Computational and Structural Biotechnology Journal 19 (2021) 759–766

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Unraveling the molecular basis of host cell receptor usage in SARS-CoV-2 and other human pathogenic β-CoVs

Camila Pontes [a,b,1], Victoria Ruiz-Serra [a,1], Rosalba Lepore [a,2,*], Alfonso Valencia [a,c,2]

[a] Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
[b] University of Brasília (UnB), 70910-900, Brasília - DF, Brazil
[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ARTICLE INFO

Article history:
Received 29 September 2020
Received in revised form 7 January 2021
Accepted 7 January 2021
Available online 12 January 2021

Keywords:
SARS-CoV-2
Spike protein evolution
Phylogenetic analysis
Protein subfamilies
Functional specificity
Specificity Determining Positions

ABSTRACT

The recent emergence of the novel SARS-CoV-2 in China and its rapid spread in the human population has led to a public health crisis worldwide. Like in SARS-CoV, horseshoe bats currently represent the most likely candidate animal source for SARS-CoV-2. Yet, the specific mechanisms of cross-species transmission and adaptation to the human host remain unknown. Here we show that the unsupervised analysis of conservation patterns across the β-CoV spike protein family, using sequence information alone, can provide valuable insights on the molecular basis of the specificity of β-CoVs to different host cell receptors. More precisely, our results indicate that host cell receptor usage is encoded in the amino acid sequences of different CoV spike proteins in the form of a set of specificity determining positions (SDPs). Furthermore, by integrating structural data, *in silico* mutagenesis and coevolution analysis we could elucidate the role of SDPs in mediating ACE2 binding across the Sarbecovirus lineage, either by engaging the receptor through direct intermolecular interactions or by affecting the local environment of the receptor binding motif. Finally, by the analysis of coevolving mutations across a paired MSA we were able to identify key intermolecular contacts occurring at the spike-ACE2 interface. These results show that effective mining of the evolutionary records held in the sequence of the spike protein family can help tracing the molecular mechanisms behind the evolution and host-receptor adaptation of circulating and future novel β-CoVs.

1. Introduction

The emergence of the novel SARS-CoV-2 and its ability to infect humans underscores the epidemic potential of coronaviruses, with the ongoing outbreak being the third documented event in humans in the last two decades [1,2]. To date (January 2021), there have already been more than 84 million reported cases of COVID-19 worldwide and over 1,850,000 deaths (https://covid19.who.int/) [3].

Coronaviruses (CoVs) are enveloped, positive-sense RNA viruses known for infecting a wide range of hosts [4]. Together with SARS-CoV, MERS-CoV, OC43 and HKU1 from the β-CoV genus, and 229E and NL63 from the α-CoV genus, SARS-CoV-2 is the seventh confirmed CoV able to infect humans. While 229E, OC43, NL63 and HKU1 widely circulate in the human population and mostly cause mild disease manifestations in immunocompetent individuals, SARS and MERS CoVs are mainly spread in zoonotic reservoirs, with different intermediate host putatively involved in human transmission [4]. CoV entry into the host cells is mediated by the spike glycoprotein (S), a membrane-anchored homotrimer consisting of two distinct subunits: S1, responsible of viral-host recognition and S2, promoting virus-cell membrane fusion [5,6]. Similar to SARS-CoV, SARS-CoV-2 S has been reported to bind the human angiotensin converting enzyme 2 (hACE2) as a cellular entry receptor via the receptor binding domain (RBD), located in the C-terminal region of the S1 subunit [7].

Abbreviations: CoVs, Coronaviruses; hACE2, human angiotensin converting enzyme 2; RBD, receptor binding domain; NTD, N-terminal domain; SDPs, specificity determining positions; MSA, multiple sequence alignment; MCA, multiple correspondence analysis; MI, mutual information; APC, average product correction; EV, evolutionary rate; RBM, receptor binding motif.

* Corresponding author.
E-mail address: alba.lepore@bsc.es (R. Lepore).

[1] These authors contributed equally.
[2] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2021.01.006

Both SARS-CoV-2 and SARS-CoV belong to the Sarbecovirus subgenus, their spike proteins share more than 75% sequence identity and show similar RBD architecture and mode of binding to the human receptor [8]. However, structural and biophysical evidence showed that SARS-CoV-2 S exhibits a 10- to 20-fold increased affinity to hACE2 compared to its SARS counterpart [7] and, compared to other β-CoVs, possesses unique features such as a putative O-glycosylation site and a polybasic cleavage site [9]. The latters are invoked as key factors to be monitored towards understanding the differential virulence of SARS-CoV-2, while determinants of host adaptation and cellular tropism are mainly found within the S1 subunit [7,10].

Like in SARS-CoV, horseshoe bats represent the most likely candidate animal source for SARS-CoV-2 so far [11]. Yet, the specific mechanisms of cross-species transmission and adaptation to the human host remain unknown. Here, we survey the sequence conservation patterns of the β-CoV spike protein family to elucidate the molecular events involved in the differential host-pathogen interaction patterns observed across β-CoVs. Our results show that receptor specificity is encoded in the amino acid sequences of different β-CoVs spike proteins in the form of a set of specificity determining positions (SDPs).

SDPs are long-established predictors of protein functional sites and a variety of computational methods have been developed to help their identification systematically, based on sequence information alone (for a review see [12,13]). As opposed to fully conserved sites in MSA, which generally point to residues that mediate a common function to all members of a given protein family, SDPs correspond to residues that modulate the functional specificity of different protein subfamilies, and their implication for catalytic activity, ligand binding and protein interactions has been experimentally validated in a number of cases [14–16]. In this study, we show that the identification of SDPs can provide a valuable tool for tracing the molecular mechanisms behind the evolution and host-receptor adaptation of circulating and future novel coronaviruses.

2. Materials & methods

2.1. Sequence dataset and multiple sequence alignment

A BLAST search was performed against the NCBI nr database using the SARS-CoV-2 spike protein (NCBI Reference Sequence: YP_009724390.1) as a query [17]. The top 1000 significant hits (e-value < 0.05) were selected and clustered to identify non redundant sequences using CD-Hit [18] with the following parameters: -s 0.90 -c 0.98. Cluster representatives were used to build a multiple sequence alignment (MSA) using the MAFFT software [19]. The final alignment contained 135 spike protein sequences which were manually reviewed to annotate information on viral strain, host organism and cell receptors using information extracted from the NCBI Protein database, UniprotKB and the literature.

2.2. Detection of SDPs and spike protein subfamilies.

The S3Det method [20] uses multiple correspondence analysis (MCA) to identify differentially conserved positions and sequence subfamilies within a given MSA, and has been shown to outperform similar methods in the latter task. MSA positions that follow the subfamily segregation are defined as SDPs of the family. Here, the unsupervised mode of S3Det was used in a two-level decomposition analysis to identify SDPs linked to the spike protein family segregation between and within β-CoV subgroups. Phylogenetic analysis was performed both on the full β-CoV MSA and for individual subfamilies using the PhyML method [21] with default parameters.

2.3. Coevolution analysis

Coevolving MSA positions were identified by computing the MI-APC [22]. MI-APC is a mutual information (MI)-based score corrected by the average product correction (APC) of the background noise and phylogenetic signal. To ensure robust statistics, MSA columns were filtered according to percentage of gaps (>= 50%) and Shannon entropy (<= avg - std), computed as follows:

$$H_i = -\sum_{a \in A} f_i(a) ln(f_i(a))$$

where H_i is the Shannon entropy of the i-th MSA position, $f_i(a)$is the frequency of amino acid a in the i-th MSA position, and A is the alphabet of all possible amino acids.

2.4. Protein domain annotation and enrichment analysis

Domain enrichment analyses were performed by hypergeometric testing and p-values computed as follows:

$$p(x) = \frac{C(m,x) \cdot C(n,k-x)}{C(m+n,k)}$$

where $\times$ is the number of SDPs observed in a given domain of interest, k is the number of SDPs in the test set, m is the length of the domain of interest and $n + m$ is the size of the protein. The analysis was performed both at the subunit and domain level. Annotations on subunits and domain boundaries of the SARS-CoV-2 spike glycoprotein were retrieved from the literature [23] and mapped to the other human β-CoVs spike sequences used in this analysis (SARS-CoV, MERS, OC43 and HKU1) based on the MSA.

2.5. Evolutionary rate analysis

Per-site evolutionary rates (EV) were computed by Rate4Site [24] using as input the full MSA of the β-CoV spike family and phylogenetic tree. SARS-CoV-2 sequence was set as the reference sequence. Raw rate values were used to compute relative rates by normalizing the values to the mean of 1 [25].

2.6. Structural dataset

All the structures employed in this study were retrieved from the PDB (https://www.rcsb.org) using the following PDB codes: 6VSB [7], 6VXX [26], 6LZG [27], 5X5B [28], 5X58 [28], 2AJF [29], 6OHW [30], 6NZK [30], 5X5F [28], 4L72 [31] and 5I08 [32].

3. Results

3.1. SDPs as predictors of receptor specificity across the β-CoV lineage

Here, we aim at analysing the effect of evolutionary constraints in shaping the organisation of the spike protein family across the β-CoV lineage. To this aim, we use a multivariate analysis based protocol which allows the automatic and simultaneous detection of both the protein family segregation and associated amino acid variations, i.e. SDPs.

A phylogenetic analysis was performed based on a MSA of the full-lenght spike sequences from SARS-CoV-2 and other representative β-CoVs. The resulting tree, shown in Fig. 1C, reflects the taxonomic classification of β-CoVs into five subgenera, namely Sarbecovirus, Hibecovirus, Nobecovirus, Embecovirus and Merbecovirus [33]. A first S3Det analysis was performed on individual

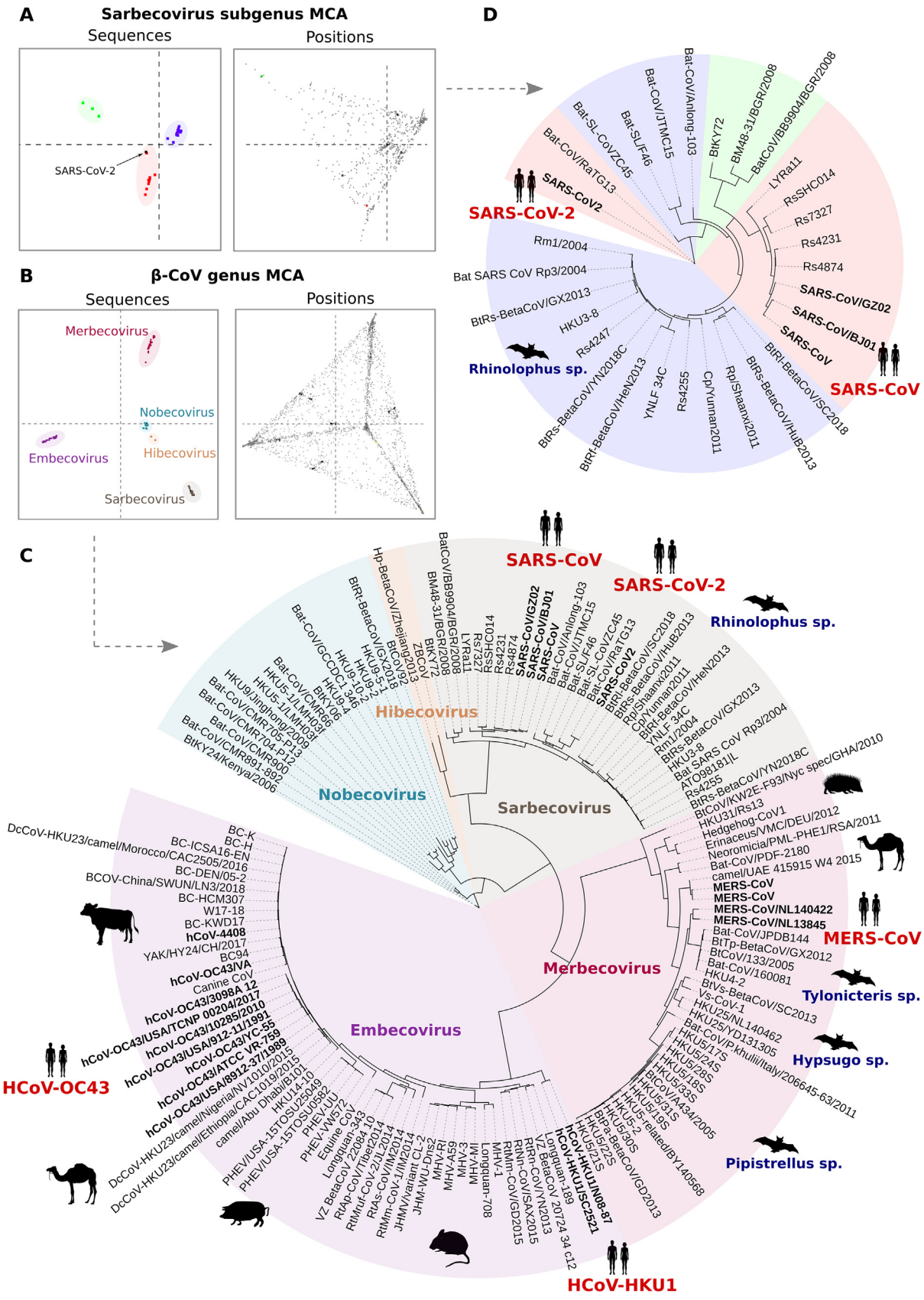

Fig. 1. Results of the S3Det MCA analysis based on the full β-CoVs family and Sarbecovirus subgroup. (A-B) Results of the S3Det MCA analysis showing the subfamily segregation and associated amino acid positions obtained for the Sarbecovirus subgroup and for the full β-CoV family, respectively. (C) Phylogenetic tree obtained for the complete β-CoV spike protein family. S3Det subfamilies are shown in different colors and reflect the phylogenetic classification of Betacoronavirus into five subgenera. (D) Phylogenetic tree obtained for the Sarbecovirus subgroup. S3Det clusters are highlighted in red, green and blue. Both SARS-CoV-2 and RaTG13 are clustered together with SARS-CoV and other members of Sarbecovirus clade 1. Phylogenetic trees were built using PhyML [35]. Spike protein sequences from human pathogenic CoVs are indicated in bold. Host species are shown for some of the nodes as dark silhouettes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

protein subfamilies from the Sarbecovirus, Embecovirus and Merbecovirus subgroups. In the case of Sarbecovirus, we identified three different clusters, two of which correspond to known Sarbecovirus clades [34]. Notably, in contrast to what observed based on phylogenetic analysis, both SARS-CoV-2 and RaTG13 are clustered together with members of Sarbecovirus clade 1, which includes human SARS-CoV and bat SARS-like sequences (Fig. 1-A,D and Supplementary Fig. S1). This result is confirmed by the analysis of similarity scoring matrices (Supplementary Fig. S2) where SARS-CoV-2 and SARS-CoV sequences cluster together and are closer to the green cluster based on the similarity of SDPs, but not when the full-length sequence is considered. Within the Embecovirus group, we identified three clusters: a first cluster corresponding to murine CoVs (MHV), a second containing rat (RtCoVs) and human CoVs (CoV-HKU1), and a third cluster containing the human CoV hCoV-OC43 and other mammalian CoVs (Supplementary Fig. S2-A,C). Within the Merbecovirus, we identified four clusters: a first cluster containing MERS and MERS-like bat CoVs, a second cluster containing hedgehog and bat CoVs, a third cluster containing HKU5 bat CoVs, and a fourth cluster containing one single bat CoV isolate (KW2E-F93) from the Nycteria species (Supplementary Fig. S3-B, D).

The SDPs associated with the subfamily segregation within these three subgenera display a strong domain enrichment within the spike S1 subunit (Fig. 2, Supplementary Table S1). Specifically, the SDPs (n = 28, Supplementary Table S3) of Sarbecovirus subfamilies, containing both SARS-CoV and SARS-CoV-2 spike sequences, are enriched in the RBD whereas in hCoV-OC43 and hCoV-HKU1, the human-pathogenic species belonging to Embecovirus, SDPs (n = 12) fall mainly in an upstream region of the spike, with a significant enrichment in the N-terminal domain (NTD). Also in MERS-CoV we observe a significant enrichment within the S1 subunit. However, as shown in Fig. 2, the SDPs (n = 33) are almost evenly distributed across the NTD and RBD regions, with a significant enrichment in the latter (Supplementary Table S2). Notably, the distribution of the SDPs shows a clear relationship with the cell receptor usage observed among β-CoVs. In particular, hCoV-HKU1 and hCoV-OC43 are known to bind sialic acid receptors on the host cells via the spike NTD [36], SARS-CoV and SARS-CoV-2 recognize the ACE2 receptors via the RBD [37,38], while MERS-CoV has been reported to use both DPP4 [39] and sialic acid receptors [40] via the RBD and NTD domains, respectively.

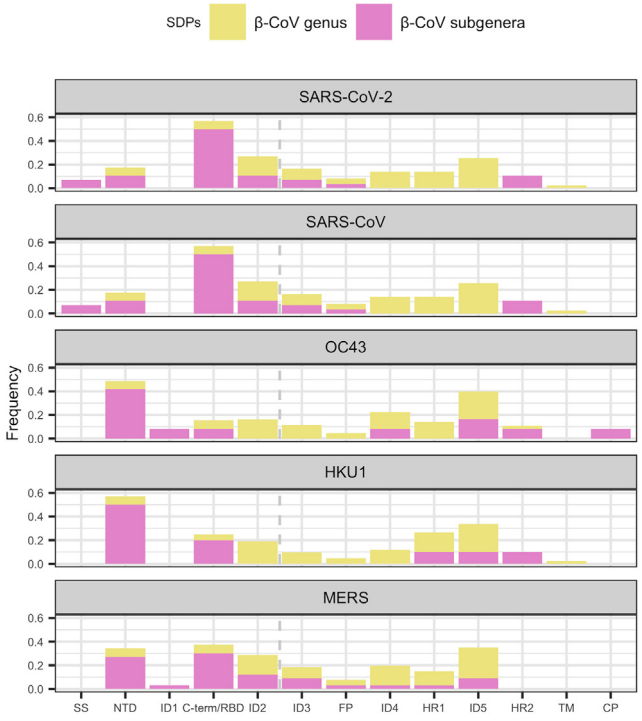

Fig. 2. Frequency distribution of SDPs across different domains of the spike sequence from five human pathogenic β-CoVs. Protein domains are denoted as follows: *SS*, signal sequence; *NTD*, N-terminal domain; *RBD*, receptor binding domain; *FP*, fusion peptide; *HR1*, heptad repeat 1; *HR2*, heptad repeat 2. Interdomain regions are denoted by *ID* followed by an integer according to the order in which they appear in the sequence. Dashed vertical lines denote S1/S2 subunits boundaries.

In summary, the SDPs found within these β-CoV subgenera define a specific region of the receptor binding domains: they are part of, or in direct contact with, the ACE2 interacting surface (Fig. 3, Supplementary Fig. S4); have a relative low impact on protein stability (Supplementary Fig. S5); and have evolved quite recently (Supplementary Fig. S6). These characteristics point to a key role of SDPs in mediating the functional specificity of the protein subfamily, i.e. the recognition of the host-cell receptor.

A second S3Det analysis was performed on the full β-CoV MSA. As shown in Fig. 1A-D, the identified sequence subfamilies are consistent with the phylogenetic classification of Betacoronavirus into five subgenera [33]. In this case, the SDPs linked to the full β-CoV spike family segregation are mainly located in the S2 protein subunit, with a statistically significant enrichment in regions corresponding to interdomain 4, domain HR1 and interdomain 5 (Fig. 2, Supplementary Table S1-S2-S4). These SDPs show characteristics of sites under structural constraints [41], i.e. slowly evolving sites (Supplementary Fig. S6) buried within the 3D structure of the protein (Supplementary Fig. S4) and potentially destabilizing (see results of *in silico* mutagenesis experiments in Supplementary Fig. S5).

Collectively, these results indicate that SDPs capture the functional diversification observed within the individual protein subfamilies, whereby host-cell receptor specificity arises in the context of a structural framework that is specific to each β-CoV phylogenetic group.

3.2. Relationship between SDPs and ACE2 binding across the Sarbecovirus lineage

Structural and mutagenesis studies have shown that the spike RBD of Sarbecovirus contains all the necessary information for host receptor binding and that a few amino acid substitutions in this region can lead to efficient cross-species transmission [34,42]. Binding to ACE2 is clade-specific and occurs at the carboxy-terminal region of the RBD, by an extended concave loop subdomain which forms the interaction interface with the ACE2 N-terminal helix. Notably, both the ACE2-contacting residues and the surrounding amino acids, collectively referred to as the receptor binding motif (RBM), are required to impart human receptor usage within the Sarbecovirus lineage [34].

Consistently, we observe that several SDPs associated with the Sarbecovirus subfamily fall within the RBM, i.e., Y451/Y438, L461/L448, T470/N457, C488/C474, P491/P477, G496/G482, G502/G488, P507/P493, Y508/Y494 on SARS-CoV-2/SARS-CoV sequences, respectively (Fig. 3). Of these, Y494 has been previously reported as critical for ACE2 binding [43]. Two SDPs, namely G496 and G502, fall within the receptor interface forming two hydrogen bonds with the ACE2 K353. Other SDPs, such as L461, T470, C488, P491 and G496 make direct contact with ACE2 contacting residues. Hence, these SDPs are likely to play an important role in ACE2 binding by affecting the local orientation of ACE2 contacting residues. This hypothesis is further supported by results from *in silico* mutagenesis and coevolution analysis (details in next section). Specifically, we tested the effect of amino acid mutations across

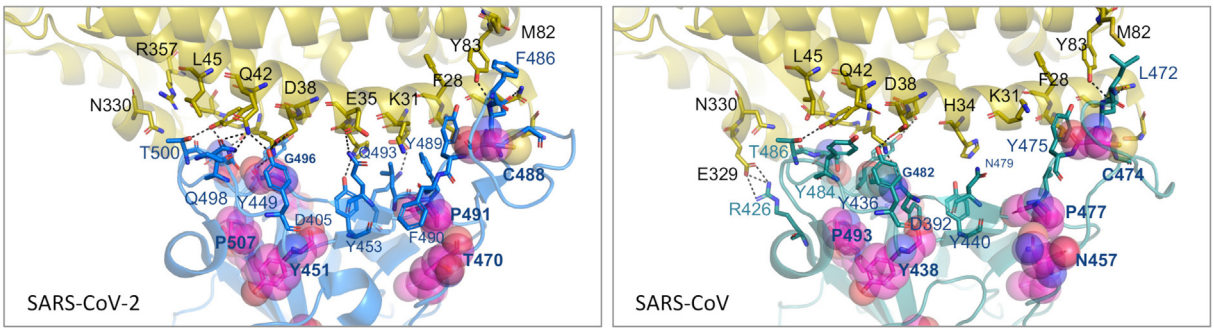

Fig. 3. Structural localization of SDPs. 3D structure of the spike protein from SARS-CoV-2 (blue; PDB ID: 6LZG) and SARS-CoV (green; PDB ID: 2AJF) in complex with the human ACE2 cell receptor (yellow). Amino acid residues at the interface are shown as sticks. Intermolecular contacts are shown as dashed black lines. SDPs are highlighted as spheres. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the RBM by mutating the SARS-CoV-2 sequence to the consensus spike sequence from clade 2, i.e. a clade known to be incompatible with hACE2 usage [34]. Notably, while most substitutions are predicted to have a destabilizing effect on the Spike-hACE2 complex, mutations of SDP residues are predicted to have a significantly larger impact compared to non-SDP residues (Fig. 4).

3.3. Molecular coevolution analysis

Computational methods exploiting coevolution signals in MSAs of protein families are widely used to infer features such as molecular interactions and functional sites [12,20,44,45]. Such signals arise from the specific adaptation between correlated amino acid sites, where changes in one site are potentially compensated by changes in the other. In the case at hand, coevolution signatures are used as markers for the study of the physical interactions occurring between different sites of the spike as well as between the spike and their cognate host cell receptor ACE2. As it can be observed in Supplementary Fig. S7, the strongest intramolecular coevolution signal, considering the top-500 predictions, is observed over the RBD region of the spike (the overall precision of the method is reported in Supplementary Fig. S8). Fig. 5 shows in detail the RBM region, which presents above average precision and recall values of 46% and 4.6%, respectively. Among the SDPs (highlighted in green), the precision is even higher, around 62%, and the recall is 8.7%.

Notably, 18% of SDPs found within the Sarbecovirus subgenus show a coevolution signal with ACE2 contacting residues, namely Y489 (coevolving with C488, P491), Q493 (coevolving with L461, T470, C488, P491) and Q498 (coevolving with G496). These three positions are hubs on the interface with ACE2, making direct contact with positions T27, F28, K31 and Y83, positions K31, H34 and E35, and positions D38, Y41, Q42 and L45 of ACE2, respectively [27]. Particularly, position Q493/N479 in SARS-CoV-2/SARS-CoV has been described to be critical for high affinity binding of both SARS-CoV and SARS-CoV-2 to ACE2 [46,47]. Furthermore, it is interesting to notice that the C488/C474 SDP in SARS-CoV-2/SARS-CoV is an important position for the stability of the RBM as a whole and a complete loss of hACE2 binding *in vitro* has been described when this position is mutated to Alanine in SARS-CoV [48].

We next performed a coarse-grained coevolutionary analysis on a concatenated MSA containing eight spike proteins and their cognate ACE2 receptors. Contact predictions were obtained by computing the MI-APC score for every inter-protein pair of alignment positions, considering the RBD region of the spike protein and the whole ACE2. Interestingly, three RBM positions were found coupled to ACE2 positions among the top-10 MI-APC scores (Supplementary Fig. S9). Specifically, the ACE2 residue H34 was coupled to L455, S494 and Q498 on the spike protein. Additionally, the spike positions R346 and L455 were coupled to ACE2 residues Q24, N61, Q81, M82 and to E23, T27 and H34, respectively. Among these predictions, T27-L455, H34-L455 and H34-S494 correspond to true contacts within 8A distance, while the couplings between R346 and ACE2 positions could be related to long-range effects.

Notably, positions E23 and H34 have been described as crucial to SARS-CoV binding to ACE2 [49]. Also, L455 was described as important for the stabilisation of a binding hotspot between SARS-CoV-2 and ACE2 [47]. Interestingly, a recent study reported that ACE2 variants E23K and T27A are more susceptible to SARS-

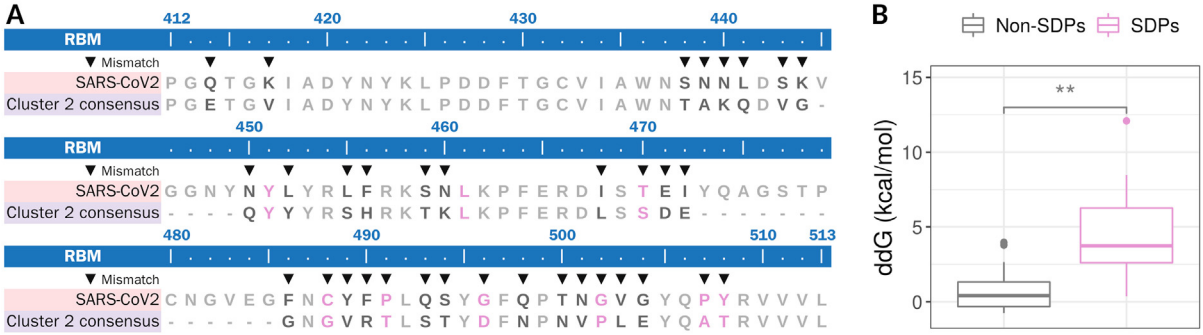

Fig. 4. Mutational impact at SDPs across the RBM. (A) Pairwise sequence alignment of SARS-CoV-2 RBM and consensus sequence of Sarbecovirus clade 2. Black triangles indicate amino acid mismatches. SDP positions are depicted as pink letters. (B) Boxplot distributions of ΔΔG values resulting from mutating SDPs and non-SDPs using FoldX (PDB ID: 6LZG). Significant differences were computed using a Wilcoxon unpaired two-sample test (p-value < 0.01). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

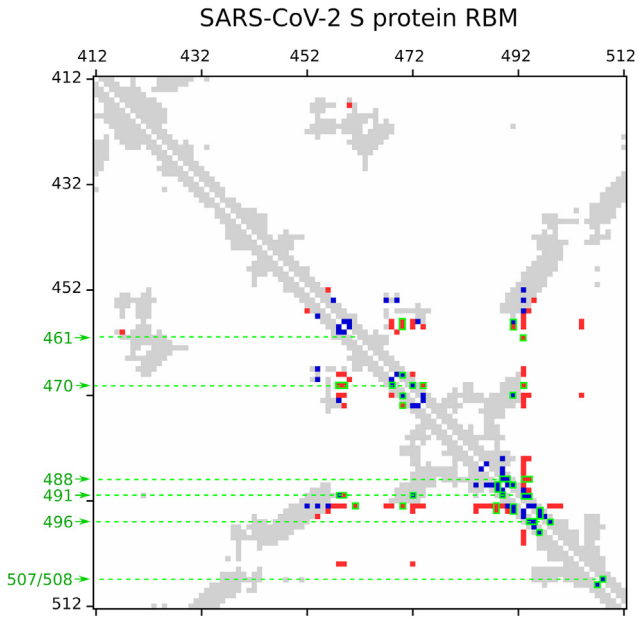

Fig. 5. Coevolution analysis within the RBM. Contact map (8A distance cutoff, any atom) over the RBM of the SARS-CoV-2 spike protein. MI-APC contact predictions (among top 500 scores) are shown in blue (true positives) and in red (false positives). SDPs are highlighted in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

CoV-2, while variant H34R decreases SARS-CoV-2 affinity [50]. These results, despite the limited number of sequences in the concatenated alignment, point to at least three specific RBM viral positions (L455, S494 and Q498 in SARS-CoV-2) likely adapted to their species-specific counterparts.

3.4. Evolution of the SARS-CoV-2 spike protein during the spread in human populations

Since the beginning of the COVID-19 outbreak and the isolation of the SARS-CoV-2 virus, laboratories around the world are continuously isolating viral genomic sequences with unprecedented speed, enabling nearly real-time data sharing of more than 274,000 genomic sequences so far [51]. After discarding partial sequences, a multiple sequence alignment was built based on a total of 273,432 SARS-CoV-2 full-lenght spike sequences isolated from human samples in 187 countries.

Our analysis of missense amino acid variations confirmed earlier reports [52] that most mutations occur within the S1 subunit, with a dominant variant observed at position 614, where more than 50% of samples carry the D614G mutation, followed by mutations A222V and L18F appearing in ~ 11% and ~ 5% of the samples, respectively. Within the RBD, the top frequent mutations are S477N, N439K and N501Y, found in 3,5%, 1%, and 0.5% of the samples, respectively. Notably, these positions fall within the RBM, forming a surface-exposed loop that is proximal to the ACE2 binding surface, and that is absent in Sarbecovirus clades 2 and 3 [34]. Although previous experiments indicate that this loop is *per se* not sufficient to impart ACE2 receptor usage, deletions of this region are associated with reduced spike expression and loss of cell entry [34]. Hence, mutations in this region are expected to impact the stability of the protein, rather than its affinity to the receptor.

Finally, the frequency of variants at SDPs within the RBD is very low so far, with an overall variability that is comparable to that observed in ACE2-contacting residues (Supplementary Fig. S10). Mutations are observed in 12 out 14 SDPs, with the top frequent mutations being T385I (0.01%), Y508H (0.008%) and T470A (0.003%). Mostly, these variants are predicted to be neutral or to destabilize the binding to the receptor (Supplementary Fig. S11B). A similar picture is observed for mutations at SDPs falling outside the RBD, with one notable exception. This is the case of V70, for which at least five independent events of the ΔH69/ΔV70 double deletion have been observed in multiple countries [53], with an overall frequency of 2.4%. Independent studies have reported the ΔH69/ΔV70 double deletion to be potentially associated with immune escape mechanisms [53], lower efficiency of PCR-based tests [54] as well as enhanced transmissibility [53,55], even if the precise mechanisms are still unclear at present. Furthermore, the localization of H69/V70, which lie on a surface exposed loop within the NTD, and the structural analysis of the effect of the double deletion, also suggest a potential alteration on the loop flexibility, and possible long-range, allosteric effects on the neirby RBD region. In line with this, it is worth noticing that the ΔH69/ΔV70 double deletion is frequently co-occurring with other amino acid replacements within the RBD (N439K, Y453F, N501Y, A570), some of which bear additional concerning features, such as increased affinity to ACE2.

In summary, while further studies are vitally needed to unravel the specific mechanisms of viral mutations on virus transmission and potential immune escape mechanisms, rapid identification and monitoring of new variants, specifically those occurring at functionally relevant sites, including the SDPs identified here, remains a priority to inform surveillance worldwide.

4. Discussion

The relationship between protein family segregation and their functional organisation has been extensively investigated for decades and a variety of computational methods have been developed to infer their evolutionary link at the residue level [56–58]. It is therefore relatively straightforward to identify the amino acid positions that modulate the functional specificity of a given enzyme towards a substrate or cofactor [20] or the binding specificity of a protein–ligand or protein–protein interaction [59,60] by the analysis of the differential conservation patterns within the MSA of a protein family. Here we apply this concept to the analysis of the SARS-CoV-2 spike in the context of a MSA of homologous sequences belonging to the β-CoV genus. The SDP analysis is based on a vectorial representation of protein sequences and amino acid positions in a multidimensional space to simultaneously identify the family segregation and the residue positions that better explain the sources of variation of the family [20,61] .

On one hand, the analysis performed at the β-CoV family level led to the identification of five sequence subfamilies, reflecting the known phylogenetic classification of Betacoronavirus into five subgenera [33] (Fig. 1C). On the other hand, the analysis performed on individual β-CoV subgenera, i.e. Sarbecovirus, Merbecovirus and Embecovirus subgroups, allowed a fine-grained classification into subfamily clusters that clearly reflect the functional diversification of the spike protein family, that is, the specificity to different host-cell receptors (Figs. 2-3). Indeed, both the clustering and domain enrichment results of the associated SDPs consistently reproduce the known cell receptor specificities observed across the different β-CoV lineages [31,62–64]. At the level of the Sarbecovirus group, for example, both SARS-CoV-2 and RaTG13 are clustered together with SARS-CoV and other SARS-like sequences from bats (Fig. 1A, D), reflecting the ability of these members of the Sarbecovirus group to bind the ACE2 cell receptor [65]. Notably, the proximity of SARS-CoV-2 and RaTG13 and other SARS-CoV sequences based on key SDP positions is different to what seen based on full sequence phylogenetic analysis, where they form distinct clades. A proximity that is driven by their shared SDPs and it is interpreted

here in terms of their shared ability to bind the human ACE2 receptor.

As it is often the case, functional constraints arise from the requirement of maintaining the interaction of proteins with other macromolecules or ligands. Such constraints translate into specific roles and properties of individual amino acids, or protein sites. In the case at hand, the analysis of the physicochemical, structural and conservation properties of the SDPs of the different subfamilies highlights a pattern that is typical of protein functional sites, as they show high conservation across the protein family, are solvent exposed, and are enriched in the receptor binding domains (Supplementary Figs. S4-6 and Supplementary Table S3). Hence, in order to assess the role of the SDPs in mediating ACE2 receptor usage across the Sarbecovirus group, we set up an *in silico* mutagenesis study and analysed the effect of amino acid mutations across the RBM and their impact on ACE2 binding. Notably, while our results are in line with previous observations [34], they point to specific positions across the RBM that might exert a critical role, either by engaging the receptor through direct intermolecular contacts or by affecting the local orientation of ACE2 contacting residues and the stability of the RBM as a whole.

Collectively, these results point to a key role of SDPs in mediating host cell receptor specificity across β-CoVs and provide, at the same time, a framework for monitoring the evolution of the SARS-CoV-2 specificity to hACE2, as well as the emergence of novel potential cross-species transmission events. As such, it is important to notice that from the analysis of amino acid variations across the circulating SARS-CoV-2 virus, SDPs found in the RBD tend to mutate with a very low frequency, similar to what is seen at ACE2 contacting sites (Supplementary Fig. S10) [66]. This is of relevance, as our results suggest that mutations in SDPs can significantly impact the receptor-binding ability of the spike. Furthermore, the experience in other scenarios has shown that mutating SDPs is, in general, sufficient to transform the properties between two groups of proteins of the same family, i.e. the interchange of the residues occupying the SDPs between two families implies a change in the associated biological properties [14,67–70]. Notable examples include the production of switch-of-function mutants of small GTPases with changed selectivity, or the change of transport specificity between MIP channel proteins by few amino acid substitutions [68,69]. In line with this reasoning, it can be argued that other members of the Sarbecovirus group might have the potential to acquire ACE2 binding ability, as they share substantial similarity in terms of SPDs. This is especially the case of members of the Sarbecovirus clade 3, which despite being phylogenetically distant to SARS-CoV-2, display identical residues in 10 out of 14 SDP positions within the RBD, making them potential candidates for new human infections. In conclusion, the results presented here show that the identification of evolutionary patterns based on the analysis of sequence information alone can provide meaningful insights on the molecular basis of host-pathogen interactions and adaptation. We believe that both the methodology and results presented in this work can provide the basis for follow-up studies analysing the potential routes of mutations that could lead to new adaptation to human hosts and ultimately contribute to better understanding and monitoring of events that are critical to public health concerns worldwide.

Funding statement

This work has received funding from the EXSCALATE4CoV project, from the European Union's Horizon 2020 Research and Innovation Programme, under grant agreement N. 101003551. Camila Pontes was supported by a PhD fellowship awarded by the Brazilian agency CAPES. Victoria Ruiz-Serra was supported by La Caixa Junior Leader Fellowship from Fundació Bancaria La Caixa (LCF/BQ/PI18/11630003). The funding sources had no involvement in the design of the study, data collection, analysis and decision to submit this manuscript for publication.

CRediT authorship contribution statement

Camila Pontes: Formal analysis, Investigation, Data curation. **Victoria Ruiz-Serra:** Formal analysis, Investigation, Data curation. **Rosalba Lepore:** Conceptualization, Supervision, Data curation, Writing - original draft, Writing - review & editing. **Alfonso Valencia:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful to all other members of the Computational Biology group for their valuable feedback and useful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.01.006.

References

[1] Hung LS. The SARS epidemic in Hong Kong: what lessons have we learned?. JRSM 2003;96(8):374–8.
[2] Aleanizy FS, Mohmed N, Alqahtani FY, El Hadi Mohamed RA. Outbreak of Middle East respiratory syndrome coronavirus in Saudi Arabia: a retrospective study. BMC Infect Dis 2017;17(1).
[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20(5):533–4.
[4] Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol 2016;24(6):490–502.
[5] Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol 2020;17(7):765–7.
[6] Li F. Evidence for a common evolutionary origin of coronavirus spike protein receptor-binding subunits. J Virol 2012;86(5):2856–8.
[7] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367(6483):1260–3.
[8] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020. https://doi.org/10.1038/s41586-020-2180-5.
[9] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26(4):450–2.
[10] Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol 2016;3(1):237–61.
[11] Boni MF, Lemey P, Jiang X, Lam T-T-Y, Perry B, Castoe T, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Evol Biol 2020:83.
[12] de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14(4):249–61.
[13] Chakraborty A, Chakrabarti S. A survey on prediction of specificity-determining sites in proteins. Briefings Bioinf 2015;16(1):71–88.
[14] Bauer B, Mirey G, Vetter IR, Garcıa-Ranea JA, Valencia A, Wittinghofer A, et al. Effector Recognition by the Small GTP-binding Proteins Ras and Ral. J. Biol. Chem. 1999;274(25):17763–70.
[15] Morillas M, Gómez-Puertas P, Bentebibel A, Sellés E, Casals N, Valencia A, et al. Identification of conserved amino acid residues in rat liver carnitine palmitoyltransferase I critical for malonyl-CoA inhibition. Mutation of methionine 593 abolishes malonyl-CoA inhibition. J Biol Chem 2003;278(11):9058–63.
[16] Cordente AG, López-Viñas E, Vázquez MI, Swiegers JH, Pretorius IS, Gómez-Puertas P, et al. Redesign of carnitine acetyltransferase specificity by protein engineering. J Biol Chem 2004;279(32):33899–908.

[17] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.
[18] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.
[19] Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 2019;20:1160–6.
[20] Rausell A, Juan D, Pazos F, Valencia A. Protein interactions and ligand binding: From protein subfamilies to functional specificity. Proc Natl Acad Sci USA 2010;107(5):1995–2000.
[21] Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–21.
[22] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–40.
[23] Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res 2020;30:343–55.
[24] Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18 (Suppl 1):S71–7.
[25] Sydykova DK, Jack BR, Spielman SJ, Wilke CO. Measuring evolutionary rates of proteins in a structural context. F1000Res 2017;6:1845.
[26] Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020;181(2):281–292.e6.
[27] Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell 2020;181(4):894–904.e9.
[28] Yuan Y, Cao D, Zhang Y, Ma J, Qi J, Wang Q, et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nat Commun 2017;8(1).
[29] Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 2005;309:1864–8.
[30] Tortorici MA, Walls AC, Lang Y, Wang C, Li Z, Koerhuis D, et al. Structural basis for human coronavirus attachment to sialic acid receptors. Nat Struct Mol Biol 2019;26(6):481–9.
[31] Wang N, Shi X, Jiang L, Zhang S, Wang D, Tong P, et al. Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. Cell Res 2013;23(8):986–93.
[32] Kirchdoerfer RN, Cottrell CA, Wang N, Pallesen J, Yassine HM, Turner HL, et al. Pre-fusion structure of a human coronavirus spike protein. Nature 2016;531 (7592):118–21.
[33] Lu R, Zhao X, Li J, Niu P, Yang Bo, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 2020;395(10224):565–74.
[34] Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol 2020;5(4):562–9.
[35] Guindon S, Delsuc F, Dufayard J-F, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 2009;537:113–37.
[36] Hulswit RJG, Lang Y, Bakkers MJG, Li W, Li Z, Schouten A, et al. Human coronaviruses OC43 and HKU1 bind to 9–acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. Proc Natl Acad Sci U S A 2019;116:2681–90.
[37] Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3.
[38] Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 2003;426(6965):450–4.
[39] Song W, Wang Y, Wang N, Wang D, Guo J, Fu L, et al. Identification of residues on human receptor DPP4 critical for MERS-CoV binding and entry. Virology 2014;471-473:49–53.
[40] Li W, Hulswit RJG, Widjaja I, Raj VS, McBride R, Peng W, et al. Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. Proc Natl Acad Sci USA 2017;114(40): E8508–17.
[41] Toth-Petroczy A, Tawfik DS. Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci 2011;108(27):11151–6.
[42] Menachery VD, Yount Jr BL, Debbink K, Agnihothram S, Gralinski LE, Plante JA, et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 2015;21(12):1508–13.
[43] Chakraborti S, Prabakaran P, Xiao X, Dimitrov DS. The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization. Virol J 2005;2:1–10.
[44] Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 2014:3. https://doi.org/10.7554/eLife.03430.
[45] Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. Proc Natl Acad Sci USA 2016;113(52):15018–23.
[46] Li W, Zhang C, Sui J, Kuhn JH, Moore MJ, Luo S, et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J 2005;24(8):1634–43.
[47] Shang J, Ye G, Shi Ke, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature 2020;581(7807):221–4.
[48] Wong SK, Li W, Moore MJ, Choe H, Farzan M. A 193-Amino Acid Fragment of the SARS Coronavirus S Protein Efficiently Binds Angiotensin-converting Enzyme 2. J. Biol. Chem. 2004;279(5):3197–201.
[49] Han DP, Penn-Nicholson A, Cho MW. Identification of critical determinants on ACE2 for SARS-CoV entry and development of a potent entry inhibitor. Virology 2006;350(1):15–25.
[50] Stawiski EW, Diwanji D, Suryamohan K, Gupta R, Fellouse FA, Fah Sathirapongsasuti J, et al. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility n.d. 10.1101/2020.04.07.024752.
[51] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. Global Challenges 2017;1(1):33–46.
[52] Laha S, Chakraborty J, Das S, Manna SK, Biswas S, Chatterjee R. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. Infect Genet Evol 2020;85:104445.
[53] Kemp SA, Harvey WT, Datir RP, Collier DA, Ferreira I, Carabelli AM, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70 n.d. 10.1101/2020.12.14.422555.
[54] Washington NL, White S, Schiabor Barrett KM, Cirulli ET, Bolze A, Lu JT. S gene dropout patterns in SARS-CoV-2 tests suggest spread of the H69del/V70del mutation in the US n.d. 10.1101/2020.12.24.20248814.
[55] McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, et al. Natural deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape n.d. 10.1101/2020.11.19.389916.
[56] Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–43.
[57] Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Mol Biol 1995;2(2):171–8.
[58] Laskowski RA, Watson JD, Thornton JM. Protein Function Prediction Using Local 3D Templates. J Mol Biol 2005;351(3):614–26.
[59] Lichtarge O, Bourne HR, Cohen FE. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. J Mol Biol 1996;257(2):342–58.
[60] del Sol Mesa A, Pazos F, Valencia A. Automatic Methods for Predicting Functionally Important Residues. J Mol Biol 2003;326(4):1289–302.
[61] Pappalardo M, Juliá M, Howard MJ, Rossman JS, Michaelis M, Wass MN. Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. Sci Rep 2016;6(1). https://doi.org/10.1038/srep23743.
[62] Lu G, Hu Y, Wang Q, Qi J, Gao F, Li Y, et al. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature 2013;500 (7461):227–31.
[63] Wang Q, Qi J, Yuan Y, Xuan Y, Han P, Wan Y, et al. Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. Cell Host Microbe 2014;16:328–37.
[64] Yang Y, Du L, Liu C, Wang L, Ma C, Tang J, et al. Receptor usage and cell entry of bat coronavirus HKU4 provide insight into bat-to-human transmission of MERS coronavirus. Proc Natl Acad Sci 2014;111(34):12516–21.
[65] Mou H, Quinlan BD, Peng H, Guo Y, Peng S, Zhang L, et al. Mutations from bat ACE2 orthologs markedly enhance ACE2-Fc neutralization of SARS-CoV-2 n.d. 10.1101/2020.06.29.178459.
[66] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics 2007;23:1875–82.
[67] Bradley D, Beltrao P. Evolution of protein kinase substrate recognition at the active site. PLoS Biol 2019;17. e3000341.
[68] Lagrée V, Froger A, Deschamps S, Hubert J-F, Delamarche C, Bonnec G, et al. Switch from an Aquaporin to a Glycerol Channel by Two Amino Acids Substitution. J Biol Chem 1999;274:6817–9.
[69] Do Heo W, Meyer T. Switch-of-Function Mutants Based on Morphology Classification of Ras Superfamily Small GTPases. Cell 2003;113(3):315–28.
[70] Rodriguez GJ, Yao R, Lichtarge O, Wensel TG. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc Natl Acad Sci USA 2010;107(17):7787–92.