PNAS RESEARCH ARTICLE | MICROBIOLOGY
BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Correlated substitutions reveal SARS-like coronaviruses recombine frequently with a diverse set of structured gene pools

Asher Preska Steinberg[a], Olin K. Silander[b], and Edo Kussell[a,c,1]

Edited by Eugene Koonin, National Institutes of Health, Bethesda, MD; received April 26, 2022; accepted September 29, 2022

Quantifying SARS-like coronavirus (SL-CoV) evolution is critical to understanding the origins of SARS-CoV-2 and the molecular processes that could underlie future epidemic viruses. While genomic analyses suggest recombination was a factor in the emergence of SARS-CoV-2, few studies have quantified recombination rates among SL-CoVs. Here, we infer recombination rates of SL-CoVs from correlated substitutions in sequencing data using a coalescent model with recombination. Our computationally-efficient, non-phylogenetic method infers recombination parameters of both sampled sequences and the unsampled gene pools with which they recombine. We apply this approach to infer recombination parameters for a range of positive-sense RNA viruses. We then analyze a set of 191 SL-CoV sequences (including SARS-CoV-2) and find that ORF1ab and S genes frequently undergo recombination. We identify which SL-CoV sequence clusters have recombined with shared gene pools, and show that these pools have distinct structures and high recombination rates, with multiple recombination events occurring per synonymous substitution. We find that individual genes have recombined with different viral reservoirs. By decoupling contributions from mutation and recombination, we recover the phylogeny of non-recombined portions for many of these SL-CoVs, including the position of SARS-CoV-2 in this clonal phylogeny. Lastly, by analyzing >400,000 SARS-CoV-2 whole genome sequences, we show current diversity levels are insufficient to infer the within-population recombination rate of the virus since the pandemic began. Our work offers new methods for inferring recombination rates in RNA viruses with implications for understanding recombination in SARS-CoV-2 evolution and the structure of clonal relationships and gene pools shaping its origins.

SARS-CoV-2 | recombination | coronavirus | phylogeny | RNA viruses

Significance

Quantifying the population genetics of SARS-like coronavirus (SL-CoV) evolution is vital to deciphering the origins of SARS-CoV-2 and pinpointing viruses with epidemic potential. While Bayesian approaches can quantify recombination for these pathogens, the required simulations of recombination networks do not scale well with the massive amounts of sequences available in the genomics era. Our approach circumvents this by measuring correlated substitutions in sequences and fitting these data to a coalescent model with recombination. This allows us to analyze hundreds of thousands of sample sequences, and infer recombination rates for unsampled viral reservoirs. Our results provide insights into both the clonal relationships of sampled SL-CoV sequence clusters and the evolutionary dynamics of the gene pools with which they recombine.

Recombination can enable viruses to rapidly adapt to selective pressures (1–4) and to avoid accumulation of deleterious mutations that can lead to viral decline and extinction (5–7). Positive-sense single-stranded RNA ((+)ssRNA) viruses display highly variable levels of recombination (8, 9), with some species such as *West Nile* and *Yellow fever viruses* showing scant evidence of recombination (10) and others such as those of the *Coronaviridae* family showing evidence of frequent recombination (11). During the ongoing COVID-19 pandemic, population genomics has played an invaluable role in tracking the spread of SARS-CoV-2 and its variants (12–14), as well as understanding correlations between genomic substitutions and transmission patterns (15–19). However, a quantitative, population genomics-based understanding of the relative contributions of recombination and mutation to the evolution of SARS-CoV-2 and other SARS-like coronaviruses (SL-CoVs) is still being developed (20–27). Such knowledge will be important to understand the emergence of past and future viruses at the source of major epidemics.

The majority of tools for studying recombination in RNA viruses are phylogeny-based, where recombination breakpoints are assessed by examining phylogenetic incongruence and Bayesian and Markov chain Monte Carlo techniques are used to infer recombination parameters (20–25, 28). These approaches have been successful at identifying instances of recombination, yet their application to large-scale population genomics data remains challenging due to the computational demands of these methods. Importantly, the inferred recombination parameters rely only on the observed (i.e., sampled) sequences, while recombination within the much larger, unobserved gene pools with which these branches interact is not captured by these models. Here, to infer the recombination parameters of (+)ssRNA viruses, we adapt our non-phylogenetic, computationally-efficient *mcorr* method, which we originally developed to measure homologous recombination rates in bacteria (29–31). In contrast to previous approaches which focus on recombination within sampled sequences, we infer recombination parameters for both sampled sequences and the larger gene pools they recombine with, revealing that SL-CoVs recombine with a diverse set of gene pools which have high levels of recombination.

Author contributions: A.P.S., O.K.S., and E.K. designed research; A.P.S., O.K.S., and E.K. interpreted the results; A.P.S. and E.K. contributed new reagents/analytic tools; A.P.S. performed the bioinformatic and data analysis; A.P.S. wrote the code for measurement of correlated mutations and inference of recombination rates for RNA viruses used in this paper; and A.P.S., O.K.S., and E.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: edo.kussell@nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2206945119/-/DCSupplemental.

Published January 24, 2023.

Results

Using Correlated Substitutions to Infer Recombination Rates in RNA Viruses. Our primary aim is to infer the mutation and recombination rate within a set of sampled viral sequences, as well as the diversity and recombination rate of the unsampled pool of viruses with which the sampled viruses recombine. We model a "sample" of viral lineages with mean coalescence time $\overline{T}_{sample}$, which replicate, mutate (at rate μ), and recombine (at rate γ) with a much larger "pool" of lineages that have a mean coalescence time $\overline{T}_{pool}$ (Fig. 1*A*). We originally applied this model to infer recombination rates for bacteria (30), with the key difference here being introduced by the process of "copy-choice" recombination in RNA viruses (described in more detail below). The model predicts the conditional probability of a synonymous substitution at a genomic site $i + l$ given a substitution at site i, which we refer to as the "correlation profile," $P(l)$, where l is the distance between sites in nucleotides (nt). In a highly recombining viral population, this profile should decline rapidly as the distance between sites increases, while in a non-recombining population, this profile should be largely flat (see Fig. 1*B* for schematic) (29). To measure correlation profiles, we use sets of whole genome sequences (WGS) as our samples and use alignments of coding regions (CDS) to determine synonymous substitutions for all possible sequence pairs. For a sequence pair within the sample, we assign a binary variable σ_i a value of 1 for a substitution and a 0 for identity at position i (we refer to σ_i as the substitution profile). We exclusively consider third-position codon sites which are fourfold degenerate. The correlation profile is obtained by $P(l) \equiv P(\sigma_{i+l}|\sigma_i = 1)$, where the conditional probability is computed over all possible sequences pairs and averaged over all positions i.

The model has two free parameters which are determined by fitting the predicted $P(l)$ to its measured values from viral genome sequences. From the fit, we then calculate several useful quantities,

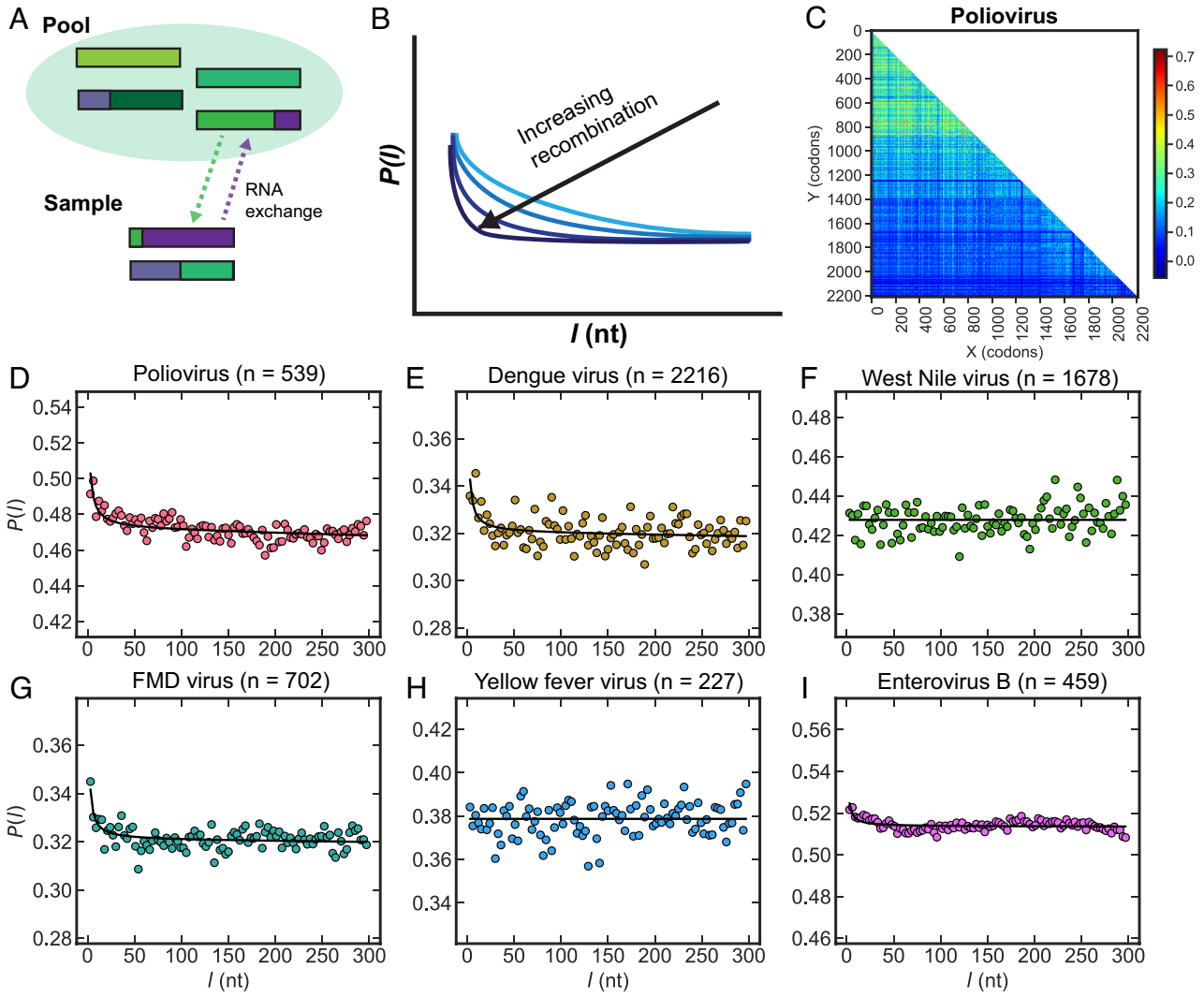

Fig. 1. Correlated substitutions in RNA viruses. (*A*) Schematic depicting the exchange of homologous RNA between a set of lineages (i.e., the "sample") and a larger "pool" of lineages via a copy-choice mechanism. (*B*) Schematic depicting different correlation profiles of synonymous substitutions with various levels of recombination. (*C*) A heatmap of Pearson's correlation coefficient ($\rho(X, Y)$) of synonymous substitutions across the coding region of the *Poliovirus* genome calculated using 539 *Poliovirus* genomes (see *Materials and Methods* for more details). Each position (X, Y) displays the corresponding value of $\rho(X, Y)$ as a color for a pair of codons located at genomic positions X and Y (given in codons). Color bar indicates the value of $\rho(X, Y)$. Monomorphic sites are assigned $\rho(X, Y) = 0$. (*D–I*) Correlation profiles of synonymous substitutions for a range of (+)ssRNA viruses. Markers correspond to the correlation profile $P(l)$ for a given separation distance l (given in nucleotides, *nt*). The fit is shown as a solid line and is performed under the assumption that only complete RNA strands are exchanged during template-switching events (i.e., we used the "template-switching model" described in *Materials and Methods* and the main text). Model selection was performed with the Akaike Information Criterion (AIC) to determine if a coalescent model with or without recombination best fit the data (see *Materials and Methods* for details). Parameters of homologous recombination are given in Table 1. n is the number of WGS analyzed. "FMD" virus stands for "Foot-and-mouth disease virus."

including the pool's mutational and recombinational divergence (given by $\theta_{pool} \equiv 2\mu\overline{T}_{pool}$ and $\phi_{pool} \equiv 2\gamma\overline{T}_{pool}$), which, respectively, correspond to the expected number of synonymous substitution and recombination events per site since coalescence of the pool; the sample's recombination coverage (c_{sample}), which is the fraction of sites that recombined since coalescence of the sample; the sample's mutational divergence ($\theta_{sample} \equiv 2\mu\overline{T}_{sample}$), which is proportional to the average age of the clonal (i.e., non-recombined) genomic portions of the sample; and the relative recombination rate of the pool ($\phi_{pool}/\theta_{pool} = \gamma/\mu$), which we denote as $(\gamma/\mu)_{pool}$. Unlike the synonymous substitutions, which we assume to be largely neutral, recombination events may be due to selective pressure or neutral drift.

Recombination in RNA viruses is thought to occur via a "copy-choice" mechanism, in which RNA-dependent RNA polymerase (RdRP) switches from one RNA template to another during RNA synthesis while remaining bound to the nascent RNA strand, creating RNA with hybrid ancestry (8, 9, 32). Here, we focus on copy-choice recombination resulting in homologous recombination; non-homologous or "illegitimate" recombination (i.e., insertion/deletion events) has previously been studied in SL-CoVs (23, 33) and RNA viruses (8) and is thought to be comparatively rare. This template-switching process occurs in the model at rate γ per site per viral replication (i.e., generation) and yields a hybrid viral genome consisting of a left arm from one genome and a right arm from another genome, joined at the recombination breakpoint (see *SI Appendix*, Fig. S1*A*). Experiments performed with murine hepatitis virus (a *betacoronavirus*) indicate that recombination can occur during negative or positive strand synthesis, and that template-switching events do not exclusively occur when two live viruses co-infect a cell but can also occur with transfected RNA fragments as small as 450 nt when RdRP switches from template to fragment (or vice versa) during synthesis (34). This latter scenario is similar to homologous recombination in bacteria, where DNA fragments of average size $\overline{f}$ are taken up by the cell at rate γ per site per generation and incorporated within a genome, replacing the homologous sites (see *SI Appendix*, Fig. S1*B*). In both the fragment incorporation model and the template-switching model, the predicted form of $P(l)$ depends on the total recombination rate (r) at a given locus, with $r = \gamma\overline{f}$ in the fragment-incorporation model and $r = \gamma L$ in the template-switching model, where L is the size of the genome (see *SI Appendix* for functional forms). We note that the fragment-incorporation model has an extra fitting parameter ($\overline{f}$).

We first analyzed correlated substitutions in *Poliovirus,* as this virus is known to have undergone substantial recombination during its evolution (9, 35–38). A genome-wide plot of the Pearson's correlation coefficient for all pairwise synonymous substitutions (which is the square root of the classic linkage disequilibrium metric 'r^2' but uses paired differences instead of allelic values (39)) across the CDS region of all major serotypes of *Poliovirus* (539 WGS used) shows that while substitutions tend to be more strongly correlated in the first ~800 codons, statistically significant correlations are found across the entire genome (Fig. 1*C*). Fitting the correlation profile for *Poliovirus*, we inferred the parameters of homologous recombination (Fig. 1*D* and Table 1). To estimate the range of these parameters, we calculated 95% bootstrap confidence intervals by sampling the 539 genomes with replacement to create bootstrap replicates (Table 1). Consistent with the literature, we found that *Poliovirus* has recombined substantially. We then proceeded to compute correlation profiles and infer recombination parameters for 12 other (+)ssRNA viruses (Fig. 1 *E*–*I*, Table 1, and *SI Appendix*, Fig. S2 and Table S1). We found results consistent with the literature, with viruses known to recombine showing evidence of recombination, e.g., *Dengue virus*, *Foot-and-mouth disease virus,* and *Enterovirus B* (9, 10, 40–49), while others where little or no recombination has been reported such as *West Nile* and *Yellow fever virus* (10, 50, 51) did not show signatures of recombination. We fit the correlation profiles in Fig. 1 using the template-switching model (Table 1) or the fragment-incorporation model (*SI Appendix*, Table S2) and found similar results; in particular, the predicted mean fragment size is generally on the order of the genome size and model selection does not strongly favor one model over the other. We therefore use the two-parameter template-switching model in all our analyses below.

Correlated Substitutions Show Evidence of Recombination Across Specific Genes in SARS-Like Betacoronaviruses. We used the 191 WGS used in the current *Nextstrain* build for SL-CoVs (52–54) and aligned these to the NCBI reference genome for SARS-CoV-2 (see *Materials and Methods* for details).

Table 1. Parameters of homologous recombination for viruses shown in main text

Virus	Gene	# of seqs	d_{sample}	θ_{pool}	ϕ_{pool}	$(\gamma/\mu)_{pool}$	$L(nt)$	Evidence ratio (w_t/w_2)
Poliovirus	Full genome	539	0.369 [0.359, 0.376]	1.29 [1.25, 1.29]	2.68 [2.42, 2.99]	2.11 [1.93, 2.32]	7.50E+03	9.96E+14
Dengue virus	Full genome	2216	0.256 [0.252, 0.260]	0.559 [0.543, 0.575]	2.93 [2.82, 3.04]	5.24 [4.97, 5.57]	1.10E+04	2.78E+04
West Nile virus	Full genome	1678	0.109	n/a	n/a	n/a	n/a	1.14E-06
Foot-and-mouth disease virus	Full genome	702	0.290 [0.284, 0.295]	0.560 [0.548, 0.572]	3.21 [3.02, 3.41]	5.73 [5.37, 6.14]	8.30E+03	1.68E+08
Yellow fever virus	Full genome	227	0.224	n/a	n/a	n/a	n/a	7.54E-03
Enterovirus B	Full genome	459	0.509 [0.506, 0.509]	1.63 [1.61, 1.64]	10.2 [9.37, 11.1]	6.25 [5.76, 6.79]	7.40E+03	3.89E+06
SL-CoV	orf1a	191	0.102 [0.0768, 0.126]	0.460 [0.400, 0.516]	2.04 [1.74, 2.34]	4.45 [4.03, 4.84]	3.00E+04	2.98E+07
SL-CoV	orf1b	191	0.0976	n/a	n/a	n/a	n/a	6.61E+00
SL-CoV	S protein	191	0.132 [0.101, 0.161]	0.561 [0.512, 0.605]	1.13 [0.915, 1.37]	2.02 [1.73, 2.35]	3.00E+04	3.91E+07

If the entire genome was fit, the gene is listed as "full genome." "SL-CoV" refers to the SARS-like betacoronaviruses. "orf1a" refers to the orf1ab CDS region before the -1 ribosomal frameshift and "orf1b" refers to the region after this frameshift. Parameters are given as the values inferred from the data, followed by the 95% bootstrap CI in square brackets (see *Materials and Methods* for calculation). We used model selection with AIC to determine if the profile was better fit with a coalescent model with or without recombination (see *Materials and Methods*). For those profiles which were better fit with the coalescent model with recombination, we assumed that only template-switching occurs (i.e., we used the "template-switching model" described in *Materials and Methods*). For those profiles better fit by the model without recombination, coverage was set to $c = 0$ and no bootstrapping was performed. L is the length of the genome in the template-switching model, w_t/w_2 is the Akaike weight of the template switching model (w_t) over the weight of the model without recombination (w_2; see *Materials and Methods*). All other parameters are described in main text.

This included SL-CoVs from bats (BtCoVs), SARS-associated coronavirus or SARS-CoV-1 (SARS-CoV-1), and SARS-CoV-2 (SARS-CoV-2). We first examined whether there were hotspots of correlated substitutions along the length of the SL-CoV genome (Fig. 2 *A* and *B*) in light of various reports of recombination hotspots in the SL-CoV and SARS-CoV-2 genomes, some of which suggest a correlation between adaptation and recombination in these regions (20, 22–25). We found correlated substitutions across the genome, with what visually appeared to be an accumulation of correlated substitutions in the coding sequence (CDS) region of orf1ab preceding the -1 ribosomal frameshift and the spike protein (throughout the paper, we will refer to the CDS region of orf1ab before the frameshift as "orf1a" and that after as "orf1b").

We next calculated correlation profiles and inferred recombination parameters across each gene (Fig. 2 *C–E*, Table 1, and *SI Appendix*, Fig. S3 and Table S3) and found strong evidence for recombination in orf1a (Fig. 2*C*) and the spike protein (Fig. 2*E*). The CDS regions of the orf3a and N proteins also displayed evidence of recombination (*SI Appendix*, Fig. S3), yet the shape of the decay in correlations was not nearly as apparent as those shown in Fig. 2. We observed orf1b had a similar synonymous diversity (d_{sample}) to the ORF regions adjacent to it (*SI Appendix*, Table S3), yet its correlation profile was flat (Fig. 2*D*); we hypothesize that the template-switching events occurring in the adjacent ORFs swapped out the entire orf1b CDS region, which would confer high diversity and a lack of recombination breakpoints. One CDS region which showed distinct patterns of correlated substitutions that cannot be adequately described by our model is that of orf8 (see *Discussion*). Overall, the inferred parameters suggest that the genes which show evidence of recombination are recombining frequently (Table 1 and *SI Appendix*, Table S3); when considering $(\gamma/\mu)_{pool}$ inferred for the orf1a and S genes, the pools these samples have exchanged RNA with are rapidly recombining at rates ranging from ~2–5 recombination events per synonymous substitution.

Using every complete genome assembly for SARS-CoV-2 in the NCBI database (444,145 sequences at the time of this analysis), we measured correlated substitutions across the SARS-CoV-2 genome (Fig. 2*F*). In contrast to what we had observed in the SL-CoVs (Fig. 2*B*), we only detected very weak correlated substitutions across the SARS-CoV-2 genome. Correlation profiles across individual genes appeared to be largely flat (Fig. 2 *G–I* and *SI Appendix*, Fig. S4; inferred parameters in *SI Appendix*, Table S4); this included the CDS regions for orf1a and the spike protein (Fig. 2 *G–I*), which showed signatures of recombination in the SL-CoV dataset (Fig. 2 *C–E*). The pronounced difference in overall scale of $P(l)$ between Fig. 2 *C–E* and *G–I* reflects the differences in sample diversity between the SL-CoVs and SARS-CoV-2 (for these genes, $d_{sample} = 9.8 \times 10^{-2} - 1.3 \times 10^{-1}$ for the SL-CoVs and $d_{sample} = 6.0 \times 10^{-4} - 1.2 \times 10^{-3}$ for SARS-CoV-2). As new subvariants of SARS-CoV-2 arose during the peer review of this manuscript, we performed an updated analysis in July 2022 using *Nextstrain*'s subsampling of SARS-CoV-2 sequences from across the globe from the last 6 mo (*SI Appendix*, Fig. S5). The correlation profile measured across the genome of these sequences was still flat, with $d_{sample} = 1.2 \times 10^{-3}$.

Recent work has suggested that SARS-CoV-2 experiences rate heterogeneity across the genome (27, 55), with specific genomic positions across the phylogeny exhibiting elevated mutation rates for $G \to U$ and $C \to U$ transitions, possibly related to APOBEC and ROS activity (55). This could cause individual sites to become "saturated" (i.e., many identical mutations occurring at the same site across the tree) and specific genomic regions to exhibit anomalously high diversity, giving the appearance of recombination from a highly diverged source. If this effect were substantial, we would expect that the SARS-CoV-2 analysis presented in Fig. 2 and *SI Appendix*, Figs. S4 and S5 would show signatures of recombination, which they do not. To determine whether such effects impact our inference of recombination rates in other datasets, such as the SL-CoV dataset, we ran simulations using *phastSim* (56), which includes hypermutability models developed to simulate observed rate heterogeneity in SARS-CoV-2 (*SI Appendix*, Fig. S6; details in *SI Appendix*). In the simulations, we allowed a proportion of sites to be "hypermutable" and have highly elevated transition rates, with both the proportion of sites and the rates set to be equal to or exceed what has been estimated for SARS-CoV-2 (see *SI Appendix* for details). We found that while sliding window averages of synonymous diversity increased in both magnitude and variability as expected (*SI Appendix*, Fig. S6 *A* and *B*), the correlation profiles we measured were consistently flat, correctly indicating that no recombination had occurred (*SI Appendix*, Fig. S6 *C* and *D*). These simulations suggest that heterogenous mutation rates, at least over a range which is biologically relevant to SARS-CoV-2, do not confound our ability to infer recombination rates using correlated synonymous substitutions.

Clonal Structure of the SARS-Like Betacoronaviruses. We sought to understand if a sufficient clonal signal remained in the SL-CoV samples which could be used to elucidate clonal relationships. We began by measuring genome-wide pairwise synonymous diversity (d_{sample}) across the 191 SL-CoVs and clustering these sequences using the average linkage algorithm (57) to create a dendrogram (Fig. 3*A*; see *SI Appendix* for details). We then split this tree into 11 flat clusters, where SARS-CoV-1 and SARS-CoV-2 each consisted of distinct clusters and the BtCoVs were broken into several clusters. The non-singleton BtCoV clusters were generally composed of sequences collected during the same time period and from the same geographic area; as examples, cluster 5 was almost entirely composed of samples from bats collected near Hong Kong and Gaungdong between 2005 and 2011 (58, 59), and cluster 6 was primarily samples collected near Yunnan Province from 2011 to 2014 (60, 61). As previously suggested (21, 25), it appears that on average across the genome SARS-CoV-2 is most closely related to a sequence cluster of BtCoVs (labeled "BtCoV (cluster 1)" in the legend shown in Fig. 3*A*). These two SL-CoVs were collected from bats in Zhejiang Province between 2015 and 2017 (62). Additional information pertaining to individual clusters is given with the sequence metadata (provided as a supplemental file).

We then determined whether a statistically significant clonal signal remains in the sampled genomes by comparing the pool's diversity (d_{pool}), inferred from the correlation profile, to the sample's diversity (d_{sample}), which is measured from the sequencing data. In this case, we can use the mutational divergence (θ_{sample}), which is proportional to the age of clonal portions, as a measure of clonal divergence. We computed the difference between d_{pool} and d_{sample} with respect to the variability in our measurement of d_{sample}, a quantity which we refer to as the residual clonality (RC) effect size (see *SI Appendix*, Eq. **S3** and description in *SI Appendix*). For the 11 SL-CoV clusters, we first inferred recombination parameters for pairs of clusters (i.e., samples composed of sequence pairs in which neither sequence is from the same cluster). Fifty-one out of fifty-five cluster pairs showed evidence of recombination as determined by model selection (see *SI Appendix*). We then plotted the recombination coverage and θ_{sample} for these cluster pairs against the RC effect size (*SI Appendix*, Fig. S7) and determined that, when θ_{sample} was greater than $\sim 10^{-3}$, the RC effect size was generally < 1. Therefore, for this dataset, we are able to infer values of $\theta_{sample} < 10^{-3}$, while for cluster pairs having RC <1 we can confidently conclude only that $\theta_{sample} > 10^{-3}$.

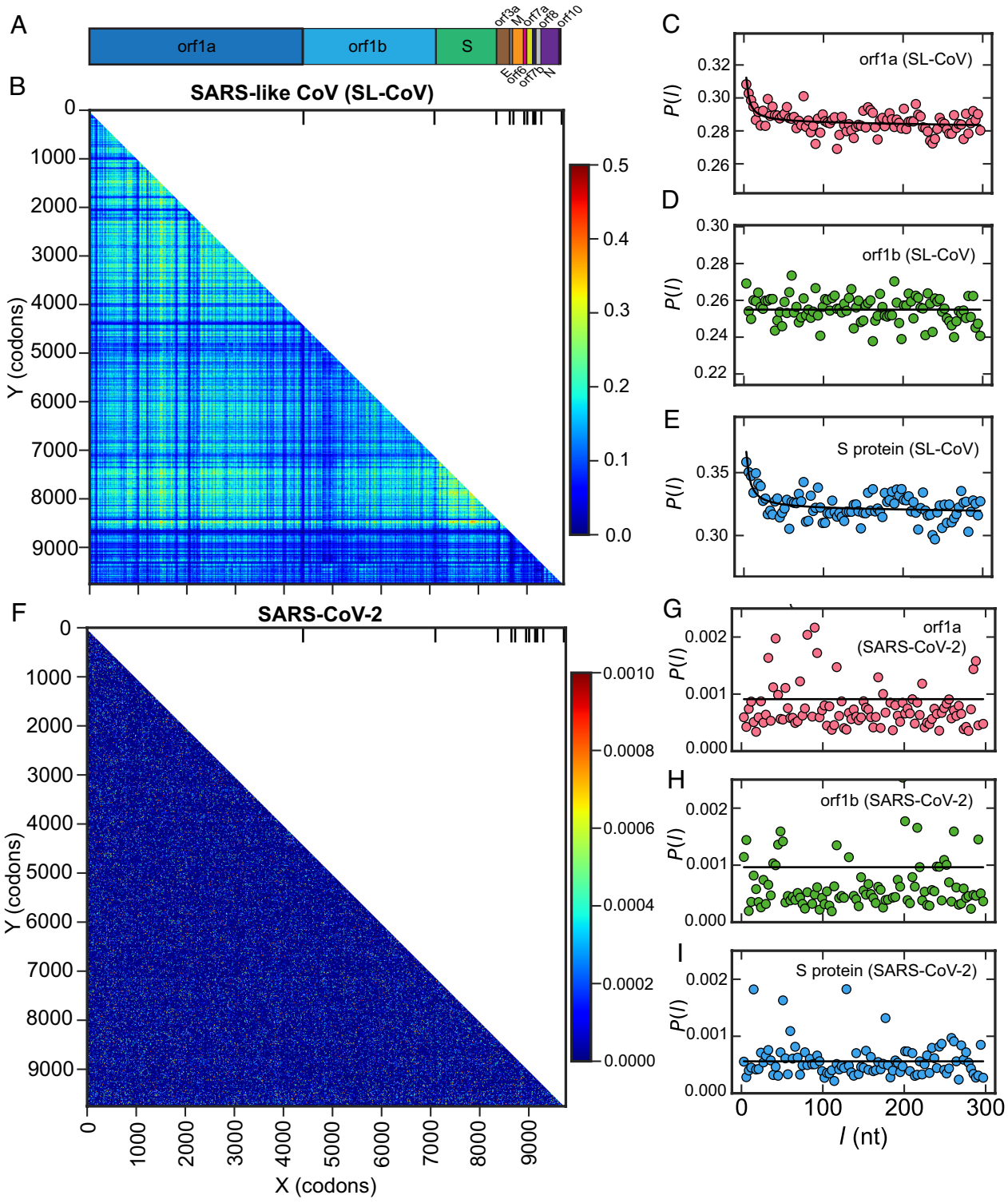

Fig. 2. Correlated substitutions in SARS-like betacoronaviruses and SARS-CoV-2. (*A*) A schematic of the SARS-CoV-2 genome. (*B*) A heatmap of Pearson's correlation coefficient ($\rho(X, Y)$) of synonymous substitutions across the coding regions of the SARS-like coronavirus genome constructed using 191 SARS-like betacoronavirus sequences (see *Materials and Methods* for more details). Each position (X, Y) displays the corresponding value of $\rho(X, Y)$ as a color for a pair of codons located at genomic positions X and Y (given in codons). Coding regions are ordered genomically. Color bar indicates the value of $\rho(X, Y)$. Ticks on the upper x-axis of the heatmap indicate where each CDS region begins and end, corresponding to the schematic of the SARS-CoV-2 genome above. "Orf1a" refers to the CDS region of orf1ab before the -1 ribosomal frameshift and "orf1b" refers to the CDS region after the frameshift. (*C*–*E*) Correlation profiles for the CDS regions of the orf1ab (*C* and *D*) and spike proteins (*E*) for SARS-like betacoronaviruses. Parameters of homologous recombination are given in Table 1. (*F*) A heatmap of $\rho(X, Y)$ (analogous to *B*) but for 444,145 SARS-CoV-2 whole genome assemblies from NCBI (all available assemblies when the analysis was conducted). (*G*–*I*) Correlation profiles for the CDS regions of orf1ab (*G* and *H*) and spike proteins (*I*) for SARS-CoV-2. Inferred parameters are given in *SI Appendix*, Table S4. In both heatmaps, monomorphic sites are assigned $\rho(X, Y) = 0$.

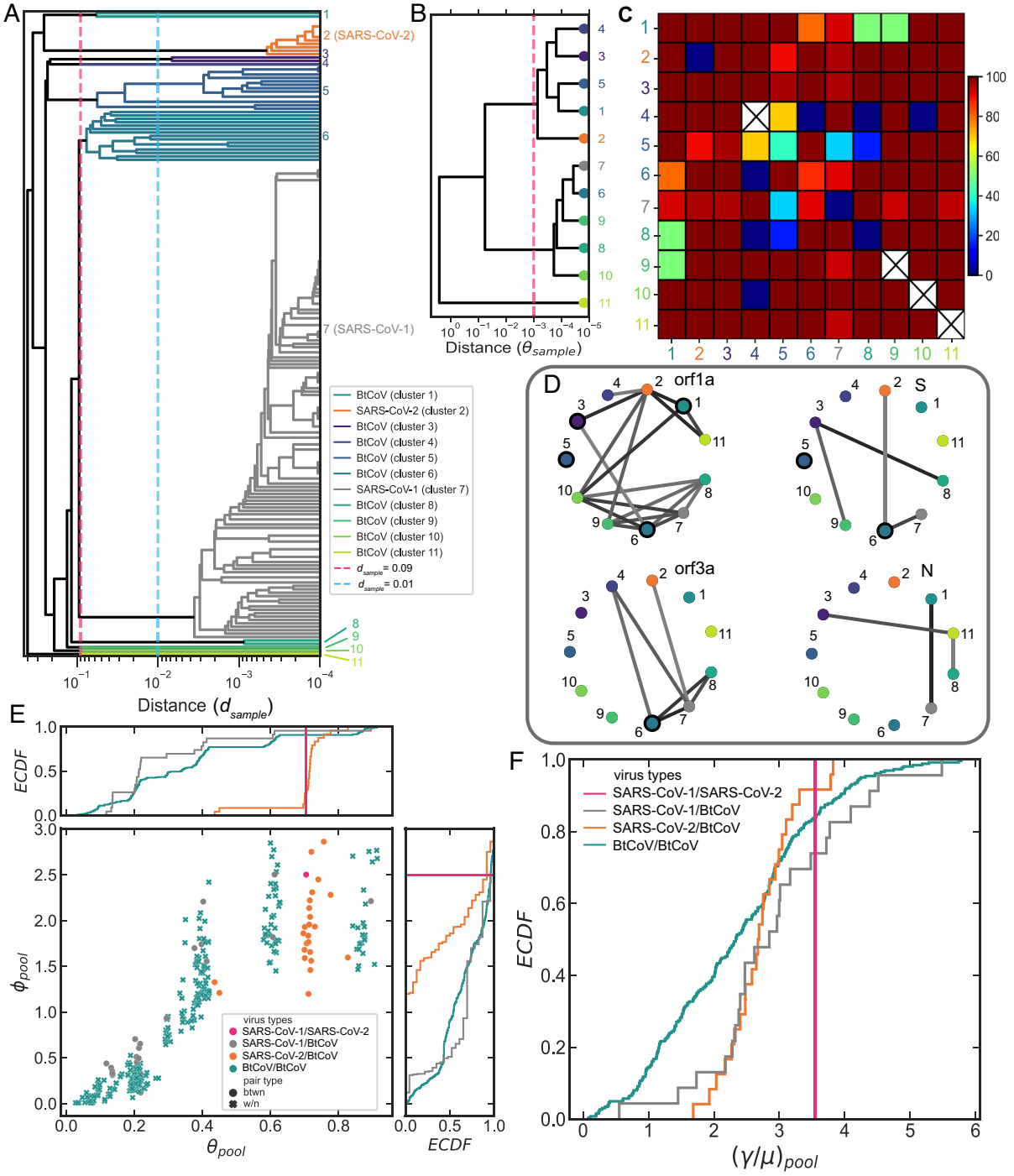

Fig. 3. Pairwise analysis of recombination among SARS-like betacoronaviruses. (*A*) Dendrogram of the 191 SARS-like betacoronavirus sequences analyzed in both Figs. 2 and 3. The tree was created using the average linkage algorithm with whole-genome pairwise synonymous diversity (d_{sample}) as the distance metric (see *SI Appendix*). The magenta, vertical dashed line depicts the distance at which the tree was cut to make the flat clusters shown in *B* and *C* (d_{sample} = 0.09). The branch colors correspond to these clusters, as does the legend. The blue, vertical dashed line depicts the cut at d_{sample} = 0.01made for the 27 flat clusters in *D* and *E*. Cluster numbers are shown along the vertical axis for the 11 flat clusters resulting from the cut made at d_{sample} = 0.09. Horizontal axis is log-scale. Clusters composed of SARS-like coronaviruses from bats are labeled "BtCoV." (*B*) Dendrogram of the 11 flat clusters from *A* created using the average linkage algorithm with θ_{sample} as the distance metric. The red dashed line at $\theta_{sample} = 10^{-3}$ indicates the maximum value beyond which the inference of θ_{sample} is obscured due to recombination from the pool (as described in the Main text and *SI Appendix*). (*C*) Heatmap depicting the percentage of total sequence pairs for a given pair of clusters which have recombined with a shared pool corresponding to the recombination network graph shown in *SI Appendix*, Fig. S9. Sequence pairs were determined to have recombined with a shared pool by computing correlation profiles across the whole genome and fitting these profiles to the model. Diagonal depicts recombination between sequence pairs within the cluster. Clusters with single sequences have crosses through their diagonal cells. (*D*) Recombination networks for individual genes computed using correlation profiles calculated across each gene for pairs of clusters. Nodes are the 11 clusters from *A* and *B*, edges connect cluster pairs which have recombined with a shared pool. Black halos around nodes indicate sequence pairs within the cluster have recombined. (*E* and *F*) Pool parameter distributions inferred from correlation profiles computed across the whole genome for pairs of sequence clusters. Clusters were made by cutting the dendrogram in *A* at d_{sample} = 0.01(depicted as vertical, blue dashed line), resulting in 27 flat clusters. Distributions are separated by virus type; those distributions in which both clusters are within the same virus type are denoted as "w/n," those which are between two virus types are denoted as "btwn." In panel *E,* the main plot shows the pool recombinational divergence (ϕ_{pool}) plotted against the pool mutational divergence (θ_{pool}). Marginal plots show *ECDF*s of each pair's divergence values. Panel *F* shows *ECDF*s of the relative recombination rate of the pool ($(\gamma/\mu)_{pool}$). All panels used the same fitting procedure as Figs. 1 and 2 (see *Materials and Methods*). For the recombination networks in *D*, if model selection suggested the profile was better fit with the null-recombination model, no edge was assigned for the cluster pair.

We constructed an average linkage tree based on θ_{sample} for the 11 SL-CoV clusters and demarcated the value above which θ_{sample} is not well-determined (Fig. 3*B*). We found that there is sufficient RC in the data to infer the clonal structure for most of the SL-CoV lineages (Fig. 3*B*), revealing key differences with respect to the dendrogram based on genome-wide pairwise distances (Fig. 3*A*). While the tree in Fig. 3*A* suggests that SARS-CoV-2 shares its most recent common ancestor (MRCA) with the BtCoVs of cluster 1, the clonal tree in Fig. 3*B* indicates that SARS-CoV-2 actually shares a MRCA with clusters 1, 3, 4, and 5, and moves clusters 3-5 farther from SARS-CoV-1 than SARS-CoV-2. Further, the timescales for these MRCAs are dramatically different, with the clonal tree based on θ_{sample} indicating a much more recent split.

To determine whether standard phylogenetic methods which assume no homologous recombination recover this clonal structure, we used *IQ-Tree* (63), which relies on maximum likelihood estimation for phylogenetic inference, to reconstruct the phylogeny for the sequences depicted in Fig. 3*A* using a generalized time reversible (GTR) model (*SI Appendix* and Fig. S8). We found that the structure of the tree matched Fig. 3*A*, but not that of Fig. 3*B*. This may suggest that, for the SL-CoVs, standard phylogenetic inference yields phylogenies which are likely obscured by recombination; the resultant phylogenies reflect both recombination from the pool and mutations within the sample.

Correlated Substitutions Reveal the Gene Pool Structure of SARS-Like Betacoronaviruses. To determine which members of the SL-CoVs have recombined with shared gene pools, we calculated correlation profiles across the whole genome for all possible sequence pairs and fit each profile to the model. We then visualized recombination between sequence pairs as a network graph (*SI Appendix*, Fig. S9), where each strain is a node, and edges connect strains which have recombined with a shared pool. Visually, the network appears to be highly connected, suggesting that many pairs have recombined with shared gene pools. To quantify network connectivity, we created a matrix of the percentage of sequence pairs which have recombined with a shared pool for a given cluster pair (Fig. 3*C*). This analysis reveals that the clusters are less interconnected than they appear, as not all clusters share pools equally; and in some cases, we find clusters that may not share the same gene pool at all. This indicates that distinct, structured gene pools exist despite a high degree of recombination and gene pool sharing across the SL-CoV lineages.

Because our analysis of the SL-CoVs revealed that individual genes have different recombination parameters, we tested whether the four genes which showed recombination signatures in *SI Appendix*, Fig. S3 each had distinct networks of recombination events. For each of the four genes, we measured correlation profiles for pairs of sequence clusters and used these profiles to determine if the cluster pair showed evidence of recombination with a shared pool (Fig. 3*D*). As profiles measured over single genes account for fewer genomic sites compared to profiles measured over whole genomes, the recombination signal for individual genes needs to be stronger relative to random correlations to allow for detection; this accounts for the slight discrepancies between Fig. 3 *C* and *D*. Each of the genes had a unique recombination network, with the most recombination occurring in orf1a. Furthermore, the analysis suggested SARS-CoV-1 and SARS-CoV-2 both underwent recombination events with the same pool in the orf3a CDS region.

To further examine the structure of the SL-CoV gene pools, we cut the tree in Fig. 3*A* at $d_s = 0.01$ (yielding 27 clusters) and computed correlation profiles for each cluster and cluster pair across the entire genome and inferred their recombination parameters. We first investigated the degree of clonality of each sample (a cluster or cluster pair) by computing its RC effect size (see *Methods* and previous section). As we have a larger sample distribution (i.e., more cluster pairs) than in the previous section, we can adopt an even stricter criterion for the RC effect size here; if we specify that the RC effect size must be greater than 2 to infer θ_{sample}, by plotting θ_{sample} versus the RC effect size we determined that 82% of samples had a sufficient RC effect size (*SI Appendix*, Fig. S10). For these samples, θ_{sample} ranged from 1.8e-5 to 3.8e-3 (we note the upper bound is similar to that used in the previous section), with a median of 1.6e-4, while c_{sample} ranged from 27 to 97% with a median of 87% (*SI Appendix*, Fig. S10). For the remaining 18% of samples where RC effect size ≤ 2, recombination coverage was comparatively higher, with c_{sample} ranging from 92% to 100%, and a median of 98%, and for these samples, we estimate a lower bound of $\theta_{sample} \sim$ 3.6e-4. Across all SL-CoV samples, we find the median θ_{sample} = 1.9e-4 and median c_{sample} = 92%. To test that the inferred parameters provide reasonable estimates for θ_{sample} and c_{sample}, we took several sequence pairs with varying levels of c_{sample} and looked at sliding window averages of diversity ($\overline{\sigma}_X$) across the genome (*SI Appendix*, Fig. S11). Our estimates of θ_{sample} suggest that clonal regions should exhibit diversity levels in the range $\sim$ 1e-5 to 1e-3, and we found that the genomic fraction with $\overline{\sigma}_X$ in this range roughly matched the inferred clonal fraction, $1 - c_{sample}$. The remainder of the genome has been recombined from the pool, and we find that the average diversity of these regions is close to our estimate of the pool diversity. Moreover, we can use the distribution of zero-SNP block lengths in these sequence pairs to provide an alternative estimate of c_{sample}, and find that this roughly matches the values inferred from the coalescent model (*SI Appendix*, Fig. S12; see *SI Appendix* or details). We conclude that despite the extensive recombination across the SL-CoV lineages, a substantial clonal signal remains in the data and a wide distribution of clonal divergence times, spanning at least two orders of magnitude, can be detected.

We then sorted each of the samples according to its virus type (e.g., a cluster pair in which one cluster is comprised of BtCoV sequences and the other is SARS-CoV-1 sequences is labeled "SARS-CoV-1/BtCoV") and examined the distributions of ϕ_{pool} and θ_{pool} across samples. We found that each pair of virus types had a distinct parameter distribution (Fig. 3*E*). Moreover, the divergence distributions are multi-modal (particularly the BtCoV/BtCoV distribution), suggesting that subsets of these samples interact with gene pools with distinct evolutionary dynamics. To remove the dependence on coalescence time, we plotted the distributions of $(\gamma/\mu)_{pool}$ for each cluster pair as empirical cumulative distribution functions to assess the relative recombination rates of these pools (Fig. 3*F*). This revealed that within the BtCoV lineages' pools, there is a wide and relatively uniform distribution of recombination rates, while each of the SARS-CoV-1 and SARS-CoV-2 recombine with shared BtCoV pools with similar characteristic rates, and exhibit narrower, unimodal distributions. Nearly all cluster pairs have $(\gamma/\mu)_{pool} > 1$, meaning that in SL-CoV gene pools there are multiple recombination events occurring per synonymous substitution, indicating that recombination plays a major role in the evolution of SL-CoVs.

Discussion

We adapted the non-phylogenetic, computationally-efficient *mcorr* method [originally developed for analysis of bacterial genomes (29, 30)] to infer the parameters of homologous recombination for SL-CoVs and other RNA viruses. The methodological advances reported here include the use of a two-parameter template-switching model, the introduction of the RC effect size as a

tool for clonal inference, the ability to analyze single genes, the measurement of correlation profiles across whole genomes (vs. gene-averaged profiles), and the improved efficiency of the method such that >400,000 sequences can be analyzed. We first demonstrated that this method is generally applicable to (+)ssRNA viruses by inferring recombination parameters for viruses which have known histories of recombination using datasets consisting of hundreds to thousands of WGS. We then applied this to understand recombination in SL-CoVs. We found strong signatures of recombination in the CDS regions of orf1a and the spike protein. While previous studies of SL-CoVs have yielded estimates of recombination rates among analyzed sample sequences (20) and others have suggested that SL-CoVs have recombined with unsampled pools (21, 24), here we infer recombination rates and parameters for both the sample sequences and the unsampled gene pools with which they recombine.

Our gene-by-gene analysis of recombination for the SL-CoVs revealed that orf1ab and the S protein show strong signatures of recombination and suggested that these parts of the genome recombine at high rates, ranging from ~2–5 recombination events per synonymous substitution (Table 1). Interestingly, when we fit the correlation profile for orf1a with the fragment-incorporation model, we found that the 95% bootstrap confidence interval for the mean fragment size ranges from ~11,000 to 30,000 nt, suggesting that the SL-CoVs may take up fragments via recombination, consistent with previous experimental observations with betacoronaviruses (34) (for additional discussion, see section on zero-SNP blocks in *SI Appendix*). We note that orf8 showed distinct patterns of correlated substitutions that cannot be adequately described by our model (*SI Appendix*, Fig. S3). This region is thought to be highly variable, to contain several stem-loops which could lead to correlations between distant sites, and to have undergone recombination (64–66). Furthermore, the RNA secondary structure in this region could lead to selection on synonymous sites resulting in codon bias (67). Therefore, we speculate that this combination of RNA secondary structure and recombination has left the orf8 gene with an uncharacteristic decay in correlated substitutions; additionally, the decay we observe could be impacted by the many deletions and nonsense substitutions in this region (65, 66). Because we cannot reliably infer parameters using our recombination model for this CDS region, we also cannot assume that evolution has proceeded clonally in this region, for orf8 we simply list d_{sample} for this gene (given in *SI Appendix*, Table S3).

By measuring correlated substitutions between SL-CoV sequence clusters and decoupling the contributions of mutation and recombination to the sample diversity, we were able to recover much of the clonal structure for the analyzed SL-CoV clusters (Fig. 3*B*). We show that due to both high recombination coverage and variability in the measurement of d_{sample}, the residual clonal signal in the data only allows sample ages up to $\theta_{sample} \sim 0.001$ to be determined. Nevertheless, this is sufficient to yield important insights into the clonal relationships between the sequence clusters. Our inference of clonal relationships indicates that SARS-CoV-2 shared an MRCA with BtCoVs from clusters 1, 3, 4, and 5 (Fig. 3*B*), whereas clustering based on d_{sample} (Fig. 3*A*) and standard methods for phylogenetic inference such as maximum likelihood estimation (*SI Appendix*, Fig. S8) would suggest that SARS-CoV-2 shares its MRCA with cluster 1 only. Further, this split is relatively recent in our inferred clonal tree (θ_{sample} <0.001) while the split predicted based on genome-wide synonymous diversity is much more distant (d_{sample} >0.1). Whereas phylogenetic methods attempt to identify ancestral relations by inferring or simulating individual recombination and mutation events that took place in different portions of the genome at different times in the past (see below), our approach determines the RC of a pair of clusters directly from its correlation profile. In this respect, our method of clonal inference is more direct; however, future analyses using broader sets of SL-CoV sequences, in combination with simulations and experiments, will be needed to determine when each approach may be most advantageous. By measuring correlated substitutions between individual sequence pairs, we inferred a recombination network for the SL-CoVs and found that the majority of sequences have recombined with a pool during their evolutionary history (*SI Appendix*, Fig. S9). Counting the number of recombined pairs (Fig. 3*C*) shows that some clusters appear to have only recombined with subsets of sequences from other clusters (e.g., cluster 5), indicating that while the SL-CoV recombination network is dense, it is also heterogeneous. We further found that each gene had a unique recombination network (Fig. 3*D*) suggesting each region has been shaped by different sets of recombination events along the evolutionary trajectories of the samples.

The observed heterogeneity in the connectivity of sequence clusters in these recombination networks led us to hypothesize that the gene pools which SL-CoVs recombine with are partitioned or structured. We tested this hypothesis by inferring recombination parameters of the SL-CoV gene pools and found these to be diverse and characterized by high recombination rates (Fig. 3 *E* and *F*). Our inference of the corresponding θ_{sample} distribution (*SI Appendix*, Fig. S10*B*) shows that generally θ_{pool} is orders of magnitude higher, which suggests that the SL-CoV gene pools constitute a diverse and largely unsampled reservoir of viral sequences. Moreover, the c_{sample} distribution (*SI Appendix*, Fig. S10*A*) indicates that the set of SL-CoV genomes have been substantially impacted by recombination, with a median of $c_{sample} \sim 87\%$ (for RC effect size >2). The diversity and structure of the gene pools for a given microorganism can vary widely; we can imagine a scenario in which different samples from a microbial population interact with a discrete set of gene pools or, alternatively, these gene pools could overlap, leading to recombination events occurring between a sample and multiple pools (Fig. 4*A*). What sets the "softness" of gene pool boundaries in microorganisms is unclear, and will undoubtably be the focus of future investigations. In the case of SL-CoVs, it seems unlikely, based on the literature (24, 25, 68, 69), that the molecular mechanisms underlying recombination and mutation are so unique to each member of this group of viruses that it would result in the diverse distributions of pool parameters which we observe (Fig. 3 *E* and *F*). However, SL-CoVs can exist in a broad range of hosts with different sets of selective pressures (11, 69), and these diverse environments with unique selective pressures may strongly influence the observed recombination rates.

A potential confounding factor in our analysis is heterogeneous mutation rates, as it has been demonstrated that SARS-CoV-2 experiences rate heterogeneity across the genome (27, 55). By analyzing single genes (Fig. 2 *C*–*E* and *SI Appendix*, Fig. S3), we account for variability in mutation rates across large genomic regions. By running simulations with heterogenous mutation rates and hypermutable sites, we control for finer scale heterogeneity in mutation rate, and our simulations suggest that, at least over a biologically relevant range, these effects do not confound our analysis (*SI Appendix*, Fig. S6). Additionally, it has been shown that many human pathogenic RNA viruses exhibit heterogeneous coalescence times as a consequence of variation in selection over time (71). We have controlled for this by separately inferring pool recombination rates for individual SL-CoV sequence clusters or cluster pairs (Fig. 3*F*), for which coalescent times are much more tightly distributed, or across the entire sample phylogeny for which coalescence times are heterogenous (tree in Fig. 3*A*, recombination rates in Table 1 for orf1a and the spike protein); the inferred pool

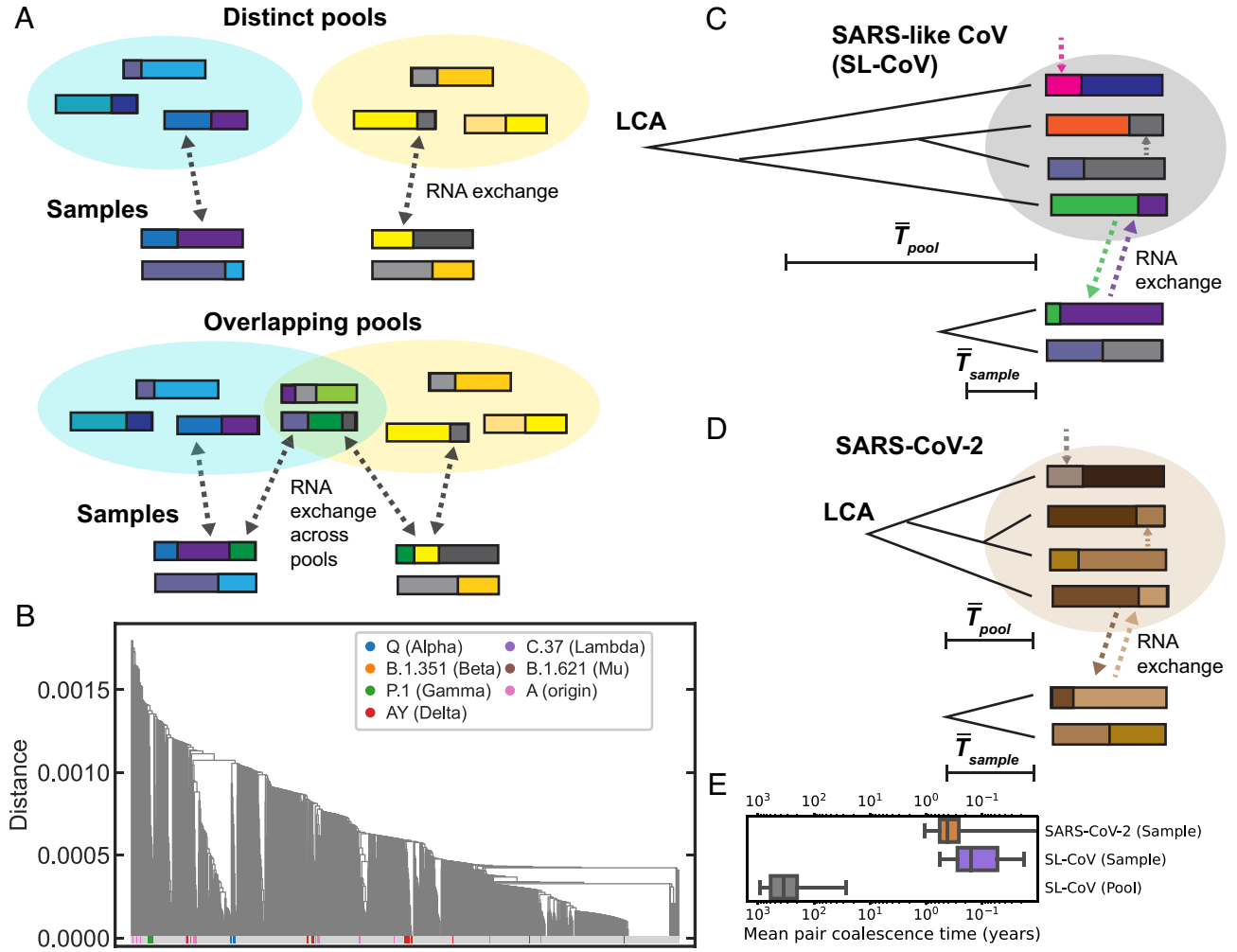

Fig. 4. Gene pool structure and masking of recombination by sequence similarity. (*A*) Schematic depicting samples recombining with distinct gene pools which have no overlap, and overlapping gene pools where recombination can occur between multiple pools and a given sample. (*B*) Dendrogram of SARS-CoV-2 sequences from the NCBI database, where one sequence from each Pango lineage was randomly selected to represent that lineage. Pairwise distances were computed using genome-wide synonymous diversity, and clustering was performed with the average linkage algorithm (see *SI Appendix*). Lineages whose parent lineages are World Health Organization designated Variants of Concern and Interest [as of October 27, 2021 (70)] have colored tips, all other tips are colored gray. (*C* and *D*) Schematics illustrating hypothesis for why detecting recombination using correlated substitutions is not possible using just SARS-CoV-2 sequences. Panel *C* shows SARS-like coronavirus (SL-CoV) samples recombining with a diverse pool, and panel *D* shows SARS-CoV-2 recombining with a pool of SARS-CoV-2 sequences. LCA is last common ancestor. (*E*) Tukey boxplots of the distributions of coalescence times of sequence pairs for samples and pools corresponding to the schematics in *C* and *D*. For the SARS-like coronaviruses (SL-CoV), the distributions are the mean coalescence times for sequence pairs within each of the clusters and cluster pairs shown in Fig. 3*E*. For the calculation of $\bar{T}_{sample}$ of the SL-CoVs, only clusters and cluster pairs with RC effect size >2 were used (see Main text and *SI Appendix*, Eq. **S3**). For SARS-CoV-2, the boxplot depicts the distribution of coalescence times for the sequence pairs of the tree shown in Fig. 4*B*. The line bisecting the box is the 50th percentile, the upper and lower edges of the box are the 25th and 75th percentile, respectively, and the whiskers are 1.5*IQR. Horizontal axis is logarithmic (the lower whisker for SARS-CoV-2 extends to zero).

recombination rates are similar, indicating that variability in coalescence time does not substantially impact our inference of recombination rates for gene pools. It is possible that selection on synonymous sites, e.g., relating to codon usage bias and preferences for CpG dinucleotide frequency and GC richness, could affect our analysis (55, 72–75). However, the strength of such selection in SL-CoV is unclear; recent work found primarily statistically insignificant patterns with regard to selection relating to CpG and GC content (55), and there have been results suggesting there is selection both against (76) or for (55) U content. In previous work, we ran simulations with an analogous model which showed that selection at linked sites, which can act to reduce diversity at synonymous sites, minimally affects our analysis (29).

Our analysis of the SL-CoV samples indicates that recombination occurs at least as often as mutation in nearly all lineages $[(\gamma/\mu)_{pool} > 1$; see Fig. 3*F*], a result that differs substantially from inference based on Bayesian MCMC phylogenetic simulations on a similar SL-CoV dataset, which found that recombination events occur 200 times less frequently than mutations (20). Phylogenetic methods typically attempt to infer the ancestry of each piece of DNA within the sampled genomes by modeling all possible recombination and mutation events that could have occurred since its coalescence. Inference of the maximum likelihood set of recombination events relies on the existence of inconsistencies in the inferred tree of different pieces of DNA across the sample. As branches are joined going backward in time, the size of the trees decreases monotonically, and there is progressively less evidence to call recombination events. Such methods are thus expected to underestimate recombination rates in strongly clustered samples with very deep branches, such as the SL-CoV dataset (Fig. 3*A*). In contrast, our approach accounts for recombination events that occur in the external, unsampled pool; these events cannot be individually inferred, however their signature is the correlation profile. These correlations develop over long timescales under the

combined effect of historical recombination and mutation events in the pool (29) and may enable more accurate measurements of $(\gamma/\mu)_{pool}$ for deeply branched samples (30). Additionally, the recombination rate estimated in ref. 20 is limited to recombination during co-infection events between distinct viral lineages, while the pool recombination rate we estimate here is based on all recombination events that occur in the ancestry of the viral gene pool, including within the same lineage.

We analyzed every complete genome assembly for SARS-CoV-2 from human hosts in NCBI (444,145 at the time of analysis) and created a map of correlated substitutions across the genome (Fig. 2*F*). We did not observe any regions with strongly correlated substitutions, nor did we find that any genes had correlation profiles which indicated the presence of recombination (Fig. 2 *G–I* and *SI Appendix*, Fig. S4). At first, this may give the impression that SARS-CoV-2 has recombined little since it entered the human population in late 2019. However, given i) the high recombination rates of related SL-CoV strains, ii) the high levels of ancestral recombination we measured between SARS-CoV-2 and SL-CoV, and iii) the conserved molecular mechanism of RNA replication which underlies template switching, we hypothesize that recombination is most likely occurring among SARS-CoV-2 in human hosts yet insufficient time has passed for SARS-CoV-2 to accumulate enough diversity to allow for detection of recombination via correlated substitutions. A simple comparison of an average linkage dendrogram sampling across all major lineages of SARS-CoV-2 (Fig. 4*B*) shows that the overall diversity levels are orders of magnitude lower for SARS-CoV-2 as compared to the SARS-like coronavirus dendrogram (Fig. 3*A*). Other studies have previously suggested that the current lack of SARS-CoV-2 diversity impedes the ability to detect recombination (77, 78) resulting in what these investigators suggest are potentially large underestimates of recombination levels in these pandemic datasets (79).

We can further this argument by estimating differences in coalescence times for SARS-CoV-2 versus the SL-CoVs as a whole. We use a standard maximum likelihood phylodynamic approach (*TreeTime*; (54)) to estimate the mutation rate as $\mu \approx 9.8 \times 10^{-4} bp^{-1} \cdot year^{-1}$ for SARS-CoV-2 (see *Methods* for details); for the SL-CoVs, previous studies have inferred $\mu \approx 5.0 \times 10^{-4} bp^{-1} \cdot year^{-1}$ (20, 21). We use these mutation rates along with θ_{sample} and θ_{pool} to estimate the mean coalescence times for pairs in the sample and pool for the SL-CoVs and SARS-CoV-2. For the SL-CoVs, if we use the median values of θ_{sample} and θ_{pool} for the parameter distributions of the 27 clusters appearing in Fig. 3 *E* and *F*, we find that $\overline{T}_{sample} \sim 1.6 \times 10^{-1}$ y and $\overline{T}_{pool} \sim 3.5 \times 10^{2}$ y (Fig. 4*E*; for $\overline{T}_{sample}$, we use cluster pairs for which RC effect size > 2, as described in *Results*). This suggests that the SL-CoV samples recombined with pools which have been accumulating diversity for much longer times than the samples (Fig. 4*C*). For SARS-CoV-2, we can use the θ_{sample} distribution of the SARS-CoV-2 sequence pairs in the dendrogram in Fig. 4*B* to compute the median coalescence time for pairs as: $\overline{T}_{sample} \sim 4.2 \times 10^{-1}$ y (Fig. 4*E*; θ_{sample} was computed for each sequence pair using the classic population genetics expression for pairwise heterozygosity (80) given as Eq. S8 in the *SI Appendix*). For this dataset, which is comprised solely of SARS-CoV-2 sequences from human hosts, opportunities for recombination have almost exclusively stemmed from co-infection events involving multiple strains from local transmission chains between humans, in which every RNA sequence is highly similar (Fig. 4*D*). Furthermore, in the case of this SARS-CoV-2 dataset, we know that our sample has effectively the same coalescence time and rate of synonymous substitution as the pool, because sequences in the sample and pool are highly overlapping (the pool here being un-sequenced SARS-CoV-2 strains in the human population). If we therefore use $\overline{T}_{sample}$ as our estimate of $\overline{T}_{pool}$ in SARS-CoV-2, this suggests that, consistent with expectation, the SL-CoV pools have had much longer to accumulate diversity, which allows us to differentiate between two sequences which have swapped to create a new hybrid when analyzing correlated substitutions. While we cannot predict when sufficient diversity will have accumulated to allow for the detection of recombination via correlated substitutions, $\overline{T}_{pool}$ for the SL-CoVs suggests this may be on the order of $\sim 10^2$ y. This analysis therefore further emphasizes the current need for computationally-efficient approaches to sift through the massive amounts of available sequencing data to pinpoint recombinant SARS-CoV-2 sequences (e.g., refs. 78 and 81–83).

While previous studies have examined recombination in SL-CoVs via the analysis of phylogenetic incongruence and Bayesian inference (20–25), our work here offers unique advantages and insights; it makes no assumptions of phylogenetic structure or evolutionary parameters (i.e., specification of prior distributions for parameters which is necessary for Bayesian inference), and we infer population genetic parameters of the larger, unsampled viral reservoirs that the SL-CoVs have recombined with. Moreover, we use differences in diversity levels between a sample and its pool to determine the residual clonal signal in the data. Our methodology enables analysis of the massive datasets that have become the norm in COVID-19 epidemiology, which are prohibitively large for current Bayesian simulation-based approaches. Our work yields a new set of tools to analyze recombination in positive-sense RNA viruses and reveals the parameters of homologous recombination of the diverse set of gene pools with which SL-CoVs recombine. This may aid in understanding how the interplay among population structure, selection, and recombination acts to mold the unique genetic architecture of viruses at the center of major epidemics.

Materials and Methods

Generation of Multi-Sequence Alignment Files. For all RNA viruses studied, we used reference-guided alignment to build consensus genomes by taking whole genome assemblies and aligning them to a reference genome from NCBI (Genbank accessions for reference genomes listed in *SI Appendix*, Table S5) using the program *ViralMSA* (84) with *Minimap2* as the aligner (85). We then used our in-house program *splitFasta* to split the multi-FASTA file generated by *ViralMSA* into separate FASTA files for each genome, and used our program *CollectGeneAlignments* to extract CDS regions and generate an XMFA file. We filtered out any gene alignment with >10% gaps using our program *FilterGaps*. For calculations of correlation profiles across single genes, our program *geneMSA* was used to split the XMFA file including all gene CDS regions into separate multi-fasta files for each gene, which could then be analyzed with our program *mcorr-gene-aln* (described below). We created the program *CollectGeneAlignments* previously (used in refs. 30 and 31), and it can be found here: https://github.com/kussell-lab/ReferenceAlignmentGenerator. All other in-house programs can be found here: https://github.com/kussell-lab/viral-mcorr.

Measurement of Correlation Coefficient of Synonymous Substitutions. We computed Pearson's correlation coefficient for synonymous substitutions along the length of the genome using the following expression:

$$\rho(X,Y) = \frac{\langle \sigma_X \sigma_Y \rangle - \langle \sigma_X \rangle \langle \sigma_Y \rangle}{\sqrt{\langle \sigma_X \rangle (1 - \langle \sigma_X \rangle) \langle \sigma_Y \rangle (1 - \langle \sigma_Y \rangle)}}, \quad [1]$$

where $\rho(X, Y)$ is Pearson's correlation coefficient for a pair of codons at genomic positions X and Y, and σ_i is the substitution profile for a site i (as described in the main text, this is a binary variable assigned a 1 for difference and 0 for identity at genomic position i). The substitution profile is measured at each fourfold

degenerate, third-position codon for every sequence pair k, then averaging over all pairs. For further details, see *SI Appendix*.

Measurement of Sample Correlation Profiles for Single Genes and Whole Genomes. Using the whole genome alignments of our sample sequences we measure the "substitution profile" ($\sigma_i(k)$) at each fourfold degenerate, third-codon position i for every sequence pair k. When calculating correlation profiles for single genes, we do this separately for each gene's CDS region, where all viruses appearing in this paper except for the coronaviruses only code for a single polyprotein. When computing correlation profiles across the genome of the coronaviruses (e.g., *SI Appendix*, Fig. S9 and Fig. 3 *E* and *F*), we first sort the CDS regions by the position they appear in the genome into one continuous sequence of codons and then compute correlation profiles across the entire genome. We compute the pairwise synonymous diversity of a CDS region as:

$$d_s = \langle \overline{\sigma_i(k)} \rangle, \qquad [\mathbf{2}]$$

in which the bar signifies averaging over sequence pairs k and the bracket signifies averaging over positions i. We compute the joint probability of synonymous substitutions for a pair of sites separated by l nt as:

$$Q_s(l) = \langle \overline{\sigma_i(k)\sigma_{i+l}(k)} \rangle. \qquad [\mathbf{3}]$$

The correlation profile is then calculated as $P(l) = Q_s(l)/d_s$. For further details see *SI Appendix*.

Fitting Procedure for Correlation Profiles and Model Selection. The fitting procedure used here is largely described in refs. 30 and 31. We used the LMFIT python package version 0.9.7 (86) to fit the analytical form of $P(l)$ appearing in *SI Appendix*, Eq. **S2** (link to package here: https://lmfit.github.io/lmfit-py/). To infer recombination parameters, we fit the data with *SI Appendix*, Eq. **S2** by either varying the parameters θ_s, ϕ_s, and fixing $\bar{f}$ to be the length of the genome, or by varying all three parameters. The former fit is the "template-switching model," which assumes only complete RNA templates are exchanged during template-switching events, and the latter is the "fragment-incorporation model," which assumes that template-switching events can involve incomplete RNA templates (i.e., fragments). Both are described in the main text, and all data shown are the results of fitting with the template-switching model except for the parameters shown in *SI Appendix*, Table S2. To distinguish between profiles with distinct signatures of recombination and instances of either no recombination or unclear signals of recombination, we compared the fits from the template-switching and fragment-incorporation models to what we refer to as the "null-recombination" model. In the null-recombination model, we simply set $c_{s,1} = c_{s,2} = 0$ in *SI Appendix*, Eq. **S2**, yielding $P(l) = d(2\theta_s) = 2\theta_s/(1 + 2\theta_s\tilde{a})$. This gives a correlation profile independent of l, which is fit by averaging the measured values of $P(l)$, yielding a single parameter: θ_s. We then perform model selection using the Akaike information criterion (AIC) to determine which of the three models best predicts the data. For further details see *SI Appendix*.

Data, Materials, and Software Availability. Code data have been deposited in https://github.com/kussell-lab/viral-mcorr (https://doi.org/10.1101/2022.08.26.505425). All study data are included in the article and/or *SI Appendix*. Previously published data were used for this work. All sequences used in this study are publicly available through NCBI GenBank. Accession numbers for genome assemblies are provided as supplementary files. Accession numbers for reference genomes are provided in *SI Appendix*, Table S5.

ACKNOWLEDGMENTS. This work was supported by NIH grant R01-GM-097356 (to E.K.) and grant 20/1041 from the Health Research Council of New Zealand (to O.K.S). Asher Preska Steinberg is a Simons Foundation Awardee of the Life Sciences Research Foundation. We gratefully acknowledge the New York University (NYU) high-performance computing cluster for resources, and its staff for technical support.

Author affiliations: [a]Department of Biology and Center for Genomics and Systems Biology, New York University, New York, NY 10003; [b]School of Natural Sciences, Massey University, Auckland 0745, New Zealand; and [c]Department of Physics, New York University, New York, NY 10003

1. T. Nora *et al.*, Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J. Virol.* **81**, 7620–7628 (2007).
2. L. Moutouh, J. Corbeil, D. D. Richman, Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6106–6111 (1996).
3. E. van der Walt *et al.*, Rapid host adaptation by extensive recombination. *J. Gen. Virol.* **90**, 734–746 (2009).
4. K. Yusa, M. F. Kavlick, P. Kosalaraksa, H. Mitsuya, HIV-1 acquires resistance to two classes of antiviral drugs through homologous recombination. *Antiviral Res.* **36**, 179–189 (1997).
5. Y. Xiao *et al.*, RNA recombination enhances adaptability and is required for virus spread and virulence. *Cell Host Microbe* **19**, 493–503 (2016).
6. L. Chao, T. Tran, C. Matthews, Muller's ratchet and the advantage of sex in the RNA virus phi6. *Evolution (N. Y).* **46**, 289–299 (1992).
7. L. Chao, T. T. Tran, T. T. Tran, The advantage of sex in the RNA virus phi6. *Genetics* **147**, 953–959 (1997).
8. E. Simon-Loriere, E. C. Holmes, Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617–626 (2011).
9. M. M. C. Lai, "Genetic recombination in RNA viruses" in *Genetic Diversity of RNA Viruses*, J. J. Holland, Ed. (Springer, Berlin Heidelberg, 1992), pp. 21–32.
10. S. S. Twiddy, E. C. Holmes, The extent of homologous recombination in members of the genus Flavivirus. *J. Gen. Virol.* **84**, 429–440 (2003).
11. K. V. Holmes, "Coronaviruses (Coronaviridae)" in *Encyclopedia of Virology*, Granoff Webster, Eds. (Academic Press, San Diego, ed. 2, 1999), pp. 291–298.
12. T. Bedford *et al.*, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
13. J. R. Fauver *et al.*, Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell* **181**, 990–996.e5 (2020).
14. E. B. Hodcroft *et al.*, Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
15. N. G. Davies *et al.*, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
16. C. A. Pearson *et al.*, Estimates of severity and transmissibility of novel South Africa SARS-CoV-2 variant 501Y.V2. *Preprint* **50**, 1–4 (2021).
17. E. Volz *et al.*, Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
18. E. C. Sabino *et al.*, Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet* **397**, 452–455 (2021).
19. H. Tegally *et al.*, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
20. N. F. Müller, K. E. Kistler, T. Bedford, A Bayesian approach to infer recombination patterns in coronaviruses. *Nat. Commun.* **13**, 4186 (2022).
21. M. F. Boni *et al.*, Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
22. M. Nikolaidis, P. Markoulatos, Y. Van de Peer, S. G. Oliver, G. D. Amoutzias, The neighborhood of the Spike gene is a hotspot for modular intertypic homologous and non-homologous recombination in Coronavirus genomes. *Mol. Biol. Evol.* **6**, msab292 (2021).
23. B. S. Chrisman *et al.*, Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Min.* **14**, 1–16 (2021).
24. S. A. Goldstein, J. Brown, B. S. Pedersen, A. R. Quinlan, N. C. Elde, Extensive recombination-driven coronavirus diversification expands the pool of potential pandemic pathogens. *Genome Biol. Evol.* **14**, evac161 (2022)
25. X. Li *et al.*, Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6**, 1–12 (2020).
26. H. Wang, S. L. K. Pond, A. Nekrutenko, R. Nielsen, Testing recombination in the pandemic SARS-CoV-2 strains (2020) (February 3, 2022).
27. N. D. Rochman, Y. I. Wolf, G. Faure, F. Zhang, E. V. Koonin, Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104241118 (2021).
28. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).
29. M. Lin, E. Kussell, Correlated mutations and homologous recombination within bacterial populations. *Genetics* **205**, 891–917 (2017).
30. M. Lin, E. Kussell, Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* **16**, 199–204 (2019).
31. A. Preska Steinberg, M. Lin, E. Kussell, Core genes can have higher recombination rates than accessory genes within global microbial populations. *Elife* **11**, e78533 (2022).
32. K. Bentley, D. J. Evans, Mechanisms and consequences of positive-strand RNA virus recombination. *J. Gen. Virol.* **99**, 1345–1356 (2018).
33. S. K. Garushyants, I. B. Rogozin, E. V. Koonin, Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring. *Commun. Biol.* **4**, 1–9 (2021).
34. C. L. Liao, M. M. Lai, RNA recombination in a coronavirus: Recombination between viral genomic RNA and transfected RNA fragments. *J. Virol.* **66**, 6117–6124 (1992).
35. N. Ledinko, Genetic recombination with poliovirus type 1. *Virology* **20**, 107–119 (1963).
36. G. K. Hirst, Genetic recombination with Newcastle disease virus, polioviruses, and influenza. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 303–309 (1962).
37. A. P. Gmyl *et al.*, Nonreplicative RNA recombination in Poliovirus. *J. Virol.* **73**, 8958–8965 (1999).
38. C. Savolainen-Kopra, S. Blomqvist, Mechanisms of genetic variation in polioviruses. *Rev. Med. Virol.* **20**, 358–371 (2010).
39. M. Slatkin, Linkage disequilibrium–Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).

40. A. N. Lukashev *et al.*, Recombination in circulating human enterovirus B: Independent evolution of structural and non-structural genome regions. *J. Gen. Virol.* **86**, 3281–3290 (2005).
41. M. S. Oberste, K. Maher, M. A. Pallansch, Evidence for frequent recombination within species human enterovirus B Based on complete genomic sequences of all thirty-seven serotypes. *J. Virol.* **78**, 855–867 (2004).
42. E. C. Holmes, M. Worobey, A. Rambaut, Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.* **16**, 405–409 (1999).
43. M. Worobey, A. Rambaut, E. C. Holmes, Widespread intra-serotype recombination in natural populations of dengue virus. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 7352–7357 (1999).
44. H. J. G. Tolou *et al.*, Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome suquences. *J. Gen. Virol.* **82**, 1283–1290 (2001).
45. N. Y. Uzcategui *et al.*, Molecular epidemiology of dengue type 2 virus in Venezuela: Evidence for in situ virus evolution and recombination. *J. Gen. Virol.* **82**, 2945–2953 (2001).
46. A. M. Q. King, D. McCahon, K. Saunders, J. W. I. Newman, W. R. Slade, Multiple sites of recombination within the RNA genome of Foot-and-mouth disease virus. *Virus Res.* **3**, 373–384 (1985).
47. D. McCahon, W. Slade, A. Priston, J. Lake, An extended genetic recombination map for Foot-and-mouth disease virus. *J. Gen. Virol.* **35**, 555–565 (1977).
48. L. Ferretti *et al.*, Within-host recombination in the Foot-and-mouth disease virus genome. *Viruses* **10**, 1–14 (2018).
49. S. M. Jamal *et al.*, Evidence for multiple recombination events within Foot-and-mouth disease viruses circulating in West Eurasia. *Transbound. Emerg. Dis.* **67**, 979–993 (2020).
50. B. E. Pickett, E. J. Lefkowitz, Recombination in West Nile Virus: Minimal contribution to genomic diversity. *Virol. J.* **6**, 1–7 (2009).
51. C. E. McGee *et al.*, Stability of yellow fever virus under recombinatory pressure as compared with chikungunya virus. *PLoS One* **6**, e23247 (2011).
52. T. Bedford, E. B. Hodcroft, Phylogeny of SARS-like betacoronaviruses including novel coronavirus SARS-CoV-2. *J. Mol. Biol.* **432**, 3309–3325 (2020).
53. J. Hadfield *et al.*, NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
54. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, 1–9 (2018).
55. N. De Maio *et al.*, Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, (2021).
56. N. De Maio *et al.*, phastSim: Efficient simulation of sequence evolution for pandemic-scale datasets. *PLoS Comput. Biol.* **18**, e1010056 (2022).
57. T. J. Wheeler, J. D. Kececioglu, Multiple alignment by aligning alignments. *Bioinformatics* **23**, 559–568 (2007).
58. S. K. P. Lau *et al.*, Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related rhinolophus bat coronavirus in china reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. *J. Virol.* **84**, 2808–2819 (2010).
59. S. K. P. Lau *et al.*, Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14040–14045 (2005).
60. B. Hu *et al.*, Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, 1–27 (2017).
61. X. Y. Ge *et al.*, Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
62. D. Hu *et al.*, Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.* **7**, 154 (2018).
63. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
64. J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
65. F. Pereira, Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infect. Genet. Evol.* **85**, 104525 (2020).
66. L. Zinzula, Lost in deletion: The enigmatic ORF8 protein of SARS-CoV-2. *Biochem. Biophys. Res. Commun.* **538**, 116–124 (2021).
67. J. B. Plotkin, G. Kudla, Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).
68. J. Gribble *et al.*, The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathog.* **17**, 1–28 (2021).
69. P. V'kovski, Kratzel, Steiner, Stalder, Thiel, Coronavirus biology and replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.*, 10.1038/s41579-020-00468-6 (2020).
70. WHO, Tracking SARS-CoV-2 variants *(World Health Organization*, 2021).
71. P. Mutz *et al.*, Human pathogenic RNA viruses establish noncompeting lineages by occupying independent niches. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2121335119 (2022), 10.1073/pnas.2121335119.
72. G. M. Jenkins, E. C. Holmes, The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **92**, 1–7 (2003).
73. P. C. Y. Woo, B. H. L. Wong, Y. Huang, S. K. P. Lau, K. Y. Yuen, Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. *Virology* **369**, 431–442 (2007).
74. T. Mourier *et al.*, Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.* **538**, 35–39 (2021).
75. M. Dilucca, S. Forcelloni, A. G. Georgakilas, A. Giansanti, A. Pavlopoulou, Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses* **12**, 498 (2020).
76. A. M. Rice *et al.*, Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: Implications for vaccine design. *Mol. Biol. Evol.* **38**, 67–83 (2021).
77. A. Ignatieva, J. Hein, P. A. Jenkins, Evidence of ongoing recombination in sars-cov-2 through genealogical reconstruction. *Mol. Biol. Evol.* **39**, 1–11 (2022).
78. D. VanInsberghe, A. S. Neish, A. C. Lowen, K. Koelle, Recombinant SARS-CoV-2 genomes are currently circulating at low levels. *Virus Evol.* **7**, 1–12 (2021).
79. Y. Turakhia *et al.*, Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609**, 994–997 (2022), 10.1038/s41586-022-05189-9.
80. J. Wakeley, *Coalescent Theory: An Introduction* (Macmillan Learning, ed. 1, 2009).
81. H. Yi, 2019 Novel coronavirus is undergoing active recombination. *Clin. Infect. Dis.* **71**, 884–887 (2020).
82. D. Haddad *et al.*, SARS-CoV-2: Possible recombination and emergence of potentially more virulent strains. *PLoS One* **16**, 1–20 (2021).
83. A. Varabyou, C. Pockrandt, S. L. Salzberg, M. Pertea, Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *Genetics* **218**, iyab074 (2021).
84. N. Moshiri, ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* **37**, 714–716 (2021).
85. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
86. M. Newville, T. Stensitzki, D. B. Allen, A. Ingargiola, LMFIT, Non-linear least-square minimization and curve-fitting for python. *Zenodo*, 10.5281/zenodo.11813 (2014).