

Supplementary Material

Recurrent recombination and insertion–deletion events shape the genome-wide evolutionary history of the *Coronaviridae* family

The supplementary material includes the Tables S1-S5 and Figures S1-S3.

Supplementary Tables

Table S1. Representative *Coronaviridae* genomes included in the study dataset. The table lists the analyzed whole-genome sequences, which represent all *Coronaviridae* genera available in NCBI as of January 2024. For each virus genome, the table provides the corresponding viral species, a short identifier (Nickname) used in this study, the NCBI accession code, sampling date and country, taxonomic genus, genome size in nucleotides, and the number of annotated proteins. The nickname begins with a letter indicating the genus (*A* = *Alphacoronavirus*, *B* = *Betacoronavirus*, *G* = *Gammacoronavirus*, *D* = *Deltacoronavirus*, *U* = *Unclassified in NCBI*), followed by an abbreviation of the species name (i.e., “A-HCoV-229E” refers to *Alphacoronavirus* human coronavirus 229E, and “B-SARS-CoV-2” refers to *Betacoronavirus* SARS-CoV-2). NA indicates that information is not available.

Species	Nickname	Genus	NCBI code	Sampling date	Country	Genome size (nt)	Number of annotated proteins
Bat coronavirus	U-BatCoV1	Unclassified	NC_048212	Dec-2013	Cameroon	28975	9
Bat coronavirus	U-BatCoV2	Unclassified	NC_034440	20-Feb-2013	Uganda	29642	11
Rodent coronavirus	U-RodentCoV	Unclassified	NC_046954	Aug-2014	China	31393	7
Alphacoronavirus Bat-CoV/P.kuhlii/Italy/3398-19/2015	A-BatCoV3	Alpha	NC_046964	2015	Italy	28128	6
Bat Hp-betacoronavirus/Zhejiang2013	B-BatCoV4	Beta	NC_025217	29-Apr-2013	China	31491	10
Bat coronavirus 1A	A-BatCoV5	Alpha	NC_010437	NA	Hong Kong	28326	7
Bat coronavirus BM48-31/BGR/2008	B-BatCoV6	Beta	NC_014470	2008	Bulgaria	29276	10
Bat coronavirus CDPHE15/USA/2006	A-BatCoV7	Alpha	NC_022103	07-Aug-2006	USA: Boulder County, CO	28035	7
Beluga whale coronavirus SW1	G-BelugaCoV	Gamma	NC_010646	NA	NA	31686	14
Betacoronavirus England 1	B-EnCoV1	Beta	NC_038294	11-Sept-2012	United Kingdom	30111	11
Middle East respiratory syndrome-related coronavirus	B-MERSCoV	Beta	NC_019843	13-Jun-2012	NA	30119	11
Betacoronavirus	B-						
Erinaceus/VMC/DEU/2012	ErinaceusCoV	Beta	NC_039207	2012	Germany	30148	12
Betacoronavirus HKU24	B-HKU24	Beta	NC_026011	17-May-2012	China	31249	11
Bovine coronavirus	B-BovineCoV	Beta	NC_003045	NA	NA	31028	12

Human coronavirus OC43	B-OC43	Beta	NC_006213	NA	NA	30741	10
BtMr-AlphaCoV/SAX2011	A-BtMrCoV	Alpha	NC_028811	2011	China	27935	6
BtNv-AlphaCoV/SC2013	A-BtNvCoV	Alpha	NC_028833	2013	China	27783	6
BtRf-AlphaCoV/HuB2013	A-BtRfCoV1	Alpha	NC_028814	2013	China	27608	7
BtRf-AlphaCoV/YN2012	A-BtRfCoV2	Alpha	NC_028824	2012	China	26975	6
Bulbul coronavirus HKU11-934	D-HKU11-934	Delta	NC_011547	Jan-2007	Hong Kong	26487	9
Camel alphacoronavirus	A-CamelCoV	Alpha	NC_028752	Mar-2015	Saudi Arabia	27395	7
	A-HumanCoV-229E	Alpha	NC_002645	NA	NA	27317	8
Human coronavirus 229E	G-GooseCoV	Gamma	NC_046965	Aug-2017	Canada	28539	16
Canada goose coronavirus	D-HKU21	Delta	NC_016996	2007	Hong Kong	26223	10
Common moorhen coronavirus HKU21	A-AcCoV-JC34	Alpha	NC_034972	Oct-2011	China	27682	9
Coronavirus AcCoV-JC34	G-DuckCoV	Gamma	NC_048214	2014	China	27754	12
Duck coronavirus	G-IBV1	Gamma	NC_001451	NA	NA	27608	10
Infectious bronchitis virus	G-IBV2	Gamma	NC_048213	2003	India	27464	10
Infectious bronchitis virus	G-TurkeyCoV	Gamma	NC_010800	NA	Canada	27657	11
Turkey coronavirus	A-FIPV	Alpha	NC_002306	NA	NA	29355	9
Feline infectious peritonitis virus	A-SwineCoV	Alpha	NC_028806	04-Sept-2009	Italy	28111	9
Swine enteric coronavirus	A-TGEV	Alpha	NC_038861	NA	USA, Indiana	28586	8
Transmissible gastroenteritis virus	A-FerretCoV	Alpha	NC_030292	2010	Netherlands	28434	9
Ferret coronavirus	B-HKU1	Beta	NC_006577	NA	NA	29926	9
Human coronavirus HKU1	A-NL63	Alpha	NC_005831	NA	ND	27553	7
Human coronavirus NL63	A-RatCoV1	Alpha	NC_032730	2013	China	28763	5
Lucheng Rn rat coronavirus	D-HKU18	Delta	NC_016993	2007	Hong Kong	26689	10
Magpie-robin coronavirus HKU18	A-HKU8	Alpha	NC_010438	NA	Hong Kong	28773	8
Miniopiterus bat coronavirus HKU8	A-MinkCoV	Alpha	NC_023760	01-Jan-1998	USA: Wisconsin	28941	10
Mink coronavirus strain WD1127	D-HKU13-3514	Delta	NC_011550	Apr-2007	Hong Kong	26552	9
Munia coronavirus HKU13-3514	B-MHV1	Beta	NC_048217	NA	NA	31335	11
Murine hepatitis virus	B-MHV2	Beta	NC_001846	NA	NA	31357	8
Murine hepatitis virus							

Rat coronavirus Parker	B-RatCoV2	Beta	NC_012936	NA	NA	31250	10
NL63-related bat coronavirus	A-BatCoV8	Alpha	NC_048216	21-Jun-2010	Kenya	28679	8
NL63-related bat coronavirus	A-BatCoV9	Alpha	NC_032107	21-Jun-2010	Kenya	28363	8
Night heron coronavirus HKU19	D-HKU19	Delta	NC_016994	2007	Hong Kong China, Guangdong province	26077	8
Pipistrellus bat coronavirus HKU5	B-HKU5	Beta	NC_009020	NA	Hong Kong	30482	11
Porcine coronavirus HKU15	D-HKU15	Delta	NC_039208	2010	NA	25425	7
Porcine epidemic diarrhea virus	A-PEDV	Alpha	NC_003436	NA	China	28033	6
Rabbit coronavirus HKU14	B-HKU14	Beta	NC_017083	2006	NA	31100	12
Rhinolophus bat coronavirus HKU2	A-HKU2	Alpha	NC_009988	NA	China	27165	8
Rousettus bat coronavirus	B-BatCoV10	Beta	NC_030886	28-May-2014	China	30161	10
Rousettus bat coronavirus HKU10	A-HKU10	Alpha	NC_018871	Oct-2005	China, Guangdong province	28494	9
Rousettus bat coronavirus HKU9	B-HKU9	Beta	NC_009021	NA	Canada:	29114	9
SARS coronavirus Tor2	B-SARS-CoV-Tor2	Beta	NC_004718	NA	Toronto	29751	15
Severe acute respiratory syndrome coronavirus 2	B-SARS-CoV-2	Beta	NC_045512	Dec-2019	China	29903	12
Scotophilus bat coronavirus 512	A-BatCoV11	Alpha	NC_009657	NA	NA	28203	6
Shrew coronavirus	U-ShrewCoV1	Unclassified	NC_046955	Dec-2014	China	27102	5
Sparrow coronavirus HKU17	D-HKU17	Delta	NC_016992	2007	Hong Kong	26083	9
Thrush coronavirus HKU12-600	D-HKU12-600	Delta	NC_011549	Jan-2007	Hong Kong China, Guangdong province	26396	9
Tylonycteris bat coronavirus HKU4	B-HKU4	Beta	NC_009019	NA	China	30286	11
Wencheng Sm shrew coronavirus	A-ShrewCoV2	Alpha	NC_035191	2015	China	25995	5
Wencheng Sm shrew coronavirus	A-ShrewCoV3	Alpha	NC_048211	2015	China	25984	5
White-eye coronavirus HKU16	D-HKU16	Delta	NC_016991	2007	Hong Kong	26041	8
Wigeon coronavirus HKU20	D-HKU20	Delta	NC_016995	2008	Hong Kong	26227	11

Table S2. Prior distributions used for the parameters of the indel models. These prior distributions were employed to sample parameter values for simulating genetic data under each indel model in the approximate Bayesian computation method. The parameters ‘q’ and ‘r’ correspond to the parameters of the Binomial indel-length model as implemented in *INDELible* (Fletcher and Yang 2009), applied to both insertion and deletion events. The parameter “a” in the Zipf model corresponds to the scaling exponent of the distribution. Sequence length is denoted by “l”. These prior distributions follow those used in previous studies, such as (Loewenthal et al. 2021).

Parameter	Binomial model	Zipf truncated model
Insertion rate	U (0 – 0.05)	U (0 – 0.05)
Deletion rate	U (0 – 0.05)	U (0 – 0.05)
q for insertion	U (0.1 - 0.5)	-
q for deletion	U (0.1 - 0.5)	-
r for insertion	U (1 – 20)	-
r for deletion	U (1 – 20)	-
a for insertion	-	U (1.1 – 2)
a for deletion	-	U (1.1 – 2)
t (truncation)	-	50
Root length	$0.8 \times l - 1.1 \times l$	$0.8 \times l - 1.1 \times l$

Table S3. Summary statistics used in the approximate Bayesian computation method for indel model selection and parameter estimation. SS and MSA refer to summary statistics and multiple sequence alignment, respectively. These summary statistics were selected based on previous studies, such as (Loewenthal et al. 2021).

SS number	SS specification
1	Total number of indel blocks in the alignment
2	Total number of unique indel blocks in the alignment
3	Average indel block length
4	Average unique indel block length
5	Number of indel blocks of length one
6	Number of indel blocks of length two
7	Number of indel blocks of length three
8	Number of indel blocks of length four or more
9	Alignment length
10	Minimum length of sequence in the alignment
11	Maximum length of sequence in the alignment
12	Number of MSA columns with zero indel
13	Number of MSA columns with one indel
14	Number of MSA columns with two indel
15	Number of MSA columns with $n-1$ indels
16	Number of indels of length one that appear only in one sequence
17	Number of indels of length one that are shared between exactly two sequences
18	Number of indels of length one that are shared between exactly $n-1$ sequences
19	Number of indels of length two that appear only in one sequence
20	Number of indels of length two that are shared between exactly two sequences
21	Number of indels of length two that are shared between exactly $n-1$ sequences
22	Number of indels of length three that appear only in one sequence
23	Number of indels of length three that are shared between exactly two sequences
24	Number of indels of length three that are shared between exactly $n-1$ sequences
25	Number of indels of length at least four that appear only in one sequence
26	Number of indels of length at least four that are shared between exactly two sequences
27	Number of indels of length at least four that are shared between exactly $n-1$ sequences

Table S4. Insertion and deletion rates used to evaluate the accuracy of parameter estimation with the approximate Bayesian computation method. The table shows various combinations of parameters (e.g., insertion rate = 0 and deletion rate = 0.01, or insertion rate = 0 and deletion rate = 0.03) that were used to evaluate the approximate Bayesian computation method in the analysis of the *Coronaviridae* family and SARS-CoV-2 datasets. These parameter values were evenly distributed within the prior distributions. Note that the parameters differ between the coronavirus family and SARS-CoV-2 datasets to provide proper fitting with the expectations (i.e., the number of observed indels is much higher in the *Coronaviridae* family dataset than in the SARS-CoV-2 dataset).

	<i>Coronaviridae</i> family								SARS-CoV-2		
Insertion rate	0	0	0	0.01	0.03	0.05	0.025	0.02	0.04	0.002	0.004
Deletion rate	0.01	0.03	0.05	0	0	0	0.025	0.04	0.02	0.004	0.002

Table S5. High-confidence recombination events detected in *Coronaviridae* genomes. For each recombination event (row), the table includes the following information: breakpoint position in the alignment of genome sequences, gene/s affected by the recombination event, recombinant sequences involved in the event (A-/B-/G-/D-/U- refer to *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, *Deltacoronavirus*, unclassified, respectively), and the major/minor parental lineages of the recombination event (Unknown or parentheses indicate cases where parents are not present in the study data and are inferred based on similarity to the actual parental sequence). Only recombination events that were statistically supported (p -value < 0.05) by four or more well-established recombination detection methods (RDP, GENECONV, MaxChi, BootScan and SiScan) implemented in RDP5 (Martin et al. 2021) are included.

Breakpoint begin	Breakpoint end	Affected genes	Recombinant genome/s	Minor parent/s genomes	Major parent/s genomes
690	929	5'UTR	D-HKU15, D-HKU17	(D-HKU21), Unknown	D-HKU18
2000	2244	ORF1ab	G-GooseCoV	A-CamelCoV, A- Human CoV- 229E	G-IBV1, G-IBV2, G-TurkeyCoV A-AcCoV-JC34,
2095	4313	ORF1ab	A-ShrewCoV2, A-ShrewCoV3	B-HKU1	A-RatCoV1
2416	2759	ORF1ab	G-IBV1, G-IBV2, G-TurkeyCoV A-BatCoV5, A-BatCoV7, A-BtMrCoV, A-BtRfCoV1, A- BtRfCoV2, A-CamelCoV, A-HumanCoV-229E, A-AcCoV- JC34, A-FIPV, A-SwineCoV, A-TGEV, A-FerretCoV, A- NL63, A-RatCoV1, A-HKU8, A-MinkCoV, A-BatCoV8, A- BatCoV9, A-PEDV, A-HKU2, A-HKU10, A-BatCoV11	B-HKU5	G-GooseCoV
3387	3611	ORF1ab		(B-BovineCoV), Unknown, Unknown (B- OC43)	B-HKU24
4334	6000	ORF1ab	G-GooseCoV	A-HumanCoV- 229E, B-HKU4	G-TurkeyCoV
4350	8820	ORF1ab	U-RodentCoV, B-HKU24, B-BovineCoV, B-OC43, B-HKU1, B-MHV1, B-MHV2, B-RatCoV2, B-HKU14	A-ShrewCoV2, A-ShrewCoV3 (U-ShrewCoV1), Unknown	(A-RatCoV1), Unknown
5708	8776	ORF1ab	A-ShrewCoV2, A-ShrewCoV3	Unknown	A-BtRfCoV2
11028	12004	ORF1ab	A-ShrewCoV2, A-ShrewCoV3	B-HKU1	A-FerretCoV (D-HKU18),
11032	12850	ORF1ab	D-HKU21	D-HKU12-600 U-BatCoV2, B- EnCoV1, B- MERSCoV, B-	Unknown
11228	14459	ORF1ab	G-DuckCoV, G-IBV1, G-IBV2, G-TurkeyCoV		G-GooseCoV

				ErinaceusCoV, B-HKU5	
12058	14505	ORF1ab	U-RodentCoV, B-HKU24, B-BovineCoV, B-OC43, B-HKU1, B-MHV1, B-MHV2, B-RatCoV2, B-HKU14	A-ShrewCoV2, A-ShrewCoV3 (D-HKU11-934), Unknown	U-BatCoV2, B-EnCoV1, B-MERSCoV, B-HKU5, B-HKU4
12848	13508	ORF1ab	U-BatCoV2, B-EnCoV1, B-MERSCoV, B-ErinaceusCoV, B-HKU5, B-HKU4	(A-BatCoV7), Unknown	B-SARS-CoV-Tor2
15505	15704	ORF1ab	A-FIPV, A-SwineCoV, A-TGEV	(B-HKU14), Unknown	A-FerretCoV
16497	16887	ORF1ab	G-BelugaCoV, G-DuckCoV, G-IBV1, G-IBV2, G-TurkeyCoV	Unknown	A-PEDV
18113	18229	ORF1ab	D-HKU11-934, D-HKU21, D-HKU18, D-HKU13-3514, D-HKU19, D-HKU15, D-HKU17, D-HKU12-600, D-HKU16, D-HKU20	(B-HKU4), Unknown	A-HumanCoV-229E
18743	19728	ORF1ab	U-RodentCoV, B-HKU24, B-BovineCoV, B-OC43, B-HKU1, B-MHV1, B-MHV2, B-RatCoV2, B-HKU14	A-ShrewCoV2, A-ShrewCoV3 (U-BatCoV2), Unknown, Unknown(B-HKU4)	U-BatCoV1
18991	19761	ORF1ab	G-GooseCoV		G-IBV1, G-IBV2, G-TurkeyCoV
19092	19532	ORF1ab	A-BatCoV3, A-BatCoV5, A-BatCoV7, A-BtMrCoV, A-BtNvCoV, A-BtRfCoV1, A-AcCoV-JC34, A-FIPV, A-SwineCoV, A-TGEV, A-FerretCoV, A-NL63, A-RatCoV1, A-HKU8, A-MinkCoV, A-BatCoV8, A-BatCoV9, A-PEDV, A-HKU2, A-HKU10, A-BatCoV11	(D-HKU21), Unknown	A-ShrewCoV2, A-ShrewCoV3
19983	20069	ORF1ab	B-BatCoV4, B-BatCoV6, B-SARS-CoV-Tor2, B-SARS-CoV-2	(A-RatCoV1), Unknown	B-HKU1
20337	20487	ORF1ab	D-HKU21, D-HKU18, D-HKU19, D-HKU17, D-HKU16	(B-ErinaceusCoV), Unknown	A-FerretCoV (A-BatCoV7), Unknown
20482	20826	ORF1ab	A-AcCoV-JC34, A-RatCoV1	A-BtRfCoV2, A-HKU2	
22185	22368	ORF1ab	A-BatCoV3, A-BatCoV5, A-BatCoV7, A-BtMrCoV, A-BtNvCoV, A-BtRfCoV1, A-CamelCoV, A-HumanCoV-229E,	D-HKU21	A-ShrewCoV2, A-ShrewCoV3

			A-AcCoV-JC34, A-FIPV, A-TGEV, A-FerretCoV, A-NL63, A-RatCoV1, A-HKU8, A-MinkCoV, A-BatCoV8, A-BatCoV9, A-PEDV, A-HKU2, A-HKU10, A-BatCoV11		
23670	24068	ORF1ab	A-CamelCoV, A-HumanCoV-229E	G-BelugaCoV	A-BatCoV3
23946	24178	ORF1ab	A-BtRfCoV2, A-HKU2	U-ShrewCoV1	A-BtRfCoV1
			G-BelugaCoV, G-DuckCoV, G-IBV1, G-IBV2, G-TurkeyCoV, D-HKU13-3514, D-HKU15, D-HKU17, D-HKU12-600, D-HKU16	(B-BatCoV10), Unknown	U-BatCoV2 (G-IBV2), Unknown,
25990	26315	ORF1ab	A-BatCoV3, A-BatCoV5, A-BatCoV7, A-BtMrCoV, A-BtNvCoV, A-BtRfCoV1, A-BtRfCoV2, A-CamelCoV, A-HumanCoV-229E, A-NL63, A-HKU8, A-BatCoV8, A-BatCoV9, A-PEDV, A-HKU2, A-HKU10, A-BatCoV11	D-HKU11-934, D-HKU17	Unknown(G-IBV1) (D-HKU11-934),
26057	26390	ORF1ab	U-RodentCoV, B-HKU24, B-BovineCoV, B-OC43, B-HKU1, B-MHV1, B-MHV2, B-RatCoV2, B-HKU14	A-AcCoV-JC34 A-BtRfCoV2, A-HKU2	Unknown
26494	26777	ORF1ab			
26546	26788	ORF1ab	U-BatCoV1, B-BatCoV10, B-HKU9		G-BelugaCoV
			A-BatCoV3, A-BatCoV5, A-BatCoV7, A-BtMrCoV, A-BtNvCoV, A-BtRfCoV1, A-CamelCoV, A-HumanCoV-229E, A-AcCoV-JC34, A-FerretCoV, A-NL63, A-RatCoV1, A-HKU8, A-MinkCoV, A-BatCoV8, A-BatCoV9, A-PEDV, A-HKU10, A-BatCoV11, U-ShrewCoV1, A-ShrewCoV2, A-ShrewCoV3	U-BatCoV2, U-ShrewCoV1, B-HKU4 (D-HKU15), Unknown, Unknown (D-HKU17)	D-HKU18
26901	27560	ORF1ab		A-HumanCoV-229E	G-IBV1, G-IBV2, G-TurkeyCoV
28932	29846	ORF1ab	G-GooseCoV		
29145	29451	ORF1ab	D-HKU20		D-HKU19
30888	44669	ORF1ab, NS2, HE, ORF2, S	A-SwineCoV		
36096	41515	S	A-TGEV, A-FerretCoV, A-MinkCoV	A-PEDV	A-TGEV
36393	41582	S	A-HumanCoV-229E	D-HKU20	A-AcCoV-JC34, A-RatCoV1
				D-HKU20	A-ShrewCoV2, A-ShrewCoV3

38551	39113	S	U-BatCoV2	(B-HKU4), Unknown	B-EnCoV1, B- MERSCoV	B-
39268	40540	S	U-BatCoV2	(B-HKU4), Unknown	B-EnCoV1, B- MERSCoV	B-
41258	43243	S	A-BatCoV11	D-HKU11-934	A-RatCoV1	
41483	41723	S	A-BtRfCoV1	G-DuckCoV	A-BtMrCoV	
41673	43242	S	A-FIPV, A-TGEV, A-FerretCoV, A-MinkCoV D-HKU11-934, D-HKU21, D-HKU13-3514, D-HKU19, D- HKU15, D-HKU16, D-HKU20	D-HKU11-934, A-HumanCoV- 229E	A-AcCoV-JC34 A-AcCoV-JC34, A-RatCoV1	
41681	43220	S	A-BatCoV3, A-BtMrCoV, A-BtRfCoV1, A-HumanCoV- 229E, A-NL63, A-HKU8, A-BatCoV8, A-PEDV	(G-DuckCoV), Unknown	A-AcCoV-JC34, A-RatCoV1	
41724	42510	S	A-AcCoV-JC34, A-RatCoV1, U-ShrewCoV1, A-ShrewCoV2, A-ShrewCoV3	U-BatCoV2 (G-GooseCoV), Unknown	B-BovineCoV	
42520	43328	S	G-DuckCoV	(D-HKU20), Unknown	G-IBV2	
50135	50774	E, M	D-HKU21	Unknown	D-HKU13-3514	
50416	50790	M	A-BatCoV9	A-BtRfCoV1 (B-BatCoV10), Unknown	A-NL63	
60096	60380	N	A-AcCoV-JC34, A-RatCoV1	B-BatCoV6, B- SARS-CoV-2	A-FIPV	
60350	60718	N	A-PEDV		A-BtMrCoV	
60856	71499	N, Internal protein (N gene), p10, ORF7_2 and 3x-like protein, ORFs8- 11, ORFz, ORFx_2, ORFy, 3'-UTR	D-HKU11-934	D-HKU13-3514	(D-HKU17), Unknown, Unknown, (D- HKU15)	
63912	68054	ORF7 and 3x-like protein, ORF8, ORF9 ORF7 and 3x-like protein, ORF8-11, ORFz, ORFx_2, ORFy, 3'-UTR	D-HKU16	D-HKU15, D- HKU17	D-HKU11-934, D-HKU12-600	
63916	71520	ORF7 and 3x-like protein	D-HKU12-600	D-HKU13-3514	(D-HKU17), Unknown	
64358	64561		B-BatCoV10	A-HKU10	B-HKU9	

Supplementary Figures

Figure S1. Evaluation of indel model selection using the approximate Bayesian computation method. The figure shows the posterior probability of correctly identifying the true indel model using the approximate Bayesian computation method for the *Coronaviridae* family genome dataset (A) and spike coding gene dataset (B), as well as for the SARS-CoV-2 genome dataset (C) and spike coding gene dataset (D). Error bars represent the 95% confidence intervals from the mean of posterior probabilities for recovering the true indel model in the test data.

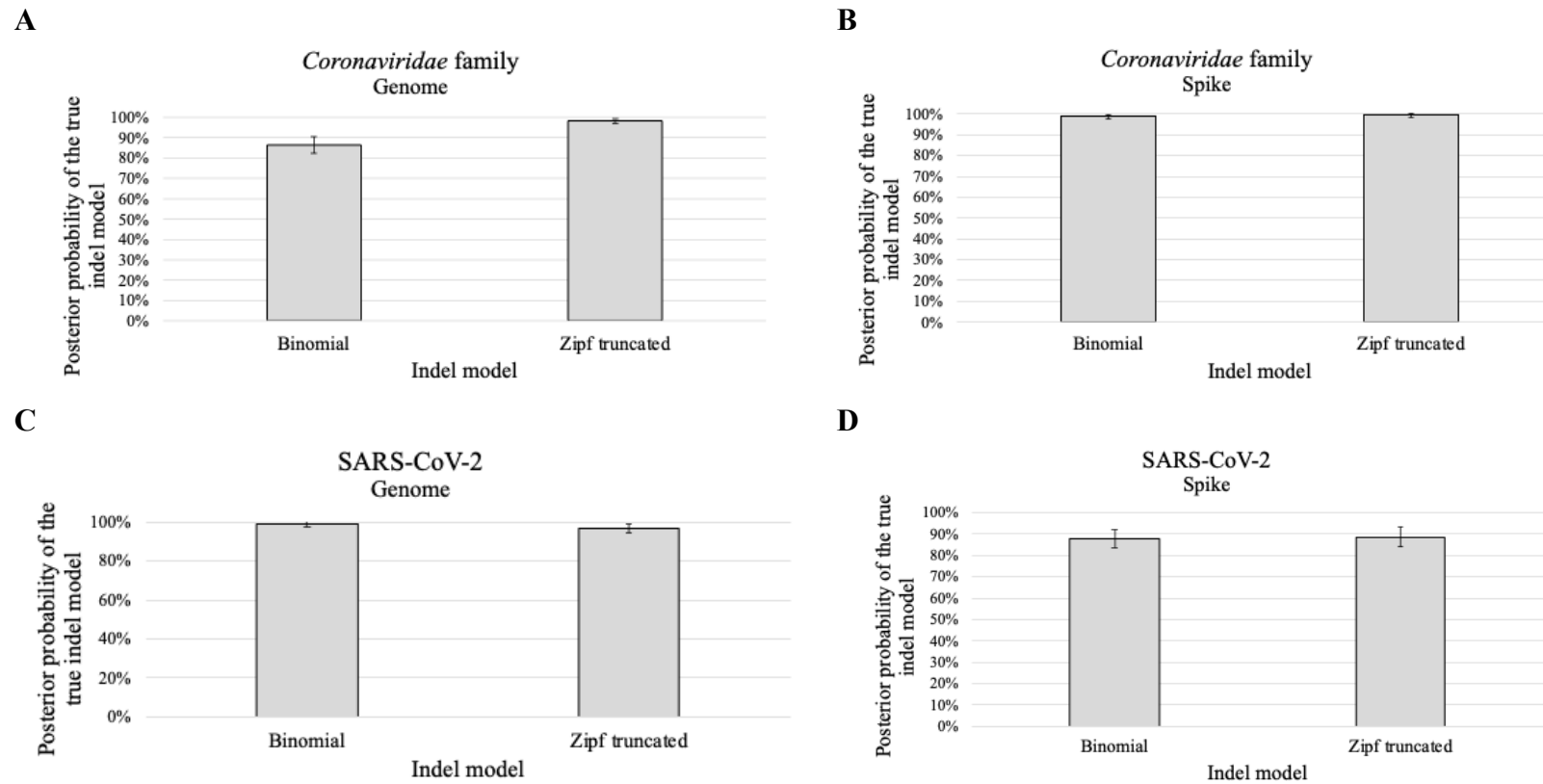
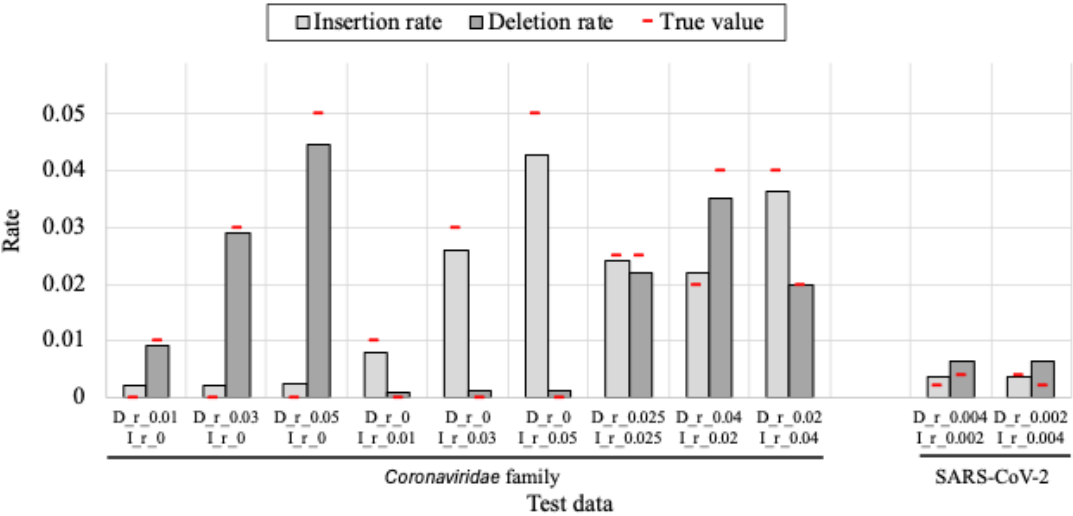


Figure S2. Evaluation of insertion and deletion rate estimation using the approximate Bayesian computation method. The figure shows the true insertion and deletion rates (red line, also indicated on the x-axis) and the corresponding estimated insertion (clear grey bars) and deletion (dark grey bars) rates for the genome (A) and spike coding gene (B) of *Coronaviridae* family and SARS-CoV-2 test data. “ D_r ” refers to the deletion rate and “ I_r ” refers to the insertion rate. In general, the true rates were accurately estimated, with the highest errors occurring at extremely high rates, where the genetic signatures can be reduced.

A



B

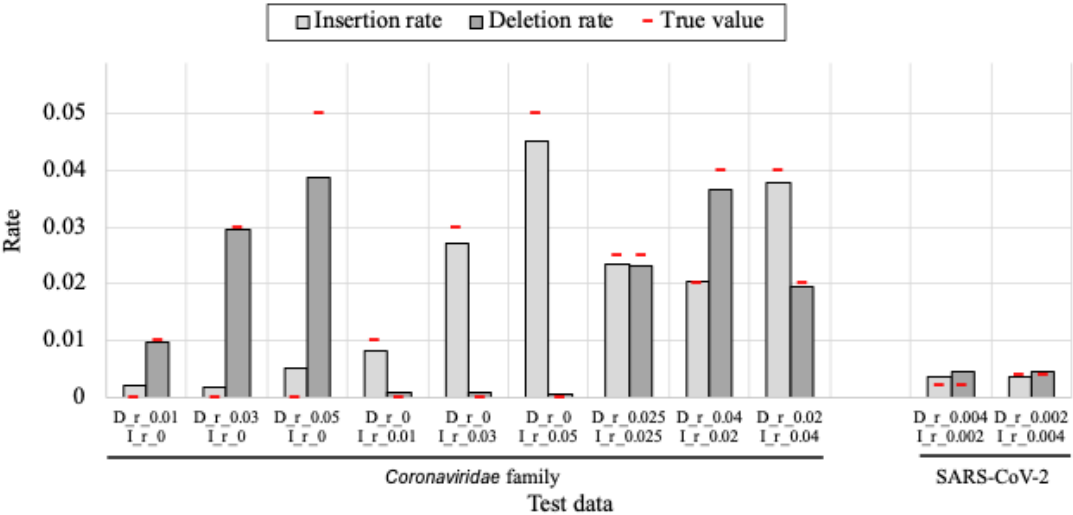


Figure S3. Evolutionary history of indel events in the *Coronaviridae* family. The figure shows all inferred indels along the genome (left panel) and indels affecting entire genes (right panel) along the *Coronaviridae* phylogeny. Indels are indicated by their genomic position and are color-coded according to the affected gene (legend). Emerging indels that are not shared with the closest common ancestor are highlighted in dark colors, while indels present in the common ancestor are shown in light colors.



Age: Young, Old
Gender: Male, Female
Marital Status: Single, Married
Income: Low, High
Occupation: Student, Teacher, Doctor, Lawyer, Other

References cited in the supplementary material

- Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* 2009;26:1879–1888. 10.1093/molbev/msp098.
- Loewenthal G, Rapoport D, Avram O, et al. A Probabilistic Model for Indel Evolution: Differentiating Insertions from Deletions. *Mol Biol Evol* 2021;38:5769–5781. 10.1093/molbev/msab266.
- Martin DP, Varsani A, Roumagnac P, et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* 2021;7:veaa087. 10.1093/ve/veaa087.