Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie

Ales Varabyou [1,2*†], Christopher Pockrandt [1,3*†], Steven L. Salzberg [1,2,3,4], and Mihaela Pertea [1,2,3*]

1 Center for Computational Biology, Johns Hopkins University, Baltimore, MD, 21211, USA

2 Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA

3 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA

4 Department of Biostatistics, Johns Hopkins University, Baltimore, MD, 21205, USA

† These authors contributed equally to this work.

*Corresponding author.

Email: ales.varabyou@jhu.edu, pockrandt@jhu.edu, mpertea@jhu.edu

Running title: SARS-CoV-2 inter-clade recombination

Keywords: SARS-CoV-2, COVID-19, recombination, coronavirus

Abstract

The ability to detect recombination in pathogen genomes is crucial to the accuracy of phylogenetic analysis and consequently to forecasting the spread of infectious diseases and to developing therapeutics and public health policies. However, in case of the SARS-CoV-2, the low divergence of near-identical genomes sequenced over a short period of time makes conventional analysis infeasible. Using a novel method, we identified 225 anomalous SARS-CoV-2 genomes of likely recombinant origins out of the first 87,695 genomes to be released, several of which have persisted in the population. Bolotie is specifically designed to perform a rapid search for inter-clade recombination events over extremely large datasets, facilitating analysis of novel isolates in seconds. In cases where raw sequencing data was available, we were able to rule out the possibility that these samples represented co-infections by analyzing the underlying sequence reads. The Bolotie software and other data from our study are available at https://github.com/salzberg-lab/bolotie.

Introduction

Since the beginning of 2020, the COVID-19 pandemic caused by a newly emerged strain of a betacoronavirus, SARS-CoV-2, has been responsible for over 950,000 deaths and over 30 million infections to date (Dong *et al.* 2020). The strain has been hypothesized to have emerged as a result of a recombination event between strains of betacoronavirus endemic to certain species of bats and pangolins (Zhang *et al.* 2020), although its precise origin is not yet known. To date, the genetic diversity of the SARS-CoV-2 has been increasing slowly compared to other RNA viruses, with 5 to 7 major circulating clades being identified based on multiple variants common to large numbers of isolates in the GISAID database (Shu and McCauley 2017; Hadfield *et al.* 2018). This relative stability of the genetic content of the circulating forms of the virus is promising for the development of vaccines and therapeutics, as well as general understanding of the biology and pathology of SARS-CoV-2.

However, as with other RNA viruses, coronaviruses are known to undergo mutations at high rates (Drake and Holland 1999). Inter- and intra-host recombinations are also well-studied and occur frequently (Su *et al.* 2016). As more mutations and lineages of SARS-CoV-2 get fixed in the population, a recombination event caused by a co-infection of a single patient with particles of distinct clades may lead to emergence of novel lineages, posing risks to the efficacy of future treatments. In fact, several accounts of recombination events in SARS-CoV-2 have been reported in recent months (VanInsberghe *et al.* 2020). As such, rapid and consistent surveillance of the sequenced genomes of SARS-CoV-2 for both novel mutations and recombinations is critical to the development of effective treatments and vaccines (Demİr *et al.* 2020).

Multiple computational methods have been developed to detect recombination in microbial genomes and have been used in studies of HIV-1 mutagenesis, bacterial evolution, and other applications (Posada 2002). Some popular methods, such as 3seq, analyze every possible triplet of a set of genomes and statistically evaluate the triplets (Lam *et al.* 2018). Other methods, like PhiPack, are designed to work for low-divergence genomes but still require a significant number of variants (1-5%) to perform statistical analysis (Bruen *et al.* 2006). Another limitation of some of these methods is that the algorithms are computationally intensive and require significant resources and time to perform analysis, particularly for the amount of data (nearly 300,000 complete high-coverage genomes to date) that has been generated for SARS-CoV-2. There exist over 4 quadrillion unique triplets of sequences for the currently available SARS-CoV-2 genomes, and a similarity analysis for each triplet would be computationally infeasible. A more efficient approach is necessary if we want to be able to detect recombinants in a realistic amount of time.

In this work we present Bolotie, a new algorithm designed to conduct mutational analysis and to detect recombinant forms and other anomalies among a very large set of viral sequences. The methods presented are also designed such that novel sequences can be analyzed efficiently without the need to rerun the entire protocol.

We applied Bolotie to search for recombination events in 87,695 complete genomes of SARS-CoV-2 currently available in the GISAID database (Shu and McCauley 2017). In our analysis we identified multiple unique cases of recombination between 4 prominent clades of the virus. Several of the identified recombination events, including some previously reported (VanInsberghe *et al.* 2020), appear in multiple isolates, suggesting transmission in the population.

Lastly, we propose a methodology for distinguishing true recombination events from cases of mis-assembly of isolates from a host co-infected with several distinct lineages of a pathogen. The proposed method can be applied by to verify future SARS-CoV-2 genomes prior to database submission.

Materials and Methods

In this work we present Bolotie, a collection of methods that enables rapid alignment, variant calling, inter-clade recombination detection, and parent sequence search for large sets of assembled viral genomes. Our method is robust even when the divergence of collected genomes is very small, and it is designed for ultra-fast detection of anomalies. Because Bolotie utilizes a probability index that can be used for a one-against-all analysis, alignments and indices do not have to be recomputed when evaluating novel sequences as recombinants, which greatly increases the efficiency of single-sequence analysis.

Data

304,811 complete high-coverage genomes were obtained from GISAID (Shu and McCauley 2017) using the provided interface. The detailed analysis and evaluation of the software presented in this work is based on a subset of this dataset of 87,695 genomes collected before October 2020. SARS-CoV-2 reference genome isolate Wuhan-Hu-1 (GenBank, accession no. MN908947) was obtained from NCBI and used to guide the alignment, variant calling and consensus sequence generation in the protocol.

Clade assignments for all available genomes were obtained from GISAID (Shu and McCauley 2017). Lineages S, L, V and O were grouped together, similarly to NextStrain (Hadfield *et al.* 2018), and based on the close distances observed in our independent phylogenetic analyses. Mappings between clade IDs in our analysis, those defined by GISAID, and those defined by NextStrain are provided in Supplementary Table 1.

To test genomes for the possibility that the putative recombinant isolates might represent a co-infection with two or more distinct SARS-CoV-2 isolates, we searched the SRA database of raw read data using all GISAID and lab-assigned identifiers.

Computing Alignments

A global alignment method to the reference using the implementation from the KSW2 library (Li 2018; Suzuki and Kasahara 2018) was developed to facilitate efficient and parallelized variant calling for each query sequence. Because sequence divergence is very low in the collection of SARS-CoV-2 genomes, the reference-guided approach used instead of the conventional multiple sequence alignment is 1) significantly faster, 2) allows simple addition of new sequences, and 3) does not cause explosions in gaps at 3’ and 5’ ends of the viral genomes. Alignment was performed with a DNAFULL scoring matrix, a gap penalty of 12 and gap extension penalty of 4.

Computing Consensus Sequences

After the pairwise alignment step, we construct for every genome in the input set a consensus sequence by substituting reference genome alleles for the high-frequency variants called by the aligner. Not only does this approach allow us to filter variants used to search for recombinants, but also produces a set of genomes with a standardized set of coordinates which can be used as a multiple sequence alignment (MSA) for phylogenetic analysis.

Because the 3' and 5' ends of viral genomes are notoriously difficult to assemble correctly, we chose to force the first and last 200 bases in the consensus sequence of each genome to be identical to the reference genome in accordance with previous studies (VanInsberghe *et al.* 2020).

Next, since Bolotie is designed to work for genomes with very few variants it is particularly sensitive to ambiguous nucleotides. To avoid biases caused by uncalled bases, any such instances were treated as an unknown base (N). Furthermore, to avoid bias in our predictions for all sequences in the dataset we replace nucleotides at a position of low-frequency variants with the reference allele making such positions equally probable for any clade (clade neutral). We define a low-frequency SNV as one that has fewer than 100 genome sequences that differ from the reference sequence at that position.

Because we did not use structural variants when constructing consensus sequences, the final collection of filtered genomes represents a MSA. This not only allows us to use it directly for phylogenetic analysis but also to compute distances among sequences efficiently.

Identifying Recombinants

Based on the provided clade information for each sequence, a model is created to evaluate each sequence as a potential recombinant. From the MSA the conditional probability for each nucleotide position is computed; i.e., the probability of this position belonging to clade C_i given the nucleotide b. However, to account for differences in clade sizes, we multiply the base counts at each position for each clade with the reciprocal of the number of sequences in that clade. The algorithm also ensures that any ambiguous character is assigned a neutral conditional probability of $\frac{1}{|C|}$. For every position in the MSA we now have the normalized conditional probabilities of that base position belonging to a certain clade given the base observed in the consensus sequence.

Let $C = c_1, c_2, \dots, c_k$ be the clades (in this paper: k=4) with $c_i \in C$. We define two sets of bases$B_4 = \{A, C, G, T\}$ and B_{16} being all IUPAC characters. $n_{c_i,b}$ denotes the number of sequences in clade c_i that have base $b \in B_4$ at the position of interest. Now the weighted conditional probability$Pr(c_i|b)$of the sequence belonging to clade c_i given the observed base $b \in B_{16}$ at a certain position is defined as:

$$Pr(c_i \mid b) = \begin{cases} \frac{n_{c_i,b} * w_{c_i}}{\sum_{c_j \in C} n_{c_j,b} * w_{c_j}}, & b \in \mathcal{B}_4 \\ \frac{1}{|C|}, & otherwise\ (i.e., b\ is\ ambiguous) \end{cases} \quad (1)$$

$$w_{c_i} = \frac{1}{\sum_{b \in \mathcal{B}_4} n_{c_i,b}} \quad (2)$$

Now we determine for each sequence whether it might be a recombinant. To do this we need to determine for each base the most likely clade. This can be modeled as an HMM and solved efficiently (in both time and space) with the Viterbi algorithm in $O(nc^2)$ time, where *n* is the sequence length and *c* is the number of clusters (Forney 1973). The Viterbi algorithm determines the most likely sequences of clades that should be assigned to the positions in the genome, by finding the sequence of assignments that maximizes the posterior probability across the space of all possible clade assignments. This probability depends on the prior probabilities computed for each position to be in a clade, and the probability of switching from one clade to another. We assigned a very small value to this probability (0.0001, shown in Figure 1), on the assumption that a recombinant is an unlikely event. We empirically tried several larger and smaller values and found that they made almost no difference in the results.

At each position of the MSA, each state is representing one clade with the conditional probabilities for every four bases. The Viterbi algorithm then finds the path with the highest likelihood given the sequence of an isolate. The model is identical for every isolate and only depends on the MSA. Figure 1 shows the approach, but only illustrates the conditional probability of the observed base of the sequence in each state.

This method cannot detect the exact breakpoint location for a recombinant, because it relies on discrete SNV differences between clades; thus, it can only narrow down the breakpoint to the region between two clade-specific SNVs.

Searching for Closest Parental Genomes

Previous methods for recombination detection examine triplets of sequences to detect a potential recombinant genome and its parents. Unlike such methods, our algorithm relies on conditional probabilities computed for clades – a strategy that provides additional statistical power and also reduces the complexity of the problem.

To identify potential parents of recombinant sequences we have implemented a dedicated method. For each recombined segment of the sequence, as inferred by our algorithm, we compute a Kimura distance matrix (Kimura 1980) to other sequences in the clade of the corresponding segment. Kimura distance is a distance metric that scores transitions (A <-> G and C <-> T) differently than transversions (interchange of purine for pyrimidine bases). To be more "accurate" we also include low-frequency SNVs that we neglected in the previous step in the consensus sequences. All sequences with the lowest distance score are reported as most likely parents that have contributed the segment in the recombination event.

Investigating sequencing data

To test our recombination candidates for signs of co-infection, we aligned available reads with Bowtie2 (Langmead and Salzberg 2012) against the Wuhan-Hu-1 reference genome using the "--very-sensitive-local" option and otherwise default parameters. The mappings were further sorted and indexed using samtools (Li *et al.* 2009). Counts for individual nucleotides were obtained using bam-readcount software (https://github.com/genome/bam-readcount) and positions with high conditional probabilities were extracted and summarized.

Phylogenetic analysis

To test how well information is preserved in our consensus sequences, we obtained a pre-computed tree from NextStrain (Hadfield *et al.* 2018) which contained a total of 4,494 representative genomes chosen by NextStrain. Only 4,039 genomes from those in the

NextStrain tree were available on GISAID at the time when we obtained genome assemblies for our analysis.

First, we re-built the tree using the set of 4,039 genomes using the general time-reversible (GTR) model as used by the NextStrain platform and allowing IQ-TREE (Minh *et al.* 2020) to automatically choose the precise model using its ModelFinder package (Kalyaanamoorthy *et al.* 2017). The final tree was generated under the GTR model with 1000 rounds of bootstrapping. The same approach was taken to build the phylogenetic tree that included all identified anomalous sequences.

Results

We first aligned all genomes to the Wuhan-Hu-1 reference genome (GenBank accession MN908947), from which we detected 84,322 single-nucleotide variants (SNVs) at 29,503 sites. After removing all variants that appear in fewer than 100 sequences, we retained a set of 659 SNVs at 411 unique sites. Alignments also revealed 1,349 unique structural variants (934 deletions and 415 insertions). While 2 deletions and 1 insertion were present in over 100 genomes, for the purpose of computational efficiency we did not consider them further.

Using Bolotie on the set of well-supported variants to search for recombination events between sequences in the 4 major clades of SARS-CoV-2, we identified 225 possibly recombinant genomes. Figure 2 illustrates that many of the identified recombination events were represented by a single genome. However, several lineages with near-identical sequences were observed. In Figure 2 these lineages appear as broad red bands with a high density of outgoing arcs. Additionally, several smaller groups of near-identical recombinant signatures have been observed in which genomes differed by one or two variants. Those genomes were often found to be neighbors in the computed maximum likelihood (ML) trees and often had the same or neighboring inferred parental sequences.

Of the 225 recombinant genomes, 109 were labeled in the original GISAID data as belonging to clade #0, 111 in clade #1 and 5 in clade #2 (Figure 2). Recombination events happened between members of all 4 clades, with 171 parental genomes identified in clade #0, 41 parental genomes in clade #1, 148 in clade #2 and 90 in clade #3. Additionally, 15 out of 225 potential recombinants were found present in the set of 4,039 representative genomes used by NextStrain (Supplementary Table 2).

Of the 225 identified recombinants, most of the recombinant signatures had 1 or 2 breakpoints like the ones shown in Figure 3A, 3B and 3C. However, at least 6 genomes including the one depicted in Figure 3D exhibited more complex patterns of mosaicism with 3 breakpoints.

While in majority of cases the path-finding algorithm of Bolotie relied on clade-defining variants with high conditional probabilities > 0.9, several positions exhibited an inverse pattern and were also useful in the analysis. For example, in Figure 3A (further detailed in Table 1), mutations of cytosine (C) to thymine (T) at position 14,407 and adenine (A) to guanine (G) at position 23,402 are not characteristic of clade 1 (yellow) since they have the same conditional probability of ~0.33 of defining clades 2 and 3. However, these positions are informative in an inverse way, namely that observing a C and an A at those positions is very unlikely if the sequence originated in clade 0. This adds additional evidence to the recombinant origins of the genomes.

Even though most anomalous sequences reported by Bolotie had clean separations between two parental clades, some mutational signatures contained admixtures in conditional probabilities from other clades. As illustrated in Figure 3C, at position 240 the conditional probability has a greater affinity towards a blue clade. A signature like that could be indicative of a random mutation, sequencing or assembly artifacts, and Bolotie resolved the parental clades in a seemingly parsimonious way.

Finally, we have applied the same analysis to the most recent set of genomes available through GISAID. As of February 2021 a total of 304,811 complete high-coverage genomes were available on GISAID with classifications into 5 clades. We analyzed new data with Bolotie using the same set of parameters, except for increasing the minimum allele frequency from 100 to 500 genomes to account for the larger total dataset size. Despite identifying 775 anomalous sequences in the data, the rate at which anomalous sequences are being sequenced and deposited on GISAID remains constant (Supplementary Figure 2, Supplementary Tables 4,6) and no widely spread clusters of recombinant genomes were identified.

Sequencing and assembly artifacts

Although recombination has been extensively observed within and between members of the coronaviridae family, such observations were in the past characterized based on years of accumulated variants in a relatively small collection of genomes. The bulk of the SARS-CoV-2 sequences currently deposited in GISAID and GENBANK were collected and sequenced between the months of March and May of 2020 and show very high degree of similarity, which combined with the large number of available genomes, makes evolutionary analysis very challenging.

Another complication is that an apparent recombinant strain might instead be the result of a co-infection. Suppose a sample was collected from a patient co-infected with two distinct lineages of the virus, where one lineage contains two SNVs while the other lineage contains two distinct SNVs at the same positions. Such a sample would be amplified and sequenced in a single batch. Reads representing both alleles would be provided to an assembler to produce the final genome. Due to differences in coverage of each allele, it is possible that an assembly algorithm would produce a sequence with alleles from both clades, creating an artifactual recombinant. Re-analysis of raw sequencing reads by mapping them against the genome should reveal such artifacts, because reads from a patient co-infected by multiple clades would reveal both alleles at the corresponding sites.

To evaluate this hypothesis, we obtained sets of raw sequence reads deposited at NCBI/SRA or ENA for some of the recombinant genomes identified with Bolotie. Unfortunately, GISAID does not require authors to submit raw data, and only a limited number of submitters have placed their data in public archives with corresponding GISAID identifiers. Thus, we were only able to recover a limited number of datasets for our analysis.

Reads for the EPI_ISL_439137 isolate (Figure 3A) were obtained from the European Bioinformatics Institute's ENA database. As summarized in Table 1, all positions with high conditional affinities for a clade had a homogenous composition, indicating that the data did not derive from two distinct isolates, but instead it was likely a single isolate containing variants from two parental lineages.

Searching the list of recombinant genomes for similar signatures, we identified another isolate, EPI_ISL_489588 also from Scotland, dated one week earlier, which contained the same variants. Another isolate, EPI_ISL_510303 from Spain, had all but one variant (at position 28,143) matching the recombinant signature of the isolates from Scotland. Given the rapid mutation rate in RNA viruses, it is possible that an independent mutation occurred at that position, or that the reference allele is an assembly artifact, however we could not find raw data corresponding to the Spanish isolate and were unable to further investigate possible reasons for the missing variant.

Estimation of false positives

To provide a simple estimate of the false discovery rate of Bolotie, we generated a large set of simulated SARS-CoV-2 genomes that contained no recombination events. First, we computed a consensus genome for each of the 4 main clades shown in Figure 2. Then for each of the four genomes we generated 25,000 descendants by incorporating minor alleles at the frequencies found in the data, as well as random mutations at the rate of $6*10^{-4}$, as reported by others (van Dorp *et al.* 2020). We then ran Bolotie on these 100,000 genomes, and found only 4 false positive results. Further details can be found in the Supplementary Methods.

Performance

Complete analysis of the 87,695 genomes using Bolotie including alignment and index construction took a total of ~5.5 hours using 36 threads on two Intel Xeon E5-2680 v2 processors with 10 cores each. Using the conditional probability table provided with the software, analyzing a single additional genome takes on average ~30 seconds.

Discussion

The method and experiments presented in this work demonstrate that recombination has occurred between the four existing major clades of SARS-CoV-2. While some of the inferred events may be homoplasies or technical artifacts, our analysis shows that at least some of the genomes likely represent true cases of recombination.

Of the 225 recombination events identified in our analysis, the majority were represented by single isolates, suggesting that the event was not established in the population. However, because two-thirds of the available genomes were sequenced between late March and early May, it is possible that more data will reveal additional recombinant lineages.

The 225 inferred recombinant genomes comprise less than 1% of all sequences analyzed. It is possible that many more anomalous genomes could be detected by lowering the variant frequency threshold. For example, if we require 50 sequences to confirm a variant rather than 100, the number of informative sites increases more than two-fold from 411 to 996, possibly allowing detection of events that are rarer, such as those which involve smaller emerging lineages within the 4 clades. However, due to decreased specificity such an approach might require stricter manual inspection as the false positive rate is expected to increase substantially.

While overall all events detected by Bolotie passed manual verification, a possibility of mis-assembly in cases of co-infection by particles from different clades could also explain the presence of multiple clade-specific variants within a single genome. Although we were unable to obtain raw read data for all recombinants, our analysis of the EPI_ISL_439137 isolate (Figure 3A, Table 1) shows that at least in one case the recombinant origins of the genome can be validated. Several other recombinants identified in our analysis (EPI_ISL_468407, EPI_ISL_452334, EPI_ISL_475584, EPI_ISL_464547) have also been previously identified by other groups (VanInsberghe *et al.* 2020).

Due to the differences in library preparation, sequencing technology and assembly protocols, the need for raw data and independent validation is very high. We urge researchers to submit raw sequencing data so that any future studies can verify their findings, not only when studying recombination events, but also individual rare variants, transmission patterns, clade prevalence in different populations, etc.

Because our method relies heavily on the accuracy of variant calls, we sought to compare how well available phylogenetic trees agree with trees built using consensus sequences constructed from the alignment data we obtained. The trees shown in figures 4A and 4B are very similar, confirming that consensus sequences constructed by Bolotie preserve essential information sufficient for accurate phylogenetic analysis. In our analysis, Bolotie identified 15 out of 4,039 sequences used by NextStrain (Hadfield *et al.* 2018) as recombinants (Supplementary Table 2). Minor differences between the NextStrain tree and the tree computed from Bolotie consensus sequences are to be expected since consensus sequences have 200 bases replaced with the reference at both 3' and 5' ends and do not include any structural variants. Additionally, since NextStrain tree includes 455 isolates submitted to GISAID after we downloaded our latest set, those sequences are also expected to slightly alter the topology. Lastly, differences in software versions and randomized methods inherent in the tree-building software are expected to produce trees with minor differences on each iteration.

On the other hand, the introduction of 225 recombinant or otherwise anomalous genomes produced a mildly distorted tree with multiple outliers, shown in Figure 4C and Figure 2. In both illustrations most of the recombinant genomes were assigned a clade different from the non-recombinant neighbors by GISAID. Furthermore, several groups of potential recombinant genomes with identical or highly similar mutational signatures were identified by Bolotie (Figure 2). Such lineages are especially important for phylogenetic analysis, as they may affect the topology of the trees more significantly. Even minor perturbations to the topology of the tree in the presence of misclassified outliers may have adverse effects on the studies of dynamics and transmission of the pathogen lineages in the population (Awadalla 2003). This once again illustrates the importance of properly handling anomalous sequences in phylogenetic analysis. It must also be noted that different clade assignments for the SARS-CoV-2 genomes currently exist (Shu and McCauley 2017; Hadfield *et al.* 2018), at least in part due to differences in tree-building strategies (Supplementary Table 1). However, even though discrepancies in clade assignment may present a challenge to Bolotie, the results will still be of use for the refinement of phylogenetic trees and ultimately clade assignments of the genomes. Grouping together conflicting clades and smaller clades should increase the specificity of the results.

To the best of our knowledge, there currently exists one other recent evaluation of recombination in SARS-CoV-2 genomes (VanInsberghe *et al.* 2020). Several studies hinted at the possibility of recombination occurring, but the results were inconclusive due to very small sample size at the time (Yi 2020). Bolotie adds 221 candidate recombinants to the 5 proposed by VanInsberghe et al. Upon comparison of our results we noted that isolate EPI_ISL_464547 was reported by VanInsberghe et al. but was absent in the output of Bolotie. Of the 5 genomes reported in that study, EPI_ISL_464547 had the shortest length of the second segment. Upon closer investigation of conditional probabilities computed by Bolotie for that genome, we found that none of the variants are clade-defining, nor does the dominant clade 0 ever drop in probability below the baseline of 0.25 at which all genomes are equally likely. The only exception was a variant at the tail end of the sequence, which was equally likely for clades 0, 1 and 3 at 0.33 probability (Supplementary Figure 1). Thus, we concluded that the assignment done by Bolotie was likely correct for that genome.

However, it is possible that differences in partitioning of clades 0, 1 and 3 between our method and that used by VanInsberghe et al. would result in different conditional probabilities at that variant position. As shown in Supplementary Table 1, one such discrepancy does exist between clades 1 and 3, where NextStrain classification adds 334 sequences from clade GH to clade G. However, clade 0 would still have a conditional probability at that position greater than the baseline. As a result, we find it difficult to provide a conclusive assessment of the EPI_ISL_464547 isolate given currently available data.

Our work focused primarily on the development of a highly efficient, scalable general-purpose method for detecting recombination events in viral genomes irrespective of the divergence rates in the pool of collected isolates. An additional purpose was to search for convincing evidence that recombination does indeed happen in SARS-CoV-2. Hence, we imposed multiple conservative criteria, such as assessment of only 4 clades. However, it is important to note that recombination likely happens within lineages of the same clade. While not targeted in our analysis, future studies may choose to evaluate these intra-clade events, possibly yielding much higher numbers of recombination events.

Conclusion

Given how recently this novel strain of coronavirus appeared, much remains to be learned about SARS-CoV-2 and how it may change over time. Recombination events, which may have been responsible for the initial emergence of SARS-CoV-2 (Zhang *et al.* 2020), may have significant impact on future transmission and virulence of the virus. As such, our ability to detect recombination events in a timely manner is crucial in the ongoing efforts to find a solution to the pandemic and prevent additional casualties.

Utilizing an enormous collection of SARS-CoV-2 isolates sequenced by thousands of researchers around the globe (Shu and McCauley 2017), we were able to develop a method that can reliably detect sequences with anomalous mutation patterns which are indicative of recombination events. Using the proposed method, we identified 225 high likelihood recombinant sequences. Our findings suggest that recombination in SARS-CoV-2 is much more common than previously reported and that several recombinant lineages may have become established in the population.

We hope that the software presented here along with provided pre-built indices will help to detect future recombination events quickly and reliably, and aid in efforts to track the spread of the SARS-CoV-2 virus.

Data Availability Statement

The core method is implemented in C++ and based on the SeqAn (Reinert *et al.* 2017) and KSW2 (Li 2018; Suzuki and Kasahara 2018) libraries, the tree building is performed by IQ-TREE (Minh *et al.* 2020). The code and test data are available for download on GitHub: https://github.com/salzberg-lab/bolotie. The SARS-CoV-2 index built using the genomes in our analyses is also available for download at ftp://ftp.ccb.jhu.edu/pub/data/bolotie_sars_cov_2/.

A wrapper script is provided in the GitHub repository to run all steps of the protocol. This script, while convenient, is intended for replicability and testing and lacks some of the available features of Bolotie.

Supplemental Material available at figshare: https://doi.org/10.25386/genetics.14553696

Acknowledgements

We acknowledge the work of all the authors, originating and submitting laboratories who contributed nucleotide sequences to the GISAID EpiCov Database (87,695 isolates, 2 September 2020). We would also like to thank Dr. Martin Steinegger for the helpful discussions on algorithms for clustering large collections of viral genomes.

Funding

This work was supported in part by Fast Grants (part of Emergent Ventures at George Mason University) and by the US National Institutes of Health [grants R01-AI141009 and R35-GM130151].

Author Contributions

A.V. and C.P. developed and implemented the methods, ran experiments. A.V., C.P., S.L.S and M.P. conceptualized the study, methods and wrote the manuscript.

Competing Interests

The authors have no conflicts of interest to declare.

References

Awadalla P., 2003 The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. 4: 50–60.

Bruen T. C., H. Philippe, and D. Bryant, 2006 A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665–2681.

Demİr A. B., D. Benvenuto, Y. H. ABACIOĞLU, S. Angeletti, and M. Ciccozzi, 2020 Identification of the nucleotide substitutions in 62 SARS-CoV-2 sequences from Turkey. Turkish J. Biol. 44: 178–184.

Dong E., H. Du, and L. Gardner, 2020 An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20: 533–534.

Dorp L. van, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, *et al.*, 2020 No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat. Commun. 11: 1–8.

Drake J. W., and J. J. Holland, 1999 Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. 96: 13910–13913.

Forney G. D., 1973 The viterbi algorithm. Proc. IEEE 61: 268–278.

Hadfield J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, *et al.*, 2018 Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34: 4121–4123.

Kalyaanamoorthy S., B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin, 2017 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14: 587–589.

Kimura M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111–120.

Lam H. M., O. Ratmann, and M. F. Boni, 2018 Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol. Biol. Evol. 35: 247–251.

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357.

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Minh B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, *et al.*, 2020 IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37: 1530–1534.

Posada D., 2002 Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19: 708–717.

Reinert K., T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, *et al.*, 2017 The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. J. Biotechnol. 261: 157–168.

Shu Y., and J. McCauley, 2017 GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22: 30494.

Su S., G. Wong, W. Shi, J. Liu, A. C. K. Lai, *et al.*, 2016 Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24: 490–502.

Suzuki H., and M. Kasahara, 2018 Introducing difference recurrence relations for faster semi-global alignment of long sequences. BMC Bioinformatics 19: 33–47.

VanInsberghe D., A. S. Neish, A. C. Lowen, and K. Koelle, 2020 Identification of SARS-CoV-2 recombinant genomes. bioRxiv.

Yi H., 2020 2019 novel coronavirus is undergoing active recombination. Clin. Infect. Dis.

Zhang T., Q. Wu, and Z. Zhang, 2020 Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr. Biol.

Figure 1. *The maximum conditional probability for each nucleotide is highlighted in gray, while the path with the maximum likelihood is highlighted in bold. By penalizing switching of clades, insignificant differences in probabilities between clades as well as short windows representing a switch to a different clade are avoided. For clarity transitions between nodes on non-optimal paths are indicated in gray without labeled probabilities.*

Figure 2*. An unrooted topological cladogram of 4,249 SARS-CoV-2 genomes including 225 recombinants labeled as red bars. Arcs link each recombinant to both inferred parental genomes. The color of the arc corresponds to the color of the clade to which a recombinant was clustered within the tree. Clades correspond to the GISAID clades GR (****0****), GH (****1****), G (****2****) and all minor lineages combined (****3****).*

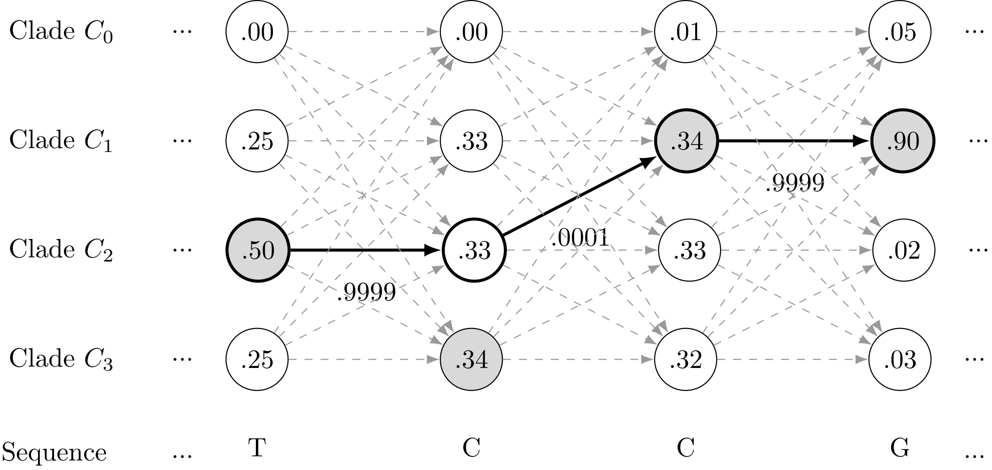
Figure 3. *Four examples of inferred recombinant sequences: A. EPI_ISL_439137; B. EPI_ISL_468407; C. EPI_ISL_509874; D. EPI_ISL_417420. The top section of each plot shows conditional probabilities of a clade given a nucleotide at each position. Bars are plotted for the two parent clades and the other clades are shown as dots of the corresponding color. Each peak >0.1 above the baseline (0.25) is labeled with the number of genomes it appears in. An average is reported whenever there are multiple variants in close proximity on the plot, listing the number of averaged variants in parentheses. The three lower panels of each plot show the*

frequency of variants at each position for parental clades (top and bottom rows) and variants observed on the recombinant genome (middle row).

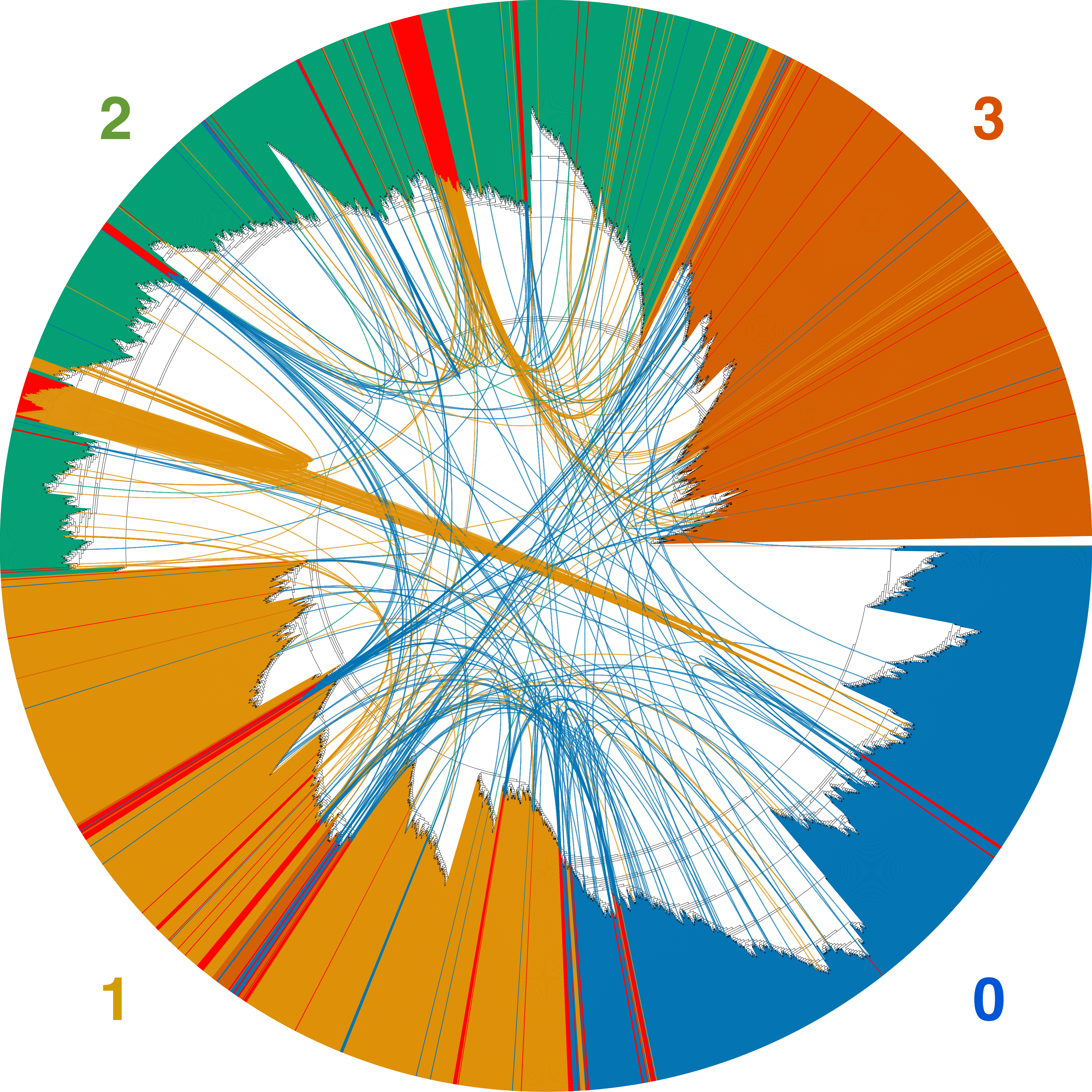
Figure 4. *Effects of sequence composition on the topology of the phylogenetic tress for SARS-CoV-2. A tree obtained directly from NextStrain (**A**) is first compared to (**B**) the tree computed using Bolotie consensus sequences for the same set of isolates. **(C)** Shows a tree computed for the same set of isolates with 210 additional recombinant sequences as identified by Bolotie. Leaf nodes that correspond to recombinant genomes are labeled with red dots.*

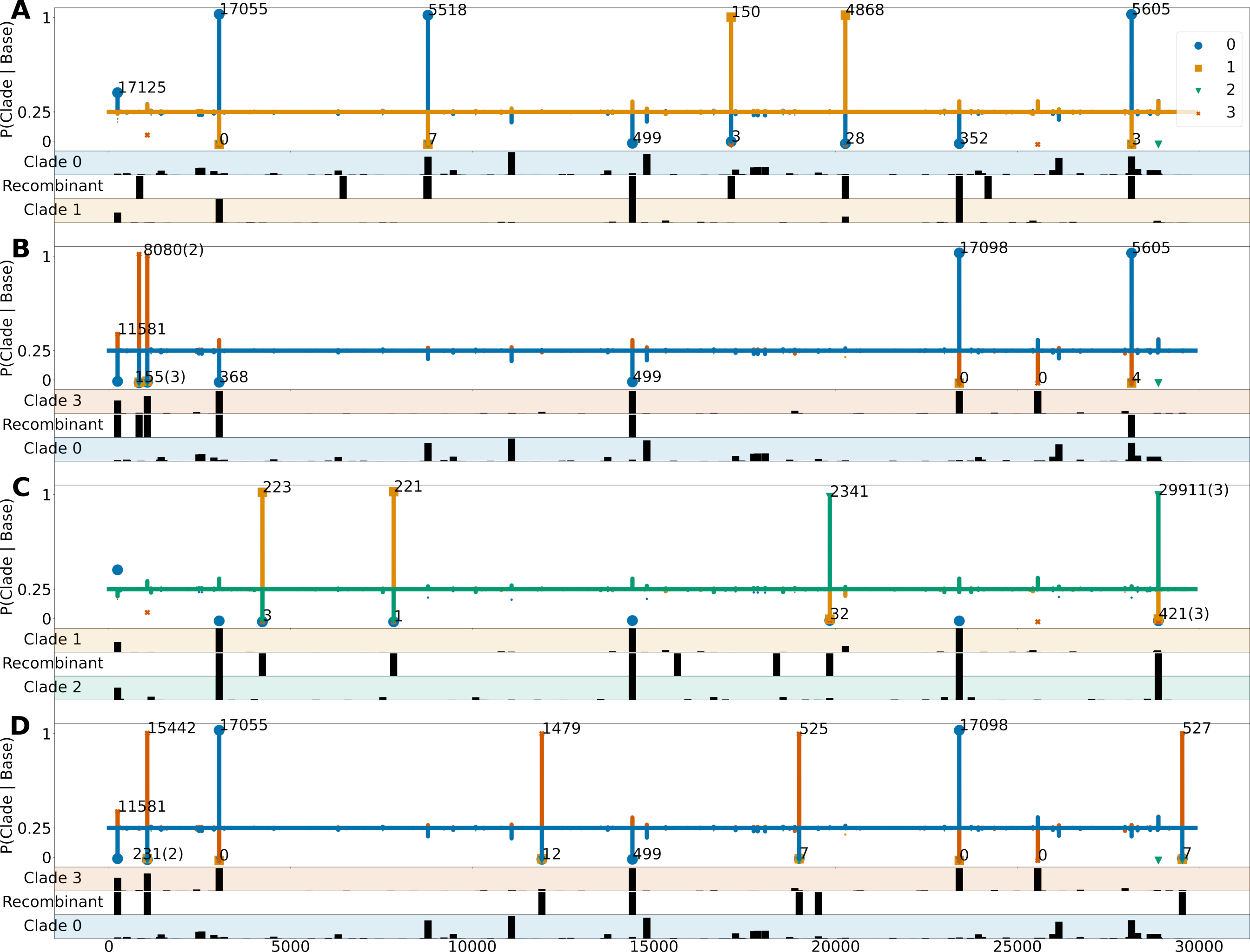
Clade	Position	Reference	Observed	P(0|Base)	P(1|Base)	A	C	G	T
0	240	C	C	0.3975	0.2358	4	6492	5	1
	3036	C	C	0.9998	0.0001	1	1422	1	9
	8781	C	T	0.9929	0.0012	0	0	0	140
1	14407	C	T	0.0092	0.3297	1	3	2	7359
	17125	T	C	0.0217	0.9783	1	7991	0	6
	20267	A	G	0.0061	0.9935	0	0	13	0
	23402	A	G	0.0066	0.3311	0	0	196	0
0	28143	T	C	0.9988	0.0005	2	8004	0	4

Table 1. *Mutational signature of the EPI_ISL_439137 recombinant isolate (Figure 3A). The table shows all positions with defining conditional probabilities for each of the parental clades. Read counts extracted from the data deposited in EBI are provided to illustrate the likely single-isolate origin of the genome.*

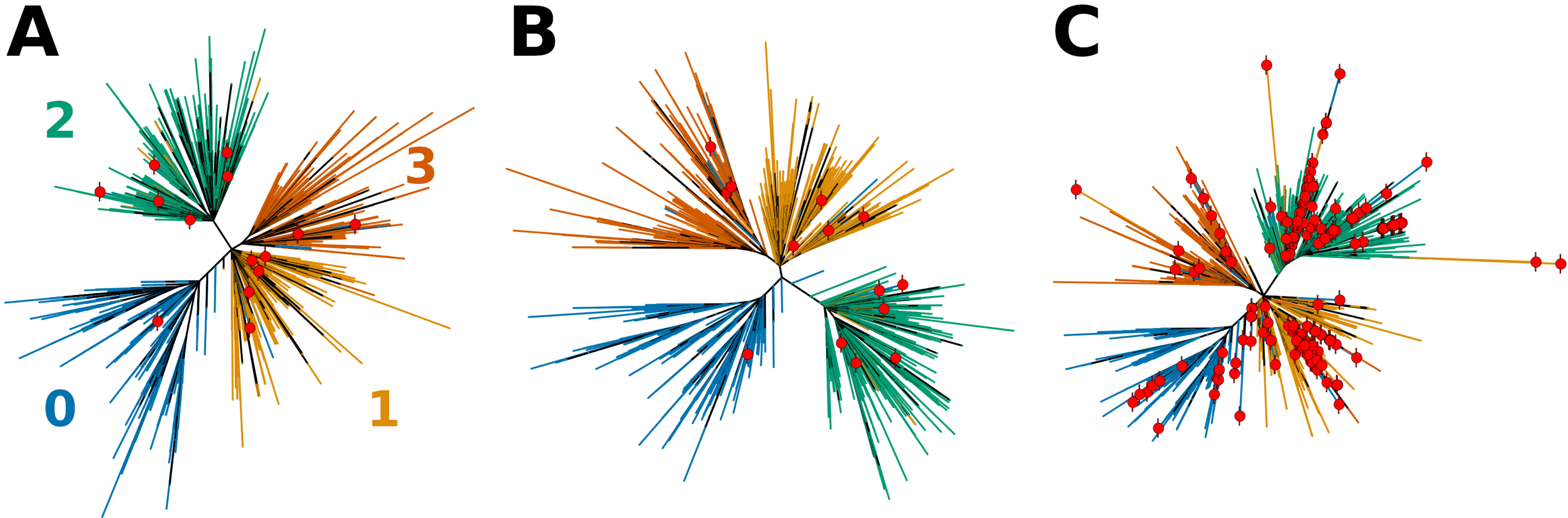


Clade C_0
Clade C_1
Clade C_2
Clade C_3
Sequence
...
.00
.25
.50
.25
.00
.33
.33
.34
.01
.34
.33
.32
.05
.90
.02
.03
.9999
.0001
.9999
T
C
C
G


2
3
1
0



A
B
C
D
P(Clade | Base)
Clade 0
Recombinant
Clade 1
Clade 3
Clade 2
17055
5518
150
4868
5605
17125
0
7
499
3
28
352
8080(2)
17098
11581
155(3)
368
4
223
221
2341
29911(3)
1
32
421(3)
15442
1479
525
527
231(2)
12
0
5000
10000
15000
20000
25000
30000
0.25
1


A
2
3
0
1
B
C