Molecular analyses of over hundred sixty clinical isolates of SARS-CoV-2: Insights on likely origin, evolution and spread, and possible intervention

Supplementary Information

Figures

Figure S1

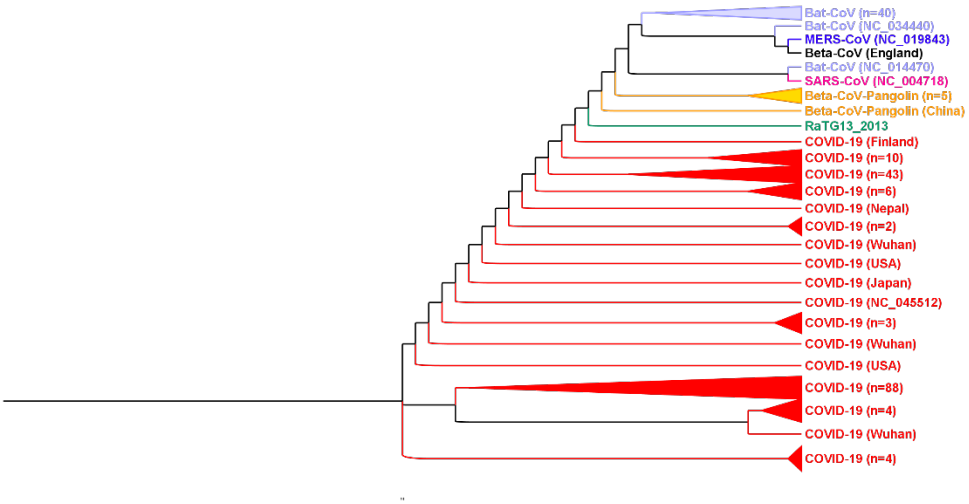

Figure S1. Whole-genome phylogenetic tree of SARS-CoV-2. Phylogenetic relationship based on whole-genome comparisons of 167 clinical isolates with bat coronavirus RaTG13, SARS-CoV and MERS-CoV.

Figure S2

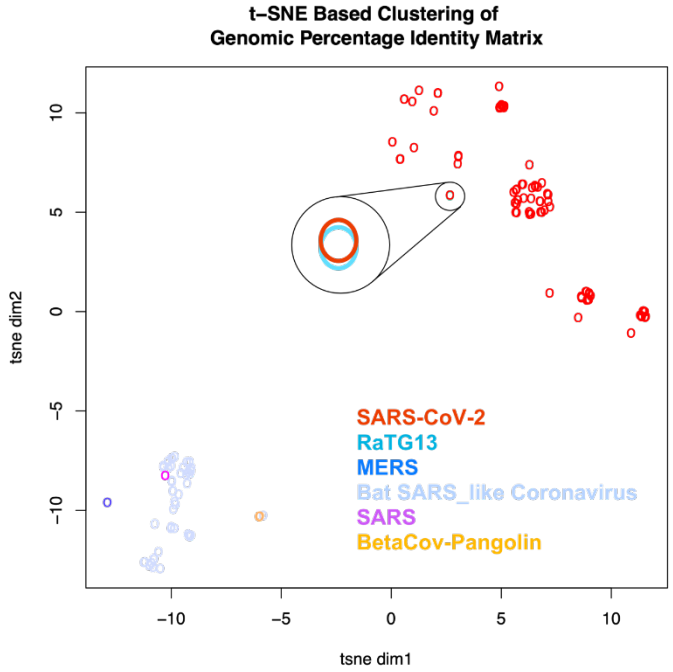

Figure S2. Multiple Sequence Alignment (MSA) of SARS-CoV-2 (n=106) with bat coronavirus RaTG13, SARS-CoV and MERS-CoV was performed using MAFFT on the genomes and Percentage Identity Matrix (PIM) was computed. Machine learning algorithm t-SNE (distributed Stochastic Neighbour Embedding) was used to visualize high-dimensional data and highlight the genomic similarity between genomes. t-SNE converts similarities

between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

Figure S3

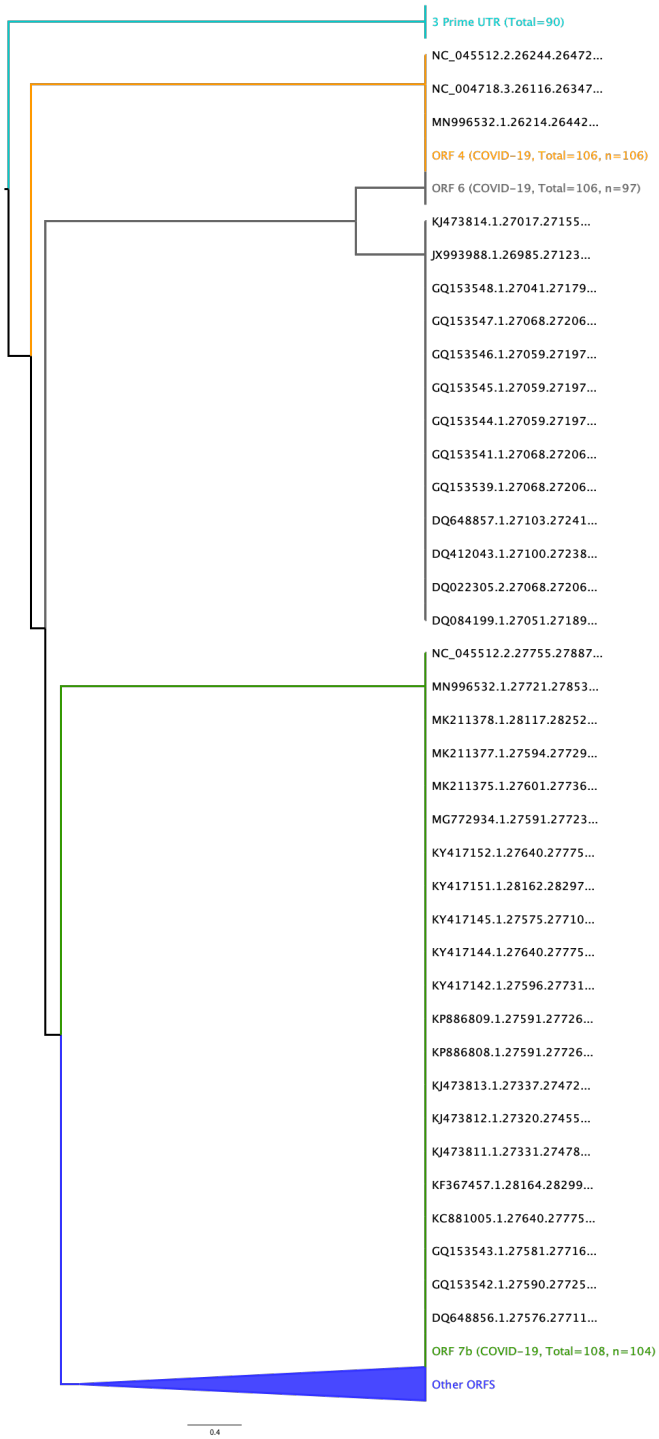

Figure S3. ORF based phylogenetic relationship of COVID-2019 (n=106) clinical isolates with RaTG13, SARS-CoV, MERS and other coronaviruses. The clustering shows sequence based specificity of some of the ORF's, 4 (orange colour), 6 (grey colour) and 7b (green colour), in SARS-CoV-2 clinical isolates. The other ORFs (in blue colour) were found to be

mostly conserved through all the genomes. Total represents the number of coronaviruses falling in one cluster whereas n represents the number of clinical isolates in that cluster.

Figure S4

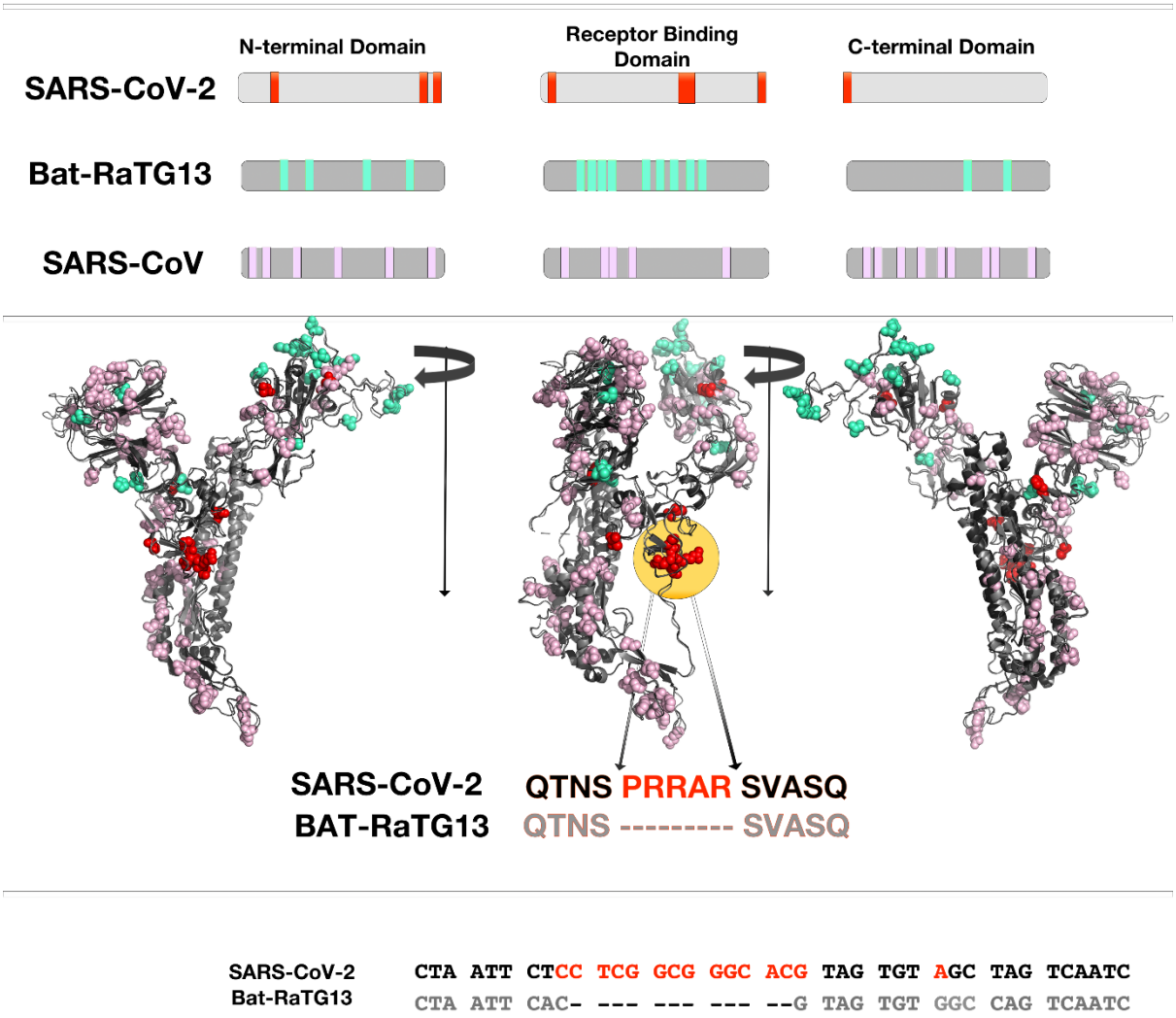

Figure S4. Comparative mutations in Spike glycoprotein among RaTG13, SARS-COV and SARS CoV-2 clinical isolates (n=106). Single point amino acid mutations upon a comparison of RaTG13 and SARS-CoV with SARS-CoV-2 are shown in cyan green and light pink spheres, respectively. Mutations among the clinical isolates are shown in red. Comparative structural analysis between and RaTG13 and SARS-CoV-2 highlights crucial mutations restricted to ACE-2 receptor binding domain. Four amino acid inserts PRRA were observed in all clinical isolates and absent in RaTG13; this is highlighted as green spheres. The corresponding nucleotide alignment is also shown highlighting gaps. The insert in the disordered region appears at a solvent accessible site which could play a crucial role in receptor binding.

Figure S5

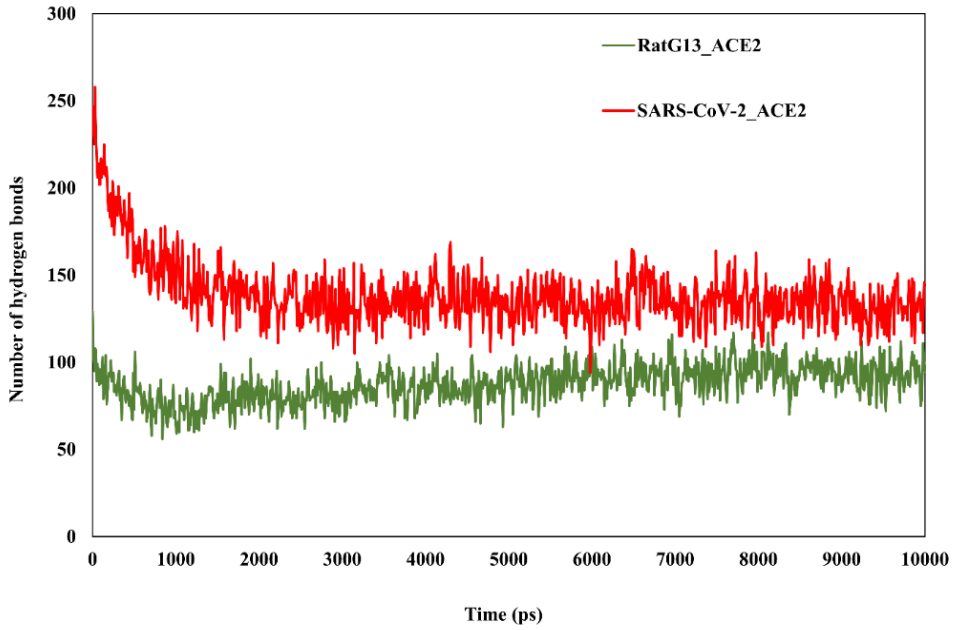

Figure S5. Timeline of ACE-2 human receptor interactions with spike protein of Bat RatG13 and SARS-CoV-2 coronavirus. The higher number of hydrogen bonds in ACE2_SARS-CoV-2 complex shows its preferential binding compared to RatG13 spike protein.

Figure S6

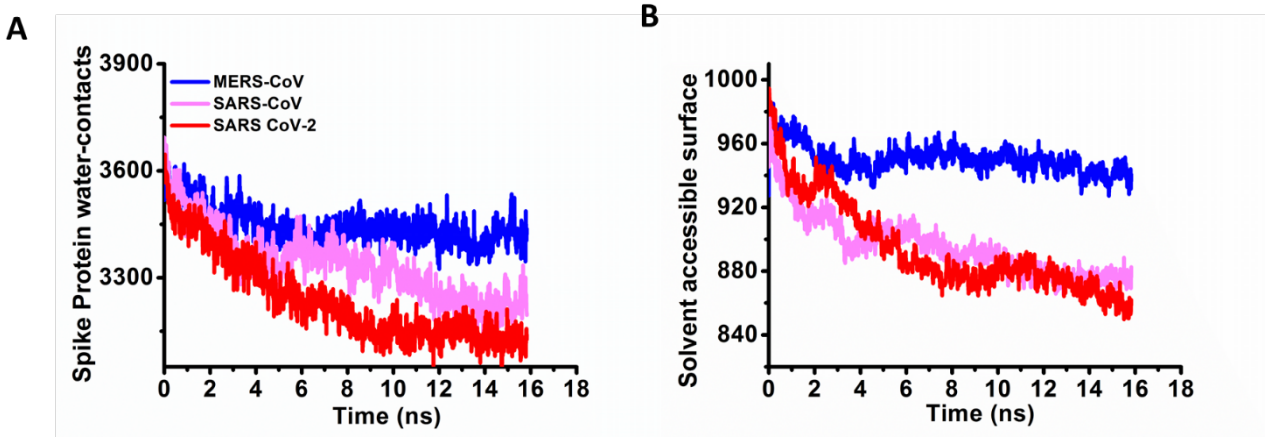

Figure S6. Timeline of MD simulations of MERS, SARS and SARS-CoV-2 spike proteins. (A) Protein-water contacts of spike proteins over simulation trajectory show the gradual lowering of solvent-spike protein surface interactions owing to less number of polar residues 590 than SARS and MERS. (B) Solvent accessibility of SARS-COV-2 also decreased more compared to others. This indicates relatively more hydrophobicity at SARS-CoV-2 surface,

which may promote viral aggregate formation through interactions at the Spike protein surface.

Figure S7

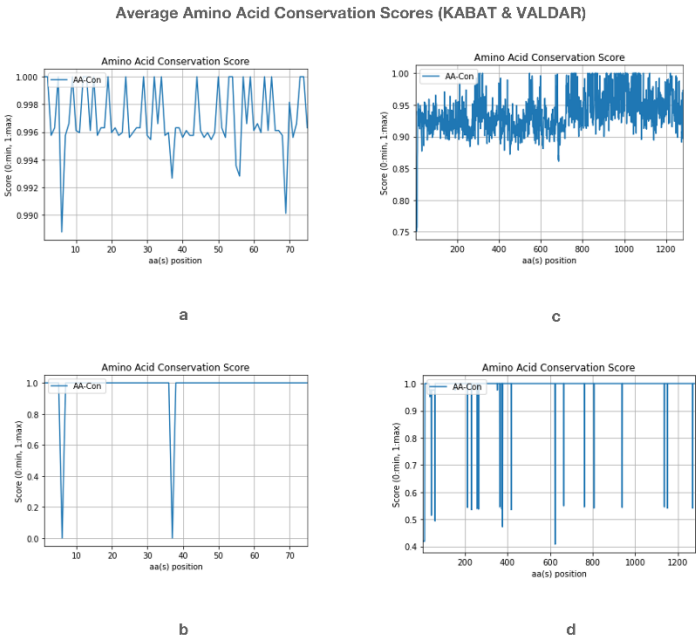

Figure S7. Amino acid conservation analysis of envelope and spike proteins using an average of KABAT and VALDAR score. **(a)** The conservation of amino acid(s) in the envelope proteins of the Bat coronavirus genomes and SARS-CoV-2 clinical isolates. **(b)** The conservation of amino acid(s) in the envelope proteins of SARS-CoV-2 clinical isolate. **(c)** The conservation of amino acids in the spike proteins of the Bat coronavirus genomes and SARS-CoV-2 clinical isolates. **(d)** The conservation of amino acids in the spike proteins of SARS-CoV-2 clinical isolate.

Figure S8

Figure S8. (A) Cartoon representation of motifs in the spike protein of SARS-CoV2, RaTG13 and BetaCoV-Pangolins. The motif was absent in all other bat betacoronaviruses. An interesting feature of this motif was its folding at the solvent exposed disordered region in the N-terminal domain of spike protein. Mutations near this site had been shown to confer extended host range in murine coronaviruses. (B) Multiple sequence alignment shows SYLTPGKLES (highlighted black) and another motif QTQTNS (highlighted green) present only in Bat RaTG13 and SARS-CoV-2 strains and absent in spike proteins in BetaCoV-pangolins (sequence submitted in Feb 2020).

Molecular analyses of over hundred sixty clinical isolates of SARS-CoV-2: Insights on likely origin, evolution and spread, and possible intervention

Tables

Table S1. NCBI nomenclature of ORFs in SARS-CoV-2, RaTG13, SARS-CoV and MERS.

SARS-CoV-2 NCBI Accession No.	SARS-CoV-2 Sequence	SARS-CoV-2 No. of Residues	SARS-CoV-2 Genome position	Name (SARS-CoV-2)	Name (RaTG13)	Name (SARS-CoV)	Name (MERS)
YP_009724389.1	>MESLVPGFNE... VISSDVLVNN	7096aa	266..13468	ORF1ab	ORF1ab	ORF1ab	ORF1ab
YP_009725295.1	>MESLVPGFNE..... AQSFLNGFAV	4405aa	13468..21555	ORF1a		ORF1a	ORF1a
YP_009724390.1	>MFVFLVLLPL..... VLKGVKLHYT	1,273aa	21563..25384	Spike (S)	Spike (S)	Spike (S)	Spike (S)
YP_009724391.1	>MDLFMRIFTI..... EPTTTTSVPL	275aa	25393..26220	ORF3a	NS3	ORF3a	
YP_009724392.1	>MYSFVSEETG..... NSSRVPDLLV	75aa	26245..26472	ORF4	Envelope (E)	Envelop e (E)	Envelope (E)
YP_009724393.1	>MADSNGTITV.... SSDNIALLVQ	222aa	26523..27191	Membrane (M)	Membrane (M)	Matrix (M)	Membrane (M)
YP_009724394.1	>MFHLVDFQVT... LDEEQPMEID	61aa	27202..27387	ORF6	NS6	ORF6	
YP_009724395.1	>MKIILFLALI..... TLCFTLKRKTE	121aa	27394..27759	ORF7a	NS7a	ORF7a	
YP_009725296.1	>MIELSLIDFY..... LQDHNETCHA	43aa	27756..27887	ORF7b	NS7b	SARS7b	
YP_009724396.1	>MKFLVFLGII.....	121aa	27894..28259	ORF8	NS8	ORF8b	

	HDVRVVLDFI						
YP_009724397.2	>MSDNGPQNQR.... SMSSADSTQA	419aa	28274..29533	Nucleocapsid (N)	Nucleocapsid (N)	Nucleocapsid (N)	Nucleocapsid (N)
YP_009725255.1	>MGYINVFAFP..... QVDVVNFNLT	38aa	29558..29674		NA		

Table S2. Pre-MD protein-protein interactions between spike protein of SARS-CoV-2 (chain B) and RaTG13 (chain B) and host receptor ACE2 (chain A). Note the overall increase in hydrogen bonding and hydrophobic interactions between SARS-CoV-2:ACE2 as compared to RaTG13:ACE2.

Spike-RatG13:ACE2	**Spike-SARS-CoV-2:ACE2**
Hydrogen bonds	
Gly425(B)-Lys68(A)	Tyr449(B)-Asp38(A)
Lys426(B)-Glu75(A)	Gln498(B)-Asp38(A)
Lys426(B)-Glu35(A)	Gln498(B)-His34(A)
Tyr504(B)-Lys31(A)	Gln493(B)-Gln76(A)
Tyr430(B)-Asp38(A)	Gln493(B)-Lys31(A)
Tyr499(B)-His34(A)	Ser494(B)-Lys31(A)
Tyr482(B)-Lys353(A)	Glu484(B)-Lys74(A)
	Gly446(B)-Gln42(A)
	Tyr489(B)-Gln81(A)
	Lys417(B)-Glu23(A)
Hydrophobic interactions	
Leu495(B), Asn496(B), Ile427(B), Tyr462(B), Leu464(B), Tyr502(B), Phe465(B), His514(B), Ala484(B)	Gly447(B), Pro499(B), Thr500(B), Asn501(B), Val483(B), Gly496(B), Tyr505(B), Arg403(B), Tyr453(B), Tyr495(B), Tyr473(B), Leu492(B), Phe486(B), Pe456(B), Phe490(B)
Ala386(A), Ala387(A), Gly354(A), Leu39(A), Phe72(A), Asp30(A), Gln42(A), Glu37(A), Leu79(A), Phe28(A)	Leu39(A), Lys353(A), Asp30(A), Glu110(A), Glu35(A), Phe72(A), Met82(A), Leu79(A), Asn103(A), Gln24(A), Val107(A), Ser19((A), Thr78(A)

Table S3. Post-MD protein-protein interactions between spike protein of SARS-CoV-2 (chain B) and RaTG13 (chain B) and host receptor ACE2 (chain A). Note the overall increase in hydrogen bonding and hydrophobic interactions between SARS-CoV-2:ACE2 as compared to RaTG13:ACE2.

Spike-RatG13:ACE2	**Spike-SARS-CoV-2:ACE2**
Hydrogen bonds	
Tyr462(B)-Glu35(A)	Gln498(B)-Glu35(A) (3)
Arg466(B)-Lys353(A)	Ser494(B)-Lys31(A)
Tyr498(B)-Tyr498(B)	Tyr505(B)-Thr27(A)
	Gly502(B)-His34(A)
	Thr500(B)-His34(A)
	Lys417(B)-Ser19(A)
	Tyr421(B)-Gln24(A)
	Tyr489(B)-Gln81(A)
Hydrophobic interactions	
Lys426(B), Thr424(B), Gln423(B), Tyr430(B), Ile427(B), Leu464(B), Ser486(B), Tyr502(B), Phe465(B), Ala484(B), Tyr482(B)	Gly496(B), Phe497(B), Ile418(B), Asn501(B), Gln493(B), Leu455(B), Phe456(B), Tyr473(B), Phe486(B), Glu484(B), Gly485(B)
Leu39(A), Lys68(A), Asn64(A), Asp38(A), Gly326(A), His34(A), Gly354(A)	Asp38(A), Leu39(A), Phe28(A), Asp30(A), Glu37(A), Leu79(A), Glu23(A), Tyr83(A), Met82(A), Thr78(A)

Table S4. Top-scoring drugs that can serve as a probable candidates against spike protein of SARS-CoV-2. We targeted RBD of spike protein of SARS-CoV-2 and performed virtual screening using FDA approved drug library to identify probable inhibitors of COVID-2019.

DrugBank ID	**Drug name**	**Docking score (kcal/mol)**	**Description**
DB00157	NADH	-11.31	NADH plays essential metabolic roles and has been used to combat chronic fatigue syndrome. It is also being explored to be used against dementia and improving mental health.
DB01698	Rutin	-9.94	An existing US FDA approved drug used for strengthening weakened capillaries. Also, it has powerful antioxidant with potential biological effect in reducing post-thrombotic syndrome, veins insufficiency or endothelial dysfunction.