Supplementary Material

S1 Supplementary Text

Selection of *Sarbecovirus* strains

Our choice of the 54 *Sarbecovirus* strains was driven by two considerations. First, we wanted to sample broadly from available *Sarbecovirus* whole genomes in order to adequately capture strain diversity, and second, we wanted to include strains with potential relevance for SARS-CoV-2 evolution. Accordingly, we started with the collection of 44 broadly sampled *Sarbecovirus* strains (one SARS-CoV-2 strain, one SARS-CoV strain, and 42 strains from bats) used in Jungreis et al. (2021), and augmented that collection with strains hosted in civet cats and pangolins due to their proposed role as zoonotic origins for the SARS (2003) (Guan et al., 2003) and SARS-CoV-2 (2019) (Zhang et al., 2020) pandemics. A complete list of the chosen strains is available in Supplementary Table S1. For each strain, the complete genome sequence was obtained from the NCBI sequence database (NCBI Resource Coordinators, 2018).

Throughout this study, we reference results of Boni et al. (2020) and Makarenkov et al. (2021). Boni et al. include 19 strains in their analysis that we leave out, including 279_2005, JL2012, JTMC15, SX2013, Rs4874, RsSHC014, Rs3367, Longquan_140, HKU3-[2-6,8-11,13], and Pangolin-CoV. As these strains have a close relative HKU-3-[1,7,12] that is included in our analysis, and the strains do not represent new hosts, their exclusion should not alter results substantially. We include two additional strains, 16BO133 and 273_2005. Makarenkov et al. includes 6 strains that we leave out, including Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019, HKU3-6, and three SARS-CoV-2 strains (Australia VIC231 2020, USA UT 00346 2020, Hu Italy TE4836 2020). As mentioned in Makarenkov et al., the SARS-CoV-2 strains are very similar and therefore do not add further information to the analysis.

Strain tree reconstruction using BEAST

We estimated a dated strain tree for each of the three aligned regions/whole genome using BEAST v.1.10.4 (Suchard et al., 2018). Following Boni et al. (2020), we used a GTR+Γ substitution model and an uncorrelated relaxed clock model with a log-normal distribution. We used a normal distribution with mean 0.00078 and standard deviation 0.0003 as an informative rate prior, based on estimated rates for MERS-CoV (Boni et al., 2020), and ran BEAST until chains were sufficiently mixed, generally for more than 10 million iterations, with effective sample sizes greater than 100 for branch lengths and root ages. We rooted each strain tree using the outgroup containing the strains from Bulgaria 2008 [BM48-31] and Kenya 2007 [BtKY72], which were identified in Boni et al. (2020) as the most evolutionarily distant strains. Given the rooted strain trees, we again ran BEAST until chains were sufficiently mixed, using topology-preserving operations only to estimate divergence times for each ancestral strain.

Construction of gene families

For each annotated gene, we extracted and used the longest protein sequence for that gene. For strains that did not have all 11 genes annotated, we aligned their full genome to the longest annotated gene sequence among other strains and extracted the overlapping alignment. Almost all gene families, with the exception of *ORF1ab*, *spike*, and *nucleocapsid*, were unannotated in at least one strain. Using this approach, we were able to confidently identify the missing genes for an additional 5 gene families, resulting in a total of 8 complete gene families (with one gene from each of these gene families present in each strain). Two of the gene families, *ORF10* and *ORF6*, could not be detected in strains KJ473815.1 and KJ473816.1. Finally, *ORF8*, which was initially annotated in only 31 of the 54 strains, could not be detected in 10 strains.

Orthogonal verification of *spike* HGTs using Simplot

We assessed the accuracy of additional recombination events inferred through virDTL using SimPlot. Specifically, we used SimPlot to analyse the donor, recipient, and recipient-sister strains to orthogonally verify each of the five other highly supported HGTs identified by virDTL in the *spike* gene. The *spike* gene is a good candidate for such a SimPlot analysis since it is sufficiently long for recombinations to be easily visible and interpretable. We find a clear signal for recombination in the *spike* gene for each of the five cases (Figure S3). In three of the cases (F46 to Rf4092, Rs4081 to YN2018D, and Anlong-103 to YN2013), the recombination affects predominantly the *spike* gene region. In the other two cases (Jiyuan_84 to HeB2013 and Rs9401 to Rs7327), the donor sequence is more similar to the recipient sequence along the majority of the genome. Possible explanations include that the *spike* gene HGT may be part of a larger recombination event, that the *spike* gene HGT may be an artifact of incorrect placement of the affected strains in the strain tree, or that the recipient sister sequence has undergone rapid evolution. Analysis of the sequences and dated species tree suggest the latter is more likely for the transfer from Jiyuan_84 to HeB2013, while the transfer from Rs9401 to Rs7327 is more likely due to a multi-gene recombination.

S2 Supplementary Figures and Tables

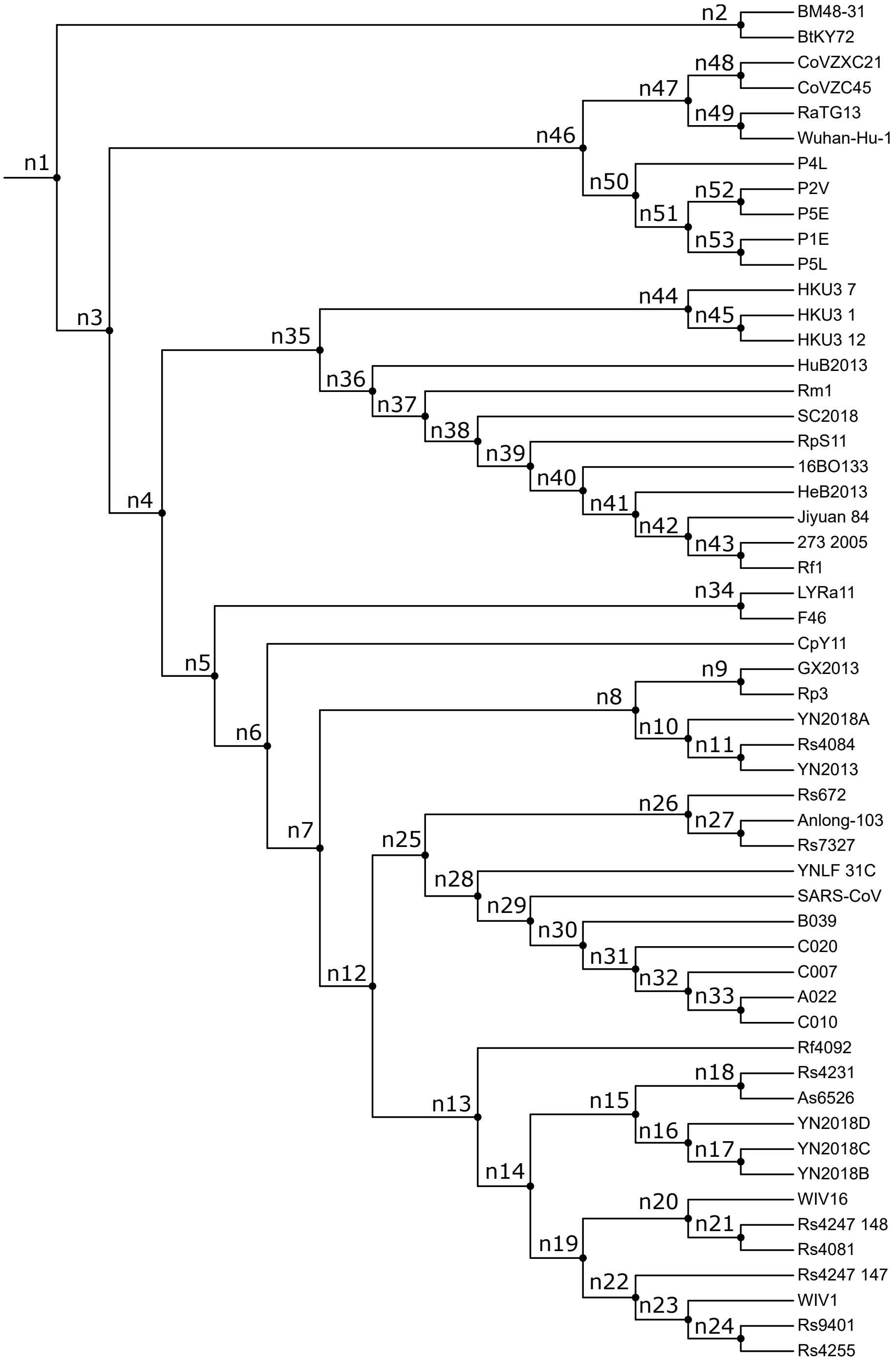


Figure S1: **Full strain tree (NRR-B) and internal node labels.**

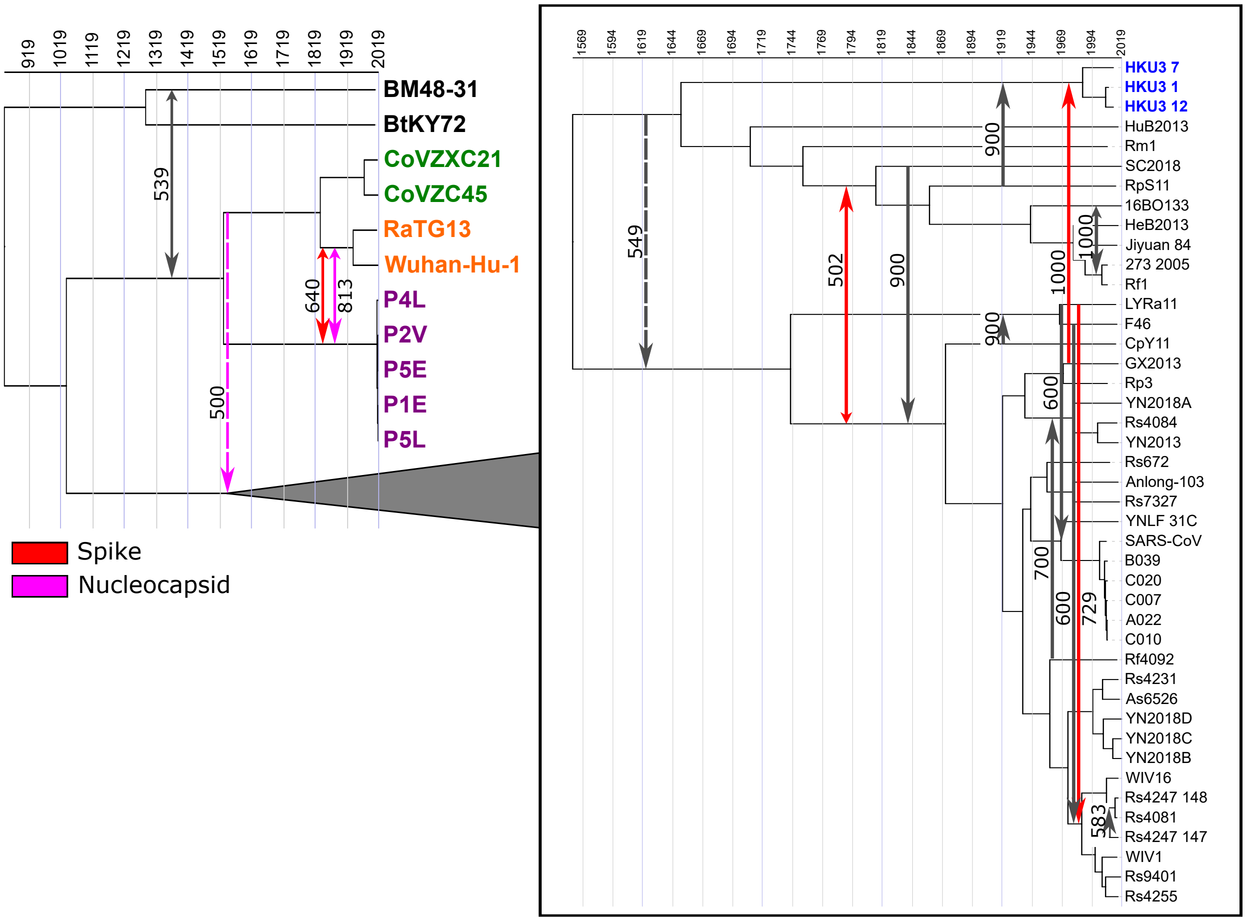
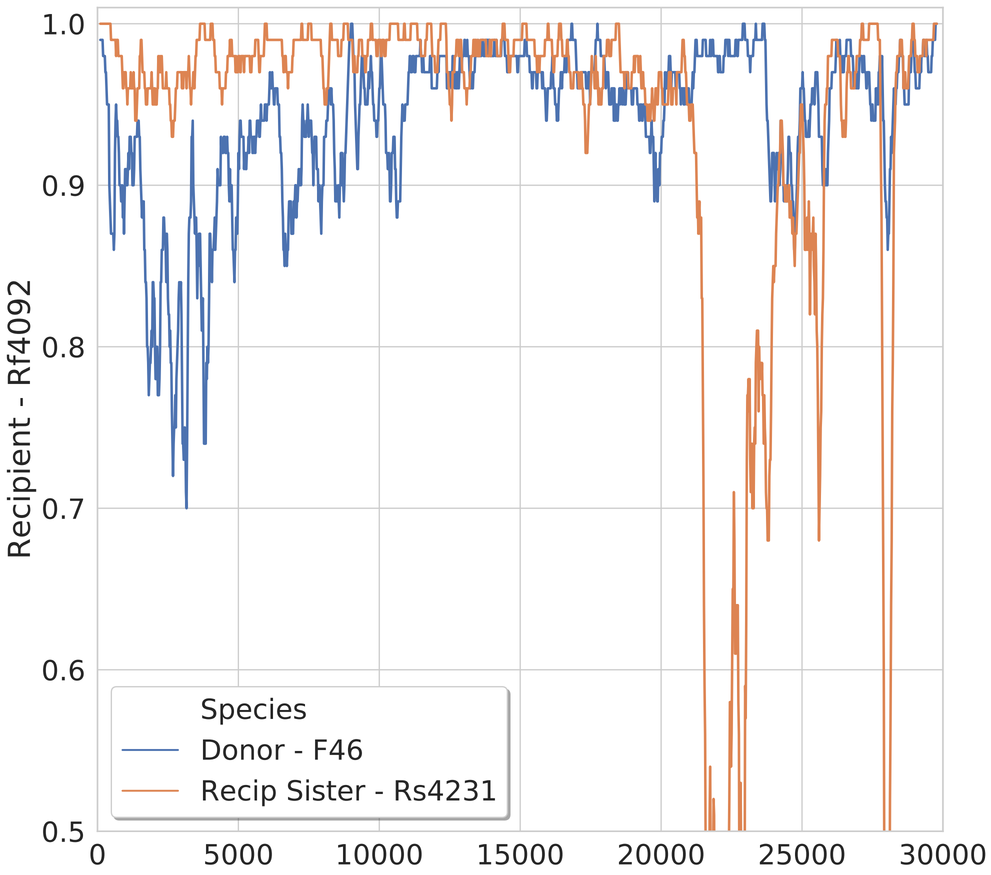
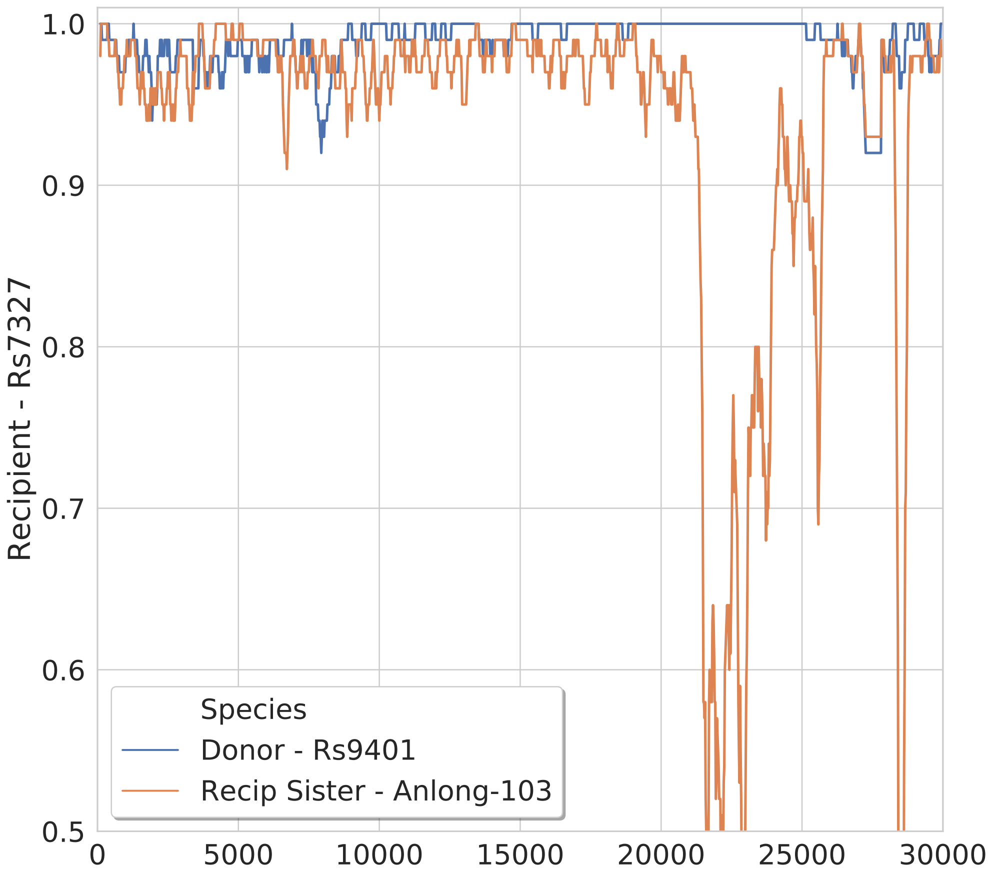


Figure S2: **Highly-supported time-consistent HGTs in the *Sarbecovirus* subgenus.** Time-consistent HGTs with an ancestral recipient and greater than 500 support are shown on a dated strain tree. Support values are shown for OptRoot-rooted gene trees, with transfers in the *spike* (red) and *nucleocapsid* genes (pink) highlighted. Smaller arrows indicate there also exists an HGT with at least 100 support in the reverse direction, suggesting directional uncertainty. All HGTs shown are also supported using MAD-rooted gene trees except one transfer in the *nucleocapsid* and one in the *membrane* (dashed lines).



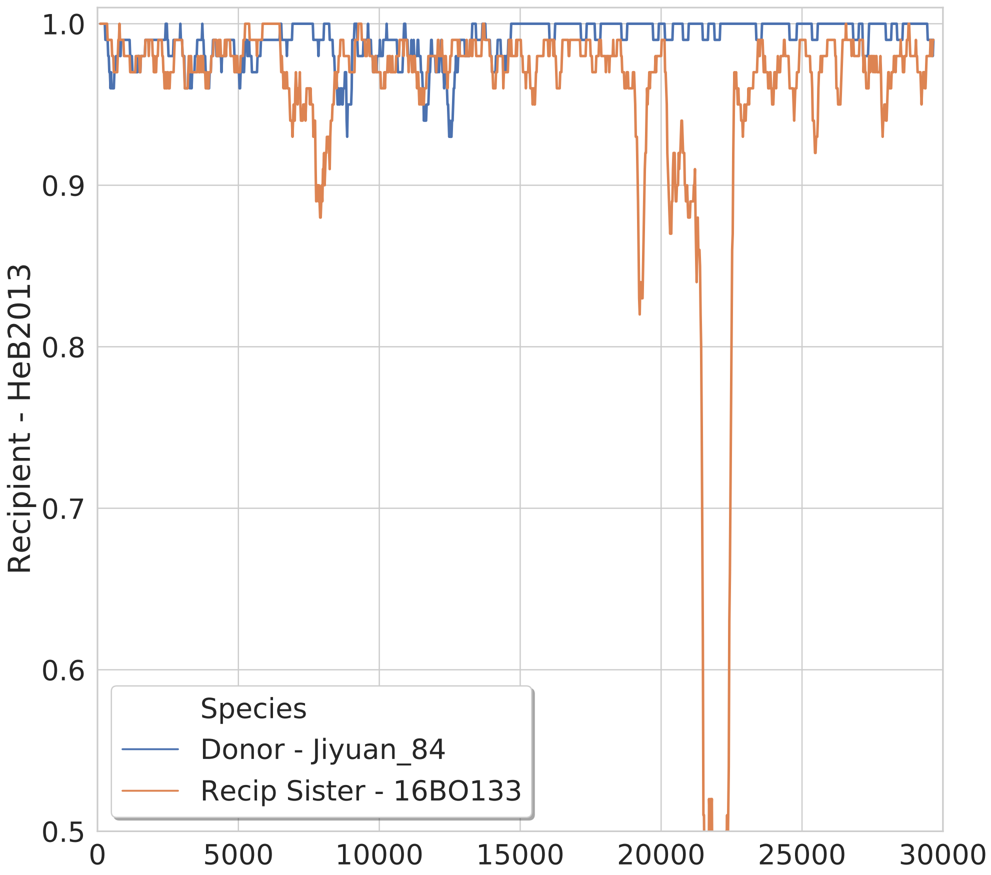
(a) F46 to Rf4092 (894 support)

Figure S3: **SimPlot Validation of leaf-to-leaf HGTs.** Part (a): SimPlot for highly supported leaf-to-leaf HGT in the *spike* gene family from F46 to Rf4092. Figure continued on next page.



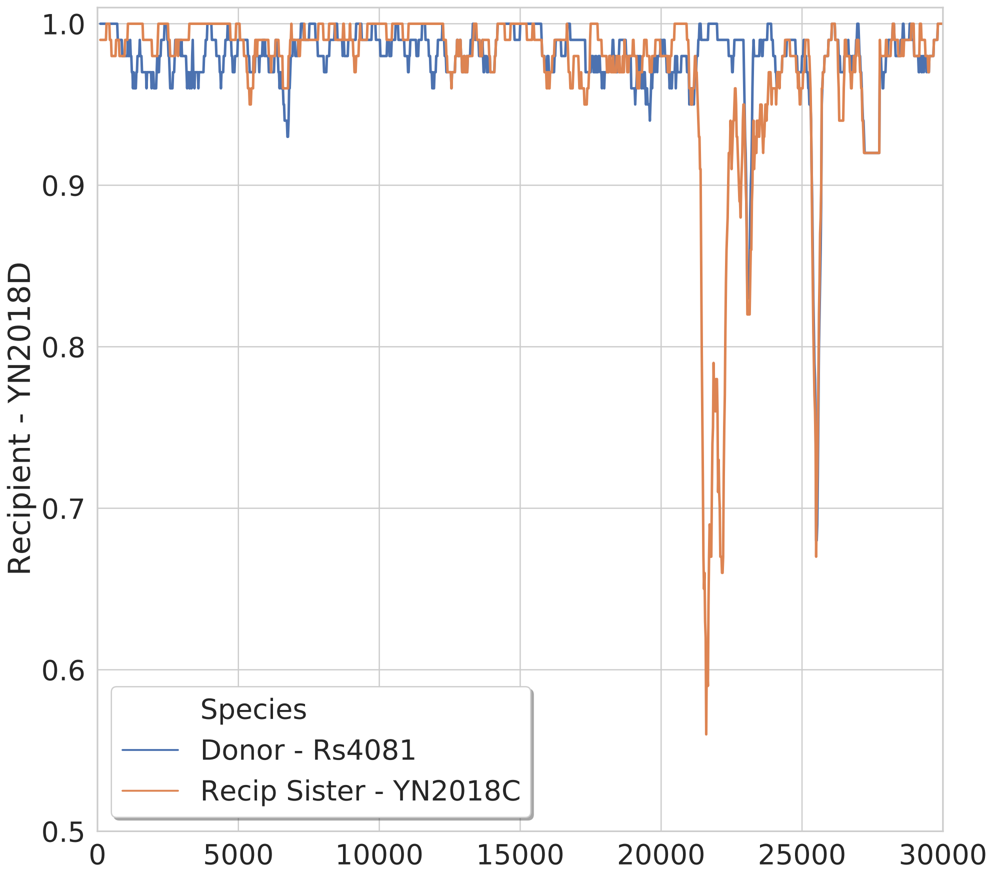
(b) Rs9401 to Rs7327 (777 support)

Figure S3: **SimPlot Validation of leaf-to-leaf HGTs.** Part (b): SimPlot for highly supported leaf-to-leaf HGT in the *spike* gene family from Rs9401 to Rs7327. Figure continued on next page.



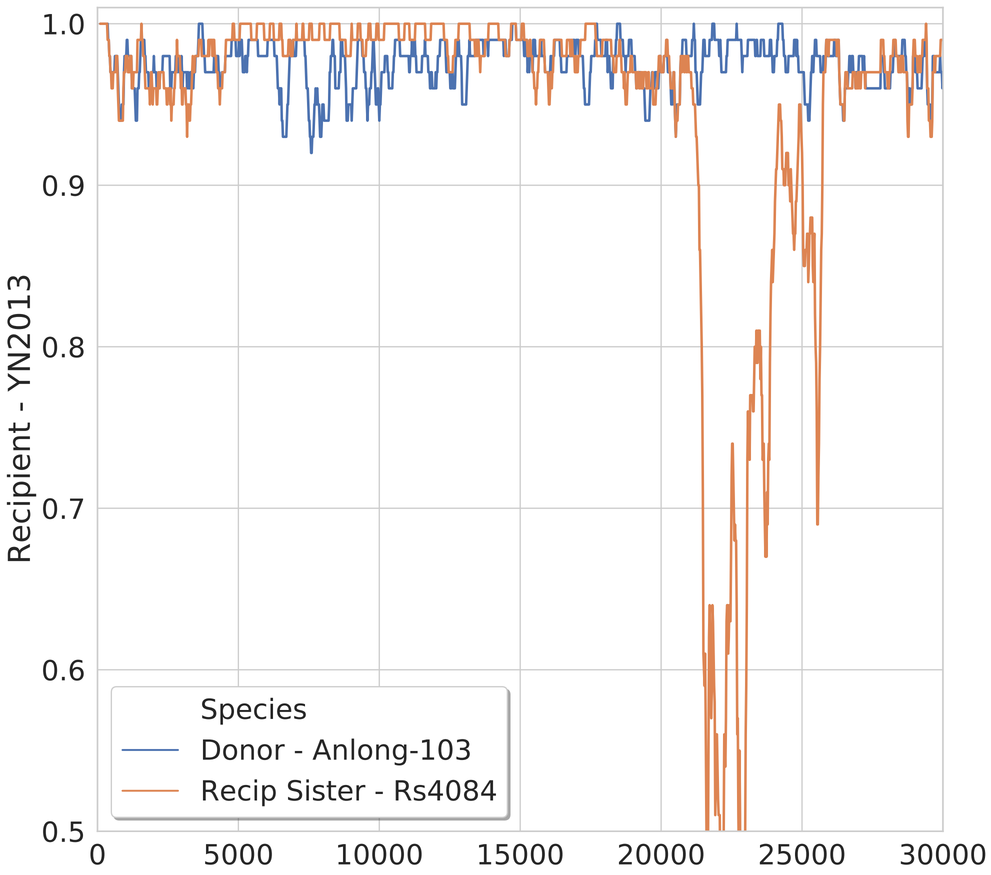
(c) Jiyuan 84 to HeB2013 (535 support)

Figure S3: **SimPlot Validation of leaf-to-leaf HGTs.** Part (c): SimPlot for highly supported leaf-to-leaf HGT in the *spike* gene family from Jiyuan_84 to HeB2013. Figure continued on next page.



(d) Rs4081 to YN2018D (511 support)

Figure S3: **SimPlot Validation of leaf-to-leaf HGTs.** Part (d): SimPlot for highly supported leaf-to-leaf HGT in the *spike* gene family from Rs4081 to YN2018D. Figure continued on next page.



(e) Anlong-103 to YN2013 (503 support)

Figure S3: **SimPlot Validation of leaf-to-leaf HGTs.** Part (e): SimPlot for highly supported leaf-to-leaf HGT in the *spike* gene family from Anlong-103 to YN2013.

Table S1: **List of strains included in analysis.**

Short ID	**NCBI Accession**	**Host Species**	**Collection Region**	**Collection Year**
Wuhan-Hu-1	NC045512	Human	Hubei	2019
SARS-CoV	NC004718	Human	Toronto	2003
RaTG13	MN996532	Bat	Yunnan	2013
CoVZC45	MG772933	Bat	Zhejiang	2017
CoVZXC21	MG772934	Bat	Zhejiang	2015
WIV16	KT444582	Bat	Yunnan	2013
Rs4231	KY417146	Bat	Yunnan	2013
YN2018B	MK211376	Bat	Yunnan	2016
Rs7327	KY417151	Bat	Yunnan	2014
Rs9401	KY417152	Bat	Yunnan	2015
Rs4084	KY417144	Bat	Yunnan	2012
WIV1	KF367457	Bat	Yunnan	2012
F46	KU973692	Bat	Yunnan	2012
Rf4092	KY417145	Bat	Yunnan	2012
YN2013	KJ473816	Bat	Yunnan	2013
Anlong-103	KY770858	Bat	Guizhou	2013
Rs4081	KY417143	Bat	Yunnan	2012
Rs4255	KY417149	Bat	Yunnan	2013
YN2018D	MK211378	Bat	Yunnan	2016
Rs672	FJ588686	Bat	Guizhou	2006
YN2018C	MK211377	Bat	Yunnan	2016
As6526	KY417142	Bat	Yunnan	2014
Rs4247_147	KY417147	Bat	Yunnan	2013
Rs4247_148	KY417148	Bat	Yunnan	2013
YN2018A	MK211375	Bat	Yunnan	2016
Rp3	DQ071615	Bat	Guangxi	2004
YNLF_31C	KP886808	Bat	Yunnan	2013
GX2013	KJ473815	Bat	Guangxi	2013
LYRa11	KF569996	Bat	Yunnan	2011
CpY11	JX993988	Bat	Yunnan	2011
SC2018	MK211374	Bat	Sichuan	2016
HuB2013	KJ473814	Bat	Hubei	2013
Rm1	DQ412043	Bat	Hubei	2004
16BO133	KY938558	Bat	South Korea	2016
Rf1	DQ412042	Bat	Hubei	2004
273_2005	DQ648856	Bat	Hubei	2004
HeB2013	KJ473812	Bat	Hebei	2013
Jiyuan_84	KY770860	Bat	Henan	2012
RpS11	JX993987	Bat	Shaanxi	2011
HKU3_7	GQ153542	Bat	Hong Kong	2009
HKU3_1	DQ022305	Bat	Hong Kong	2005
HKU3_12	GQ153547	Bat	Hong Kong	2009
BtKY72	KY352407	Bat	Kenya	2007
BM48-31	NC014470	Bat	Bulgaria	2008
P2V	MT072864	Pangolin	Guangxi	2018
P5E	MT040336	Pangolin	Guangxi	2017
P5L	MT040335	Pangolin	Guangxi	2017
P1E	MT040334	Pangolin	Guangxi	2017
P4L	MT040333	Pangolin	Guangxi	2017
C007	AY572034	Palm Civet	Guangdong	2004
A022	AY686863	Palm Civet	Guangdong	2004
C020	AY572038	Palm Civet	Guangdong	2004
C010	AY572035	Palm Civet	Guangdong	2004
B039	AY686864	Palm Civet	Guangdong	2004

Table S2: **Three candidate strain trees are highly divergent.** (*Top*) The relatively high Robinson-Foulds (RF) and subtree prune and regraft (SPR) distances between pairs of strain trees indicates substantial differences between tree topologies, and suggest that substantial recombination has occurred throughout the *Sarbecovirus* subgenus. This result motivates the need for constructing a reliable strain tree using a non-recombinant region. (*Bottom*) We report the average normalized RF distance and SPR distance between trees constructed within a genomic region. We divided the genome into 5000-base pair regions and divided each region into 1000-base pair windows with a 500-base pair offset, reconstructed a phylogeny on each such window using RAxML, and computed all pairwise RF and SPR distances between the windows within each region. We show the average internal pairwise RF and SPR distances for each 5000-base pair region and compare to the average internal pairwise RF and SPR distances for the two putative non-recombinant regions. **NRR-B is more internally consistent than other genomic regions**, which suggests less recombination in this region and motivates its use as our strain tree for reconciliation. SPR distances were estimated using the treedist package (https://rdrr.io/cran/phangorn/man/treedist.html).

Genome Region	Normalized RF Distance	SPR Distance
Whole Genome vs. NRR-B	0.653	18
Whole Genome vs. NRR-A	0.615	14
NRR-A vs. NRR-B	0.788	19
0 - 5000	0.521	12.57
4000 - 9000 (NRR-B)	**0.487**	**11.98**
5000 - 10000	0.522	12.79
10000 - 15000	0.567	13.60
13000 - 18000 (NRR-A)	0.595	13.82
15000 - 20000	0.580	13.56
20000 - 25000	0.550	13.08
25000 - 30000	0.565	13.29

Table S3: **All HGTs with** ≥ 100 **support found by our analysis.** See separate spreadsheet.

Table S4: **Number of HGTs per gene family.** For each gene family, we show the alignment length and the number of HGT events found at varying levels of support: at least 100 (61.6 percentile), at least 500 (94.9 percentile), and at least 808 (98.4 percentile).

Gene Family	Alignment		HGTs	
	Length (bp)	$\geq$ 100 support	$\geq$ 500 support	$\geq$ 808 support
ORF1ab	21,465	42	11	3
spike	3,896	54	13	4
ORF3a	836	47	10	6
envelope	231	51	0	0
membrane	687	63	11	3
ORF6	186	50	6	2
ORF7a	376	65	12	4
ORF7b	135	48	0	0
ORF8	389	68	4	1
nucleocapsid	1,271	74	9	2
ORF10	117	26	2	0
Total		588	78	25

Table S5: **Inferred HGTs in adjacent gene families likely recombined in a single event.** We randomly sampled 500,000 strain pairs from our data and randomly permuted the gene family ordering to estimate the probability π that a pair of strains with at least t supported HGTs has at least one window of size w with t HGTs in it by random association. We then performed a one-sided binomial test with probability of success π, k strain pairs in our data that fit the (w, t) window condition, and n strain pairs that have at least t supported HGTs. Bold text indicates significance at $\alpha = 0.007$, after Bonferroni correction for 7 hypotheses tested. We find that when HGTs are inferred between the same pair of strains for two adjacent gene families, they were likely transferred together, but fail to make similar claims for larger numbers of grouped genes.

w	t	π	n	k	p
2	2	0.2906	39	85	**0.0007**
3	2	0.4810	44	85	0.2848
3	3	0.1115	5	23	0.1053
4	3	0.2582	9	23	0.1136
4	4	0.0507	2	6	0.0336
5	4	0.1637	3	6	0.0595
5	5	0.0258	1	2	0.0509

Table S6: **Nodes with top 5% of HGTs in subtree rooted at node (normalized for size of subtree).**

Node	Node In	Node Out	Tree In (norm)	Tree Out (norm)
Rs4084	14	26	14.00	26.00
n26	9	10	16.33	17.67
n18	4	10	16.00	17.50
Rs7327	15	17	15.00	17.00
n34	8	3	12.50	19.00
SC2018	20	11	20.00	11.00
n24	5	10	17.00	14.00

References

M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, T. A. Castoe, A. Rambaut, and D. L. Robertson. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nature Microbiology*, 5(11):1408–1417, Nov. 2020. ISSN 2058-5276. URL `https://doi.org/10.1038/s41564-020-0771-4`.

Y. Guan, B. Zheng, Y. He, X. Liu, Z. Zhuang, C. Cheung, S. Luo, P. Li, L. Zhang, Y. Guan, et al. Isolation and characterization of viruses related to the sars coronavirus from animals in southern china. *Science*, 302(5643):276–278, 2003.

I. Jungreis, R. Sealfon, and M. Kellis. Sars-cov-2 gene content and covid-19 mutation impact by comparing 44 sarbecovirus genomes. *Nature Communications*, 12(2642), 2021. doi: https://doi.org/10.1038/s41467-021-22905-7.

V. Makarenkov, B. Mazoure, G. Rabusseau, and P. Legendre. Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin. *BMC Ecol Evo*, 21(5), 2021. URL `https://doi.org/10.1186/s12862-020-01732-2`.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1):D8–D13, Jan. 2018. doi: 10.1093/nar/gkx1095.

M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(29942656): vey016, 2018. ISSN 2057-1577. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007674/`.

T. Zhang, Q. Wu, and Z. Zhang. Probable pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Current Biology*, 30(7):1346–1351, 2020.