Supplementary Figures and Tables

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,1,*] , Alfonso Valencia[1,c,1]

[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

[*]Corresponding author: alba.lepore@bsc.es
[1]These authors contributed equally
[2]These authors contributed equally

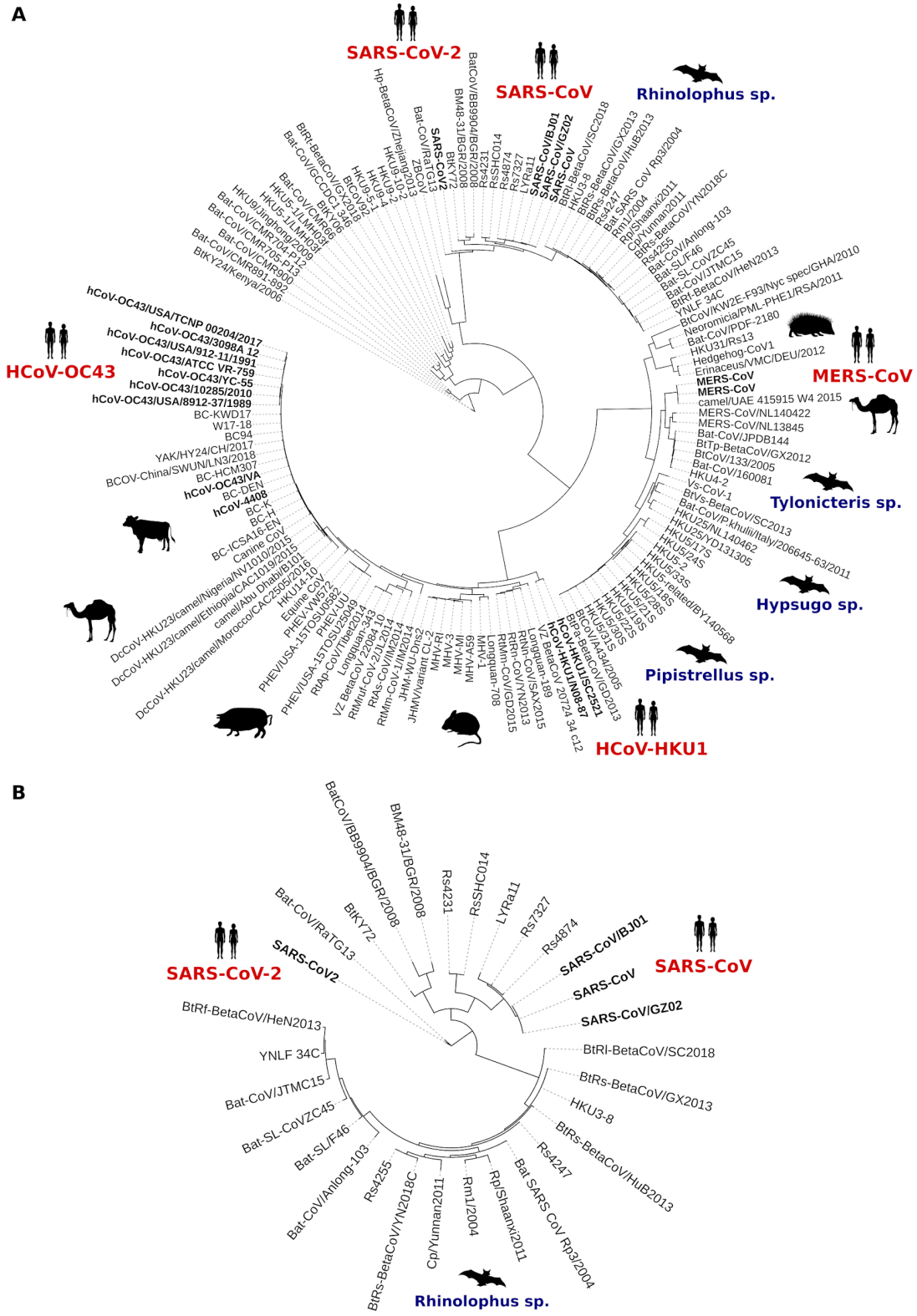

Figure S1. RBD-based phylogeny. Phylogenetic trees of Betacoronavirus genus (A) and Sarbecovirus subgenus (B) based on spike protein RBD sequences. Both trees were obtained using PhyML. Spike protein sequences from human pathogenic CoVs are in bold. Host species are shown for some of the nodes as dark silhouettes.

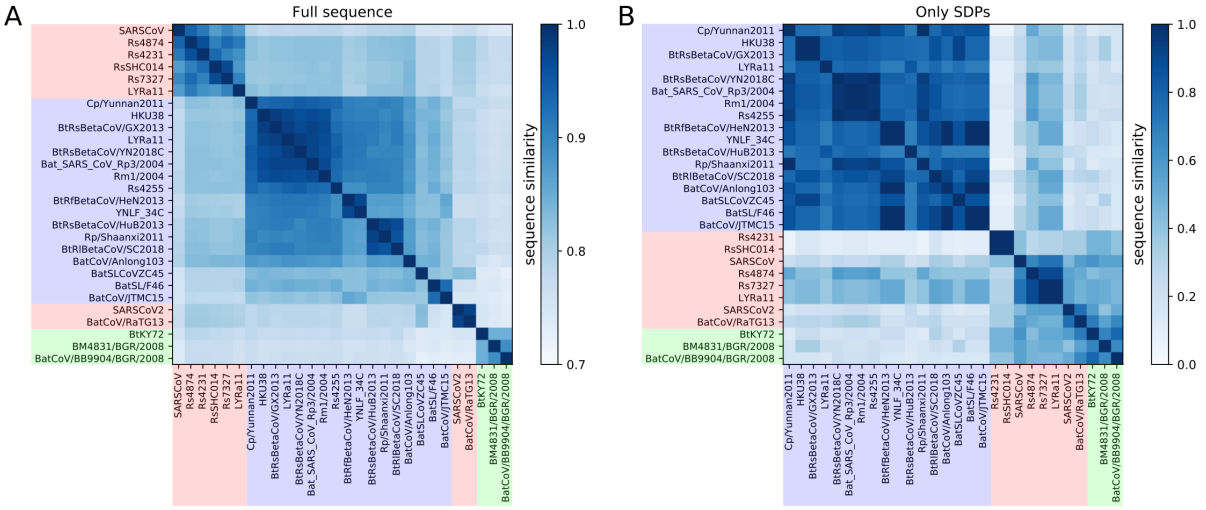

Figure S2. All-against-all sequence similarity computed based on Blosum64 substitution matrix. Sequence similarity matrix computed based on the full-length protein sequences (A) and SDPs (B). Sequence labels are colored according to their respective S3Det cluster.

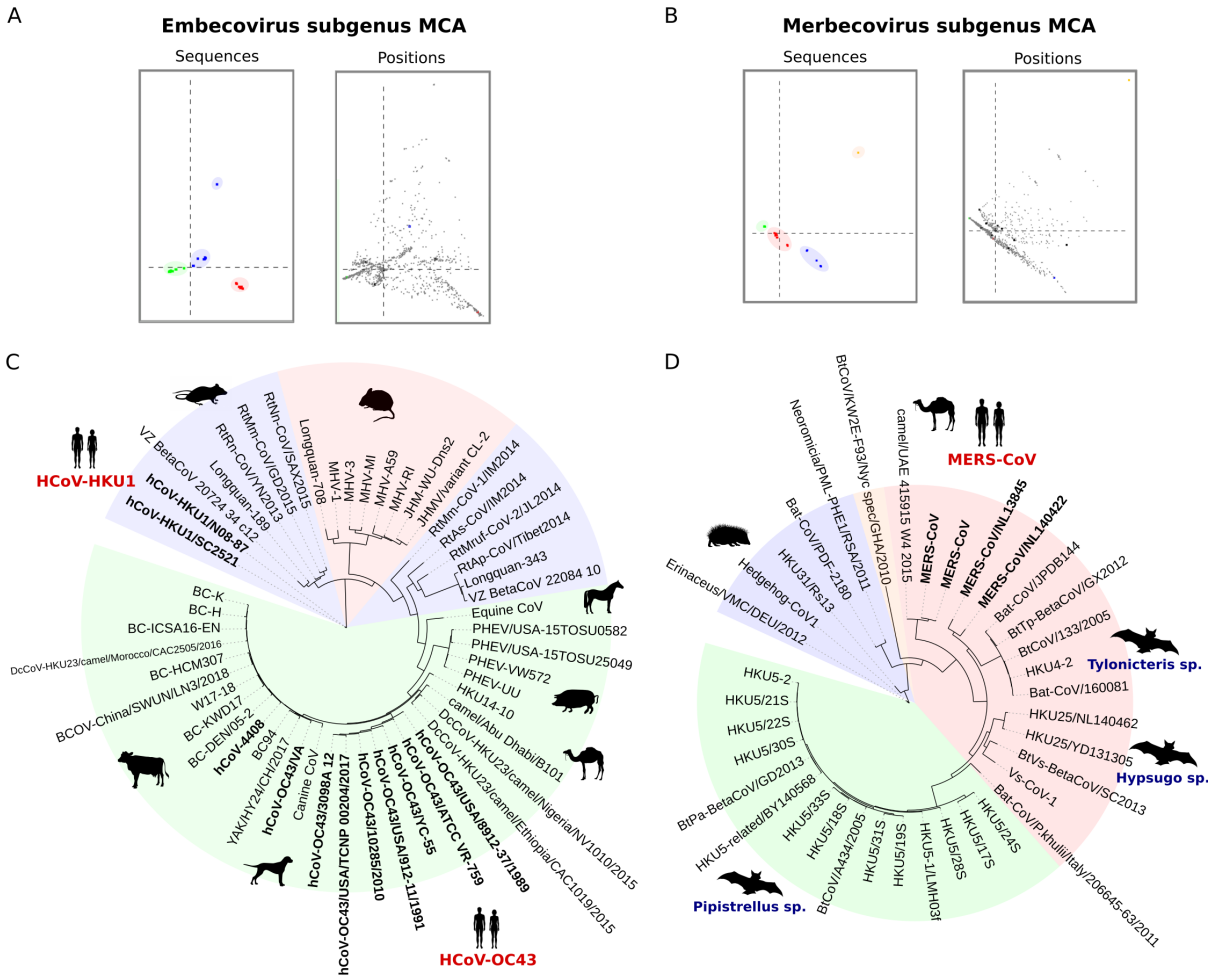

Figure S3. Results of the S3Det MCA analysis for the Embecovirus and Merbecovirus subgroups. Subfamily segregation and associated amino acid positions obtained for Embecovirus (A) and Merbecovirus (B). Phylogenetic tree of Embecovirus (C) and Merbecovirus (D) spike protein sequences obtained using PhyML21. S3Det clusters are highlighted in red, blue, green and orange. Spike protein sequences from human pathogenic CoVs are indicated in bold. Host species are shown for some of the nodes as dark silhouettes.

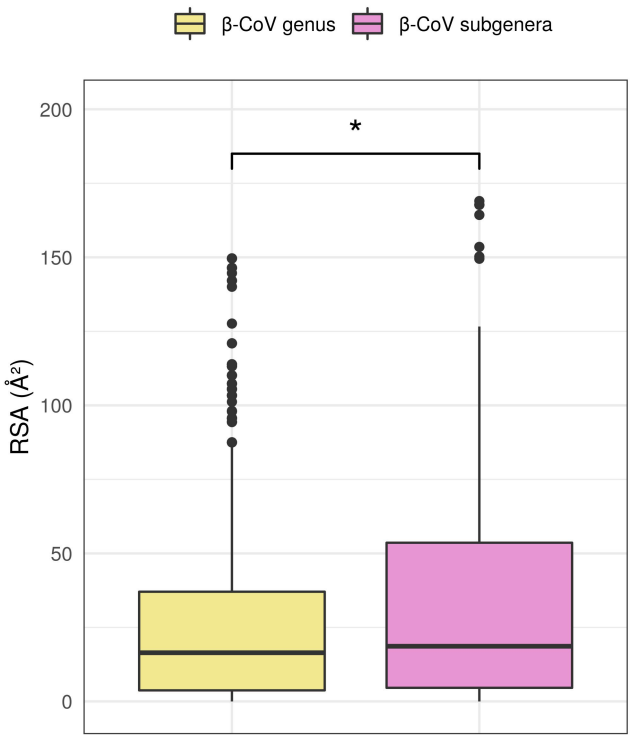

Figure S4. Boxplot distribution of per-residue relative solvent accessible area at SDPs. SDPs belonging to the β-CoV genus and subgenera are indicated in yellow and pink, respectively. Significant differences were computed using a Wilcoxon unpaired two-sample test (p-value < 0.05). RSA, relative solvent accessible area.

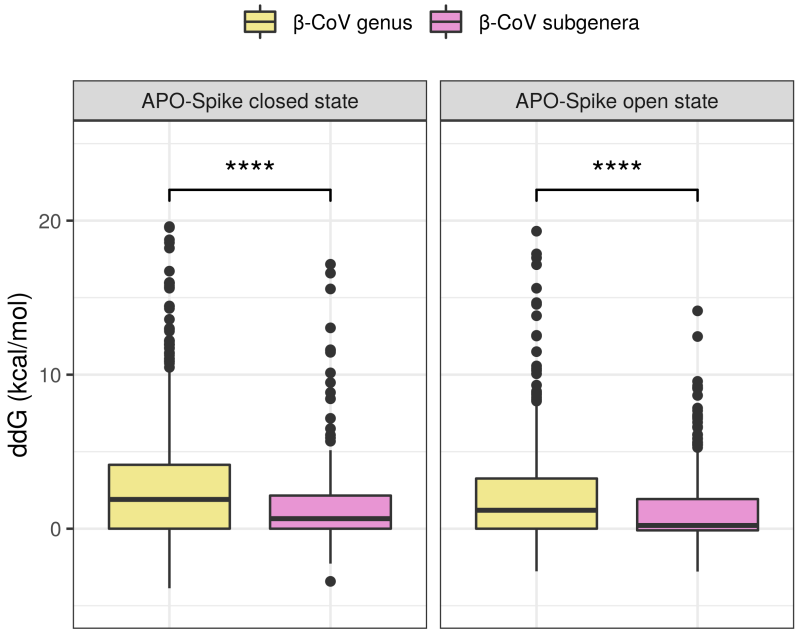

Figure S5. Boxplot distribution of ΔΔG values upon mutation of SDPs. SDPs belonging to the β-CoV genus and subgenera are indicated in yellow and pink, respectively. ΔΔG values were computed using FoldX (PDB ID: 6VXX, 6VSB). Significant differences were computed using a Wilcoxon unpaired two-sample test (p-value < 0.0001).

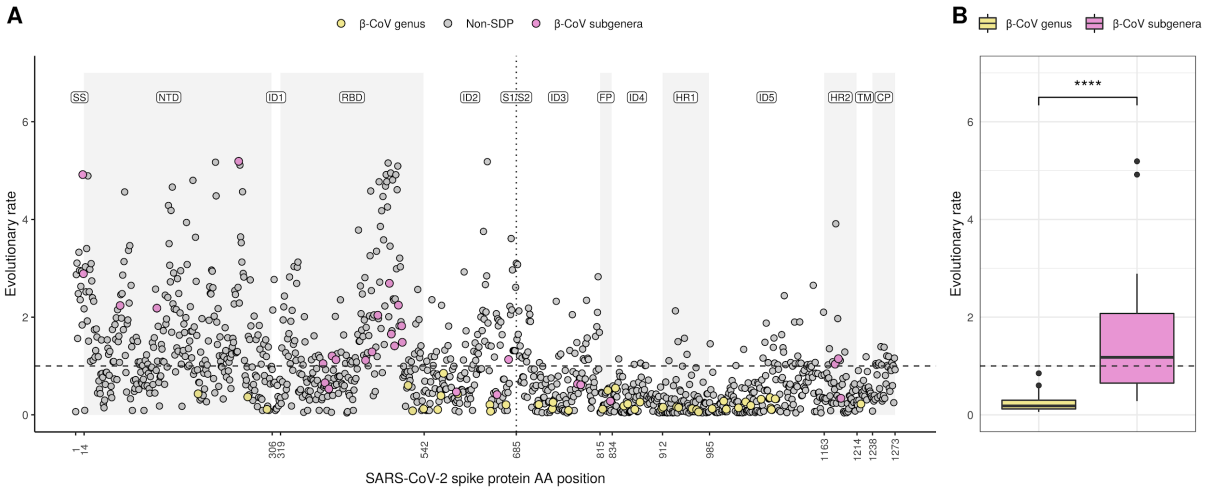

Figure S6. Evolutionary rate analysis of the SARS-CoV-2 spike protein. (A) Site-specific evolutionary rate of the SARS-CoV-2 spike protein as computed by Rate4Site. The dashed horizontal line indicates the mean EV rate. Yellow and pink dots indicate SDPs linked to the β-CoV genus and subgenera, respectively. Protein domains are highlighted by a grey shade and 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; TM, transmembrane region; CP, cytoplasmic. Interdomain regions are denoted by ID followed by an integer according to the order in which they appear in the sequence. The dotted vertical line denotes the S1/S2 subunits boundary. (B) Boxplot distributions of site-specific evolutionary rates computed on the two sets of SDPs. Significant differences were computed using a Wilcoxon unpaired two-sample test (p-value < 0.0001).

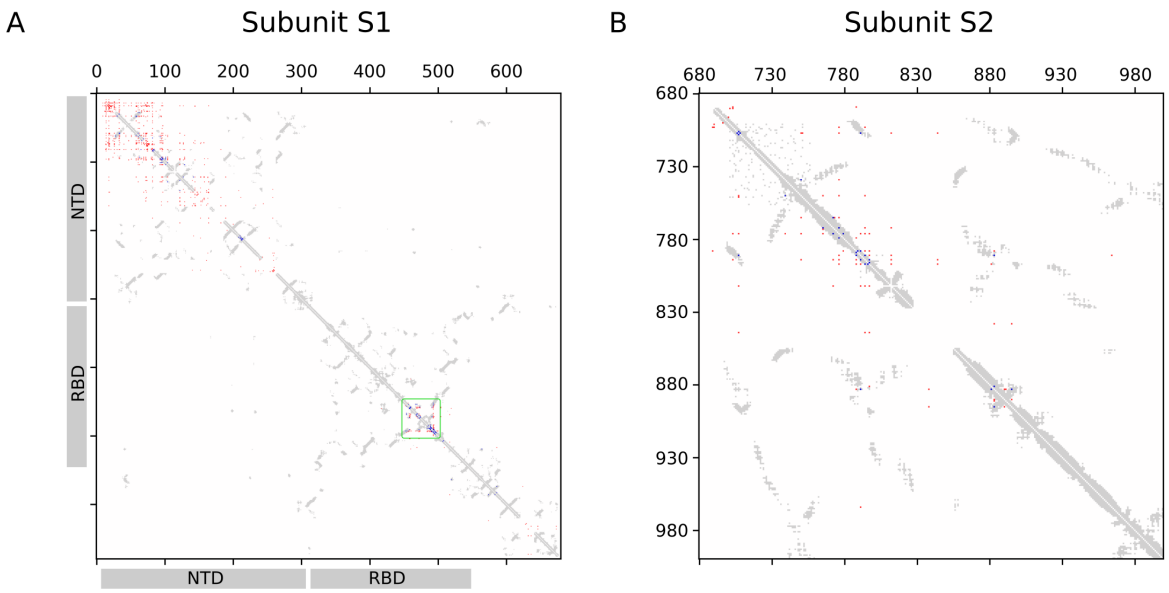

Figure S7. Coevolution analysis across the SARS-CoV-2 spike protein subunits. Contact map computed for subunit S1 (A) and S2 (B). Contact map ground truth was established considering 8A distance cutoff between any atom. Top-500 MI-APC contact predictions are shown in blue (true positives) and in red (false positives). The RBM is highlighted in green.

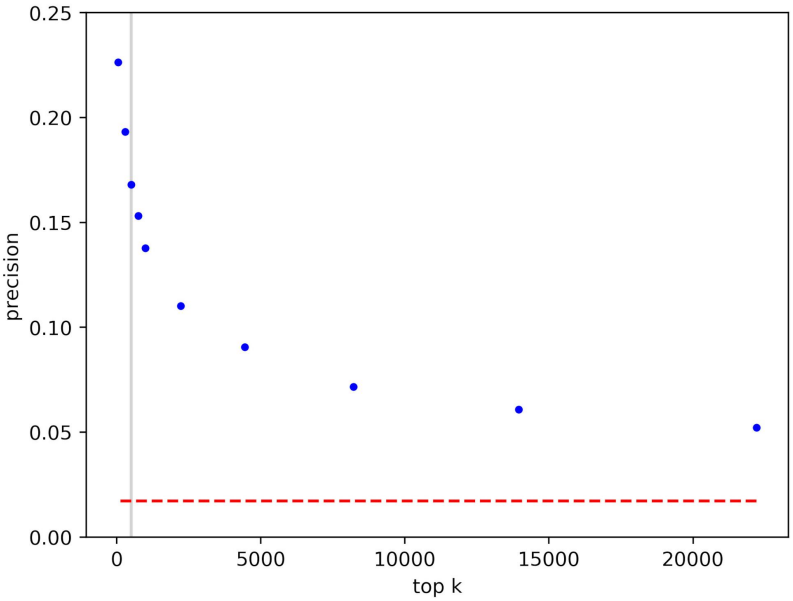

Figure S8. Probability of finding a true contact among the top-k intra-protein MI-APC predictions for the spike protein. Contact map ground truth was established considering 8A distance cutoff between any atom. Gray line: top-500, precision = 16.8%. Red dashed line: null model precision = 1.7%.

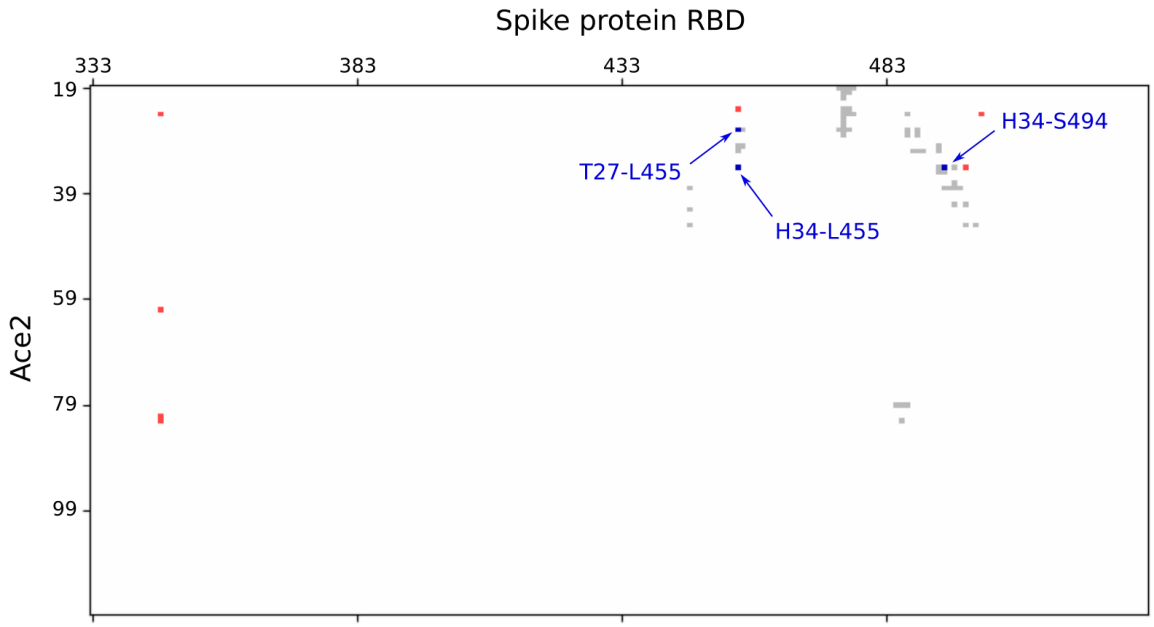

Figure S9. Coevolution analysis between SARS-CoV-2 spike protein RBD and hACE2. Contact map ground truth was established considering 8A distance cutoff between any atom. The top-10 MI-APC contact predictions are shown in blue (true contacts) and in red (false positives). The contact map was obtained from PDB structure 6LZG.

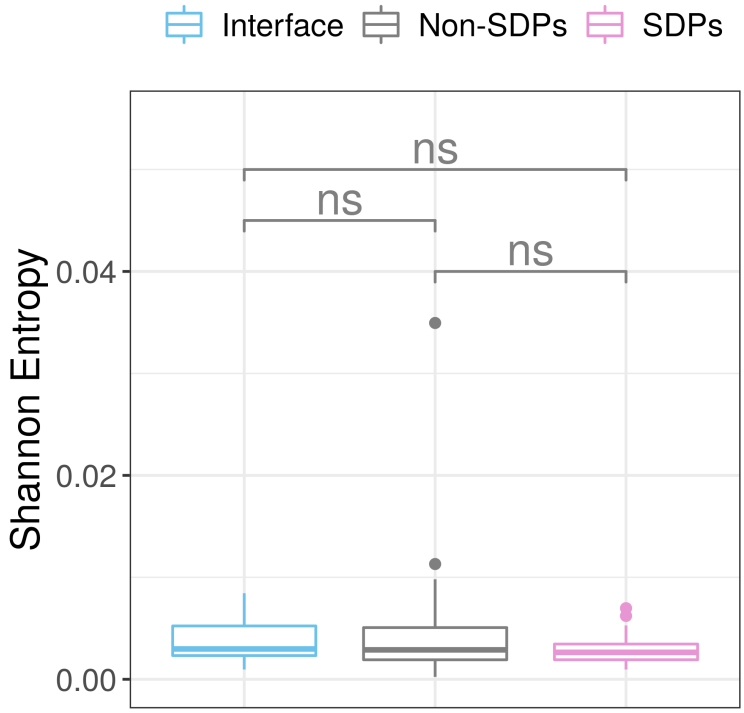

Figure S10. Shannon entropy analysis across the RBD of the circulating SARS-CoV-2 viruses. Shannon entropy distributions are reported for SARS-CoV-2 residues at the interface with the hACE2 (blue), non-SDPs (grey) and SDPs (pink). Significant differences were computed using a Wilcoxon unpaired two-sample test; ns, not significant.

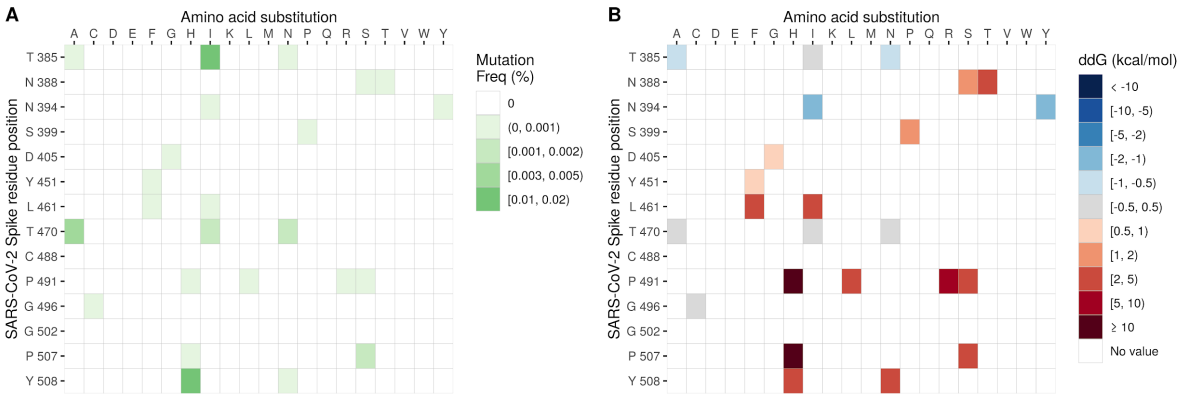

Figure S11. Frequency of mutations at SDPs among the circulating SARS-CoV-2 and predicted mutational impact. (A) Frequency of amino acid variations at SDP positions within the RBD. (B) Predicted effect of mutations in terms of ΔΔG values computed by Foldx (PDB code: 6LZG).

Table S1. Subunit-level enrichment analysis.

		p-value hypergeometric test				
Subunit	**SDP**	**SARS-CoV-2**	**SARS-CoV**	**OC43**	**HKU1**	**MERS**
S1	β-CoV genus	0.999574	0.999418	0.999871	0.999742	0.999698
	β-CoV subgenera	**0.005667**	**0.004632**	0.561269	0.284334	**0.031552**
S2	β-CoV genus	**0.001289**	**0.001716**	**0.000429**	**0.000819**	**0.000947**
	β-CoV subgenera	0.998436	0.998753	0.665795	0.891131	0.987258

Numbers in bold indicate p-value < 0.05.

Table S2. Domain-level enrichment analysis.

		p-value hypergeometric test				
Domain	**SDP**	**SARS-CoV-2**	**SARS-CoV**	**OC43**	**HKU1**	**MERS**
SS	β-CoV genus	-	-	-	-	-
	β-CoV subgenera	**0.031339**	**0.04214**	-	-	-
NTD	β-CoV genus	0.998889	0.998348	0.997644	0.995925	0.999234
	β-CoV subgenera	0.972201	0.965056	**0.090286**	**0.03567**	0.406908
ID1	β-CoV genus	-	-	-	-	-
	β-CoV subgenera	-	-	0.109207	-	0.276512
C-Term/RBD	β-CoV genus	0.988181	0.988995	0.998751	0.999818	0.987616
	β-CoV subgenera	**0.000069**	**0.00077**	0.95516	0.713181	**0.053881**
ID2	β-CoV genus	0.204935	0.195065	0.118511	0.047175	0.137046
	β-CoV subgenera	0.630695	0.620801	-	-	0.453568
ID3	β-CoV genus	0.649939	0.661564	0.49089	0.671802	0.650804
	β-CoV subgenera	0.794561	0.801416	-	-	0.683688
FP	β-CoV genus	0.132685	0.135816	0.118811	0.115631	0.116061
	β-CoV subgenera	0.346639	0.35069	-	-	0.377577
ID4	β-CoV genus	**0.043054**	**0.04567**	**0.025754**	**0.075013**	**0.009901**
	β-CoV subgenera	-	-	0.490168	-	0.870504
HR1	β-CoV genus	**0.032391**	**0.034422**	**0.024234**	**0.005911**	0.072107
	β-CoV subgenera	-	-	-	0.427303	0.844344

ID5	β-CoV genus	**0.028881**	**0.031754**	**0.046297**	**0.04176**	**0.023204**
	β-CoV subgenera	-	-	0.481971	0.759839	0.858372
HR2	β-CoV genus	-	-	0.805111	-	-
	β-CoV subgenera	0.098164	0.101424	0.363018	0.320312	-
TM	β-CoV genus	0.564982	0.570284	-	0.534476	-
	β-CoV subgenera	-	-	-	-	-
CP	β-CoV genus	-	-	-	-	-
	β-CoV subgenera	-	-	0.249255	-	-

Numbers in bold indicate p-value < 0.05.

Table S3. Sarbecovirus SDPs

Pos SARS-CoV2	AA SARS-CoV2	Pos SARS-CoV	AA SARS-CoV	Pos RaTG13	AA RaTG13
12	S	14	S	12	S
13	S	15	D	13	S
70	V	74	H	70	V
127	V	124	V	127	V
254	S	241	A	254	S
385	T	372	T	385	T
388	N	375	N	388	N
394	N	381	N	394	N
399	S	386	S	399	S
405	D	392	D	405	D
451	Y	438	Y	451	Y
461	L	448	L	461	L
470	T	457	N	470	T
488	C	474	C	488	C
491	P	477	P	491	P
496	G	482	G	496	G
502	G	488	G	502	G
507	P	493	P	507	P
508	Y	494	Y	508	Y
592	F	578	F	592	F
655	H	641	H	655	H
673	S	659	S	673	S
780	E	762	E	776	E
785	V	767	V	781	V
831	A	813	A	827	A
1180	Q	1162	Q	1176	Q
1185	R	1167	R	1181	R
1189	V	1171	V	1185	V

Table S4. β-CoV genus SDPs

Pos SARS-CoV2	AA SARS-CoV2	Pos SARS-CoV	AA SARS-CoV	Pos RaTG13	AA RaTG13
191	E	184	E	191	E
268	G	255	G	268	G
298	E	285	E	298	E
517	L	503	L	517	L
524	V	510	V	524	V
541	F	527	F	541	F
563	Q	549	Q	563	Q
568	D	554	D	568	D
572	T	558	F	572	T
594	G	580	G	594	G
644	Q	630	Q	644	Q
645	T	631	T	645	T
669	G	655	G	669	G
720	I	702	I	716	I
741	Y	723	Y	737	Y
742	I	724	I	738	I
766	A	748	A	762	A
819	E	801	E	815	E
827	T	809	T	823	T
839	D	821	E	835	D
852	A	834	A	848	A
858	L	840	L	854	L
869	M	851	M	865	M
877	L	859	L	873	L
911	V	893	V	907	V
913	Q	895	Q	909	Q
934	I	916	I	930	I
937	S	919	S	933	S

958	A	940	A	954	A
965	Q	947	Q	961	Q
968	S	950	S	964	S
989	A	971	A	985	A
1007	Y	989	Y	1003	Y
1011	Q	993	Q	1007	Q
1030	S	1012	S	1026	S
1041	D	1023	D	1037	D
1049	L	1031	L	1045	L
1065	V	1047	V	1061	V
1079	P	1061	P	1075	P
1080	A	1062	A	1076	A
1081	I	1063	I	1077	I
1087	A	1069	A	1083	A
1220	F	1202	F	1216	F