

# 1 Recombination patterns in coronaviruses

2 Nicola F. Müller<sup>a,1</sup>, Kathryn E. Kistler<sup>a,b</sup>, Trevor Bedford<sup>a,b</sup>,

3 <sup>a</sup>Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, United States

4 <sup>b</sup>Molecular and Cellular Biology Program, University of Washington, Seattle, United States

5

6 <sup>1</sup>Corresponding author

7  
8 **Contact:** nicola.felix.mueller@gmail.com

9 **Abstract:** As shown during the SARS-CoV-2 pandemic, phylogenetic and phylodynamic methods are essential  
10 tools to study the spread and evolution of pathogens. One of the central assumptions of these methods is that  
11 the shared history of pathogens isolated from different hosts can be described by a branching phylogenetic  
12 tree. Recombination breaks this assumption. This makes it problematic to apply phylogenetic methods to  
13 study recombining pathogens, including, for example, coronaviruses. Here, we introduce a Markov chain Monte  
14 Carlo approach that allows inference of recombination networks from genetic sequence data under a template  
15 switching model of recombination. Using this method, we first show that recombination is extremely common in  
16 the evolutionary history of SARS-like coronaviruses. We then show how recombination rates across the genome  
17 of the human seasonal coronaviruses 229E, OC43 and NL63 vary with rates of adaptation. This suggests that  
18 recombination could be beneficial to fitness of human seasonal coronaviruses. Additionally, this work sets the  
19 stage for Bayesian phylogenetic tracking of the spread and evolution of SARS-CoV-2 in the future, even as  
20 recombinant viruses become prevalent.

## 21 Main

22 Since its emergence, genetic sequence data has been applied to study the evolution and spread of SARS-CoV-  
23 2. Genetic sequences have, for example, been used to determine the origins SARS-CoV-2 (Andersen *et al.*,  
24 2020), when SARS-CoV-2 was introduced into the US (Bedford *et al.*, 2020) as well as to investigate whether  
25 genetic variants differ in viral fitness (Volz *et al.*, 2021). These analyses often rely on phylogenetic and phyl-  
26 dynamic approaches, at the heart of which are phylogenetic trees. Such trees denote how viruses isolated from  
27 different individuals are related and contain information about the transmission dynamics connecting these  
28 infections (Grenfell *et al.*, 2004).

29 Alongside mutations introduced by errors during replication, different recombination processes contribute  
30 to genetic diversity in RNA viruses (reviewed by Simon-Loriere and Holmes, 2011). Reassortment in segmented  
31 viruses (generally negative-sense RNA viruses), such as influenza or rotaviruses, can produce offspring that carry  
32 segments from different parent lineages (McDonald *et al.*, 2016). In other RNA viruses (generally positive-sense  
33 RNA viruses), such as flaviviruses and coronaviruses, homologous recombination can combine different parts  
34 of a genome from different parent lineages in absence of physically separate segments on the genome of those  
35 viruses (Su *et al.*, 2016). The main mechanism of this process is thought to be via template switching (Lai,  
36 1992), where the template for replication is switched during the replication process. Recombination breakpoints  
37 in experiments appear to be largely random, with selection selecting recombination breakpoints in some areas  
38 of the genome (Banner and Mc Lai, 1991). Recent work shows that recombination breakpoints occur more  
39 frequently in the spike region of betacoronaviruses, such as SARS-CoV-2 (Bobay *et al.*, 2020).

40 Recombination poses a unique challenge phylogenetic methods, as it violates the very central assumption  
41 that the evolutionary history of individuals can be denoted by branching phylogenetic trees. Recombination  
42 breaks this assumption and requires representation of the shared ancestry of a set of sequences as a network.

43 Not accounting for this can lead to biased phylogenetic and phylodynamic inferences (Posada and Crandall,  
44 2002; Müller et al., 2020).

45 An analytic description of recombination is provided by the coalescent with recombination, which models  
46 a backwards in time process where lineages can coalesce and recombine (Hudson, 1983). Backwards in time,  
47 recombination of a single lineage results in two lineages, with one parent lineage carrying the genetic material  
48 of one side of a random recombination breakpoint and the other parent lineage carrying the genetic material of  
49 the side of this breakpoint. This equates to the backwards in time equivalent of template switching where there  
50 is one recombination breakpoint per recombination event.

51 Currently, some Bayesian phylogenetic approaches exist that infer recombination networks, or ancestral  
52 recombination graphs (ARG), but are either approximate or do not directly allow for efficient model-based  
53 inference. Some approaches consider tree-based networks (Didelot et al., 2010; Vaughan et al., 2017), where the  
54 networks consist of a base tree where recombination edges always attach to edges on the base tree. Alternative  
55 approaches rely on approximations to the coalescent with recombination (Rasmussen et al., 2014; McVean and  
56 Cardin, 2005), consider a different model of recombination (Müller et al., 2020), or seek to infer recombination  
57 networks absent an explicit recombination model (Bloomquist and Suchard, 2010). There is, however, a gap for  
58 Bayesian inference of recombination networks under the coalescent with recombination that can be applied to  
59 study pathogens, such as coronaviruses.

60 In order to fill this gap, we here develop a Markov chain Monte Carlo (MCMC) approach to efficiently  
61 infer recombination networks under the coalescent with recombination for sequences sampled over time. This  
62 framework allows joint estimation of recombination networks, effective population sizes, recombination rates  
63 and parameters describing mutations over time from genetic sequence data sampled through time. We explicitly  
64 do not make additional approximation to characterize the recombination process, other than those of the  
65 coalescent with recombination (Hudson, 1983), such as, for example, the approximation of tree based networks.  
66 We implemented this approach as an open source software package for BEAST2 (Bouckaert et al., 2018). This  
67 allows incorporation of the various evolutionary models already implemented in BEAST2.

68 We first apply the coalescent with recombination to study the evolutionary history of SARS-like corona-  
69 viruses. Doing so, we show that the evolutionary history of SARS-like coronaviruses is extremely complex and  
70 has little resemblance to tree-like evolution. Additionally, we show that recombination only occurred between  
71 closely related SARS-like viruses in the recent history of SARS-CoV-2. Next, we reconstruct the evolutionary  
72 histories of MERS-CoV and three seasonal human coronaviruses to show that recombination also frequently  
73 occurs in human coronaviruses at rates that are comparable to reassortment rates in influenza viruses. Next, we  
74 show that recombination breakpoints in human coronaviruses vary with rates of adaptation across the genomes,  
75 suggesting recombination events being positively or negatively selected based on where breakpoints occur.

## 76 Rampant recombination in SARS-like coronaviruses

77 Recombination has been implicated at the beginning of the SARS-CoV-1 outbreak (Hon et al., 2008) and has  
78 been suggested as the origin of the receptor binding domain in SARS-CoV-2 (Li et al., 2020). While this strongly  
79 suggests non-tree-like evolution, the evolutionary history of SARS-like viruses has, out of necessity, mainly been  
80 denoted using phylogenetic trees.

81 We here reconstruct the recombination history of SARS-like viruses, which includes SARS-CoV-1 and SARS-  
82 CoV-2 as well as related bat (Ge et al., 2013, 2016; Zhou et al., 2020) and pangolin (Lam et al., 2020) corona-  
83 viruses. To do so, we infer the recombination network of SARS-like viruses under the coalescent with recom-  
84 bination. We assumed that the rates of recombination and effective population sizes were constant over time  
85 and that the genomes evolved under a GTR+Γ<sub>4</sub> model. Similar to the estimate in Boni et al. (2020), we used  
86 a fixed evolutionary rate of  $5 \times 10^{-4}$  per nucleotide and year. We fixed the evolutionary rate, since the time  
87 interval of sampling between individual isolates in relatively short compared to the time scale of the evolutionary  
88 history of SARS-like viruses. This means that the sampling times themselves therefore offer little insight into  
89 the evolutionary rates and in absence of other calibration points, there is therefore little information about the

90 evolutionary rate in this dataset. This in turn, means that if the evolutionary rate we used here is inaccurate  
91 then the timings of common ancestors will also be inaccurate. Therefore, exact timings and calendar dates in  
92 this analyses should be taken as guide posts rather than formal estimates.

93 As shown in Figure 1A, the evolutionary history of SARS-like viruses is characterized by a frequent recombination  
94 events. Consequently, characterizing evolutionary history of SARS-like viruses by a single genome-wide  
95 phylogeny is bound to be inaccurate and potentially misleading. We infer the recombination rate in SARS-like  
96 viruses to be approximately  $2 \times 10^{-6}$ , which is about 200 times lower than the evolutionary rate. These recombi-  
97 nation events were not evenly distributed across the genome and instead were largely concentrated in areas  
98 outside those coding for ORF1ab (Fig. S1). Additionally, we find some evidence for elevated rates of recombi-  
99 nation on spike subunit 1 compared to subunit 2 (Fig. S1). If we assume that during the replication of the  
100 genome of coronaviruses, template shifts occur randomly on the genome (Banner and Mc Lai, 1991), differences  
101 in observed recombination rates could be explained by selection favoring recombination events 3' to ORF1ab.

102 We next investigate when different viruses last shared a common ancestor (MRCA) along the genome (see  
103 Fig. 1B and Fig. S2). RmYN02 (Zhou et al., 2020) shares the MRCA with SARS-CoV-2 on the part of the  
104 genome that codes for ORF1ab (Fig. 1B). We additionally find strong evidence for one or more recombination  
105 events in the ancestry of RmYN02 at the beginning of the spike protein (Fig. 1B). This recent recombination  
106 event is unlikely to have occurred with a recent ancestor of any of the coronaviruses included in this dataset, as  
107 the common ancestor of RmYN02 with any other virus in the dataset is approximately the same (Fig. S3A). In  
108 other words, large parts of the spike protein of RmYN02 are as related to SARS-CoV-2 as SARS-CoV-2 is to  
109 SARS-CoV-1. The common ancestor timings of P2S across the genome are equal between RaTG13 and SARS-  
110 CoV-2 (Fig. S3B). RaTG13 on the other hand is more closely related to SARS-CoV-2 than P2S (Fig. S3B)  
111 across the entire genome. This suggests that no recombination events occurred in the ancestry of SARS-CoV-1,  
112 RaTG13 and P2S with distantly related viruses.

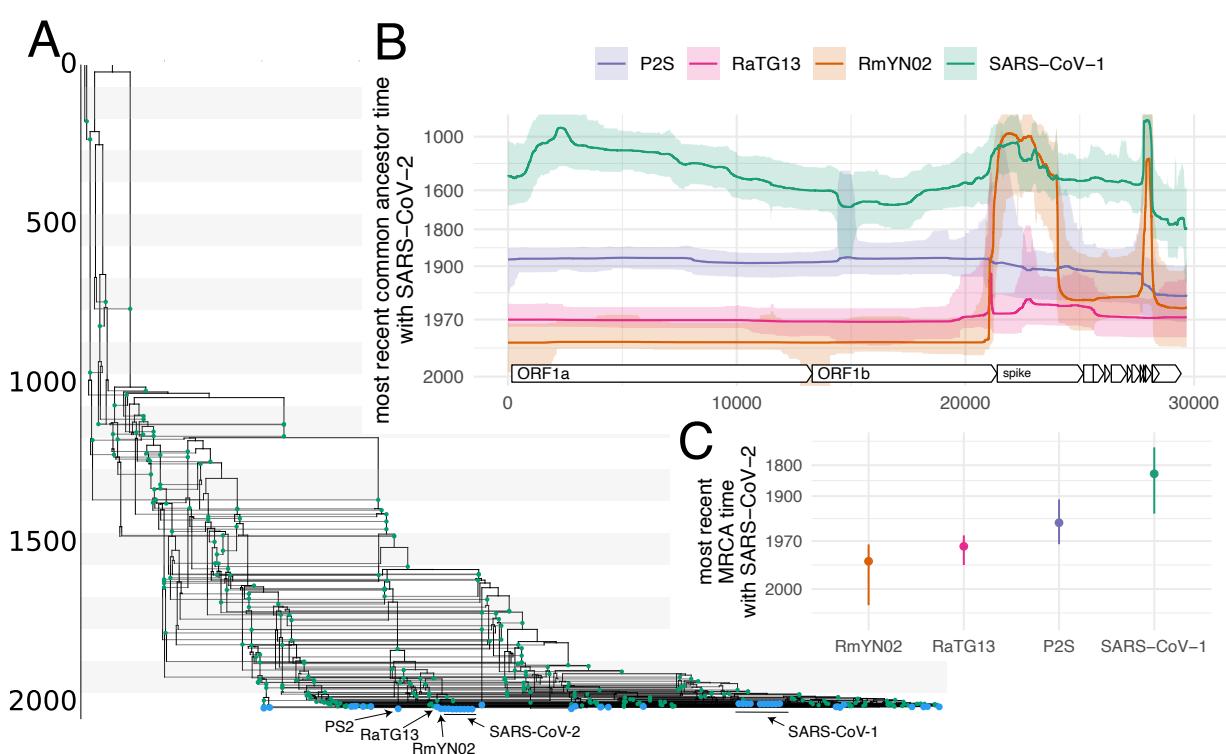
113 When looking at when different viruses last shared a common ancestor anywhere on the genome, or in other  
114 words, when looking at when the ancestral lineages of two viruses last crossed paths, we find that RmYN02 has  
115 the most recent MRCA with SARS-CoV-2 (Fig. S3C). The median estimate of the most recent MRCA with  
116 RmYN02 is 1986 (95% CI: 1973–2005), with RaTG13 to be 1975 (95% CI: 1988–1964), with P2S to be 1949  
117 (95% CI: 1907–1973) and with SARS-CoV-1 to be 1834 (95% CI: 1707–1935). These estimates are contingent  
118 on a fixed evolutionary rate of  $5 \times 10^{-4}$  per nucleotide and year.

## 119 Rates of recombination vary with rates of adaptation in human seasonal coro- 120 naviruses

121 We next investigate recombination patterns in MERS-CoV with over 2500 confirmed cases in humans as well  
122 as human seasonal coronaviruses 229E, OC43 and NL63 with widespread seasonal circulation in humans.

123 As for the SARS-like viruses, we jointly infer recombination networks, rates of recombination and population  
124 sizes for these viruses. We assumed that the genomes evolved under a GTR+Γ<sub>4</sub> model and, in contrast to the  
125 analysis of SARS-like viruses, inferred the evolutionary rates. We observe frequent recombination in the history  
126 of all 4 viruses, wherein genetic ancestry is described by network rather than a strictly branching phylogeny  
127 (Fig. 2A-D).

128 The human seasonal coronaviruses all have recombination rates around  $1 \times 10^{-5}$  per site and year (Fig. S5).  
129 This is around 10 to 20 times lower than the evolutionary rate (Fig. S6). In contrast to the recombination rates,  
130 the evolutionary rates vary greatly across the human seasonal coronaviruses, with rates between a median of  
131  $1.3 \times 10^{-4}$  (CI  $1.1 - 1.5 \times 10^{-4}$ ) for NL63 and median rate of  $2.5 \times 10^{-4}$  (CI  $2.2 - 2.7 \times 10^{-4}$ ) and  $2.1 \times 10^{-4}$   
132 (CI  $1.9 - 2.3 \times 10^{-4}$ ) for 229E and OC43 (Fig. S6). These evolutionary rates are substantially lower than those  
133 estimated for SARS-CoV-2 ( $1.1 \times 10^{-3}$  substitutions per site and year (Duchene et al., 2020)), which are more in  
134 line with our estimates for the evolutionary rates of MERS with a median rate of  $6.9 \times 10^{-4}$  (CI  $6.0 - 7.9 \times 10^{-4}$ ).  
135 Evolutionary rate estimates can be time dependent, with datasets spanning more time estimating lower rates of  
136 evolution than those spanning less time (Duchêne et al., 2014). In turn, this means that the evolutionary rates



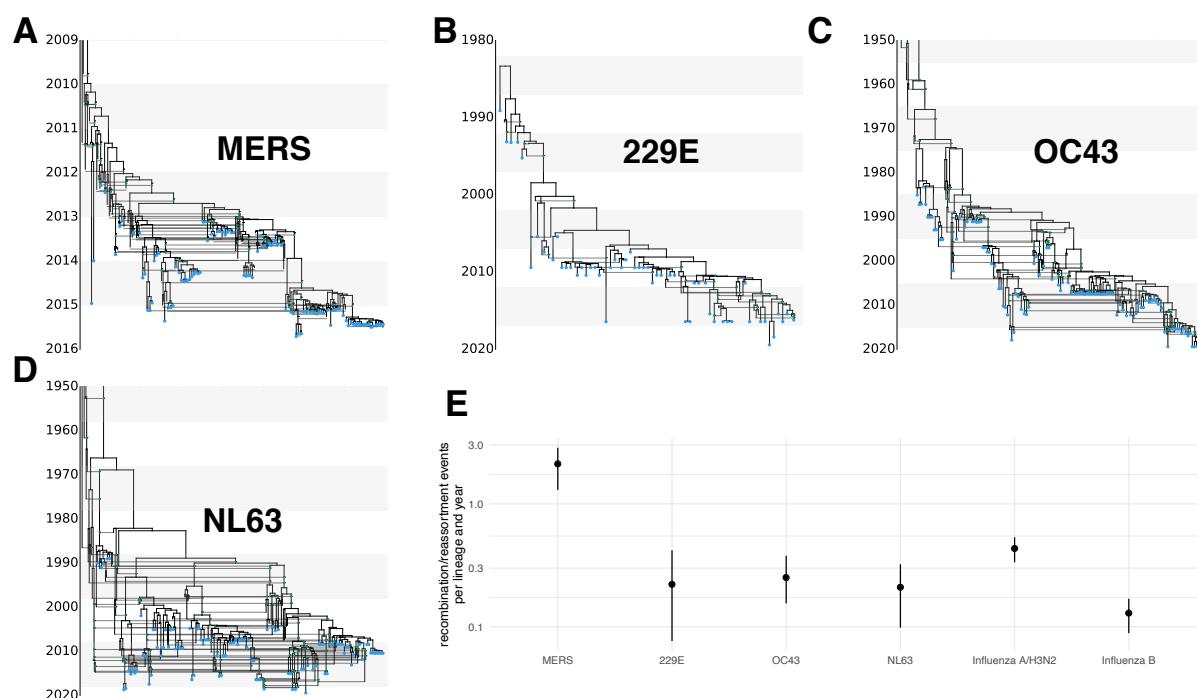
**Figure 1: Evolutionary history of SARS-like viruses.** **A** Maximum clade credibility network of SARS-like viruses. Blue dots denote samples and green dots recombination events. **B** Common ancestor times of Wuhan-Hu1 (SARS-CoV-2) with different SARS-like viruses on different positions of the genome. The y-axis denote common ancestor times in log scale. **C** Most recent time anywhere on the genome that Wuhan-Hu1 shared a common ancestor with different SARS-like viruses

estimates for SARS-CoV-2 will likely be lower the more time passes. It is unclear though, whether it is going to approximate the evolutionary rates of other seasonal coronaviruses.

On a per-lineage basis this estimated recombination rate translates into around 0.1–0.3 recombination events per lineage and year (Fig. 2E). Recombination events defined here are a product of co-infection, recombination and selection of recombinant viruses. Interestingly, the rate at which recombination events occur is highly similar to the rate at which reassortment events occur in human influenza viruses (Fig. 2D, and Müller et al. (2020)). If we assume similar selection pressures for recombinant coronaviruses compared to reassortant influenza viruses, this would indicate similar co-infection rates in influenza and coronaviruses. The incidence of coronaviruses in patients with respiratory illness cases over 12 seasons in western Scotland have been found to be lower (7% – 17%) than for influenza viruses (13%–34% but to be of the same order of magnitude (Nickbakhsh et al., 2020). Considering that seasonal coronaviruses typically are less symptomatic than influenza viruses, it is not unreasonable to assume that annual incidence and therefore likely the annual co-infection rates are comparable between influenza and coronaviruses.

Compared to human seasonal coronaviruses, recombination occurs around 3 times more often for MERS-CoV (Fig. 2E). MERS-CoV mainly circulates in camels and occasionally spills over into humans (Dudas et al., 2018). Infections of camels with MERS-CoV are highly prevalent, with close to 100% of adult camels showing antibodies against MERS-CoV (Reusken et al., 2014). Higher incidence and thus higher rates of co-infection could therefore account for higher rates of recombination in MERS-CoV compared to the human seasonal

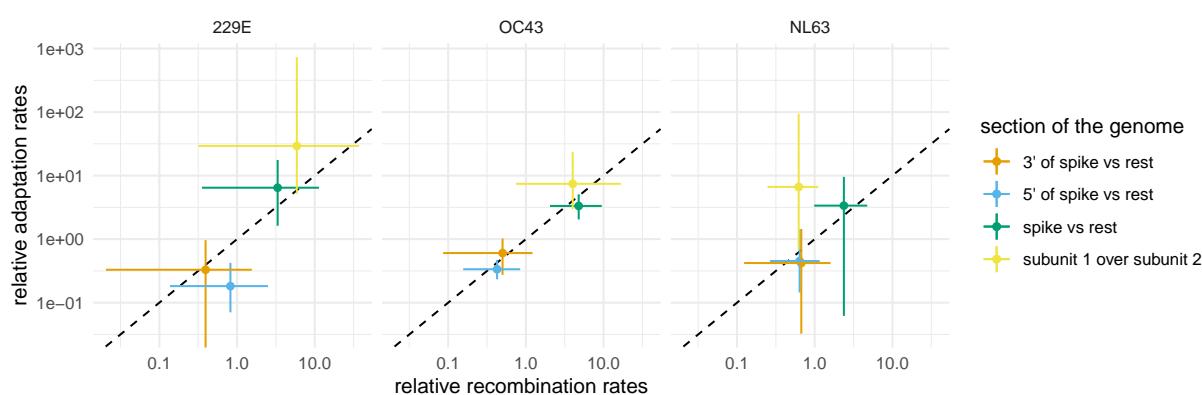
155 coronaviruses.



**Figure 2: Recombination networks and rates for coronaviruses MERS, 229E, OC43 and NL63.** Recombination networks for MERS (A) and seasonal human coronaviruses 229E (B), OC43 (C) and NL63 (D). E Recombination rates in per lineage and year for the different coronaviruses compared to reassortment rates in seasonal human influenza A/H3N2 and influenza B viruses as estimated in Müller *et al.* (2020). For OC43 and NL63, the parts of the recombination networks that stretch beyond 1950 are not shown to increase readability of more recent parts of the networks.

156 The evolutionary purpose of recombination in RNA viruses, as well as whether recombination provides  
 157 a fitness benefit is unclear (Simon-Loriere and Holmes, 2011). To investigate whether recombination benefits  
 158 fitness in human coronaviruses, we next tested whether rates of recombination differed on different parts of the  
 159 genome. To do so, we allowed everything 5' of the spike protein, i.e. mostly ORF1ab, the spike protein itself and  
 160 everything 3' of the spike protein to have a different relative rate of recombination. We computed recombination  
 161 rate ratios on each section as the recombination rate on that section divided by the mean rate on the other two  
 162 sections. We infer that recombination rates are elevated in the spike protein of all human seasonal coronaviruses  
 163 considered here (Fig. 3). This is consistent with other work estimating higher rates of recombination on the  
 164 spike protein of betacoronaviruses (Bobay *et al.*, 2020).

165 We next tested whether recombination rates are elevated on parts of the genome that also show strong signs  
 166 of adaptation. To do so, we computed the rates of adaption on different parts of the genomes of the seasonal  
 167 human coronaviruses using the approach described in (Bhatt *et al.*, 2011) and Kistler and Bedford (2021). This  
 168 approach does not explicitly consider trees to compute the rates of adaptation on different parts of the genomes  
 169 and is not affected by recombination (Kistler and Bedford, 2021). We computed adaptation rate ratios on each  
 170 section as the adaptation rate on that section divided by the mean rate on the other two sections. We find that  
 171 sections of the genome with relatively higher rates of adaptation correspond to sections of the genome with  
 172 relatively higher rates of recombination (Fig. 3). In particular, recombination and adaptation are elevated on  
 173 the section of the genome that codes for the spike protein and lower elsewhere.



**Figure 3: Comparison of recombination rates with rates of adaptation on different parts of the genomes of seasonal human coronaviruses 229E, OC43 and NL63.** Here, we show the relations of estimated relative recombination rate (x-axis) and relative adaptation rate (y-axis). The relative rates are shown for three different seasonal human coronaviruses 229E, OC43 and NL63. These estimates are shown for different parts of the genome, indicated by the different colors. These results are from two different types of analysis, one using the spike protein only (subunit 1 over subunit 2) and one using the full genome. The rate ratios denote the rate on a part of the genome divided by the average rate on the two other parts of the genome.

174 We next investigated whether these trends hold when looking at the spike protein only. The spike protein is  
 175 made up of two subunits. Subunit 1 (S1) binds to the host cell receptor, while subunit 2 (S2) facilitates fusion  
 176 of the viral and cellular membrane (Walls *et al.*, 2020). S1 contains the receptor binding domain and rates of  
 177 adaptation have been shown to be high in S1 for 229E and OC43 (Kistler and Bedford, 2021). While the rates of  
 178 adaptation are relatively low overall for NL63, there is still some evidence that they are elevated in S1 compared  
 179 to S2 (Kistler and Bedford, 2021).

180 To test whether recombination rates vary with rates of adaptation on the subunits as well, we inferred the  
 181 recombination rates from the spike protein only, allowing for different rates of recombination on S1 from the rest  
 182 of the spike protein. We find that the rates of recombination are elevated on S1 for 229E and OC43 compared  
 183 to the rest of the spike protein (Fig. 3). This is consistent with strong absolute rates of adaptation on S1 on  
 184 these two viruses. For NL63, we find weak evidence for the rate on S2 to be slightly higher than on S1 (Fig. 3),  
 185 even though the rates of adaptation are inferred to be higher on S1. The absolute rate of adaptation on the  
 186 spike protein of NL63 are, however, substantially lower than for 229E or OC43. Additionally, the uncertainty  
 187 around the estimates on adaptation rate ratios between the two subunits for NL63 are rather large and include  
 188 no difference at all. Overall, these results suggest that recombination events are either positively or negatively  
 189 selected for. Elevated rates of recombination in areas where adaptation is stronger have been described for other  
 190 organisms (reviewed here (Nachman, 2002)).

191 Recombination is generally selected for, if breaking up the linkage disequilibrium is beneficial (Barton, 1995).  
 192 Recombination can help purge deleterious mutations from the genome, such as proposed by the mutational-  
 193 deterministic hypothesis (Feldman *et al.*, 1980). Recombination can also increase the rate at which fit combi-  
 194 nation of mutations occur, such as stated by the Robertson-Hill effect (Hill and Robertson, 1966).

195 To further investigate this, we next computed the rates of recombination on fitter and less fit parts of the  
 196 recombination networks of 229E, OC43 and NL63. To do so, we first classify each edge of the inferred posterior  
 197 distribution of the recombination networks into fit and unfit based on how long a lineage survives into the future.  
 198 Fit edges are those that have descendants at least one, two, five or ten years into the future and unfit edges  
 199 those that do not. We then computed the rates of recombination on both types of edges for the entire posterior  
 200 distribution of networks. We overall do not find that fit edges show relatively higher rates of recombination (see

201 figure S7). The simplest explanation is that we do not have enough data points to measure recombination rates  
202 on unfit edges, meaning to measure recombination rates on part of the recombination network where selection  
203 had too little time to shape which lineages survive and which go extinct. An alternative explanation to why  
204 we see elevated rate or recombination in the spike protein, but do not observe a population level fitness benefit  
205 could be that most (outside of spike) recombinants could be detrimental to fitness with few (on spike) having  
206 little fitness effect at all.

## 207 Conclusion

208 Though not yet highly prevalent, evidence for recombination in SARS-CoV-2 has started to appear (VanIns-  
209 bergh et al., 2020). As such, it is crucial to know the extent to which recombination is expected to shape  
210 SARS-CoV-2 in the coming years, and to have methods to identify recombination and to perform phylogenetic  
211 reconstruction in the presence of recombination. The results shown here indicate that some recombination are  
212 either positively or negatively selected for. Estimating the deleterious load of viruses before and after recombi-  
213 nation using ancestral sequence reconstruction (Yang et al., 1995) could help shed light on which sequences are  
214 favored during recombination. Additionally, having additional sequences to reconstruct recombination patterns  
215 the seasonal coronaviruses should clarify the role recombination plays in the long term evolution of these viruses.

216 The likely rise of future SARS-CoV-2 recombinants will further necessitate methods that allow to perform  
217 phylogenetic and phylodynamic inferences in the presence of recombination (Neches et al., 2020). In absence  
218 of that, recombination has to be either ignored, leading to biased phylogenetic and phylodynamic reconstruc-  
219 tion (Posada and Crandall, 2002). Alternatively, non-recombinant parts of the genome have to be used for  
220 analyses, reducing the precision of these methods. Our approach addresses this gap by providing a Bayesian  
221 framework to infer recombination networks. To facilitate easy adaptation, we implemented the method such that  
222 setting up analyses follows the same workflow as regular BEAST2 (Bouckaert et al., 2018) analyses. Extending  
223 the current suite of population dynamic models, such as birth-death models (Stadler, 2009) or models that  
224 account for population structure (Hudson et al., 1990; Lemey et al., 2009), will further increase the applicability  
225 of recombination models to study the spread of pathogens.

## 226 Materials and Methods

### 227 Coalescent with recombination

228 The coalescent with recombination models a backwards in time coalescent and recombination process (Hudson,  
229 1983). In this process, three different events are possible: sampling, coalescence and recombination. Sampling  
230 events happen at predefined points in time. Recombination events happen at a rate proportional to the number  
231 of coexisting lineages at any point in time. Recombination events split the path of a lineage in two, with  
232 everything on one side of a recombination breakpoint going in one ancestry direction and everything on  
233 the other side of a breakpoint going in the other direction. As shown in Figure 4, the two parent lineages after  
234 a recombination event each “carry” a subset of the genome. In reality the viruses corresponding to those two  
235 lineages still “carry” the full genome, but only a part of it will have sampled descendants. In other words,  
236 only a part of the genome carried by a lineage at any time may impact the genome of a future lineage that  
237 is sampled. The probability of actually observing a recombination event on lineage  $l$  is proportional to how  
238 much genetic material that lineage carries. This can be computed as the difference between the last and first  
239 nucleotide position that is carried by  $l$ , which we denote as  $\mathcal{L}(l)$ . Coalescent events happen between co-existing  
240 lineages at a rate proportional to the number of pairs of coexisting lineages at any point in time and inversely  
241 proportional to the effective population size. The parent lineage at each coalescent event will “carry” genetic  
242 material corresponding to the union of genetic material of the two child lineages.

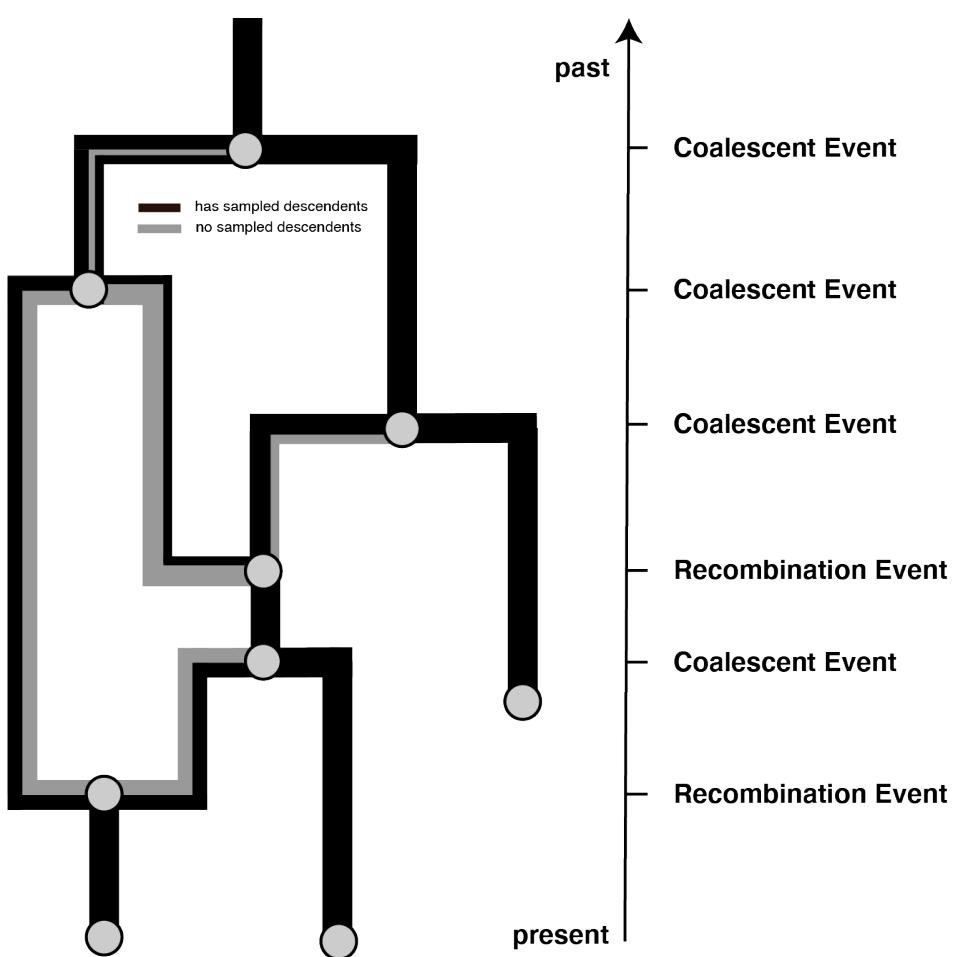


Figure 4: **Example recombination network.** Events that can occur on a recombination network as considered here. We consider events to occur from present backwards in time to the past (as is the norm when looking at coalescent processes. Lineages can be added upon sampling events, which occur at predefined points in time and are conditioned on. Recombination events split the path of a lineage in two, with everything on one side of a recombination breakpoint going in one and everything on the other side of a breakpoint going in the other direction.

## 243 Posterior probability

In order to perform joint Bayesian inference of recombination networks together with the parameters of the associated models, we use a MCMC algorithm to characterize the joint posterior density. The posterior density is denoted as:

$$P(N, \mu, \theta, \rho | D) = \frac{P(D|N, \mu)P(N|\theta, \rho)P(\mu, \theta, \rho)}{P(D)},$$

244 where  $N$  denotes the recombination network,  $\mu$  the evolutionary model,  $\theta$  the effective population size and  $\rho$   
245 the recombination rate. The multiple sequence alignment, that is the data, is denoted  $D$ .  $P(D|N, \mu)$  denotes  
246 the network likelihood,  $P(N|\theta, \rho)$ , the network prior and  $P(\mu, \theta, \rho)$  the parameter priors. As is usually done in  
247 Bayesian phylogenetics, we assume that  $P(\mu, \theta, \rho) = P(\mu)P(\theta)P(\rho)$ .

## 248 Network Likelihood

While the evolutionary history of the entire genome is a network, the evolutionary history of each individual position in the genome can be described as a tree. We can therefore denote the likelihood of observing a sequence alignment (the data denoted  $D$ ) given a network  $N$  and evolutionary model  $\mu$  as:

$$P(D|N, \mu) = \prod_{i=1}^{\text{sequence length}} P(D_i|T_i, \mu),$$

249 with  $D_i$  denoting the nucleotides at position  $i$  in the sequence alignment and  $T_i$  denoting the tree at position  $i$ .  
250 The likelihood at each individual position in the alignment can then be computed using the standard pruning  
251 algorithm (Felsenstein, 1981). We implemented the network likelihood calculation  $P(D_i|T_i, \mu)$  such that it allows  
252 making use of all the standard site models in BEAST2. Currently, we only consider strict clock models and do  
253 not allow for rate variations across different branches of the network. This is because the number of edges in  
254 the network changes over the course of the MCMC, making relaxed clock models complex to implement. We  
255 implemented the network likelihood such that it can make use of caching of intermediate results and use unique  
256 patterns in the multiple sequence alignment, similar to what is done for tree likelihood computations.

## 257 Network Prior

258 The network prior is denoted by  $P(N|\theta, \rho)$ , which is the probability of observing a network and the embedding  
259 of segment trees under the coalescent with recombination model, with effective population size  $\theta$  and per-lineage  
260 recombination rate  $\rho$ . It essentially plays the same role that tree prior plays in standard phylodynamic analyses.

We can calculate  $P(N|\theta, \rho)$  by expressing it as the product of exponential waiting times between events (i.e., recombination, coalescent, and sampling events):

$$P(N|\theta, \rho) = \prod_{i=1}^{\#events} P(\text{event}_i|L_i, \theta, \rho) \times P(\text{interval}_i|L_i, \theta, \rho),$$

261 where we define  $t_i$  to be the time of the  $i$ -th event and  $L_i$  to be the set of lineages extant immediately prior to  
262 this event. (That is,  $L_i = L_t$  for  $t \in [t_{i-1}, t_i]$ .)

Given the coalescent process is a constant size coalescent and given the  $i$ -th event is a coalescent event, the event contribution is denoted as:

$$P(\text{event}_i|L_i, \theta, \rho) = \frac{1}{\theta}.$$

If the  $i$ -th event is a recombination event and assuming constant rates of recombination over time, the event contribution is denoted as:

$$P(\text{event}_i|L_i, \theta, \rho) = \rho * \mathcal{L}(l).$$

The interval contribution denotes the probability of not observing any event in a given interval. It can be computed as the product of not observing any coalescent, nor any recombination events in interval  $i$ . We can therefore write:

$$P(\text{interval}_i|L_i, \theta, \rho) = \exp[-(\lambda^c + \lambda^r)(t_i - t_{i-1})],$$

where  $\lambda^c$  denotes the rate of coalescence and can be expressed as:

$$\lambda^c = \binom{|L_i|}{2} \frac{1}{\theta},$$

and  $\lambda^r$  denotes the rate of observing a recombination event on any co-existing lineage and can be expressed as:

$$\lambda^r = \rho \sum_{l \in L_i} \mathcal{L}(l).$$

In order to allow for recombination rates to vary across  $s$  sections  $\mathcal{S}_s$  of the genome, we modify  $\lambda^r$  to differ in each section  $\mathcal{S}_s$ , such that:

$$\lambda^r = \sum_{s \in \mathcal{S}} \rho_s \sum_{l \in L_i} \mathcal{L}(l) \cap \mathcal{S}_s,$$

with  $\mathcal{L}(l) \cap \mathcal{S}_s$  denoting the amount of overlap between  $\mathcal{L}(l)$  and  $\mathcal{S}_s$ . The recombination rate in each section  $s$  is denoted as  $\rho_s$ .

## 265 MCMC Algorithm for Recombination Networks

266 In order to explore the posterior space of recombination networks, we implemented a series of MCMC operators.  
267 These operators often have analogs in operators used to explore different phylogenetic trees and are similar to  
268 the ones used to explore reassortment networks (Müller et al., 2020). Here, we briefly summarize each of these  
269 operators.

270 **Add/remove operator.** The add/remove operator adds and removes recombination events. An extension  
271 of the subtree prune and regraft move for networks (Bordewich et al., 2017) to jointly operate on segment  
272 trees as well. We additionally implemented an adapted version to sample re-attachment under a coalescent  
273 distribution to increase acceptance probabilities.

274 **Loci diversion operator.** The loci diversion operator randomly changes the location of recombination  
275 breakpoints on a recombination event.

276 **Exchange operator.** The exchange operator changes the attachment of edges in the network while keeping  
277 the network length constant.

278 **Subnetwork slide operator.** The subnetwork slide operator changes the height of nodes in the network  
279 while allowing to change the topology.

280 **Scale operator.** The scale operator scales the heights of individual nodes or the whole network without  
281 changing the network topology.

282 **Gibbs operator.** The Gibbs operator efficiently samples any part of the network that is older than the  
283 root of any segment of the alignment and is thus not informed by any genetic data.

284 **Empty loci preoperator.** The empty segment preoperator augments the network with edges that do not  
285 carry any loci for the duration of a move, to allow larger jumps in network space.

286 One of the issues when inferring these recombination networks is that the root height can be substantially  
287 larger than when not allowing for recombination events. This can cause computational issue when performing  
288 inferences. To circumvent this, we truncate the recombination networks by reducing the recombination rate some  
289 time after all positions of the sequence alignment have reached their common ancestor height. We validate the  
290 implementation of the coalescent with recombination network prior as well as all operators in the supplement S8.  
291 We also show that truncating the recombination networks does not affect the sampling of recombination networks  
292 prior to reaching the common common ancestor height of all positions in the sequence alignment.

293 We then tested whether we are able to infer recombination networks, recombination rates, effective population  
294 sizes and evolutionary parameters from simulated data. To do so, we randomly simulated recombination networks  
295 under the coalescent with recombination. On top of these, we then simulated multiple sequence alignments.  
296 We then re-infer the parameters used to simulate using our MCMC approach. As shown in Figure S9, these  
297 parameters are retrieved well from simulated data with little bias and accurate coverage of simulated parameters  
298 by credible intervals.

299 Additionally, we compared the effective sample size values from MCMC runs inferring recombination net-  
300 works for the MERS spike protein to treating the evolutionary histories as trees. We find that although the  
301 effective sample size values are lower when inferring recombination networks, they are not orders of magnitude  
302 lower (see fig S10).

### 303 Recombination network summary

304 We implemented an algorithm to summarize distributions of recombination networks similar to the maximum  
305 clade credibility framework typically used to summarize trees in BEAST (Heled and Bouckaert, 2013). In short,  
306 the algorithm summarizes over individual trees at each position in the alignment. To do so, we first compute  
307 how often we encountered the same coalescent event at every position in the alignment during the MCMC. We  
308 then choose the network that maximizes the clade support over each position as the maximum clade credibility  
309 (MCC) network.

310 The MCC networks are logged in the extended Newick format (Cardona et al., 2008) and can be visualized  
311 in [icytree.org](http://icytree.org) (Vaughan, 2017). We here plotted the MCC networks using an adapted version of baltic  
312 (<https://github.com/evogytis/baltic>).

### 313 Software

314 The Recombination package is implemented as an addon to the Bayesian phylogenetics software platform  
315 BEAST2 (Bouckaert et al., 2018). All MCMC analyses performed here, were run using adaptive parallel  
316 tempering (Müller and Bouckaert, 2020). The source code is available at <https://github.com/nicfel/Recombination>. We additionally provide a tutorial on how to setup and postprocess an analysis at <https://github.com/nicfel/Recombination-Tutorial>. The MCC networks are plotted using an adapted version of baltic (<https://github.com/evogytis/baltic>). All other plots are done in R using ggplot2 (Wickham, 2016). The scripts to setup analyses and to plot the results in this manuscript are available from  
321 <https://github.com/nicfel/Recombination-Material>.

### 322 Sequence data

323 The genetic sequence data for OC43, NL63 and 229e were obtained from ViPR (<http://www.viprbrc.org>) as  
324 described in Kistler and Bedford (2021). All virus sequences were isolated from a human host. The sequence  
325 data for the MERS analyses were the same as described in Dudas et al. (2018), but using a randomly down  
326 sampled dataset of 100 sequences. For the SARS like analyses, we used several different deposited SARS-like  
327 genomes, mostly originating from bats, as well as humans and one pangolin derived sequence.

### 328 Rates of adaptation

329 The rates of adaptation were calculated using a modification of the McDonald-Kreitman method, as designed  
330 by Bhatt et al. (2011), and implemented in Kistler and Bedford (2021). Briefly, for each virus, we aligned the  
331 sequence of each gene or genomic region. Then, we the alignment into sliding 3-year slices, each containing a  
332 minimum of 3 sequenced isolates. We used the consensus sequence at the first time point as the outgroup. A  
333 comparison of the outgroup to the alignment of each subsequent temporal yielded a measure of synonymous  
334 and non-synonymous fixations and polymorphisms at each position in the alignment. We used proportional  
335 site-counting for these estimations (Bhatt et al., 2010). We assumed that selectively neutral sites are all silent  
336 mutations as well as replacement polymorphisms occurring at frequencies between 0.15 and 0.75 (Bhatt et al.,  
337 2011). We identified adaptive substitutions as non-synonymous fixations and high-frequency polymorphisms  
338 that exceed the neutral expectation. We then estimated the rate of adaptation (per codon per year) using linear  
339 regression of the number of adaptive substitutions inferred at each time point. In order to compute the 5' of  
340 spike and 3' of spike rates of adaptation we used the weighted average of all coding regions to the left (upstream)  
341 or right (downstream) of the spike gene, respectively, using the length of the individual sections as weights. We  
342 estimated the uncertainty by running the same analysis on 100 bootstrapped outgroups and alignments.

343 **Acknowledgments**

344 We would like to thanks Timothy G. Vaughan for helpful insights into the implementation of the software. NFM  
345 is funded by the Swiss National Science Foundation (P2Ezp3\_191891). KEK is a NSF GRFP Fellow (DGE-  
346 1762114) TB is a Pew Biomedical Scholar and is supported by NIH R35 GM119774. The Scientific Computing  
347 Infrastructure at Fred Hutch is supported by NIH ORIP S10OD028685

## 348 References

- 349 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat Med*, **26**,  
350 450–452.
- 351 Banner, L. R. and Mc Lai, M. (1991). Random nature of coronavirus rna recombination in the absence of selection pressure. *Virology*,  
352 **185**(1), 441–445.
- 353 Barton, N. (1995). A general model for the evolution of recombination. *Genetics Research*, **65**(2), 123–144.
- 354 Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie,  
355 H., et al. (2020). Cryptic transmission of sars-cov-2 in washington state. *Science*, **370**(6516), 571–575.
- 356 Bhatt, S., Katzourakis, A., and Pybus, O. G. (2010). Detecting natural selection in rna virus populations using sequence summary statistics.  
357 *Infection, Genetics and Evolution*, **10**(3), 421–430.
- 358 Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011). The genomic rate of molecular adaptation of the human influenza a virus. *Molecular  
359 biology and evolution*, **28**(9), 2443–2451.
- 360 Bloomquist, E. W. and Suchard, M. A. (2010). Unifying vertical and nonvertical evolution: a stochastic arg-based framework. *Systematic  
361 biology*, **59**(1), 27–41.
- 362 Bobay, L.-M., O'Donnell, A. C., and Ochman, H. (2020). Recombination events are concentrated in the spike protein region of betacoron-  
363 aviruses. *PLoS Genetics*, **16**(12), e1009272.
- 364 Boni, M. F., Lemey, P., Jiang, X., Lam, T. T.-Y., Perry, B. W., Castoe, T. A., Rambaut, A., and Robertson, D. L. (2020). Evolutionary  
365 origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nature Microbiology*, **5**(11), 1408–1417.
- 366 Bordewich, M., Linz, S., and Semple, C. (2017). Lost in space? generalising subtree prune and regraft to spaces of phylogenetic networks.  
367 *Journal of theoretical biology*, **423**, 1–12.
- 368 Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D.,  
369 De Maio, N., et al. (2018). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *BioRxiv*, page 474296.
- 370 Bouckaert, R. R. (2010). Densitree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**(10), 1372–1373.
- 371 Cardona, G., Rosselló, F., and Valiente, G. (2008). Extended newick: it is time for a standard representation of phylogenetic networks.  
372 *BMC bioinformatics*, **9**(1), 1–8.
- 373 Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome  
374 sequences. *Genetics*, **186**(4), 1435–1449.
- 375 Duchêne, S., Holmes, E. C., and Ho, S. Y. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in  
376 rate estimates. *Proceedings of the Royal Society B: Biological Sciences*, **281**(1786), 20140732.
- 377 Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the  
378 phylodynamic threshold of sars-cov-2. *Virus evolution*, **6**(2), veaa061.
- 379 Dudas, G., Carvalho, L. M., Rambaut, A., and Bedford, T. (2018). Mers-cov spillover at the camel-human interface. *Elife*, **7**, e31257.
- 380 Feldman, M. W., Christiansen, F. B., and Brooks, L. D. (1980). Evolution of recombination in a constant environment. *Proceedings of the  
381 National Academy of Sciences*, **77**(8), 4838–4841.
- 382 Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**(6),  
383 368–376.
- 384 Ge, X.-Y., Li, J.-L., Yang, X.-L., Chmura, A. A., Zhu, G., Epstein, J. H., Mazet, J. K., Hu, B., Zhang, W., Peng, C., et al. (2013). Isolation  
385 and characterization of a bat sars-like coronavirus that uses the ace2 receptor. *Nature*, **503**(7477), 535–538.
- 386 Ge, X.-Y., Wang, N., Zhang, W., Hu, B., Li, B., Zhang, Y.-Z., Zhou, J.-H., Luo, C.-M., Yang, X.-L., Wu, L.-J., et al. (2016). Coexistence  
387 of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virologica Sinica*, **31**(1), 31–40.
- 388 Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological  
389 and evolutionary dynamics of pathogens. *science*, **303**(5656), 327–332.
- 390 Heled, J. and Bouckaert, R. R. (2013). Looking for trees in the forest: summary tree from posterior samples. *BMC evolutionary biology*,  
391 **13**(1), 1–11.

- 392 Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, **8**(3), 269–294.
- 393 Hon, C.-C., Lam, T.-Y., Shi, Z.-L., Drummond, A. J., Yip, C.-W., Zeng, F., Lam, P.-Y., and Leung, F. C.-C. (2008). Evidence of the  
394 recombinant origin of a bat severe acute respiratory syndrome (sars)-like coronavirus and its implications on the direct ancestor of sars  
395 coronavirus. *Journal of virology*, **82**(4), 1819–1826.
- 396 Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, **23**(2), 183–201.
- 397 Hudson, R. R. et al. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, **7**(1), 44.
- 398 Kistler, K. E. and Bedford, T. (2021). Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses oc43 and  
399 229e. *Elife*, **10**, e64509.
- 400 Lai, M. (1992). Rna recombination in animal and plant viruses. *Microbiology and Molecular Biology Reviews*, **56**(1), 61–79.
- 401 Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B., Liao, Y.-S., et al.  
402 (2020). Identifying sars-cov-2-related coronaviruses in malayan pangolins. *Nature*, **583**(7815), 282–285.
- 403 Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput Biol*, **5**(9),  
404 e1000520.
- 405 Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020).  
406 Emergence of sars-cov-2 through recombination and strong purifying selection. *Science Advances*, **6**(27), eabb9153.
- 407 McDonald, S. M., Nelson, M. I., Turner, P. E., and Patton, J. T. (2016). Reassortment in segmented rna viruses: mechanisms and outcomes.  
408 *Nature Reviews Microbiology*, **14**(7), 448.
- 409 McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society  
410 B: Biological Sciences*, **360**(1459), 1387–1393.
- 411 Müller, N. F. and Bouckaert, R. R. (2020). Adaptive metropolis-coupled mcmc for beast 2. *PeerJ*, **8**, e9473.
- 412 Müller, N. F., Stoltz, U., Dudas, G., Stadler, T., and Vaughan, T. G. (2020). Bayesian inference of reassortment networks reveals fitness  
413 benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences*, **117**(29), 17104–17111.
- 414 Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Current opinion in genetics &  
415 development*, **12**(6), 657–663.
- 416 Neches, R. Y., McGee, M. D., and Kyrpides, N. C. (2020). Recombination should not be an afterthought. *Nature Reviews Microbiology*,  
417 **18**(11), 606–606.
- 418 Nickbakhsh, S., Ho, A., Marques, D. F., McMenamin, J., Gunson, R. N., and Murcia, P. R. (2020). Epidemiology of seasonal coronaviruses:  
419 establishing the context for the emergence of coronavirus disease 2019. *The Journal of infectious diseases*, **222**(1), 17–25.
- 420 Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, **6**(1), 7–11.
- 421 Posada, D. and Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution*,  
422 **54**(3), 396–402.
- 423 Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS  
424 Genet*, **10**(5), e1004342.
- 425 Reusken, C. B., Messadi, L., Feyisa, A., Ularamu, H., Godeke, G.-J., Danmarwa, A., Dawo, F., Jemli, M., Melaku, S., Shamaki, D., et al.  
426 (2014). Geographic distribution of mers coronavirus among dromedary camels, africa. *Emerging infectious diseases*, **20**(8), 1370.
- 427 Simon-Loriere, E. and Holmes, E. C. (2011). Why do rna viruses recombine? *Nature Reviews Microbiology*, **9**(8), 617–626.
- 428 Stadler, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of  
429 theoretical biology*, **261**(1), 58–66.
- 430 Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C., Zhou, J., Liu, W., Bi, Y., and Gao, G. F. (2016). Epidemiology, genetic recombination, and  
431 pathogenesis of coronaviruses. *Trends in microbiology*, **24**(6), 490–502.
- 432 VanInsberghe, D., Neish, A. S., Lowen, A. C., and Koelle, K. (2020). Identification of sars-cov-2 recombinant genomes. *BioRxiv*.
- 433 Vaughan, T. G. (2017). Icytree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, **33**(15), 2392–2394.
- 434 Vaughan, T. G., Welch, D., Drummond, A. J., Biggs, P. J., George, T., and French, N. P. (2017). Inferring ancestral recombination graphs  
435 from bacterial genomic data. *Genetics*, **205**(2), 857–870.

- 436 Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., et al.  
437 (2021). Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity. *Cell*, **184**(1), 64–75.
- 438 Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the  
439 sars-cov-2 spike glycoprotein. *Cell*, **181**(2), 281–292.
- 440 Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
- 441 Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**(4),  
442 1641–1650.
- 443 Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., et al. (2020). A novel bat coronavirus closely  
444 related to sars-cov-2 contains natural insertions at the s1/s2 cleavage site of the spike protein. *Current Biology*, **30**(11), 2196–2203.

445 **Supplementary material**



Figure S1: **Recombination rate ratios of SARS-like viruses on different parts of the genome.** Here, we show the recombination rate ratios for SARS-like viruses based on two different analyses, one using the full genome (left) and one using the spike protein only (right). The rate ratios denote the rate on a part of the genome divided by the average rate on the two other parts of the genome. s1 over s2 denotes the rate ratio on subunit 1 over subunit 2.

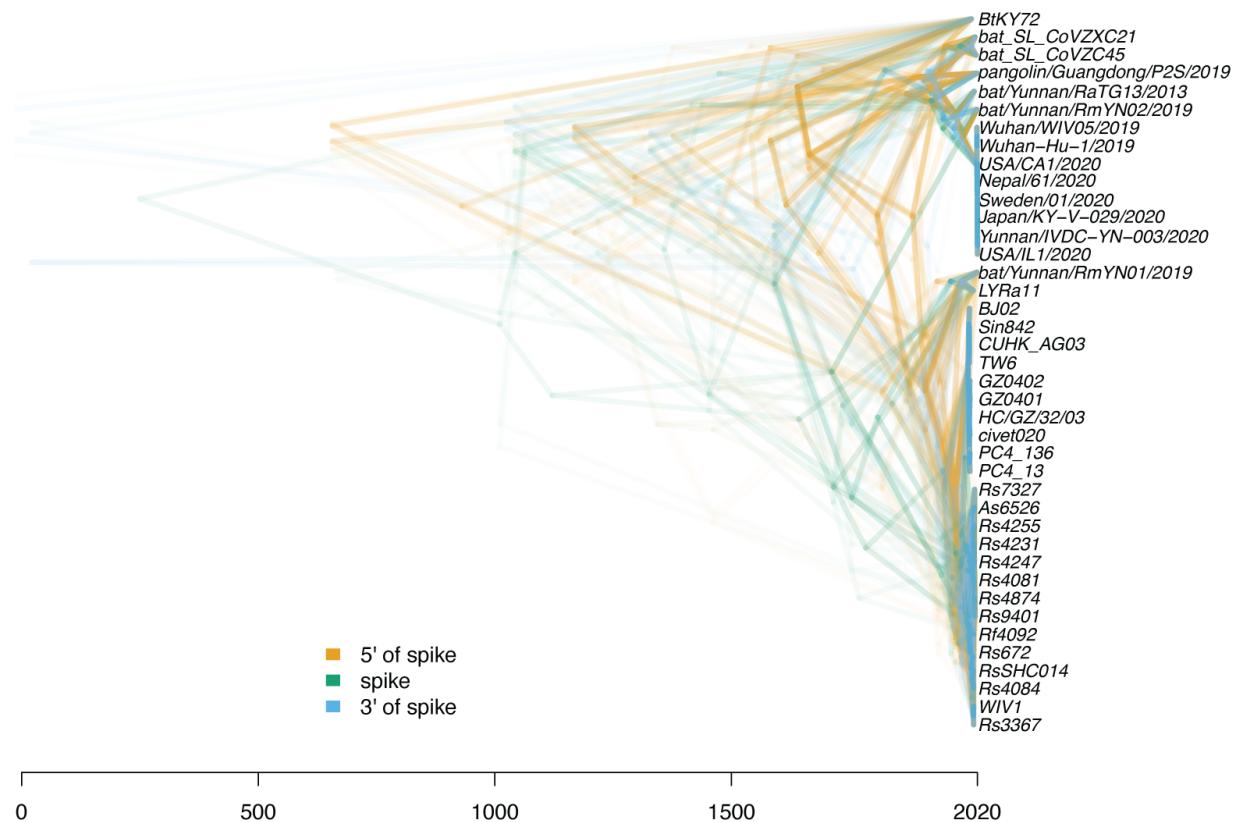
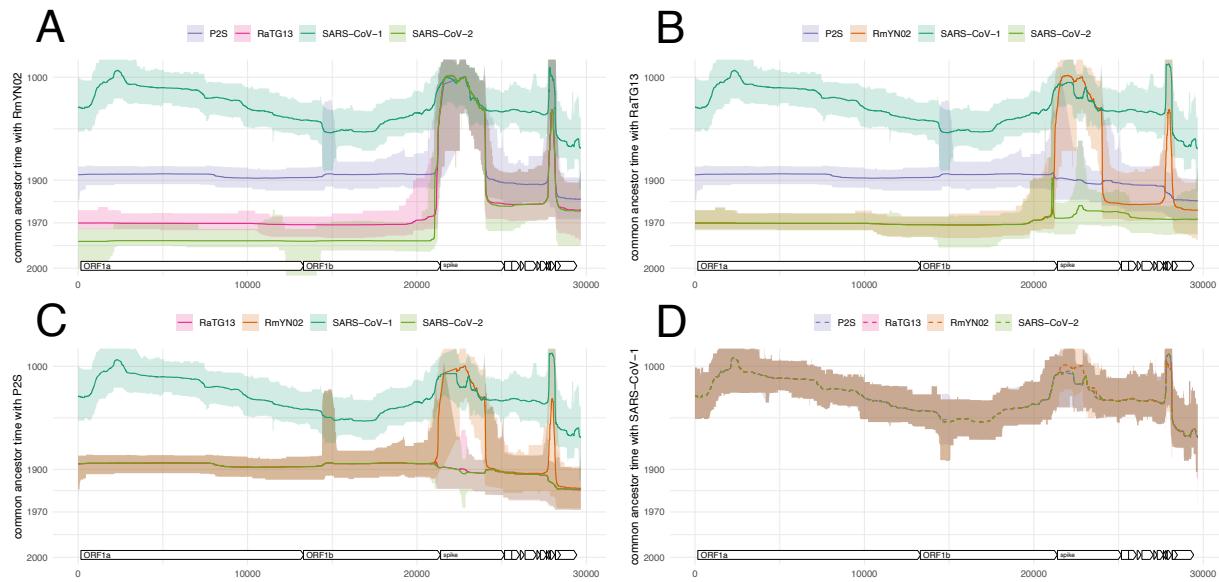


Figure S2: **Plot of the local trees of SARS-like virus on different positions across the genome.** Here, we show a densitree (Bouckaert, 2010) plot of local trees in the mcc network of SARS-like viruses. The local trees are shown for every 100th position in the genome and are computed from the mcc network shown in Fig. 1A. The different colors represent whether a local trees was towards the 5' or 3' end relative to the region that codes for the spike protein, or whether it was on the spike protein itself.



**Figure S3: Common ancestor times between sequences of the SARS-CoV-2 clade, as well as SARS-CoV-1.** Estimate of common ancestor times of RmYN02 **(A)**, RaTG13 **(B)**, P2S **(C)** and SARS-CoV-1 **(D)** with each other and with SARS-CoV-2. The estimates of the common ancestor times assume an evolutionary rate of  $5 \times 10^{-4}$ . Lower rates would push the common ancestor times further into the past, while higher rates would bring the closer to the present.

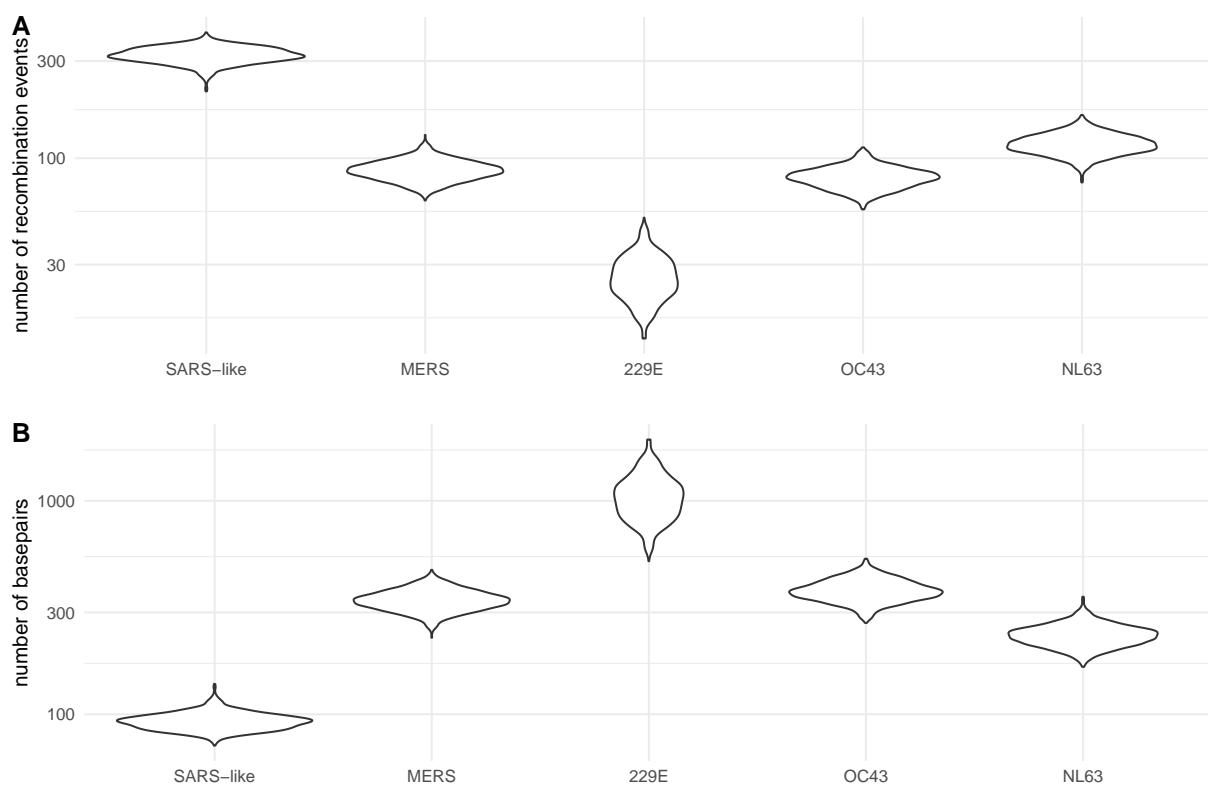


Figure S4: **Number of observable recombination events and average length of genomic segment coding for the same tree.** **A** Number of recombination events that impact the genome of sampled viruses in the dataset. **B** Average length of a segment in the genome of sampled viruses in the dataset that code for the same phylogenetic tree. That is the average length of a part of the genome that is not broken up by recombination events.

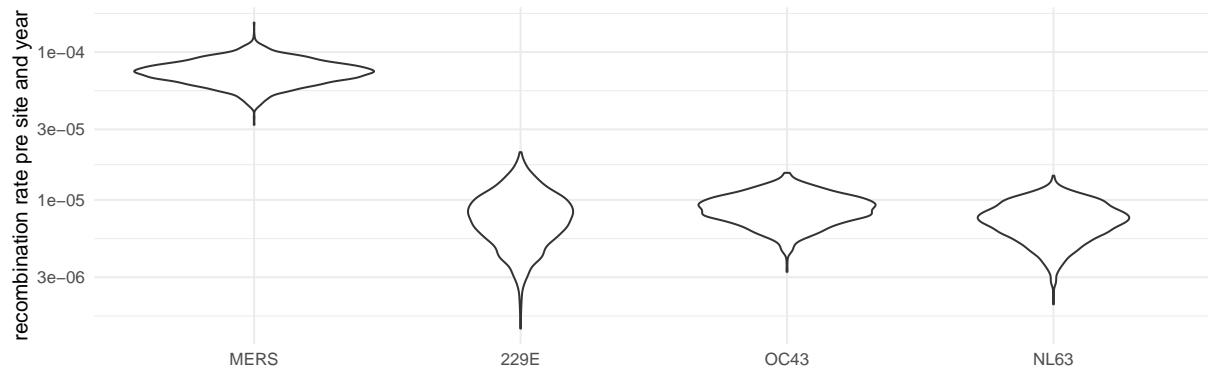


Figure S5: **Inferred recombination rates for the different coronaviruses.** Here we show the posterior distribution of recombination rates per year and per pair of adjacent nucleotides.

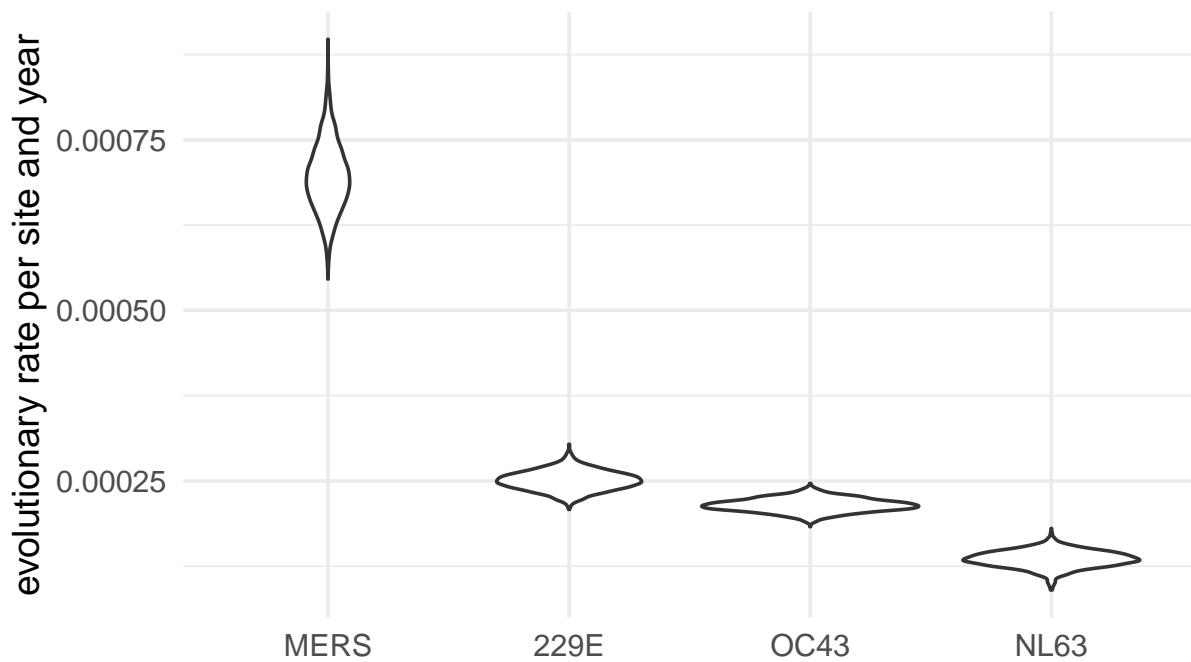
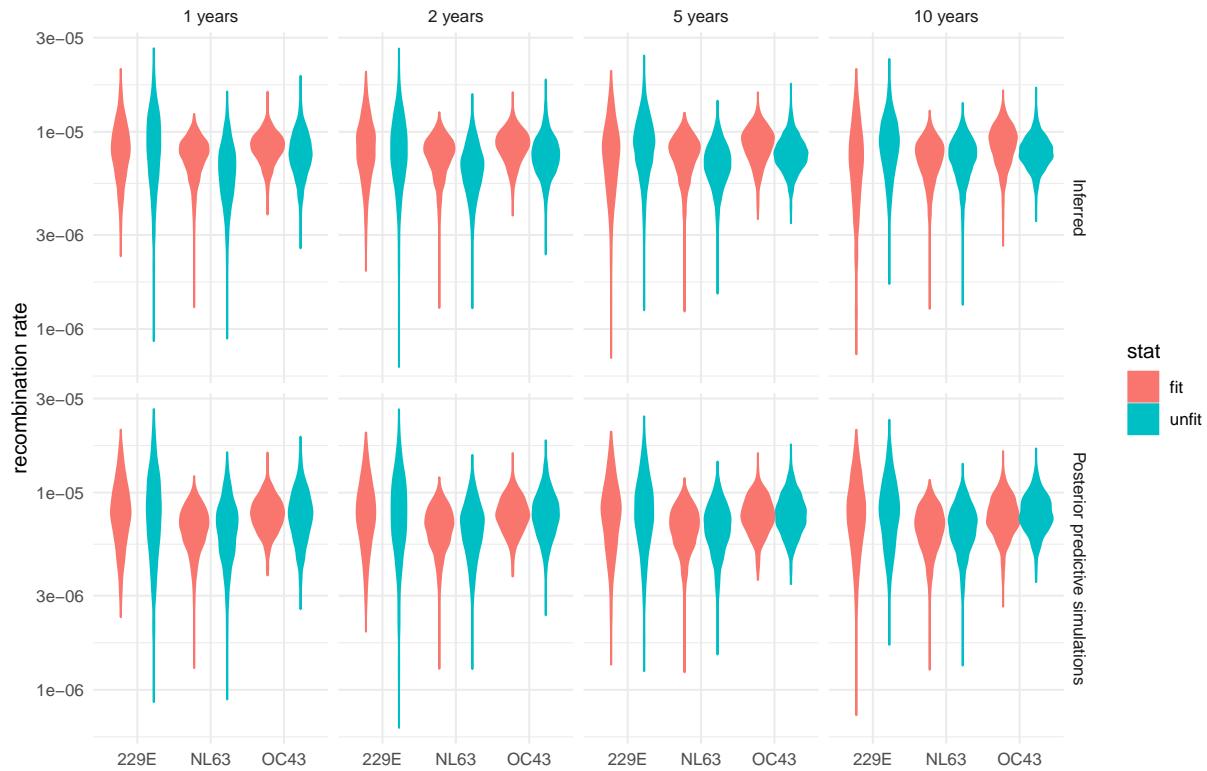
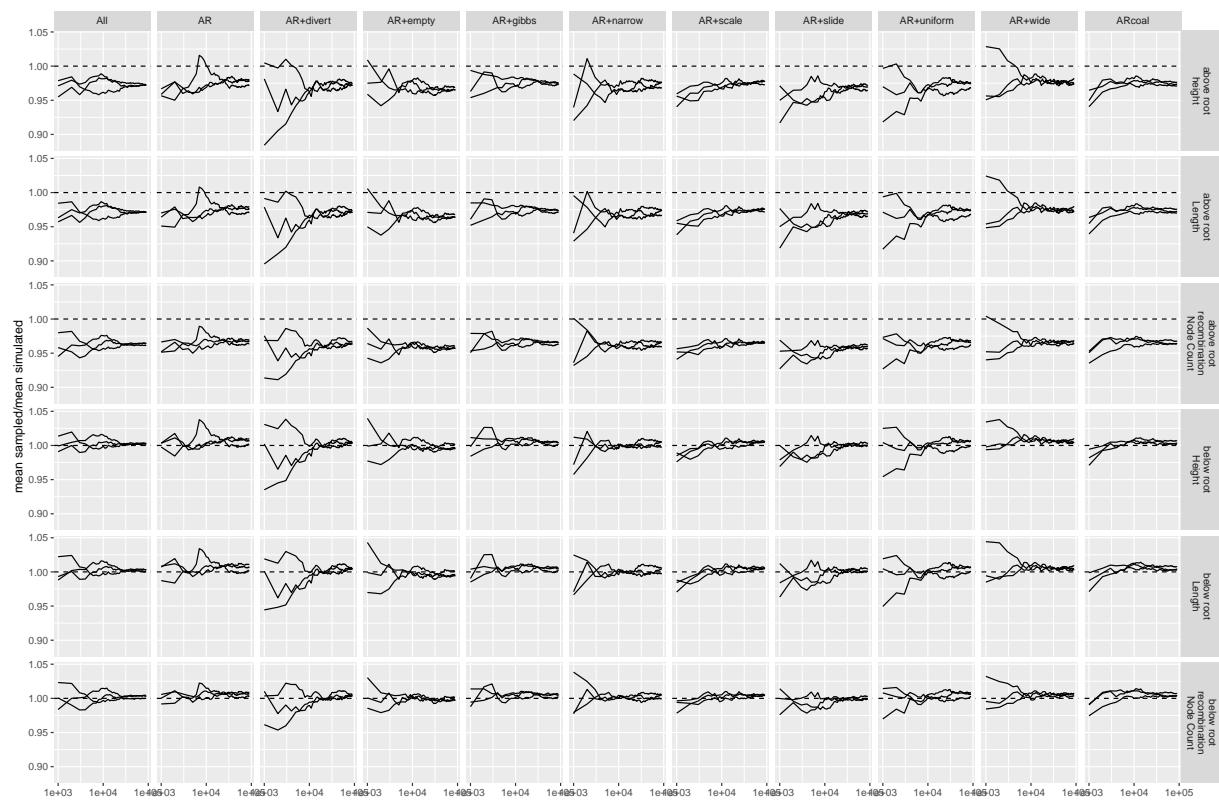


Figure S6: **Inferred evolutionary rates for the different coronaviruses.** Here we show the posterior distribution of evolutionary rates per year and nucleotide.



**Figure S7: Recombination rates of different parts of the recombination networks.** Here, we compute the recombination rates of different parts of the network based on how long lineages persist for into the future. To do so, we classified each edge of the recombination network in the posterior distribution of the different dataset into fit and unfit. Fit are edges that persist for at least 1, 2, 5 or 10 years into the future (plots from left to right). We then compute the rates of recombination on these edges as well as on those who go extinct more rapidly. We repeat the same for posterior predictive recombination networks that we simulated from the given sampling times, the inferred effective population sizes and the inferred recombination rates under the coalescent with recombination.



**Figure S8: Comparison of network statistics when simulating under the coalescent with recombination compared to sampling under the truncated coalescent with recombination.** We here compare the posterior distributions of network height, length and the number of recombination nodes when simulating recombination networks under the coalescent with recombination and when MCMC sampling under the implementation of coalescent with recombination. We compare this for all the different MCMC operators implemented. For MCMC operators which are not universal (cannot reach every point in the posterior distribution by themselves), we tested the operator jointly with the Add/remove operator. The statistics "above the root" take into account the full distribution of networks. The statistics "below the root" only take into account the parts of the network that are below (more recent) than the oldest root of any individual position in the alignment. These are the parts of the network that directly impact the likelihood.

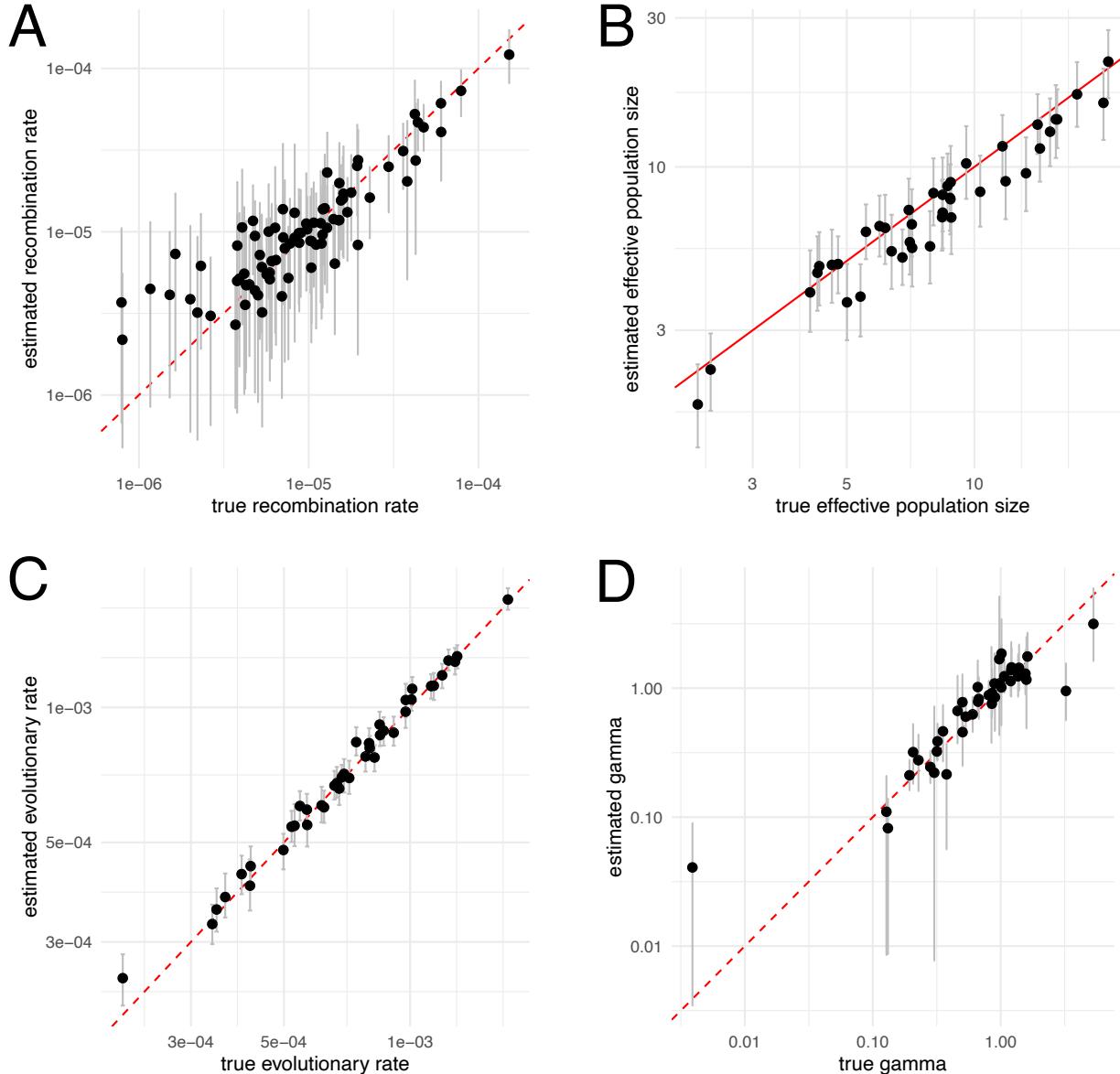


Figure S9: **Inferred vs. true rates based on simulated data.** Here, we simulated recombination networks and sequence alignment using the randomly drawn values on the x-axis and then re-inferred these parameters on the y-axis.

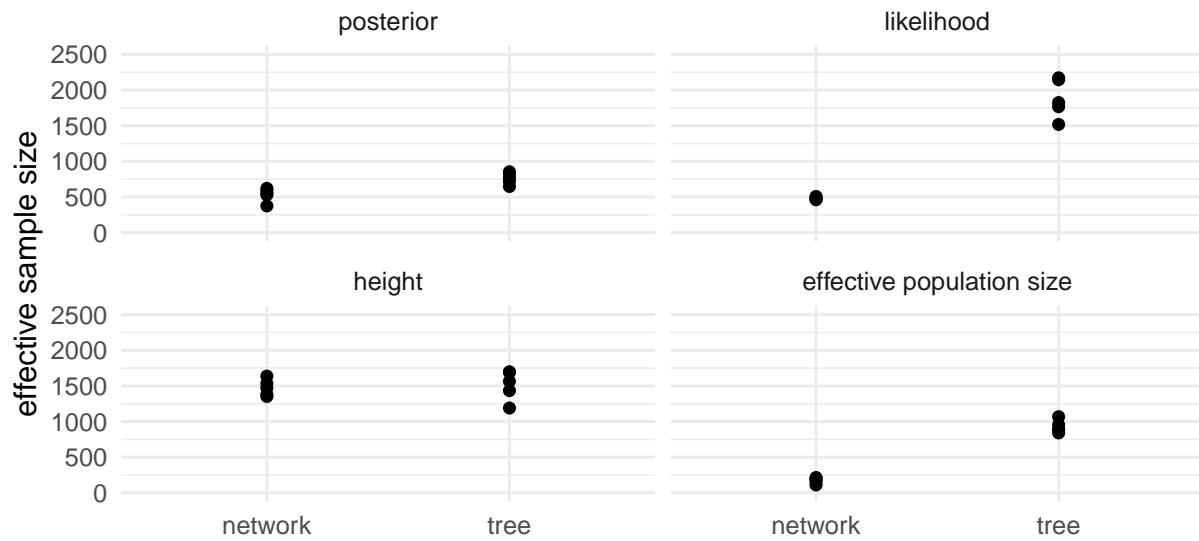


Figure S10: **Effective sample Sizes of MERS MCMC runs using the spike protein only.** Here, we compare ESS values after 25 Million MCMC iterations when inferring either networks or considering trees only for 100 MERS spike sequences. The operator weights for the inference of recombination networks is the same as used in the other coronaviruses in this manuscript. For the tree inferences, we used the default operator weights. We computed the effective sample size values computed using coda (Plummer et al., 2006) for posterior probabilities, network/tree likelihood values, network/tree root heights and effective population sizes.