Recombination patterns in coronaviruses

Nicola F. Müller^{a,1}, Kathryn E. Kistler^{a,b}, Trevor Bedford^{a,b,c},

^{a}Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
^{b}Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA
^{c}Howard Hughes Medical Institute, Seattle, WA, USA

^{1}Corresponding author

Contact: nicola.felix.mueller@gmail.com

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

Main

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

Along with mutations introduced by errors during replication or by anti-viral molecules (for example (Kim *et al.*, 2014)), different recombination processes contribute to genetic diversity in RNA viruses (reviewed by Simon-Loriere and Holmes, 2011). Reassortment in segmented viruses (generally negative-sense RNA viruses), such as influenza or rotaviruses, can produce offspring that carry segments from different parent lineages (McDonald *et al.*, 2016). In other RNA viruses (generally positive-sense RNA viruses), such as flaviviruses and coronaviruses, homologous recombination can combine different parts of a genome from different parent lineages in absence of physically separate segments on the genome of those viruses (Su *et al.*, 2016). The main mechanism of this process is thought to be via template switching (Lai, 1992), where the template for replication is switched during the replication process. Recombination breakpoints in experiments appear to be largely random, with selection selecting recombination breakpoints in some areas of the genome (Banner and Mc Lai, 1991). Recent work shows that recombination breakpoints occur more frequently in the spike region of betacoronaviruses, such as SARS-CoV-2 (Bobay *et al.*, 2020). While the evolutionary purpose of recombination in RNA viruses is not completely understood (Simon-Loriere and Holmes, 2011), there are different explanations of why recombination may be beneficial. In general, recombination is selected for if breaking up the linkage disequilibrium is beneficial (Barton, 1995). Recombination can help purge deleterious mutations from the genome, such as proposed by the mutational-deterministic hypothesis (Feldman *et al.*, 1980). It can also increase the rate at which fit combination of mutations occur, such as stated by the Robertson-Hill effect (Hill and Robertson, 1966). Alternatively, recombination in RNA viruses may also just be a by-product of the processivity the viral polymerase (Simon-Loriere and Holmes, 2011).

Recombination poses a unique challenge to phylogenetic methods, as it violates the very central assumption that the evolutionary history of individuals can be denoted by branching phylogenetic trees. Recombination breaks this assumption and requires representation of the shared ancestry of a set of sequences as a network. Not accounting for this can lead to biased phylogenetic and phylodynamic inferences (Posada and Crandall, 2002; Müller *et al.*, 2020). An analytic description of recombination is provided by the coalescent with recombination, which models a backwards in time process where lineages can coalesce and recombine (Hudson, 1983). When recombination is considered backwards in time, a single lineage results in two parent lineages, with one parent lineage carrying the genetic material from one side of a random recombination breakpoint and the other parent lineage carrying the genetic material of the other side of this breakpoint. This equates to the backwards in time equivalent of template switching where there is one recombination breakpoint per recombination event.

Currently, some Bayesian phylogenetic approaches exist that infer recombination networks, or ancestral recombination graphs (ARG), but are either approximate or do not directly allow for efficient model-based inference. Some approaches consider tree-based networks (Didelot *et al.*, 2010; Vaughan *et al.*, 2017), where the networks consist of a base tree with recombination edges that always attach to edges on the base tree. Alternative approaches rely on approximations to the coalescent with recombination (Rasmussen *et al.*, 2014; McVean and Cardin, 2005), consider a different model of recombination (Müller *et al.*, 2020), or seek to infer recombination networks absent an explicit recombination model (Bloomquist and Suchard, 2010). Bayesian and maximum likelihood methods have also been used to account for gene transfer events when describing the evolutionary history of species from multiple loci (for example (Meng and Kubatko, 2009; Yu *et al.*, 2014). Additionally, methods have been used to describe non-tree-like evolution using split trees (Bryant and Moulton, 2004; Huson and Bryant, 2006). There is, however, a gap for Bayesian inference of recombination networks under the coalescent with recombination that can be applied to study pathogens, such as coronaviruses.

In order to fill this gap, we here develop a Markov chain Monte Carlo (MCMC) approach to efficiently infer recombination networks under the coalescent with recombination for sequences sampled over time. This framework allows joint estimation of recombination networks, effective population sizes, recombination rates and parameters describing mutations over time from genetic sequence data sampled through time. We explicitly do not make additional approximation to characterize the recombination process, other than those of the coalescent with recombination (Hudson, 1983), such as, for example, the approximation of tree based networks. We implemented this approach as an open source software package for BEAST2 (Bouckaert *et al.*, 2018). This allows incorporation of the various evolutionary models already implemented in BEAST2. Using a Bayesian approach has several advantages. In particular, it allows us to account for uncertainty in the parameter and network estimates. Additionally, it allows balancing different sources of information against each other. The coalescent with recombination model, for example, will tend to favor networks with fewer recombination events. This ”cost” of adding more recombination events depends on the recombination rate. At lower rates of recombination, adding new recombination events is more costly and the information coming from the sequence alignment in support of a recombination event needs to be greater.

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

Rampant recombination in SARS-like coronaviruses

Recombination has been implicated at the beginning of the SARS-CoV-1 outbreak (Hon *et al.*, 2008) and has been suggested as the origin of the receptor binding domain in SARS-CoV-2 (Li *et al.*, 2020), though Boni *et al.* (2020) report that recombination is unlikely to be the origin of SARS-CoV-2. While this strongly suggests non-tree-like evolution, the evolutionary history of SARS-like viruses has, out of necessity, mainly been denoted using phylogenetic trees.

We here reconstruct the recombination history of SARS-like viruses, which includes SARS-CoV-1 and SARS-CoV-2 as well as related bat (Ge *et al.*, 2013, 2016; Zhou *et al.*, 2020) and pangolin (Lam *et al.*, 2020) coronaviruses. To do so, we infer the recombination network of SARS-like viruses under the coalescent with recombination. We assumed that the rates of recombination and effective population sizes were constant over time and that the genomes evolved under a GTR+Γ_4 model. Similar to the estimate in Boni *et al.* (2020), we used a fixed evolutionary rate of 5×10^{-4} per nucleotide and year. We fixed the evolutionary rate, since the time interval of sampling between individual isolates is relatively short compared to the time scale of the evolutionary history of SARS-like viruses. This means that the sampling times themselves offer little insight into the evolutionary rates and, in absence of other calibration points, there is little information about the evolutionary rate in this dataset. This, in turn, means that if the evolutionary rate we used here is inaccurate then the timings of common ancestors will also be inaccurate. Therefore, exact timings and calendar dates in this analyses should be taken as guide posts rather than formal estimates.

As shown in Figure 1A, the evolutionary history of SARS-like viruses is characterized by frequent recombination events. Consequently, characterizing evolutionary history of SARS-like viruses by a single genome-wide phylogeny is bound to be inaccurate and potentially misleading. We infer the recombination rate in SARS-like viruses to be approximately 2×10^{-6} per site per year, which is about 200 times lower than the evolutionary rate. This rate translated to about 0.06 recombination events per lineage per year, which is slightly lower than the estimated rate of recombination for the seasonal human coronaviruses and the reassortment rates for pandemic 1918 like influenza A/H1N1 and influenza B viruses, which are all around $0.1 - 0.2$ reassortment events per lineage per year (Müller *et al.*, 2020). This recombination rate is a function of co-infection rates, probability of recombination occurring upon co-infection, and selection. As such, the recombination rate we infer here will be (possibly substantially) lower than the within host rate of recombination.

These recombination events were not evenly distributed across the genome and, instead, were relatively higher in areas outside those coding for ORF1ab (Fig. S1 & S5). Additionally, our inference suggests that rates of recombination are slightly elevated on spike subunit 1 compared to subunit 2 (Fig. S1). If we track recombination events ancestral to the SARS-CoV-2 lineage that are inferred to have happened in the last 100 years, we find evidence for recombination breakpoints occurring close to the 5' end of spike, just outside the coding region (see fig S4). Additionally, we find support for recombination breakpoints toward the 3' end of spike, near the nucleocapsid gene (see fig S4). If we assume that during genome replication in coronaviruses template shifts occur randomly on the genome (Banner and Mc Lai, 1991), differences in observed recombination rates could be explained by selection favoring recombination events 3' to ORF1ab relative to elsewhere on the genome. For instance, it has been suggested that selection has acted on multiple recombination events within spike to enhance dynamic molecular movements of the Spike protein (Tagliamonte *et al.*, 2021).

We next investigate when different viruses last shared a common ancestor (MRCA) along the genome (see Fig. 1B and Fig. S2). RmYN02 (Zhou *et al.*, 2020) shares the MRCA with SARS-CoV-2 on the part of the genome that codes for ORF1ab (Fig. 1B). We additionally find strong evidence for one or more recombination events in the ancestry of RmYN02 at the beginning of spike (Fig. 1B). This recent recombination event is unlikely to have occurred with a recent ancestor of any of the coronaviruses included in this dataset since the common ancestor of RmYN02 with any other virus in the dataset is approximately the same (Fig. S3A). In other words, large parts of the spike protein of RmYN02 are as related to SARS-CoV-2 as SARS-CoV-2 is to

SARS-CoV-1. The common ancestor timings of P2S across the genome are equal between RaTG13 and SARS-CoV-2 (Fig. S3C). RaTG13 on the other hand is more closely related to SARS-CoV-2 than P2S (Fig. S3B) across the entire genome.

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

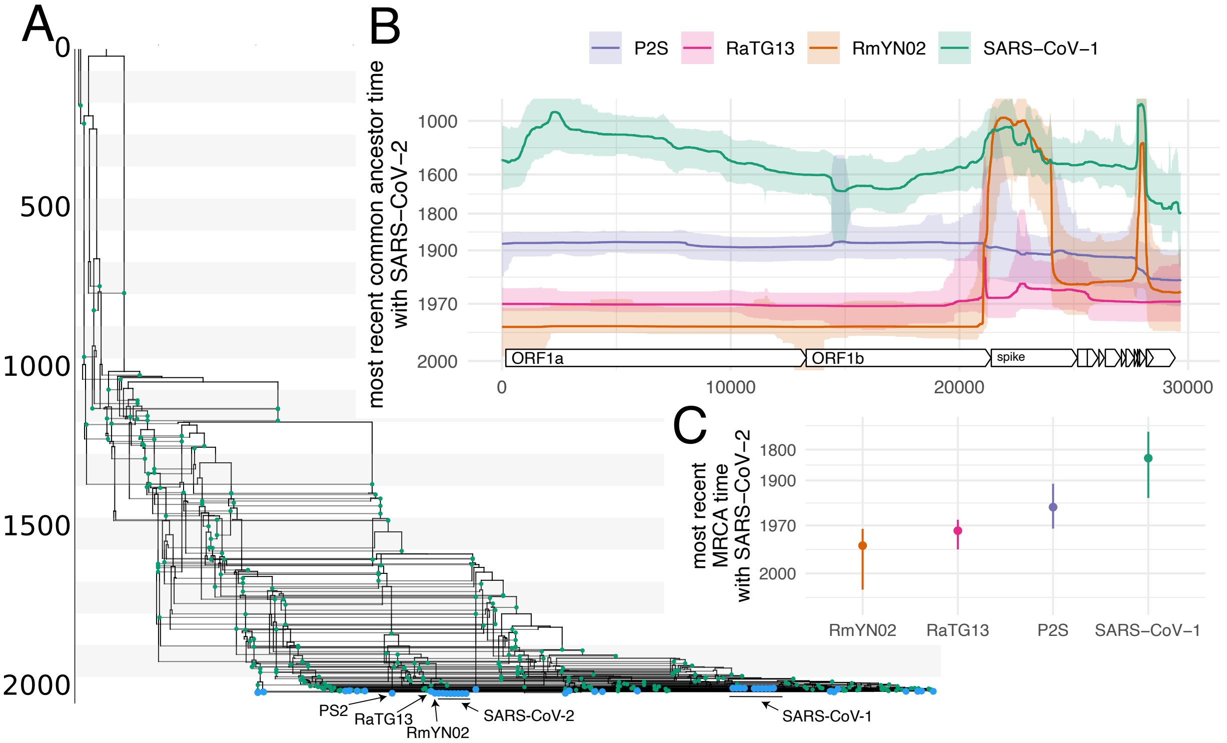

Figure 1: **Evolutionary history of SARS-like viruses. A** Maximum clade credibility network of SARS-like viruses. Blue dots denote samples and green dots recombination events. **B** Common ancestor times of Wuhan-Hu1 (SARS-CoV-2) with different SARS-like viruses on different positions of the genome. The y-axis denotes 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

Rates of recombination vary with rates of adaptation in human seasonal coronaviruses

We next investigate recombination patterns in MERS-CoV, which has over 2500 confirmed cases in humans, as well as in human seasonal coronaviruses 229E, OC43 and NL63, which have widespread seasonal circulation in humans. As for the SARS-like viruses, we jointly infer recombination networks, rates of recombination and population sizes for these viruses. We assumed that the genomes evolved under a GTR+Γ_4 model and, in contrast to the analysis of SARS-like viruses, inferred the evolutionary rates. We observe frequent recombination in the history of all 4 viruses, wherein genetic ancestry is described by network rather than a strictly branching phylogeny (Fig. 2A-D).

The human seasonal coronaviruses all have recombination rates around 1×10^{-5} per site and year (Fig. S7). This is around 10 to 20 times lower than the evolutionary rate (Fig. S8). In contrast to the recombination rates, the evolutionary rates vary greatly across the human seasonal coronaviruses, with rates between a median of 1.3×10^{-4} (CI $1.1 - 1.5 \times 10^{-4}$) for NL63 and median rate of 2.5×10^{-4} (CI $2.2 - 2.7 \times 10^{-4}$) and 2.1×10^{-4} (CI $1.9 - 2.3 \times 10^{-4}$) for 229E and OC43 (Fig. S8). These evolutionary rates are substantially lower than those estimated for SARS-CoV-2 (1.1×10^{-3} substitutions per site and year (Duchene *et al.*, 2020)), which are more in line with our estimates for the evolutionary rates of MERS with a median rate of 6.9×10^{-4} (CI $6.0 - 7.9 \times 10^{-4}$). Evolutionary rate estimates can be time dependent, with datasets spanning more time estimating lower rates of evolution that those spanning less time (Duchêne *et al.*, 2014). In turn, this means that the evolutionary rates estimates for SARS-CoV-2 will likely be lower the more time passes. It is unclear though, it will approximate the evolutionary rates of other seasonal coronaviruses.

On a per-lineage basis the estimated recombination rate for seasonal coronaviruses 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). MERS-CoV infections are highly prevalent in camels, 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 coronaviruses.

The evolutionary purpose of recombination in RNA viruses, as well as whether recombination provides a fitness benefit is unclear (Simon-Loriere and Holmes, 2011). To investigate whether recombination benefits fitness in human coronaviruses, we next tested whether rates of recombination differed on different parts of the genome. To do so, we allowed for different relative rates of recombination within the region 5' of spike (i.e. mostly ORF1ab), spike itself, and everything 3' of spike. We computed recombination rate ratios on each of these 3 sections of the genome as the recombination rate on that section divided by the mean rate on the other two sections. We infer that recombination rates are elevated in the spike protein of all human seasonal coronaviruses considered here (Fig. 3, S10 & S11). This is consistent with other work estimating higher rates of recombination on the spike protein of betacoronaviruses (Bobay *et al.*, 2020).

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

We next investigated whether these trends hold when looking only at spike. The spike protein is made up of two subunits. Subunit 1 (S1) binds to the host cell receptor, while subunit 2 (S2) facilitates fusion of the viral and cellular membrane (Walls *et al.*, 2020). S1 contains the receptor binding domain and rates of adaptation have

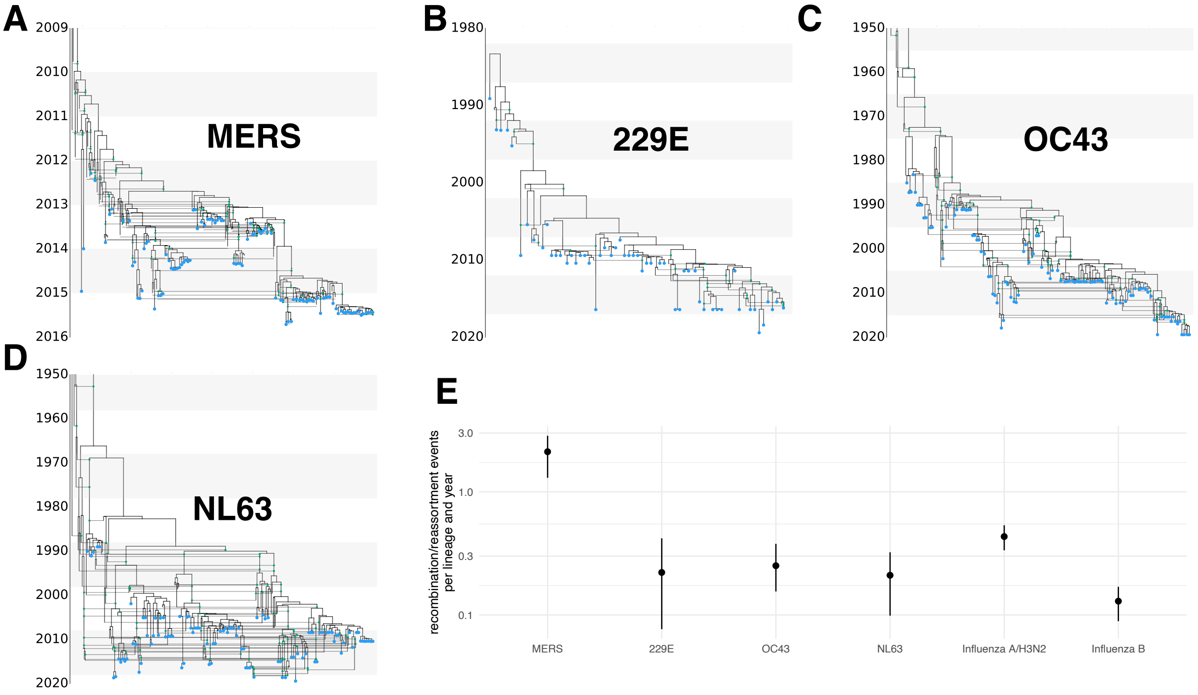

Figure 2: **Recombination networks and rates for coronaviruses MERS, 229E, OC43 and NL63.** Recombination networks for MERS (**A**) and seasonal human coronaviruses 229E (**B**), OC43 (**C**) and NL63 (**D**). **E** Recombination rates (per lineage and year) for the different coronaviruses compared to 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.

been shown to be high in S1 for 229E and OC43 (Kistler and Bedford, 2021). While the rates of adaptation are relatively low overall for NL63, there is still some evidence that they are elevated in S1 compared to S2 (Kistler and Bedford, 2021).

To test whether recombination rates vary with rates of adaptation on the subunits of spike as well, we inferred the recombination rates from spike only, allowing for different rates of recombination on S1 versus the rest of spike. We find that the rates of recombination are elevated on S1 for 229E and OC43 compared to the rest of the spike gene (Fig. 3). This is consistent with strong absolute rates of adaptation on S1 on these two viruses. For NL63, we find weak evidence for the rate on S2 to be slightly higher than on S1 (Fig. 3), even though the rates of adaptation are inferred to be higher on S1. The absolute rate of adaptation in S1 of NL63 is, however, substantially lower than for 229E or OC43. Additionally, the uncertainty around the estimates on adaption rate ratios between the two subunits for NL63 are rather large and include no difference at all. Overall, these results suggest that recombination events are either positively or negatively selected for. Elevated rates of recombination in areas where adaptation is stronger have been described for other organisms (reviewed here (Nachman, 2002)). Alternatively, higher rates of recombination could also be due to mechanistic reasons, as has been suggested in the case of SARS-CoV-2 (Turakhia *et al.*, 2021).

To further investigate this, we next computed the rates of recombination on fitter and less fit parts of the recombination networks of 229E, OC43 and NL63. To do so, we first classify each edge of the inferred posterior distribution of the recombination networks into fit and unfit based on how long a lineage survives into the future. Fit edges are those that have descendants at least one, two, five or ten years into the future and unfit edges

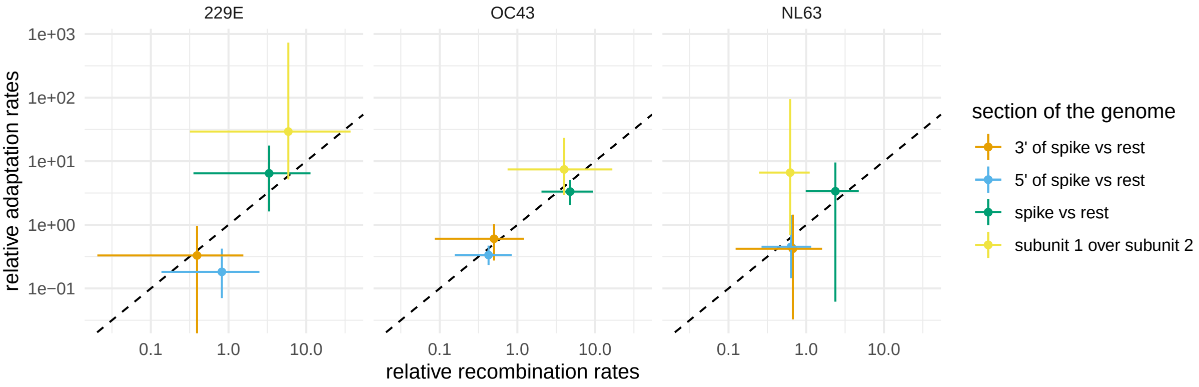

Figure 3: **Comparison of recombination rates with rates of adaptation on different parts of the genomes of seasonal human coronaviruses 229E, OC43 and NL63.** Relationship between estimated relative recombination rate (x-axis) and relative adaptation rate (y-axis) for three different seasonal human coronaviruses: 229E, OC43 and NL63. These estimates are shown for different parts of the genome, indicated by the different colors. These result from two different types of analysis: one using spike only (subunit 1 over subunit 2, shown in yellow) and one using the full genome (shown in orange, blue and green). 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.

those that do not. We then computed the rates of recombination on both types of edges for the entire posterior distribution of networks. Overall, we do not find that fit edges show relatively higher rates of recombination (see figure S9). The simplest explanation is that we do not have enough data points to measure recombination rates on unfit edges, meaning to measure recombination rates on part of the recombination network where selection had too little time to shape which lineages survive and which go extinct. An alternative explanation to why we see elevated rate or recombination in the spike protein, but do not observe a population level fitness benefit could be that most (outside of spike) recombinants could be detrimental to fitness with few (within spike) having little fitness effect at all.

Conclusion

Though not yet highly prevalent, evidence for recombination in SARS-CoV-2 has started to appear (VanInsberghe *et al.*, 2020; Jackson *et al.*, 2021; Varabyou *et al.*, 2021; Ignatieva *et al.*, 2021). As such, it is crucial to know the extent to which recombination is expected to shape SARS-CoV-2 in the coming years, to have methods to identify recombination and to perform phylogenetic reconstruction in the presence of recombination. The results shown here indicate that some recombinants are either positively or negatively selected for. Estimating the deleterious load of viruses before and after recombination using ancestral sequence reconstruction (Yang *et al.*, 1995) could help shed light on which sequences are favored during recombination. Furthermore, having additional sequences to reconstruct recombination patterns the seasonal coronaviruses should clarify the role recombination plays in the long term evolution of these viruses.

While their impact on the evolutionary dynamics of SARS-CoV-2 remains unclear, the likely rise of future SARS-CoV-2 recombinants will further necessitate methods that allow phylogenetic and phylodynamic inferences to be performed in the presence of recombination (Neches *et al.*, 2020). In absence of that, recombination has to be either ignored, leading to biased phylogenetic and phylodynamic reconstruction (Posada and Crandall, 2002), or non-recombinant parts of the genome have to be used for analyses, reducing the precision of these methods. Our approach addresses this gap by providing a Bayesian framework to infer recombination networks. To facilitate easy adaptation, we implemented the method so that analyses can be set up following the same workflow as regular BEAST2 (Bouckaert *et al.*, 2018) analyses. Extending the current suite of population dynamic models, such as birth-death models (Stadler, 2009) or models that account for population structure (Hudson *et al.*, 1990; Lemey *et al.*, 2009), will further increase the applicability of recombination models to study the spread of pathogens.

Materials and Methods

Coalescent with recombination

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

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:

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

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

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 μ as:

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

with D_i denoting the nucleotides at position i in the sequence alignment and T_i denoting the tree at position i. The likelihood at each individual position in the alignment can then be computed using the standard pruning

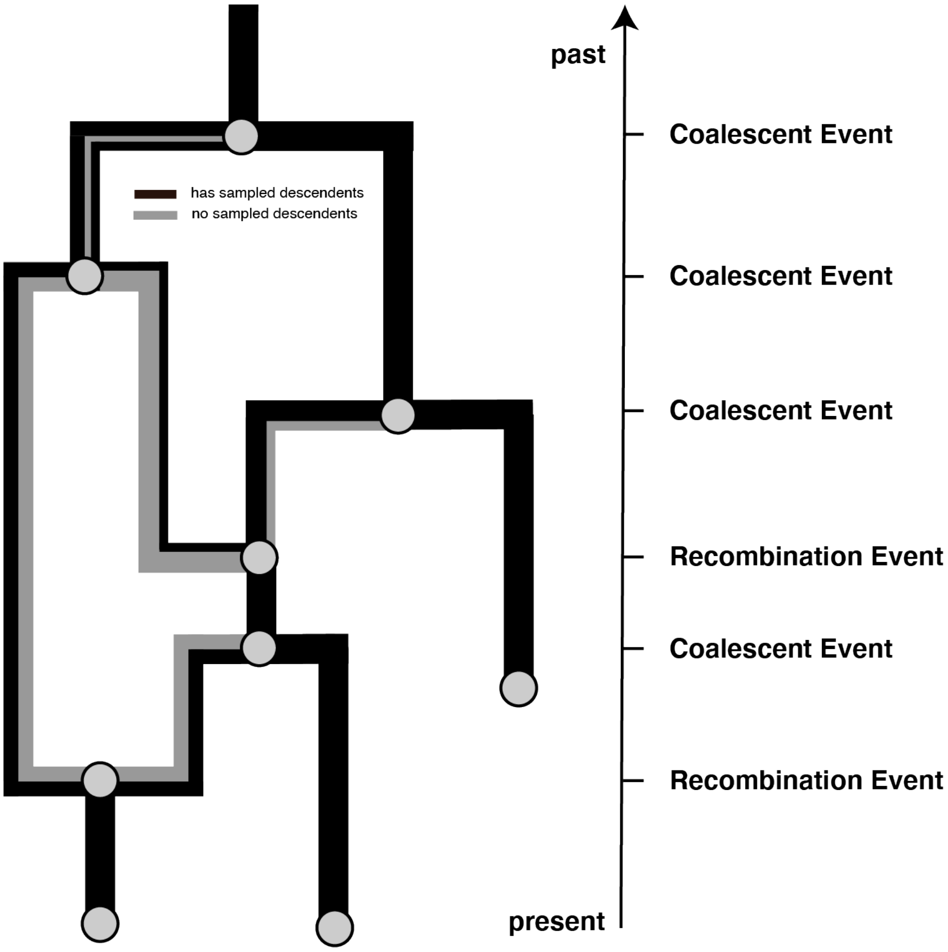

Figure 4: **Example recombination network**. Events that can occur on a recombination network as considered here. We consider events to occur from present backwards in time to the past (as is the norm when looking at coalescent processes). Lineages can be added upon sampling events, which occur at predefined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

algorithm (Felsenstein, 1981). We implemented the network likelihood calculation $P(D_i|T_i, \mu)$ such that it allows making use of all the standard site models in BEAST2. Currently, we only consider strict clock models and do not allow for rate variations across different branches of the network. This is because the number of edges in the network changes over the course of the MCMC, making relaxed clock models complex to implement. We implemented the network likelihood such that it can make use of caching of intermediate results and use unique patterns in the multiple sequence alignment, similar to what is done for tree likelihood computations.

Network Prior

The network prior is denoted by $P(N|\theta, \rho)$, which is the probability of observing a network and the embedding of segment trees under the coalescent with recombination model, with effective population size θ and per-lineage recombination rate ρ. 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):

$$P(N|\theta,\rho) = \prod_{i=1}^{\#events} P(event_i|L_i,\theta,\rho) \times P(interval_i|L_i,\theta,\rho),$$

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 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:

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

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

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

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:

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

where λ^c denotes the rate of coalescence and can be expressed as:

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

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

MCMC Algorithm for Recombination Networks

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

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

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

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

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

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

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

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

One of the issues when inferring these recombination networks is that the root height can be substantially larger than when not allowing for recombination events. This can cause computational issue when performing inferences. To circumvent this, we truncate the recombination networks by reducing the recombination rate some time after all positions of the sequence alignment have reached their common ancestor height.

Validation and testing

We validate the implementation of the coalescent with recombination network prior as well as all operators in the supplement S12. We also show that truncating the recombination networks does not affect the sampling of recombination networks prior to reaching the common common ancestor height of all positions in the sequence alignment.

We then tested whether we are able to infer recombination networks, recombination rates, effective population sizes and evolutionary parameters from simulated data. To do so, we randomly simulated recombination networks under the coalescent with recombination. On top of these, we then simulated multiple sequence alignments. We then re-infer the parameters used to simulate using our MCMC approach. As shown in Figure S13, these parameters are retrieved well from simulated data with little bias and accurate coverage of simulated parameters by credible intervals.

We next tested how well we can retrieve individual recombination events. To do so, we plot the location and timings simulated recombination events for the first 9 out of 100 simulations. We then plot the density of recombination events in the posterior distribution of networks, based on timing and location of the inferred breakpoint on the genome. As shown in figure S14, we are able to retrieve the true (simulated) recombination events well.

We next tested how the speed of inference scales with the number of recombination events, the number of samples in the dataset and the evolutionary rate. To do so, we simulated 300 recombination networks and sequence alignment of length 10,000 under a Jukes Cantor model with between 10 and 200 leafs and a recombination rate between 1×10^{-5} and 2×10^{-5} recombination events per site per year. This means that for each simulation, there were between 0 and 100 recombination events, allowing us to investigate how the inference scales in different settings. As shown in figure S16, the ESS per hour decreases with the number of recombination events and samples, but not the evolutionary rates. In particular, the ESS per hour decreases much faster with the number of recombination events in a dataset than the number of samples. This suggest that the methods can currently be used more easily to analyze dataset with large number of samples over large number of recombination events.

We next tested how the choice of the prior distribution on the recombination rate impacts the recombination rate estimate. To do so, we simulate 20 recombination networks and sequence alignment of length 10000 under a Jukes Cantor model with 100 leafs and a recombination rate drawn randomly from a log-normal distribution. We then infer the recombination rates using 5 different recombination rate priors as shown in figure S15F that put some or a lot of weight on the wrong parameters. As shown in figures S15A-E, we are able to infer recombination rates, even with the wrong priors.

Additionally, we compared the effective sample size values from MCMC runs inferring recombination networks for the MERS spike protein to treating the evolutionary histories as trees. We find that although the effective sample size values are lower when inferring recombination networks, they are not orders of magnitude lower (see fig S17).

Recombination network summary

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

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

Sequence data

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

Rates of adaptation

The rates of adaptation were calculated using a modification of the McDonald-Kreitman method, as designed by Bhatt *et al.* (2011), and implemented in Kistler and Bedford (2021). Briefly, for each virus, we aligned the sequence of each gene or genomic region. Then, we split the alignment into 3-year sliding windows, each containing a minimum of 3 sequenced isolates. We used the consensus sequence at the first time point as the outgroup. A comparison of the outgroup to the alignment of each subsequent temporal yielded a measure of synonymous and non-synonymous fixations and polymorphisms at each position in the alignment. This approach requires having sequence data gathered over relatively long time periods where the consensus genome allows for accurate description of the long term evolutionary patterns and, as such, would not be adequate for a pathogen with relatively short evolutionary history, such as for SARS-CoV-2. We used proportional site-counting for these estimations (Bhatt *et al.*, 2010). We assumed that selectively neutral sites are all silent mutations as well as replacement polymorphisms occurring at frequencies between 0.15 and 0.75 (Bhatt *et al.*, 2011). We identified adaptive substitutions as non-synonymous fixations and high-frequency polymorphisms that exceed the neutral expectation. We then estimated the rate of adaptation (per codon per year) using linear regression of the number of adaptive substitutions inferred at each time point. In order to compute the 5' of spike and 3' of spike rates of adaptation we used the weighted average of all coding regions to the left (upstream) or right (downstream) of the spike gene, respectively, using the length of the individual sections as weights. We estimated the uncertainty by running the same analysis on 100 bootstrapped outgroups and alignments.

Code availability

The Recombination package is implemented as an addon to the Bayesian phylogenetics software platform BEAST2 (Bouckaert *et al.*, 2018). All MCMC analyses performed here, were run using adaptive parallel 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) and ggenes (Wilkins, 2019).

Data availability

The BEAST2 input xml files for all coronavirus analyses in this manuscript, as well as the files used to post process these analyses are available from `https://github.com/nicfel/Recombination-Material`. The xml files include the sequence data and exact input specification of the coronavirus analyses performed in this manuscript

Acknowledgments

We would like to thank three anonymous reviewers for their helpful comments that improved the manuscript. We would also like to thanks Timothy G. Vaughan for helpful insights into the implementation of the software. NFM is funded by the Swiss National Science Foundation (P2EZP3_191891). KEK is a NSF GRFP Fellow (DGE-1762114) TB is a Pew Biomedical Scholar and is supported by NIH R35 GM119774. The Scientific Computing Infrastructure at Fred Hutch is supported by NIH ORIP S10OD028685

Conflict of interest

The authors declare no conflict of interest

Author Contributions

NFM and TB conceived and designed the experiments. NFM and KEK performed statistical analysis and analysed the data. NFM implemented the software. NFM, KEK and TB wrote the paper.

References

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**, 450–452.

Banner, L. R. and Mc Lai, M. (1991). Random nature of coronavirus rna recombination in the absence of selection pressure. *Virology*, **185**(1), 441–445.

Barton, N. (1995). A general model for the evolution of recombination. *Genetics Research*, **65**(2), 123–144.

Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., *et al.* (2020). Cryptic transmission of sars-cov-2 in washington state. *Science*, **370**(6516), 571–575.

Bhatt, S., Katzourakis, A., and Pybus, O. G. (2010). Detecting natural selection in rna virus populations using sequence summary statistics. *Infection, Genetics and Evolution*, **10**(3), 421–430.

Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011). The genomic rate of molecular adaptation of the human influenza a virus. *Molecular biology and evolution*, **28**(9), 2443–2451.

Bloomquist, E. W. and Suchard, M. A. (2010). Unifying vertical and nonvertical evolution: a stochastic arg-based framework. *Systematic biology*, **59**(1), 27–41.

Bobay, L.-M., O'Donnell, A. C., and Ochman, H. (2020). Recombination events are concentrated in the spike protein region of betacoronaviruses. *PLoS Genetics*, **16**(12), e1009272.

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 origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nature Microbiology*, **5**(11), 1408–1417.

Bordewich, M., Linz, S., and Semple, C. (2017). Lost in space? generalising subtree prune and regraft to spaces of phylogenetic networks. *Journal of theoretical biology*, **423**, 1–12.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D., De Maio, N., *et al.* (2018). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *BioRxiv*, page 474296.

Bouckaert, R. R. (2010). Densitree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**(10), 1372–1373.

Bryant, D. and Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, **21**(2), 255–265.

Cardona, G., Rosselló, F., and Valiente, G. (2008). Extended newick: it is time for a standard representation of phylogenetic networks. *BMC bioinformatics*, **9**(1), 1–8.

Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, **186**(4), 1435–1449.

Duchêne, S., Holmes, E. C., and Ho, S. Y. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B: Biological Sciences*, **281**(1786), 20140732.

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the phylodynamic threshold of sars-cov-2. *Virus evolution*, **6**(2), veaa061.

Dudas, G., Carvalho, L. M., Rambaut, A., and Bedford, T. (2018). Mers-cov spillover at the camel-human interface. *Elife*, **7**, e31257.

Feldman, M. W., Christiansen, F. B., and Brooks, L. D. (1980). Evolution of recombination in a constant environment. *Proceedings of the National Academy of Sciences*, **77**(8), 4838–4841.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**(6), 368–376.

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 and characterization of a bat sars-like coronavirus that uses the ace2 receptor. *Nature*, **503**(7477), 535–538.

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 of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virologica Sinica*, **31**(1), 31–40.

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 and evolutionary dynamics of pathogens. *science*, **303**(5656), 327–332.

Heled, J. and Bouckaert, R. R. (2013). Looking for trees in the forest: summary tree from posterior samples. *BMC evolutionary biology*, **13**(1), 1–11.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, **8**(3), 269–294.

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 recombinant origin of a bat severe acute respiratory syndrome (sars)-like coronavirus and its implications on the direct ancestor of sars coronavirus. *Journal of virology*, **82**(4), 1819–1826.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, **23**(2), 183–201.

Hudson, R. R. *et al.* (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, **7**(1), 44.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, **23**(2), 254–267.

Ignatieva, A., Hein, J., and Jenkins, P. A. (2021). Ongoing recombination in sars-cov-2 revealed through genealogical reconstruction. *bioRxiv*.

Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., *et al.* (2021). Generation and transmission of interlineage recombinants in the sars-cov-2 pandemic. *Cell*.

Kim, E.-Y., Lorenzo-Redondo, R., Little, S. J., Chung, Y.-S., Phalora, P. K., Maljkovic Berry, I., Archer, J., Penugonda, S., Fischer, W., Richman, D. D., *et al.* (2014). Human apobec3 induced mutation of human immunodeficiency virus type-1 contributes to adaptation and evolution in natural infection. *PLoS pathogens*, **10**(7), e1004281.

Kistler, K. E. and Bedford, T. (2021). Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses oc43 and 229e. *Elife*, **10**, e64509.

Lai, M. (1992). Rna recombination in animal and plant viruses. *Microbiology and Molecular Biology Reviews*, **56**(1), 61–79.

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.* (2020). Identifying sars-cov-2-related coronaviruses in malayan pangolins. *Nature*, **583**(7815), 282–285.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput Biol*, **5**(9), e1000520.

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). Emergence of sars-cov-2 through recombination and strong purifying selection. *Science Advances*, **6**(27), eabb9153.

McDonald, S. M., Nelson, M. I., Turner, P. E., and Patton, J. T. (2016). Reassortment in segmented rna viruses: mechanisms and outcomes. *Nature Reviews Microbiology*, **14**(7), 448.

McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459), 1387–1393.

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical population biology*, **75**(1), 35–45.

Müller, N. F. and Bouckaert, R. R. (2020). Adaptive metropolis-coupled mcmc for beast 2. *PeerJ*, **8**, e9473.

Müller, N. F., Stolz, U., Dudas, G., Stadler, T., and Vaughan, T. G. (2020). Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences*, **117**(29), 17104–17111.

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Current opinion in genetics & development*, **12**(6), 657–663.

Neches, R. Y., McGee, M. D., and Kyrpides, N. C. (2020). Recombination should not be an afterthought. *Nature Reviews Microbiology*, **18**(11), 606–606.

Nickbakhsh, S., Ho, A., Marques, D. F., McMenamin, J., Gunson, R. N., and Murcia, P. R. (2020). Epidemiology of seasonal coronaviruses: establishing the context for the emergence of coronavirus disease 2019. *The Journal of infectious diseases*, **222**(1), 17–25.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, **6**(1), 7–11.

Posada, D. and Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution*, **54**(3), 396–402.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet*, **10**(5), e1004342.

Reusken, C. B., Messadi, L., Feyisa, A., Ularamu, H., Godeke, G.-J., Danmarwa, A., Dawo, F., Jemli, M., Melaku, S., Shamaki, D., *et al.* (2014). Geographic distribution of mers coronavirus among dromedary camels, africa. *Emerging infectious diseases*, **20**(8), 1370.

Simon-Loriere, E. and Holmes, E. C. (2011). Why do rna viruses recombine? *Nature Reviews Microbiology*, **9**(8), 617–626.

Stadler, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of theoretical biology*, **261**(1), 58–66.

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 pathogenesis of coronaviruses. *Trends in microbiology*, **24**(6), 490–502.

Tagliamonte, M. S., Abid, N., Borocci, S., Sangiovanni, E., Ostrov, D. A., Kosakovsky Pond, S. L., Salemi, M., Chillemi, G., and Mavian, C. (2021). Multiple recombination events and strong purifying selection at the origin of sars-cov-2 spike glycoprotein increased correlated dynamic movements. *International Journal of Molecular Sciences*, **22**(1), 80.

Turakhia, Y., Thornlow, B., Hinrichs, A. S., Mcbroome, J., Ayala, N., Ye, C., De Maio, N., Haussler, D., Lanfear, R., and Corbett-Detig, R. (2021). Pandemic-scale phylogenomics reveals elevated recombination rates in the sars-cov-2 spike region. *bioRxiv*.

VanInsberghe, D., Neish, A. S., Lowen, A. C., and Koelle, K. (2020). Identification of sars-cov-2 recombinant genomes. *BioRxiv*.

Varabyou, A., Pockrandt, C., Salzberg, S. L., and Pertea, M. (2021). Rapid detection of inter-clade recombination in sars-cov-2 with bolotie. *Genetics*, **218**(3), iyab074.

Vaughan, T. G. (2017). Icytree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, **33**(15), 2392–2394.

Vaughan, T. G., Welch, D., Drummond, A. J., Biggs, P. J., George, T., and French, N. P. (2017). Inferring ancestral recombination graphs from bacterial genomic data. *Genetics*, **205**(2), 857–870.

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.* (2021). Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity. *Cell*, **184**(1), 64–75.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, **181**(2), 281–292.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wilkins, D. (2019). gggenes: draw gene arrow maps in 'ggplot2'. r package version 0.4. 0.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**(4), 1641–1650.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, **111**(46), 16448–16453.

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 related to sars-cov-2 contains natural insertions at the s1/s2 cleavage site of the spike protein. *Current Biology*, **30**(11), 2196–2203.

530 Supplementary material

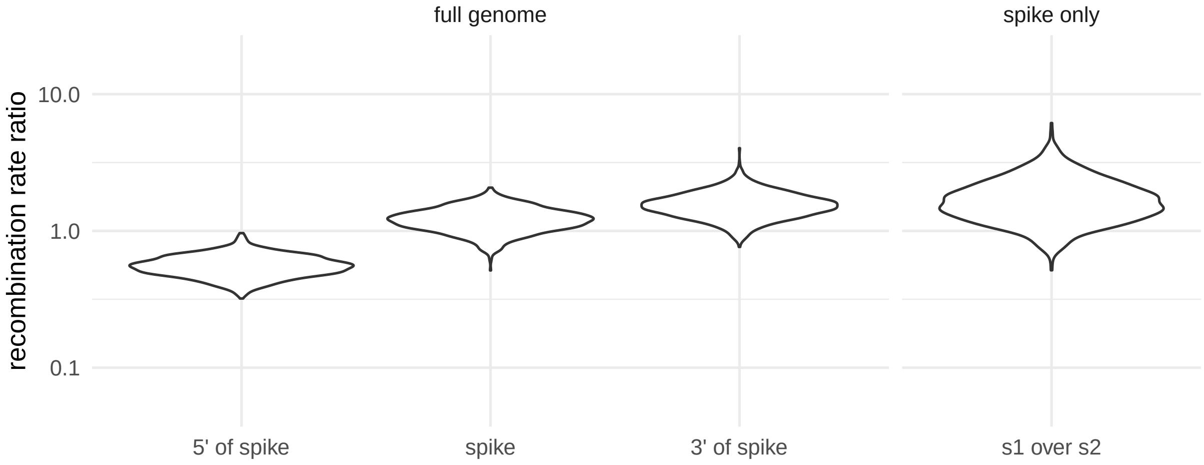

Figure S1: **Recombination rate ratios of SARS-like viruses on different parts of the genome.** Recombination rate ratios for SARS-like viruses based on two different analyses: one using the full genome (left) and one using the spike protein only (right). The rate ratios denote the rate on a part of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

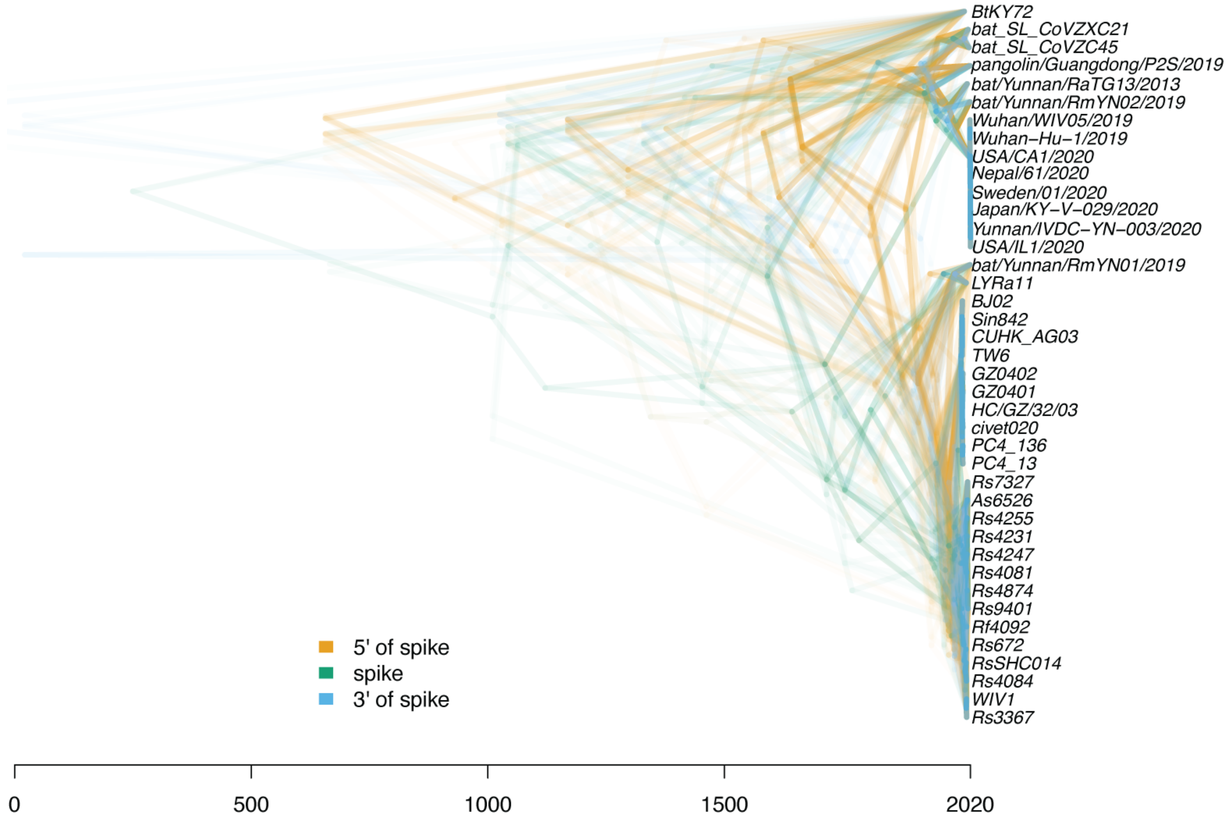

Figure S2: **Plot of the local trees of SARS-like virus on different positions across the genome.** Densitree (Bouckaert, 2010) plot of local trees in the mcc network of SARS-like viruses. The local trees are shown for every 100th position in the genome and are computed from the mcc network shown in 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 spike itself.

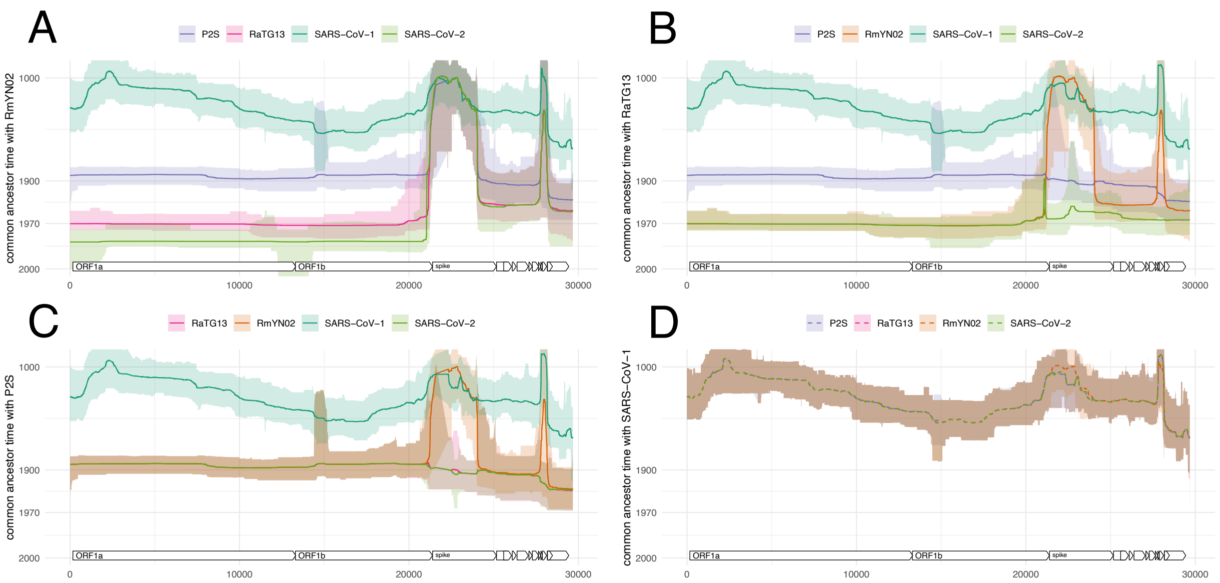

Figure S3: **Common ancestor times between sequences of the SARS-CoV-2 clade, as well as SARS-CoV-1.** Estimate of common ancestor times of RmYN02 (**A**), RaTG13 **B**, P2S **C** and SARS-CoV-1 **D** with each other and with SARS-CoV-2. The estimates of the common ancestor times assume an evolutionary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

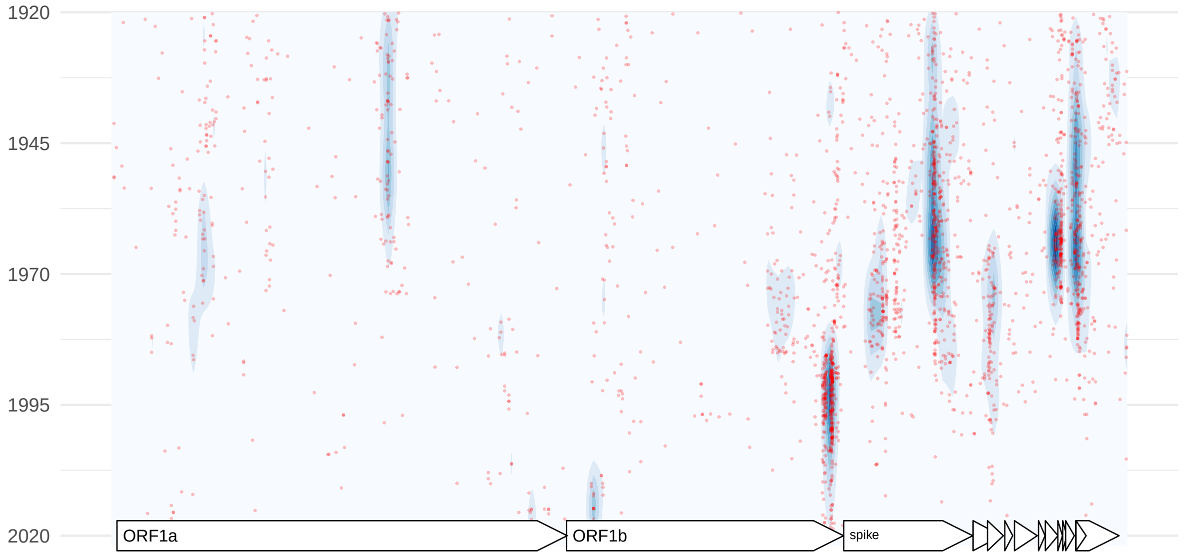

Figure S4: **Inferred timings and locations of recombination events ancestral to SARS-CoV-2 in the last one hundred years.** Timings and positions of inferred recombination events ancestral to the SARS-CoV-2 lineage are plotted. Each red dot denotes one event in the posterior distribution with the genome position on the x-axis and the year on the y-axis. The density of these events is shown by a contour plot.

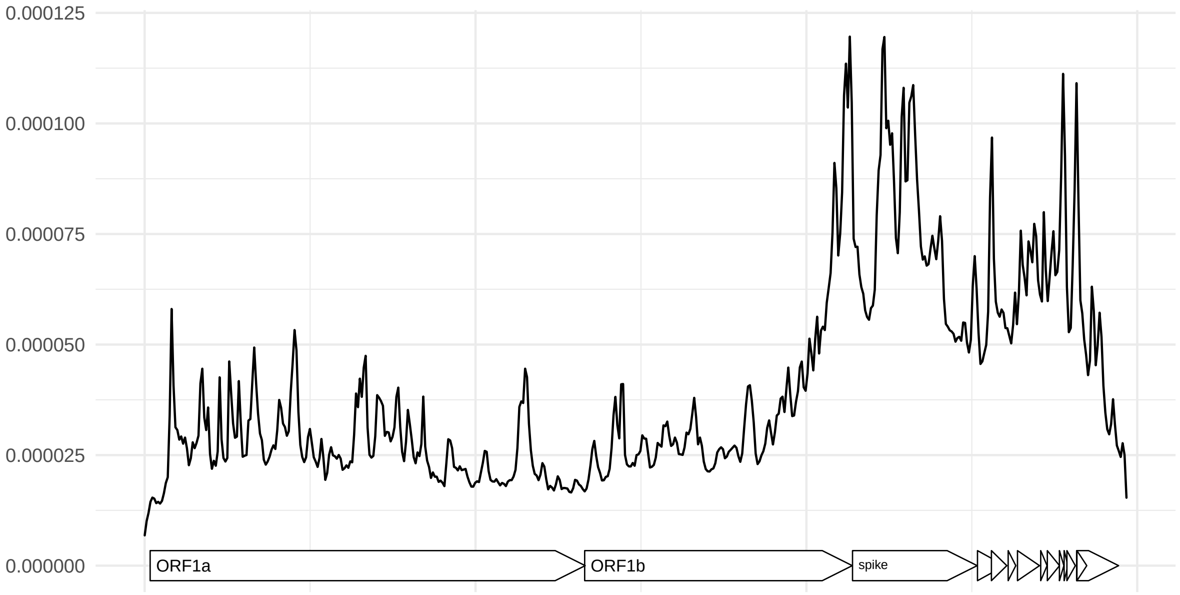

Figure S5: **Inferred locations of recombination events in the SARS-like dataset.** Here, we show the probability density of recombination events (on the y-axis) along the genome (on the x-axis) of SARS-like viruses.

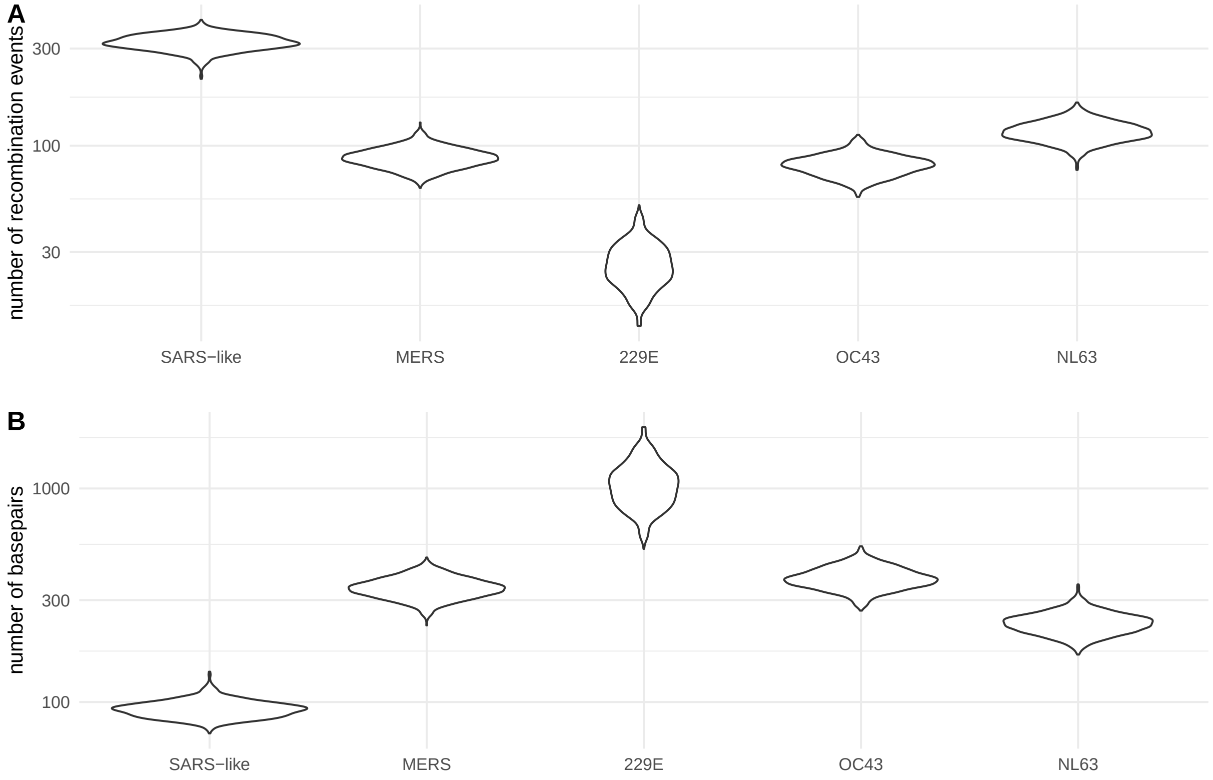

Figure S6: **Number of observable recombination events and average length of genomic segment coding for the same tree. A** Number of recombination events that impact the genome of sampled viruses in the dataset. **B** Average length of a segment in the genome of sampled viruses in the dataset that code for the same phylogenetic tree. That is the average length of a part of the genome that is not broken up by recombination events.

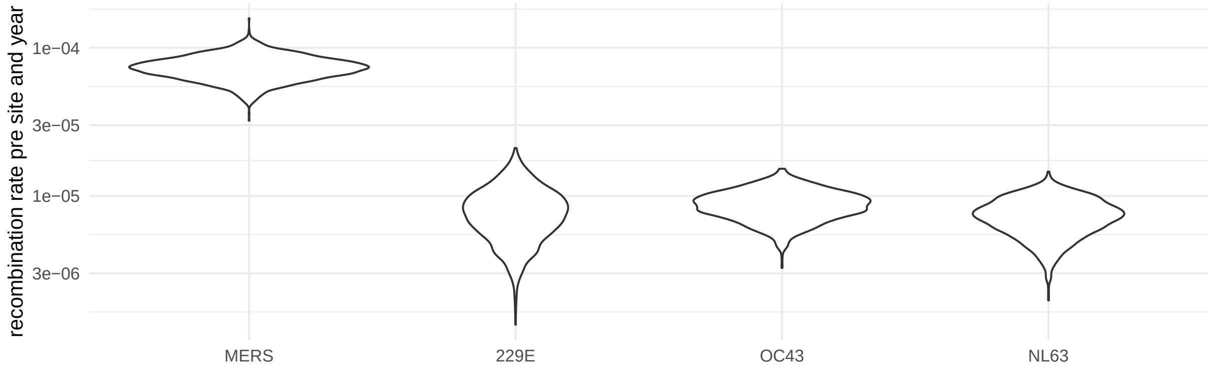

Figure S7: **Inferred recombination rates for the different coronaviruses.** Posterior distribution of recombination rates per year and per pair of adjacent nucleotides.

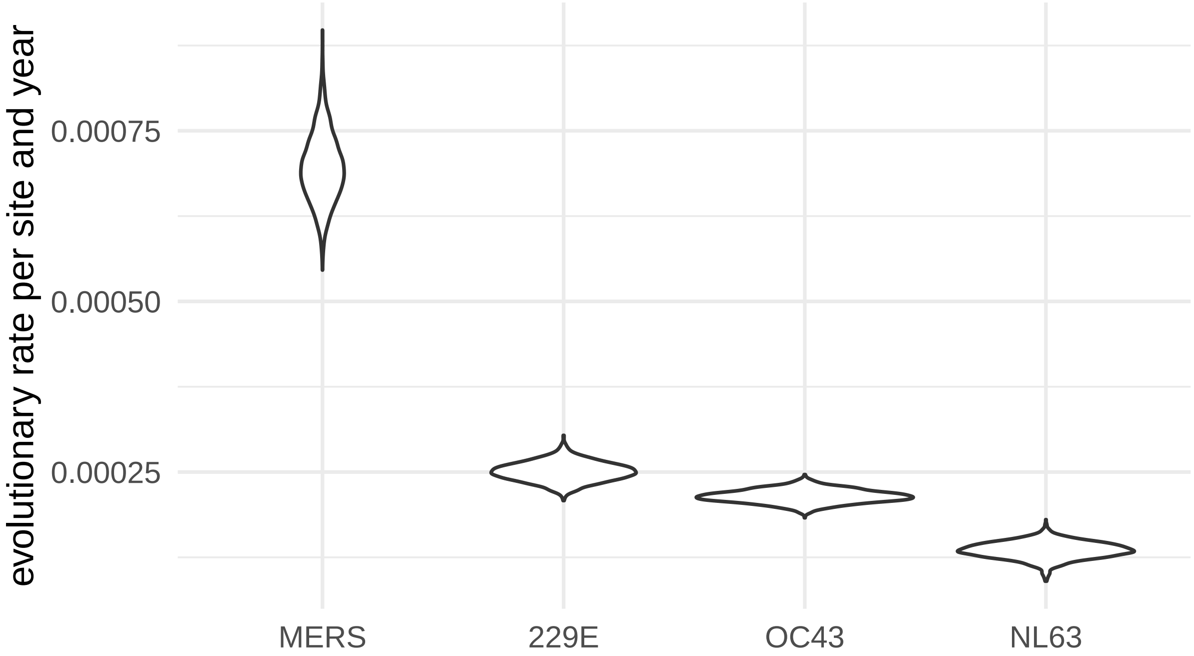

Figure S8: **Inferred evolutionary rates for the different coronaviruses.** Posterior distribution of evolutionary rates per year and nucleotide.

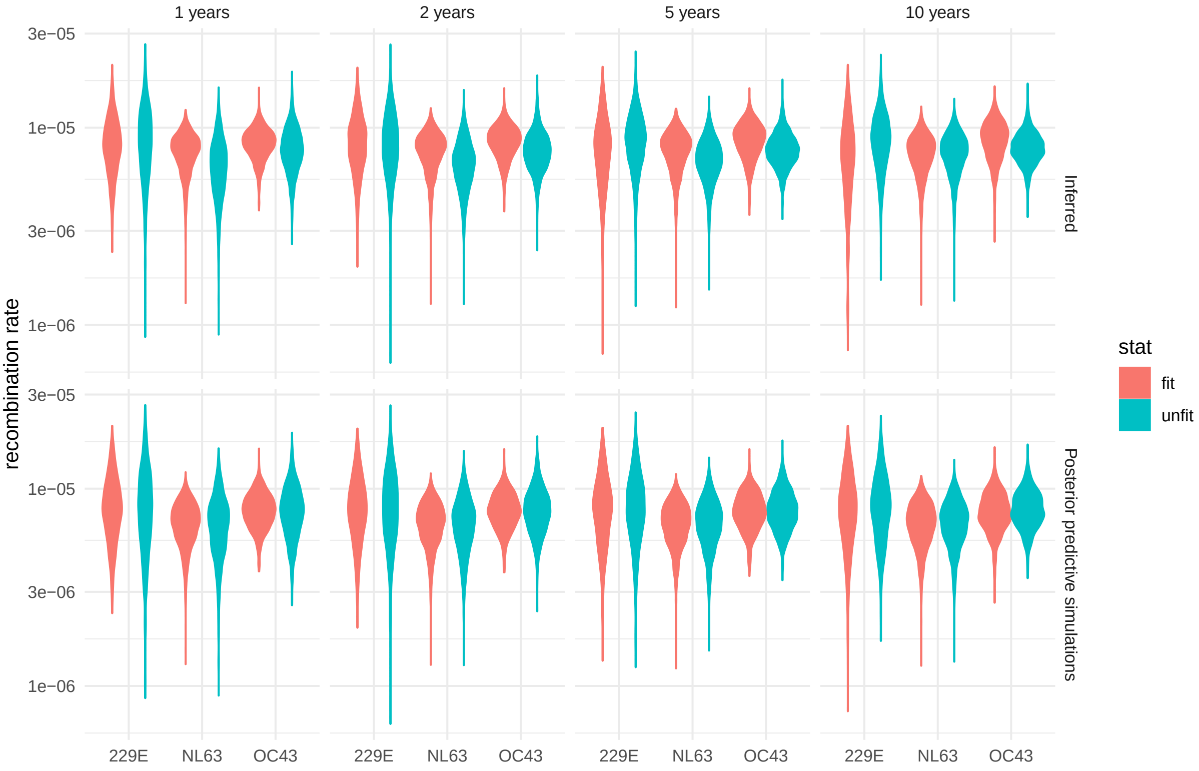

Figure S9: **Recombination rates of different parts of the recombination networks.** Recombination rates are computed for different parts of the network based on how long lineages persist for into the future. For this analysis, we classified each edge of the recombination network in the posterior distribution 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 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.

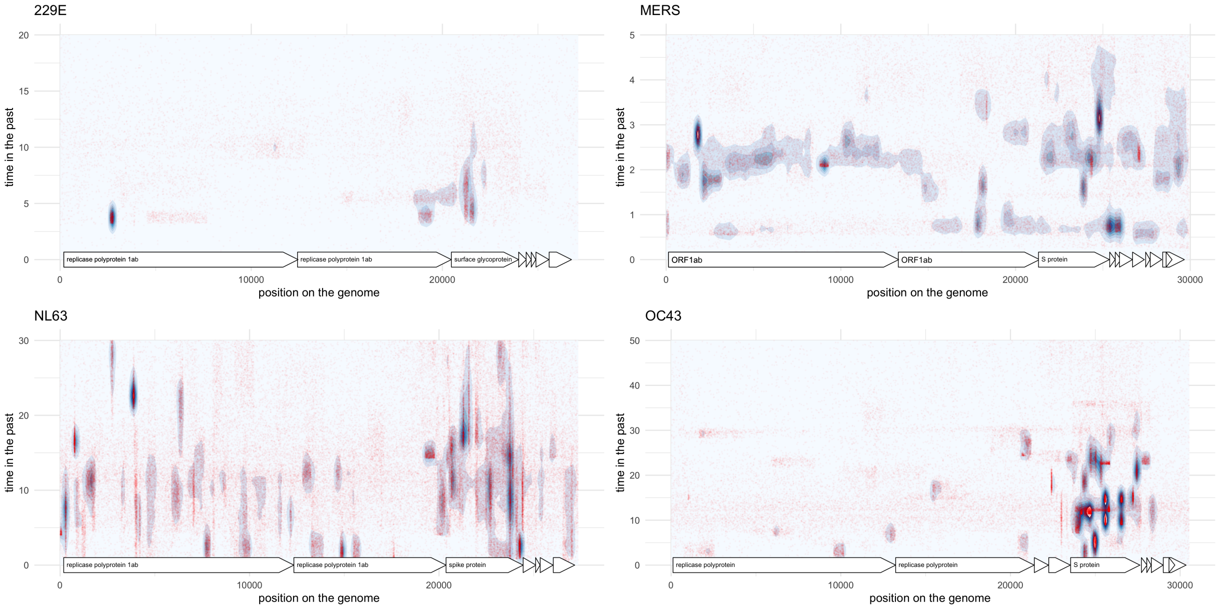

Figure S10: **Inferred timings and locations of recombination events of MERS, 229E, OC43 and NL63.** Each red dot denotes one event in the posterior distribution with the genome position on the x-axis and the year on the y-axis. The density of these events is shown by a contour plot.

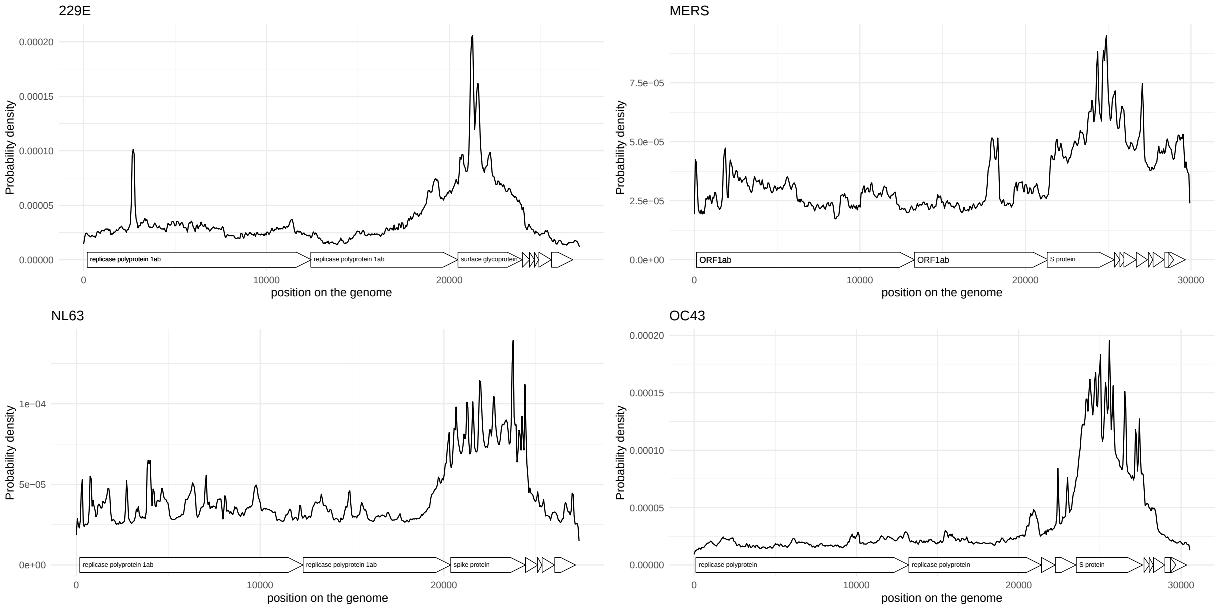

Figure S11: **Inferred locations of recombination events of MERS, 229E, OC43 and NL63.** Here, we show the probability density of recombination events (on the y-axis) along the genome (on the x-axis) for different coronaviruses.

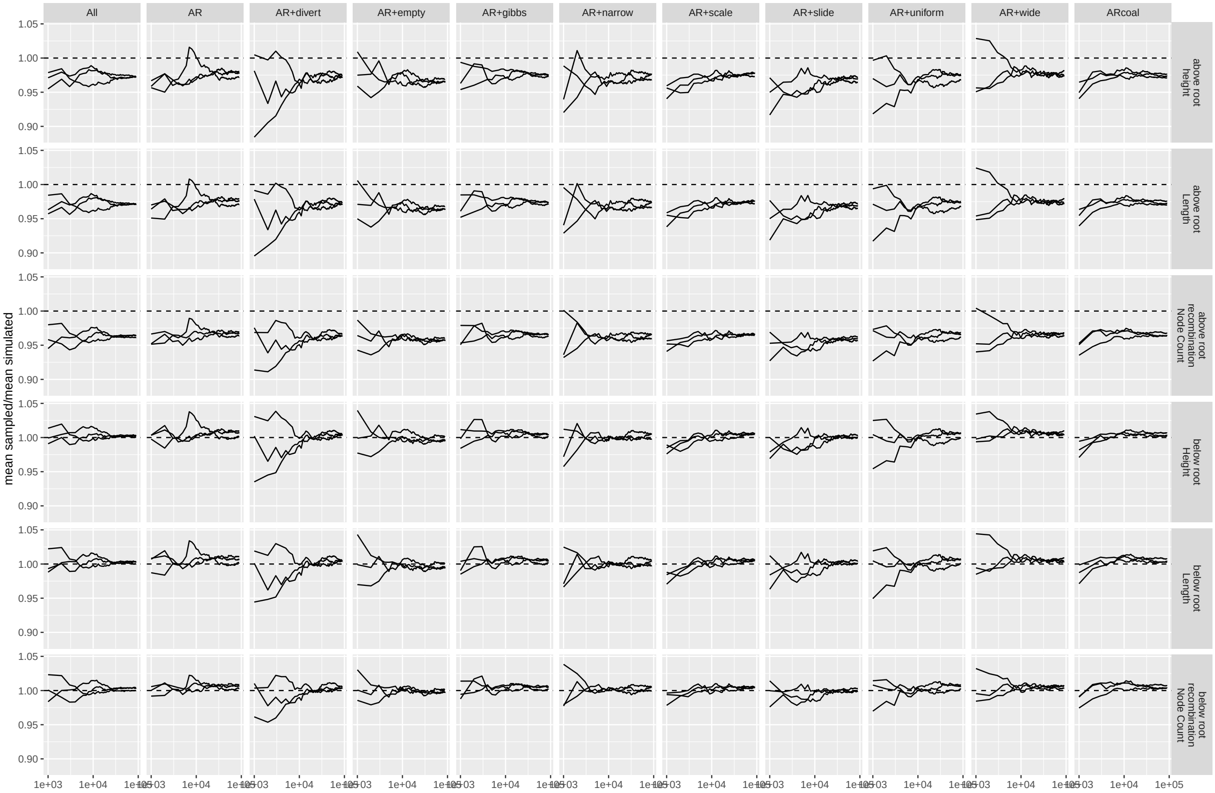

Figure S12: **Comparison of network statistics when simulating under the coalescent with recombination compared to sampling under the truncated coalescent with recombination.** We compare the posterior distributions of network height, length and the number of recombination nodes when simulating recombination networks under th 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.

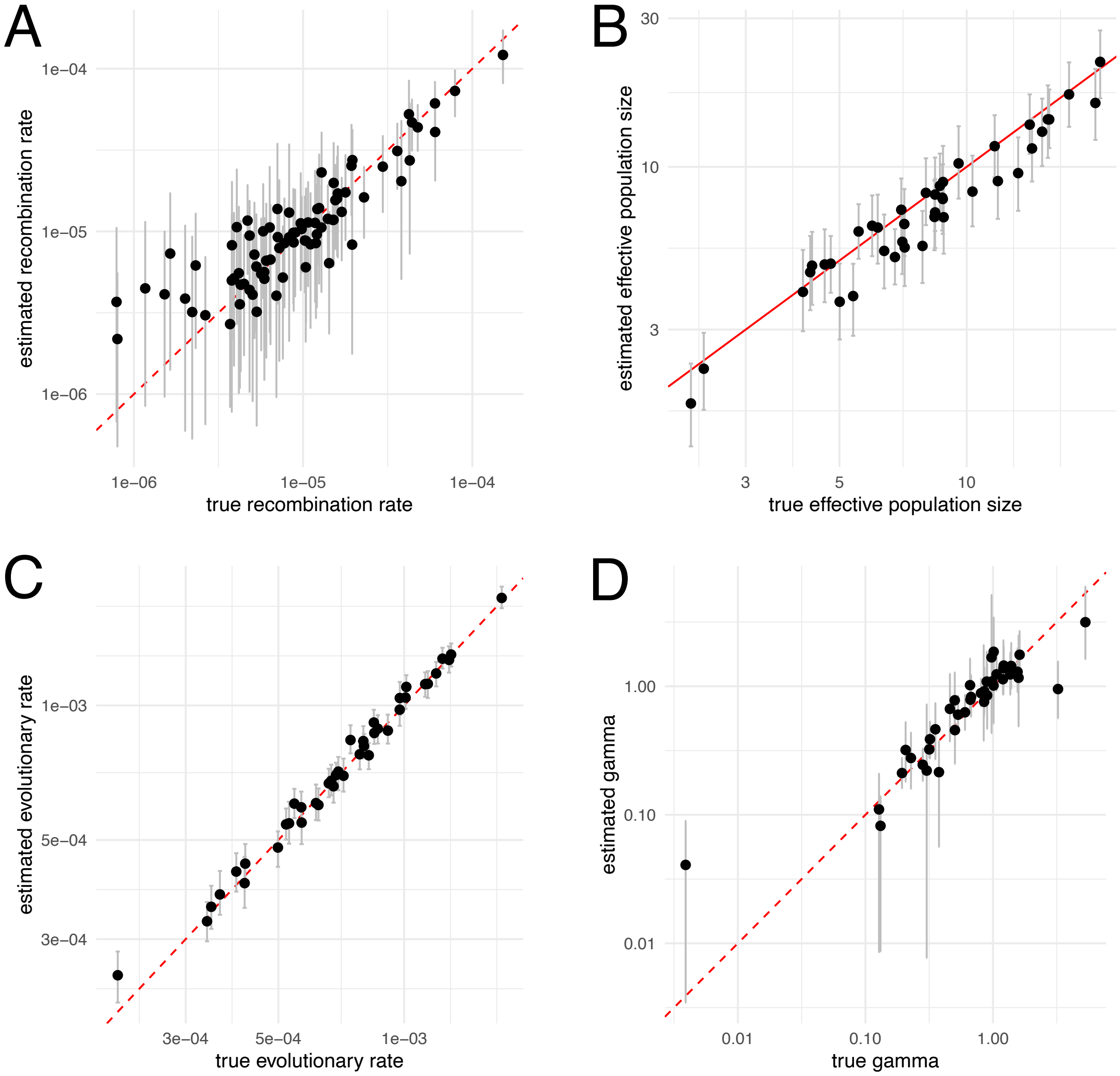

Figure S13: **Inferred vs. true rates based on simulated data.** We simulated recombination networks and sequence alignment using the randomly drawn values on the x-axis and then re-inferred these parameters on the y-axis. The size of the cross is scaled by the product of the recombination length and the amount of genetic information that recombined. The contour plot shows the distribution of inferred recombination events by location on the genome and time computed from the inferred posterior distribution of networks.

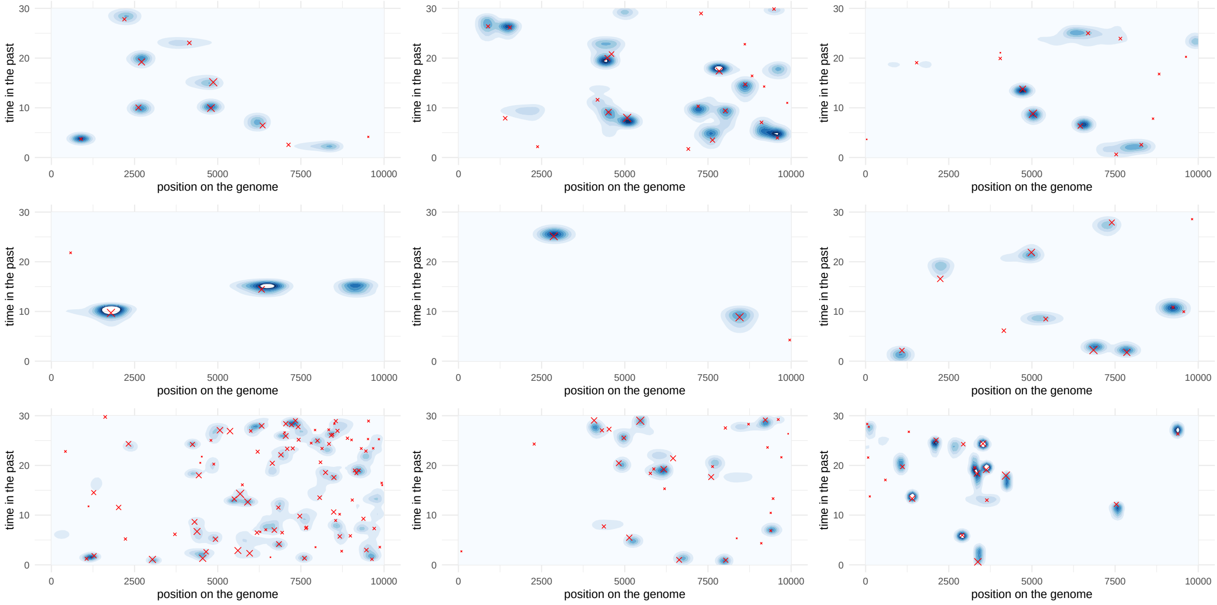

Figure S14: **Inferred vs. simulated recombination events.** Here, we show the recombination events for the simulated networks (red cross) with the position on the genome on the x-axis and the timing of the event on the y-axis. The contour plots show the density for inferred recombination events for the first 9 iterations of the simulation study. The time in the past is limited to the duration of sampling, i.e. the time when samples were taken.

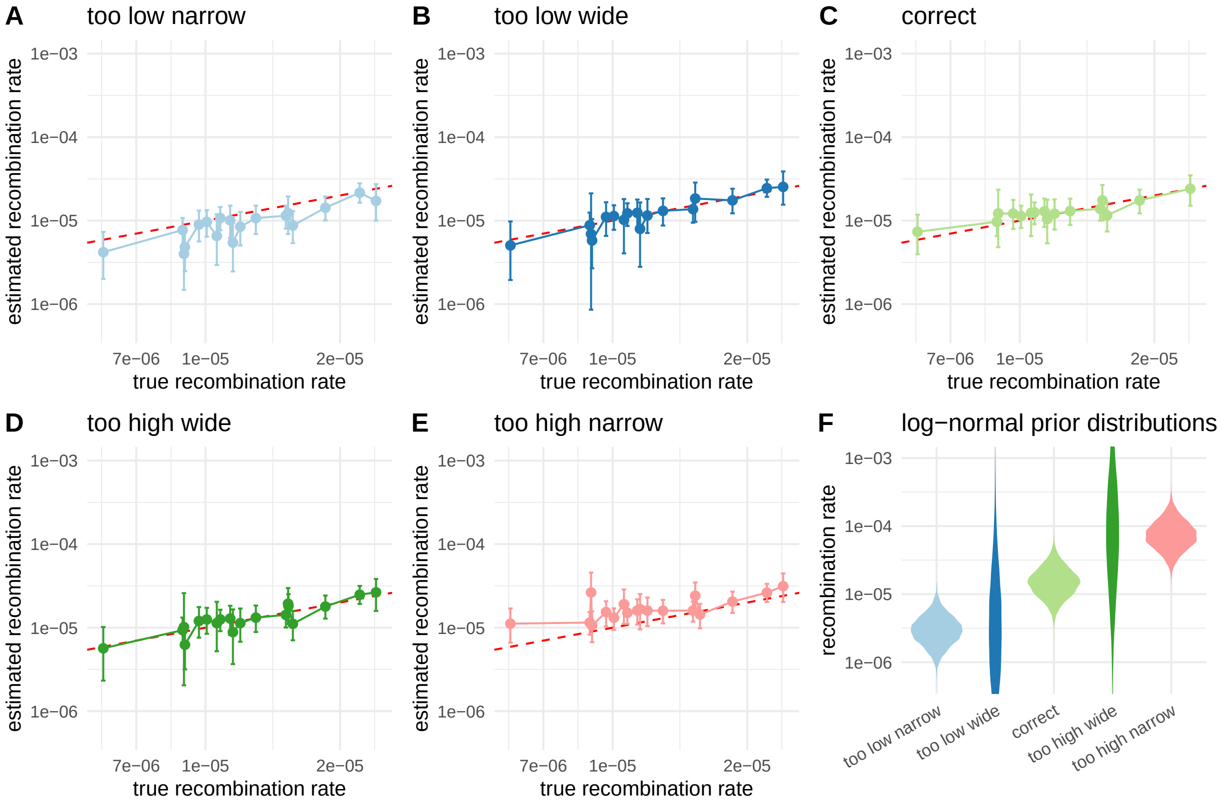

Figure S15: **Impact of the recombination rate prior distribution on the inferred recombination rates.** Here, we compare then inferred recombination rates when using different prior distributions that differed from the distributions from which the rates for simulations were sampled. The rates for simulations were sampled from a log-normal distribution with $\mu = -11.12$ and $\sigma = 0.5$. In **A**, we shows the inferred rates when using a prior distribution with $\mu = -12.74$ and $\sigma = 0.5$ (leading to a 5 times lower mean in real space than the correct prior). In **B**, we shows the inferred rates when using a prior distribution with $\mu = -12.74$ and $\sigma = 2$. In **C**, we shows the inferred rates when using the same prior distribution as was sampled under. In **D**, we shows the inferred rates when using a prior distribution with $\mu = -9.72$ and $\sigma = 2$. In **E**, we shows the inferred rates when using a prior distribution with $\mu = -9.72$ and $\sigma = 0.5$ (leading to a 5 times higher mean in real space than the correct prior). Figure **F** shows the corresponding density plots for all log normal distributions used as prior distributions on the recombination rates.

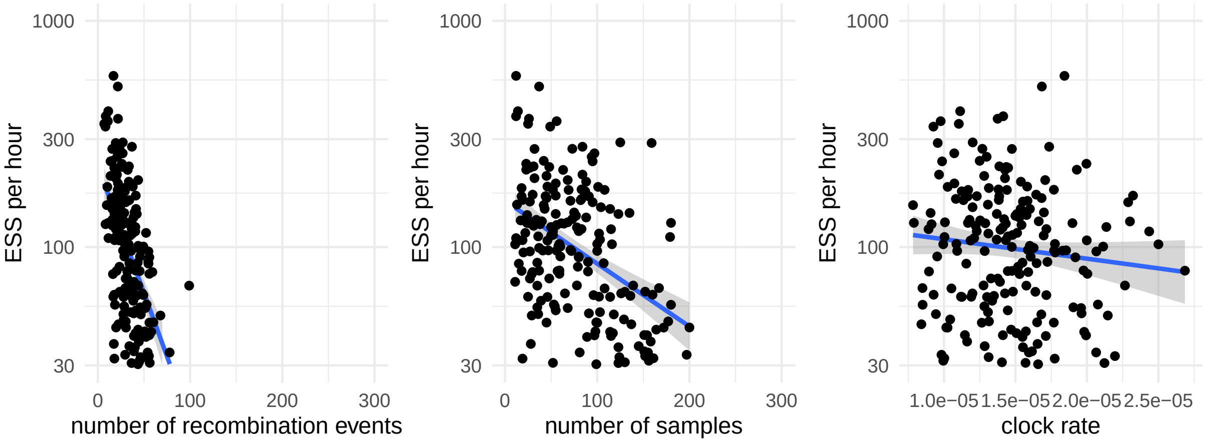

Figure S16: **Effective sample sizes per megasample.** We computed the effective sample size values computed using coda (Plummer *et al.*, 2006) for posterior probabilities, network/tree likelihood values, network/tree root heights and effective population sizes.

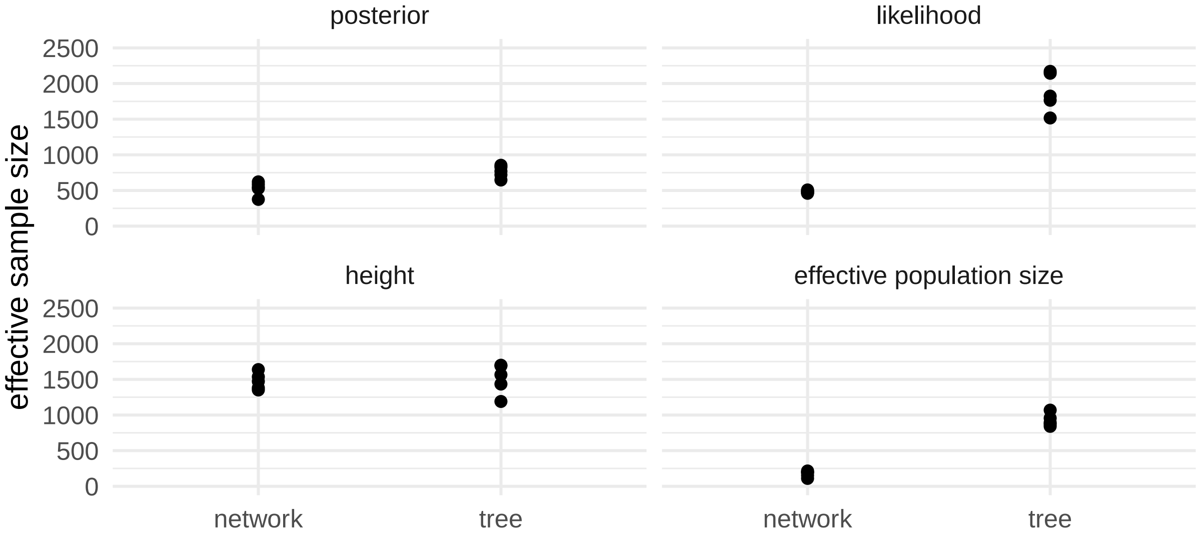

Figure S17: **Effective sample Sizes of MERS MCMC runs using the spike protein only.** Here, we compare ESS values after 25 Million MCMC iterations when inferring either networks or considering trees only for 100 MERS spike sequences. The operator weights for the inference of recombination networks is 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.