Supplemental Material For Publication

Origin, phylogeny, variability and epitope conservation of SARS-CoV-2 worldwide

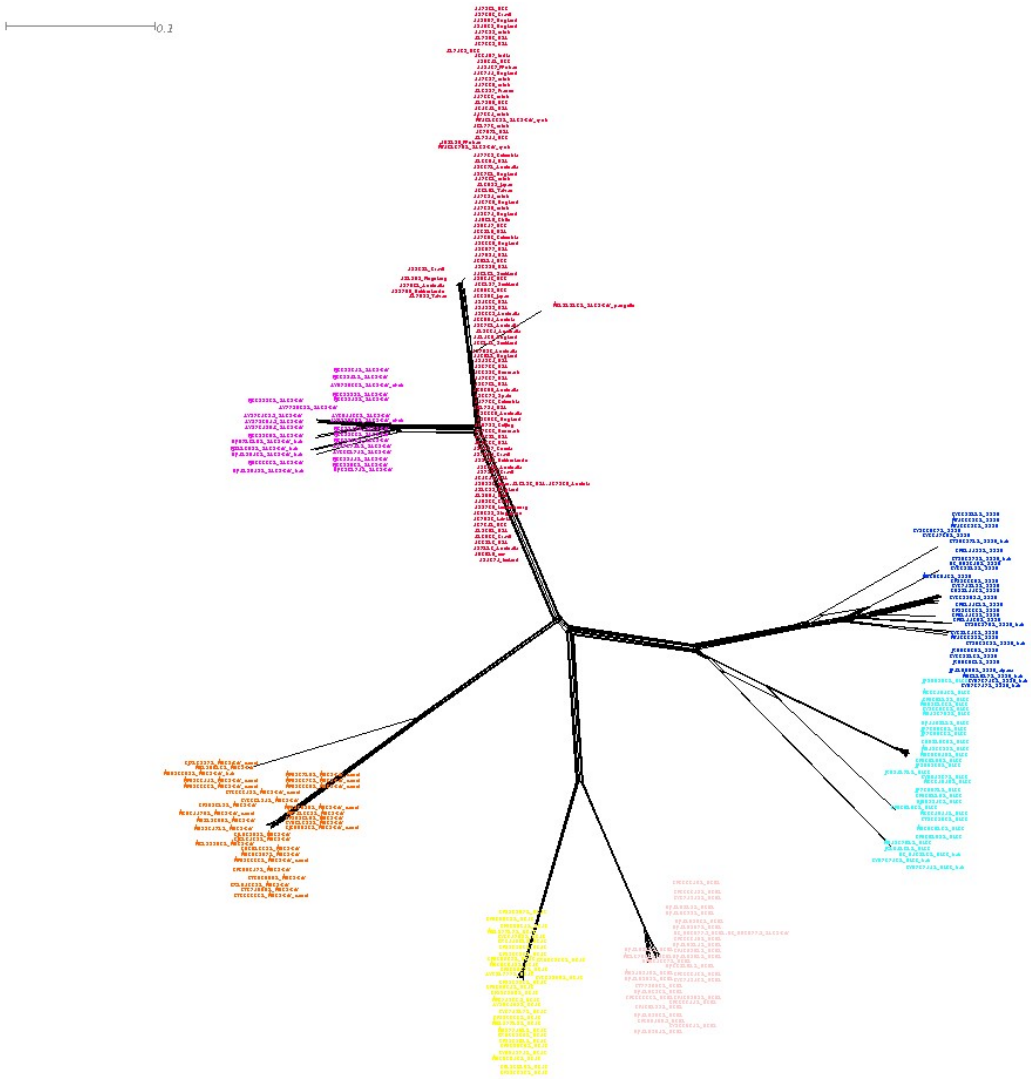


Figure S1. SplitsTree network of coronavirus infecting humans using SplitsTree 4.1. Taxa in Red – SARS-CoV-2; Blue – 229E; Cyan – NL63; Green-brown – HKU1; Yellow – OC43; Orange – MERS; Magenta – SARS-CoV.

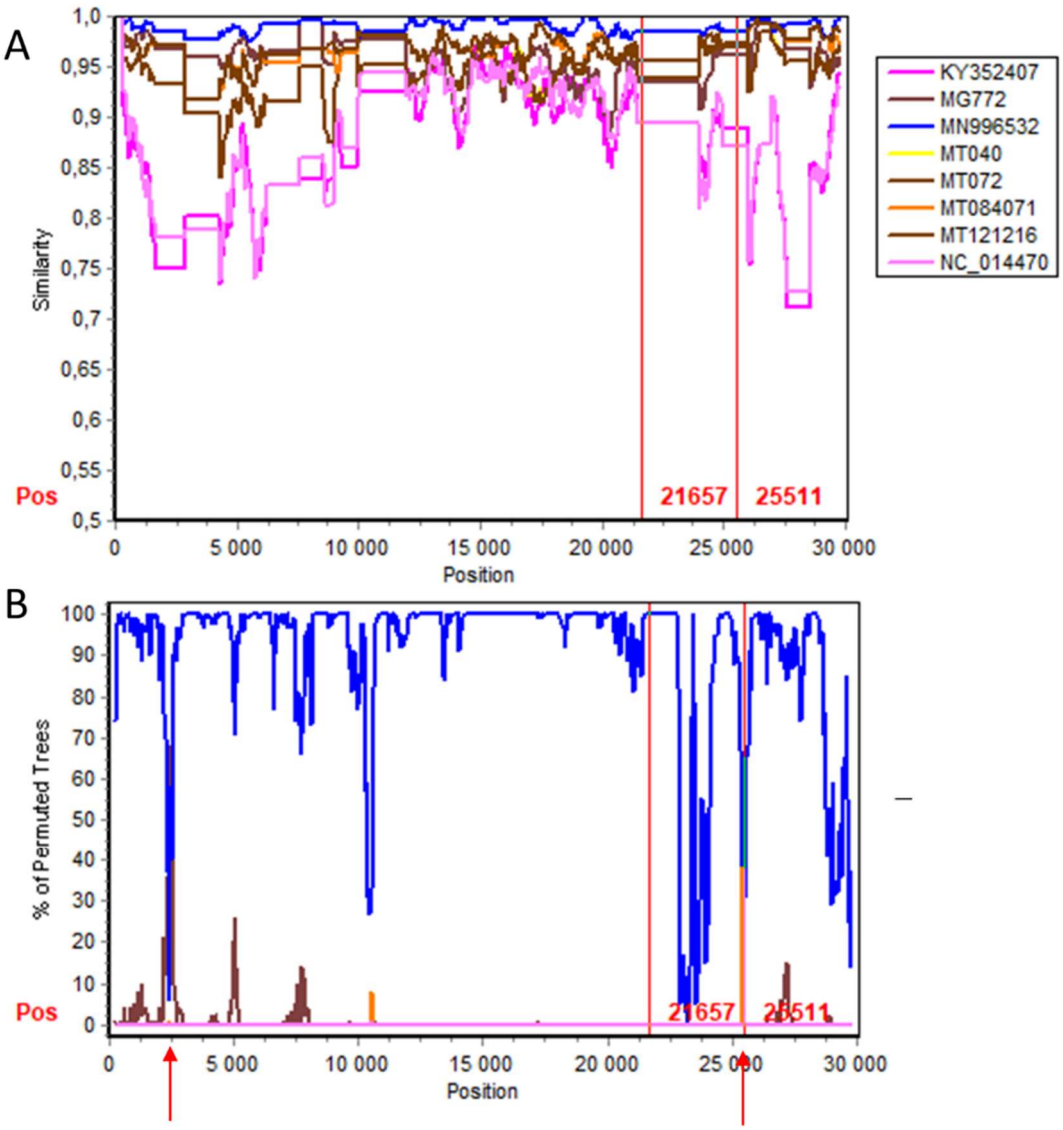


Figure S2. A) Similarity plot along the reference genome of SARS-CoV-2 versus the sequences of the non-human-SARS-CoV-2 emerging cluster, using SimPlot version3.5.1. Each curve is a comparison between the genome being analyzed and the reference genome. B) Bootscan plot querying the SARS-CoV-2 reference genome with the genomes of the non-human-SARS-CoV-2 emerging cluster. The bootscan graph indicates two recombinant regions (red arrows) with breakpoint positions at approximate sequence positions around ~2350 to ~2400 (region of orf 1ab) and ~25400 to ~25500 (region of spike gene) between the bat coronavirus genome RaTG13 and another group of bat genomes named bat-SL-CoVZXC21 and bat-SL-CoVZC45; and between bat coronavirus genome RaTG13 and the pangolin coronavirus genome MP789, respectably. The parameters of SimPlot similarity plot and bootscan analysis were set as default except for a window size of 500 and step size of 50. The breakpoint positions refer to the coordinates of the spike gene.

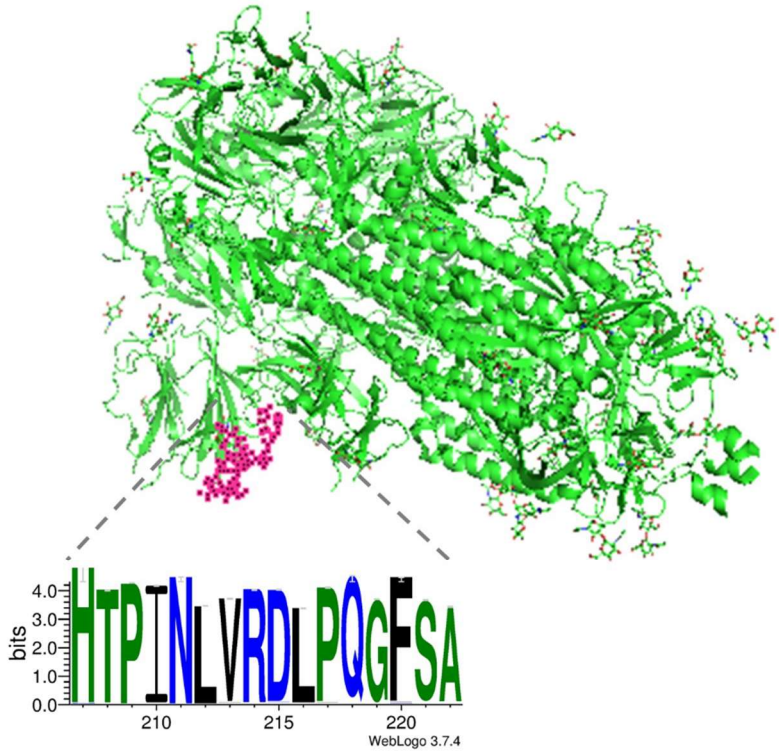


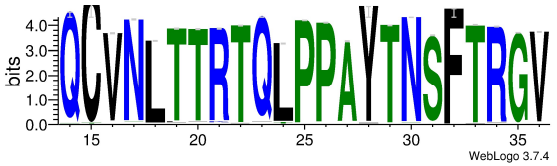
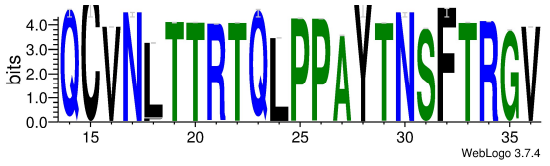
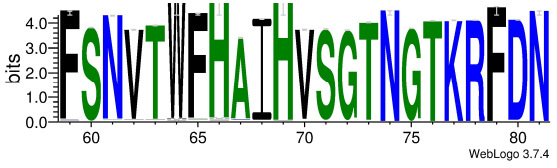
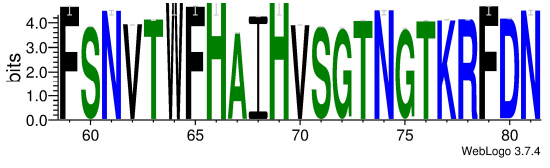
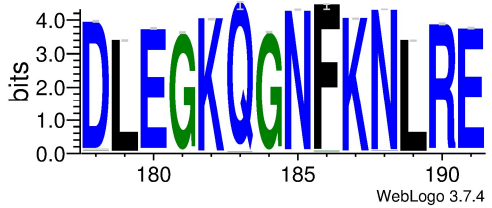
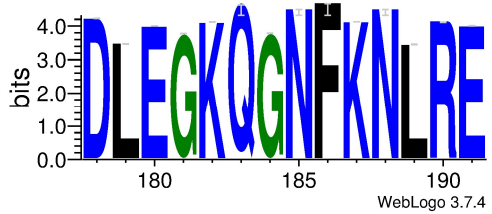
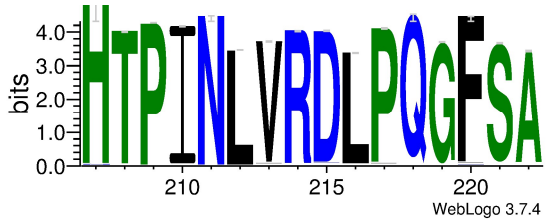
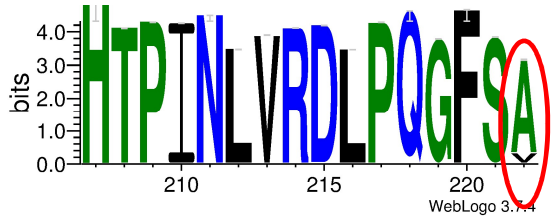
Figure S3. SARS-CoV-2 S glycoprotein structure (pdb: 6vxx). Figure caption using PyMol version 2.0. Dashed line highlight conservation of a putative epitope region (amino acids residues positions ranging from 207 to 222) and correspondent sequence logo from 19471 SARS-CoV-2 genome alignment, evidencing that the putative epitope is an exposed region of the protein.

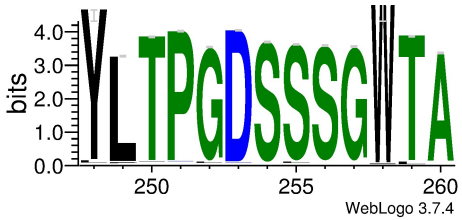
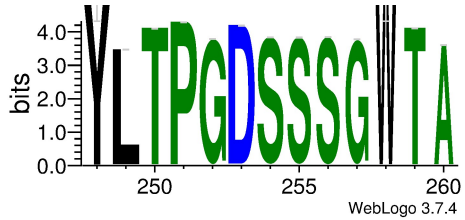
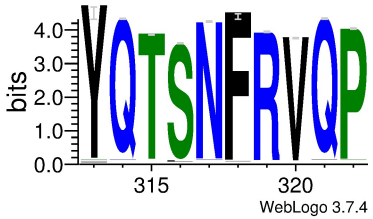
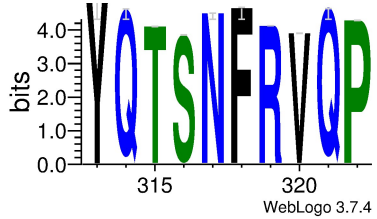
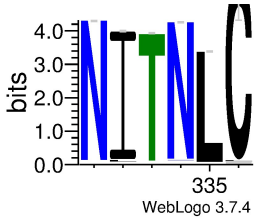
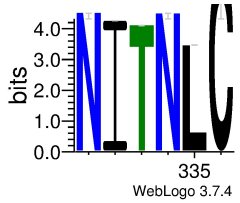
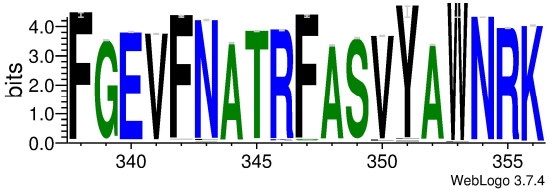
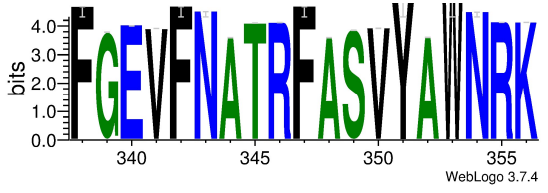
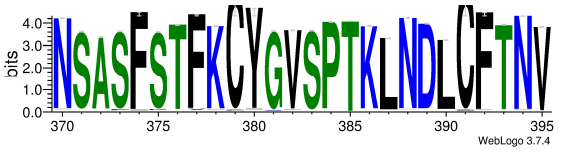
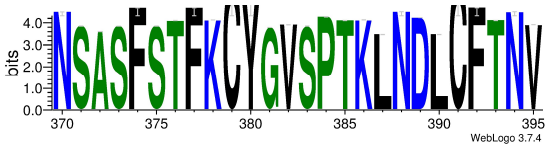
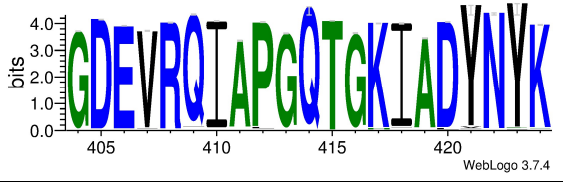
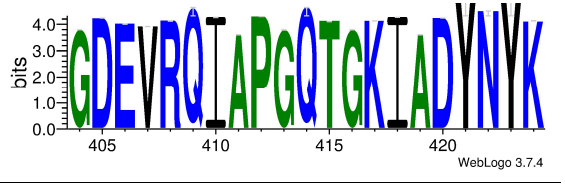
Table S1. Variant density across SARS-CoV-2 Open reading frames (ORFs).

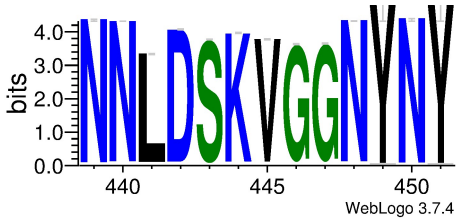
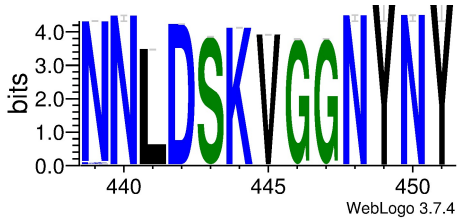
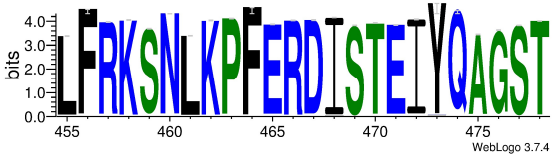
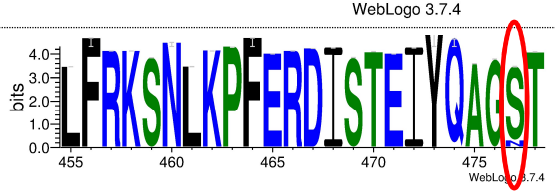
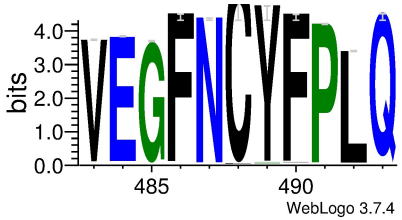
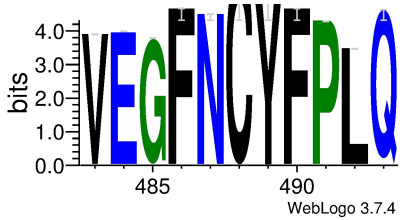
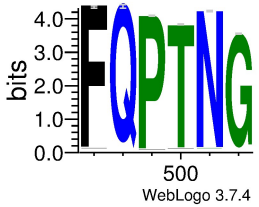
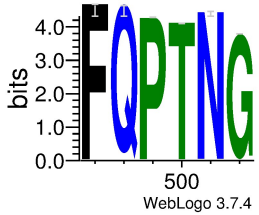
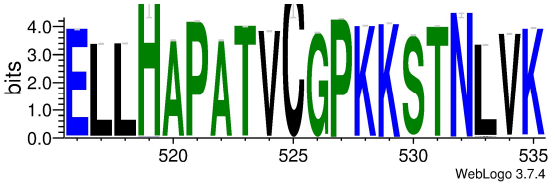
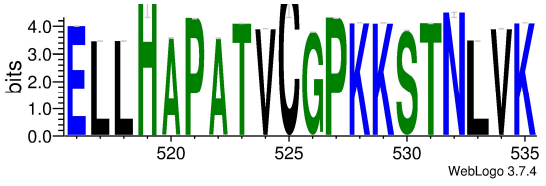
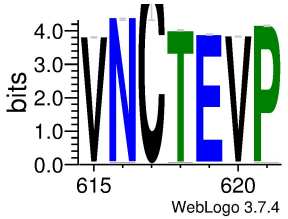
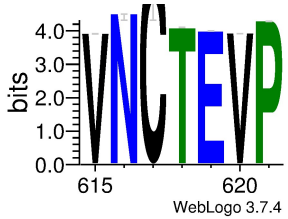
SARS-CoV-2 open Reading frames	Variant density
orf1ab	0,30
mature peptide -leader protein (both pp1a and pp1ab)	**0,40**
mature peptide -nsp2 (both pp1a and pp1ab)	**0,37**
mature peptide - nsp3 (both pp1a and pp1ab)	0,31
mature peptide - ns4 (both pp1a and pp1ab)	0,30
mature peptide - 3C-like proteinase (both pp1a and pp1ab)	0,26
mature peptide - nsp6 (both pp1a and pp1ab)	0,29
mature peptide - nsp7 (both pp1a and pp1ab)	0,33
mature peptide - nsp8 (both pp1a and pp1ab)	0,29
mature peptide - nsp9 (both pp1a and pp1ab)	0,28
mature peptide - nsp10 (both pp1a and pp1ab)	0,25
mature peptide - RNA-dependent RNA polymerase (produced by pp1ab only)	0,27
mature peptide - helicase (produced by pp1ab only)	0,25
mature peptide - 3'-to-5' exonuclease (produced by pp1ab only)	0,29
mature peptide - endoRNAse (produced by pp1ab only)	0,29
mature peptide -2'-O-ribose methyltransferase (produced by pp1ab only)	0,26
orf1a	0,31
mature peptide - nsp11 (produced by pp1a only)	0,26
spike glycoprotein	**0,31**
SARS-CoV-like_Spike_S1_NTD	**0,39**
Spike_NTD	**0,37**
SARS-CoV-2_Spike_S1_RBD	0,26
SARS-CoV-like_Spike_SD1-2_S1-S2_S2	0,29
Corona_S2	0,28
CoV_S2_C	0,26
0rf3a	**0,47**
gene E	0,29
gene M	0,29
ORF6	**0,41**
ORF7a	**0,49**
ORF7b	**0,40**
ORF8	**0,45**
gene N	**0,42**
ORF10	**0,38**

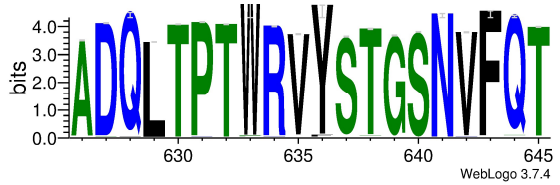
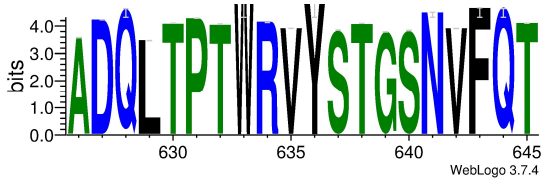
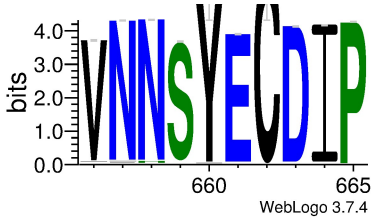
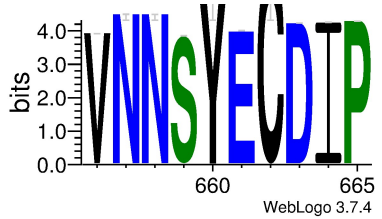
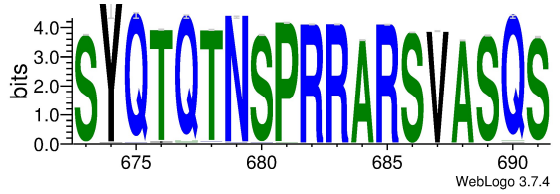
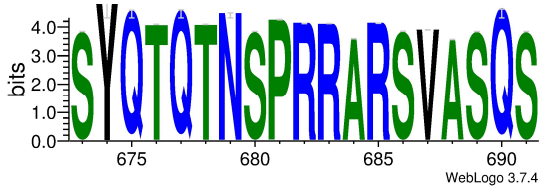
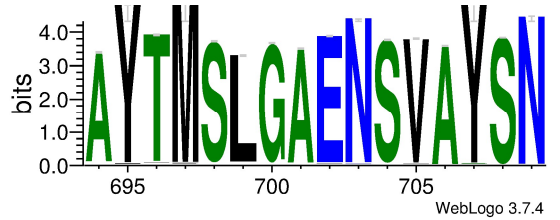
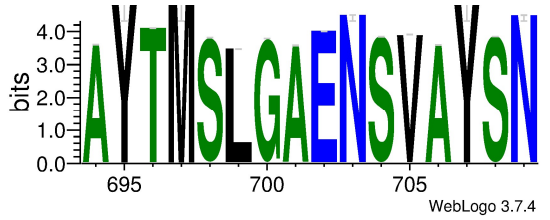
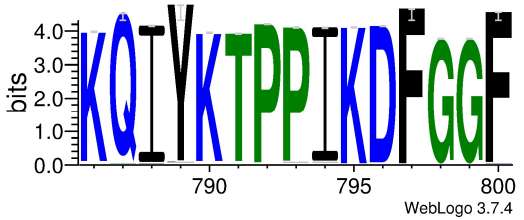
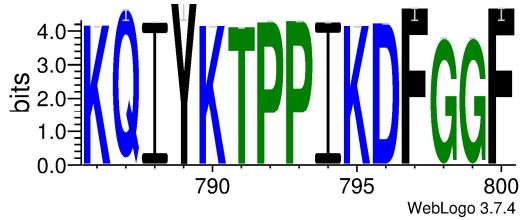
Note: orf1ab and orf1a are translated into either polyprotein ab or pp1a. Some mature peptides are produced by both orf1ab and orfa, while others are produced only by one of the polyproteins as depicted in the table. The variant density of the conserved domains of spike glycoprotein is also presented.

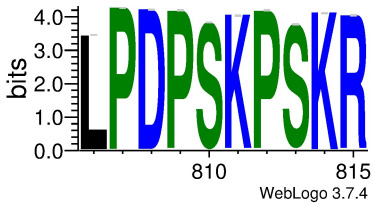
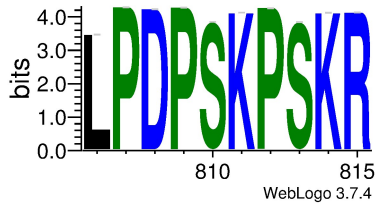
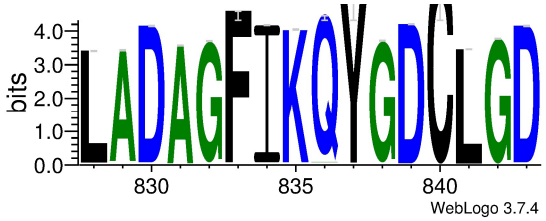
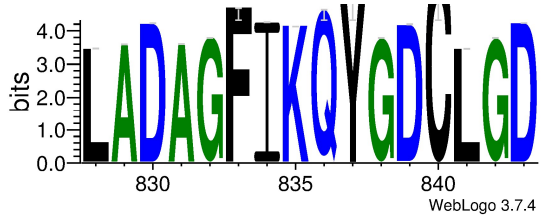
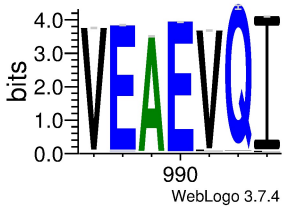
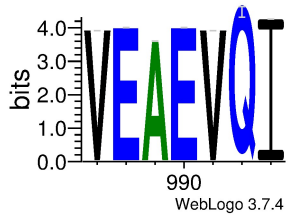
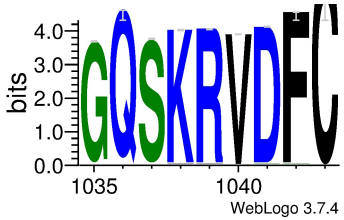
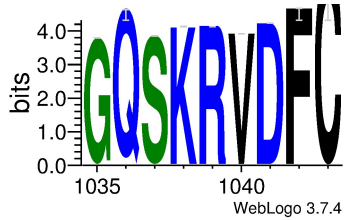
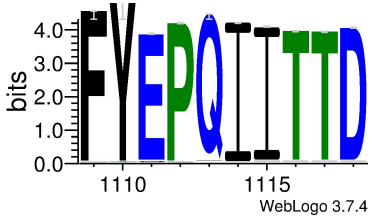
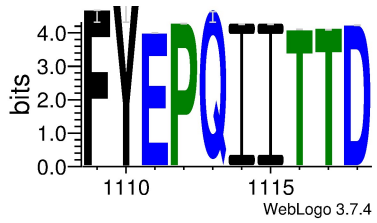
Table S2. Predicted epitopes from spike (S) glycoprotein of SARS-CoV-2.

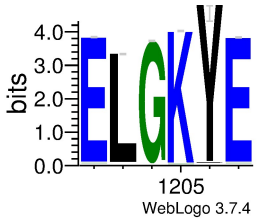
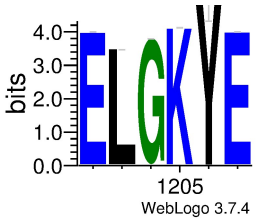
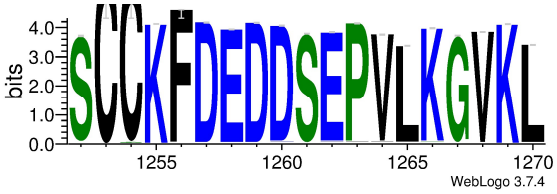
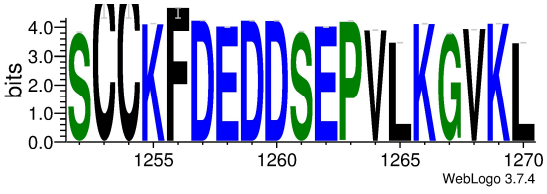
Sequence	Size	% of exposed residues	Average epitope probability	Exposed epitope 3D structure [a]	Start residue	End residue	Epitope conservancy in all SARS-CoV-2 (%)	Sequence logo of 19471 S glycoprotein aligned sequences	Sequence log of 31323 unique S glycoprotein aligned sequences from a set of 199984 sequences
QCVNLTTRTQLPPAYTNSFTRGV	23	0.522	0.581		14	36	89.8		
FSNVTWFHAIHVSGTNGTKRFDN	23	0.478	0.559		59	81	96.7		
LGVYYHKNNKSWMESEFRVYSSA	23	0.565	0.570		141	163	96.9		
DLEGKQGNFKNLRE	14	0.571	0.546		178	191	98.0		
HTPINLVRDLPQGFSA	16	0.750	0.574	Yes	207	222	85.7		

YLTPGDSSSGWTA	13	0.692	0.595		248	260	87.9		
YQTSNFRVQP	10	0.500	0.536	Yes	313	322	88.5		
NITNLC	6	0.667	0.517	Yes	331	336	99.4		
FGEVFNATRFASVYAWNRK	19	0.632	0.552	Yes	338	356	98.7		
NSASFSTFKCYGVSPTKLNDLCFTNV	26	0.500	0.563	Yes	370	395	98.3		
GDEVRQIAPGQTGKIADYNYK	21	0.286	0.518	Yes	404	424	98.5		

NNLDSKVGGNYNY	13	0.615	0.567		439	451	94.1		
LFRKSNLKPFERDISTEIYQAGST	24	0.458	0.561		455	478	89.8		
VEGFNCYFPLQ	11	0.636	0.533		483	493	96.1		
FQPTNG	6	0.667	0.506		497	502	96.1		
ELLHAPATVCGPKKSTNLVK	20	0.700	0.564	Yes	516	535	95.1		
VNCTEVP	7	0.429	0.522		615	621	99.8		

ADQLTPTWRVYSTGSNVFQT	20	0.450	0.576		626	645	99.1		
VNNSYECDIP	10	0.700	0.540	Yes	656	665	99.2		
SYQTQTNSPRRARSVASQS	19	0.789	0.618		673	691	97.9		
AYTMSLGAENSVAYSN	16	0.813	0.608	Yes	694	709	98.8		
KQIYKTPPIKDFGGF	15	0.600	0.560	Yes	786	800	99.1		

LPDPSKPSKR	10	0.700	0.563	Yes	806	815	99.1		
LADAGFIKQYGDCLGD	16	0.563	0.529		828	843	98.8		
VEAEVQI	7	0.429	0.525	Yes	987	993	99.6		
GQSKRVDFC	9	0.222	0.530	Yes	1035	1043	99.6		
FYEPQIITTD	10	0.600	0.541	Yes	1109	1118	99.5		

ELGKYE	6	0.500	0.516		1202	1207	99,5		
SCCKFDEDDSEPVLKGVKL	19	0.632	0.537		1252	1270	98,6		

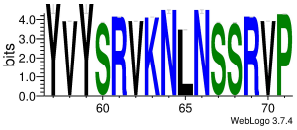
[a] - exposed epitope according to 3D structure visualized with PyMOL, pdb: 6vxx (Schrodinger LLC 2015). Red circle – evidence of two enriched amino acids (A and V; S and N).

Table S3. Conservation of the predicted epitopes from spike (S) glycoprotein of SARS-CoV-2 worldwide and by continent.

Continent Number of analyzed sequences	World * 199984		Africa 1667		Asia 13015		Europe 122533		North America 45222		South America 2240		Oceania 15307	
Putative epitope	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %	Analyzed sequences	Conserved %
QCVNLTTRTQLPPAYTNSFTRGV	179508	89.8	1463	87.8	12487	95.9	105243	85.9	43513	96.2	2098	93.7	14704	96.1
FSNVTWFHAIHVSGTNGTKRFDN	193426	96.7	1634	98.0	12735	97.8	117402	95.8	44429	98.2	2201	98.3	15025	98.2
LGVYYHKNNKSWMESEFRVYSSA	193732	96.9	1579	94.7	12584	96.7	119786	97.8	43897	97.1	2139	95.5	13747	89.8
DLEGKQGNFKNLRE	195929	98.0	1597	95.8	12665	97.3	120784	98.6	44483	98.4	2129	95.0	14271	93.2
HTPINLVRDLPQGFSA	171307	**85.7**	1590	95.4	12670	97.3	96504	**78.8**	44264	97.9	2157	96.3	14122	92.3
YLTPGDSSSGWTA	175793	87.9	1351	81.0	12371	95.1	109355	89.2	38890	**86.0**	1739	77.6	12087	79.0
YQTSNFRVQP	176961	88.5	1338	**80.3**	12268	**94.3**	108940	88.9	39869	88.2	1699	**75.8**	12847	83.9
NITNLC	198862	99.4	1638	98.3	12925	99.3	121904	99.5	45018	99.5	2178	97.2	15199	99.3
FGEVFNATRFASVYAWNRK	197428	98.7	1630	97.8	12841	98.7	120886	98.7	44857	99.2	2170	96.9	15044	98.3
NSASFSTFKCYGVSPTKLNDLCFTNV	196527	98.3	1534	92.0	12785	98.2	120200	98.1	44830	99.1	2187	97.6	14991	97.9
GDEVRQIAPGQTGKIADYNYK	197048	98.5	1617	97.0	12870	98.9	120600	98.4	44780	99.0	2203	98.3	14978	97.9
NNLDSKVGGNYNY	188256	94.1	1535	92.1	12692	97.5	115710	94.4	43670	96.6	2126	94.9	12523	81.8
LFRKSNLKPFERDISTEIYQAGST	179646	89.8	1529	91.7	12627	97.0	115597	94.3	43611	96.4	2113	94.3	4169	**27.2**
VEGFNCYFPLQ	192178	96.1	1547	92.8	12688	97.5	119173	97.3	43822	96.9	2111	94.2	12837	83.9
FQPTNG	192093	96.1	1546	92.7	12676	97.4	119102	97.2	43895	97.1	2115	94.4	12759	83.4
ELLHAPATVCGPKKSTNLVK	190093	95.1	1528	91.7	12537	96.3	118456	96.7	43189	95.5	2131	95.1	12252	80.0
VNCTEVP	199555	**99.8**	1662	**99.7**	12974	**99.7**	122310	**99.8**	45151	**99.8**	2231	**99.6**	15227	**99.5**
ADQLTPTWRVYSTGSNVFQT	198274	99.1	1654	99.2	12955	99.5	121539	99.2	44995	99.5	2224	99.3	14907	97.4
VNNSYECDIP	198310	99.2	1658	99.5	12854	98.8	121426	99.1	44998	99.5	2208	98.6	15166	99.1
SYQTQTNSPRRARSVASQS	195698	97.9	1629	97.7	12609	96.9	119857	97.8	44368	98.1	2188	97.7	15047	98.3
AYTMSLGAENSVAYSN	197658	98.8	1654	99.2	12823	98.5	121207	98.9	44837	99.1	2200	98.2	14937	97.6
KQIYKTPPIKDFGGF	198101	99.1	1609	96.5	12773	98.1	121705	99.3	44815	99.1	2153	96.1	15046	98.3
LPDPSKPSKR	198217	99.1	1621	97.2	12771	98.1	121781	99.4	44841	99.2	2181	97.4	15022	98.1
LADAGFIKQYGDCLGD	197582	98.8	1632	97.9	12686	97.5	121164	98.9	44879	99.2	2188	97.7	15033	98.2
VEAEVQI	199202	99.6	1640	98.4	12911	99.2	122210	99.7	45095	99.7	2176	97.1	15170	99.1
GQSKRVDFC	199129	99.6	1635	98.1	12892	99.1	122177	99.7	45072	99.7	2178	97.2	15175	99.1
FYEPQIITTD	199039	99.5	1660	99.6	12929	99.3	122089	99.6	45071	99.7	2224	99.3	15066	98.4
ELGKYE	199014	99.5	1641	98.4	12920	99.3	122057	99.6	45036	99.6	2174	97.1	15186	99.2
SCCKFDEDDSEPVLKGVKL	197284	98.6	1646	98.7	12871	98.9	120663	98.5	44881	99.2	2208	98.6	15015	98.1

* Incudes S glycoprotein sequences of the SARS-CoV-2 isolated from humans, with more than 1250 amino acids, available at GISAID. The smallest and the highest conservancy percentage is presented in bold for easy interval reading.

Table S4. Predicted epitopes from envelope (E) protein of SARS-CoV-2.

Sequence	Size	% of exposed residues	Average epitope probability	Exposed epitope 3D structure [a]	Start residue	End residue	Epitope conservancy in all SARS-CoV-2 (%)	Sequence logo
YVYSRVKNLNSSRVP	15	66,667	0,561		57	71	99.3	

[a] - exposed epitope according to 3D structure visualized (pdb: 5X29) using PyMOL (Schrodinger LLC 2015).

Table S5. Predicted epitopes from membrane (M) glycoprotein of SARS-CoV-2.

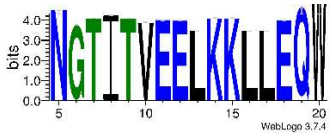
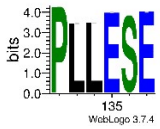
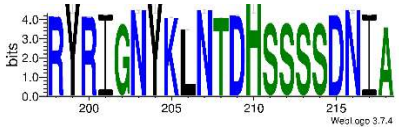
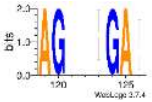
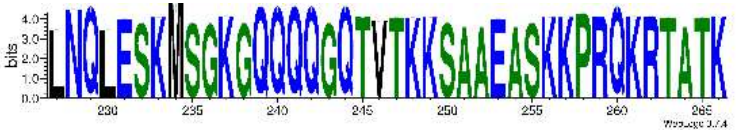
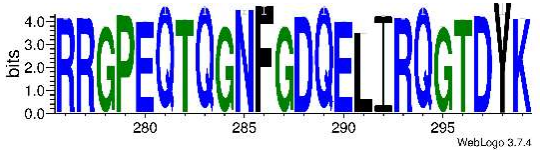
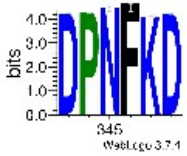
Sequence	Size	% of exposed residues	Average epitope probability	Start residue	End residue	Epitope conservancy in all SARS-CoV-2 (%)	Sequence logo
NGTITVEELKKLLEQW	16	56,250	0,530	5	20	99.5	
PLLESE	6	66,667	0,518	132	137	99.0	
KLGASQRVAGDS	12	83,333	0,548	180	191	99.0	
RYRIGNYKLNTDHSSSSDNIA	21	71,429	0,614	198	218	98.8	

Table S6. Predicted epitopes from nucleocapsid (N) phosphoprotein of SARS-CoV-2.

Sequence	Size	% of exposed residues	Average epitope probability	Exposed epitope 3D structure [a]	Start residue	End residue	Epitope conservancy in all SARS-CoV-2 (%)	Sequence logo
NGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNN	45	81,250	0,638		4	48	96.8	
HGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLS	47	65,957	0,562	Yes	59	105	98.6	
AGLPYGAN	8	75,000	0,541	Yes	119	126	99.5	
GALNTPKDHIGTRNPANNAAI	21	76,190	0,595	Yes	137	157	99.2	
TLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD	51	88,235	0,662		166	216	70.0	
LNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK	40	85,000	0,614		227	266	98.7	
RRGPEQTQGNFGDQELIRQGTDYK	24	58,333	0,553		276	299	99.0	
DPNFKD	6	66,667	0,533		343	348	99.2	

DAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQ QSMSSADS	59	88,136	0,614		358	416	96.7

[a] - exposed epitope according to 3D structure visualized (pdb: 6VYO) using PyMOL (Schrodinger LLC 2015).