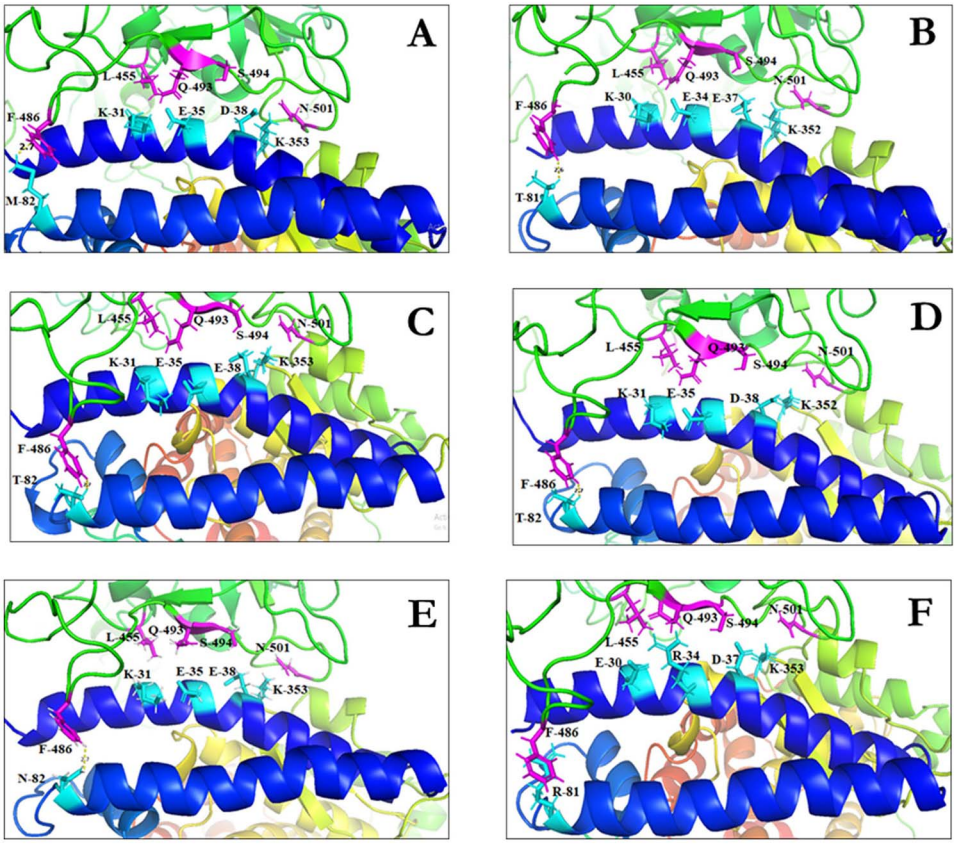


Figure 2. Homology-based model structures of SARS-CoV-2 spike protein with ACE2 from human, dog, tiger and cattle. (**A**) Human ACE2-'S' protein of SARS-CoV-2, (**B**) Dog ACE2-'S' protein of SARS-CoV-2, (**C**) Tiger ACE2-'S' protein of SARS-CoV-2, (**D**) Bovine ACE2-'S' protein of SARS-CoV-2, (**E**) Pangolin ACE2-'S' protein of SARS-CoV-2, (**F**) Chicken ACE2-'S' protein of SARS-CoV-2. The homology modeling analysis was carried out with the template of the SARS-CoV-Human ACE2 crystal structure using the PDB ID: 6ACG. All the ACE2 receptors shown with blue color and SARS-CoV-2 spike protein binding chain is shown in green color. The five critical contact residues in ACE2 and SARS-CoV-2 are depicted in the figure. Distance between F486 SARS-CoV-2 residue and ACE2 is shown in yellow color.

Table 1. A summary of key residues in the ACE2 receptor of human and different animal species that interact with S-protein of SARS-CoV-2. Original positions are given in the bracket, and aligned position is provided in the first row (ACE2). Color coding has been done for identical residues (KEDMK) based on the human ACE2 hotspots where K(31) is depicted by red, E by green, D by yellow, M by blue and K(353) by orange

S. N.	Species_ACE2	31	35	38	82	353	Binding
Affinity							
1	Human	K	E	D	M	K	+
2	Dog	K(30)	E(34)	E(37)	T(81)	K(352)	+
3	Cat	K	E	E	T	K	+
4	Bovine	K	E	D	T	K(352)	+
5	Pig	K	E	D	T	K	+
6	Chicken	E(30)	R (34)	D(37)	R(81)	K(353)	–
7	Rat	K	E	D	N	H	–
8	Mouse	N	E	D	S	H	–
9	African green monkey	K	E	D	M	K	+
10	Orangutan	K	E	D	M	K	+
11	Tiger	K	E	E	T	K	+
12	Ferret	K	E	E	T	K	+
13	Pangolin	K	E	E	N	K	+
14	Palm Civet	T	E	E	T	K	
15	Horseshoe bat	K	K	D	N	K	+

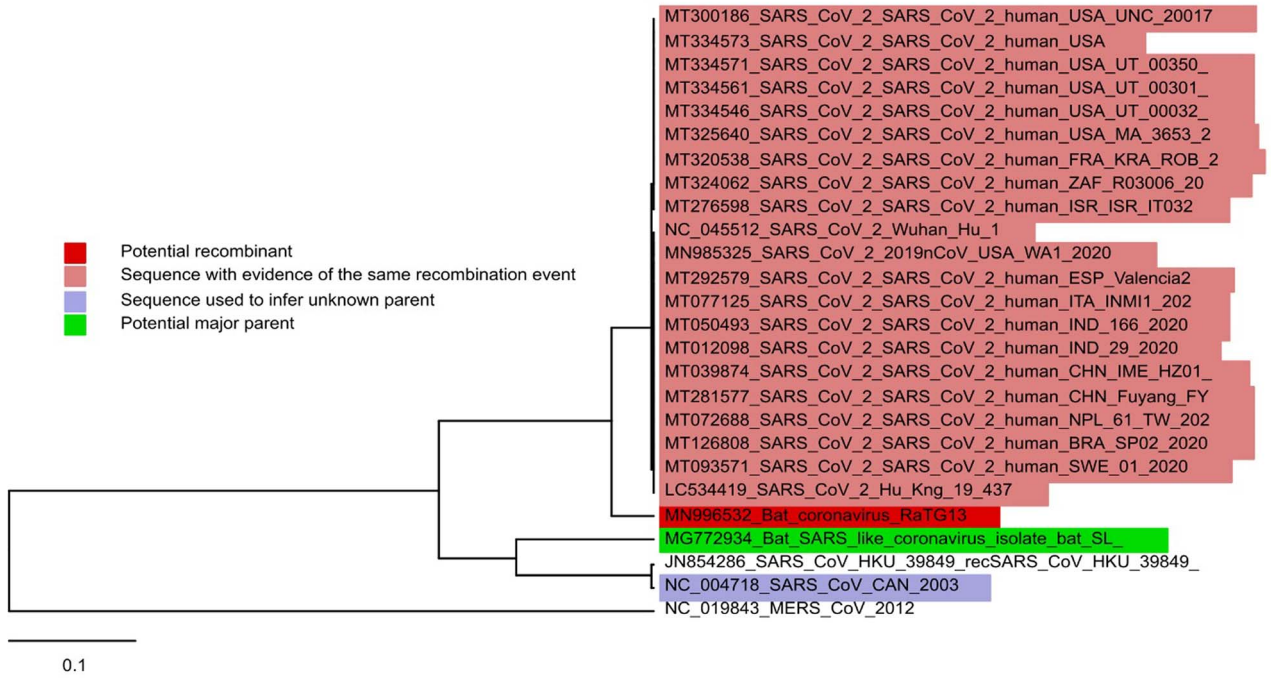


Figure 3. UPGMA tree derived from the minor parent (NC_004718). The probable recombinant is highlighted in red, whereas all other sequences that show similar recombination events are highlighted in pink. The parent strains inferred for identifying recombination are highlighted in green (major parent) and ash (unknown/minor parent).

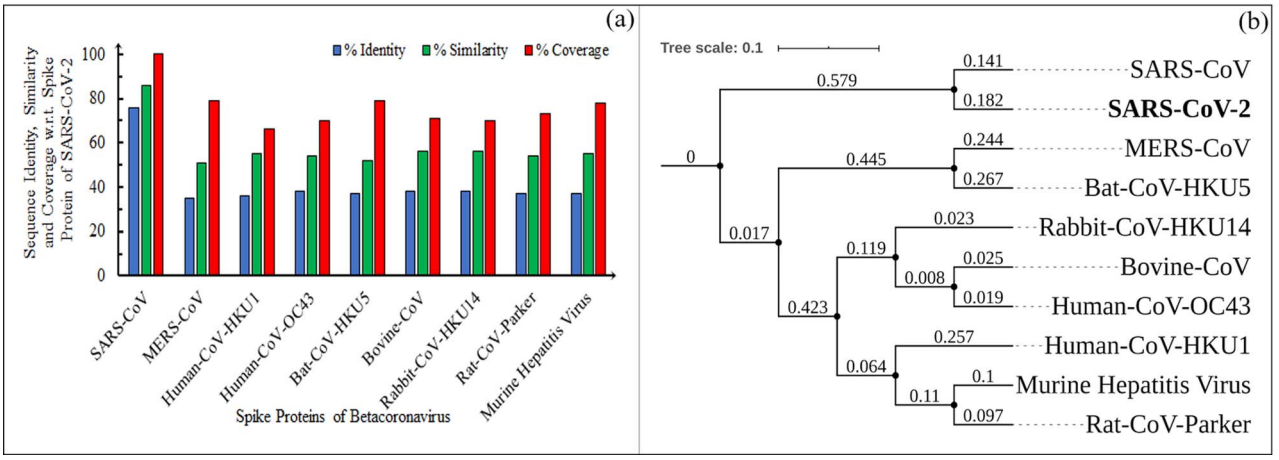


Figure 4. A comparison of S-gene (proteins and nucleotide) sequences of beta-coronaviruses. (**A**). Pairwise percent protein sequence identity, similarity and alignment coverage of spike protein of SARS-CoV-2 with other beta-coronaviruses are shown. (**B**) A phylogenetic tree of S-gene of beta-coronaviruses is depicted to show the relative relatedness among S-genes.

provide its fitness to the primary host and reservoirs. The origin and natural history of SARS-CoV-2 is not yet well established, and a comprehensive CUA might be useful in providing novel insights.

Gene and protein sequence level analysis

The gene sequences for S (spike)-gene of 10 different beta-coronaviruses along with corresponding protein sequences were extracted from NCBI web resource. The RefSeq identifier, protein identifier and primary host information for all the beta-coronaviruses considered in the study are provided in Table S2. The virus–host information is adopted from Virus-Host Database [52]. A pairwise alignment of Spike protein of SARS-CoV-2 was performed with the Spike proteins of all other beta-coronaviruses (as accounted in Table S2), and percentage sequence identity, similarity, and coverage of alignments are computed as shown in Figure 4A. It is observed that the Spike protein of SARS-CoV-2 shares a very high sequence identity (76%), sequence similarity (86%) and coverage of pairwise alignment (100%) with the Spike protein of SARS-CoV, while the Spike protein of all other beta-coronaviruses has relatively low sequence identity (35–38%), low sequence similarity (51–56%) and low coverage of pairwise alignment (66–79%). Furthermore, a multiple sequence alignment of all the S-gene sequences from the ten beta-coronaviruses was performed by using the multiple sequence alignment tool, MUSCLE [53]. The multiple sequence

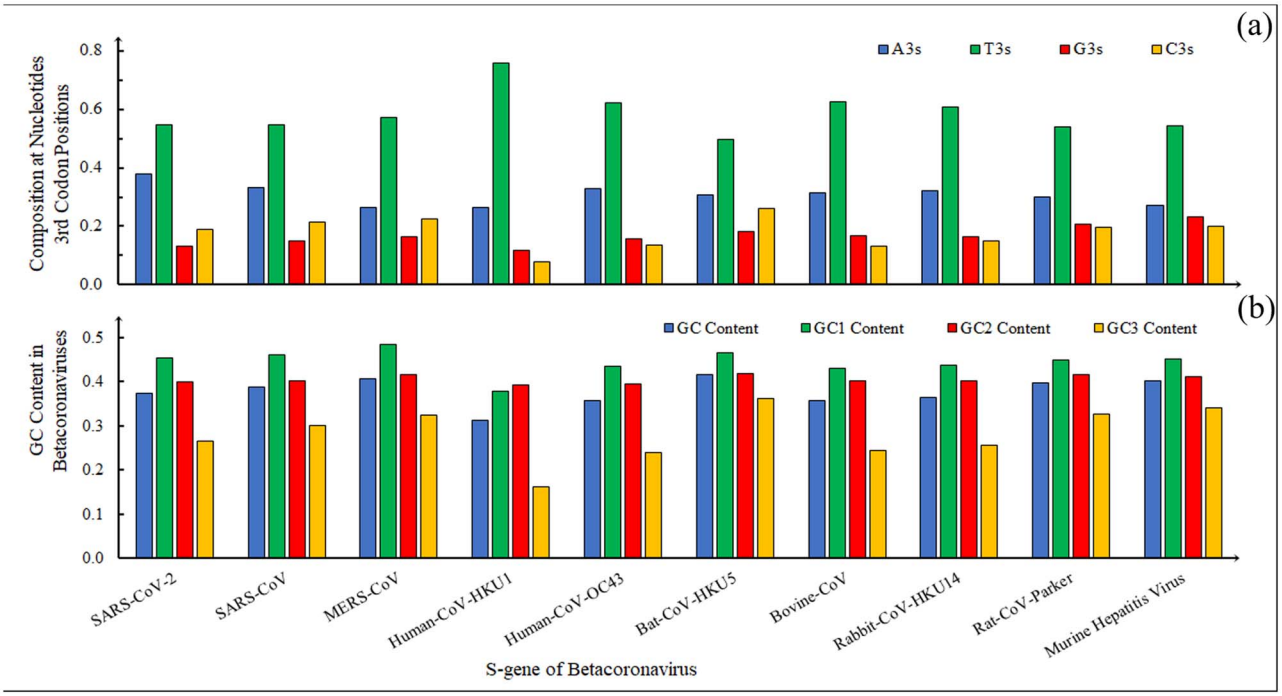


Figure 5. A comparison of composition analysis of the S-gene sequence of SARS-CoV-2 with other beta-coronaviruses. (**A**) Comparison of compositions of different nucleotides in the S-gene sequence of SARS-CoV-2 at the third synonymous codon position with other beta-coronaviruses. (**B**) A comparison of overall GC content and GC content at first, second and third codon positions of S-gene sequence of SARS-CoV-2 with other beta-coronaviruses.

alignment was further used for the phylogenetic tree (tree scale = 0.1) analysis, as depicted in Figure 4B.

Composition analysis of S-gene sequences

The nucleotide composition at the third synonymous codon position (T3s, C3s, A3s, G3s), overall GC content and the GC content at the first, second and third codon positions (G1, G2, G3) were calculated for all S-gene nucleotide sequences. The analysis of S-gene sequences in ten different beta-coronaviruses was performed to explore any compositional bias in the S-gene of SARS-CoV2. It is observed that the composition of adenine at the third synonymous codon position (A3s) is the highest for S-gene of SARS-CoV-2 as compared to other beta-coronaviruses considered here. In addition, the composition of purines at third synonymous codon positions (G3s and C3s) is observed to be very low as compared to the composition of pyrimidines at third synonymous codon positions (A3s and T3s). It is clear that the T3s content is noticeably high, and G3s are the lowest when compared to other nucleotide contents in the third position. For instance, in S-gene of SARS-CoV-2, the purine composition (G3s+C3s) is 0.32 as compared to 0.93 of the pyrimidines (A3s+T3s). Despite a very high protein sequence similarity (86%) in spike proteins (product of S-gene) among SARS-CoV-2 and SARS, the G3s+C3s of SARS-CoV is considerably higher than SARS-CoV-2 (0.36 for SARS-CoV and 0.32 for SARS-CoV-2). The difference in contents of different nucleotides for synonymous codon positions indicates a preferential codon usage among S-genes of SARS-CoV-2 and SARS-CoV. The comparison of compositional bias in the S-gene sequence of SARS-CoV-2 and other beta-coronaviruses is shown in Figure 5.

Effective number of codons and ENC-GC3s plot analysis

The effective number of codons (ENCs) reflects a simplistic measure of codon usage, which ranges from 20 to 61. A lower value of ENCs (i.e. ENCs $<$ 40) specifies a strong codon usage bias, while an ENCs value of 20 signifies that only single codon is effectively used for each amino acid, and a value of 61 reflects that all synonymous codons contributed equally in coding corresponding amino acids [54, 55]. The ENC value for S-gene of SARS-CoV-2 is 44.16, which is slightly on a lower side as compared to S-gene of SARS-CoV (ENC = 45.73) and MERS-CoV (ENC = 45.97). Overall, the ENC of S-gene of SARS-CoV-2 is the third lowest among the S-genes of 10 beta-coronaviruses, superseding Human-CoV-HKU1 (ENC = 32.77) and bovine-CoV (ENC = 43.86). Since the ENC value for S-gene of SARS-CoV-2 is closer to the cut-off for strong codon bias (ENC $<$ 40), it may be inferred that there is some extent of codon bias, which is higher than S-genes of SARS-CoV and MERS-CoV. The percent GC content at third synonymous codon position (%GC3s) for S-gene of SARS-CoV-2 (GC3s = 25.2%) is also less than its counterparts SARS-CoV (GC3s = 28.3%) and MERS-CoV (GC3s = 30.8%).

The effect of mutation pressure and natural selection on codon usage of a gene is commonly estimated with the help of ENC plots [56, 57]. The observed and expected distribution of GC3s for a gene is compared, where the ENC (expected) may be calculated using equation (1).

$$ENC(expected) = 2 + S + \frac{29}{S^2 + \left(1 - S^2\right)} \quad (1)$$

For different fraction GC content (S), varying from 0 to 1, the values for ENC (expected) were calculated. The data point for ENC (observed) and GC3s for S-genes were plotted with a standard curve for ENC (expected). The S-gene data points close to the standard curve of ENC (expected) suggest mutational pressure as one of the key factors in determining its codon usage bias. In contrast, the S-gene data points distant to the

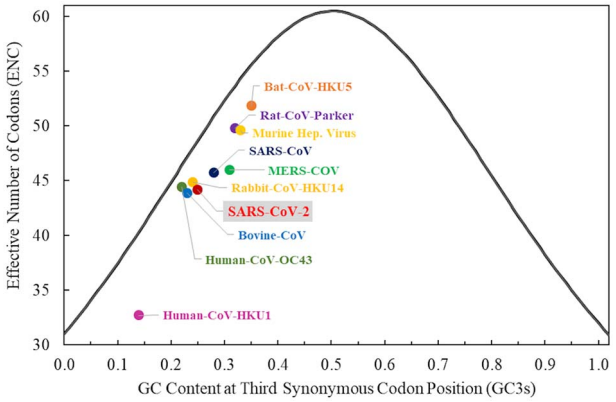


Figure 6. ENC-GC3s plot for S-genes of different beta-coronaviruses. The standard curve for expected ENC for different GC composition levels at the third synonymous codon position is shown as a bell-shaped line. The S-genes of all the beta-coronaviruses that fall close to the standard curve at different distances indicate the influence of mutational pressure in codon usage bias.

standard curve of ENC (expected) reflects the role of natural selection along with other factors in dictating the codon usage bias [48, 51, 58–60]. To further investigate the codon usage bias in S-gene of SARS-CoV-2 and its benchmarking with S-genes of other beta-coronaviruses, the ENC-GC3s plot derived from S-genes is superposed over the ENC (expected)-GC3s curve. The values for ENC(expected) were calculated using equation (1) at different GC3s contents. The ENC-GC3s data points were labeled in accordance with the beta-coronaviruses, as shown in Figure 6. In this figure, the curve denotes the ENC(expected) with the only factor of mutation pressure. The ENC-GC3s data points for all the beta-coronaviruses are close to the standard curve of ENC(expected), which indicates that the codon bias for S-gene is significantly influenced by mutational pressure along with other factors [61, 62]. From the ENC-GC3s plot, it may also be inferred that there is higher influence of mutational pressure in codon bias in S-gene of SARS-CoV-2 ($\Delta ENC_{(GC3s = 0.252)} = 4.64$) as compared to SARS-CoV ($\Delta ENC_{(GC3s = 0.283)} = 5.36$) and MERS-CoV ($\Delta ENC_{(GC3s = 0.308)} = 6.88$), where $\Delta ENC_{(GC3s = n)}$ is deviation in ENC value of S-gene from ENC(expected) at a given value (n) of GC3s content. A lower value of deviation is an indicator of more influence of mutational pressure than other factors and vice versa [59, 63–65]. The GC3s, ENC(observed), ENC(expected) and ΔENC for S-genes of all beta-coronaviruses are provided in Table S3. The ENC(observed) values for S-gene of beta-coronaviruses is found to correlate very highly with overall GC content ($r = 0.92$), GC1 content ($r = 0.81$), GC2 content ($r = 0.78$) and GC3 content ($r = 0.94$).

Relative synonymous CUA

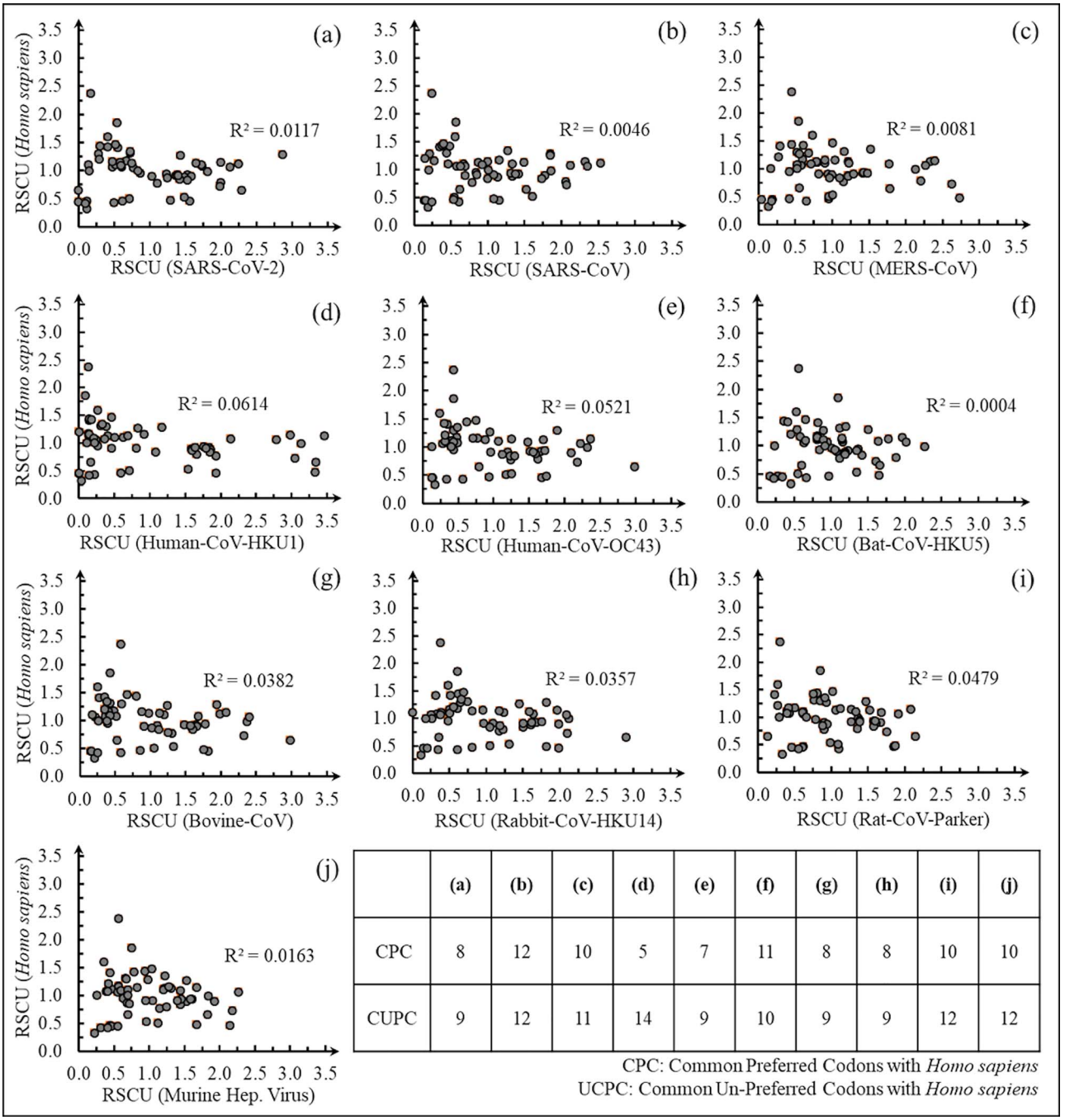
To explore the codon usage, the Relative Synonymous Codon Usage (RSCU) calculation was performed for the estimation of synonymous codon usage for each codon. The RSCU is a ratio of the over-served frequency of a codon to the expected frequency of the same codon and may be interpreted as RSCU (+ve bias) > 1, RSCU (no bias) = 1 and RSCU (−ve bias) < 1. It is worth mentioning that the amino acid residues, which are coded by single codon (Methionine and Tryptophan) and termination codons, are not included in the codon-wise RSCU analysis [56, 59, 60, 66].

The RSCU analysis was performed to compare RSCU values of S-gene of SARS-CoV-2 with other beta-coronaviruses. A summary of RSCU values for 59 codons (excluding UGG, AUG, UAG, UAA and UGA) for different beta-coronaviruses is provided in Supplementary Table S4. The Pearson correlation coefficient for RSCU values of S-gene of SARS-CoV-2 with S-genes of other beta-coronaviruses is calculated to estimate the degree of relatedness among them. It is observed that the RSCU values are highly correlated to S-gene of SARS-CoV ($r = 0.90$), followed by S-genes of Bovine-CoV ($r = 0.87$), Rabbit-CoV-HKU14 ($r = 0.86$), Human-CoV-OC43 ($r = 0.86$), Bat-CoV-HKU5 ($r = 0.84$), Human-CoV-HKU1 ($r = 0.79$), Rat-CoV-Parker ($r = 0.77$), Murine hepatitis virus ($r = 0.71$) and MERS-CoV ($r = 0.71$). For S-gene of SARS-CoV-2, out of 59 synonymous codons, 27 are identified as preferred codons, having RSCU > 1. In SARS-CoV, 28 codons are identified as preferred codons, while in MERS-CoV there are 24 preferred codons.

A summary of the number of preferred codons, number of preferred codons ending with A/U, number of unpreferred codons ending with A/U, number of preferred codons ending with G/C, number of un-preferred codons ending with G/C is provided in Supplementary Table S5. To further investigate the RSCU values of S-genes of different beta-coronaviruses, these values were compared with those of *Homo sapiens*. The RSCU values of all the hosts (apart from *H. sapiens*) are provided in Supplementary Table S6, which includes *Camelus dromedarius*, *Pipistrellus abramus*, *Bos taurus*, *Oryctolagus cuniculus*, *Rattus norvegicus* and *Mus musculus*. To further investigate the RSCU values of S-genes of beta-coronaviruses, a correlation analysis of RCSU values for S-genes with those of *H. sapiens* was performed (Figure 7). The embedded table in Figure 7 accounts for the preferred codons (RSCU ≥ 1) of S-genes of different beta-coronaviruses, which are also preferred in *H. sapiens*. Likewise, it also takes account of the unpreferred codon (RSCU < 1) of S-genes of different beta-coronaviruses, which are also unpreferred in *H. sapiens*. A higher number of common preferred and unpreferred codons among virus and its host indicates the better suitability of host for the virus [50, 64, 67–70].

Codon adaptation index and tRNA adaptation index analysis

A relative adaptation of a gene for codon usage of its host can be estimated with the help of the Codon Adaptation Index (CAI). The CAI of a gene for its host varies from 0 to 1 where a higher value reflects the usage of most abundant codons [59, 65, 71, 72]. It is a widely accepted measure for quantification of similarities between codon usage of a gene and a reference dataset to explore the synonymous codon usage for nucleotide sequences. Additionally, the CAI has also been used in the approximation of gene expressivity, for identifying the factors governing synonymous codon usage bias at the genome level, and for investigating horizontal genes transfer [65, 66, 73, 74]. The CAI for S-genes of different beta-coronaviruses was calculated using CAIcal (http://genomes.urv.es/CAIcal), which utilizes pre-compiled codon usage of host organisms [75]. The CAI values were calculated using the respective primary hosts (as shown in Table S2) and *H. sapiens* (if *H. sapiens* is not the reported primary host). The CAI for S-gene of SARS-CoV-2 is computed for five different hosts, viz. *H. sapiens, R. norvegicus, M. musculus, B. taurus* and *O. cuniculus*. The codon usage statistics of these hosts were adopted from the Codon Usage Database [76]. The CAI values are observed to be close to 0.70 with the codon usage statistics of *H. sapiens* for all the beta-coronaviruses, which indicate that the S-gene is well adapted to *H. sapiens* as a host. Likewise, in the comparison of ENC (observed) for S-genes and CAI for different hosts, it is observed that the CAI values of S-gene for *O. cuniculus* correlates well with ENC (observed) of S-gene ($r = 0.79$), followed by CAI values for *B. taurus* ($r = 0.64$),



	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
CPC	8	12	10	5	7	11	8	8	10	10
CUPC	9	12	11	14	9	10	9	9	12	12

CPC: Common Preferred Codons with *Homo sapiens*
UCPC: Common Un-Preferred Codons with *Homo sapiens*

Figure 7. Pairwise correlation analysis of RSCU values of 59 codons of S-genes of beta-coronaviruses with corresponding RSCU values of *H. sapiens*. The R-squared values of linear regression analysis are shown in respective plots. The embedded table denotes the number of common preferred (RSCU ≥ 1) codons and unpreferred (RSCU < 1) codons for the S-genes of the different beta-coronaviruses with *H. sapiens*.

R. norvegicus (r = 0.59), *M. musculus* (r = 0.43) and *H. sapiens* (r = 0.33).

Furthermore, the tRNA Adaptation Index (tAI) of S-gene of 10 different beta-coronaviruses was calculated for the respective primary host (if *H. sapiens* is not the primary host) and *H. sapiens*. The tAI measures the translational efficiency of each codon–anticodon pairing for a gene by utilizing the intracellular tRNA copy number of host species. It is believed that the gene expression level directly correlates with adaptation to intracellular tRNAs [63, 77, 78]. The efficiency quantification of a coding sequence to be recognized by the intra-cellular tRNA pool may be performed by using tAI, which considers weights that identify the wobble interactions between codons and tRNA [51, 54, 55, 79]. The tAI of S-gene of SARS-CoV-2 for *H. sapiens* is found to be 0.30, while the tAI of S-genes of other beta-coronaviruses for *H. sapiens* as the host is also very similar to it. Interestingly, the S-gene of SARS-CoV-2 and other beta-coronaviruses showed a high tAI for *H. sapiens* and *O. cuniculus* as compared to *R. norvegicus*, *M. musculus* and *B. taurus*. The CAI and tAI values of S-gene of SARS-CoV-2 for five different hosts are plotted and compared to S-gene of other beta-coronaviruses for the same five hosts in Figure 8.

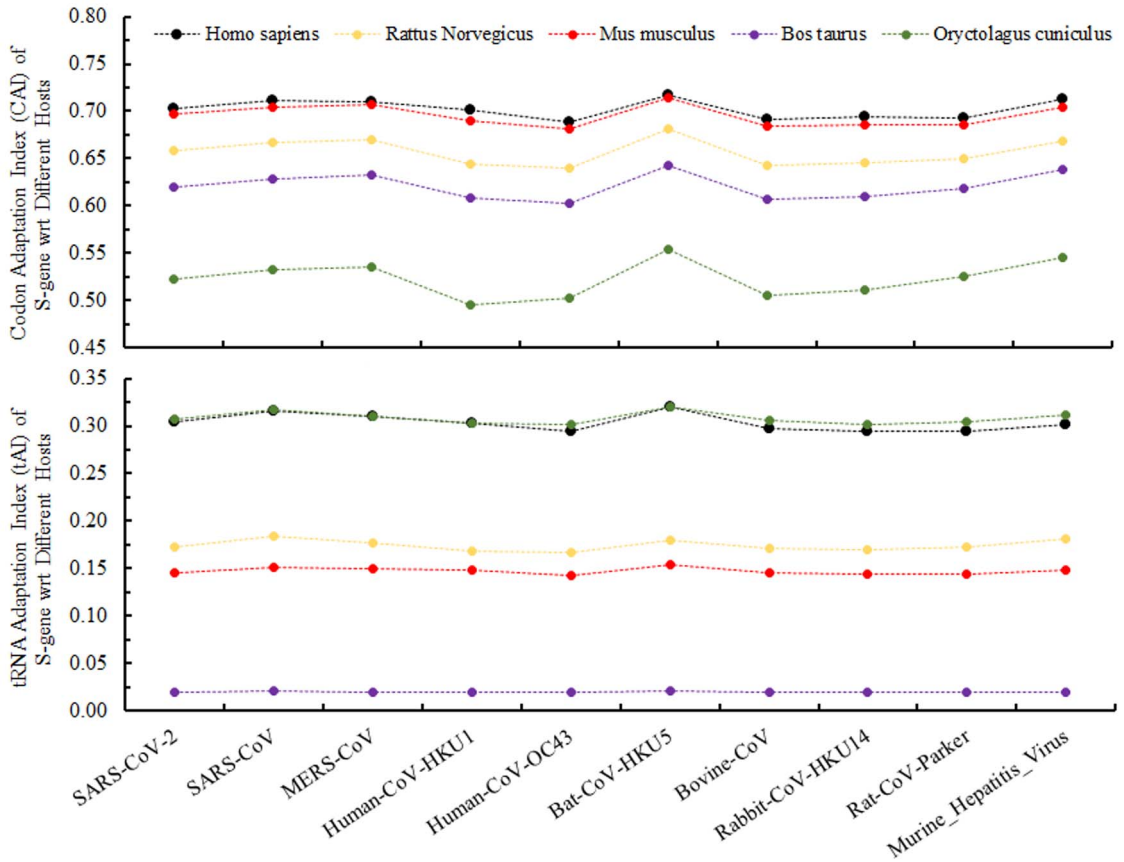


Figure 8. A depiction of the CAI and tAI of S-gene of SARS-CoV-2 with respect to five different hosts and its comparison with S-gene of other beta-coronaviruses. The S-gene of all beta-coronaviruses shows better adaptable codon usage in the case of *H. sapiens* as compared to other studied hosts.

Relative Codon Deoptimization Index analysis

An estimation of the resemblance of the codon frequencies of a particular gene against a reference genome is generally achieved by using the Relative Codon Deoptimization Index (RCDI). The RCDI helps in understanding host–virus phylogenetic relationships and in deducing the possible host range of a virus [59, 80, 81]. An expected RCDI (eRCDI) was calculated by considering randomly generated sequences with similar G+C and amino acid contents to the query gene sequence. A comparison of RCDI and eRCDI may signify the implications of RCDI values for the gene of interest [80]. The high ratio of RCDI to eRCDI is an indication of similar codon usage of the virus and the host. The RCDI for the S-gene of different beta-coronaviruses are computed for respective primary hosts (if *H. sapiens* is not the primary host) and *H. sapiens* as the potential host. There is an inverse relationship between RCDI and degree of adaption to the host.

The RCDI may reveal the synergistic impact of codon bias on gene expression and the potential co-evolution of the virus and its host genomes. The ratio of RCDI (expected) to the RCDI (observed) gives a direct measure of virus–host adaptability, where the ratio close to 1 or higher represent better adaptability as compared to lower values [49, 82–84]. The RCDI value of S-gene of SARS-CoV-2 for *H. sapiens* as a host is observed to be 1.57 as compared to 1.48 and 1.53 for S-genes of SARS-CoV and MERS-CoV, respectively. However, the RCDI values of S-gene for the other two beta-coronaviruses having *H. sapiens* as primary hosts, i.e. Human-CoV-HKU1 (RCDI = 2.27) and Human-CoV-OC43 (RCDI = 1.74), have considerably higher RCDI values than the RCDI value of S-gene of SARS-CoV-2. From the RCDI values, it may be inferred that S-gene of SARS-CoV-2 is better suited to humans as host than SARS-CoV and MERS-CoV while S-gene of Human-CoV-HKU1 is very well adapted to the human genome. The observed RCDI values for S-genes of different beta-coronaviruses for five different hosts are provided in Table 2.

Neutrality plot (neutral evolution) analysis

In the absence of any external pressure, the mutation at first, second or third codon position is supposed to be equally likely. However, the mutations in the first and second codon position may lead to a change in amino acid (non-synonymous mutations), while the third codon position mutations mostly result into the same amino acid (synonymous mutations) [62, 65, 85–87]. The neutrality plot analysis was implemented to examine the mutation-selection equilibrium in determining codon usage bias. The average GC content at first and second positions (GC12) is plotted against the average GC content at the third codon position (GC3) to compare the impacts of natural selection and mutation pressure on codon usage of protein-coding sequences. The values for GC12 and GC3s for S-gene of beta-coronaviruses are calculated using EMBOSS (cusp module) [88]. The slope of the regression line represents the evolutionary speed of mutation pressure (slope = 0 suggests no effect of mutational pressure) and natural selection (slope = 1 indicates complete neutrality) while the regression coefficient against GC3 represents mutation-selection equilibrium coefficient [58, 64, 83]. Here, the neutrality plot offered an overall evolutionary

Table 2. A summary of observed RCDI values of S-gene of different beta-coronaviruses with respect to five different hosts. In parenthesis, the ratio of RCDI (observed) to the RCDI (expected) is provided where the ratio close to one or higher indicates better viral adaptability to the corresponding host(s)

Beta coronavirus	*H. sapiens*	*R. norvegicus*	*M. musculus*	*B. Taurus*	*O. cuniculus*
SARS-CoV-2	1.57 (0.87)	1.65 (0.86)	1.58 (0.87)	1.72 (0.87)	1.97 (0.85)
SARS-CoV	1.48 (0.82)	1.55 (0.81)	1.49 (0.82)	1.61 (0.81)	1.83 (0.79)
MERS-CoV	1.53 (0.84)	1.59 (0.83)	1.53 (0.84)	1.65 (0.83)	1.88 (0.82)
Human-CoV-HKU1	2.27 (1.25)	2.39 (1.25)	2.29 (1.26)	2.49 (1.26)	2.91 (1.26)
Human-CoV-OC43	1.74 (0.96)	1.82 (0.95)	1.74 (0.96)	1.90 (0.96)	2.20 (0.96)
Bat-CoV-HKU5	1.34 (0.74)	1.38 (0.73)	1.34 (0.74)	1.44 (0.73)	1.61 (0.70)
Bovine-CoV	1.76 (0.97)	1.85 (0.97)	1.77 (0.97)	1.92 (0.97)	2.23 (0.97)
Rabbit-CoV-HKU14	1.70 (0.94)	1.80 (0.94)	1.72 (0.95)	1.86 (0.94)	2.16 (0.94)
Rat-CoV-Parker	1.48 (0.82)	1.55 (0.81)	1.49 (0.82)	1.60 (0.81)	1.83 (0.79)
Murine hepatitis virus	1.49 (0.82)	1.58 (0.83)	1.51 (0.83)	1.61 (0.81)	1.84 (0.80)

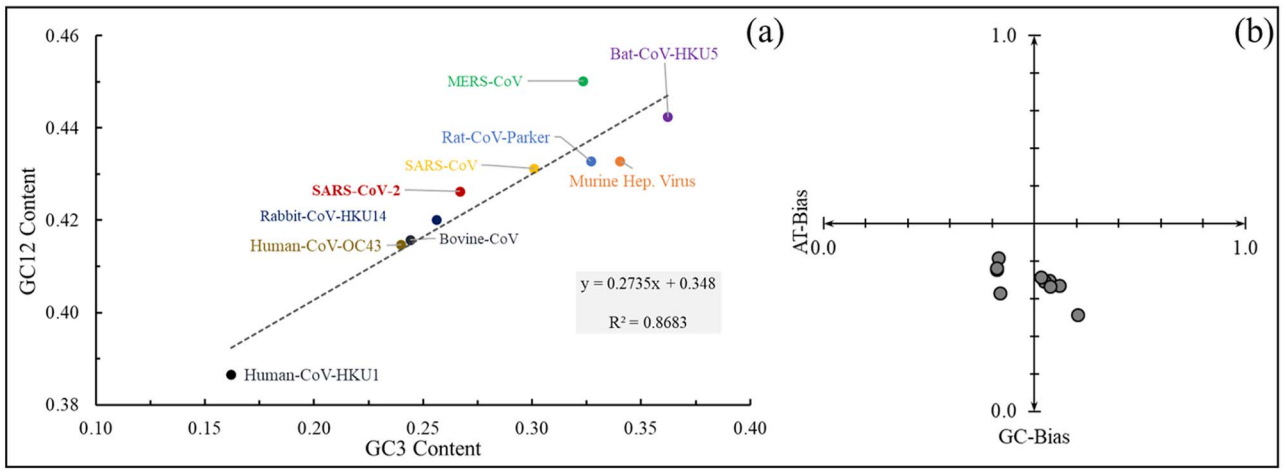


Figure 9. (**A**) Neutrality Plot analysis for S-gene in different beta-coronaviruses. The average GC content at first and second codon positions (GC12) plotted against GC content at the third codon position. The slope of the linear regression line quantifies the influence of mutational pressure in governing the codon usage bias in S-genes of beta-coronaviruses. (**B**) PR2-bias plot, using (A3/(A3+T3)) and (G3/(G3+C3)) for S-genes of different beta-coronaviruses.

speed of S-genes and mutation-selection equilibrium of S-genes in different beta-coronaviruses.

The scatter plot for GC3 (*x*-axis) and GC12 (*y*-axis) contents for S-genes of 10 beta-coronaviruses was used to assess an evolutionary speed of S-genes and mutation-selection equilibrium of S-genes via neutrality plot analysis. A narrow distribution of GC contents indicates the bias of natural selection against mutational pressure. A high correlation between GC3 and GC12 signifies the influence of mutational force at all the codon positions [65, 69, 70, 89, 90]. The slope of the regression line (slope = 0.2735) indicates that the codon usage bias of S-gene in different beta-coronaviruses is governed partly through the mutational pressure (27.35%) and majorly through the natural selection and other factors (72.65%). A very high correlation coefficient (R^2 = 0.87, $P < 0.001$) further supports the considerable influence of mutational pressure. The neutrality plot for S-genes of different beta-coronaviruses is shown in Figure 9A.

PR2-bias plot analysis

The composition of each nucleotide at the third codon position (A3, T3, C3 and G3) was calculated for the S-gene of beta-coronaviruses. These compositions were further used to calculate AT-bias [A3/(A3+T3)] and GC-bias [G3/(G3 + C3)]. Parity Rule 2 (PR2) bias was analyzed by plotting the AT-bias as the ordinate and GC-bias as the abscissa. The PR2-bias plot indicates the relationship among purines (A and G) and pyrimidines (T and C) in codon composition [59, 72, 85, 91, 92].

In mutation pressure analysis, it was observed previously that the AT and GC appear in pairs at third codon positions. It is found that the S-gene of SARS-CoV-2 has almost equal AT- (0.408) and GC-bias (0.414). However, in other beta-coronaviruses, the S-gene has unequal AT- and GC-bias. For instance, in SARS-CoV (AT-bias = 0.410 and GC-bias = 0.377), MERS-CoV (AT-bias = 0.418 and GC-bias = 0.316), human-CoV-HKU1 (AT-bias = 0.602 and GC-bias = 0.258), human-CoV-OC43 (AT-bias = 0.536 and GC-bias = 0.347), bat-CoV-HKU5 (AT-bias = 0.410 and GC-bias = 0.383), bovine-CoV (AT-bias = 0.558 and GC-bias = 0.335), rabbit-CoV-HKU14 (AT-bias = 0.524 and GC-bias = 0.346), rat-CoV-Parker (AT-bias = 0.516 and GC-bias = 0.358) and murine hepatitis virus (AT-bias = 0.538 and GC-bias = 0.333) showed higher deviations in AT- and GC-bias as compared to S-gene of SARS-CoV-2. The differences in AT and GC contents in protein-coding genes are indicative of the contribution of mutation pressure in codon usage bias. The factors resulting in the deviation from neutrality may be further explored by investigating the directionality of AT- and GC-bias. Parity Rule 2 (PR2) plot is used to quantify these biases in S-genes of different beta-coronaviruses [64, 91–93]. All data points lie between 0.25 and 0.60, which suggests an overall low bias in S-gene of beta-coronaviruses. A division

of PR2-plot into four quadrants at 0.5 intersections of AT- and GC-bias showed that AT- and GC-bias for S-gene of four beta-coronaviruses fell into the quadrant-III. For the S-gene of six other beta-coronaviruses, it fell into the quadrant-IV, as shown in Figure 9B.

Discussion

The CoV genome contains various open-reading frames (ORFs) that have genes transcribed by different TRS (as depicted in Figure 1). Genes encoding non-structural proteins (ORF1.1 to ORF1.6) are present at the 5'-UTR, whereas genes encoding structural proteins (N, M, E and S) are present at the 3'-UTR. Between these genes, several accessory genes, encoding accessory proteins, are present. There are a specific number of accessory genes for each virus type. Since the number of ORF genes, and consequently the number of polypeptide chains, is variable in different virus types, in each virus replication cycle, a differential frequency of non-structural protein production occurs as transcription begins at various TRS in each sub-genomic mRNA. For instance, SARS-CoV produces ORF1.1 and ORF1.2 genes, whereas SARS-CoV-2 produces ORF1.1 to ORF1.6 genes, which are involved in the synthesis of several groups of fusion proteins [94–96]. Additionally, CoVs utilize a leaky scanning mechanism (shunting) to produce proteins from overlapping ORFs, translating various proteins from the same mRNA [97]. The surface spike (S) in some CoVs is produced by proteolytic cleavage of a spike precursor [18]. The SARS-CoV has two spike precursors (Sp1 and Sp2), which are proteolytically cleaved to produce two surface glycosylated spikes (S1 and S2) and a protease fragment (S0) [96].

The S-protein gene is highly divergent, but at the nucleotide level, SARS-CoV-2 and SARS-CoV show 72% S-protein nucleotide sequence identity [98]. Lv [99] reported that the SARS-CoV-2 sequence was 50, 79, 96 and 88%, similar to the MERS-CoV, SARS-CoV, bat-SARSr-CoV RaTG13 and two bat CoV (bat-SL-CoVZC45 and bat-SL-CoVZXC21 sequences), respectively. In comparison with the S-protein of SARS-CoV, Wuhan-Hu-1 showed more mutations with a ~19% divergence at the amino acid level [100]. Furthermore, on comparing the RBDs of SARS-CoV and SARS-CoV-2 (Wuhan-Hu-1), 73% of conserved amino acid regions in Wuhan-Hu-1 were observed [39]. Furin and other endogenous proteases play a role in viral host range and infectivity [101]. The S1 and S2 junctions of SARS-CoV-2 possesses a polybasic cleavage [14] site (RRAR) that is cleaved by furin. However, no polybasic cleavage sites were observed in humans, pangolin and bat beta-CoVs [102]. The cleavage site in the MERS-CoV S-protein allows bat MERS-like-CoVs to infect human cells [103]. The polybasic cleavage site also contains a unique proline (PRRAR) insertion before the furin cleavage site. The cleavage site flanked by proline is a unique attribute of SARS-CoV-2 [104]. Since mid-2019, the cleavage site S1/S2 in SARS-like-CoV (RmYN02) in the Rhinolophus bat in the Yunnan province also showed a vital insertion of phenylalanine amino acids, suggesting that these insertions occurred due to CoV natural evolution [7] and possibly not laboratory escape. However, RmYN02 is more divergent and has a sequence similarity of ~72% with that of the SARS-CoV-2 S-protein.

Biochemical interaction studies and crystal structure analysis demonstrated that the SARS-CoV S-protein binds to human ACE2 (hACE2) with a very high affinity [36]. ACE2 is the functional SARS-CoV receptor *in vitro* [105] and *in vivo* [106]. It is required for host cell entry and consequent viral replication. Overexpression of hACE2 increased disease severity in a SARS-CoV mouse model, exhibiting that viral entry into the host cell is an essential step [107]. Additionally, infusing the SARS-CoV S-protein into the mice intensified lung injury. Fundamentally, this injury was constricted due to the obstruction of the renin-angiotensin pathway by ACE2 expression [108]. It is reported that ACE2, from different species, overexpressed in HeLa cells with hACE2, pig ACE2 and civet ACE2 (but not mouse ACE2) permitted SARS-CoV-2 infection and replication, thus directly showing that SARS-CoV-2 used ACE2 as a cellular entry receptor [7]. A diverse type of intermediate host, including pangolins but not mice and rats, for SARS-CoV-2 may exist [109]. Furthermore, the study proved that SARS-CoV-2 did not use the receptors, the aminopeptidase N and dipeptidyl peptidase receptors, which were used by other CoVs [7]. Therefore, the SARS-CoV-2 S-protein has directly bound with the host cell surface ACE2 receptor, facilitating virus entry and replication [110]. The crystal structure of the SARS-CoV-2 CTD S-protein-hACE2 complex revealed an hACE2-binding mode that was similar to that in SARS-CoV.

The analysis of the key residues (K31, E35, D38, M82 and K353) on hACE2, which are responsible for interaction with S-protein in SARS-CoV-2, showed that the primate species like African Green Monkey and Orangutan possesses exactly similar key amino acid residues, which are present in humans (Figure 2 and Table 1). The HawkDock binding free energy in African green monkey (−45.77 kcal/mol) is higher than for human ACE2 (−49.36 kcal/-mol) interaction [41]. In contrast to D38 and M82 in hACE2, the corresponding amino acid in dog ACE2 are E37 and T81, and for remaining species like a cat, tiger and ferret ACE2 are E38 and T82, respectively (Figure 2). The inter-residue distance between F486 of SARS-CoV-2 S-protein and the corresponding contact amino acid (M/T82) in ACE2 is 2.7 Å for humans, 2.6 Å for a dog and 2.7 Å for a tiger, respectively. The predicted distance between F486 of SARS-CoV-2 S-protein and corresponding ACE2 amino acid residues for dog and tiger is $\leq$ 2.7 Å (Figure 2A–D). This distance is shorter/similar to that of hACE2 and SARS2-CoV-2 interaction. The HawkDock binding free energy of the hACE2-S protein complex is −49.36 kcal/mol, while for other species such as a dog and tiger, it is −52.49 and −36.33 kcal/mol, respectively. It may be inferred that that dog ACE2 receptor shows higher binding energy compared to hACE2 (Figure 2 A–D).

Species like pangolin possess identical RBM signatures, which also ascertains the intense speculation regarding pangolins being the intermediate host for SARS-CoV-2. In the case of pangolin ACE2, for hotspot M82 is replaced by N82 and N82 showed a closer contact of 1.62 Å with F486 of S-protein in SARS-CoV-2. In our homology modeling analysis, the inter-residue distance between N82 and F486 was observed to be 2.7 Å (Figure 2 E and F)), and HawkDock Binding free energy was found to be −36.17 kcal/mol. This modeling analysis revealed a similar distance between N82 and F486 while comparing it with M82 and F486 of the hACE2-SARS-CoV-2 interaction. In farms, the animal species like ferrets and pigs, the only change in hotspot 82 was observed where M (Met) changed to T (Thr) with a distance of 2.7 Å (Figure 2 A–D). The docking binding energy of bovine ACE2–S-protein was observed as −41.97 kcal/mol, which is lower than the hACE2 spike protein binding energy [41]. In contrast to mammalian farm species, chicken ACE2 possess a significantly different amino acid signature, viz. E30, R34, D37, R81 and K353 for the corresponding hotspot points concerning hACE2, which does not favor the interaction of chicken ACE2 with SARS-CoV-2 S-protein (Figure 2 E and F)). Furthermore, ACE2 of lab animal species like mouse and rat was not found to interact with SARS-CoV-2 due to a purely different amino acid signature (K, E, D, N, H and N, E, D, S, H, respectively) as contact points. Palm

civet and horseshoe bats have been known to harbor SARS-CoV infection, which led to the SARS-CoV outbreak in 2002–2004 [34]. The key residues K31, K38 and K353 acted as a hotspot and were found intact in horseshoe bats, whereas K31 changed to T31 at hotspot in palm civet ACE2, which formed a strong hydrogen bond with Y442 of SARS-CoV Spike protein during 2002–2004 outbreak [34]. These observations further strengthen the speculation that palm civet and horseshoe bat species can also favor the transmission of closely related SARS-CoV-2 virus to humans. This was further explored with the help of detection of potential recombination events among different strains of coronaviruses. In the current study, results showed that the recombination events in the newly emerged SARS-CoV-2 that are similar to recombination events occurred in SARS-CoV-1 during its emergence. The sequences were identical to those found in horseshoe bats, Himalayan palm civets and raccoons [111, 112]. It was found that all the SARS-CoV-2 sequences fell in the same clade of its most probable animal counterpart (RaTG13 strain of SARS-CoV-2) (Figure 3). The breakpoint analysis showed that irrespective of the place of origin, all SARS-CoV-2 sequences showed one recombinant pattern. This was also apparent for the RaTG13 strain sequence, which possessed a more extended region that was involved in recombination. This indicated a common origin of SARS-CoV-2, its predecessor SARS-CoV-1 (NC_004718) and bat SARS-like-CoV (MG772934) (Figure S1). It was worth noting that the major parent (bat SARS-like-CoV) in the study was, in turn, a recombinant of the minor parent (SARS-CoV-1), making the recombination events more complex to understand. A constant exchange of gene segments among co-infecting viruses in the possible reservoir and human hosts occurred before SARS-CoV-2 emerged to cause the COVID-19 pandemic.

Furthermore, the phylogenetic analysis of spike-proteins in different betacoronaviruses showed that all the beta-coronaviruses share a very distant relationship with the S-gene of SARS-CoV-2 except the S-gene of SARS-CoV. The compositional analysis of S-genes of beta-coronaviruses are rich in A or T at the third codon position (0.72 ± 0.06), which is in accordance with the AT-rich human genome. A similarity in nucleotide compositions may help in imparting the highly efficient adaptability of the virus into the host for replication [58, 64, 65, 90, 93]. The various indices of CUA were performed to analyze the codon usage bias among S-genes of different beta-coronaviruses. The ENC-GC3s plot indicated that the codon usage bias of S-genes in beta-coronaviruses is significantly affected by mutation pressure among certain other factors as the distribution of ENC(observed) is very close to the standard curve of ENC(expected) at different GC3s values [61, 62]. Among S-genes of SARS-CoV-2, SARS-CoV and MERS-CoV, the influence of mutation pressure is more in SARS-CoV-2 (ΔENC = 4.64). The quantification of the role of mutation pressure in shaping the codon usage through neutrality plot suggested a 27% influence of mutation pressure. At the same time, the rest is contributed by natural selection along with other factors. This assumption is further supported by the presence of a very high correlation between GC12 and GC3 compositions ($R^2 = 0.87$, $P < 0.001$). The RSCU values for 59 codons for S-gene of SARS-CoV-2 showed a very high correlation with S-genes of other beta-coronaviruses, suggesting a similar extent of codon usage bias is seen among different beta-coronaviruses. Moreover, the number of preferred codons having A/U at the third synonymous codon position varies from 20 to 25 (except 18 in MERS-CoV). In S-gene of SARS-CoV-2, 25 out of 27 preferred codons have A/U at the third codon position (among the highest fraction for all beta-coronaviruses), which suggests that S-gene of SARS-CoV-2 is inclined to use A/U ending codons and the nucleotide composition changes resulting from different factors may play a pivotal feature in the codon usage bias. The comparison of RSCU values of S-gene of different beta-coronaviruses with corresponding RSCU values in *H. sapiens* suggested a significantly similar codon usage. For instance, the S-gene of SARS-CoV-2 and *H. sapiens* share eight similar preferred codons, the S-gene of SARS-CoV and *H. sapiens* shares 12 similar preferred codons, and the S-gene of MERS-CoV and *H. sapiens* shares 10 similar preferred codons. The number of the commonly preferred codons among viruses and hosts indicates the better adaptability and replication efficiency of a virus into the host [65, 66, 73, 93, 113]. Also, it is observed that there are not very considerable deviations in the values of CAI among the S-genes of beta-coronaviruses for the given hosts. The viruses use a controlled expression of viral proteins to achieve replicative suitability. The means of attaining replicative suitability varies among viruses as some use optimized codon usage while others escape the host immune system [59, 61, 72, 114]. The high CAI values reflect the optimal codon usage to achieve replication efficiency in the host [75, 90, 115]. The S-gene of SARS-CoV-2 showed relatively better adaptability in *H. sapiens* when compared to *R. norvegicus*, *M. musculus*, *B. taurus* and *O. cuniculus*. However, the S-gene of other beta-coronaviruses also showed a similar trend over all the hosts. Overall, it may be inferred that the S-gene of beta-coronaviruses uses the most abundant codons.

The RCDI analysis suggested better adaptability of S-gene of different beta-coronaviruses in five other hosts. The S-gene of SARS-CoV-2 showed a better concordance with all the five hosts as compared to its counterparts in SARS-CoV and MERS-CoV. Also, the ratio of observed RCDI to the expected RCDI values further supported the assumption of better adaptability of S-gene of SARS-CoV-2 than SARS-CoV and MERS-CoV. The high RCDI values indicate that the spike-protein may be expressed in dormant stages, and the virus may be present at a low replication rate in the hosts. The neutrality plot analysis showed the dominance of natural selection (selection pressure) over the mutational pressure. The slope of the regression line indicates that the codon usage bias of S-gene in different beta-coronaviruses is governed partly through the mutational pressure (27.35%) and majorly through the natural selection and other factors (72.65%). A very high correlation coefficient ($R^2 = 0.87$, $P < 0.001$) further supports the considerable influence of mutational pressure. An analysis of AT- and GC-biases using Parity Rule 2 (PR2) suggested an overall low bias in S-genes of beta-coronaviruses. However, the AT- and GT-biases are very balanced in S-gene of SARS-CoV-2 as compared to other beta-coronaviruses.

Overall, in this study, we have presented an exhaustive analysis on structure of CoV genomes, spike-protein configuration, spike-protein and hACE2 mediated entry of SARS-Cov-2 to human host, recombination analysis of spike-protein among different strains of the virus, and CUA of S-gene in different beta-coronaviruses and its compatibility with different hosts.

Conclusions and future prospects

The new global coronavirus disease 2019 (COVID-19) pandemic, by severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), has infected over 17 million people all over the world and causing nearly 0.8 million deaths. The spike-protein of the SARS-CoV-2 is present on the virus envelope and plays a key role in gaining the viral entry into the human cell with the help of its receptor-binding domain. The S-protein is fundamental

for SARS-CoV-2 because the detachment of the S1 subunit destabilizes the trimer structure, exposing the RBD to host cell membranes. Compared to its predecessors, SARS-CoV, the S-glycoprotein of SARS-CoV-2, has unique characteristics that help in the spread of the virus. Studies support that SARS-CoV-2 showed nearly similar or higher human receptor-binding ability than SARS-CoV due to the presence of the S-protein, leading to human-to-human transmission. Although the recurrence of recombination is higher in the S-glycoprotein, the 5′ end of the gene may act as a recombination coldspot. While recombination in the glycoprotein RBD may extend the infection tropism, negative selection pressure and recombination coldspots limit its changeability to maintain the right structure and function. RBD is a topnotch region for drug development, whereas areas in coldspots or negative pressure might be more useful for vaccine development.

The codon usage bias analysis of SARS-CoV-2 and its benchmarking with nine other beta-coronaviruses indicated that there is significant codon usage bias among the S-gene of different beta-coronaviruses and is majorly governed by natural selection and mutational pressure along with certain other compositional constraints. The codon usage adaptability of these beta-coronaviruses is also explored by considering five different hosts. Various indices of CUA of S-gene of SARS-CoV-2 helped in quantifying its adaptability in different hosts. The CAI and tAI of S-gene of SARS-CoV-2 indicate that it is best adapted to *H. sapiens,* followed by *M. musculus* (according to CAI) and *O. cuniculus* (according to tAI). The study may help in the identification of potential experimental animal model for investigating pathogenicity for drug and vaccine development experiments. With the second-highest CAI value (after *H. sapiens*) *M. musculus* may be the most appropriate model for pathogenicity directed animal experiments for drug and vaccine development regimes. Analytical studies have shown that SARS-CoV-2 possibly emerged due to recombination events that involved ancestral strains circulating among bats and pangolins. However, the recombinant sequences of the parent strains could not be distinguished. Furthermore, sampling of CoVs from various wildlife species is required to understand the mechanism that led to human transmission. Future emergence of new CoVs should be a constant concern, and continuous surveillance of the viral population must be performed to understand the dynamics of viruses and thus help in the prevention or control of new zoonotic outbreaks.

Data availability statement

The data is provided in the supplementary. Any other additional data related to this study will be provided on request through corresponding author.

Acknowledgments

All the authors acknowledge and thank their respective Institutes and Universities. Y.S.M. acknowledge the Education Division, Indian Council of Agricultural Research for National Fellowship.

Funding

The information compiled and analyzed in this article did not require any substantial funding to be stated.

References

1. Jiang S, Hillyer C, Du L. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. *Trends Immunol* 2020;**41**:355–9.
2. Drosten C, Günther S, Preiser W, *et al*. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1967–76.
3. Ksiazek TG, Erdman D, Goldsmith CS, *et al*. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1953–66.
4. Zaki AM, Van Boheemen S, Bestebroer TM, *et al*. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* 2012;**367**:1814–20.
5. Zhu N, Zhang D, Wang W, *et al*. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727–33.
6. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 2020;**395**:497–506.
7. Zhou P, Yang X, Wang X, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.
8. Hu B, Zeng L-P, Yang X-L, *et al*. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* 2017;**13**:e1006698.
9. Yang X-L, Hu B, Wang B, *et al*. Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of severe acute respiratory syndrome coronavirus. *J Virol* 2016;**90**:3253–6.
10. Kan B, Wang M, Jing H, *et al*. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J Virol* 2005;**79**:11892–900.
11. Wang M, Yan M, Xu H, *et al*. SARS-CoV infection in a restaurant from palm civet. *Emerg Infect Dis* 2005;**11**:1860.
12. Haagmans BL, Al Dhahiry SH, Reusken CB, *et al*. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect Dis* 2014;**14**:140–5.
13. Memish ZA, Mishra N, Olival KJ, *et al*. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg Infect Dis* 2013;**19**:1819.
14. Walls AC, Park Y-J, Tortorici MA, *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020;**181**:281–292.e286.
15. Li F. Structure, function, and evolution of coronavirus spike proteins. *Annual Review of Virology* 2016;**3**:237–61.
16. Madu IG, Roth SL, Belouzard S, *et al*. Characterization of a highly conserved domain within the severe acute respiratory syndrome coronavirus spike protein S2 domain with characteristics of a viral fusion peptide. *J Virol* 2009;**83**:7411–21.
17. Millet JK, Whittaker GR. Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. *Virus Res* 2015;**202**:120–34.
18. Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc Natl Acad Sci* 2009;**106**:5871–6.
19. Park J-E, Li K, Barlan A, *et al*. Proteolytic processing of Middle East respiratory syndrome coronavirus spikes expands virus tropism. *Proc Natl Acad Sci* 2016;**113**:12262–7.

20. Yang Y, Peng F, Wang R, *et al*. The deadly coronaviruses: the 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. *J Autoimmun* 2020;**109**:102434.
21. Brian DA, Hogue BG, Kienzle TE. The coronavirus hemagglutinin esterase glycoprotein. In: *The Coronaviridae*. Boston, USA: Springer, 1995, 165–79.
22. De Groot RJ. Structure, function and evolution of the hemagglutinin-esterase proteins of corona-and toroviruses. *Glycoconj J* 2006;**23**:59–72.
23. Sethna PB, Hung S-L, Brian DA. Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. *Proc Natl Acad Sci* 1989;**86**:5626–30.
24. Irigoyen N, Firth AE, Jones JD, *et al*. High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. *PLoS Pathog* 2016;**12**:e1005473.
25. Wrapp D, Wang N, Corbett KS, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;**367**:1260–3.
26. Yuan M, Wu NC, Zhu X, *et al*. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* 2020;**368**:630–3.
27. Bosch BJ, Van der Zee R, De Haan CA, *et al*. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J Virol* 2003;**77**:8801–11.
28. Gui M, Song W, Zhou H, *et al*. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res* 2017;**27**:119–29.
29. Shang J, Zheng Y, Yang Y, *et al*. Cryo-electron microscopy structure of porcine deltacoronavirus spike protein in the prefusion state. *J Virol* 2018;**92**(4):e01556–17.
30. Tian H, Tao P. Deciphering the protein motion of S1 subunit in SARS-CoV-2 spike glycoprotein through integrated computational methods. *J Biomol Struct Dyn*. 2020. doi: 10.1080/07391102.2020.1802338.
31. Wu F, Zhao S, Yu B, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**:265–9.
32. Tagliamonte MS, Abid N, Ostrov DA, *et al*. Recombination and purifying selection preserves covariant movements of mosaic SARS-CoV-2 protein S. *bioRxiv* 2020. doi: 2020.2003.2030.015685.
33. Lu R, Zhao X, Li J, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 2020;**395**:565–74.
34. Wan Y, Shang J, Graham R, *et al*. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol* 2020;**94**:e00127–20.
35. Xu X, Chen P, Wang J, *et al*. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci China Life Sci* 2020;**63**:457–60.
36. Li F, Li W, Farzan M, *et al*. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* 2005;**309**:1864–8.
37. Wu K, Peng G, Wilken M, *et al*. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. *J Biol Chem* 2012;**287**:8904–11.
38. Wang Q, Zhang Y, Wu L, *et al*. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 2020;**181**:894–904.e899.
39. Waterhouse A, Bertoni M, Bienert S, *et al*. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;**46**:W296–303.
40. DeLano WL. Pymol: an open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography* 2002;**40**:82–92.
41. Weng G, Wang E, Wang Z, *et al*. HawkDock: a web server to predict and analyze the protein–protein complex based on computational docking and MM/GBSA. *Nucleic Acids Res* 2019;**47**:W322–30.
42. Hirst GK. Genetic recombination with Newcastle disease virus, polioviruses, and influenza. In: *Cold Spring Harbor Symposia on Quantitative Biology*. New York, USA: Cold Spring Harbor Laboratory Press, 1962, 303–9.
43. Lai MM, Baric RS, Makino S, *et al*. Recombination between nonsegmented RNA genomes of murine coronaviruses. *J Virol* 1985;**56**:449–56.
44. Cong Y, Zarlenga DS, Richt JA, *et al*. Evolution and homologous recombination of the hemagglutinin–esterase gene sequences from porcine torovirus. *Virus Genes* 2013;**47**:66–74.
45. Martin DP, Murrell B, Golden M, *et al*. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evolution* 2015;**1**(1):vev003.
46. Alexaki A, Kames J, Holcomb DD, *et al*. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant Gene Design. *J Mol Biol* 2019;**431**:2434–41.
47. Komar AA. The Yin and Yang of codon usage. *Hum Mol Genet* 2016;**25**:R77–85.
48. Kumar N, Bera BC, Greenbaum BD, *et al*. Revelation of influencing factors in overall codon usage bias of equine influenza viruses. *PLoS One* 2016;**11**:e0154376.
49. Wang H, Liu S, Zhang B, *et al*. Analysis of synonymous codon usage bias of Zika virus and its adaption to the hosts. *PLoS One* 2016;**11**:e0166260.
50. Gu W, Zhou T, Ma J, *et al*. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. *Virus Res* 2004;**101**:155–61.
51. Kumar N, Kulkarni DD, Lee B, *et al*. Evolution of codon usage bias in Henipaviruses is governed by natural selection and is host-specific. *Viruses* 2018;**10**:604.
52. Mihara T, Nishimura Y, Shimizu Y, *et al*. Linking virus genomes with host taxonomy. *Viruses* 2016;**8**:66.
53. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* 2005;**102**:10557–62.
54. Guan D-L, Ma L-B, Khan MS, *et al*. Analysis of codon usage patterns in Hirudinaria manillensis reveals a preference for GC-ending codons caused by dominant selection constraints. *BMC Genomics* 2018;**19**:542.
55. Khandia R, Singhal S, Kumar U, *et al*. Analysis of Nipah virus codon usage and adaptation to hosts. *Front Microbiol* 2019;**10**:886.
56. Wright F. The 'effective number of codons' used in a gene. *Gene* 1990;**87**:23–9.
57. Fuglsang A. The 'effective number of codons' revisited. *Biochem Biophys Res Commun* 2004;**317**:957–64.
58. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* 2002;**19**:1390–4.
59. Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol* 1998;**47**:268–74.

60. Hershberg R, Petrov DA. Selection on codon bias. *Annu Rev Genet* 2008;**42**:287–99.
61. Sharp PM, Stenico M, Peden JF, *et al*. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* 1993;**21**:835–41.
62. Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* 2001;**2**:research0010.0011.
63. Md R, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 2004;**32**:5036–44.
64. Lehmann J, Libchaber A. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *RNA* 2008;**14**:1264–9.
65. Belalov IS, Lukashev AN. Causes and implications of codon usage bias in RNA viruses. *PLoS One* 2013;**8**:e56642.
66. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. *Trends Biotechnol* 2004;**22**:346–53.
67. Zhou H, Wang H, Huang L, *et al*. Heterogeneity in codon usages of sobemovirus genes. *Arch Virol* 2005;**150**:1591–605.
68. Lara-Ramírez EE, Salazar MI, MdJ L-L, *et al*. Large-scale genomic analysis of codon usage in dengue virus and evaluation of its phylogenetic dependence. *Biomed Res Int* 2014;**2014**.
69. Cristina J, Moreno P, Moratorio G, *et al*. Genome-wide analysis of codon usage bias in ebolavirus. *Virus Res* 2015;**196**:87–93.
70. Butt AM, Nasrullah I, Qamar R, *et al*. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerging Microbes & Infections* 2016;**5**:1–14.
71. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987;**15**:1281–95.
72. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci* 1988;**85**:2653–7.
73. Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA expression. *PLoS Genet* 2006;**2**:e221.
74. Carbone A, Zinovyev A, Képes F. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 2003;**19**:2005–15.
75. Puigbò P, Bravo IG, Garcia-Vallve S. CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* 2008;**3**:1–8.
76. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* 2000;**28**:292–2.
77. Dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res* 2003;**31**:6976–85.
78. Sabi R, Volvovitch Daniel R, Tuller T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* 2017;33, 589–91.
79. Song H, Liu J, Song Q, *et al*. Comprehensive analysis of codon usage bias in seven epichloe species and their peramine-coding genes. *Front Microbiol* 2017;**8**:1419.
80. Sharp PM, Li W-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 1986;**24**:28–38.
81. Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 1986;**14**:5125–43.
82. Puigbò P, Aragonès L, Garcia-Vallvé S. RCDI/eRCDI: a web-server to estimate codon usage deoptimization. *BMC Res Notes* 2010;**3**:87.
83. Kunec D, Osterrieder N. Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep* 2016;**14**:55–67.
84. Castells M, Victoria M, Colina R, *et al*. Genome-wide analysis of codon usage bias in bovine coronavirus. *Virol J* 2017;**14**:115.
85. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 1991;**129**:897–907.
86. Archetti M. Codon usage bias and mutation constraints reduce the level of ErrorMinimization of the genetic code. *J Mol Evol* 2004;**59**:258–66.
87. Chen SL, Lee W, Hottes AK, *et al*. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci* 2004;**101**:3480–5.
88. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;**16**(6):276–7. doi: 10.1016/s0168-9525(00)02024-2.
89. Butt AM, Nasrullah I, Tong Y. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. *PLoS One* 2014;**9**:e90905.
90. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res* 2013;**41**:2073–94.
91. Sueoka N. Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+ C content of third codon position. *Gene* 1999;**238**:53–8.
92. Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. *J Mol Evol* 2001;**53**:469–76.
93. Dietel A-K, Merker H, Kaltenpoth M, *et al*. Selective advantages favour high genomic AT-contents in intracellular elements. *PLoS Genet* 2019;**15**:e1007778.
94. van Boheemen S, de Graaf M, Lauber C, *et al*. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio* 2012;**3**(6):e00473–12.
95. Bock LV, Caliskan N, Korniy N, *et al*. Thermodynamic control of– 1 programmed ribosomal frameshifting. *Nat Commun* 2019;**10**:1–11.
96. Pascual MR. Coronavirus SARS-CoV-2: analysis of subgenomic mRNA transcription, 3CLpro and PL2pro protease cleavage sites and protein synthesis. 2020.
97. Nakagawa K, Lokugamage K, Makino S. Viral and cellular mRNA translation in coronavirus-infected cells. *Advances in Virus Research Elsevier* 2016;**96**:165–92.
98. Zhang Y-Z, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* 2020;**181**:223–7.
99. Lv H, Wu NC, Tsang OT-Y, *et al*. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections. *Cell Rep* 2020;**31**:107725.
100. Rehman SU, Shafique L, Ihsan A, *et al*. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. *Pathogens* 2020;**9**(3):240.
101. Nao N, Yamagishi J, Miyamoto H, *et al*. Genetic predisposition to acquire a polybasic cleavage site for highly pathogenic avian influenza virus hemagglutinin. *MBio* 2017;**8**:e02298–16.
102. Chan C-M, Woo PC, Lau SK, *et al*. Spike protein, S, of human coronavirus HKU1: role in viral life cycle and application in antibody detection. *Exp Biol Med* 2008;**233**:1527–36.
103. Menachery VD, Dinnon KH, Yount BL, *et al*. Trypsin treatment unlocks barrier for zoonotic bat coronavirus infection. *J Virol* 2020;**94**:e01774–19.

104. Andersen KG, Rambaut A, Lipkin WI, *et al*. The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**:450–2.
105. Li W, Moore MJ, Vasilieva N, *et al*. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003;**426**:450–4.
106. Kuba K, Imai Y, Rao S, *et al*. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nat Med* 2005;**11**:875–9.
107. Yang X-h, Deng W, Tong Z, *et al*. Mice transgenic for human angiotensin-converting enzyme 2 provide a model for SARS coronavirus infection. *Comp Med* 2007;**57**:450–9.
108. Imai Y, Kuba K, Rao S, *et al*. Angiotensin-converting enzyme 2 protects from severe acute lung failure. *Nature* 2005;**436**:112–6.
109. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol* 2020;**30**:1346–1351.e1342.
110. Zhang H, Penninger JM, Li Y, *et al*. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med* 2020;**46**:586–90.
111. Lau SKP, Feng Y, Chen H, *et al*. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. *J Virol* 2015;**89**:10532–47.
112. Guan Y, Zheng BJ, He YQ, *et al*. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 2003;**302**:276.
113. Brandão PE. Avian coronavirus spike glycoprotein ectodomain shows a low codon adaptation to Gallus gallus with virus-exclusive codons in strategic amino acids positions. *J Mol Evol* 2012;**75**:19–24.
114. Smith DW. Problems of translating heterologous genes in expression systems: the role of tRNA. *Biotechnol Prog* 1996;**12**:417–22.
115. Piovesan A, Pelleri MC, Antonaros F, *et al*. On the length, weight and GC content of the human genome. *BMC Res Notes* 2019;**12**:106.