nature

https://doi.org/10.1038/s41586-022-05189-9

Accelerated Article Preview

Pandemic-Scale Phylogenomics Reveals The SARS-CoV-2 Recombination Landscape

Received: 13 December 2021

Accepted: 3 August 2022

Accelerated Article Preview

[redacted]

Cite this article as: Turakhia, Y. et al. Pandemic-Scale Phylogenomics Reveals The SARS-CoV-2 Recombination Landscape. *Nature* https://doi.org/10.1038/s41586-022-05189-9 (2022)

Yatish Turakhia, Bryan Thornlow, Angie Hinrichs, Jakob McBroome, Nicolas Ayala, Cheng Ye, Kyle Smith, Nicola De Maio, David Haussler, Robert Lanfear & Russell Corbett-Detig

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Title:

Pandemic-Scale Phylogenomics Reveals The SARS-CoV-2 Recombination Landscape

Authors:

Yatish Turakhia[1,2,3,*,†], Bryan Thornlow[1,2,†], Angie Hinrichs[2], Jakob McBroome[1,2], Nicolas Ayala[1,2], Cheng Ye[3], Kyle Smith[4], Nicola De Maio[5], David Haussler[1,2,6], Robert Lanfear[7], Russell Corbett-Detig[1,2,*]

Affiliations:

[1] Department of Biomolecular Engineering, University of California, Santa Cruz; Santa Cruz, CA 95064, USA
[2] Genomics Institute, University of California, Santa Cruz; Santa Cruz, CA 95064, USA
[3] Department of Electrical and Computer Engineering, University of California, San Diego; San Diego, CA 92093, USA
[4] Department of Biological Sciences, University of California, San Diego; San Diego, CA 92093, USA
[5] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus; Cambridge CB10 1SD, UK
[6] Howard Hughes Medical Institute, University of California, Santa Cruz; Santa Cruz, CA 95064, USA
[7] Department of Ecology and Evolution, Research School of Biology, Australian National University; Canberra, ACT 2601, Australia
*Correspondence to: rucorbet@ucsc.edu and yturakhia@ucsd.edu
†Denotes equal contribution

Summary Paragraph:

Accurate and timely detection of recombinant lineages is crucial for interpreting genetic variation, reconstructing epidemic spread, identifying selection and variants of interest, and accurately performing phylogenetic analyses [1–4]. During the SARS-CoV-2 pandemic, genomic data generation has exceeded the capacities of existing analysis platforms, thereby crippling real-time analysis of viral evolution [5]. Here, we use a novel phylogenomic method to search a nearly comprehensive SARS-CoV-2 phylogeny for recombinant lineages. In a 1.6M sample tree from May 2021, we identify 589 recombination events, which indicate that approximately 2.7% of sequenced SARS-CoV-2 genomes have detectable recombinant ancestry. Recombination breakpoints are inferred to occur disproportionately in the 3' portion of the genome that contains the spike protein. Our results highlight the need for timely analyses of recombination for pinpointing the emergence of recombinant lineages with the potential to increase transmissibility or virulence of the virus. We anticipate that this approach will empower comprehensive real time tracking of viral recombination during the SARS-CoV-2 pandemic and beyond.

Main Text:

Recombination is a primary contributor of novel genetic variation in many prevalent pathogens, including *betacoronaviruses* [6], the clade that includes SARS-CoV-2. By mixing genetic material from diverse genomes, recombination can produce novel combinations of mutations that have potentially important phenotypic effects [7]. For example, recombination is thought to have played an important role in the recent evolutionary histories of MERS [8], SARS-CoV [9–12]. Recombination might also have the potential to generate viruses with zoonotic potential in the future [13]. Therefore, accurate and timely characterization of recombination is foundational for understanding the evolutionary biology and infectious potential of established and emerging pathogens in human, agricultural, and natural populations.

Now that substantial genetic diversity is present across SARS-CoV-2 populations [14] and co-infection with different SARS-CoV-2 variants has been known to sometimes occur [15], recombination is expected to be an important source of new genetic variation during the pandemic. Whether or not there is a detectable signal for recombination events in the SARS-CoV-2 genomes has been fiercely debated since the early days of the pandemic [13]. Nonetheless, several apparently genuine recombinant lineages have been identified using *ad hoc* approaches [16] and semi-automated methods that cope with vast SARS-CoV-2 datasets by reducing the search space for possible pairs of recombinant ancestors [16,17]. Because of the importance of timely and accurate surveillance of viral genetic variation during the ongoing SARS-CoV-2 pandemic, new approaches for detecting and characterizing recombinant haplotypes are needed to evaluate new variant genome sequences as quickly as they become available. Such rapid turnaround is essential for driving an informed and coordinated public health response to novel SARS-CoV-2 variants.

We developed a novel method for detecting recombination in pandemic-scale phylogenies, Recombination Inference using Phylogenetic PLacEmentS (RIPPLES, Fig. 1). Because recombination violates the central assumption of many phylogenetic methods, *i.e.,* that a single evolutionary history is shared across the genome, recombinant lineages arising from diverse genomes will often be found on "long branches" which result from accommodating the divergent evolutionary histories of the two parental haplotypes (Fig. 1). Note that as long as recombination is relatively uncommon, phylogenetic inference is expected to remain accurate even when branch lengths are artifactually expanded [18]. RIPPLES exploits that signal by first identifying long branches on a comprehensive SARS-CoV-2 mutation-annotated tree [19,20]. RIPPLES then exhaustively breaks the potential recombinant sequence into distinct segments and replaces each onto a global phylogeny using maximum parsimony. RIPPLES reports the two parental nodes – hereafter termed donor and acceptor – that result in the highest parsimony score improvement relative to the original placement on the global phylogeny (Text S1). Our approach therefore leverages phylogenetic signals for each parental lineage as well as the spatial correlation of markers along the genome. We establish significance using a null model conditioned on the inferred site-specific rates of *de novo* mutation (Text S2-S3).

Substantial testing via simulation indicates that RIPPLES is efficient, sensitive, and can confidently identify recombinant lineages (Text S4-S6). As expected [21], when recombination occurs towards the edges of the genome or between genetically similar sequences, it is harder to detect using RIPPLES (Figs. S1-S2).

Nonetheless, RIPPLES detects simulated recombinants with 75.8% sensitivity. Among the simulated samples detected as recombinants, RIPPLES accurately identifies 90% of simulated breakpoints. (Extended Data Table 1, Text S6). Furthermore, RIPPLES is able to detect all highly confident recombinants identified in a previous analysis[16] (Text S6). Recombination analysis using RIPPLES on a global phylogeny of approximately 1.6 million SARS-CoV-2 genomes reveals that a significant fraction of the sequenced SARS-CoV-2 genomes belong to detectable recombinant lineages. To mitigate the impacts of sequencing and assembly errors, we exclude all nodes with only a single descendant, we applied conservative filters to remove potentially spurious samples from the recombinant sets flagged by RIPPLES, and we manually confirmed mutations in a subset of putative recombinant samples using raw sequence read data (Text S7-S8, Extended Data Table 2, Extended Data Fig. 3). After this, we retained 589 unique recombination events, which have a combined total of 43,104 descendant samples (Extended Data Table 3). This means that approximately 2.7% of total sampled SARS-CoV-2 genomes are inferred to belong to detectable recombinant lineages. *Post hoc* statistical analysis yields an empirical false discovery rate estimate of 11.0% for our statistical thresholds (Text S9, Extended Data Table 4). Additionally, excess similarity of geographic location and date metadata among the descendants of donor and acceptor nodes supports the notion that many ancestors of recombinant genomes co-circulated within human populations (Text S10-S11, Extended Data Fig. 4-5). Because recombination events that occur between genetically similar viral lineages are challenging to detect (Extended Data Fig. 2), ours is expected to be a potentially large underestimate of the overall frequency of recombination. As a result, the RIPPLES estimate is likely conservative with respect to the global frequency of recombination in the SARS-CoV-2 population.

RIPPLES uncovered a strikingly non-uniform distribution of recombination breakpoint positions across the SARS-CoV-2 genome, consistent with previous analyses in *betacoronaviruses* [11,22]. In particular, among putative recombination events there is an excess of recombination breakpoints towards the 3' end of the SARS-CoV-2 genome relative to expectations based on random breakpoint positions ($p < 1\times10^{-7}$; permutation test; Text S12). Importantly, no such bias is apparent when we simulate recombination breakpoints following a uniform distribution (Text S13, Extended Data Fig. 1). Change-point analysis identifies an increase in the frequency of recombination breakpoints immediately 5' of the Spike protein region (20,875 bp; Text S14), and this pattern is consistent when restricting ourselves to putative nodes with the largest numbers of descendants and among diverse data sources further suggesting that it is not artefactual (Text S15, Extended Data Table 5). The rate of putative recombination breakpoints is approximately three times higher towards the 3' of the change-point than the 5' interval (Fig. 2) – which is similar to the relative recombination rates in the genomes of other human coronaviruses [11].

Several lines of evidence suggest that the skewed distribution of recombination breakpoint positions is not a consequence of positive selection at the level of between-host transmission dynamics. First, many of these recombinant clades have existed for a relatively short period of time, and might already be extinct. The mean timespan between the earliest and latest dates of observed descendants of detected recombinant nodes is just 37 days. Second, of the subset of recombination events that we inferred to occur between Variants of Concern (VOC; lineages B.1.1.7, B.1.351, B.1.617.2, and P.1 [23]) and other lineages, VOCs contribute slightly fewer Spike protein mutations than non-VOC lineages on average (60 out of 125 VOC/non-VOC recombinants, P = 0.48, sign test). Third, recombinant clade size does not greatly differ from the remaining clade sizes, which would be expected if recombinant lineages experienced strong selection (P = 0.8470, permutation test). Therefore, although natural selection on between-host transmission dynamics of recombinant lineages could also impact the observed distribution of recombinant breakpoint positions [11], our data indicates that other biases shape the distribution of recombination events across the SARS-CoV-2 genome. These could include a neutral mechanistic bias affecting the distribution of recombination breakpoints.

Although not yet widespread among circulating SARS-CoV-2 genomes, recombination has measurably contributed to the genetic diversity within SARS-CoV-2 lineages. The ratio of variable positions contributed by recombination versus those resulting from *de novo* mutation, R/M, is commonly used to summarize the relative impacts of these two sources of variation [22]. Using our dataset of putative recombination events, we estimate that R/M = 0.00264 in SARS-CoV-2 (Text S16). This is low for a coronavirus population (*e.g.,* for MERS, R/M is estimated to be 0.25-0.31, [22]), which presumably reflects the extremely low genetic diversity among possible recombinant ancestors during the earliest phases of the pandemic and the conservative nature of our approach. As SARS-CoV-2 populations accumulate genetic diversity and co-infect hosts with other species of viruses, recombination will play an increasingly large role in generating functional genetic diversity and this ratio could increase [24]. RIPPLES is therefore poised to play a primary role in detecting novel recombinant lineages and quantifying their impacts on viral genomic diversity as the pandemic progresses.

Our extensively optimized implementation of RIPPLES allows it to search the entire phylogenetic tree and detect recombination both within and between SARS-CoV-2 lineages without *a priori* defining a set of lineages or clade-defining mutations. This is a key advantage of our approach relative to other methods that cope with the scale of SARS-CoV-2 datasets by reducing the search space for possible recombination events (*e.g.,* [16,17,25]). RIPPLES discovers 223 recombination events within branches of the same Pango lineages. Our results also include 366 inter-lineage recombination events (Extended Data Table 3). Additionally, we find evidence that recombination has influenced the Pangolin SARS-CoV-2 nomenclature system [23]. Specifically, we discover that the root of B.1.355 lineage might have resulted from a recombination event between nodes belonging to the B.1.595 and B.1.371 lineages (Fig 3, Extended Data Table 3). These diverse recombination events highlight the versatility and strengths of the approach taken in RIPPLES.

The detection of increased recombination rates in the 3' portion of the SARS-CoV-2 genome, which contains the Spike protein, highlight the utility of ongoing surveillance. The Spike protein is a primary location of functional novelty for viral lineages as they adapt to transmission within and among human hosts. Our discovery of the excess of recombination events specifically around the Spike protein, as well as and the relatively high levels of recombinants currently in circulation, underline the importance of monitoring the evolution of new viral lineages that arise through mutation or recombination through real-time analyses of viral genomes. Our work also emphasizes the impact that explicitly considering phylogenetic networks will have for accurate interpretation of SARS-CoV-2 sequences [11].

Beyond SARS-CoV-2, recombination is a major evolutionary force driving viral and microbial adaptation. It can drive the spread of antibiotic resistance [7], drug resistance [1], and immunity and vaccine escape [2]. Identification of recombination is an essential component of pathogen evolutionary analyses pipelines, since recombination can affect the quality of phylogenetic, transmission and phylodynamic inference [3]. For these reasons, computational tools to detect microbial recombination have become very popular and important in recent years [4]. The SARS-CoV-2 pandemic has driven an unprecedented surge of pathogen genome sequencing and data sharing, which has in turn highlighted some of the limitations of current software in investigating large genomic datasets [5]. RIPPLES was built for pandemic-scale datasets and is sufficiently optimized to exhaustively search for recombination in one of the largest phylogenies ever inferred in 40 minutes (Text S17). We expect RIPPLES to perform best on densely sampled genomic datasets, which will likely become the norm for many globally distributed pathogens, but we caution that it has not yet been validated on other species. To facilitate real-time analysis of recombination among tens of thousands of new SARS-CoV-2 sequences being generated by diverse research groups worldwide each day [26–28], RIPPLES provides an option to evaluate evidence for recombination ancestry in any user-supplied samples within minutes (Text S17). RIPPLES therefore opens the door for rapid analysis of recombination in heavily sampled and rapidly evolving pathogen populations, as well as providing a tool for real-time investigation of recombinants during a pandemic.

References and Notes

1. Moutouh, L., Corbeil, J. & Richman, D. D. 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).
2. Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat. Genet.* **44**, 352–355 (2012).
3. Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891 (2000).
4. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
5. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* vol. 591 30–33 (2021).
6. Forni, D., Cagliani, R. & Sironi, M. Recombination and Positive Selection Differentially Shaped the Diversity of Betacoronavirus Subgenera. *Viruses* **12**, (2020).
7. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322 (2010).
8. Dudas, G. & Rambaut, A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. *Virus Evol* **2**, vev023 (2016).
9. Lau, S. K. P. *et al.* Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. *Journal of Virology* vol. 89 10532–10547 (2015).
10. Holmes, E. C. & Rambaut, A. Viral evolution and the emergence of SARS coronavirus. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 1059–1065 (2004).
11. Müller, N. F., Kistler, K. E. & Bedford, T. Recombination patterns in coronaviruses. *bioRxiv* (2021) doi:10.1101/2021.04.28.441806.
12. Bobay, L.-M., O’Donnell, A. C. & Ochman, H. Recombination events are concentrated in the spike protein region of Betacoronaviruses. *PLoS Genet.* **16**, e1009272 (2020).
13. Li, X. *et al.* Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. *Science*

Advances **6**, (2020).

14. De Maio, N. *et al.* Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biology & Evolution* (2021) doi:10.1101/2021.01.14.426705.
15. Taghizadeh, P. *et al.* Study on SARS-CoV-2 strains in Iran reveals potential contribution of co-infection with and recombination between different strains to the emergence of new strains. *Virology* **562**, 63–73 (2021).
16. Jackson, B. *et al.* Generation and transmission of inter-lineage recombinants in the SARS-CoV-2 pandemic. *Cell* **184**, 5179–5188 (2021).
17. VanInsberghe, D., Neish, A. S., Lowen, A. C. & Koelle, K. Recombinant SARS-CoV-2 genomes are currently circulating at low levels. *bioRxiv* (2021) doi:10.1101/2020.08.05.238386.
18. Hedge, J. & Wilson, D. J. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* **5**, e02158 (2014).
19. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
20. McBroome, J. *et al.* A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021).
21. Stephens, J. C. On the frequency of undetectable recombination events. *Genetics* **112**, 923–926 (1986).
22. Patiño-Galindo, J. Á., Filip, I. & Rabadan, R. Global Patterns of Recombination across Human Viruses. *Mol. Biol. Evol.* **38**, 2520–2531 (2021).
23. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403–1407 (2020).
24. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. *JAMA* **323**, 2085–2086 (2020).
25. Varabyou, A., Pockrandt, C., Salzberg, S. L. & Pertea, M. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *Genetics* (2021) doi:10.1093/genetics/iyab074.
26. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* vol. 22 (2017).

27. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **49**, D92–D96 (2021).

28. COVID-19 Genomics UK (COG-UK) Consortium. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020).

29. Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).

30. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

31. Killick, R. & Eckley, I. changepoint:an R package for changepoint analysis. *J. Stat. Softw.* **58**, 19 (2014).

32. Turakhia, Y. *et al. Supplement to Pandemic-Scale Phylogenomics Reveals A Landscape of SARS-CoV2 Recombination*. (2022). doi:10.5281/zenodo.6717378.

33. Turakhia, Y. *et al. yatisht/usher: v0.5.6*. (2022). doi:10.5281/zenodo.6709991.

Methods

RIPPLES uses the space-efficient data structure of mutation-annotated trees (MATs) [20], in which the branches of the phylogenetic tree are annotated with mutations that have been inferred to have occurred on them, to identify recombination events. Fig. 1 illustrates the underlying algorithm. RIPPLES identifies putative recombinant nodes containing at least the number of mutations specified by the user, and infers the set of mutations that have occurred on its corresponding sequence by accounting for all mutations annotated on the branches on its path from the root. RIPPLES then adds one or two breakpoints on mutation sites and assesses parsimony score improvement using partial placements compared to the starting parsimony. For more details, see Text S1. To determine whether putative recombinants were significant, we developed a null model by selecting nodes at random and adding k additional mutations drawn from the actual mutation spectra in our global tree. We then placed these samples on the tree and used RIPPLES to determine their parsimony score improvements (Text S2). For each putative recombinant in our global tree, we compared its parsimony score improvement to the distribution of null parsimony score improvements for the same initial parsimony score (Text S3). We developed our starting tree by first taking the May 28 2021 public tree [19,20], masking all problematic sites [29], and pruning samples with fewer than 28,000 non-N nucleotides as well as those with 2 or more non-[ACGTN-] nucleotides (Text S5). After this, we optimized this tree by running matOptimize (Text S4) twice with an SPR radius of 10 and 40 in subsequent rounds with the masked VCF as an input. Instructions for using RIPPLES are available at https://usher-wiki.readthedocs.io/en/latest/tutorials.html. We ran RIPPLES on the n2d-highcpu-224 Google Cloud Platform (GCP) instance containing 224 vCPUs (Text S18).

To test RIPPLES' sensitivity, we simulated recombinant samples by choosing 2 random internal nodes from our phylogeny with at least 10 descendants and choosing breakpoints at random across the genome. We generated 1,000 simulations each for one and two breakpoint recombinants with 0, 1, 2, and 3 additional mutations added to the sequence after the recombination event, using scripts available at https://github.com/bpt26/recombination/. These combinations yielded 2,000 total simulated recombinant lineages. We then measured the ability of RIPPLES to detect breakpoints as a function of the position of the breakpoint and the minimum genetic distance from the recombinant node to either parent (Text S6, genetic distance is estimated based on the number of mutations inferred to separate the focal samples, lineages or nodes). We also evaluated the sensitivity of RIPPLES by ensuring that it detected each of the high-confidence recombinant SARS-CoV-2 clusters of Jackson et al. [16].

We applied several *post hoc* filters to remove putative recombinant nodes that may be false positives resulting from several possible sources of error. For each internal node from each trio (putative recombinant, donor, and acceptor nodes) that comprised a recombinant event, we downloaded the consensus genome sequence for the nearest descendants of each node, from COG-UK, GenBank, GISAID, and the China National Center for Bioinformatics. We then aligned the sequences of all descendants for each trio using MAFFT [30], focusing specifically on recombination-informative sites, i.e. where the allele of the recombinant node matched one parent node but not the other. If recombination-informative mutations were near to indels or missing bases (Ns), or if the entire basis for recombination was a single cluster of mutations in a 20-nucleotide span (Text S7). We also confirmed sequence quality by manually examining raw reads for 10 samples where we could confidently link the raw sequence read data to a given consensus genome (Text S8). To estimate the false discovery rate (FDR) associated with our specific approach and statistical threshold selected, we computed a *post hoc* empirical FDR. We obtained the number of internal nodes that we tested and which were associated with a given parsimony score. Then, for each initial parsimony score and parsimony score improvement, we obtained the expected number of internal nodes that would display that parsimony score improvement under the null model. Our FDR (Extended Data Table 4) is the ratio of expected nodes for a given initial and final parsimony score to the number of detected recombinant nodes with the same initial and final parsimony score (Text S9).

We also performed *post hoc* analysis using sample metadata to determine if the ancestors of the recombinant nodes had higher spatial or temporal overlap than expected by chance. We computed geographic overlap as the joint probability of choosing a sample from the same country from the descendants of the donor and the acceptor nodes. For temporal overlap, we recorded intervals from the earliest to the most recent sample descended from the donor and acceptor, respectively, and calculated the minimum number of days separating the two intervals (with 0 for overlapping intervals). We generated a null distribution for both categories by selecting, for each detected trio, two random internal nodes from the tree with a number of descendants equal to the real donor and acceptor respectively. We then calculated geographic and temporal overlap in the same way for this random set (Extended Data Fig. 4, Text S10).

To determine whether identified recombination breakpoints are significantly shifted towards the 3' end of the genome, we performed a permutation test comparing the difference of the mean of the distribution of uniformly simulated breakpoints with the mean of the detected breakpoint position distribution in the true set (Text S12). We also conducted a change-point analysis using the changepoint R package [31] and fit a Poisson model to the count of recombination prediction interval midpoints. We then computed the mean rate of recombination breakpoints within the intervals on either side of the identified change-point to estimate the fold increase in recombination rate in the 3' portion of the genome (Text S13). To estimate R/M, we found the decrease in parsimony score associated with each detected recombination event as an estimate of R. We then calculated M by taking this value and subtracting it from the total number of mutations observed across our entire phylogeny (Text S16). R/M is the ratio of these values.

Acknowledgments: We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared via GISAID (Table S1) [26], China National Center for Bioinformation (Table S2), COVID-19 Genomics UK (COG-UK) [28] (Table S3), and National Center for Biotechnology Information database [27] (Table S4), on which this research is based. Additionally, the authors thank Shoh Mollenkamp for assisting with the code development.

Funding: B.T., J.M., and R.C.-D. were funded by National Institutes of Health grant R35GM128932. R.C.-D. was funded by Alfred P. Sloan Foundation fellowship and University of California Office of the President Emergency COVID-19 Research Seed Funding Grant R00RG2456. B.T. and J.M. were funded by the National Institutes of Health grant T32HG008345. B.T. was funded by the National Institutes of Health grant F31HG010584. N.D.M. was funded by the European Molecular Biology Laboratory. R.L. was funded by Australian Research Council grant DP200103151 and Chan-Zuckerberg Initiative grant. Additional funding for this project was provided by Eric and Wendy Schmidt by recommendation of the Schmidt Futures program

Author contributions: R.C.-D. and Y.T. developed the approach and wrote the manuscript. R.C.-D., Y.T., B.T., and R.L. designed experiments. Y.T., B.T., A.H., N.D.M. conducted experiments. Y.T., B.T., A.H., J.M., N.A., C.Y. developed code. R.C.-D. and D.H. supervised group. Y.T., B.T., A.H., J.M., N.A., C.Y., N.D.M., D.H., R.L., and R.C.-D. edited manuscript.

Competing interests: R.L. works as an advisor to GISAID. The remaining authors declare no competing interests.

Data Availability: All data is available in the manuscript or the supplementary materials. Dataset 1 (containing the phylogeny analyzed for recombination in this study in Newick format) and Dataset 2 (containing a list of descendant samples of recombinant nodes identified through RIPPLES) are available at https://doi.org/10.5281/zenodo.6717378 [32].

Code Availability: RIPPLES software is available under the MIT license as part of the UShER package at https://github.com/yatisht/usher. We provide a reproducible Google Cloud Platform (GCP) workflow for RIPPLES under https://github.com/yatisht/usher/tree/master/scripts/recombination. An archived version of the specific code and workflow used in this study is available from https://doi.org/10.5281/zenodo.6709991 [33]. We distribute RIPPLES with UShER because it uses the same underlying data objects and UShER is required to infer the input MAT. Documentation for RIPPLES and associated utilities can be found at https://usher-wiki.readthedocs.io/en/latest/.

Figure Legends

Fig. 1. RIPPLES exhaustively searches for optimal parsimony improvements using partial interval placements. (A): A phylogeny with 6 internal nodes (labeled a-f), in which node f is the one being currently investigated as a putative recombinant. The initial parsimony score of node f is 4, according to the multiple sequence alignment below the phylogeny, which displays the variation among samples and internal nodes. Note that internal nodes may not have corresponding sequences in reality, but test for recombination using reconstructed ancestral genomes. **(B-D)**: Three partial placements given breakpoints are shown with their resulting parsimony scores. Arrows mark sites that increase the sum parsimony of the two partial placements of f. The optimal partial placement and breakpoint prediction for node f is in the center (C), with one breakpoint after site 9 and with partial placements both as a sibling of node c and as a descendant of node d.

Fig. 2. RIPPLES detects an excess of recombination in the Spike protein region. (A): The distribution of midpoints of each breakpoint's prediction interval are shown as a density plot, with the underlying recombination prediction intervals plotted as individual lines in gray. We used the midpoint of the breakpoint prediction interval because recombination events can only be localized to prediction intervals which are the regions between two recombination informative SNPs. A dashed vertical line at position 20,875 delimits recombination rate regions identified by change-point analysis (Text S15). The apparent lack of recombination towards the chromosome edges likely reflects a detection bias we describe above (Extended Data Fig. 2) **(B-D)**: Recombination-informative sites (i.e., positions where the recombinant node matches either but not both parent nodes) for three example recombinant trios detected by RIPPLES. The numbers to the left of each sequence correspond to the node identifiers from our MAT. B and D are examples of a recombinant with a single breakpoint (shown in dotted lines), C is an example of a recombinant with two breakpoints. Panels B-D were generated using the SNIPIT package (https://github.com/aineniamh/snipit).

Fig. 3. RIPPLES uncovered evidence that the B.1.355 lineage might have resulted from a recombination event between lineages of B.1.595 and B.1.371. (A): Sub-phylogeny consisting of all 78 B.1.355 samples (purple) and the most closely related 78 samples to nodes 94353 and 102299 from lineages B.1.371 and B.1.595, respectively, using the "k nearest samples" function in matUtils [20]. Nodes 94353 (red) and 102299 (blue) are connected by dotted lines to node 94354 (purple), the root of lineage B.1.355. Recombination-informative mutations are marked where they occur in the phylogeny, with those occurring in a parent but not shared by the recombinant sequence shown in gray. **(B)**: Recombination-informative sites (i.e. sites where the recombinant node matches either but not both parent nodes) are shown following the same format as Fig. 2B-D. B was generated using the SNIPIT package (https://github.com/aineniamh/snipit).

Extended Data Fig. 1. Histogram of inferred and simulated recombination breakpoint positions. A) True simulated breakpoints (red) are shown with all detected recombination interval midpoints (blue). Where blue bars exceed the height of red, it implies an excess rate of detection relative to the true rate of breakpoint positions. Likewise, where red bars exceed the height of blue, it implies a deficit. **B)** True simulated breakpoints (red) are shown with detected recombination interval midpoints for the 20% of the most closely related donor-acceptor pairs (blue). In both comparisons, we broke ties between equivalently improved partial phylogenetic placement parsimony scores by selecting the largest recombination intervals.

Extended Data Fig. 2. RIPPLES more easily detects breakpoints causing large changes in parsimony score. The distribution of simulated breakpoints detected for each simulated sample is shown for each sample by **A)** initial parsimony score and **B)** minimum genetic distance from simulated sample to parent. Initial parsimony (A) is dependent upon the initial placement of the recombinant node in the tree and refers to the genetic distance in mutations between the recombinant node and its direct parent in the phylogeny. Minimum genetic distance from sample to parent (**B**) refers to the number of mutations relevant to recombination that separate the recombinant node from either the donor or the acceptor, and is not dependent on -the initial phylogeny. Similarly, among the simulated samples detected by RIPPLES, the detected and undetected breakpoints are shown by **C)** initial parsimony score and **D**) minimum genetic distance to parent. Detected samples and breakpoints are shown in black and undetected samples and breakpoints are shown in red. We condition on locating the true breakpoints and observing a significant parsimony score according to our phylogenetic null model. Therefore, we exclude recombination events with minimum starting parsimony scores and genetic distances of less than 3, as these are not significant under our null model.

Extended Data Fig. 3. Examples of detected trios filtered out due to sequence quality concerns. A) Partial alignment of consensus sequences from a filtered recombinant trio of nodes 77695, 169585, and 77690, centered on site 28225, has consensus sequences of mostly 'N' spanning several sites meant to be informative of a recombination event. This can occur when many descendant samples have missing data. Mismatches between the three consensus sequences immediately flanking this region may be the result of poor sequencing quality as well. **B)** Partial alignment of consensus sequences from a filtered recombinant trio of nodes 173213, 173209, and 173274, centered on site 16846, has 7 recombination-informative mutations in an 8-nucleotide window that are unlikely to be true mutation events, but rather an alignment artifact or a complex indel event. **C)** Partial alignment of consensus sequences from a filtered recombinant trio of nodes 293461, 293460, and 211841, centered on site 29769, has 3 mismatches in a 5-nucleotide window, immediately flanked by a large gap in the alignment and are unlikely to be true mutations.

Extended Data Fig. 4. Recombinant ancestors exhibit increased spatial and temporal overlap. A) Spatial and **B)** temporal overlap for our recombinant trios (in blue) and the null distribution (in gray), with Mann-Whitney Ranked-Sum p-values for the statistical increase in overlap for the recombinant ancestors shown on the top.

Extended Data Fig. 5. Ancestors of recombinants are genetically similar. A) The initial parsimony scores for placements of putative (red) and simulated (blue) recombinant samples. **B)** The genetic distance between inferred (red) and simulated (blue) ancestor-donor pairs that gave rise to putative or simulated recombinants.

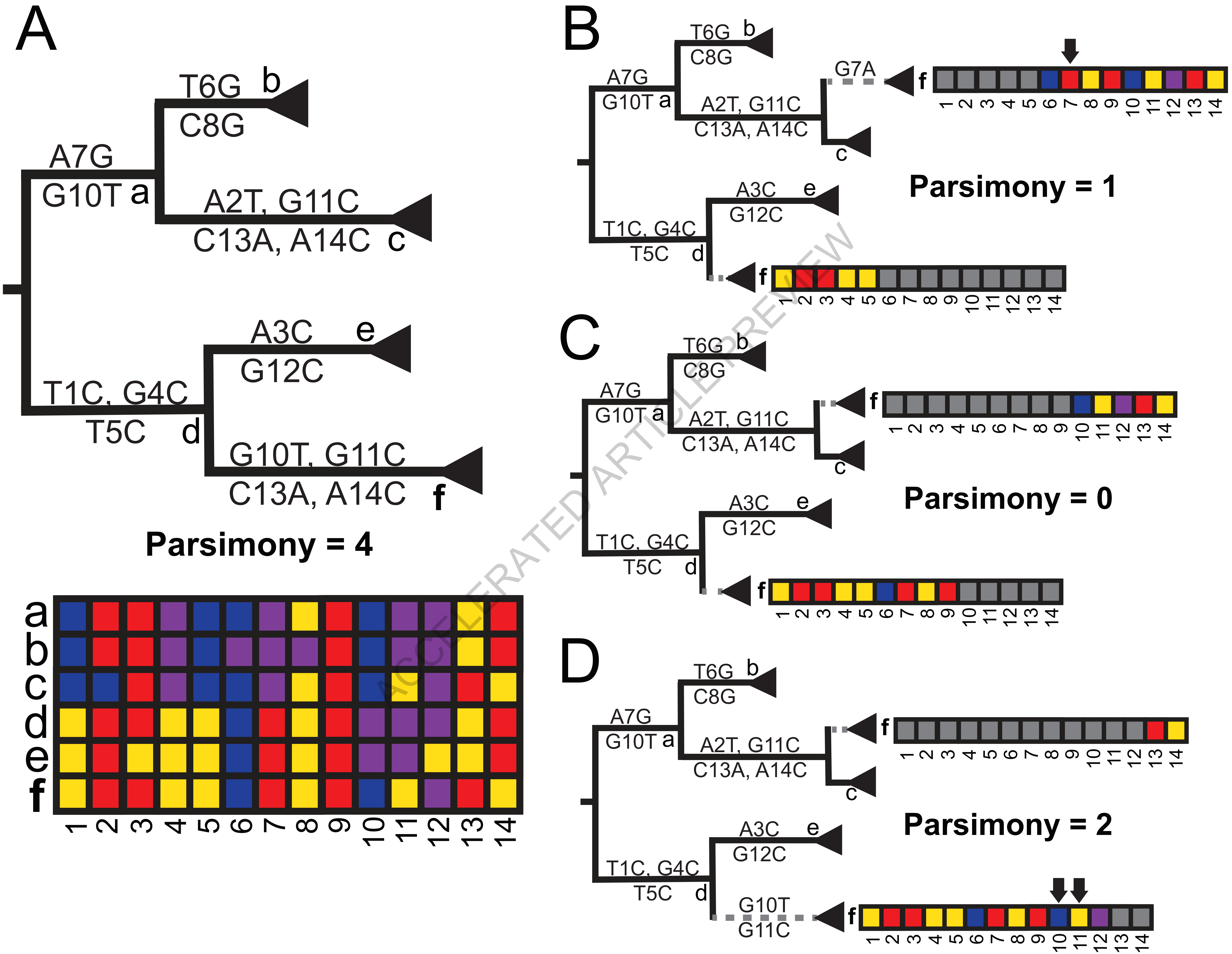
Extended Data Table 1. Summary of simulated breakpoint detection. If a simulated recombinant had only statistically insignificant parsimony improvements, it is not included here as we consider this recombination event undetectable.

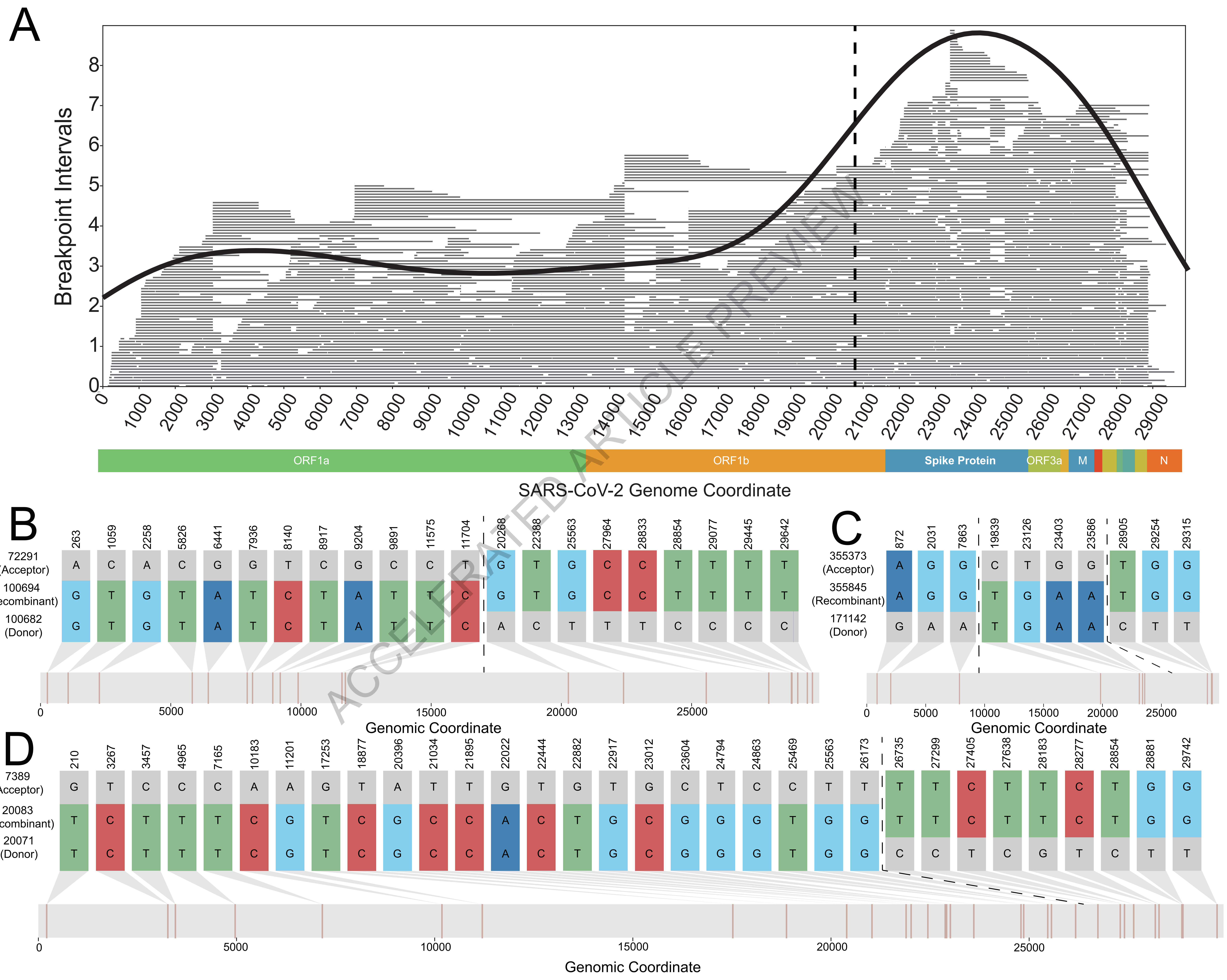
Extended Data Table 2. Raw sequence read datasets used to confirm recombination informative positions in selected recombinant samples.

Extended Data Table 3. Summary of detected recombinant nodes.

Extended Data Table 4. False discovery rate estimation for each parsimony score improvement observed in our dataset.

Extended Data Table 5. Increased rate of breakpoint interval midpoint in the 3' portion of the genome when the recombinants are subdivided by the country of origin.


A
T6G b
C8G
A7G
G10T a
A2T, G11C
C13A, A14C c
A3C e
G12C
T1C, G4C
T5C d
G10T, G11C
C13A, A14C f
Parsimony = 4
a
b
c
d
e
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
B
T6G b
C8G
A7G
G10T a
A2T, G11C
C13A, A14C
G7A
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
c
Parsimony = 1
A3C e
G12C
T1C, G4C
T5C d
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
C
T6G b
C8G
A7G
G10T a
A2T, G11C
C13A, A14C
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
c
Parsimony = 0
A3C e
G12C
T1C, G4C
T5C d
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
D
T6G b
C8G
A7G
G10T a
A2T, G11C
C13A, A14C
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14
c
Parsimony = 2
A3C e
G12C
T1C, G4C
T5C d
G10T
G11C
f
1 2 3 4 5 6 7 8 9 10 11 12 13 14



B

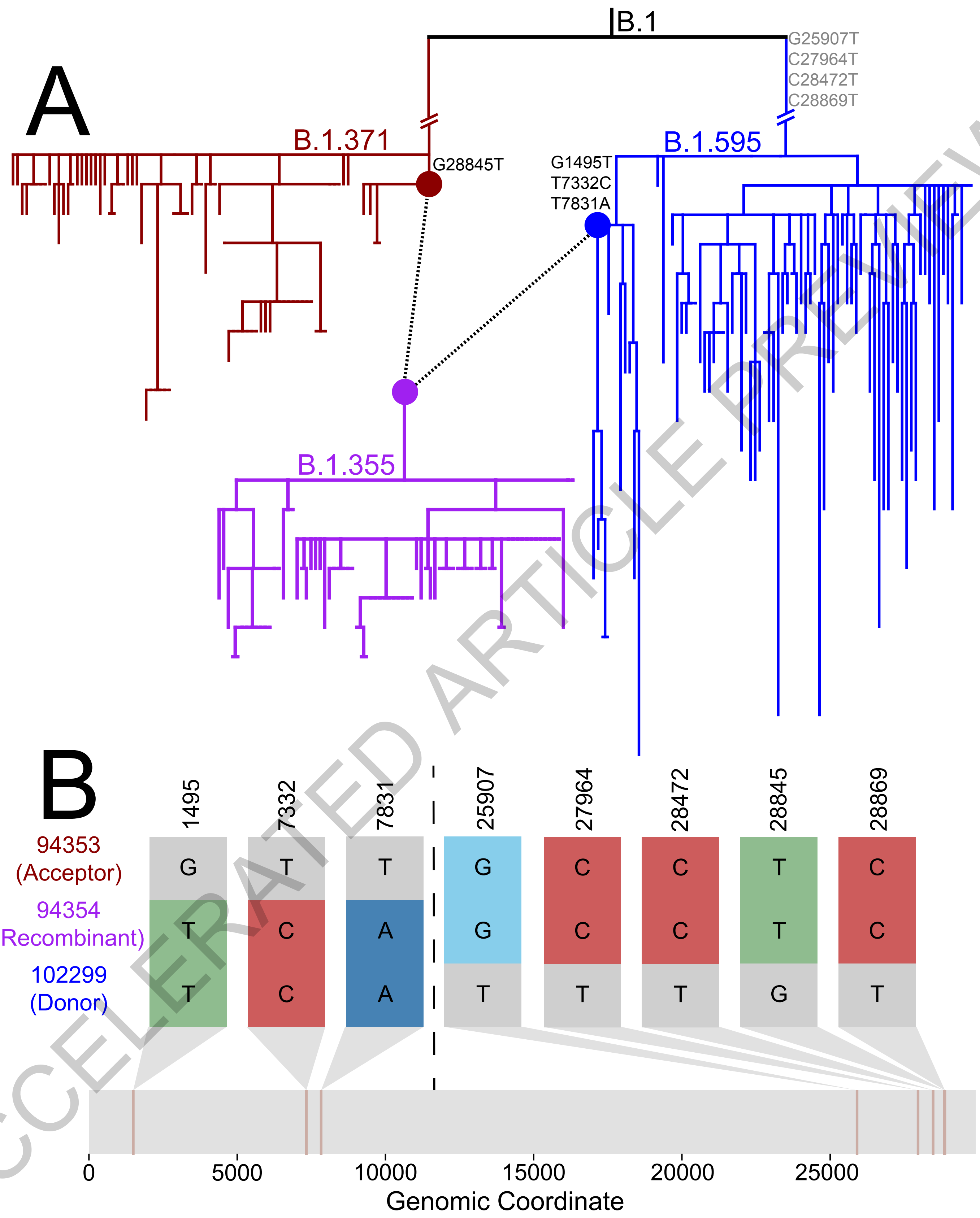
	263	1059	2258	5826	6441	7936	8140	8917	9204	9891	11575	11704	20268	22388	25563	27964	28833	28854	29077	29445	29642
72291 (Acceptor)	A	C	A	C	G	G	T	C	G	C	C	T	G	T	G	C	C	T	T	T	T
100694 (Recombinant)	G	T	G	T	A	T	C	T	A	T	T	C	G	T	G	C	C	T	T	T	T
100682 (Donor)	G	T	G	T	A	T	C	T	A	T	T	C	A	C	T	T	T	C	C	C	C

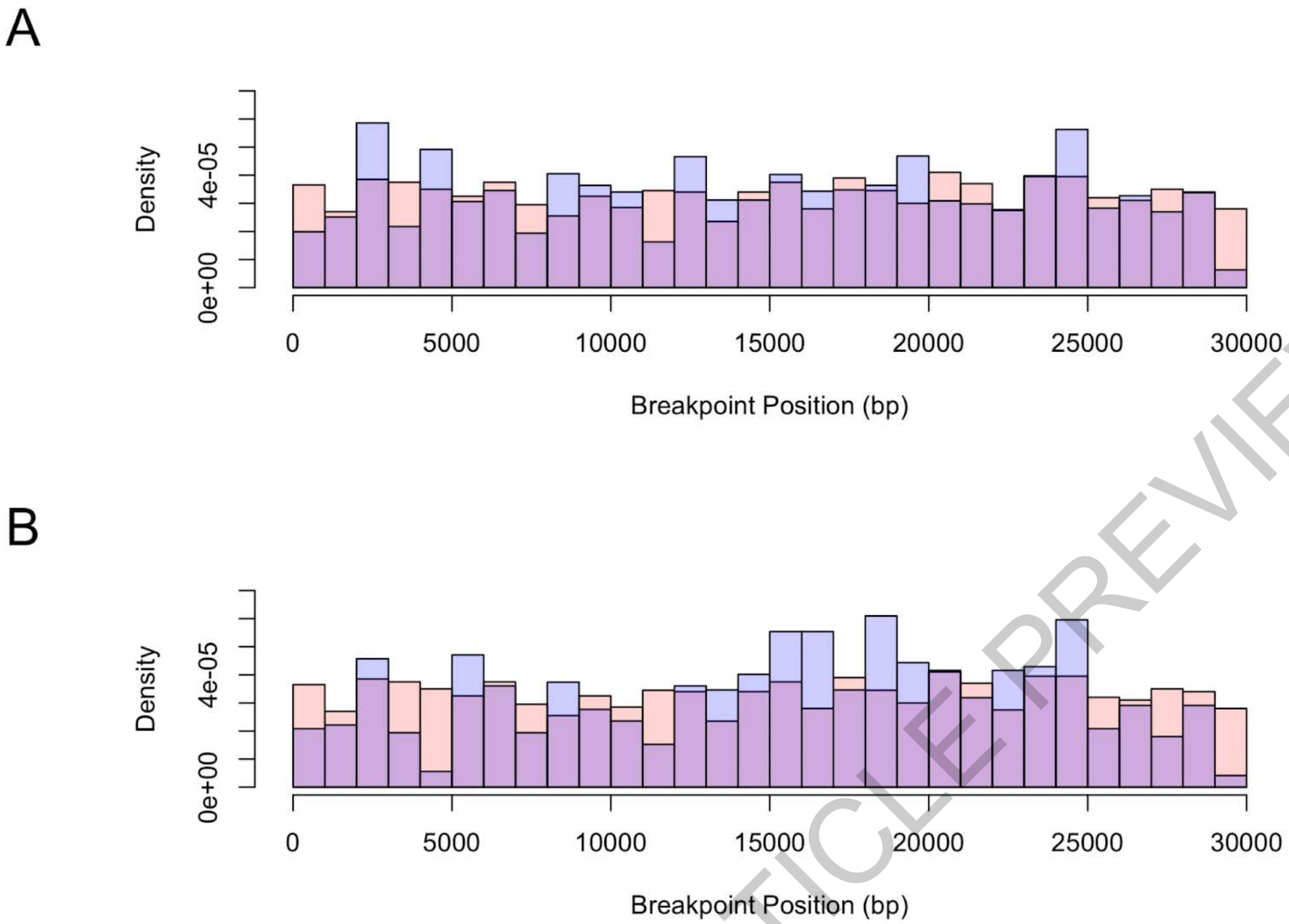
C

	872	2031	7863	19839	23126	23403	23586	28905	29254	29315
355373 (Acceptor)	A	G	G	C	T	G	G	T	G	G
355845 (Recombinant)	A	G	G	T	G	A	A	T	G	G
171142 (Donor)	G	A	A	T	G	A	A	C	T	T

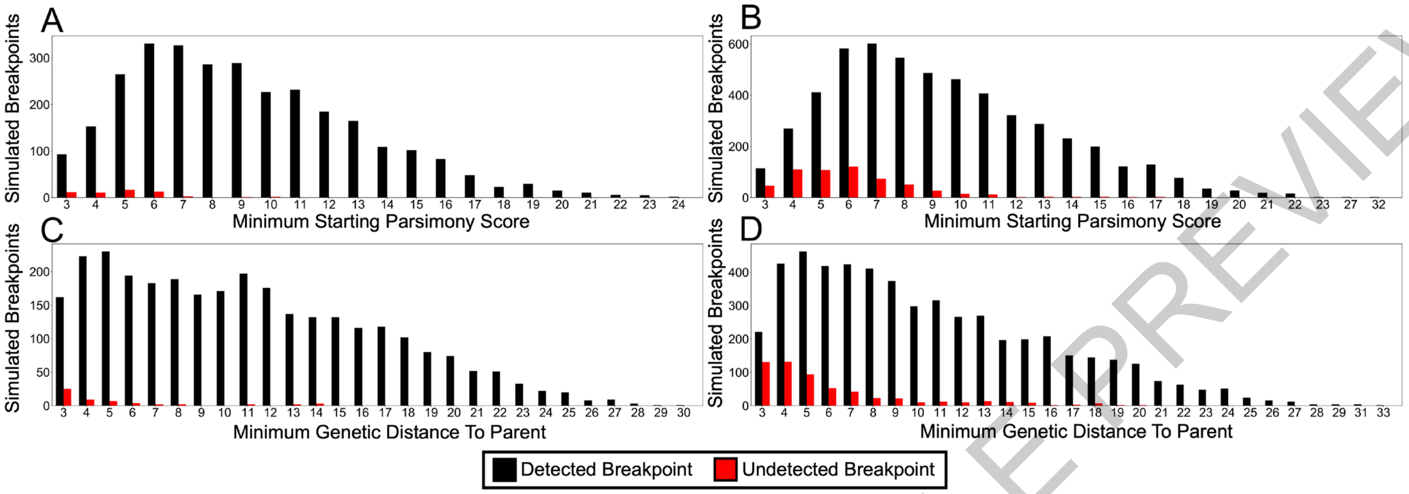
D

	210	3267	3457	4965	7165	10183	11201	17253	18877	20396	21034	21895	22022	22444	22882	22917	23012	23604	24794	24863	25469	25563	26173	26735	27299	27405	27638	28183	28277	28854	28881	29742
7389 (Acceptor)	G	T	C	C	C	A	A	G	T	A	T	T	G	T	G	T	G	C	T	C	C	T	T	T	T	C	T	T	C	T	G	G
20083 (Recombinant)	T	C	T	T	T	C	G	T	C	G	C	C	A	C	T	G	C	G	G	G	T	G	G	T	T	C	T	T	C	T	G	G
20071 (Donor)	T	C	T	T	T	C	G	T	C	G	C	C	A	C	T	G	C	G	G	G	T	G	G	C	C	T	C	G	T	C	T	T

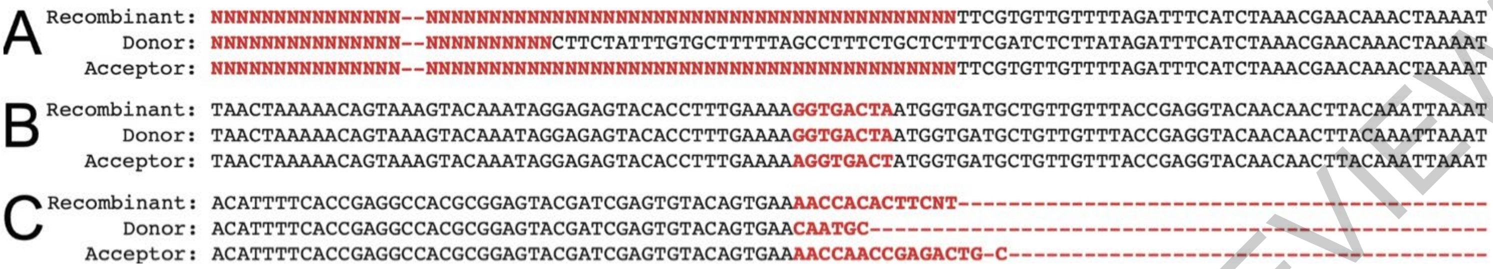

A
B.1
G25907T
C27964T
C28472T
C28869T
B.1.371
B.1.595
G28845T
G1495T
T7332C
T7831A
B.1.355
B
1495
7332
7831
25907
27964
28472
28845
28869
94353
(Acceptor)
94354
(Recombinant)
102299
(Donor)
G T T G C C T C
T C A G C C T C
T C A T T T G T
0
5000
10000
15000
20000
25000
Genomic Coordinate



Extended Data Fig. 1

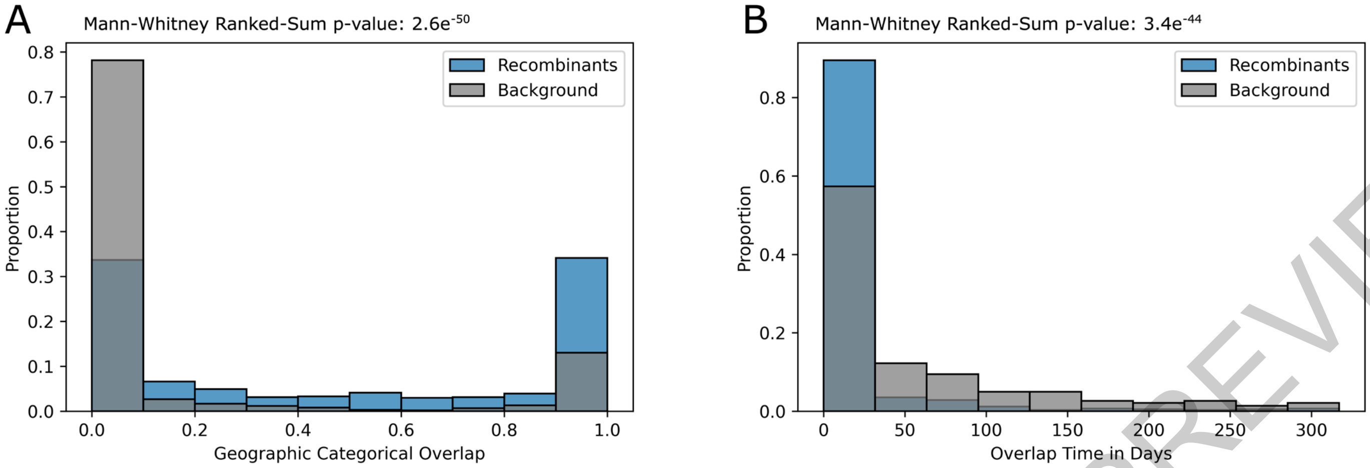


Extended Data Fig. 2

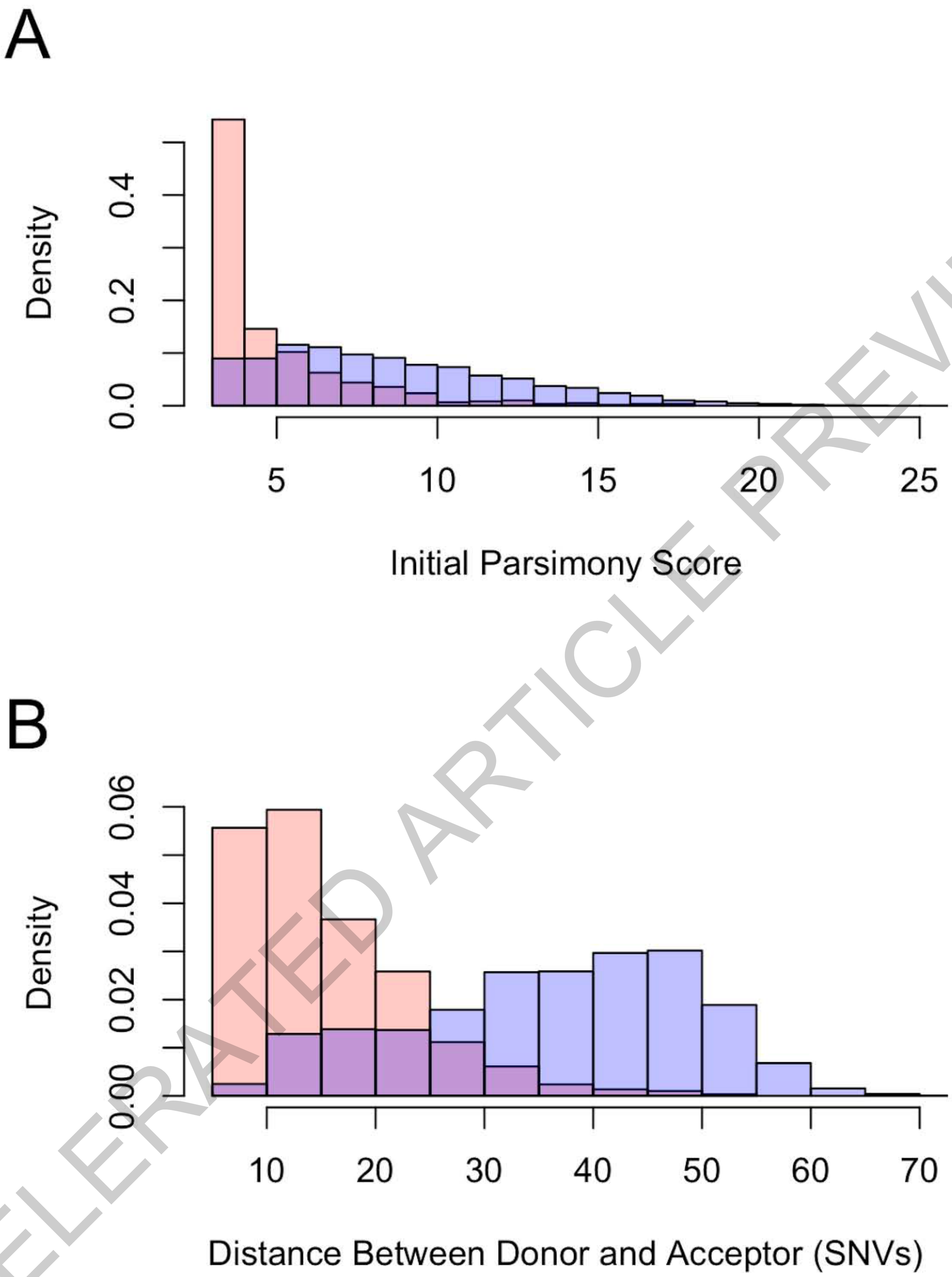


```
A Recombinant: NNNNNNNNNNNNNNN--NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCGTGTTGTTTTAGATTTCATCTAAACGAACAAACTAAAAT
        Donor: NNNNNNNNNNNNNNN--NNNNNNNNNNCTTCTATTTGTGCTTTTTAGCCTTTCTGCTCTTTCGATCTCTTATAGATTTCATCTAAACGAACAAACTAAAAT
     Acceptor: NNNNNNNNNNNNNNN--NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCGTGTTGTTTTAGATTTCATCTAAACGAACAAACTAAAAT

B Recombinant: TAACTAAAAACAGTAAAGTACAAATAGGAGAGTACACCTTTGAAAAGGTGACTAATGGTGATGCTGTTGTTTACCGAGGTACAACAACTTACAAATTAAAT
        Donor: TAACTAAAAACAGTAAAGTACAAATAGGAGAGTACACCTTTGAAAAGGTGACTAATGGTGATGCTGTTGTTTACCGAGGTACAACAACTTACAAATTAAAT
     Acceptor: TAACTAAAAACAGTAAAGTACAAATAGGAGAGTACACCTTTGAAAAAGGTGACTATGGTGATGCTGTTGTTTACCGAGGTACAACAACTTACAAATTAAAT

C Recombinant: ACATTTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAAACCACACTTCNT------------------------------------------
        Donor: ACATTTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACAGTGAACAATGC-------------------------------------------------
     Acceptor: ACATTTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAAACCAACCGAGACTG-C--------------------------------------
```

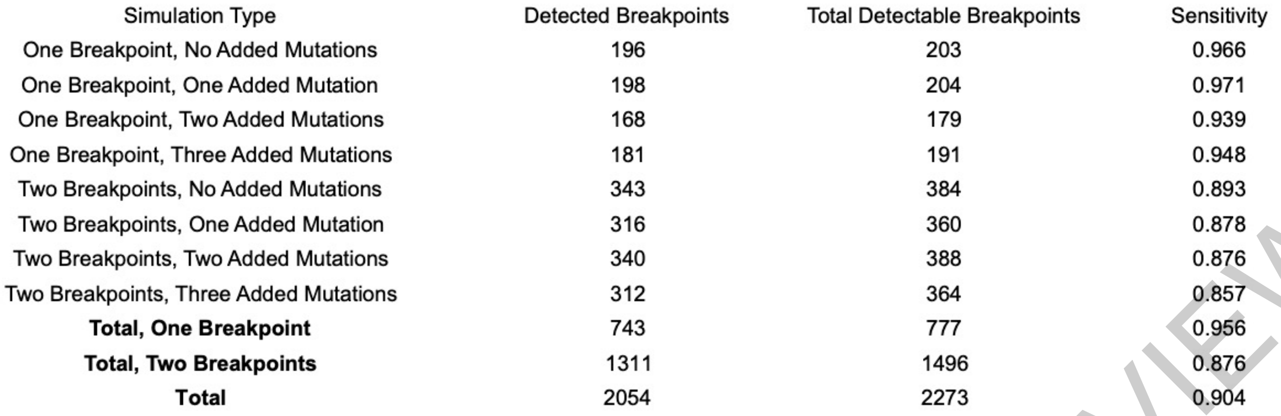
Extended Data Fig. 3



Extended Data Fig. 4

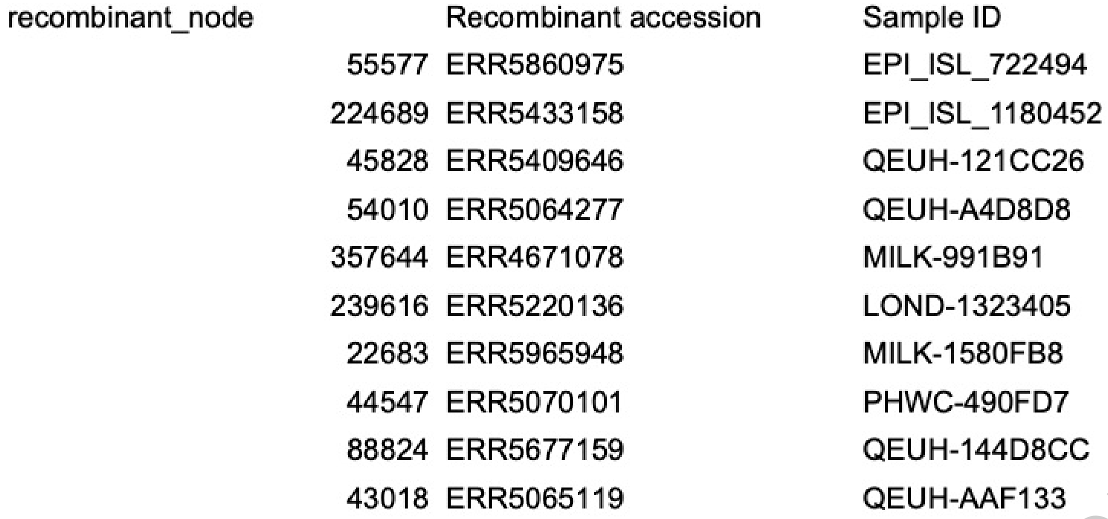


Extended Data Fig. 5



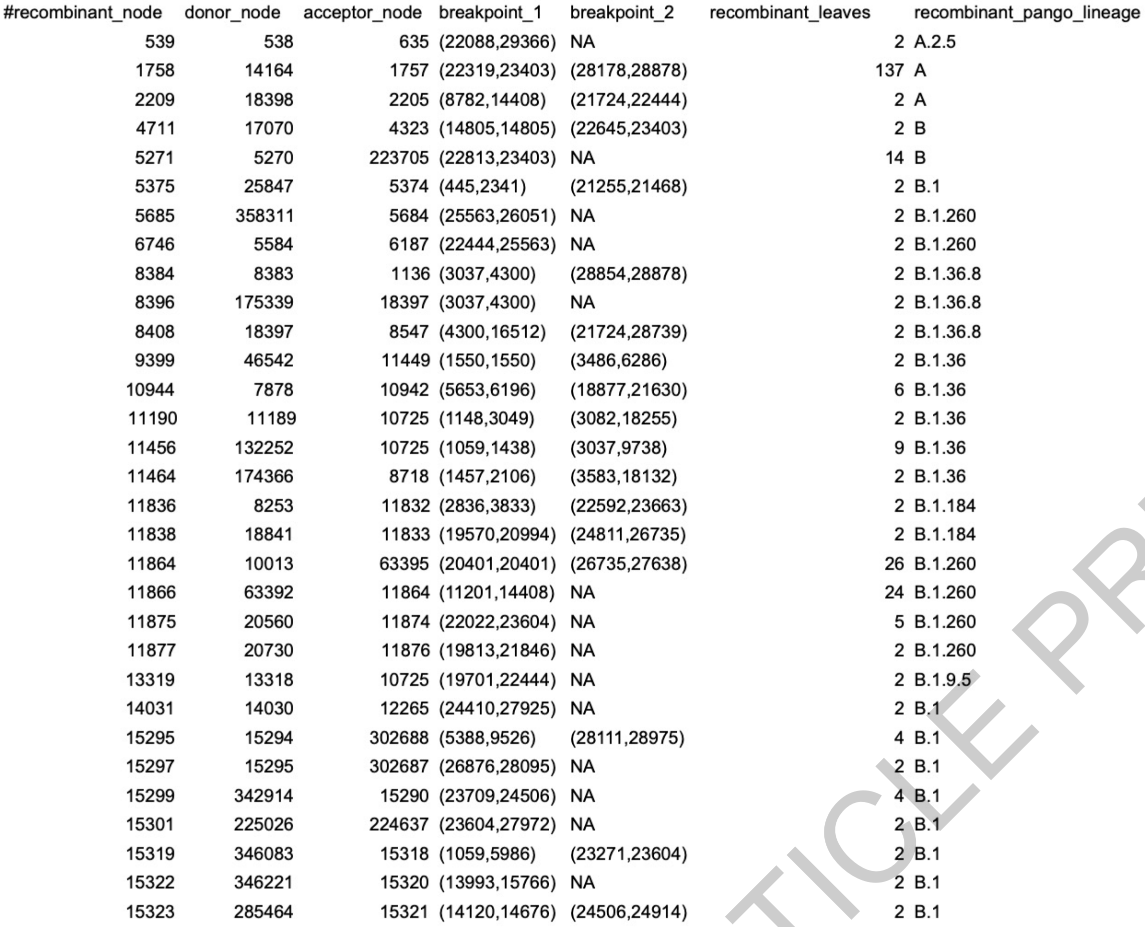
Simulation Type	Detected Breakpoints	Total Detectable Breakpoints	Sensitivity
One Breakpoint, No Added Mutations	196	203	0.966
One Breakpoint, One Added Mutation	198	204	0.971
One Breakpoint, Two Added Mutations	168	179	0.939
One Breakpoint, Three Added Mutations	181	191	0.948
Two Breakpoints, No Added Mutations	343	384	0.893
Two Breakpoints, One Added Mutation	316	360	0.878
Two Breakpoints, Two Added Mutations	340	388	0.876
Two Breakpoints, Three Added Mutations	312	364	0.857
Total, One Breakpoint	743	777	0.956
Total, Two Breakpoints	1311	1496	0.876
Total	2054	2273	0.904

Extended Data Table 1



recombinant_node	Recombinant accession	Sample ID
55577	ERR5860975	EPI_ISL_722494
224689	ERR5433158	EPI_ISL_1180452
45828	ERR5409646	QEUH-121CC26
54010	ERR5064277	QEUH-A4D8D8
357644	ERR4671078	MILK-991B91
239616	ERR5220136	LOND-1323405
22683	ERR5965948	MILK-1580FB8
44547	ERR5070101	PHWC-490FD7
88824	ERR5677159	QEUH-144D8CC
43018	ERR5065119	QEUH-AAF133

Extended Data Table 2

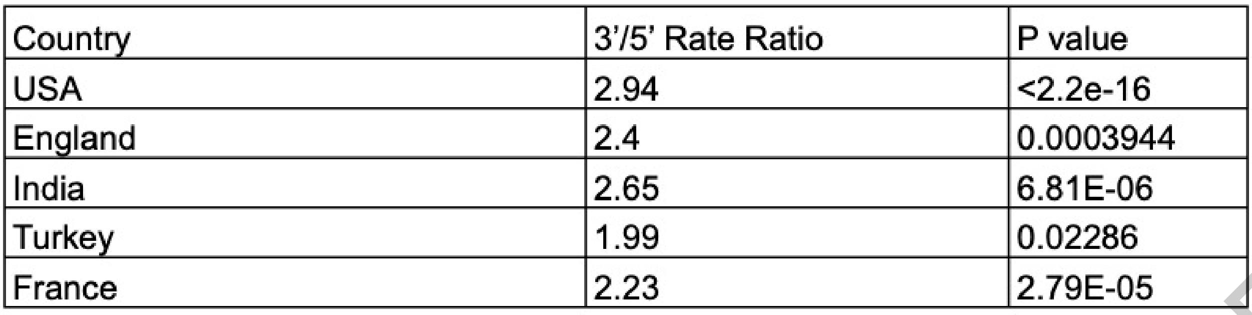


#recombinant_node	donor_node	acceptor_node	breakpoint_1	breakpoint_2	recombinant_leaves	recombinant_pango_lineage
539	538	635	(22088,29366)	NA	2	A.2.5
1758	14164	1757	(22319,23403)	(28178,28878)	137	A
2209	18398	2205	(8782,14408)	(21724,22444)	2	A
4711	17070	4323	(14805,14805)	(22645,23403)	2	B
5271	5270	223705	(22813,23403)	NA	14	B
5375	25847	5374	(445,2341)	(21255,21468)	2	B.1
5685	358311	5684	(25563,26051)	NA	2	B.1.260
6746	5584	6187	(22444,25563)	NA	2	B.1.260
8384	8383	1136	(3037,4300)	(28854,28878)	2	B.1.36.8
8396	175339	18397	(3037,4300)	NA	2	B.1.36.8
8408	18397	8547	(4300,16512)	(21724,28739)	2	B.1.36.8
9399	46542	11449	(1550,1550)	(3486,6286)	2	B.1.36
10944	7878	10942	(5653,6196)	(18877,21630)	6	B.1.36
11190	11189	10725	(1148,3049)	(3082,18255)	2	B.1.36
11456	132252	10725	(1059,1438)	(3037,9738)	9	B.1.36
11464	174366	8718	(1457,2106)	(3583,18132)	2	B.1.36
11836	8253	11832	(2836,3833)	(22592,23663)	2	B.1.184
11838	18841	11833	(19570,20994)	(24811,26735)	2	B.1.184
11864	10013	63395	(20401,20401)	(26735,27638)	26	B.1.260
11866	63392	11864	(11201,14408)	NA	24	B.1.260
11875	20560	11874	(22022,23604)	NA	5	B.1.260
11877	20730	11876	(19813,21846)	NA	2	B.1.260
13319	13318	10725	(19701,22444)	NA	2	B.1.9.5
14031	14030	12265	(24410,27925)	NA	2	B.1
15295	15294	302688	(5388,9526)	(28111,28975)	4	B.1
15297	15295	302687	(26876,28095)	NA	2	B.1
15299	342914	15290	(23709,24506)	NA	4	B.1
15301	225026	224637	(23604,27972)	NA	2	B.1
15319	346083	15318	(1059,5986)	(23271,23604)	2	B.1
15322	346221	15320	(13993,15766)	NA	2	B.1
15323	285464	15321	(14120,14676)	(24506,24914)	2	B.1

Extended Data Table 3

Starting Parsimony	Improvement	Nodes in Tree	P-value	Expected False Discoveries	Actual Discoveries
3	3	25670	0.0005373455132	13.79365932	187
4	3	10948	0.0005373455132	5.882858678	106
4	4	10948	0.0005373455132	5.882858678	27
5	3	5206	0.001590668081	8.281018028	44
5	4	5206	0.0005302226935	2.760339343	30
5	5	5206	0.0005302226935	2.760339343	12
6	3	2654	0.001143510577	3.034877073	33
6	4	2654	0.0005717552887	1.517438536	15
6	5	2654	0.0005717552887	1.517438536	9
6	6	2654	0.0005717552887	1.517438536	3
7	3	1456	0.002528445006	3.681415929	21
7	4	1456	0.0006321112516	0.9203539823	7
7	5	1456	0.0006321112516	0.9203539823	4
7	6	1456	0.0006321112516	0.9203539823	3
7	7	1456	0.0006321112516	0.9203539823	2
8	3	796	0.003248862898	2.586094867	13
8	4	796	0.0006497725796	0.5172189734	5
8	5	796	0.0006497725796	0.5172189734	4
8	6	796	0.0006497725796	0.5172189734	3
8	7	796	0.0006497725796	0.5172189734	1
9	3	455	0.002652519894	1.206896552	7
9	4	455	0.0006631299735	0.3017241379	5
9	5	455	0.0006631299735	0.3017241379	1
9	6	455	0.0006631299735	0.3017241379	3
9	7	455	0.0006631299735	0.3017241379	3
9	8	455	0.0006631299735	0.3017241379	1
9	9	455	0.0006631299735	0.3017241379	1
10	3	267	0.005365526492	1.432595573	6
10	4	267	0.0006706908115	0.1790744467	4
10	5	267	0.0006706908115	0.1790744467	1
10	6	267	0.0006706908115	0.1790744467	1

Extended Data Table 4



Country	3’/5’ Rate Ratio	P value
USA	2.94	<2.2e-16
England	2.4	0.0003944
India	2.65	6.81E-06
Turkey	1.99	0.02286
France	2.23	2.79E-05

Extended Data Table 5

nature research

Corresponding author(s): Yatish Turakhia
Russell Corbett-Detig

Last updated by author(s): August 18, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data used in this work are available from GISAID (gisaid.org), COG-UK, and Genbank, with specific sample accessions listed in Supplemental Tables 5-8.
Data analysis	The data was analyzed using code available at https://github.com/yatisht/usher and https://github.com/bpt26/recombination. All software versions are indicated where appropriate in the methods section of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in this work are available from GISAID (gisaid.org), COG-UK, and GenBank, with specific sample accessions listed in Supplemental Tables 5-8.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences   □ Behavioural & social sciences   ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In this study, we describe an efficient method that exhaustively searches a phylogeny with applications demonstrated for the current SARS-CoV-2 global phylogeny. We compared our approach to many existing methods and documented accuracy (on simulated data), consistency (with empirical data), compute time and memory usage requirements.
Research sample	Our study is based on existing dataset of SARS-CoV-2 sequences shared via GISAID (gisaid.org), GenBank, and COG-UK. The specific sample accessions are listed in Supplementary Tables 5-8.
Sampling strategy	Not relevant. We chose to work primarily with our 28/5/21 public release of the SARS-CoV-2 phylogeny, because in order to develop our software, we needed a constant tree to perform experiments on and these were the most up-to-date available at the time we began this work. We also worked with simulated data, designed to behave similarly to the real data, as described in our Methods section.
Data collection	All sequences marked as 'complete' and 'high coverage' submitted up to 28/5/21 were downloaded from GISAID (gisaid.org), as well as sequences from GenBank, and COG-UK, were used to build the global phylogeny after a few additional filtering steps (Methods). These data are from a collection of sequences obtained throughout the world during the SARS-CoV-2 pandemic. Supplementary Tables 5-8 list all individuals responsible for the primary data collection in all sequences used in this study.
Timing and spatial scale	All sequences present in the 28/5/2021 public tree were used, except for those pruned out according to our Methods section. We chose 28/5/21 because we needed a consistent sample with which to hone our methods and conduct experiments, as well as to have a "reference tree" to refer back to throughout the study.
Data exclusions	Incomplete and low-coverage sequences as well as those with known sequence issues were excluded (Methods). Our previous study and other related studies cited in the Methods demonstrate that errors can lead to false nucleotide substitutions for myriad reasons unrelated to the biology of the virus itself. We have masked these sites from our analysis and the specific criteria for exclusion are indicated in the method section.
Reproducibility	All our findings and results are completely reproducible using the code and data available from https://github.com/yatisht/usher. Simulations and filtration of sequences were conducted using code from https://github.com/bpt26/recombination.
Randomization	Not relevant. We used identical dataset for all comparative analysis hence randomization is not necessary for comparing results of the approaches used in this study.
Blinding	Blinding is not relevant because experimenter bias cannot affect the results of this analysis.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging