A Compatibility Approach to Identify Recombination Breakpoints in Bacterial and Viral Genomes

Yi-Pin Lai and Thomas R. Ioerger
Department of Computer Science and Engineering
Texas A & M University
College Station, Texas 77843
{yplai.tw,ioerger}@tamu.edu

ABSTRACT

Recombination is an evolutionary force that results in mosaic genomes for microorganisms. The evolutionary history of microorganisms cannot be properly inferred if recombination has occurred among a set of taxa. That is, polymorphic sites of a multiple sequence alignment cannot be described by a single phylogenetic tree. Thus, detecting the presence of recombination is crucial before phylogeny inference. The phylogenetic-based methods are commonly utilized to explore recombination, however, the compatibility-based methods are more computationally efficient since the phylogeny construction is not required. We propose a novel approach focusing on the pairwise compatibility of polymorphic sites of given regions to characterize potential breakpoints in recombinant bacterial and viral genomes. The performance of average compatibility ratio (ACR) approach is evaluated on simulated alignments of different scenarios comparing with two programs, GARD and RDP4. Three empirical datasets of varying genome sizes with varying levels of homoplasy are also utilized for testing. The results demonstrate that our approach is able to detect the presence of recombination and identify the recombinant breakpoints efficiently, which provides a better understanding of distinct phylogenies among mosaic sequences.

CCS CONCEPTS

•**Applied computing** →*Computational genomics; Bioinformatics;*

KEYWORDS

Phylogeny; Compatibility; Recombination; SNPs; Multiple Sequence Alignments; Bacterial Genomes; Viral Genomes

1 INTRODUCTION

The accurate inference of phylogeny for microorganisms is essential to better understand the evolutionary history of a set of species. A phylogenetic tree will be reconstructed based on polymorphisms of a multiple sequence alignment to present the evolutionary relationships of a set of isolates using methods such as maximum parsimony, maximum likelihood, or minimum evolution, etc. If there exists a single tree topology that can explain all polymorphic sites of an alignment, then the tree represents the evolutionary history of the given set of taxa. In recent decades, studies indicate that some microorganisms are relatively less mosaic because of shorter evolutionary history or complete clonality, for instance, *Mycobacterium tuberculosis*[15] and *Bacillus anthracis*[26]. However, growing evidences have shown that some microbiota exhibit homoplasy in their genomes, that is, they are not monophyletic. The appearance of homoplasy is that polymorphic sites are not from a common ancestor but arise independently in multiple branches[45]. Homoplasy could be caused by horizontal gene transfer (HGT) and homologous recombination. For bacteria, HGT is an evolutionary force that frequently occurs and results in mosaic genomes[17, 44]. The known mechanisms of HGT are transformation, transduction and conjugation. Transformation and transduction acquire a new DNA fragment from environment or other organisms via insertion while conjugation requires intercellular contact to transfer genetic variants from donor to recipient by homologous recombination. Homologous recombination exchanges genetic material between two homologous DNA sequences [4]. For viruses, main mechanisms for driving evolution are point mutation and recombination. Several bacteria and viruses have been reported that suggest recombination events occurred during evolution, including HIV-1[1, 32, 36], *Mycobacterium avium*[18], *Staphylococcus aureus*[13], *Streptococcus pneumoniae*[6] and *Salmonella enterica*[7].

A phylogenetic tree can be misleading if a sequence alignment is mosaic since some recombinant sites cannot be described by a single tree topology. If the microorganisms underwent recombination, it is essential to estimate the extent of homoplasy and identify recombination breakpoints to obtain corresponding phylogenies to model the evolution[27, 35, 40]. Current methods to detect recombination can be categorized into phylogenetic, distance, compatibility and nucleotide substitution distribution methods[33, 34]. Phylogenetic methods are commonly used in existing programs, for instance, Simplot, Plato, GARD, RDP4 and ClonalFrameML[8, 14, 23, 31, 36]. If recombinant regions exist, then the tree topologies that are constructed based on the regions will be discordant with the global topology. Among them, GARD is a phylogenetic-based method that applies a genetic algorithm to build a maximum likelihood model for recombination detection[30, 31]. RDP4 is also a phylogenetic-based method that makes use of a pairwise scanning approach to scan for the alignment[23]. In contrast, distance and compatibility methods do not require the phylogeny. For distance-based methods, pairwise distances will be computed, and distance patterns will be searched along a multiple sequence alignment. PHYPRO uses a sliding window to get the genetic statistics[47] while RAT

ACM-BCB'17, August 20-23, 2017, Boston, MA, USA

DOI: http://dx.doi.org/10.1145/3107411.3107432

utilizes thresholds of genetic distance and number of sequences to search for possible recombinant regions[12]. In addition, compatibility methods focus on site-to-site congruence and extend pairwise sites to all possible combinations of pairs within a given region. If two sites are compatible, then the relationship of taxa at both sites can be explained by the same tree topology[11, 20, 21, 24]. Nucleotide substitution distribution methods analyze the trends in substitutions or fitting degree in nucleotide distributions across the alignment[39, 43, 48].

With the advancement of next-generation sequencing techniques, the whole genomes of a number of strains can be analyzed at the same time so the size of an alignment may increase as the number of strains scales. Since phylogenetic-based methods require the phylogeny construction, they are less efficient than matrix-based or compatibility-based methods. Also, studies have shown that methods of substitution patterns and compatibility perform better than phylogenetic methods or other general methods in terms of accuracy and efficiency[5, 33, 34, 41]. Yet, only a few studies have applied the compatibility concept[3, 16]. Reticulate is a program to compute neighbor similarity score (NSS) using compatibility matrices to cluster sites that are compatible[16]. The clustering is obtained by shuffling the matrices with Monte Carlo randomization. Another recombination testing program defines a refined incompatibility score as a pairwise homoplasy index (PHI) to test for recombination[3]. The significance of PHI is computed by permutation tests. Both methods are able to detect recombination and report the informative sites, but unable to identify breakpoints to obtain non-overlapping regions.

To better understand different phylogenetic histories of evolution that involve recombination for a given set of taxa in a more computationally efficient way, we develop an average compatibility ratio approach to analyze the multiple sequence alignment of Single Nucleotide Polymorphisms (SNPs) to explore possible recombination breakpoints. The overall compatibility ratio will be calculated first to screen for the presence or absence of recombination. If it presents, then ACR approach will scan for each site along the alignment to identify the site where its upstream sites are incompatible with its downstream sites but the adjacent sites located within the same region are more jointly compatible. The potential breakpoints depending on a given threshold will be reported, and then the alignment will be divided into non-overlapping consecutive segments. The compatibility ratios and changes per site (CPS) information of each segment are also provided. The ACR approach is tested on simulated datasets of two scenarios and three empirical datasets, and the performance is compared with GARD[31] and RDP4[23]. Comparing the results of simulation scenarios with ground truths, we show that ACR approach performs better than other programs. The results of the analyses of bacterial and viral genomes demonstrate that our proposed approach is able to characterize biological sequence alignments and provide segments that reflect distinct phylogenetic histories.

2 METHODS

2.1 Compatibility Score

For a given multiple sequence alignment of n taxa ($n \geq 4$) and m sites ($m \geq 1$), a character χ is a set of states $(A_1, A_2, \ldots, A_n)$ of all strains at a given site of the alignment. The states are elements of $\{A, T, C, G\}$. Hence, for the above alignment, there are m characters and four types of states. Pairwise compatibility is defined as two characters are compatible if and only if there exists a phylogenetic tree that explains both characters. That is, no extra changes of state are required for both characters to evolve with a single tree topology. For a given set of characters, $C = (\chi_1, \chi_2, \ldots)$, all characters are jointly compatible if and only if all pairs of characters are compatible with each other, i.e., a single tree exists and fits for all of the characters[10, 11, 20, 21, 24].

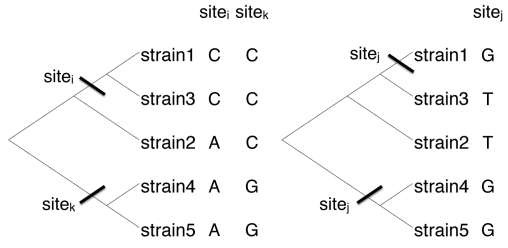
Since the single nucleotide polymorphisms in a multiple sequence alignment are informative for inferring phylogeny, the polymorphic sites are extracted. If the site has an unknown nucleotide, a gap or over two types of nucleotides, it is considered as ambiguous and excluded before determining compatibility. Therefore, each site will contain two types of nucleotides, and then the multiple sequence alignment will be converted to the binary sequence. That is, each character becomes a binary character whose states have two elements, 0 and 1.

Given a site in a binary sequence alignment of length n, the character χ_b is presented as a binary string concatenating the states of all n strains in the same order as the alignment, and the set χ'_b records the indexes of the strains with state 1 in χ_b. Two binary characters χ_1 and χ_2, which represent two sites, are compatible if and only if one of the following conditions is satisfied: $\chi'_1 \subseteq \chi'_2$, $\chi'_2 \subseteq \chi'_1$, $\overline{\chi}'_1 \subseteq \chi'_2$ or $\chi'_2 \subseteq \overline{\chi}'_1$. In other words, two binary characters of length n are compatible if and only if the set of pairs of binary states $A \in \chi_1$ and $B \in \chi_2$, $\{(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)\}$, has at most three types of binary pairs of states out of four combinations, $\{(0,0), (0,1), (1,0), (1,1)\}$. The compatibility theorem has been proved in [10, 11]. An example of SNP sequences of three characters in $site_i$, $site_j$ and $site_k$ is illustrated for the concept of compatibility. For $site_i$ and $site_j$, they are incompatible according to the definition, so there does not exist any tree that is able to accommodate characters at the $site_i$ and $site_j$. Characters at the $site_i$ and $site_k$ are compatible, so there exists a single tree that is compatible with both sites. As shown in Figure 1(a), the number of changes per site for $site_i$ or $site_k$ is 1, yet it requires at least 2 changes for $site_j$, that is, presence of homoplasy. Characters at the $site_j$ and $site_k$ are also compatible. In Figure 1(b), the tree that is able to describe both $site_j$ and $site_k$ cannot accommodate the character at the $site_i$. Figure 1 demonstrates that there does not exist a tree that is compatible with $site_i$ and $site_j$, therefore there does not exist a tree that is compatible with these 3 sites.

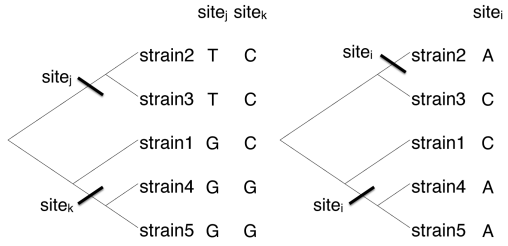
Example for demonstrating compatibility:

$$\begin{array}{c} \\ strain_1 \\ strain_2 \\ strain_3 \\ strain_4 \\ strain_5 \end{array} \begin{array}{c} \begin{array}{ccc} site_i & site_j & site_k \end{array} \\ \begin{pmatrix} C & G & C \\ A & T & C \\ C & T & C \\ A & G & G \\ A & G & G \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ strain_1 \\ strain_2 \\ strain_3 \\ strain_4 \\ strain_5 \end{array} \begin{array}{c} \begin{array}{ccc} site_i & site_j & site_k \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \end{array}$$

2.1.1 Pairwise Compatibility Score of Characters: Suppose that the number of total polymorphic sites in a sequence alignment of n taxa is m, a m x m matrix, $CompatPW$, is generated to record



(a) A tree based on SNPs at the $site_i$ and $site_k$



(b) A tree based on SNPs at the $site_j$ and $site_k$

Figure 1: Phylogenetic trees of the example

compatibility information for each pair of sites. That is, the compatibility information of pairwise characters at the $site_i$ and $site_j$, χ_i and χ_j, is recorded in $CompatPW_{ij}$, where $1 \le i, j \le m$. If two characters are compatible, the score is 1; otherwise, 0.

$$CompatPW_{ij} = \begin{cases} 1, & \text{if characters of } \chi_i \text{ and } \chi_j \text{ are compatible} \\ 0, & \text{otherwise} \end{cases}$$

$$CompatPW = \begin{bmatrix} CompatPW_{11} & CompatPW_{12} & \dots & CompatPW_{1m} \\ CompatPW_{21} & CompatPW_{22} & \dots & CompatPW_{2m} \\ \vdots & \vdots & & \vdots \\ CompatPW_{m1} & CompatPW_{m2} & \dots & CompatPW_{mm} \end{bmatrix}$$

2.1.2 Summary of Compatibility Scores of a Given Region: The summary of compatibility of a given region starting from $site_p$ to $site_q$, $CompatSum_{pq}$, where $1 \le p, q \le m$, is the summation of compatibility scores of all pairs within the region. For an alignment of length m, given a length l ($l \le m$), a m x l matrix $CompatSumMatrix$ represents the summary of compatibility of all pairs within a given region of at most length l starting from each site to its downstream sites. The $CompatSum_{pq}$ could be calculated recursively by summing $CompatSum_{p(q-1)}$ and the compatibility information of additional combinations of pairs that are contributed by adding a downstream site q.

Since the SNP sequence alignment of microorganisms may be cyclic or linear, two conditions are considered during computation. If the genome sequence is cyclic, the next downstream site of the ending site in the alignment will be the first site. For each site, the summary of compatibility will be calculated recursively starting from its next downstream site to the given length l. If the genome sequence is linear, the summary of compatibility will be computed recursively until q reaches to m or the summation of p and l, whichever occurs first.

$$CompatSum_{pq} = \begin{cases} CompatPW_{pq}, \text{if } q = p+1 \\ CompatSum_{p(q-1)} + \sum_{i=p}^{q-1} CompatPW_{iq}, \text{if } 1 \le p < q \le m \\ CompatSum_{p(q-1)} + \sum_{i=p}^{m} CompatPW_{iq} + \sum_{i=1}^{q-1} CompatPW_{iq}, \\ \text{if } 1 \le q < p \le m \text{ (cyclic)} \end{cases}$$

$$CompatSumMatrix(cyclic) = \begin{bmatrix} CompatSum_{12} & CompatSum_{13} & \dots & CompatSum_{1(1+l)} \\ CompatSum_{23} & CompatSum_{24} & \dots & CompatSum_{2(2+l)} \\ \vdots & \vdots & & \vdots \\ CompatSum_{m1} & CompatSum_{m2} & \dots & CompatSum_{ml} \end{bmatrix}$$

2.1.3 Ratio of Compatibility Scores of a Given Region: The ratio of compatibility score of a given region, $CompatRR_{pq}$, is the summation of compatibility scores of all pairs of sites to the number of all combinations of pairwise sites starting from $site_p$ to $site_q$. The $CompatRR_{pq}$ will be bounded between 0 and 1, which indicates the tendency of compatibility within the regions between $site_p$ and $site_q$. In other words, the higher ratio reflects that more characters are jointly compatible, suggesting more sites are congruent in a tree within the region. In contrast, the lower the ratio is, the less likely the recombination events happened in the region.

$$CompatRR_{pq} = \frac{CompatSum_{pq}}{\binom{r}{2}},$$

$$\begin{cases} r = q - p + 1, & \text{if } 1 \le p < q \le m \\ r = m - p + q + 1, & \text{if } 1 \le q < p \le m \text{ (cyclic)} \end{cases}$$

2.2 Characterization of Recombination Breakpoints

A sliding window size, w, is defined in terms of SNPs, regardless of genomic span (bp) or overall rate of mutations because non-polymorphic sites are non-informative and discarded. That is, the size of a sliding window is the constant number of polymorphic sites instead of nucleotides. It would get better estimates of incompatibility, especially in the regions with sparse SNPs. Given a size w, for each $site_i$, we average all of the compatibility ratios of regions that include $site_i$ and locate within $[i - w, i + w]$, i.e., $avgCompatRR_i$. The total number of combinations of regions that satisfy the criteria is $|w|^2$. If the genome sequence is linear, the sites located in the head and tail regions of size w will be ignored. The sliding window size can be adjusted according to the length of an alignment.

A site with a local minimum represents that its upstream region and downstream region within the length of $|w|$ are less jointly compatible comparing to other regions. So, sites with local minimums are the potential breakpoints for inferring phylogeny incongruence. The default value of the cutoff is the mean minus two standard deviations of all ratios, which is of the one-tailed order of a two-sigma

effect. The value can be adjusted for rare recombinant or frequently recombinant sequences. Or top k non-overlapping segments of the entire alignment can be reported by giving an intended number of k. Then, the top k-1 sites that have local minimums will be returned as breakpoints. In addition, the minimum length of a segment can be set to reduce the false positive rate since smaller regions may result in higher compatibility ratio. The consecutive sites between every two adjacent breakpoints construct a non-overlapping segment that is expected to be more congruent with a topology. Thus, a list of top potentially recombinant regions will be obtained.

$$avgCompatRR_i = \begin{cases} \dfrac{\sum_{p=i-w}^{i-1}\sum_{q=i+1}^{w} CompatRR_{pq}}{|w|^2}, \text{if } w+1 \le i \le m-w \\ \dfrac{(\sum_{p=m-(w+1-i)}^{m} + \sum_{p=1}^{i-1})\sum_{q=i+1}^{w} CompatRR_{pq}}{|w|^2}, \\ \text{if } i < w+1 \text{ (cyclic)} \\ \dfrac{\sum_{p=i-w}^{i-1}(\sum_{q=i+1}^{m} + \sum_{q=1}^{w-(m-i)+1})CompatRR_{pq}}{|w|^2}, \\ \text{if } i > m-w \text{ (cyclic)} \end{cases}$$

2.3 Build the Phylogenetic Trees and Calculate the Homoplasy Ratios for Segments:

Once the recombination segments are identified, the homoplasy ratio will be computed to evaluate the congruence/correctness, and the phylogenetic tree will be plotted for each segment. The polymorphic sites of segments will be extracted and used to build a phylogenetic tree for all strains. The ratio of the required SNPs for constructing a phylogenetic tree to the total number of selected SNPs of a segment will be calculated and defined as homoplasy ratio, that is, changes per site. The PHYLIP package of version 3.695[29] is used to infer the phylogeny of a dataset. The phylogenetic tree is reconstructed using the maximum parsimony method as the default setting. The homoplasy ratio can also be calculated using Sankoff's algorithm[37]. Sankoff's algorithm is a dynamic programming algorithm for counting the minimum cost of a tree using a cost matrix and backtracing approach. Here, the Sankoff score stands for the minimum number of state changes that are required in a given tree topology. The score of 1 represents that the character of a given site is congruent with the tree while the score above 1 stands for the extra number of changes needed for a given tree. The higher the Sankoff score, the higher the extent of incongruence of a given site, i.e., multiple changes are required to explain the tree pattern. Hence, the homoplasy ratio using Sankoff's algorithm is the summation of the Sankoff scores to the number of sites of a given region.

2.4 Computational Complexity

The Algorithms of compatibility determination, average ratio calculation and breakpoints characterization are shown in Algorithm 1 to Algorithm 3.

2.4.1 Time Complexity. Since it has to loop over all SNP sites and all isolates, the time complexity of Algorithm 2 is $O(ml * max(n, l))$, where m is the length of SNPs, n is the number of strains and l is the number of downstream sites. For Algorithm 3, the time complexity is $O(mw)$, where w is the sliding window size.

2.4.2 Space Complexity. The matrix size of Algorithm 2 is $O(ml)$, while the space of Algorithm 3 is $O(m)$.

Algorithm 1 Determination of pairwise compatibility of characters

Input: Two binary characters χ_i and χ_j at the $site_i$ and $site_j$
Let A and B be the states of characters χ_i and χ_j, respectively
$A\_subset\_B \leftarrow True$, $B\_subset\_A \leftarrow True$, $notA\_subset\_B \leftarrow True$, $B\_subset\_notA \leftarrow True$
for $i = 1 \rightarrow n$ **do**
  if $A[i] = 1$ && $B[i] = 0$ **then**
    $A\_subset\_B \leftarrow False$
  end if
  if $A[i] = 0$ && $B[i] = 1$ **then**
    $B\_subset\_A \leftarrow False$
  end if
  if $A[i] = 0$ && $B[i] = 0$ **then**
    $notA\_subset\_B \leftarrow False$
  end if
  if $A[i] = 1$ && $B[i] = 1$ **then**
    $B\_subset\_notA \leftarrow False$
  end if
end for
if A_subset_B || B_subset_A || $notA\_subset\_B$ || B_subset_notA **then**
  return $True$
else
  return $False$
end if
$True$ means χ_i and χ_j are compatible while $False$ means they are incompatible

Algorithm 2 Efficient computation of compatibility score and ratio for each site to its downstream sites of length l

Input: A binary n x m matrix of an alignment of SNPs with m sites and n strains
Require: A given length of downstream sites, l
if $CompatAll \ge$ a given threshold (99.5% as default) **then**
  Print "No recombination event detected in the alignment."
  return
end if
$CompatPW[m, l] \leftarrow 0$
for $i = 1 \rightarrow m$ **do**
  for $j = i + 1 \rightarrow l$ **do**
    $CompatPW[i, j] \leftarrow$ Algorithm 1 (Characters at the $site_i$ and $site_j$)
  end for
end for
$CompatSumMatrix[m, l] \leftarrow 0$
for $p = 1 \rightarrow m$ **do**
  for $q = p + 1 \rightarrow p + l$ **do**
    if $q = p + 1$ **then**
      $CompatSumMatrix[p, 1] \leftarrow CompatPW[p, q]$
    else if $q \le m$ **then**
      $CompatSumMatrix[p, q-p] \leftarrow CompatSumMatrix[p, q-p-1] + \sum_{i=p}^{q-1} CompatPW[i, q]$
    else
      $t \leftarrow q - m$
      $CompatSumMatrix[p, q-p] \leftarrow CompatSumMatrix[p, q-p-1] + \sum_{i=p}^{m} CompatPW[i, t] + \sum_{i=1}^{t-1} CompatPW[i, t]$
    end if
  end for
end for
$CompatRR[m, l] \leftarrow 0$
for $r = 1 \rightarrow l$ **do**
  $NumCombi[1 : m, r] \leftarrow \binom{r+1}{2}$
end for
$CompatRR \leftarrow \frac{CompatSumMatrix}{NumCombi}$
return $CompatRR$

Algorithm 3 Identification of recombination breakpoints

Input: The m x l $CompatRR$ matrix
Require: A given sliding window size of w, targeting k breakpoints or cutoff for local minimums
$avgCompatRR[m, 1] \leftarrow 0$
for $i = 1 \rightarrow m$ **do**
  if $i \geq w + 1$ && $i \leq m - w$ **then**
    $avgCompatRR[i, 1] \leftarrow \frac{\sum_{p=i-w}^{i-1} \sum_{q=i+1}^{w} CompatRR_{pq}}{|w|^2}$
  else if $i < w + 1$ **then**
    $avgCompatRR[i, 1] \leftarrow \frac{(\sum_{p=m-(w+1-i)}^{m} + \sum_{p=1}^{i-1}) \sum_{q=i+1}^{w} CompatRR_{pq}}{|w|^2}$
  else
    $avgCompatRR[i, 1] \leftarrow \frac{\sum_{p=i-w}^{i-1} (\sum_{q=i+1}^{m} + \sum_{q=1}^{w-(m-i)+1}) CompatRR_{pq}}{|w|^2}$
  end if
end for
return positions of local minimums that are lower than the given cutoff or top k local minimums as breakpoints

3 EVALUATION DATASETS

Four simulated datasets and three empirical datasets of bacterial and viral genomes are used to evaluate the ACR approach. The sequences of all strains are aligned, the sites that have low coverages or gaps are excluded, and then the sites that have polymorphisms are extracted. The input format is the same as fasta files.

3.1 Generation of Simulated Datasets

To generate datasets of recombinant sequences, our simulation algorithm consists of the generation of DNA sequences evolving along a tree topology and the construction of shuffled tree topology. For a given phylogeny, it generates (polymorphic) nucleotide sequences by generating a random ancestral sequence (with equal nucleotide prior probabilities), and then sampling changes on random branches (one per site) with probability proportional to branch length, assuming equal nucleotide transition probabilities. Also, the input tree topology can be modified by swapping any two branches. That is, the algorithm will randomly pick two internal nodes, cut off one node with all its child nodes as a branch, and attach it to another node to get a shuffled tree. Given a non-recombinant phylogenetic topology and a targeting sequence length, the sequences of strains will be simulated till reaching the given length based on the given topology and the shuffled topology, respectively. Next, recombinant sequences will be obtained by concatenating the generated sequences to get the simulated alignment. For each alignment, the positions of region boundaries will be recorded, the homoplasy score for each site will be calculated using the Sankoff's algorithm, and the scores will be averaged to get the extent of genetic diversity.

3.2 Empirical Datasets

The first empirical dataset consists of nine *Staphylococcus aureus (S. aureus)* clinical isolates[13] and aligned with the reference strain of ST8:USA300 (methicillin-resistant *S. aureus*). Mosaic structure has already been reported for these strains. The alignment has 48417 SNP sites. The second one is Human Immunodeficiency virus type 1 (HIV-1) recombinant dataset. Recombination is well-known among HIV-1 genomes. It contains an alignment of gene *pol* (DNA polymerase) of 12 strains and 1593 sites[42]. The third one is composed of 50 strains of *Mycobacterium tuberculosis (M. tuberculosis)* with 10565 SNP sites[46] that is highly congruent and have shown basically no recombination events in previous studies[2, 15].

4 RESULTS

4.1 Evaluations using Simulated Datasets

To evaluate the performance of our compatibility approach, a non-homoplastic phylogenetic tree topology of 73 taxa is utilized from a study of clinical isolates of *M. tuberculosis* in Panama [19] to generate SNP sequence alignments that are assumed cyclic. Two scenarios are set, including different levels of genetic diversity of one recombinant region and datasets of more than one recombinant regions. The first simulated scenario has two datasets of different levels of homoplasy containing 3000 SNP sites where the region[1000, 2000] is reorganized. The alignments of relatively low and high levels of homoplasy and the plot of six kinds of sliding window sizes for averaging the ratios for each site are shown in Figure 2 and Figure 3, respectively. The multiple curves show how sensitivity for detecting breakpoints depends on window size used.

The second simulated scenario has a dataset of two recombined regions with the same reorganizing topology and a dataset of two recombined regions with distinct topologies. Both datasets contain 6000 SNP sites where region[1000, 2000] and region[4000, 4500] are reorganized. Both of them are reconstructed by shuffling some internal branches, that is, these two regions have recombination and may not be described by the original tree topology. The first dataset has two distinct topologies for two recombinant regions, and the second dataset has the same topology for two recombinant regions. The plot of the Sankoff score for each site and the plot of six kinds of sliding window sizes for averaging the ratios for each site are shown in Figure 4 for dataset I and in Figure 5 for dataset II.

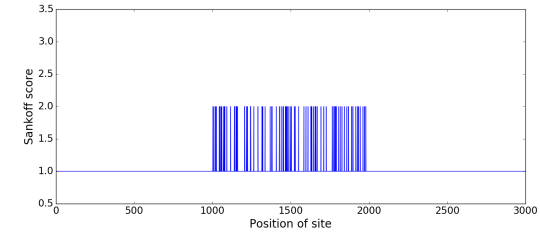
The ground truth and the results of using GARD, RDP4 and our approach for four simulated datasets of two scenarios are listed in Table 1. In scenario I, the identified breakpoints are around 1000 and 2000 with differences of at most 25 nucleotides for two datasets for three approaches, showing the capability of recombination detection. In scenario II, GARD only reported one breakpoint around 2000 for both datasets yet the expecting results should be four points that are close to 1000, 2000, 4000 and 4500, suggesting that GARD is unable to detect the recombinant regions for this kind of simulations. The breakpoints characterized by our approach are closer to the ground truth than those by RDP4.

4.2 Evaluations using Empirical Datasets

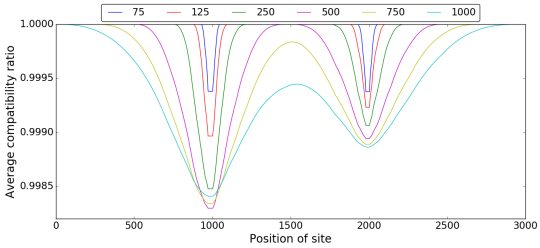
4.2.1 A Case Study on Recombinant Dataset: Staphylococcus aureus. The phylogenetic tree that is constructed using all SNP sites is shown in Figure 6. It requires a total number of 54157 SNPs to build the tree yet has 48417 SNPs, so its number of changes per site is 1.119 (54157/48417). The overall compatibility ratio is 0.978. The plot of average compatibility ratio of three window sizes is shown in Figure 7. Figure 8 illustrates that 10 breakpoints are identified using default setting and the window size of 500 for the overall alignment. Since the GARD is unable to analyze sequences of length above 12k, the alignments between 10k and 20k are taken out to evaluate three programs in the same base. Top four breakpoints are identified by our ACR approach using the window size of 500 and default cutoff. Five segments are obtained, and

Table 1: Breakpoints of four simulated datasets identified by our approach, GARD and RDP4

Breakpoints	Ground truth	ACR	GARD	RDP4
ScenarioI – DatasetI	1000, 2000	989, 2000	1025, 1999	984, 2010
ScenarioI – DatasetII	1000, 2000	998, 1998	1003, 1996	989, 1266, 2037
ScenarioII – DatasetI	1000, 2000, 4000, 4500	993, 2006, 4005, 4496	2001	978, 2000, 4005, 4515
ScenarioII – DatasetII	1000, 2000, 4000, 4500	998, 1995, 4000, 4498	2004	980, 2008, 4014, 4533

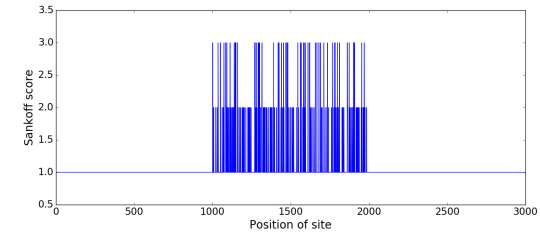


(a) Sankoff score for each site

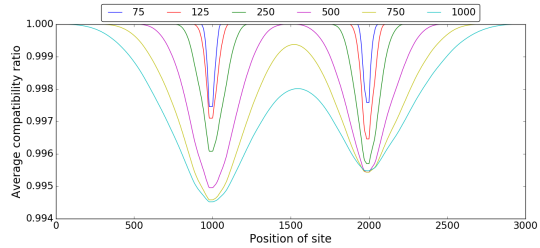


(b) Average compatibility ratio for each site using window sizes of 75, 125, 250, 500, 750 and 1000

Figure 2: Scenario I–Dataset I: Relatively low level of homoplasy

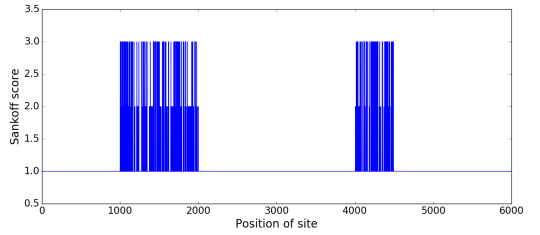


(a) Sankoff score for each site

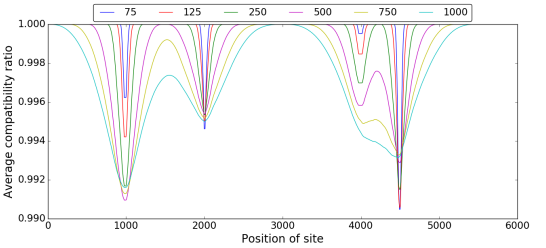


(b) Average compatibility ratio for each site using window sizes of 75, 125, 250, 500, 750 and 1000

Figure 3: Scenario I–Dataset II: Relatively high level of homoplasy

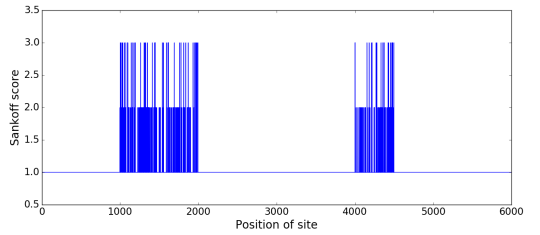


(a) Sankoff score for each site

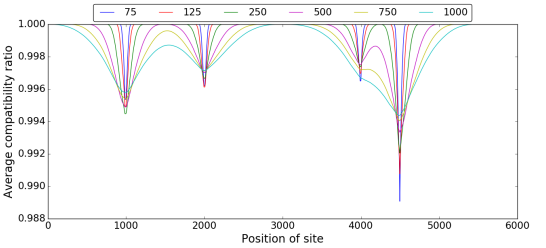


(b) Average compatibility ratio for each site using window sizes of 75, 125, 250, 500, 750 and 1000

Figure 4: Scenario II–Dataset I: Combined regions with distinct tree topologies



(a) Sankoff score for each site



(b) Average compatibility ratio for each site using window sizes of 75, 125, 250, 500, 750 and 1000

Figure 5: Scenario II–Dataset II: Combined regions with the same tree topology

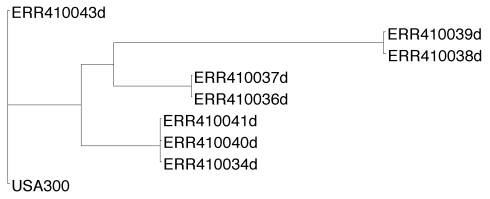


Figure 6: Phylogenetic tree of 9 strains based on all 48417 SNPs for *S. aureus*

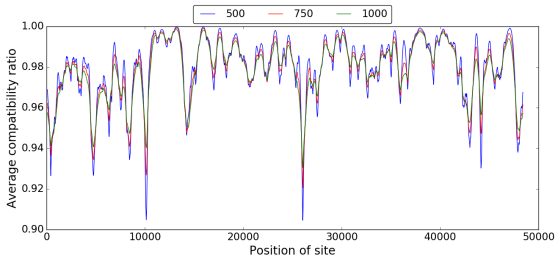


Figure 7: Average compatibility ratio for each site using window sizes of 500, 750 and 1000 for *S. aureus*

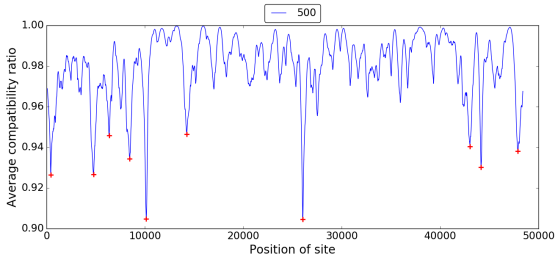
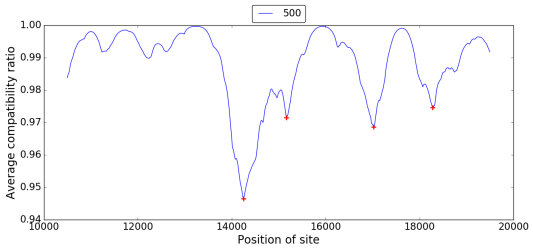


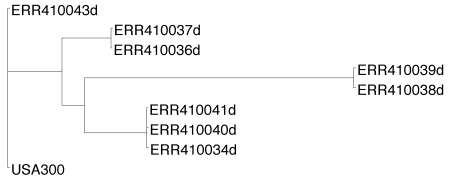
Figure 8: Identified breakpoints using window sizes of 500 for *S. aureus*

their local phylogenetic trees are shown in Figure 9. Comparing to the regional tree, phylogenetic trees of the second, third and fifth segments show that the branch of two strains, ERR410037 and ERR410036, receives the copies from parents of the branch of three strains, ERR410041, ERR410040 and ERR410034. The tree topology of the regional tree is different from the global tree. Additionally, the information of compatibility ratio (CompatRR) and changes per site for 5 segments, local region between 10k and 20k and the entire region are listed in Table 2. The analysis of the alignment using GARD with default setting was stopped before convergence while RDP4 reveals evidence of recombination in this alignment using PHI-test (p-value $< 10^{-5}$). The breakpoint distribution plot using RDP4 is shown in Figure 10. The top six breakpoints are 10147, 14410, 14975, 16990, 17285 and 19485. Besides the first and last sites, the rest are close to the sites identified by ACR, showing similar recombination detection results from ACR and RDP4.

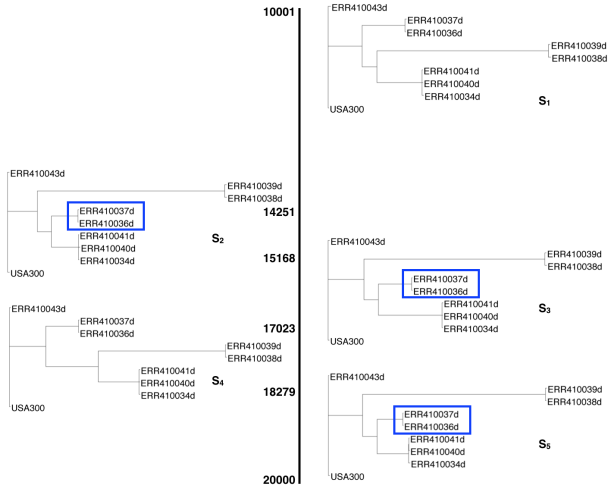
4.2.2 A Case Study on Recombinant Dataset: HIV-1 pol. The HIV-1 *pol* (DNA polymerase) dataset contains 12 strains with 1593 sites and 72 polymorphic sites. The multiple-sequence alignment is



(a) Top four breakpoints of the region[10k, 20k] using window sizes of 500 nucleotides



(b) Phylogenetic tree of the region[10k, 20k]



(c) Regional phylogenetic trees of five segments

Figure 9: Recombination segments within the region[10k, 20k] identified by ACR for *S. aureus*

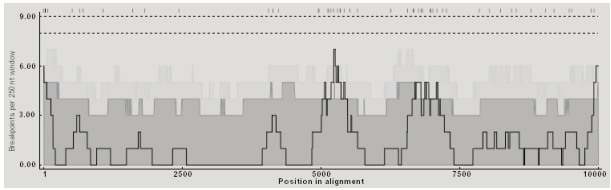


Figure 10: Breakpoints characterized by RDP4 for *S. aureus*

acyclic for the HIV-1 virus. Figure 11 describes the phylogenetic tree of using all polymorphic sites. The number of changes per site is 1.375 (99/72), and the overall compatibility ratio is 0.927. The plot of average compatibility ratio of three window sizes is shown in Figure 12. Two breakpoints, 27 and 61, are identified using default cutoff (mean-2*std) and window size 7. The results

Table 2: Compatibility ratio and changes per site of three segments within the region[10k, 20k] using ACR for *S. aureus*

	Segment	Size	Genes	CompatRR	Required sites	CPS
S_{all}	[1, 48417]	48417	*USA300HOU_0001-USA300HOU_2716*	0.978	105217	1.119
S_{local}	[10001, 20000]	10000	*USA300HOU_0555-USA300HOU_1155*	0.986	10945	1.095
S_1	[10001, 14250]	4250	*USA300HOU_0555-USA300HOU_0810*	0.990	4591	1.080
S_2	[14251, 15167]	917	*USA300HOU_0810-USA300HOU_0903*	0.953	1022	1.115
S_3	[15168, 17022]	1855	*USA300HOU_0903-USA300HOU_0983*	0.992	1987	1.071
S_4	[17023, 18278]	1256	*USA300HOU_0984-USA300HOU_1042*	0.989	1350	1.075
S_5	[18279, 20000]	1722	*USA300HOU_1043-USA300HOU_1155*	0.987	1878	1.091

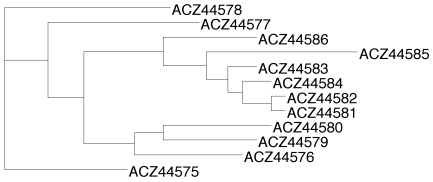


Figure 11: Phylogenetic tree of 12 strains based on all 72 SNPs for HIV-1 *pol*

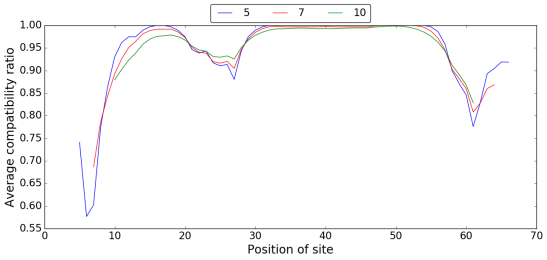
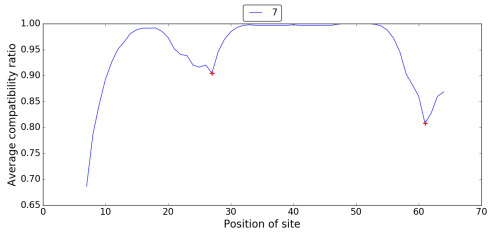


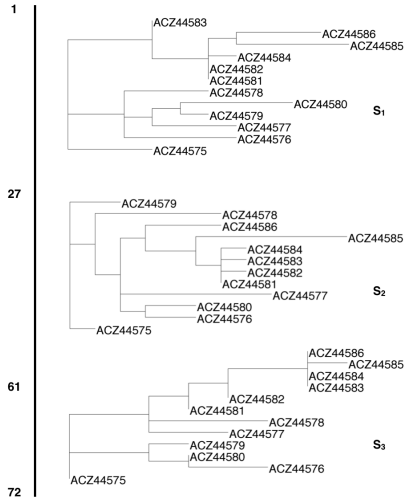
Figure 12: Average compatibility ratio for each site using window sizes of 5, 7 and 10 for HIV-1 *pol*

of three segments are shown in Figure 13 and listed in Table 3. The first segment has the lower number of changes per site and different evolutionary phylogeny. The number of changes per site of the second segment drops down to 1.088, indicating that the sites within this segment are more jointly compatible. The coordinates of 11 sites located in the third segments are from 1325 to 1586 consecutively in the *pol* gene. The polymorphic sites within this region are highly incompatible with each other and incongruent with the global tree. However, both GARD and RDP4 found no evidence of recombination events in this alignment.

4.2.3 A Case Study on Non-recombinant Dataset: Mycobacterium tuberculosis. The dataset contains 50 strains with 10565 SNPs sites. The number of changes per site is 1.006 (10633/10565). The overall compatibility ratio is 0.999, reflecting the clonal nature of *M. tuberculosis* strains worldwide. Hence, we should expect to find no recombination. The plot of average compatibility ratio of three window sizes is shown in Figure 14. Since the average compatibility ratio of the entire alignment is over 99.5%, our approach will report no combination breakpoints. In addition, both GARD and RDP4 reported that no evidence of recombination event was found in the alignment.



(a) Top two breakpoints using window sizes of 7 nucleotides



(b) Regional phylogenetic trees of four segments

Figure 13: Recombination segments identified by ACR for HIV-1 *pol*

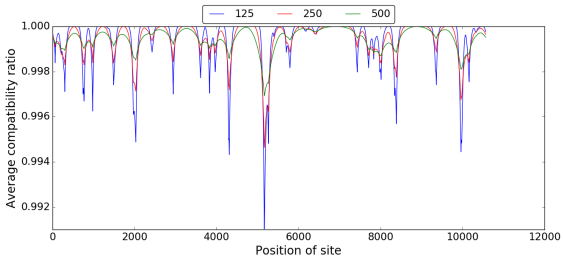


Figure 14: Average compatibility ratio for each site using window sizes of 125, 250 and 500 for *M. tuberculosis*

Table 3: Compatibility ratio and changes per site for window size of 7 using ACR for HIV-1 *pol*

	Segment	Size	CompatRR	Required sites	CPS
S_{all}	[1, 72]	72	0.927	99	1.375
S_1	[1, 27]	27	0.912	36	1.333
S_2	[28, 61]	34	0.989	37	1.088
S_3	[62, 72]	11	0.727	18	1.636

5 DISCUSSION

In the simulations, scenario I provides two levels of sequence divergences within a region in the middle of alignments. GARD, RDP4 and our approaches are all able to detect the presence of recombination. The breakpoints that are identified by our approach are closer to the ground truth than the other two programs. In scenario II, two recombinant regions with the same or distinct phylogenies are simulated. GARD only reports one breakpoint. RDP4 is able to report four breakpoints but deviates more from the ground truth than our method. Default settings of GARD and RDP4 are used for analyzing the alignments, respectively. GARD has three parameters, including nucleotide substitution bias mode, site-to-site rate variation, rate classes. The performance of GARD may be improved by adjusting different settings for parameters. RDP4 contains its own automated RDP program[22] as well as other programs. Automated RDP program and breakpoint distribution plot are applied. No parameter is required for automated RDP program while the windows size can be adjusted for screening for the alignment using breakpoint distribution plot.

Since recombination events are more complex in biological organisms and the genomes of virus are usually smaller than genomes of bacteria, recombinant empirical datasets of bacteria and virus are included for evaluation. In the virus dataset of shorter sequences, GARD and RDP4 report no recombination in the alignment of 72 polymorphic sites. In the bacteria dataset of larger size of sequences, the results of RDP4 and our approach are similar. GARD has the restriction of sequence length of 12k nucleotides. The analysis of the alignment using GARD was stopped before convergence, because the CPU time limit per job was reached.

Evidence of homologous recombination has been reported in the pathogen of *S. aureus*. Examining the polymorphic sequences of a set of isolates throughout the genomes provides an opportunity to profile the genetic exchanges that may be driven by selective forces such as antibiotic usage or drug resistance[9, 28]. The existence of several recombination hotspots has been discovered in *S. aureus*, including transposons, plasmids, phages, the staphylococcal cassette chromosome (SCC), pathogenicity genomic islands and genomic islets [13, 38]. In our analysis, the methicillin-resistance gene *mecA* (*USA300HOU_0956*) is located in the third segment in Figure 9. The *mecA* gene belongs to staphylococcal cassette chromosome *mec* (SCC*mec*) elements that are genomic islands. The third segment has distinct phylogeny, indicating genetic changes occur within the region. Therefore, identifying recombination breakpoints based on the variation of alignment would contribute to uncovering the genomic evolutionary histories in terms of phylogenetic incongruence.

It has been observed that the HIV-1 genome is highly mosaic within the regions of envelope gene (*env*), polymerase gene (*pol*) and *gag* [1, 25]. The highly chimeric genomes of HIV-1 isolates are generated by inter-subtype or inter-group recombination of divergent strains under selective pressures, selective advantages, rapid evolution or immune response after infection[36, 42]. Studies have shown that recombination is responsible for HIV-1 evolution, suggesting that exploration of the recombination patterns and hotspots is crucial for characterization of pathogenesis. In our analysis, 72 sites are polymorphic out of 1593 site of alignment, yet 22 polymorphic sites are not compatible to the global tree topology, showing highly mosaic in the sequences. The alignment is further divided to three segments using our ACR approach. As a result, the changes per site for the first and second segments are lower than the global one, and the phylogenies of them are distinct from the global tree, showing that the evolutionary histories are inconsistent. Furthermore, the last 11 consecutive polymorphic sites located within the third segment are quite incompatible with each other.

6 CONCLUSIONS

The proposed average compatibility ratio approach is able to quickly screen for multiple sequence alignments to detect the presence of recombination and then characterize the breakpoints. Non-overlapping segments divided by the identified breakpoints have higher compatibility ratios, lower homoplasy ratios and distinct phylogenies, showing that adjacent sites within a segment are more jointly compatible and are more likely to be described by a single tree. Our analyses of simulated and empirical datasets and comparisons of two programs demonstrate that our ACR approach is effective and efficient to identify recombinant regions in bacterial and viral genomes. It could potentially provide a better understanding of phylogenetic relationships and evolutionary history among a set of taxa.

REFERENCES

[1] John Archer, John W Pinney, Jun Fan, Etienne Simon-Loriere, Eric J Arts, Matteo Negroni, and David L Robertson. 2008. Identifying the important HIV-1 recombination breakpoints. *PLoS Comput Biol* 4, 9 (2008), e1000178.
[2] C Arnold. 2007. Molecular evolution of Mycobacterium tuberculosis. *Clinical microbiology and infection* 13, 2 (2007), 120–128.
[3] Trevor C Bruen, Hervé Philippe, and David Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172, 4 (2006), 2665–2681.
[4] Josephine Bryant, Claire Chewapreecha, and Stephen D Bentley. 2012. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future microbiology* 7, 11 (2012), 1283–1296.
[5] Cheong Xin Chan, Robert G Beiko, and Mark A Ragan. 2006. Detecting recombination in evolving nucleotide sequences. *BMC bioinformatics* 7, 1 (2006), 412.
[6] Nicholas J Croucher, Simon R Harris, Christophe Fraser, Michael A Quail, John Burton, Mark van der Linden, Lesley McGee, Anne von Gottberg, Jae Hoon Song,

Kwan Soo Ko, and others. 2011. Rapid pneumococcal evolution in response to clinical interventions. *science* 331, 6016 (2011), 430–434.

[7] Xavier Didelot, Rory Bowden, Teresa Street, Tanya Golubchik, Chris Spencer, Gil McVean, Vartul Sangal, Muna F Anjum, Mark Achtman, Daniel Falush, and others. 2011. Recombination and population structure in Salmonella enterica. *PLoS Genet* 7, 7 (2011), e1002191.

[8] Xavier Didelot and Daniel J Wilson. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11, 2 (2015), e1004041.

[9] Elizabeth M Driebe, Jason W Sahl, Chandler Roe, Jolene R Bowers, James M Schupp, John D Gillece, Erin Kelley, Lance B Price, Talima R Pearson, Crystal M Hepp, and others. 2015. Using whole genome analysis to examine recombination across diverse sequence types of Staphylococcus aureus. *PloS one* 10, 7 (2015), e0130955.

[10] George F Estabrook, CS Johnson, and FR McMorris. 1976. A mathematical foundation for the analysis of cladistic character compatibility. *Mathematical Biosciences* 29, 1-2 (1976), 181–187.

[11] George F Estabrook and FR McMorris. 1980. When is one estimate of evolutionary relationships a refinement of another? *Journal of Mathematical Biology* 10, 4 (1980), 367–373.

[12] Graham J Etherington, Jo Dicks, and Ian N Roberts. 2004. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics* 21, 3 (2004), 278–281.

[13] Richard G Everitt, Xavier Didelot, Elizabeth M Batty, Ruth R Miller, Kyle Knox, Bernadette C Young, Rory Bowden, Adam Auton, Antonina Votintseva, Hanna Larner-Svensson, and others. 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nature communications* 5 (2014).

[14] Nicholas C Grassly and Edward C Holmes. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution* 14, 3 (1997), 239–247.

[15] Michaela M Gutacker, James C Smoot, Cristi A Lux Migliaccio, Stacy M Ricklefs, Su Hua, Debby V Cousins, Edward A Graviss, Elena Shashkina, Barry N Kreiswirth, and James M Musser. 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 162, 4 (2002), 1533–1543.

[16] Ingrid B Jakobsen and Simon Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer applications in the biosciences: CABIOS* 12, 4 (1996), 291–295.

[17] BG Kelly, A Vespermann, and DJ Bolton. 2009. The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens. *Food and Chemical Toxicology* 47, 5 (2009), 951–968.

[18] Elzbieta Krzywinska, Jaroslaw Krzywinski, and Jeffrey S Schorey. 2004. Naturally occurring horizontal gene transfer and homologous recombination in Mycobacterium. *Microbiology* 150, 6 (2004), 1707–1712.

[19] Fedora Lanzas, Petros C Karakousis, James C Sacchettini, and Thomas R Ioerger. 2013. Multidrug-resistant tuberculosis in panama is driven by clonal expansion of a multidrug-resistant Mycobacterium tuberculosis strain related to the KZN extensively drug-resistant M. tuberculosis strain from South Africa. *Journal of clinical microbiology* 51, 10 (2013), 3277–3285.

[20] Walter J Le Quesne. 1969. A method of selection of characters in numerical taxonomy. *Systematic Biology* 18, 2 (1969), 201–205.

[21] Walter J LeQuesne. 1972. Further studies based on the uniquely derived character concept. *Systematic Biology* 21, 3 (1972), 281–288.

[22] Darren Martin and Ed Rybicki. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16, 6 (2000), 562–563.

[23] Darren P Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* 1, 1 (2015), vev003.

[24] Christopher A Meacham and George F Estabrook. 1985. Compatibility methods in systematics. *Annual Review of Ecology and Systematics* 16, 1 (1985), 431–446.

[25] Adewunmi Onafuwa-Nuga and Alice Telesnitsky. 2009. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiology and Molecular Biology Reviews* 73, 3 (2009), 451–480.

[26] Talima Pearson, Joseph D Busch, Jacques Ravel, Timothy D Read, Shane D Rhoton, Jana M U'ren, Tatum S Simonson, Sergey M Kachur, Rebecca R Leadem, Michelle L Cardon, and others. 2004. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 101, 37 (2004), 13536–13541.

[27] Hervé Philippe, Henner Brinkmann, Dennis V Lavrov, D Timothy J Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* 9, 3 (2011), e1000602.

[28] Paul J Planet, Apurva Narechania, Liang Chen, Barun Mathema, Sam Boundy, Gordon Archer, and Barry Kreiswirth. 2016. Architecture of a Species: Phylogenomics of Staphylococcus aureus. *Trends in Microbiology* (2016).

[29] DOTREE Plotree and DOTGRAM Plotgram. 1989. PHYLIP-phylogeny inference package (version 3.2). *cladistics* 5, 163 (1989), 6.

[30] Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon DW Frost. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution* 23, 10 (2006), 1891–1901.

[31] Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon DW Frost. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 24 (2006), 3096–3098.

[32] Sergei L Kosakovsky Pond, David Posada, Eric Stawiski, Colombe Chappey, Art FY Poon, Gareth Hughes, Esther Fearnhill, Mike B Gravenor, Andrew J Leigh Brown, and Simon DW Frost. 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput Biol* 5, 11 (2009), e1000581.

[33] David Posada. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular biology and evolution* 19, 5 (2002), 708–717.

[34] David Posada and Keith A Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences* 98, 24 (2001), 13757–13762.

[35] David Posada and Keith A Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution* 54, 3 (2002), 396–402.

[36] MIKA O SALMINEN, JEAN K CARR, DONALD S BURKE, and FRANCINE E McCUTCHAN. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS research and human retroviruses* 11, 11 (1995), 1423–1425.

[37] David Sankoff. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28, 1 (1975), 35–42.

[38] Célio D Santos-Júnior, António Veríssimo, and Joana Costa. 2016. The recombination dynamics of Staphylococcus aureus inferred from spA gene. *BMC microbiology* 16, 1 (2016), 143.

[39] SA Sawyer. 1999. GENECONV: A computer package for the statistical detection of gene conversion. *Distributed by the author, Department of Mathematics, Washington University in St. Louis* (1999).

[40] Mikkel H Schierup and Jotun Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156, 2 (2000), 879–891.

[41] Robert W Scotland and Mike Steel. 2015. Circumstances in which parsimony but not compatibility will be provably misleading. *Systematic biology* 64, 3 (2015), 492–504.

[42] Binshan Shi, Christina Kitchen, Barbara Weiser, Douglas Mayers, Brian Foley, Kimdar Kemal, Kathryn Anastos, Marc Suchard, Monica Parker, Cheryl Brunner, and others. 2010. Evolution and recombination of genes encoding HIV-1 drug resistance and tropism during antiretroviral therapy. *Virology* 404, 1 (2010), 5–20.

[43] John Maynard Smith. 1992. Analyzing the mosaic structure of genes. *Journal of molecular evolution* 34, 2 (1992), 126–129.

[44] Christopher M Thomas and Kaare M Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology* 3, 9 (2005), 711–721.

[45] David B Wake. 1991. Homoplasy: the result of natural selection, or evidence of design limitations? *The American Naturalist* 138, 3 (1991), 543–567.

[46] Timothy M Walker, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, and others. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases* 13, 2 (2013), 137–146.

[47] Georg F Weiller. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution* 15, 3 (1998), 326–335.

[48] Michael Worobey. 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Molecular Biology and Evolution* 18, 8 (2001), 1425–1434.