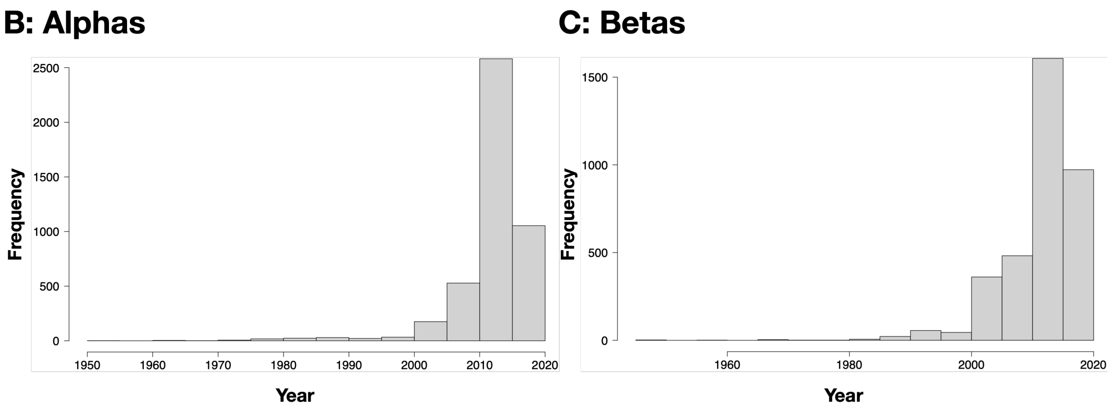
Figure S1:

Downloaded and cleaned dataset D1, by host (A) and year (B-C).

A

Alpha CoVs		Beta CoVs	
Host	**Count**	**Host**	**Count**
Alpaca	1	Antelope	4
Bat	146	Bat	199
Camel	63	Bison	1
Canine	73	Bovine	314
Feline	368	Buffalo	4
Ferret	18	Camel	383
Fox	1	Canine	19
Human	509	Chimp	7
Hyena	1	Civet	29
Mink	5	Deer	10
Murine	10	Equine	19
Porcine	3266	Ferret	1
Shrew	11	Giraffe	5
		Hedgehog	11
		Human	2384
		MERS	8
		Monkey	2
		Murine	87
		Pangolin	11
		Porcine	28
		Rabbit	8
		Raccoon Dog	3
		Tahr	2
		Waterbuck	6
		Yak	15
Total	**4472**	**Total**	**3560**



B: Alphas

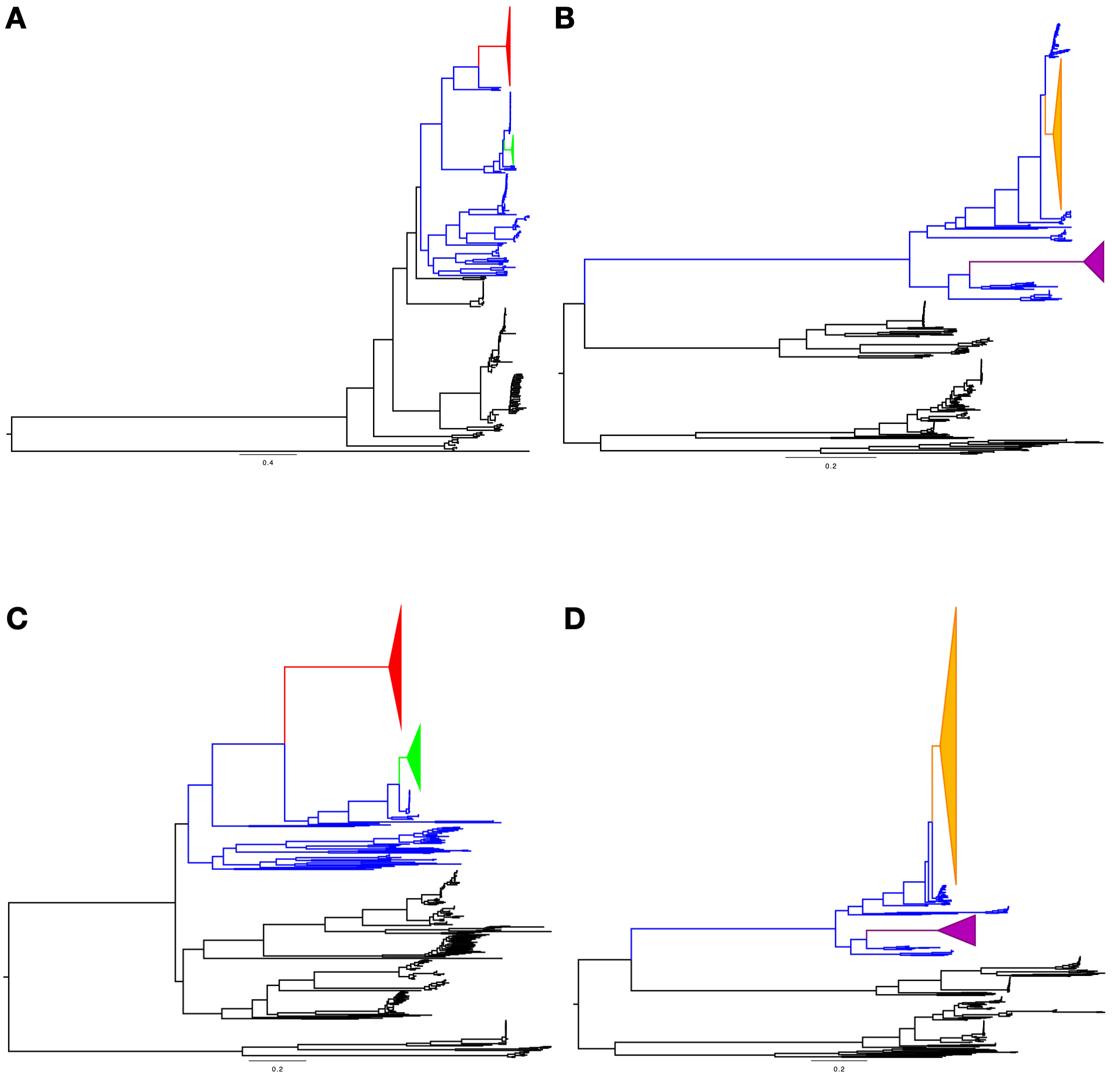
C: Betas

Figure S2:

Maximum Likelihood (ML) phylogenetic trees from dataset D2 for whole genomes (A-B), and the Spike (C-D), Nucleocapsid (E-F), Membrane (G-H) and Envelope (A-J) proteins. Branches highlighted in blue represent the non-sHCoV sequences selected for the final analyses in addition to the seasonal human coronaviruses (sHCoVs) colored in red (NL63), green (229E), orange (OC43) and purple (HKU1). On these trees, CoVs from hosts other than the sHCoVs and those that share a MRCA as per the ML trees had been subsampled to ~ 30-40 sequences per host-clade, i.e. if there were two porcine clades positioned on different parts of a tree then they were subsampled independently.

Key:

🟦 Selected sequences 🟥 NL63 🟩 229E 🟧 OC43 🟪 HKU1



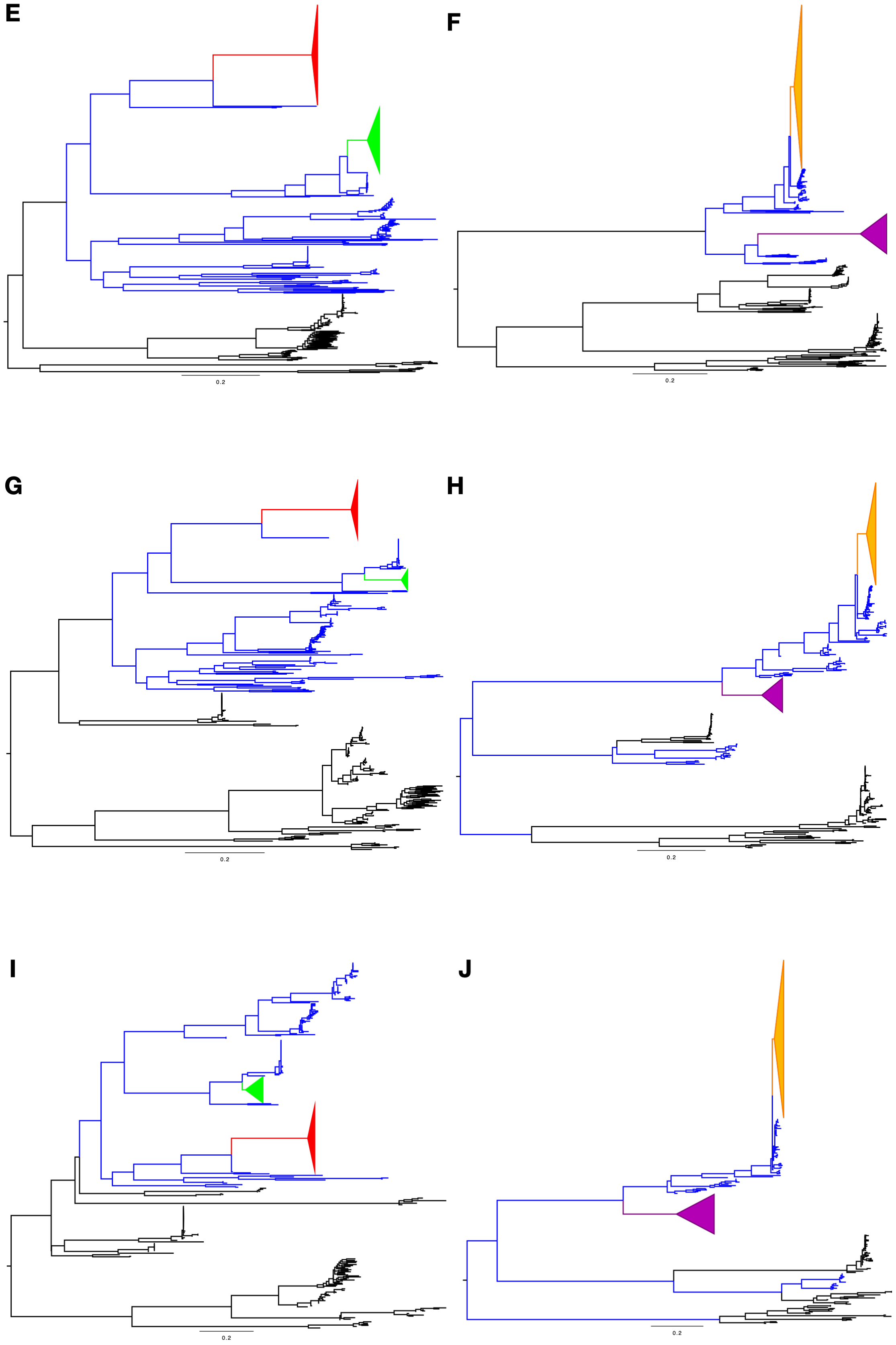

E
0.2
F
0.2
G
0.2
H
0.2
I
0.2
J
0.2

Figure S3:

Estimates of the MRCA age for full genomes and four open reading frames (dataset D5) of the seasonal human coronavirus species. The black horizontal lines represent the dates of first isolation for 229E (1966), OC43 (1967), NL63 (2004) and HKU1 (2005). The star (*) symbol shows the parents to the MRCAs of the sHCoVs. The WGS is missing data points for HKU1_all (collective for both genotypes) and genotype B as sequences for genotype B were all removed in the generation of recombination-free WGS dataset D5.

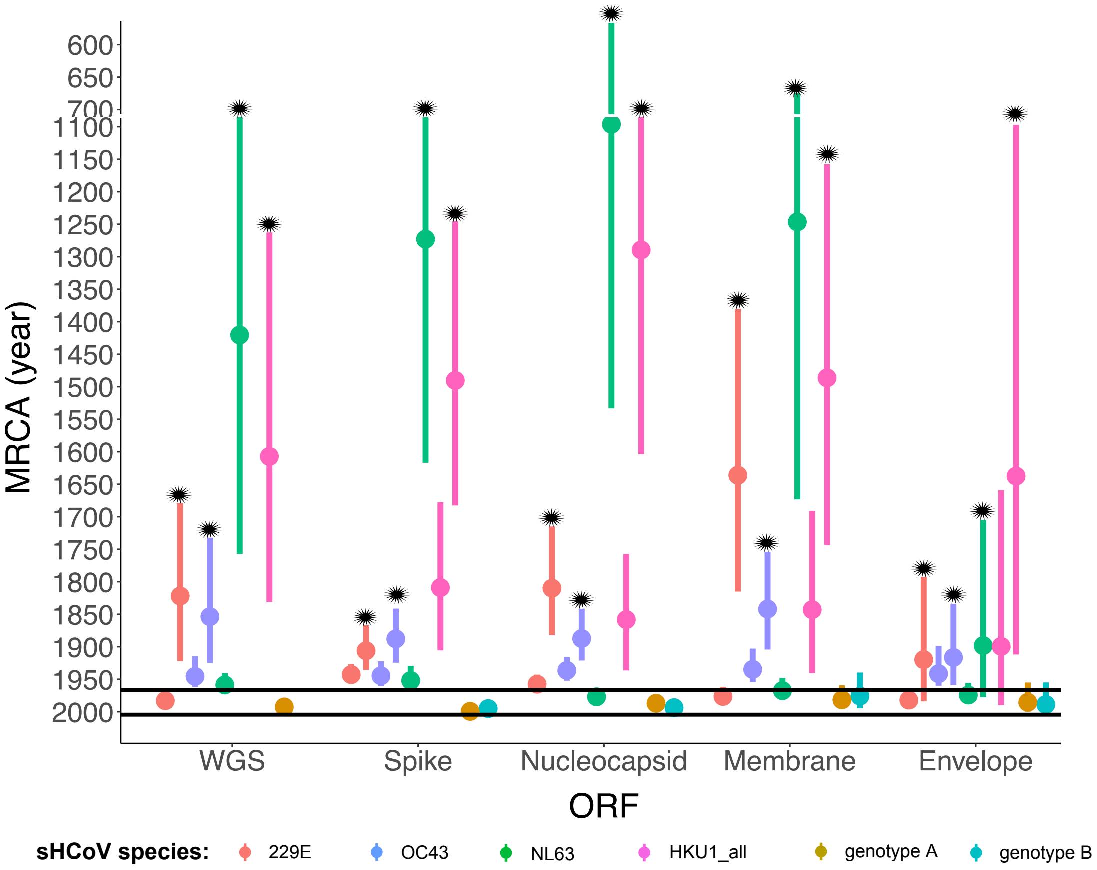


Table S1:

Mean pairwise genetic distances for various CoV host-clades. The CoV host-clade sequences selected for this analysis were based on ML tree topologies in *Supplementary Figure 1*, selecting host-clades that were sister clades to the sHCoVs. However, when the isolates that shared a recent ancestor with humans, rather than one in the more distant past, were represented by a single sequence, the mean pairwise genetic distance could not be calculated on the single sequence. We also included CoVs from hosts not closely related to the sHCoVs for comparison. For each sHCoV species and ORF/WGS, the cells are colored from the highest (red) to the lowest (green) genetic distance.

Genus	Species	CoV	WGS	Spike	Envelope	Membrane	Nucleocapsid
AlphaCoV	229E	229E	0.0048	0.0347	0.0040	0.0079	0.0121
		Camelid	0.0020	0.0056	0.0024	0.0015	0.0027
		Bat	0.0476	0.0521	NA	NA	NA
	NL63	NL63	0.0081	0.0362	0.0077	0.0082	0.0076
		Bat	NA	0.5048	NA	NA	NA
BetaCoV	OC43	OC43	0.0076	0.0268	0.0085	0.0082	0.0077
		Bovine	0.0104	0.0247	0.0081	0.0120	0.0099
		Ungulate & Canine	0.0157	0.0274	0.0091	0.0101	0.0116
		Equine	0.0087	0.0129	0.0128	0.0049	0.0075
		Rabbit	0.0051	NA	0.0028	0.0019	0.0019
		Murine	NA	0.1713	NA	NA	0.1377
	HKU1	HKU1	0.0222	0.1040	0.0795	0.0176	0.0231
		Genotype B	0.0122	0.0120	0.0072	0.0039	0.0049
		Genotype A	0.0028	0.0049	0.0011	0.0020	0.0063
		Porcine	0.0134	0.0367	0.0127	0.0102	0.0148

		Murine	0.1738	0.1424	0.0812	0.0311	0.0619
		Key: NA: Only a single sequence available					

Table S2:

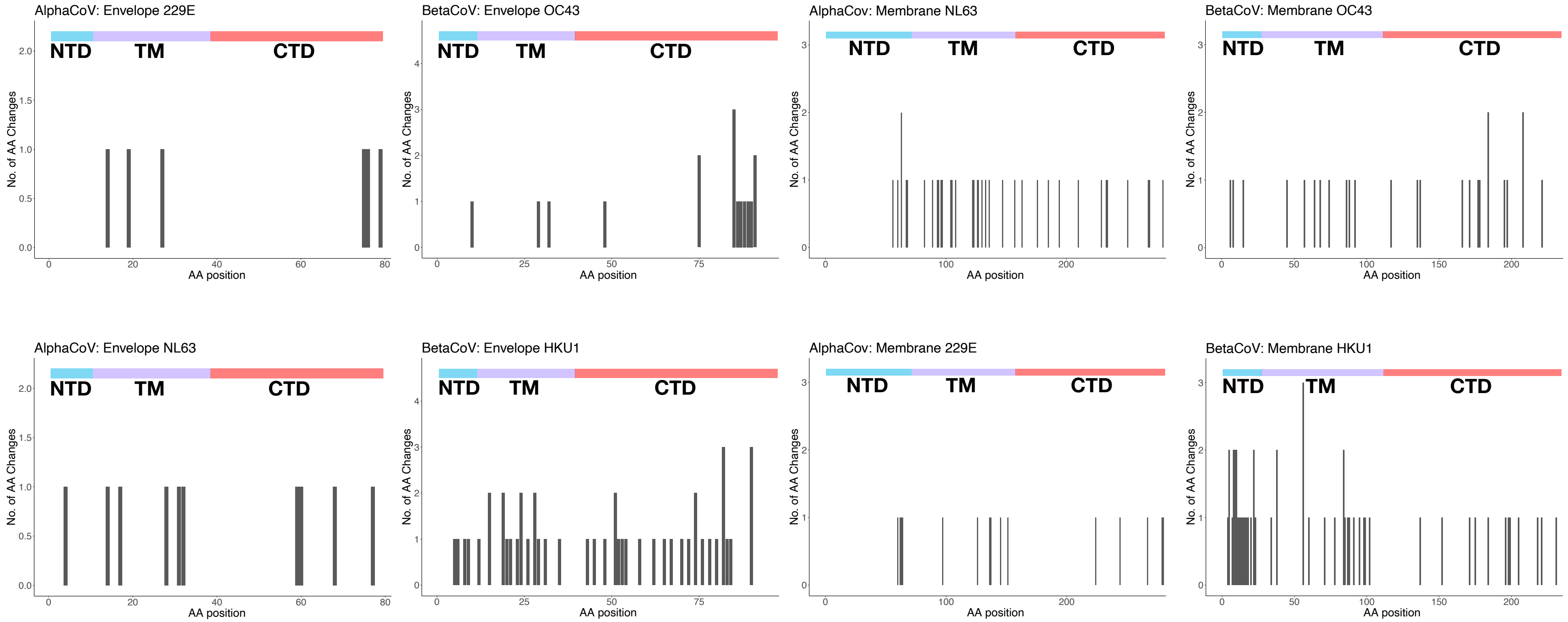
Test for positive selection in the emergence of sHCoV species from datasets D4 and D5. In bold are the host-jump branches where positive selection was inferred using the likelihood ratio test at a threshold of p≤0.05. Also shown are ω (ratio of nonsynonymous to synonymous substitutions) and the proportion of sites in each rate.

ORF	sHCoV species	aBSREL	BUSTED by sHCoV species
Spike	229E	ω_1=0.23 (100%); *p-value*=1.0	ω_1=0.03 (73.9%), ω_2=0.72 (11.1%), ω_3=2.54 (14.9%); *p-value*=0.418
	NL63	ω_1=0.00 (62%), ω_2=0.17 (29%), ω_3=3850 (9%); *p-value*=0.183	ω_1=0.01 (62.3%), ω_2=0.1 (37.1%), ω_3=1 (0.7%); *p-value*=0.5
	OC43	ω_1=0.51 (100%); *p-value*=1.0	ω_1=0.23 (26.6%), ω_2=0.37 (55.9%), ω_3=1 (17.5%); *p-value*=0.5
	HKU1_all	ω_1=0.00 (70%), ω_2=0.31 (23%), ω_3=307 (6.7%); *p-value*=0.074	ω_1=0.02 (85.3%), ω_2=0.19 (9.2%), ω_3=2.71 (5.5%); *p-value*=0.249
	HKU1 Genotype A	ω_1=0.0468 (96%), ω_2=2.48 (4%); *p-value*=0.2825	ω_1=0.00 (78.1%), ω_2=0.28 (21.4%), ω_3=12.15 (0.5%); *p-value*=0.329
	HKU1 Genotype B	ω_1=0.00 (87%), ω_2=1.0 (13%); *p-value*=0.5	ω_1=0.00 (79.2%), ω_2=0.33 (20.8%), ω_3=1.02 (0.0%); *p-value*=0.5
Nucleocapsid	229E	ω_1=0.40 (100%); *p-value*=1.0	ω_1=0.00 (17.7%), ω_2=0.02 (64.1%), ω_3=1.14 (18.2%); *p-value*=0.5
	NL63	**ω_1=0.05 (85%), ω_2=0.05 (1.7%), ω_3=62.6 (14%); *p-value*=0.000**	ω_1=0.02 (67.4%), ω_2=0.07 (21.1%), ω_3=2.55 (11.5%); *p-value*=0.425
	OC43	ω_1=0.23 (100%); *p-value*=1.0	ω_1=0.00 (0%), ω_2=0.20 (100%), ω_3=1.11 (0%); *p-value*=0.5
	HKU1_all	**ω_1=0.07 (90%), ω_2=3850 (10%); *p-value*=0.008**	ω_1=0.00 (56.5%), ω_2=0.24 (39.9%), ω_3=4000 (3.6%); *p-value*=0.063
	HKU1 Genotype A	ω_1=0.351 (100%); *p-value*=1.0	ω_1=0.99 (36.6%), ω_2=1.00 (0%), ω_3=23.58 (63.4%); *p-value*=0.254
	HKU1 Genotype B	**ω_1=0.00 (91%), ω_2=6.30 (8.6%); *p-value*=0.0305**	ω_1=0.00 (68.3%), ω_2=0.00 (16.6%), ω_3=3.72 (15.1%); *p-value*=0.122
Membrane	229E	ω_1=0.09 (98%), ω_2=13.3 (2%); *p-value*=0.2112	ω_1=0.00 (75.1%), ω_2=0.24 (21.9%), ω_3=3.81 (2.9%); *p-value*=0.414
	NL63	ω_1=0.00 (87%), ω_2=0.92 (13%); *p-value*=1.0	ω_1=0.00 (81.9%), ω_2=0.27 (18.1%), ω_3=1.00 (0%); *p-value*=0.5
	OC43	ω_1=0.43 (100%); *p-value*=1.0	ω_1=0.36 (11.7%), ω_2=0.36 (87.9%), ω_3=9999999171.60 (0.5%); *p-value*=0.168
	HKU1_all	ω_1=0.39 (100%); *p-value*=1.0	ω_1=0.03 (40.6%), ω_2=0.08 (59.4%), ω_3=1.00 (0%); *p-value*=0.5
	HKU1 Genotype A	ω_1=0.281 (100%); *p-value*=1.0	ω_1=0.00 (55.4%), ω_2=0.04 (30.2%), ω_3=1.52 (14.4%); *p-value*=0.483
	HKU1 Genotype B	ω_1=0.103 (100%); *p-value*=1.0	ω_1=0.27 (77.5%), ω_2=0.28 (22.5%), ω_3=1.00 (0%); *p-value*=0.5
Envelope	229E	ω_1=0.23 (100%); *p-value*=1.0	ω_1=0.23 (100%), ω_2=0.25 (0%), ω_3=1.00 (0%); *p-value*=0.5
	NL63	ω_1=0.09 (100%); *p-value*=1.0	ω_1=0.03 (0%), ω_2=0.07 (100%), ω_3=1.11 (0%); *p-value*=0.5
	OC43	ω_1=0.78 (100%); *p-value*=1.0	ω_1=1.00 (0%), ω_2=1.00 (0%), ω_3=14.11 (100%); *p-value*=0.457
	HKU1_all	ω_1=0.14 (100%); *p-value*=1.0	ω_1=0.00 (33.9%), ω_2=0.18 (66.2%), ω_3=1.00 (0%); *p-value*=0.5

	HKU1 Genotype A	ω_1=0.53 (100%); *p-value*=1.0	ω_1=0.28 (0%), ω_2=0.29 (100%), ω_3=1.00 (0%); *p-value*=0.5
	HKU1 Genotype B	ω_1=0.15 (100%); *p-value*=1.0	ω_1=1.00 (0%), ω_2=1.00 (0%), ω_3=3.14 (100%); *p-value*=0.463

Figure S4:

The number of inferred amino acid changes (AA) within sHCoV clades for AA positions in the envelope, membrane, nucleocapsid and spike proteins from datasets D4 and D5. At the top of each plot, the functional domains or regions of the respective proteins are shown; NTD=N-terminal domain, TM=transmembrane domain, CTD=C-terminal domain, RBD=receptor binding domain, LINK=central linker domain, LINK-Dimer=dimerization domain, S1 subunit, S2 subunit, FP=fusion peptide, IFP=internal fusion peptide, HR1=heptad repeat 1, and HR2=heptad repeat 2.



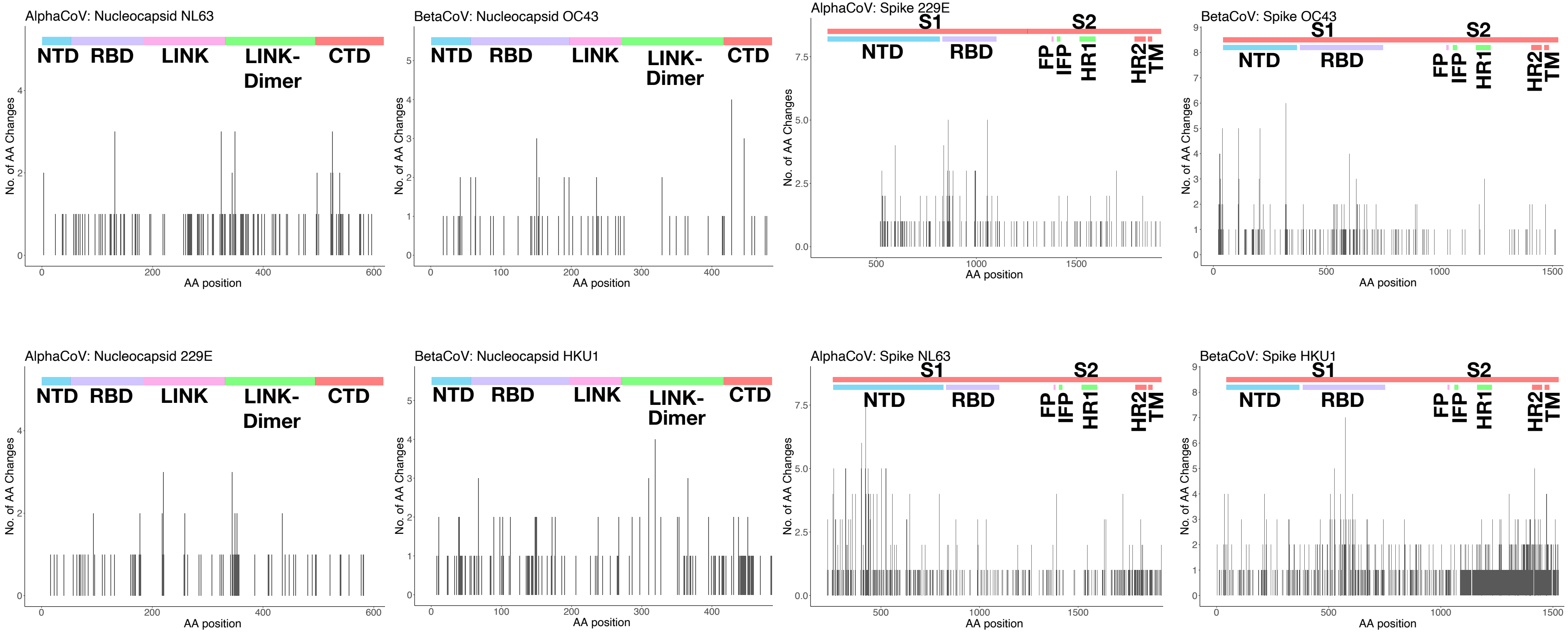

AlphaCoV: Nucleocapsid NL63
BetaCoV: Nucleocapsid OC43
AlphaCoV: Spike 229E
BetaCoV: Spike OC43
AlphaCoV: Nucleocapsid 229E
BetaCoV: Nucleocapsid HKU1
AlphaCoV: Spike NL63
BetaCoV: Spike HKU1
NTD
RBD
LINK
LINK-Dimer
CTD
S1
S2
FP
IFP
HR1
HR2
TM
No. of AA Changes
AA position
0
2
4
0
1
2
3
4
5
0.0
2.5
5.0
7.5
0
1
2
3
4
5
6
7
8
9
0
200
400
600
0
100
200
300
400
500
1000
1500

Table S3:

A select amino acid changes occurring along the host-jump branches leading to the emergence of the sHCoVs

Type of AA change along the host-jump branch	**ORF**			
	Spike	**Nucleocapsid**	**Membrane**	**Envelope**
Convergent changes to the same AA	*NL63 & HKU1*[a]*:* L/G1224F, V/C1228L *NL63 & OC43:* V/T1142I *229E & HKU1:* I/D1153S *HKU1 & OC43:* N/Y489D, A/T497P	*NL63 & 229E:* N/A402S *NL63 & HKU1:* A/P182S, A/T241S	-	-
Divergent changes to different AAs from the same AA	*NL63 & 229E:* I769L/N, T1253M/I, I1153A/S *NL63, 229E & HKU1:* D1170T/E/N *NL63 & HKU1:* T1116V/S, V1227L/F *HKU1 & OC43:* D1135L/Y, G1221A/C, D463.11S/N, Q490.7W/-, T8.21N/K, V8.27F/K, T680S/K *HKU1, OC43 & HUMAN*[b]*:* D1181Y/V/E	*NL63 & 229E:* E236S/D *229E & HKU1:* I127V/L	*NL63 & 229E:* L82I/F *229E & HKU1:* E12D/Q	229E & HKU1 V26I/F

	HKU1& HUMAN: V105T/L			
Parallel changes from-and-to the same AA	*NL63 & 229E:* A1009S, L1210I *HKU1 & OC43:* N146K *HKU1 & HUMAN:* F8.6L[c]			
Changes to AAs observed in SARS-CoV-2	*OC43:* N856K (Omicron) *HKU1:* R969K (Omicron), E339D (Omicron), F371L (Omicron) *NL63:* P80A(Beta), T375S (Wuhan-Hu-1)	*HKU1:* E63D (Wuhan-Hu-1)	*NL63:* L82I (Wuhan-Hu-1)	-
Completely reversed AA changes	*NL63 & HKU1:* I/V1225V/I N/D1165D/N	-	-	-
Partially reversed AA changes	*NL63 & HKU1:* S/N162Y/S, S/A271T/S, T/A629S/T, A/N776N/T, V/A817A/F, S/A975N/S, E/Q1154Q/H *NL63 & HKU1 & OC43:* I/S/T624T/-/S *229E & HKU1:* N/K641K/- *HKU1 & OC43:* K/I154N/K, K/T257I/K, V/I329I/K	*NL63 & 229E:* D/S158S/N NL63 & HKU1: E/D125Q/E *229E & HKU1:* D/E321V/D, S/L365-/S	*NL63 & 229E:* V/I222I/F	*NL63 & HKU1:* L/I27F/L

	HKU1 & HUMAN: S/G405R/S			

Key:

[a] Where there are two or more AAs, the order of the AAs follow the sHCoV species shown, i.e. for *NL63 & HKU1* in L/G1224F represents NL63: L1224F and HKU1: in G1224F

[b] A lone human CoV (FJ415324) that clusters with ungulate and canine CoVs

[c] Where there was an AA insertion in the sHCoVs relative to the Wuhan-Hu-1 SARS-CoV-2 reference genome, we used the X.Y positional notation where X is the reference genome position and Y is the nth sHCoV AA insertion.