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

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Decoding molecular factors shaping human angiotensin converting enzyme 2 receptor usage by spike glycoprotein in lineage B beta-coronaviruses

Sandipan Chakraborty a,b,*, Sanjana Ghosh b, Trisha Mondal b

a Center for Innovation in Molecular and Pharmaceutical Sciences (CIMPS), Dr. Reddy's Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad 500046, India
b Amity Institute of Biotechnology, Amity University, Kolkata 700135, India

ARTICLE INFO

Keywords:
ACE2 usage
Beta-coronaviruses
Recombination
Selection forces
Interfacial hydrogen bonds

ABSTRACT

Acquiring the human ACE2 receptor usage trait enables the coronaviruses to spill over to humans. However, the origin of the ACE2 usage trait in coronaviruses is poorly understood. Using a multi-disciplinary approach combining evolutionary bioinformatics and molecular dynamics simulation, we decode the principal driving force behind human ACE2 receptor recognition in coronaviruses. Genomic content, evolutionary divergence, and codon usage bias analysis reveal that SARS-CoV2 is evolutionarily divergent from other human ACE2-user CoVs, indicating that SARS-CoV2 originates from a different lineage. Sequence analysis shows that all the human ACE2-user CoVs contain two insertions in the receptor-binding motif (RBM) that directly interact with ACE2. However, the insertion sequences in SARS-CoV2 are divergent from other ACE2-user CoVs, implicating their different recombination origins. The potential of mean force calculations reveal that the high binding affinity of SARS-CoV2 RBD to human ACE2 is primarily attributed to its ability to form a higher number of hydrogen bonds than the other ACE2-user CoVs. The adaptive branch-site random effects likelihood method identifies positive selection bias across the ACE2 user CoVs lineages. Recombination and selection forces shape the spike evolution in human ACE2-using beta-CoVs to optimize the interfacial hydrogen bonds between RBD and ACE2. However, these evolutionary forces work within the constraints of nucleotide composition, ensuring optimum codon adaptation of the spike (S) gene within the host cell.

1. Introduction

Understanding the evolutionary dynamics of coronavirus is a topic of recent great importance. We are witnessing a Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) pandemic, initially reported from Wuhan, China, in December 2019. The virus evolves into various variants with enhanced receptor recognition and immune evasion abilities during the pandemic [1–3]. As a result, we have witnessed many waves of upsurge cases coming one after another, led by different variants [4,5]. Thus studying the evolutionary arms race dynamics of SARS-CoV2 for its receptor usage is very important to understand the pathway of diversifications of the virus. Identifying crucial residues for receptor recognition under strong evolutionary bias is essential in predicting future variants with improved host cell invasion abilities.

Coronaviruses belong to the family Coronaviridae in the order Nidovirales and further classified into four genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus [6]. Alpha- and beta-coronaviruses primarily infect mammals [6], while gamma-coronaviruses infect avian species [7,8]. Notably, the first evidence of avian species infection by an emerging mammalian alphacoronavirus also has been demonstrated very recently [9]. Deltacoronaviruses can infect both mammalian and avian species [10]. All CoVs encode a surface glycoprotein spike, which binds to the host-cell receptor to mediate viral entry into the cell [11,12]. Structurally, the spike glycoprotein is comprised of two domains, S1 and S2. Three monomers of spike proteins entangle to form a homo-trimeric large clover-shaped protrusion [13]. Three S1 domains entwine to form the ectodomain, and the S2 domains intertwine to create the stalk, transmembrane, and small intracellular domains [14]. The S1 region contains the receptor-binding domain (RBD), which binds the host receptor to open up the cleavage sites [15]. Cleavage by the host proteases mediates the fusion of the viral membrane to the host membrane.
Receptor usage in coronaviruses is a puzzle. There are four lineages in betacoronavirus. Lineage A, *Embecovirus*, includes OC43 and HKU1, which can cause the common cold in humans [16]. Lineage B, *Sarbecovirus*, contains two significant viruses that cause pandemics (SARS-CoV and SARS-CoV2). On the other hand, MERS-CoV of lineage C infects humans [16]. Lineage D is mostly bat coronaviruses like Eidolon bat coronavirus G704 and Rousettus bat coronavirus HKU9 [17]. However, receptor usage by coronavirus spike protein does not follow the classification pattern based on the phylogeny [18]. Phylogenetically distinct alphacoronavirus HCoV-NL63 and the betacoronavirus SARS-CoV recognize human angiotensin-converting enzyme 2 (ACE2) [19, 20]. Bat SARS-like coronavirus Rp3, a phylogenetically closer CoV to SARS-CoV, fails to bind human ACE2 [21]. On the other hand, bat SARS-like coronavirus WIV1 and WIV16 bind human ACE2 [22].

Understanding the receptor usage trait using whole-genome phylogeny can be misleading as different genomic regions undergo recombination events from different origins [18]. Genetic analysis reveals that recombination events primarily dictate the receptor utilization phenomenon in coronavirus [18, 22]. A breakpoint has been detected at the spike gene (S gene) boundary [24]. Sequence comparisons of coronavirus spike proteins reveal insertion regions in the receptor-binding domain (RBD) of the CoVs capable of infecting humans. Li et al. suggested different regions within the spike protein undergo diverse evolutionary processes. Recombination is highly evident in the receptor-binding domain (RBD), while the S2 region has been predicted to undergo purifying selection [23]. Evolutionary analysis by Wells et al. showed that the SARS-CoV2 lineage most likely evolved from the ancestral ACE2-using lineage [18]. Recombination with at least one virus from this group possibly introduced the ACE2 usage trait to the SARS-CoV 2 lineage [18]. Thus, it is essential to study the genetic diversification process in the S gene and identify the driving force for ACE2 recognition that shapes the evolution of the lineage B beta-CoVs. Notably, the RBD-ACE2 complexes of SARS-CoV and SARS-CoV2 are structurally very similar. However, SARS-CoV2 RBD binds with human ACE2 with a four-fold higher binding affinity than SARS-CoV due to the effective receptor recognition mechanism [25], which accounts for the observed higher infectivity of SARS-CoV2. Thus beyond recombination, several processes of genetic optimization of the coding sequences are involved that govern the receptor specificity and binding affinity among closely related CoVs.

Letko and co-workers developed a method to study 29 closely related CoVs spike RBD from lineage B of betacoronavirus on their ability to bind human ACE2 [26]. Their data shows that clade 1 of these closely related spike proteins can bind to human ACE2. Here, we study the genetic diversification of the S genes from SARS-CoV2 and 29 other lineage B of beta-coronaviruses, studied by Letko et al. [26]. Genetic diversification is a dynamic and complex process. Using a combination of evolutionary bioinformatics and extensive molecular dynamics simulation, we are trying to decode the driving force that shapes the evolutionary processes involved in the spike protein to gain binding affinity and specificity toward the human ACE2 receptor.

2. Methods

2.1. Preparation of the dataset

Letko and co-workers studied twenty-nine closely related lineage B beta-CoVs spike RBD on their ability to bind human ACE2 [26]. The coding region of the S genes and their respective protein sequences of those 29 CoVs and SARS-CoV2 were obtained from the NCBI sequence database. The Wuhan isolate was considered for SARS-CoV2. The accession number of all the sequences are listed in the Supplementary database. The Wuhan isolate was considered for SARS-CoV2. The genetic diversification of the SARS-CoV2 lineage, studied by Letko et al. [26]. Genetic analysis reveals that recombination events primarily dictate the receptor utilization phenomenon in coronavirus [18, 22]. A breakpoint has been detected at the spike gene (S gene) boundary [24]. Sequence comparisons of coronavirus spike proteins reveal insertion regions in the receptor-binding domain (RBD) of the CoVs capable of infecting humans. Li et al. suggested different regions within the spike protein undergo diverse evolutionary processes. Recombination is highly evident in the receptor-binding domain (RBD), while the S2 region has been predicted to undergo purifying selection [23]. Evolutionary analysis by Wells et al. showed that the SARS-CoV2 lineage most likely evolved from the ancestral ACE2-using lineage [18]. Recombination with at least one virus from this group possibly introduced the ACE2 usage trait to the SARS-CoV2 lineage [18]. Thus, it is essential to study the genetic diversification process in the S gene and identify the driving force for ACE2 recognition that shapes the evolution of the lineage B beta-CoVs. Notably, the RBD-ACE2 complexes of SARS-CoV and SARS-CoV2 are structurally very similar. However, SARS-CoV2 RBD binds with human ACE2 with a four-fold higher binding affinity than SARS-CoV due to the effective receptor recognition mechanism [25], which accounts for the observed higher infectivity of SARS-CoV2. Thus beyond recombination, several processes of genetic optimization of the coding sequences are involved that govern the receptor specificity and binding affinity among closely related CoVs.

Letko and co-workers developed a method to study 29 closely related CoVs spike RBD from lineage B of betacoronavirus on their ability to bind human ACE2 [26]. Their data shows that clade 1 of these closely related spike proteins can bind to human ACE2. Here, we study the genetic diversification of the S genes from SARS-CoV2 and 29 other lineage B of beta-coronaviruses, studied by Letko et al. [26]. Genetic diversification is a dynamic and complex process. Using a combination of evolutionary bioinformatics and extensive molecular dynamics simulation, we are trying to decode the driving force that shapes the evolutionary processes involved in the spike protein to gain binding affinity and specificity toward the human ACE2 receptor.

2.2. Analysis of genomic composition, codon usage, and selection pressure of S genes

Several nucleotide composition parameters of the coding region of all the 30 S genes were analyzed using the MAGA-X [27] version 10.1. Nucleotide composition at each codon site was also analyzed. Relative synonymous codon usage (RSCU) for the S gene of all 30 CoVs were calculated using the DAMBE 6.0 [28]. The Effective Number of Codon (ENC) Analysis for all the 30 S genes was performed using the DAMBE 6.0 [28]. Correspondence analysis was performed based on the codon usage pattern of 30 S genes using the program CodonW. Codon adaptation index (CAI) analysis is a quantitative value indicating the frequency of a preferred codon utilization and implicates translational efficiency. CAI values for the S gene of 30 related beta-CoVs were calculated using an RSCU reference set for humans with the CodonW package.

2.3. Evolutionary analysis and phylogeny

The coding regions of the S gene from the 30 beta-CoVs were analyzed to identify the best-suited nucleotide substitution model by performing Maximum Likelihood fits on 24 different substitution models. The general time-reversible (GTR) model with non-uniformity of evolutionary rates among sites modeled using a discrete Gamma distribution (+G) with a shape parameter of 0.309 was the best-suited substitution model, judged by the lowest Bayesian Information Criterion score. The evolutionary distance between each pair of the sequence was estimated by calculating the number of base substitutions per site between sequences using the maximum composite likelihood model [29]. Pattern disparities between sequences were estimated by calculating the homogeneity in the substitution patterns [30].

The codon-based multiple sequence alignment of the 30 beta-CoVs S gene was performed using the MUSCLE program implemented in DAMBE 6.0 [28]. The phylogenetic tree was constructed using the maximum likelihood (ML) method implemented in MEGA-X [27]. The general time-reversible (GTR) model with non-uniformity of evolutionary rates among sites modeled using a discrete Gamma distribution (+G) was used to construct the maximum likelihood tree using the Bootstrap resampling of 100. We used a branch-site model to detect the episodic selection on the branches of the obtained phylogenetic tree using the aBS-REL (Adaptive branch-site random effects likelihood) method [31] implemented using the DataMonkey web-server [32, 33].

2.4. Modeling the structure and dynamics of the spike RBD for different CoVs using equilibrium simulation

All the simulations were performed using GROMACS 2018.1 packages [34, 35]. The AMBER99SB-ILDN [36] force field and TIP3P water model were used in all the simulations. We recently refined the crystal structure of the SARS-CoV2 RBD-ACE2 complex (PDB ID: 6MUJ [37]) using extensive molecular dynamics simulation [38]. The structure of the RBD region of SARS-CoV2 was considered only. In addition to SARS-CoV2, we considered the spike protein of SARS-CoV, RsSHC014 as the representative of the ACE2 user clade. From non-ACE2 user clades, we choose YN2013, Rf4092, and Rs4081. Also, the spike protein of another non-ACE2-user CoV, CoVZXC21, which shares the same clade with SARS-CoV2, was considered. RBD structures of all the other beta-CoVs were modeled from the protein sequence using homology modeling implemented in the SWISS-Model web interface [39, 40] using the structure of the SARS-CoV2 RBD as a template.

The RBD structures of the spike proteins for all the seven CoVs were initially energy minimized in vacuo to remove any bad contacts. Then each RBD was immersed in a triclinic box containing the TIP3P water so that the minimum distance between any protein atom and box walls was >10 Å. Each system was then charge neutralized by adding the appropriate number of counter ions. Then energy minimization on each
system was performed in water using the steepest descent algorithm for 500 steps. Then 2 ns of position-restrained dynamics were performed for each system where the protein backbone dynamics were restrained, but water molecules were allowed to move freely. 2 ns simulation in NVT (canonical) ensemble was carried out at 298 K, followed by another 10 ns simulation in the NPT (Isothermal–Isobaric) ensemble for each RBD. In both the simulations, proteins and solvent molecules were allowed to move freely. Finally, 500 ns of production simulations were performed in the NPT ensemble for all the systems with periodic boundary conditions. The Nose–Hoover thermostat with a coupling time constant of 0.1 ps was used to maintain the temperature at 298 K throughout the simulation. The isotropic Parrinello–Rahman barostat with the time constant for coupling set to 2 ps was used to maintain pressure at 1 bar. Electrostatic interactions were calculated using the PME method with default values for grid spacing.

2.5. Modeling the structure and dynamics of the RBD-ACE2 complexes for different beta-CoVs using the equilibrium simulation

This study considered a recently reported refined structure of the SARS-CoV2 RBD-ACE2 complex obtained by using extensive molecular dynamics simulation [38]. The modeled RBD structures for different CoVs were aligned on the SARS-CoV2 RBD-ACE2 complex. Then the aligned RBD for each CoV in complexation with ACE2 was considered as the initial RBD-ACE2 docked complex for different CoVs. The structure of each complex was then refined by performing extensive molecular dynamics simulations. Each complex was initially energy minimized in vacuo and then solvated in a triclinic box with a dimension of 100 × 100 × 180 Å³ such that all the protein atoms were at least 10 Å apart from the box edges. The appropriate number of counter ions was added to each solvated box to neutralize each system. Then, each system was minimized in water using 500 steps with the steepest-descent algorithm, followed by 10 ns position-restrained dynamics. Then each system was equilibrated using 2 ns NVT simulation and another 2 ns NPT simulation. The final production simulation was performed for 100 ns for each system in the NPT ensemble at 298 K and a pressure of 1 bar. As mentioned in the previous section (Section 2.4), all the simulation parameters remained the same.

2.6. Estimation of the binding free energy of RBD of different CoVs with human ACE2: potential of mean force (PMF) calculations

The equilibrated structures of all the RBD-ACE2 complexes for different CoVs obtained after the equilibrium simulation mentioned in the previous section were used to calculate the potential of mean force (PMF) using the umbrella sampling method. The spike RBD was pulled from the RBD-ACE2 complex for each CoV along the Z-direction with an interval of 1 Å using an umbrella force constant of 500 kJ.mol⁻¹.nm⁻². The center of mass (COM) distance between RBD and ACE2 was used as the reaction coordinate. A 2 ns equilibration was performed in each umbrella window, followed by a 3 ns production run in the NPT ensemble using the same thermostat, barostat, and associated coupling parameters mentioned in the equilibrium simulation section. Twenty-four windows were considered for each system to sample the entire reaction coordinate. A weighted histogram analysis method (WHAM) [41] was used to construct the PMF profile. Sufficient overlap among all the windows was confirmed by histogram analysis.

3. Results and discussion

3.1. Relative synonymous codon usage pattern of S gene for closely-related beta-CoVs

Relative synonymous codon usage (RSCU) is an important indicator of the balance between mutational biases and selection pressure for translational optimization [42]. Optimal codon usage ensures translational efficiency and accuracy [43]. Gene length, mutation pressure, and natural selection are the critical factors that bias the codon usage pattern [44]. We have analyzed the RSCU patterns of the S gene sequences from 30 beta-CoVs. The RSCU reflects the preference for the usage of a specific codon. The RSCU values >1.0 indicate preferential usage of a codon over other synonymous codons. RSCU values of <1.0 indicate nonpreferred usage of a codon. RSCU values of exactly 1.0 signify no preference.

Interestingly, some codons with high preferential usage are G and C rich at the 1st and 2nd codon positions. Although, the genome is primarily AT/U-rich for all the CoVs [45]. Overall, the codon usage pattern is very similar among the related CoVs for the S gene. However, closer insight reveals intricate differences among different beta-CoV species. SARS-CoV and SARS-CoV2 use human ACE2 for host recognition, but the usage of GGU (G), AUU (H), CCU (L), CCU (P), GGU (R), GGU (V) codons are different among them (Fig. 1). This difference in codon usage patterns indicates different degrees of mutational biases and selection pressure on the S genes of closely related CoVs.

3.2. Genomic composition, mutational bias and selection pressure reveal a distinct pattern for SARS-CoV2 in comparison to other ACE2-user CoVs

Our previous analysis of 82 SARS-CoV2 genomes reveals that the SARS-CoV2 genomes are AT/U-rich with average AT/U content of 62.07 ± 0.004 % [46]. Here, we have analyzed the nucleotide composition of the S genes from different beta-CoVs. Interestingly, the genomic composition of the SARS-CoV2 S gene is an outlier from the rest of the S genes grouping of different CoVs. The GC content is lowest for the SARS-CoV2 S gene. Therefore, the 3-D projections constructed by the first three PCs have been used to cluster different beta-CoVs based on genomic composition. Interestingly, the ACE2 user CoVs WIV1, RsSHC014, and Rs4084 are clubbed closely within the GC-content range of 38–39 % for the S genes, where non-ACE2 user CoVs are also grouped (Fig. 2A). Thus, genomic content is not the sole driver of the human ACE2 usage in beta-CoVs.

The effective number of codons (ENC) and the GC-content at the synonymous third codon position (GC3) provide crucial insight into the dominant forces shaping the codon usage bias. ENC values vary from 20 (Highest bias, where a single codon only encodes one amino acid) to 61 (Unbiased, where all synonymous codons are equally used) [47]. ENC versus GC3 plot provides insight into the influence of base composition constraints on codon usage bias. The S gene of SARS-CoV2 is highly codon-optimized with the lowest ENC values. Among the closely related beta-CoVs, the ENC values vary from 44 to 50, while the GC content at the 3rd codon position varies from 26 to 32 % (Fig. 2B). Thus, compositional constraints strongly influence the codon usage bias for the S genes in the beta-CoV lineages. Again, the distribution reveals that the ACE2 usage trait is not correlated to codon usage bias among beta-CoVs.

The problem is essentially multi-dimensional. There are 61 codons used to code for 20 amino acids. Among them, UGG and AUG codons code for a single amino acid. The remaining 59 synonymous codons variables are used to code for different amino acids in spike proteins from 30 beta-CoVs. Principal component analysis (PCA) is an effective multi-variate analysis tool to decode patterns from highly complex data. The first three components explain ~53 % variance of the dataset. Therefore, the 3-D projections constructed by the first three PCs have been used to cluster different beta-CoVs based on codon usage diversity. Interestingly, ACE2 user CoVs WIV1, RsSHC014, and Rs4084 are clubbed together in the 3-D subspace, revealing a similar codon usage bias of the S gene in this group. SARS-CoV is associated with another human ACE2 binder CoV, Rs4231 (Fig. 2C). Thus, within the ACE2 user CoVs, genomic constraints and mutational bias may play a role in shaping the S gene evolution. However, the SARS-CoV2 S gene is distinctively different in genomic composition and codon usage bias. It is
closely clubbed with two non-ACE2 user CoVs, CoVZXC21 and CoVZC45. Therefore the evolutionary forces that shape the evolution of the S gene of SARS-CoV2 are different from the rest of the ACE2 user CoVs. Furthermore, the parity plot (Fig. 2D) reveals that all the CoVs are distributed away from the origin point (0.5, 0.5), implying the involvement of mutation pressure and evolutionary processes in shaping the evolution of the S gene in related beta-CoVs.

Codon adaptation index (CAI) is an important indicator of host adaptation and implicates the expression level of pathogenic proteins within the host. Higher CAI values indicate strong adaptation to a particular host [48,49]. We have calculated the CAI of the S gene for 30 related Beta-CoVs to the human host. Interestingly, all the beta-CoVs

Fig. 1. A heatmap representation of codon usage of the S genes from closely related CoVs.

Fig. 2. (A) Variations of genomic composition of the S genes from 30 closely related beta-CoVs. (B) ENC values are plotted against the GC3s for the S genes from 30 closely related beta-CoVs. (C) Principal Component Analysis (PCA) of the codon usage pattern of the S genes from 30 closely related beta-CoVs. Parity rule-2 bias plot (D) and Codon Adaptation Index (CAI) plot (E) of the S genes from 30 closely related beta-CoVs are shown.
capable of invading human hosts using ACE2 exhibit high CAI values. Most CoVs capable of binding human ACE2 show a CAI value of 0.7 or higher (Fig. 2E). The CAI value of LYRa11 is 0.680. Letko et al. showed that the RBD of LYRa11 can bind to human ACE2 but with the lowest affinity among all the studied human ACE2-binder lineage B beta-CoVs [26]. All the non-ACE2 user CoVs show lower CAI values (0.638–0.668). Interestingly, the CAI values for CoVZXC21 and CoVZC45 are 0.682 and 0.679, respectively. However, both of them can’t bind to human ACE2. Therefore, the S gene is highly codon-optimized in those CoVs capable of invading humans to ensure high replicative kinetics within the host cell. However, this degree of adaptation with the host does not correlate with human ACE2 receptor binding affinity. Thus the host-induced mutational bias is not dictating the ACE2 binding affinity. Therefore, the evolution of the SARS-CoV2 S gene to gain high affinity during receptor binding is not influenced by host adaptation-driven selection bias. Instead, genomic composition constraints maintain codon adaptation to ensure translational efficiency in the human host.

### 3.3. SARS-CoV2 is evolutionarily divergent from other ACE2-user CoVs

We have computed the evolutionary divergence of the S genes from 30 lineage B beta-CoVs by calculating the number of base substitutions per site between a pair of genes using the Maximum Composite Likelihood model. As evident from Fig. 3A, all the S genes from ACE2 user CoVs are closer in terms of evolutionary distance and clubbed in a group. Most of the remaining S genes from the non-ACE2 user CoVs form a separate group. Interestingly, the SARS-CoV2 S gene is evolutionarily distant from both groups. We further categorized S genes in terms of their pattern disparity index. The probability of rejecting/accepting the null hypothesis, i.e., sequences evolved with the same substitution pattern, judged by the disparity index, is shown in Fig. 3B. ACE2 user CoVs display a distinct substitution pattern compared to the non-ACE2 user CoVs. Within the ACE2 user group, CoVs display a similar substitution pattern in the S gene. Interestingly, the S gene of SARS-CoV2 evolves with a different degree of pattern disparity compared to the ACE2 user and non-ACE2 user CoVs. Thus, genomic composition, selection bias, evolutionary distance, and pattern disparity index clearly show that the SARS-CoV2 S gene evolution follows a different pattern than other ACE2-user CoVs. These observations follow Wells et al. that the SARS-CoV and SARS-CoV2 evolve differently, as evident from the time-calibrated phylogeny [18].

### 3.4. Role of recombination on human ACE2 recognition for beta-CoVs

Sequence comparisons of the spike glycoprotein from ACE2 user CoVs and SARS-CoV2 are shown in Fig. 4. We have also considered the spike sequence from CoVZXC21 and CoVZC45, two closely related CoVs to SARS-CoV2. Also, the spike sequence of a distantly related CoV, BM48/31/BGR/2008, has been considered. Here, we have focused on aligning sequences of the RBD regions. Interestingly, in the region that directly interacts with the ACE2, the receptor-binding motif (RBM), the sequences are highly conserved among ACE2 user CoVs, apart from the SARS-CoV2. SARS-CoV2 sequence diverges significantly from the other ACE2 user beta-CoVs. Compared to non-ACE2 user CoVs, the spike sequence of ACE2 users contains two insertion regions in the RBM. This observation signifies the role of recombination in the evolution of spike protein to gain human ACE2 usage. However, the sequence of insertions in the spike protein of SARS-CoV2 is different from the remaining ACE2 user CoVs. These observations show that the origin of the recombination event for the SARS-CoV2 is different compared to the other ACE2 user beta-CoVs. Thus SARS-CoV2 S gene evolution is not coherent along the pathway of natural selection among other ACE2 users CoVs. The loss of ACE2 receptor binding ability of BM48/31/BGR/2008 is due to the absence of the first insertion segment and another single residue insertion (Fig. 4). A pictorial representation of the alignment of the entire spike protein sequences from 30 beta-CoVs encoded on the spike monomer of the prefusion trimeric complex is shown on the right-hand side of Fig. 4. As evident from the figure, the tip of the RBD that interacts with ACE2 (RBM region) is highly variable, which is primarily contributed by the two insertion regions in ACE2 user CoVs and sequence divergence between ACE2 user and non-ACE2 user CoVs.

Based on the inframere codon alignment, we then performed a phylogenetic analysis of the 30 lineage B beta-CoVs. Sequences were initially aligned and then analyzed to identify the best-suited nucleotide substitution model by performing Maximum Likelihood fits on 24 different models. The general time-reversible (GTR) model with non-uniformity of evolutionary rates among sites modeled using a discrete Gamma distribution (+G) with a shape parameter of 0.309 was the best-suited substitution model judged by the lowest Bayesian Information Criterion score. The maximum likelihood method has been used to generate the phylogenetic tree using the GTR + G substitution model, and the results are shown in Fig. 5A.

Interestingly, all the ACE2 user CoVs are clubbed together in a clade. The only remaining ACE2 user, SARS-CoV2, is clubbed with the non-ACE2 user CoVs, CoVZXC21, and CoVZC45 in the phylogenetic tree. The ability of SARS-CoV2 spike protein to bind human ACE2 originated

---

**Table**: Summary of S gene index and Name

| Index | Name    | Index | Name    |
|-------|---------|-------|---------|
| 1     | SARS-CoV2 16 | YN2013 |
| 2     | WIV1 17 | R4092 |
| 3     | LYRa11 18 | J12012 |
| 4     | Rs7327 19 | H62013 |
| 5     | Rs4231 20 | HeB2013 |
| 6     | Rs4084 21 | BM48:31/BGR/2008 |
| 7     | Rs7SHC014 22 | Yunnan2011 |
| 8     | As6526 23 | Shaanxi2011 |
| 9     | Rs4237 24 | Bl_CoV273/2005 |
| 10    | Rs4081 25 | Rp3_2004 |
| 11    | Rs4247 26 | R12004 |
| 12    | HKU3-8 27 | Bl_CoV273/2005 |
| 13    | HKU3-13 28 | CoVZXC21 |
| 14    | GX2013 29 | CoVZC45 |
| 15    | Longquan-140 30 | SARS-CoV |

**Fig. 3.** (A) Evolutionary divergence of 30 S genes from different beta-CoVs is shown in terms of the number of base substitutions per site between a pair of genes. (B) Homogeneity in the substitution patterns between the 30 S genes is shown.
Fig. 4. Multiple sequence alignment (MSA) of the RBD regions of the spike proteins of a few representative beta-CoVs. The right-hand side shows the conserved and variable regions of the monomeric spike protein obtained from the MSA of spike proteins of 30 beta-CoVs encoded in the prefusion trimeric complex.

Fig. 5. (A) Maximum likelihood tree of in-frame codon alignment of S gene sequences from 30 different beta-CoVs. ACE2 user clade is defined. SARS-CoV2 is also highlighted in the tree. (B) Episodic diversifying selection across the phylogenetic tree branches determined by the Adaptive branch-site random effects likelihood method is shown.
from recombination events, and the sources of recombination are not similar to the other ACE2 user CoVs. We have further analyzed the selection pressure across the phylogenetic tree branches. To detect selection limited to particular lineages, we have used the aBS-REL (Adaptive branch-site random effects likelihood) method to independently estimate the probability of selection at each branch of the phylogeny. The aBS-REL method predicts evidence of episodic diversifying selection on 10 out of 57 branches in the phylogeny (Fig. 5B). Significance was assessed using the Likelihood Ratio Test at a $p \leq 0.05$ after correcting for multiple testing. Graphical representation of phylogenetic tree obtained using the GTR + G substitution model has been color-coded in terms of episodic diversifying selection and shown in Fig. 5B. Three nodes that experience a high degree of episodic diversifying selection are labeled in the figure. All three nodes of the phylogenetic tree belong to the ACE2 user lineage. Thus strong positive selection bias shaping the evolution of the $S$ gene in the CoVs belongs to the ACE2 user lineages. However, the node that diverges into SARS-CoV2, CoVZXC21, and CoVZC45 experience moderate selection bias.

### 3.5. Sequence divergence in two insertion regions frames the shape of the receptor binding motif (RBM): Evidence from molecular dynamics simulation

To understand how recombination and natural selection shape the spike protein's structure, dynamics, and ability to bind human ACE2, we have used extensive all-atom molecular dynamics simulation and binding free energy calculations. We choose the spike protein of SARS-CoV2 and SARS-CoV, RsSHC014 as the representative from the ACE2 user clade. From non-ACE2 user clades, we choose YN2013, RF4092, and R4081. Also, the spike protein of another non-ACE2 user CoV, CoVZXC21, which shares the same clade with SARS-CoV2, has been considered. Since the RBD domain of CoVs can fold independently, we have considered only the RBD region in this study. Structures of 7 CoV RBDs have been modeled from the sequence using the crystal structure of SARS-CoV2 RBD as a template. The structure and dynamics of each RBD have been studied using 500 ns equilibrium simulation.

Conformational dynamics of the RBD obtained from the simulation are represented in the 2-D space defined by root mean square deviation (RMSD) and radius of gyration (Rg). The structure and dynamics of the RBD of SARS-CoV2 and SARS-CoV are very similar in the 2-D space as they occupy a very identical region in the 2-D subspace. RBD dynamics of non-ACE2 user CoVs like YN2013, RF4092, and R4081 are very different. They are widely distributed in the 2-D subspace, indicating sampling of diverse conformational space during the simulation (Fig. 6A). The dynamics observed in MD simulation are very complex, where small amplitude motions are clubbed together with large-scale functional motions. Principal Component Analysis (PCA) is an efficient data reduction technique to mine large-amplitude functional motions from the complex simulations trajectory. The first few components are generally associated with high eigenvalues and represent functional motions [50,51]. We chose the first two components (PCs) to define the 2-D subspace and projected each simulation trajectory on the 2-D subspace (Fig. 6B). The figure shows that the functional large amplitude motion is more evident for the spike protein of both SARS-CoV and SARS-CoV2, whereas for other spike proteins, the dynamics are more confined in the 2-D space. Thus the observed RMSD variations of non-ACE2 user spike RBD are primarily due to loop fluctuations.

To further decode the dynamics of the RBD of different CoVs, we have performed cluster analysis on each simulation trajectory using a 0.15 nm RMSD cut-off to identify the most populated structure. The average structure from the most populated cluster is considered as the most probable solution structure of the CoV RBD. The structural alignment of the solution structures of all seven spike RBDs from different CoVs is shown in Fig. 6C. The core structure of the RBD is highly stable, and there is almost no difference in the structure of different CoVs from the ACE2 user and non-user groups in this region. However, the loop

---

**Fig. 6.** (A) 2-D representation of the conformational dynamics of the spike protein RBDs from seven different CoVs defined by root mean square deviation (RMSD) and radius of gyration (Rg). (B) 2-D projection of each simulation trajectory on the 2-D subspace defined by the first two principal components. (C) Structural alignment of the most populated solution structures of all seven spike RBDs from different CoVs obtained from the cluster analysis of simulation trajectory is shown.
dynamics are very different among different CoVs. Notably, the RBM shows distinct differences among ACE2 users and non-user CoVs (highlighted in the inset). The RBM is primarily unstructured and comprised of long loopy sections. The anti-parallel \( \beta \)-sheet at the middle of the RBM is conserved among all the CoVs irrespective of its receptor usage. There are two loopy overhangs on the two opposite sides of the \( \beta \)-sheet. The loop on the right-hand side of the \( \beta \)-sheet adopts a very similar structure for both SARS-CoV and SARS-CoV2. Another ACE2 region represents the binder, RsSCH014, also shares a similar loopy orientation. Notably, the region represents the “Insert 1” of the MSA (Fig. 4). There is a considerable variation in the conformation of the loop region for other CoVs, which can alter the packing of this region with the ACE2 binding interface. On the left-hand side of the \( \beta \)-sheet is a \( \beta \)-ridge representing the “Insert 2” region of the MSA (Fig. 4). CoVs belong to the ACE2 binder clade containing the insertion. SARS-CoV2 also includes the insertion, but this region is missing in its closest homologs from the clade CoVZXC21 and CoVZC45. There is significant sequence variation in the “Insert 2” region between SARS-CoV2 and CoVs belonging to the ACE2 user clade, leading to substantial deviation in the conformation of the loopy ridge region between them. SARS-CoV2 “Insert 2” region adopts more curvy loops than other ACE2-user CoVs. This structural deviation may lead to the alteration in the binding affinity to the human ACE2 among beta-CoVs.

3.6. Hydrogen bonding between the RBD and ACE2 dictates the binding affinity: a PMF analysis

We then computed the effect of insertion and sequence divergence of the spike protein RBM on human ACE2 receptor recognition using the potential of mean force (PMF) calculation. We have computed the binding free energy by calculating the unbinding energy of the spike from the spike-ACE2 complexes of different CoVs to infinite separation where there is no interaction between them. We have calculated the PMF for binding the spike protein of SARS-CoV, RsSCH014 as representative CoVs from the ACE2 user clade. From the SARS-CoV2 clade, we have considered SARS-CoV2 and CoVZXC21 to understand the effect of insertion on human ACE2 recognition. We have considered three other CoVs from the non-ACE2 user clade, YN2013, Rs4081 and Rf4092.

As evident from Fig. 7A, SARS-CoV2 RBD interacts with the human ACE2 with the highest affinity among all the CoVs. However, its closest homolog from the same clade, CoVZXC21, interacts with human ACE2 with significantly reduced affinity. This reduction in the binding affinity is primarily attributed to the two insertions in the SARS-CoV2 RBM. SARS-CoV and RsSCH014 belong to the ACE2-user clade also include the insertion sequences. The RBM sequences between SARS-CoV and RsSCH014 are more similar than the SARS-CoV2. Thus, the spike RBD of the two CoVs from the ACE2 user clade binds to the human ACE2 with comparable affinity. However, SARS-CoV interacts more preferably with human ACE2 than the RsSCH014. However, the binding affinities of the spike RBD to the human ACE2 for CoVs belonging to the ACE2-user clade is significantly lower than the SARS-CoV2. SARS-CoV2 and CoVs belonging to the ACE2 user clade contain the two insertion sequences, but there is considerable sequence divergence between them. Thus, the gain in binding affinity for SARS-CoV2 spike protein is contributed to the sequence divergence in the crucial region of the RBM that interacts with the ACE2 binding interface. The significant difference in the sequence of the RBM region of the spike protein indicates the different evolutionary origin of the S gene between SARS-CoV2 and other CoVs from the ACE2 user clade. Expectedly, spike RBDs from the non-ACE2 user CoVs bind with low affinities to the human ACE2. We then decode the critical factors that dictate the human ACE2 receptor usage of beta-CoVs. We have calculated the stability of spike RBD upon human ACE2 binding using the equilibrium simulation of each of the seven spike-ACE2 complexes for different CoVs (Fig. 7B). Spike RBD remains highly stable after ACE2 binding for ACE2 user CoVs like SARS-CoV, SARS-CoV2, and RsSCH014. Throughout most of the simulation trajectory for these three CoVs, RBD RMSD variations are observed within 3 Å. For non-ACE2 user CoVs, like Rs4081 and CoVZXC21, a significant increase in the RMSD has been observed for spike RBD after ACE2 binding. Thus these two CoVs do not form a stable spike-ACE2 complex. However, two other non-ACE2 users, YN2013 and Rf4092, also form stable spike-ACE2 complexes, evident from the low RMSD distribution of the RBD-ACE2 complexes throughout the simulation. Thus, forming a stable complex is not the sole dictator of ACE2 binding affinity.

![Figure 7](image_url)

**Fig. 7.** (A) Potential of mean force (PMF) calculations of CoVs RBD binding to the human ACE2 are shown. (B) RMSD distribution of RBDs obtained from the equilibrium simulation of the RBD-ACE2 complexes of different CoVs. (C) The number of hydrogen bonds between RBD and ACE2 in each RBD-ACE2 complex obtained from the equilibrium simulation is shown.
We then computed the number of hydrogen bonds between ACE2 and RBD for all the seven CoVs throughout the simulation timescale (Fig. 7C). Interestingly, the pattern is strongly coherent with the ACE2 binding affinity data. SARS-CoV2 RBD forms the highest hydrogen bonds with the receptor, followed by SARS-CoV and RsSCH014. The number of hydrogen bonds between ACE2 and spike RBD also discriminates between ACE2 and non-ACE2 user CoVs. Non-ACE2 user CoVs RBD forms an equivalent number of hydrogen bonds with ACE2, comparable to another ACE2 user CoV, RsSCH014. However, the distribution is broad for CoVZXC21, which signifies its fluctuating nature. As the RBD fluctuation for this CoV is higher upon complex formation with ACE2, evident from RMSD distribution, the RBD-ACE2 hydrogen bond network is transient. Loss of stability in hydrogen bonds between RBD and ACE2 accounts for the reduction in the binding affinity for CoVZXC21.

4. Conclusions

Acquiring the human ACE2 receptor usage is the critical event in the zoonosis of coronaviruses, particularly for those CoVs which causes pandemic in human. Letko and co-workers developed a method to study 29 closely related CoVs spike RBD from lineage B of betacoronavirus on their ability to bind human ACE2 [26]. Their data shows that clade 1 of these closely related spike proteins can bind to human ACE2. Here, spike proteins of all the 29 CoVs from lineage B beta-coronaviruses and SARS-CoV2 have been analyzed to understand the origin of human ACE2 receptor usage in this group of closely related beta-CoVs. Receptor usage in coronaviruses is a puzzle. Phylogenetically distinct HCoV-NL63 alphacoronavirus and SARS-CoV recognize human ACE2 [19]. In contrast, Bat SARS-like coronavirus Rp3, phylogenetically closer to SARS-CoV, fails to bind human ACE2 [21]. Understanding the human ACE2 usage using the whole genome phylogeny is not useful as different genome segments undergo recombination from different origins [18,23]. Previously, it has been suggested that the S gene is a better probe to evaluate the receptor usage trait in coronaviruses [18]. Detailed analysis of the genomic composition, codon usage, selection analysis, protein sequence analysis in combination with extensive molecular dynamics simulation, and binding free energy calculation with human ACE2 has been used to evaluate several evolutionary factors on ACE2 usage.

Interestingly, genomic content and codon usage bias show that SARS-CoV2 is not grouped with other human ACE2-binder CoVs. Though all the ACE2 user CoVs contain two insertions in the RBD, the insertion sequences in SARS-CoV2 and other ACE2 user CoVs are vastly divergent, impinging their different recombination origins. This observation agrees with Wells et al., which showed that the SARS-CoV2 lineage most likely evolved from the ancestral ACE2-using lineage [18]. However, often ACE2-user CoVs are clubbed together with the non-ACE2 user in ENc-plot, parity rule-2 plot, and genomic composition analysis, which indicates that the genomic composition and codon usage bias does not correlate with the ACE2 receptor usage. Surprisingly, we observe that all the beta-CoVs capable of invading the human host maintain an optimal codon adaptation index of ~0.7. Thus, genomic constraints and mutational bias ensure rapid replicative kinetics of the S gene, but not the receptor usage trait.

Two insertions in the RBM of the spike protein are very crucial in human ACE2 binding. These two insertions frame the ACE2 binding interface of spike protein and dictate ACE2 usage. However, the sequence divergence in these two regions dictates human ACE2 binding affinity. The adaptive branch-site random effects likelihood method identifies positive selection bias across the ACE2 user CoVs lineages in the phylogeny. We found both recombination and selection bias shape the structure and dynamics of the RBM region of the spike to gain the number of interfacial hydrogen bonds while interacting with the ACE2. Paola et al. recently identified an allosteric modulation region (AMR) present in both SARS-CoV and SARS-CoV2 spike protein which allosterically modulates the binding ability of spike with ACE2 [52]. However, our evolutionary analysis suggests that the selection forces predominately act on the RBM (receptor binding motif) region of the spike RBD. The RBM region is in direct contact with the peptidase domain of ACE2. Therefore any mutation/alteration in the RBM directly influences the ACE2 binding ability of the spike protein. Our data clearly shows that recombination and selection forces shape the spike evolution so that the interfacial hydrogen bonds between RBD and ACE2 are optimized in terms of number and stability.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbadis.2022.166514.

Declaration of competing interest

Authors declare no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgment

The author gratefully acknowledges Microsoft AI for Health program for providing Azure cloud capabilities to carry out the research (Microsoft AI for Health COVID-19 grant ID: 00011000243).

References

[1] W.T. Harvey, A.M. Carabelli, B. Jackson, R.K. Gupta, E.C. Thomson, E.M. Harrison, C. Ludden, R. Reeve, A. Rambaut, S.J. Peacock, D.L. Robertson, C.-19 G.U.K. (COG-U. Consortium, SARS-CoV-2 variants, spike mutations and immune escape, Nat. Rev. Microbiol. 19 (2021) 409–424, https://doi.org/10.1038/s41579-021-00573-0.
[2] D. Focos, F. Maggi, Neutralising antibody escape of SARS-CoV-2 spike protein: risk assessment for antibody-based COVID-19 therapeutics and vaccines, Rev. Med. Virol. (2021), https://doi.org/10.1002/rmv.2321 n/a.
[3] S. Chakraborty, E484K and N501Y SARS-CoV-2 spike mutants increase ACE2 recognition but reduce affinity for neutralizing antibody, Int. Immunopharmacol. 102 (2022), 108424, https://doi.org/10.1016/j.intimp.2021.108424.
[4] R.M. El-Shabasy, M.A. Nayel, M.M. Taher, R. Abdelmonem, K.R. Shoukair, E. R. Kenawy, Three wave changes, new variant strains, and vaccination effect against COVID-19 pandemic, Int. J. Biol. Macromol. (2022), https://doi.org/10.1016/j.ijbiomac.2022.01.118.
[5] Q. Yang, A.A.S. Syed, A. Fahira, Y. Shi, Structural analysis of the SARS-CoV-2 omicron variant proteins, Research 2021 (2021), 9769586, https://doi.org/10.34133/2021/9769586.
[6] J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses, Nat. Rev. Microbiol. 17 (2019) 181–192, https://doi.org/10.1038/s41579-018-0119-3.
[7] J. Miles, K. Blicharz-Domańska, Coronavirus in avian species – review with focus on epidemiology and diagnosis in wild birds, J. Vet. Res. 62 (2018) 249–255, https://doi.org/10.2478/jvres-2018-0025.
[8] D.K.W. Chu, C.Y.H. Leung, M. Gilbert, P.H. Joyner, E.M. Ng, T.M. Tse, Y. Guan, J.S. Peiris, L.L.M. Poon, Avian coronavirus in wild aquatic birds, J. Virol. 85 (2011) 12815–12820, https://doi.org/10.1128/JVI.05838-11.
[9] X.-Q. Mei, P. Qin, Y.-L. Yang, M. Liao, Q.-Z. Liang, Z. Zhao, F.-S. Shi, B. Wang, Y.-W. Huang, First evidence that an emerging mammalian alphacoronavirus is able to infect an avian species, Transbound. Emerg. Dis. (2022), https://doi.org/10.1111/tdb.14535 n/a.
[10] A.N. Vlasova, S.P. Kenney, J. Jung, Q. Wang, L.J. Saif, Delta-coronavirus evolution and transmission: current scenario and evolutionary perspectives, Front. Vet. Sci. 7 (2021) https://www.frontiersin.org/article/10.3389/fvets.2020.626785, https://www.frontiersin.org/article/10.3389/fvets.2020.626785.
[11] L. Duan, Q. Zheng, H. Zhang, Y. Niu, Y. Lou, H. Wang, The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: implications for the design of spike-based vaccine immunogens, Front. Immunol. 11 (2020) https://www.frontiersin.org/article/10.3389/fimmu.2020.576622, https://www.frontiersin.org/article/10.3389/fimmu.2020.576622.
[12] A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell 181 (2020) 281–292.e6, https://doi.org/10.1016/j.cell.2020.02.058.
[13] P.F.N. Souza, F.P. Mesquita, J.L. Amaral, P.G.C. Landim, K.R.P. Lima, M.B. Costa, I. Torresin, S. Chakraborty et al.


[14] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (2016) 237–261, https://doi.org/10.1146/annurev-virology-100115-042301.

[15] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science (80-.). 367 (2020) 1444–1448, https://doi.org/10.1126/science.aaz7626.

[16] M.J. Loeffelholz, Y.-W. Tang, Laboratory diagnosis of emerging human coronavirus infections – the state of the art, Emerg. Microbes Infect. 9 (2020) 747–756, https://doi.org/10.1002/emm.221751. https://doi.org/10.1007/s40493-020-00170-5.

[17] K. Tamura, S. Kumar, Evolutionary distance estimation under heterogeneous M. Letko, A. Marzi, V. Munster, Functional assessment of cell entry and receptor X. Xia, DAMBE6: new tools for microbial genomics, phylogenetics, and molecular C.-C. Hon, T.-Y. Lam, Z.-L. Shi, A.J. Drummond, C.-W. Yip, F. Zeng, P.-Y. Lam, F.C.- C. Leung, Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of M.J. Loeffelholz, Y.-W. Tang, Laboratory diagnosis of emerging human coronavirus infections – the state of the art, Emerg. Microbes Infect. 9 (2020) 747–756, https://doi.org/10.1002/emm.221751. https://doi.org/10.1007/s40493-020-00170-5.

[18] K. Tamura, S. Kumar, Evolutionary distance estimation under heterogeneous M. Letko, A. Marzi, V. Munster, Functional assessment of cell entry and receptor X. Xia, DAMBE6: new tools for microbial genomics, phylogenetics, and molecular C.-C. Hon, T.-Y. Lam, Z.-L. Shi, A.J. Drummond, C.-W. Yip, F. Zeng, P.-Y. Lam, F.C.- C. Leung, Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of M.J. Loeffelholz, Y.-W. Tang, Laboratory diagnosis of emerging human coronavirus infections – the state of the art, Emerg. Microbes Infect. 9 (2020) 747–756, https://doi.org/10.1002/emm.221751. https://doi.org/10.1007/s40493-020-00170-5.