Computing the Joint Distribution of Tree Shape and Tree Distance for Gene Tree Inference and Recombination Detection

Yujin Chung, Nicole T. Perna, and Cécile Ané

Abstract—Ancestral recombination events can cause the underlying genealogy of a site to vary along the genome. We consider Bayesian models to simultaneously detect recombination breakpoints in very long sequence alignments and estimate the phylogenetic tree of each block between breakpoints. The models we consider use a dissimilarity measure between trees in their prior distribution to favor similar trees at neighboring loci. We show empirical evidence in Enterobacteria that neighboring genomic regions have similar trees. The main hurdle in using such models is the need to properly calculate the normalizing function for the prior probabilities on trees. In this work, we quantify the impact of approximating this normalizing function as done in biomc2, a hierarchical Bayesian method to detect recombination based on distance between tree topologies. We then derive an algorithm to calculate the normalizing function exactly, for a Gibbs distribution based on the Robinson-Foulds (RF) distance between gene trees at neighboring loci. At the core is the calculation of the joint distribution of the shape of a random tree and its RF distance to a fixed tree. We also propose fast approximations to the normalizing function, which are shown to be very accurate with little impact on the Bayesian inference.

Index Terms—Phylogenetic tree, recombination, Robinson-Foulds distance, normalizing function, gene tree discordance

1 Introduction

Recombination occurs in the genomes of many organisms leading to exchange of genetic material. In eukaryotes, recombination is reciprocal. In prokaryotic organisms, homologous recombination leads to a unidirectional flow of genetic material from a donor to a recipient, more akin to eukaryotic gene conversion. This is one type of horizontal gene transfer (HGT) that is particularly common among closely related organisms, such as within species of Enterobacteria. Recombination events can complicate the analysis of the evolution of a group of organisms, as they can cause conflicting phylogenetic relationships between different regions of the genomes. Recently developed statistical methods simultaneously detect the location of recombination events along an alignment and infer phylogenetic histories of regions in the alignment defined by recombination breakpoints. These methods are based on the premise that discordant phylogenetic trees from different genomic regions are due to recombination events. RecPars [1] infers the most parsimonious history of substitutions on trees and recombination, and MDL [2] enables a penalty parameter to control the number of breakpoints. PLATO [3] infers the maximum-likelihood phylogenetic tree from the whole input alignment, and then detects regions whose likelihood values for this tree are relatively small. Similarly, ClonalOrigin [4] estimates the phylogenetic tree of the genome and recombination breakpoints in a two-stage hierarchical Bayesian framework. Hidden Markov models (HMM) assume that hidden states are the underlying trees of genetic regions [5], [6], [7], [8]. DualBrothers [9] is an extension of the first Bayesian method [10] to infer breakpoint positions and phylogenetic trees simultaneously, but works well on only few taxa. cBrother [11] improved the computational issues of DualBrothers, and StepBrothers [12] further infers relative times of recombination events. Biomc2 [13] incorporates correlation in tree topologies through the distance between trees at neighboring regions in a Bayesian model and is able to handle larger data sets.

Although tree topologies of regions between recombination breakpoints are different, these genomic regions share some evolutionary history before and after the recombination events. For example, in Fig. 1, the gray genomic regions in taxa C and D have a different evolutionary history than the white genomic regions: genes in the gray region in taxa C and D are more closely related to genes in taxon B than to genes in taxon E, but genes in the white region are more closely related to genes in taxon E. However, the trees of the gray and white regions share features during time periods t_1 (A is an outgroup in both regions) and t_3 (both regions have clade CD in their trees).

Models that favor similar trees at neighboring genomic regions can detect breakpoints between very similar trees [7] and inference can be more accurate [8]. As far as we know, biomc2 is one of the few methods that take into account correlation between tree topologies: topologies at adjacent genomic regions can be different but preferably similar (but see [4], [12]). Note that other methods such as cBrother and HMM uniformly prefer different topologies of adjacent genomic regions. Empirical evidence for correlation between trees at adjacent genomic regions is assessed

- *Y. Chung is with the Department of Statistics, University of Wisconsin, Madison, WI, and the Center for Computational Genetics and Genomics, Department of Biology, Temple University, Philadelphia, PA. E-mail: ychung@stat.wisc.edu.*
- *N.T. Perna is with the Genome Center and the Department of Genetics, University of Wisconsin, Madison, WI 53706. E-mail: ntperna@wisc.edu.*
- *C. Ané is with the Department of Statistics, University of Wisconsin, Madison, WI 53706. E-mail: ane@stat.wisc.edu.*

Manuscript received 22 Jan. 2013; revised 23 July 2013; accepted 3 Sept. 2013; published online 13 Sept. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2013-01-0027. Digital Object Identifier no. 10.1109/TCBB.2013.109.

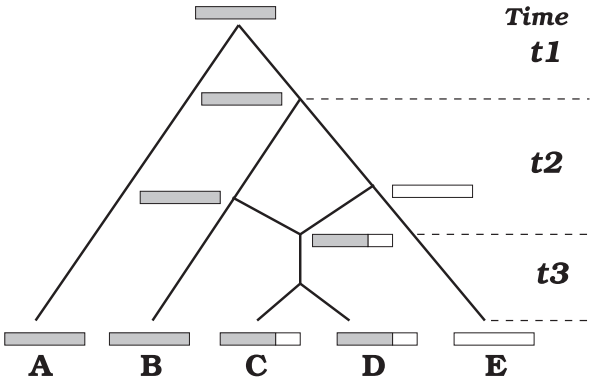

Fig. 1. Illustration of different tree topologies of genomic regions because of recombination event. The phylogenetic tree of the gray region has the clade BCD, but the white region has the clade CDE.

in Section 2. Biomc2 uses approximated subtree prune and regraft (SPR) distances ($\hat{d}_{\mathrm{SPR}}$) between tree topologies at adjacent predefined segments. The SPR distance is considered to have a truncated-Poisson distribution a priori with parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{L-1})$, using the following probability-like function on tree topologies $\mathbf{T} = (T_1, \ldots, T_L)$:

$$\tilde{P}(\mathbf{T} \mid \boldsymbol{\beta}, \mathbf{w}) = \frac{\prod_{l=1}^{L-1} \left\{ \frac{e^{-\beta_l} \beta_l^{\hat{d}_{\mathrm{SPR}}(T_l, T_{l+1})}}{\hat{d}_{\mathrm{SPR}}(T_l, T_{l+1})!} \right\}^{w_l+1}}{\tilde{\eta}(\boldsymbol{\beta}, \mathbf{w}, L)}, \tag{1}$$

where N is the number of taxa, L is the number of segments in the alignment, $\mathbf{w} = (w_1, \ldots, w_{L-1})$ are nonnegative weights,

$$\tilde{\eta}(\boldsymbol{\beta}, \mathbf{w}, L) = \prod_{l=1}^{L-1} \left[\sum_{d=0}^{D} \left\{ \frac{e^{-\beta_l} \beta_l^d}{d!} \right\}^{w_l+1} \right], \tag{2}$$

and $D = N - 3$ is the number of internal edges and an upper bound for the SPR distance. The function $\tilde{\eta}$ used by [13] is meant to normalize $\tilde{P}$ so that $\tilde{P}$ is a probability distribution: $\tilde{\eta}$ should ensure that the probabilities sum up to 1. Such a function is called a normalizing function. The parameter β_l is larger than the expected distance between trees T_l and T_{l+1} because the Poisson distribution is truncated. For convenience, however, it is interpreted here as the prior mean distance between neighboring trees. Weights $\mathbf{w}$ enable the distribution (1) to have smaller mean and variance, so as to give higher probabilities to similar trees at adjacent segments.

One difficulty for Gibbs-like distributions such as the prior distribution (1) used in biomc2 is that normalizing functions are easily overlooked or miscalculated. For example, Gibbs-like distributions have also been used for supertree estimation. A model to find the maximum-likelihood supertree from estimated smaller phylogenetic trees was proposed by [14]. Estimated gene trees can be different from the true tree on the full taxon set ("supertree") because of technical issues (e.g., incorrect orthology detection), stochastic error (e.g., estimation error), or biological processes (e.g., incomplete lineage sorting). In [14], the discrepancy between gene trees T_i and the supertree $\mathcal{T}$ is modeled using the Robinson-Foulds (RF) distance [15] d and the likelihood of $\mathcal{T}$ as

$$P_{\mathcal{T}}(T_1, \ldots, T_k) \propto \prod_{i=1}^{k} \exp[-\beta_i d(T_i, \mathcal{T})], \tag{3}$$

where $T_1, \ldots, T_k$ are the k input gene trees estimated on taxon subsets. The "maximum-likelihood supertree" proposed by [14] is

$$\arg\min_{\mathcal{T}} \sum_{i=1}^{k} \beta_i d(T_i, \mathcal{T}). \tag{4}$$

However, as pointed out in [16], the correct likelihood maximization should normalize the term (3) using

$$Z_{\mathcal{T},\boldsymbol{\beta}} = \prod_{i=1}^{k} Z_{\mathcal{T},\beta_i}^{(i)},$$

with

$$Z_{\mathcal{T},\beta_i}^{(i)} = \sum_{T:\mathcal{L}(T)=\mathcal{L}(T_i)} \exp[-\beta_i d(T, \mathcal{T})], \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, and $\mathcal{L}(T_i)$ is the set of tip labels of gene tree T_i. In [16], criterion (4) is corrected as

$$\arg\min_{\mathcal{T}} \left\{ \sum_{i=1}^{k} \beta_i d(T_i, \mathcal{T}) + \log Z_{\mathcal{T},\boldsymbol{\beta}} \right\},$$

and a polynomial-time algorithm is described to calculate the distribution of the RF distance given the tree shape of $\mathcal{T}$. In biomc2, the function $\tilde{\eta}$ used in the prior distribution (1) on trees is *not* the actual normalizing function, that is, (1) is not a probability distribution because

$$\sum_{T_1} \cdots \sum_{T_L} \tilde{P}(T_1, \ldots, T_L \mid \boldsymbol{\beta}, \mathbf{w}) \neq 1.$$

The correct normalizing function is

$$\eta(\boldsymbol{\beta}, \mathbf{w}, L) = \sum_{T_1} \cdots \sum_{T_L} \prod_{l=1}^{L-1} \left\{ \frac{e^{-\beta_l} \beta_l^{d(T_l, T_{l+1})}}{d(T_l, T_{l+1})!} \right\}^{w_l+1}. \tag{6}$$

In other words, the numerator in (1) should be summed over all possible trees, not over tree distance values. The normalizing function in biomc2 has not been corrected in its implementation or in publication as far as we know. More generally, complex computation of normalizing functions makes it difficult to embed correlations among tree topologies into statistical models.

To simultaneously detect recombination breakpoints and infer phylogenetic trees of genomic regions, we propose a method with a new Gibbs prior distribution. The Gibbs probability of a random variable X having value x is

$$P(X = x) = \frac{1}{Z(\beta)} \exp(-\beta E(x)),$$

where $E(x)$ is called the energy function of the configuration x, β is a parameter called the inverse temperature [17], and $Z(\beta)$ is the normalizing function, also called a partition function. We consider the sum of RF distances (interpreted as dissimilarities) between tree topologies at adjacent genomic regions as the energy of the phylogenetic histories of the genomic regions. We use here the RF distance to measure the presence of recombination, because a positive

RF distance between trees at adjacent sites implies the presence of one or more recombination events. However, we do not use RF distances to measure the amount of recombination (as would be done with the SPR distance) because the RF distance between trees at adjacent sites does not scale with the number of recombination events at that location [13].

Section 2 shows empirical evidence in Enterobacteria that neighboring genomic regions have similar trees. In Section 3, the impact of overlooking the normalizing function in biomc2 is investigated. In Section 4, a Bayesian model is introduced to simultaneously identify recombination breakpoints and infer phylogenetic histories. For this, we use the Gibbs distribution mentioned above as a prior distribution on tree topologies. Section 5 shows that the normalizing function of the Gibbs distribution can be calculated through the number of tree topologies with a certain shape and at a certain distance away from a given tree topology. In Section 6 we propose approximations to the normalizing function. Conclusion and discussion are in Section 7.

2 Correlation among Gene Trees in Real Data

To motivate the models considered here, we first investigate the level of spatial correlation between phylogenetic trees at neighboring loci in real data. ProgressiveMauve was applied to generate alignments of 33 *Escherichia* genomes and 8 *Shigella* genomes [18]. The longest alignment among those with nonempty sequences from all of the 41 taxa contained 52,080 base pairs (bps). We partitioned this alignment into 103 segments of 500 bps and 1 segment of 580 bps. We excluded segments from the analysis if they shared less than 4 taxa with all other segments; 76 segments remained. We applied MrBayes [19] to each segment independently. The HKY model [20] with gamma-distributed rates across sites was used with four chains, two independent runs, and 10 million generations. Trees were sampled every 100th generation and the first 10 percent were discarded. We estimated the phylogenetic tree of each segment by the greedy consensus tree with posterior probabilities on internal edges.

We modified the RF distance to account for posterior probabilities of internal edges in the trees, to give lower weight to edges with high uncertainty. Our modified weighted RF (wRF) distance is also normalized to compare subtrees on identical taxon sets,

$$wRF(T_1,T_2)=\frac{\sum_{\substack{c\in\mathcal{C}(T_{1|L})\\ c\notin\mathcal{C}(T_{2|L})}} pp_1(c)+\sum_{\substack{c\notin\mathcal{C}(T_{1|L})\\ c\in\mathcal{C}(T_{2|L})}} pp_2(c)}{\sum_{c\in\mathcal{C}(T_{1|L})} pp_1(c)+\sum_{c\in\mathcal{C}(T_{2|L})} pp_2(c)},$$

where $L=L_1\cap L_2$ is the set of taxa that are common to both trees T_1 and T_2, $T_{i|L}(i=1,2)$ is the subtree obtained from T_i after pruning taxa whose labels are not in the other tree, $\mathcal{C}(T)$ is the collection of all bipartitions in tree T, and $pp_i(c)$ is the posterior probability of c for tree T_i. Since the wRF distance was scaled between 0 and 1, it provides comparable distances between trees of different sizes. We computed the wRF distance between the consensus trees from all pairs of segments.

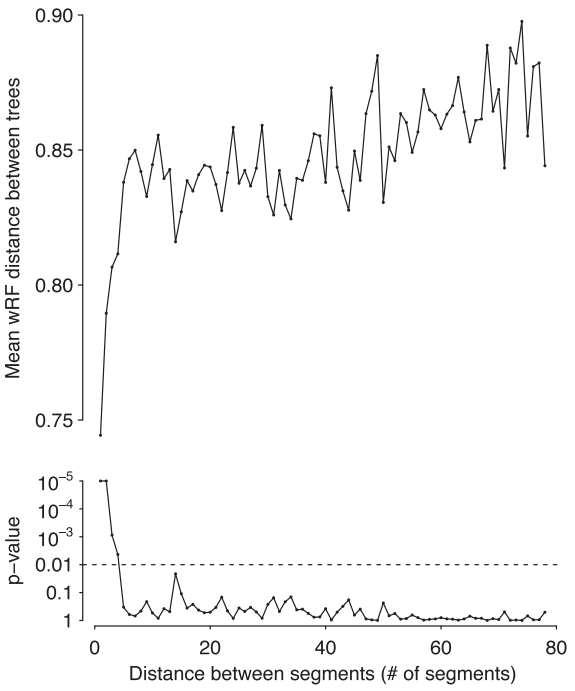

Fig. 2. Average wRF distance between trees from 500-bp segments that are a given physical distance apart in the alignment (top). For each k, a permutation test was conducted to determine if trees of loci located k segments apart are more similar than trees from randomly selected loci (bottom). The test was significant (p-value < 0.01) for $k \leq 4$ segments only (i.e., 2 kb).

To determine if trees from nearby segments are more similar (positively correlated) than trees from randomly selected segments, a permutation test was conducted on the wRF distance between trees from segments located k segments apart. We randomly shuffled the greedy consensus trees along the alignment, and then computed average wRF distances between trees located k segments away from each other. We repeated the process 100,000 times and calculated p-values by counting the number of times that the sampled average wRF distance was smaller than the observed average wRF distance.

The average wRF distance roughly increases with the physical distance between segments (see Fig. 2). At the 1 percent significance level, trees from regions no more than 2 kb (four segments) apart were significantly more similar to each other than to trees from other regions. The correlation between trees was too weak to be detected in our experiment across distances beyond 2 kb.

3 Importance of the Normalizing Function

Biomc2 is one of the few methods that take this correlation between tree topologies into account, when estimating trees $(T_1,\ldots,T_L)$ along an alignment with L predefined short segments. For the prior distribution on tree topologies, biomc2 considers the truncated-Poisson distribution in (1) parameterized by $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_L)$. Independent gamma hyperprior distributions are placed on β_l and w_l. In this section, we identify an issue with the normalizing function $\tilde{\eta}$ (2) currently implemented in biomc2.

Not fully knowing the prior distribution might not be a problem under a Markov chain Monte Carlo (MCMC) approach, where the prior distribution needs to be known

only up to a constant. Assume we want to use a prior distribution on tree topologies, $P(T_1, \ldots, T_L \mid \beta)$, that cannot be easily calculated. Suppose that P is replaced in the MCMC algorithm by $\tilde{P}$ where the product

$$\tilde{P}(T_1, \ldots, T_L \mid \beta)\tilde{f}(\boldsymbol{\beta}) = P(T_1, \ldots, T_L \mid \boldsymbol{\beta})f(\boldsymbol{\beta}),$$

is easily evaluated at each step of the MCMC. Here $f(\boldsymbol{\beta})$ is the real normalizing function for the true prior distribution P but is difficult to calculate. Instead, $\tilde{f}(\boldsymbol{\beta})$ is a pseudonormalizing function easier to calculate. If a fixed $\boldsymbol{\beta}$ is used to infer the posterior distribution of tree topologies, it is fine to use the pseudonormalizing function $\tilde{f}(\boldsymbol{\beta})$ or to simply ignore the true normalizing function $f(\boldsymbol{\beta})$. If we assume a hyperprior distribution $\pi(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$, however, then

$$\tilde{P}(T_1, \ldots, T_L \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}) = P(T_1, \ldots, T_L \mid \boldsymbol{\beta})\frac{f(\boldsymbol{\beta})}{\tilde{f}(\boldsymbol{\beta})}\pi(\boldsymbol{\beta}).$$

If $\tilde{P}$ is used instead of P in an MCMC approach to define the prior probability of trees given $\boldsymbol{\beta}$, then the hyperprior distribution *actually used* on $\boldsymbol{\beta}$ is

$$\tilde{\pi}(\boldsymbol{\beta}) = \frac{f(\boldsymbol{\beta})}{\tilde{f}(\boldsymbol{\beta})}\pi(\boldsymbol{\beta})\left[\int \frac{f(\boldsymbol{\beta}')}{\tilde{f}(\boldsymbol{\beta}')}\pi(\boldsymbol{\beta}')d\boldsymbol{\beta}'\right]^{-1}. \tag{7}$$

If the ratio $f/\tilde{f}$ is a constant, i.e., the true normalizing function is known up to a constant, then the MCMC sampling procedure is not affected by the usage of the function $\tilde{P}(T_1, \ldots, T_L \mid \beta)$ as a prior distribution. However, a problem arises when $f/\tilde{f}$ is not constant, since MCMC samples are not from the assumed posterior distribution in that case. The correct calculation of the normalizing function is then required.

To see the impact of using the pseudonormalizing function (2) rather than the true normalizing function (6) in biomc2, we compare the ratio of normalizing functions $\eta/\tilde{\eta}$ and calculate the hyperprior distribution on $\boldsymbol{\beta}$ actually used as determined by (7). Either comparison is not easy to carry out analytically, so we consider here a simple case when all w_l's are fixed to 0 and β_l's are all equal. The function used as a prior distribution on tree topologies in (1) becomes

$$\tilde{P}(T_1, \ldots, T_L \mid \beta) = \frac{1}{\tilde{\eta}(\beta, L)}\prod_{l=1}^{L-1}\frac{e^{-\beta}\beta^{d_{\mathrm{SPR}}(T_l, T_{l+1})}}{d_{\mathrm{SPR}}(T_l, T_{l+1})!}, \tag{8}$$

where d_{SPR} is the true SPR distance between tree topologies and the pseudonormalizing function is

$$\tilde{\eta}(\beta, L) = \prod_{l=1}^{L-1}\left[\sum_{d=0}^{D}\frac{e^{-\beta}\beta^{d}}{d!}\right].$$

The correct normalizing function for (8) is

$$\eta(\beta, L) = \sum_{T_1}\cdots\sum_{T_L}\prod_{l=1}^{L-1}\frac{e^{-\beta}\beta^{d_{\mathrm{SPR}}(T_l, T_{l+1})}}{d_{\mathrm{SPR}}(T_l, T_{l+1})!}.$$

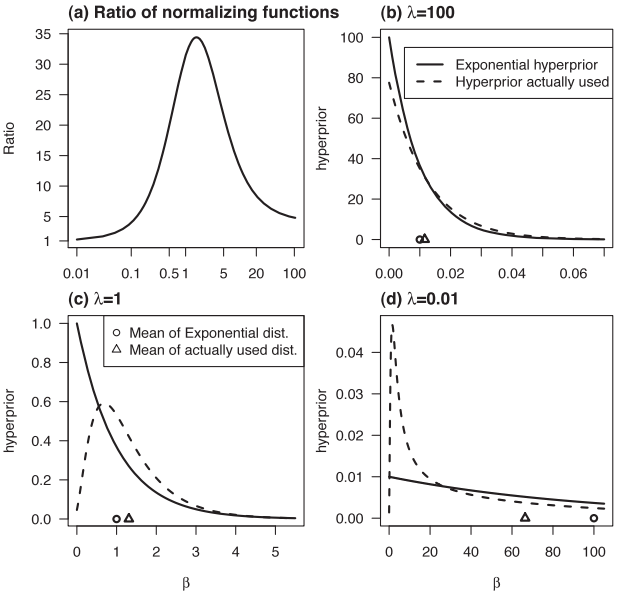

Fig. 3. Impact of using the pseudonormalizing function $\tilde{P}$ in (8). The alignment has two candidate breakpoints ($L = 3$) and $N = 5$ taxa. (a) Ratio of the true normalizing function to the pseudonormalizing function $\eta/\tilde{\eta}$. (b)-(d) The real line indicates the exponential distribution $\mathcal{E}(\lambda)$ whose mean $1/\lambda$ is indicated with a circle (○). The hyperprior density actually used $\tilde{\pi}$ is indicated with a dotted line (- -) whose mean is indicated with a triangle (Δ) when $\lambda =$ 100,1, and 0.01. Note that the axis for β in (a) is on a log scale.

We follow [21] and choose an exponential distribution $\mathcal{E}(\lambda)$ with mean $1/\lambda$ for the distribution of β, which is a special case of the gamma distribution. The hyperprior distribution actually used is then

$$\tilde{\pi}(\beta) \propto \pi(\beta)\frac{\eta(\beta, L)}{\tilde{\eta}(\beta, L)} = \lambda e^{-\lambda\beta} \times \sum_{T_1}\cdots\sum_{T_L}\left\{\frac{\beta^{\sum_{l=1}^{L-1} d_{\mathrm{SPR}}(T_l, T_{l+1})}}{\prod_{l=1}^{L-1} d_{\mathrm{SPR}}(T_l, T_{l+1})!}\right\}\Bigg/\left(\sum_{d=0}^{D}\frac{\beta^d}{d!}\right)^{L-1}.$$

When there are two candidate recombination breakpoints ($L = 3$) and $N = 5$ taxa, the ratio of the true normalizing function to the pseudonormalizing function can be calculated exactly and it is not a constant (see Fig. 3a), although the ratio converges to 4 as β increases. The hyperprior distribution actually used is

$$\tilde{\pi}(\beta) \propto \frac{1 + 24\beta + 146\beta^2 + 24\beta^3 + \beta^4}{(1 + \beta + \beta^2/2)^2}\lambda e^{-\lambda\beta}, \tag{9}$$

which differs from the targeted hyperprior distribution $\mathcal{E}(\lambda)$, as shown in Figs. 3b, 3c, and 3d for $\lambda = 100, 1$ and 0.01. The exponential density has a mode at $\beta = 0$, but the density actually used (9) has very small values near $\beta = 0$ except for $\lambda = 100$. This discrepancy might partly explain why it is recommended to use a very large λ in biomc2 and even more so when there are more candidate recombination breakpoints; or why it is recommended to use a prior distribution for the w_i values. Indeed, fixing them to 0 was shown to cause an overestimation of recombination [21]. Larger values of w_i decrease the prior mean distance between trees, which might counteract the effect of the pseudonormalizing function.

4 Gibbs Model to Infer Recombination Breakpoints and Phylogenetic Trees

In this section, we lay the mathematical foundation for a Gibbs-distribution-based method, where the issue of the normalizing function can be solved. Our approach to simultaneously infer recombination breakpoints and phylogenetic trees involves a hierarchical model with a Gibbs distribution on tree topologies given a prior frequency of recombination breakpoints, and a sequence evolution model (as in [13]) for the likelihood of sequence data given the tree topologies. We consider below a long alignment divided into L predefined arbitrary short segments, which may have different tree topologies $\mathbf{T} = (T_1, \ldots, T_L)$ because of recombination. Within a segment, all sites are assumed to have the same phylogenetic tree. Our focus here is on a new Gibbs prior distribution on $(T_1, \ldots, T_L)$ to take into account the similarity of trees across consecutive segments, and for which the normalizing function can be calculated. Tree similarity is measured by the RF distance, which is meant to detect the presence of recombination (not the amount). The RF distance between two fully resolved unrooted trees is the number of bipartitions found in only one of the two trees. It has an even value and the distance $d(\cdot,\cdot)$ used here is one-half of the RF distance. The proposed prior probability of tree topologies is then

$$P(T_1, \ldots, T_L) = \exp\left(-\beta \sum_{l=1}^{L-1} d(T_l, T_{l+1})\right) \Big/ Z_L(\beta), \qquad (10)$$

where β a nonnegative parameter and $Z_L(\beta)$ is the normalizing function

$$Z_L(\beta) = \sum_{t_1} \cdots \sum_{t_L} \exp\left(-\beta \sum_{l=1}^{L-1} d(t_l, t_{l+1})\right), \qquad (11)$$

to ensure that the probabilities in (10) sum to 1. When there is only $L = 1$ segment, $Z_1(\beta) = Z_1 = (2N-5)!!$ is the total number of tree topologies and does not depend on β.

Under this Gibbs distribution, similar trees at adjacent segments are favored. For large β, $Z_L(\beta)$ approaches Z_1 and the Gibbs distribution forces all trees to be identical:

$$P(T_1, \ldots, T_L) = \begin{cases} 1/Z_1 & \text{if } T_1 = \cdots = T_L, \\ 0 & \text{otherwise,} \end{cases}$$

as if no recombination occurred. When $\beta = 0$, $Z_L(\beta) = Z_1^L$, and the Gibbs probability $P(T_1 = t_1, \ldots, T_L = t_L) = 1/Z_1^L$ regardless of the values of $t_1, \ldots, t_L$. In other words, trees become independent, each with a uniform distribution. Between these two extreme distributions, $1/\beta$ scales with the average recombination rate per segment. We will informally call β the a priori inverse recombination rate.

The Gibbs distribution has desirable properties, such as the following Markov property, by the Hammersley-Clifford theorem [22]:

$$\begin{aligned} &P(T_j = t \mid T_1, \ldots, T_{j-1}, T_{j+1}, \ldots, T_L) \\ &\quad = P(T_j = t \mid T_{j-1}, T_{j+1}), \text{ for } j = 2, \ldots, L-1. \end{aligned} \qquad (12)$$

In other words, conditional on its neighbors, a tree T_j is independent of the trees at all other segments. Moreover, the distribution is homogeneous across the alignment: $P(T_{i+1} = t_2 \mid T_i = t_1, T_{i+2} = t_3)$ is independent of i. Additionally, the sequence of tree topologies from the Gibbs distribution is a nonstationary Markov chain: β parameterizes the transition rate between T_i and T_{i+1} with lower β resulting in a higher probability of change and the transition probabilities are generally inhomogeneous in i, as are the marginal distributions.

We define a block as the collection of all the consecutive segments located between two recombination breakpoints. In other words, segments (and sites) within a block are inferred to have the same tree topology while segments in two adjacent blocks are inferred to have different tree topologies. Knowing the prior distribution on the number of breakpoints B can be useful to choose an appropriate value for the recombination rate, or an appropriate hyperprior mean if the inverse recombination rate, β, is given a hyperprior distribution. Indeed, the following proposition (proved in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.109) links β to the expected number of recombination breakpoints.

Proposition 1. *Assume the Gibbs distribution (10). When $\beta = 0$, the number of recombination breakpoints B has a binomial distribution $\mathcal{B}(L-1, 1-1/Z_1)$ with expectation $E(B) = (L-1)(1-1/Z_1)$. If $\beta = \infty$, there is exactly one block: $B = 0$ with probability 1. In general,*

$$P(B = b) = \binom{L-1}{b} \sum_{i=0}^{b} \frac{\binom{b}{i}(-1)^{b-i} Z_{i+1}(\beta)}{Z_L(\beta)},$$

with an expected number of breakpoints

$$E(B) = (L-1)\left(1 - \frac{Z_{L-1}(\beta)}{Z_L(\beta)}\right).$$

At each segment boundary, the probability of there being a recombination breakpoint is $1 - Z_{L-1}/Z_L$. This is maximum at $1 - 1/Z_1$ when the recombination rate is very large ($\beta = 0$). When the recombination rate is small (large β), this is approximately $2(N-3)e^{-\beta}$ (see Appendix A, which can be found in the online supplemental material).

5 The Gibbs Normalizing Function

When tree topologies $(T_1, \ldots, T_L)$ of consecutive segments follow the Gibbs distribution in (10), the corresponding normalizing function $Z_L(\beta)$ in (11) depends on β. With this model, it is necessary to either compute the normalizing function exactly or to provide a good approximation for it if we want to place a hyperprior on β, such as an exponential distribution with mean $1/\lambda$. In this section, we develop an algorithm to calculate $Z_L(\beta)$ exactly. We can rewrite

$$Z_L(\beta) = \sum_{S_1 \in \mathcal{S}_N} \zeta(S_1) Z_{L,S_1}(\beta), \qquad (13)$$

where the sum goes over the set of unrooted tree shapes $\mathcal{S}_N$ on N tips, $\zeta(S) = |\{T : \mathbf{S}(T) = S\}|$, $\mathbf{S}(T)$ denotes a tree shape from tree T by discarding the terminal node labels, and

$$Z_{L,S_1}(\beta) = \sum_{T_2} \cdots \sum_{T_L} \exp\left\{-\beta \sum_{l=1}^{L-1} d(T_l, T_{l+1})\right\}$$

for any fixed T_1 of shape S_1. The value of $Z_{L,S_1}(\beta)$ can be recursively computed as

$$Z_{L,S_1}(\beta) = \sum_{S_2 \in \mathcal{S}_N} \sum_{y=0}^{N-3} \zeta_{2,S_1}(S_2, y) e^{-\beta y} Z_{L-1,S_2}(\beta), \tag{14}$$

where, for any fixed T of shape S,

$$\zeta_{2,S}(S', y) = |\{T' : d(T, T') = y \text{ and } \mathbf{S}(T') = S'\}|. \tag{15}$$

Therefore, $Z_{L,S_1}(\beta)$ in (14) and eventually $Z_L(\beta)$ can be recursively computed from the $\zeta_{2,S}(S', y)$ values. The rest of the section provides a way to calculate $\zeta_{2,S}(S', x)$ for all values of x and all shapes S, S'. In other words, the goal of the following sections is to determine the joint distribution of the shape of T_2 and $d(T_1, T_2)$ conditional on T_1 (or its shape) when T_2 has a uniform distribution. The C code for computing the joint distribution $\zeta_{2,S}(S', x)$ is available upon request.

5.1 Computing the Joint Distribution of the Robinson-Foulds Metric and Tree Shape

Computing $\zeta_{2,S}(S', x)$ in (15) is required to recursively compute the normalizing function $Z_L(\beta)$. We fix tree T with shape S in the rest of Section 4. Then, $\zeta_{2,S}(S', x)$ is the number of tree topologies with shape S' whose distance from T is x. In this section, we provide several generating functions that are linked to our target frequency $\zeta_{2,S}(S', x)$, simplified as $\zeta_S(S', x)$ here. First, we define $q_S(S', d)$ as

$$\begin{aligned} q_S(S', d) &= |\{T' : \mathbf{S}(T') = S', \\ &\qquad T \text{ and } T' \text{ share } \textit{exactly } d \text{ bipartitions}\}| \\ &= \sum_{\alpha \in \mathcal{A}: |\alpha| = d} |\{T' : \mathbf{S}(T') = S', \\ &\qquad T \text{ and } T' \text{ share } \textit{exactly} \text{ bipartitions } \alpha\}|, \end{aligned}$$

where $\mathcal{A}$ is the set of all possible bipartitions from tree T, and thereby $\zeta_S(S', x)$ can be calculated through

$$q_S(S', d) = \zeta_S(S', N - 3 - d) \text{ for } d = 0, \ldots, N - 3.$$

The generating function for $q_S(S', d)$, defined as

$$Q_{S,S'}(x) = \sum_{d=0}^{N-3} q_S(S', d) x^d,$$

is called the "exact" generating function by [23]. The "at-least" generating function for the number of tree topologies with shape S' is defined as

$$U_{S,S'}(x) = \sum_{d=0}^{N-3} u_S(S', d) x^d, \tag{16}$$

where

$$\begin{aligned} u_S(S', d) = \sum_{\alpha \in \mathcal{A}: |\alpha| = d} &|\{T' : \mathbf{S}(T') = S', T \text{ and } T' \\ &\text{share partitions } \alpha \text{ and possibly others}\}| \end{aligned} \tag{17}$$

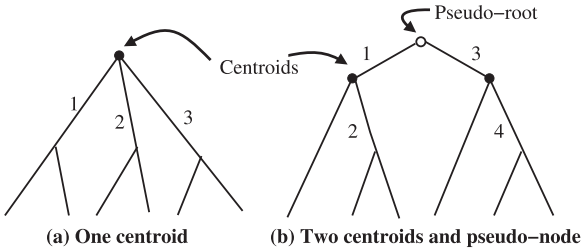

Fig. 4. Two tree shapes in LLC form with (a) one centroid and (b) two centroids and newly introduced pseudoroot. Centroids are indicated with filled circles (•) and the pseudoroot is indicated with an empty circle (○). Internal edge labels are defined using a preorder tree traversal.

and satisfies the following equation by the principle of inclusion and exclusion [23]:

$$Q_{S,S'}(x) = U_{S,S'}(x - 1).$$

Therefore, if we can determine U, then we can determine Q and all $\zeta_S(S', d)$values. The following sections present an algorithm to compute $u_S(S', d)$.

5.2 Definitions and Theorems

To compute $\zeta_{2,S}(S', x)$ in (15) through $u_S(S', d)$ in (17), we first define the terminology used in the following sections. First, we assume that all trees and tree shapes are in their left-light centered (LLC) form, which provides an unique representation and was used to rank all possible tree shapes [24]. Edges and nodes on trees or shapes in LLC form can be labeled in a unique way. To transform an unrooted tree or tree shape into its LLC form, we first determine its centroid(s). A centroid is a node that leads to no more than half of the terminal nodes. Furnas [24] showed that any binary tree has either a single or two centroid nodes, and that these two centroids must be neighbors. If there are two centroids, a new node called the "pseudoroot" is introduced on the edge connecting the two centroids (see Fig. 4) and used to root the tree. The tree is rooted at the unique centroid node otherwise. Then, every edge should lead to an equal number of or fewer terminal nodes than any sister edge on its right, for the tree to be in LLC form. Once trees and shapes are in LLC form, edges are labeled as $1, \ldots, N - 3$ following a preorder tree traversal (root to tip then left to right; see Fig. 4). Note that these edge labels do not correspond to bipartitions, but instead only depend on the tree shape.

We now define edge and node properties. A node is called "cherry" if it is directly connected to two leaves. Edges e and e' in a tree are *symmetric* if we can exchange the labels of e and e' by flipping subtrees at their most recent common ancestor (MRCA) and possibly at some of its descendant nodes while maintaining the tree in LLC form. For example, edges labeled 2 and 4 are symmetric in Fig. 4b. Two nodes are *symmetric* in a tree if their parent edges are symmetric. Two nodes in a tree are *incomparable* if one is not ancestor or descendant of the other. A set of nodes $\{\nu_1, \ldots, \nu_n\}$ in a tree is an *antichain* if the nodes are pairwise incomparable. If an antichain is not a proper subset of any other antichain, then it is a *maximal antichain*.

Let $\mathbf{e}$ be a vector of internal edge labels on tree T. Define the tree forest $T \backslash_m \mathbf{e}$ as the set of subtrees derived by disconnecting edges in $\mathbf{e}$ and by adding labels as described

next. Pseudoterminal nodes are introduced where internal edges in $\mathbf{e}$ are disconnected. The edge indices are used to label these pseudoterminal nodes. That way, the two new terminal nodes from the same original internal edge have *matching* labels. More specifically, the two new nodes obtained from cutting e_i are both labeled m_i.

If the argument of a shape function $\mathbf{S}$ is a tree forest $T\backslash_m\mathbf{e}$, then $\mathbf{S}$ generates a forest from $T\backslash_m\mathbf{e}$ by removing the (pseudo)root and terminal node labels but keeping pseudoterminal node labels. That is,

$$\mathbf{S}(T\backslash_m\mathbf{e}) = \{\mathbf{S}(F_1), \ldots, \mathbf{S}(F_{|\mathbf{e}|+1})\},$$

where the F_i's are the elements of forest $T\backslash_m\mathbf{e}$. Note that if T_1 and T_2 are different topologies but have the same shape, then $\mathbf{S}(T\backslash_m\mathbf{e}) = \mathbf{S}(T'\backslash_m\mathbf{e})$ for any edge vector $\mathbf{e}$ on T (or T').

Similarly, for any label set $\mathcal{L}$ and permutation $\boldsymbol{\sigma}_\mathcal{L}$ of these labels, we consider $\boldsymbol{\sigma}_\mathcal{L}$ as applying to trees by only permuting labels in $\mathcal{L}$. If the argument tree contains pseudoterminal nodes with matching labels, $\boldsymbol{\sigma}_\mathcal{L}$ only permutes the original node labels in $\mathcal{L}$.

Key to our formulas are two equivalence relations between vectors of edges. They are used later when edges are matched to bipartitions across the two trees, to avoid double counting.

Definition 1 (Set equivalence). *Let $\mathbf{e}$ and $\mathbf{e}'$ be vectors of edge labels on tree T. They are set-equivalent if $\mathbf{e}$ can be obtained from $\mathbf{e}'$ by permuting the order of elements in $\mathbf{e}'$. For each set-equivalence class, the representative edge vector $\mathbf{e}$ is defined as the only class member whose elements are arranged with ascending labels. The collection of all set-equivalence class representatives is denoted as $\mathcal{E}(T)$.*

Definition 2 (Subtree-shape equivalence). *Vectors of edge labels $\mathbf{e}$ and $\mathbf{e}'$ are* subtree-shape equivalent *if $\mathbf{S}(T\backslash_m\mathbf{e}) = \mathbf{S}(T\backslash_m\mathbf{e}')$. Note that this relation depends on T through its shape only. $\mathbf{e} = (e_1, \ldots, e_d)$ is defined as the representative of its subtree-shape equivalence class if it satisfies the following conditions:*

1. *$e_1 \le e'$ for any edge e' symmetric with e_1.*
2. *For $d > 1$,*
   a. *subvector $(e_1, \ldots, e_{d-1})$ is the representative of its subtree-shape equivalence class,*
   b. *$e_d \le e'$ for any $e' \notin (e_1, \ldots, e_{d-1})$ symmetric with e_d and that satisfies the following conditions: for each $e_i \in (e_1, \ldots, e_{d-1})$,* i) *$e_d$ and e' are descendants of e_i or* ii) *e_d, e' and e_i are pairwise incomparable and $\mathrm{MRCA}(e_i, e_d) = \mathrm{MRCA}(e_i, e')$.*

If T has a pseudoroot, the 2 edges e_L and e_R connected to that root represent a unique edge on the unrooted tree. Therefore, for this definition, all edges (except e_L and e_R) are considered to be descendant of the left edge e_L.

We prove in the Appendix, which can be found in the online supplemental material, that this definition identifies a unique representative of every equivalence class. $\mathring{\mathcal{E}}(T)$ is defined as the collection of all subtree-shaped equivalent class representatives.

For a vector $\mathbf{e} = (e_1, \ldots, e_h)$ of edges in a tree topology T, $\mathbf{S}(T/\bar{\mathbf{e}})$ is defined as the shape of the consensus tree obtained by contracting all edges but $e_1, \ldots, e_h$ on T, and by giving label c_i to the edge corresponding to e_i. Suppose that trees T and T' have shape S and S', respectively, and consider edge vectors $\mathbf{e}$ on T and $\mathbf{e}'$ on T'. Note that $\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}}')$ holds precisely when there exist tree topologies T_1 and T_1' with shape S and S', respectively, such that the bipartitions defined by $\mathbf{e}$ on T_1 are the same as the bipartitions defined by $\mathbf{e}'$ on T_1'.

For the remainder of this paper, we further fix a tree T' with shape S', and define ν_0 and ν_0' to be the roots of T and T' (once in LLC form). For $d \ge 0$ we define

$$\gamma_S(S', d) = \sum_{\substack{\mathbf{e}\in\mathring{\mathcal{E}}(T),\\ |\mathbf{e}|=d}} \sum_{\substack{\mathbf{e}'\in\mathring{\mathcal{E}}(T'),\\ |\mathbf{e}'|=d}} N(T'\backslash_m\mathbf{e}')\mathbb{I}_{\mathbf{S}(T/\bar{\mathbf{e}})=\mathbf{S}(T'/\bar{\mathbf{e}}')}, \tag{18}$$

where $\mathbb{I}$ is the indicator function, and

$$N(T\backslash_m\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|+1} \#\{F : \exists\boldsymbol{\sigma}_{\mathcal{L}_i} \text{ such that } \boldsymbol{\sigma}_{\mathcal{L}_i}(F) = F_i, F_i \in T\backslash_m\mathbf{e}\}. \tag{19}$$

Each term in the product is the number of trees obtained by permuting the original tip labels $\mathcal{L}_i$ on tree F_i in the forest $T\backslash_m\mathbf{e}$. Note that the $N(T\backslash_m\mathbf{e})$ values are easily calculated recursively (see the Appendix C, which can be found in the online supplemental material). We also define the generating function

$$\Gamma_{S,S'}(x) = \sum_{d=0}^{N-3} \gamma_S(S', d)x^d. \tag{20}$$

The following theorem shows that γ equals u, and hence is the object of interest to eventually compute $\zeta_{2,S}(S', x)$ (proved in Appendix B, which can be found in the online supplemental material).

Theorem 1. *$\Gamma_{S,S'}(x)$ is the "at-least" generating function for the number of tree topologies with shape S'. In other words, $\Gamma_{S,S'}(x) = U_{S,S'}(x)$ and $u_S(S', d) = \gamma_S(S', d)$ in (17).*

We are now ready to define the main object that our algorithm calculates recursively through the tree. Consider a vector V of p antichain nodes in tree T, arranged with ascending labels, and a vector V' of q antichain nodes in tree T'. Further, consider vectors D and K of p nonnegative integers, and a vector M of p 0/1 elements. Similarly, consider vectors D', K', and M' of size q with nonnegative and binary elements. Finally, H is assumed to be a set of pairs of indices, pairing elements of V with elements of V'. The following function generalizes the γ function (18):

$$\begin{aligned} &R(V, V', D, D', K, K', M, M', H) \\ &\quad= \sum_{E=(\mathbf{e}_1,\ldots,\mathbf{e}_p)} \sum_{E'=(\mathbf{e}_1',\ldots,\mathbf{e}_q')} \sum_{G'\in\mathbb{G}_{E',V',D',M',H}} \\ &\quad\quad \in\mathring{\mathcal{M}}_{V,D,K,M} \in\check{\mathcal{M}}_{V',D',K',M'} \\ &\quad\quad \left\{\prod_{i=1}^{q} N\left(T'_{v_i'}\backslash_m\mathbf{e}_i'\right) \times \mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G')\right\}, \end{aligned} \tag{21}$$

where all elements are described in the rest of this section, and such that $\gamma_S(S', d) = \gamma(d)$ is

$$\gamma(d) = \sum_{k=0}^{N} R((\nu_0), (\nu'_0), (d), (d), (k), (k), (0), (0), \emptyset). \quad (22)$$

Given $V = (v_1, \ldots, v_p)$, E, M, and K of size p, $\mathcal{T}_V = (T_{v_1}, \ldots, T_{v_p})$ is a vector of subtrees of T satisfying the following conditions: 1) T_{v_i} contains all descendants of node v_i; and 2) T_{v_i} is rooted at v_i if $m_i = 0$. If $m_i = 1$, the parent edge is included in T_{v_i} as a root edge and is considered as an internal edge. We define $\mathring{\mathcal{M}}_{V,D,K,M} = \prod_{i=1}^{p} \mathring{\mathcal{M}}_{v_i,d_i,k_i,m_i}$ with

$$\mathring{\mathcal{M}}_{v,d,k,m} = \{\mathbf{e} : |\mathbf{e}| = d + m, \mathbf{e} \in \mathring{\mathcal{E}}(T_v), |F_v| = k; \text{ the parent edge of } v \in \mathbf{e} \text{ if } m = 1\},$$

where F_v is the element of $T_v \backslash_m \mathbf{e}$ containing node v and $|F_v|$ is the number of original terminal nodes in F_v, not counting pseudoterminal nodes. Similarly, $\breve{\mathcal{M}}_{V',D',K',M'} = \prod_{j=1}^{q} \breve{\mathcal{M}}_{v'_j,d'_j,k'_j,m'_j}$ and

$$\breve{\mathcal{M}}_{v,d,k,m} = \big\{\mathbf{e} : |\mathbf{e}| = d + m, \mathbf{e} \in \breve{\mathcal{E}}(T_v), |F_v| = k; \text{ the parent edge of } v \in \mathbf{e} \text{ if } m = 1\big\}.$$

We next consider position vectors. They will be used later to merge vectors $(\mathbf{e}_1, \ldots, \mathbf{e}_p) \in \breve{\mathcal{M}}_{V',D',K',M'}$ onto a single vector $\mathbf{e}^*$ of all elements in a specific order. This order can be specified by a positioning $G = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$, to place edge $e_{i,j}$ in position $g_{i,j}$ in $\mathbf{e}^*$, that is, $e^*_{g_{i,j}} = e_{i,j}$.

Definition 3. *Given E, V, M, and D of size p, the set $\mathbb{G}_{E,V,D,M}$ of permissible positionings of edges in E is defined as the set of $G = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$ such that*

1. *for all i, $|\mathbf{g}_i| = d_i + m_i$;*
2. *$\bigcup_{i=1}^{p} \mathbf{g}_i = \{1, \ldots, \sum_{i=1}^{p}(d_i + m_i)\}$;*
3. *for all i, elements in $\mathbf{g}_i$ are arranged in ascending order;*
4. *For any symmetric sibling nodes ν_1 and ν_2 in tree T and any maximal antichain W_1 in subtree T_{ν_1} and maximal antichain W_2 in subtree T_{ν_2}, if $W_1 \subset V$ and $W_2 \subset V$, say $W_1 = \{v_{i_1}, \ldots, v_{i_r}\}$ and $W_2 = \{v_{j_1}, \ldots, v_{j_s}\}$ $(i_r < j_1)$, then*

$$\min\{\mathbf{g}_{i_1}, \ldots, \mathbf{g}_{i_r}\} \le \min\{\mathbf{g}_{j_1}, \ldots, \mathbf{g}_{j_s}\}.$$

If the pseudoroot exists and has two symmetric children ν_1 and ν_2, and if $e_{1,1} = 1$, then it is additionally required that

$$2\text{nd } \min\{\mathbf{g}_{i_1}, \ldots, \mathbf{g}_{i_r}\} \le \min\{\mathbf{g}_{j_1}, \ldots, \mathbf{g}_{j_s}\}.$$

Note that $(\mathbf{e}_1, \ldots, \mathbf{e}_p) \in \mathring{\mathcal{M}}_{V,D,K,M}$ are naturally merged onto a single edge vector by concatenation. In other words, the position of edge $e_{i,r} \in \mathbf{e}_i$ is defined as $\sum_{x=1}^{i-1}(d_i + m_i) + r$. Finally, $\mathbb{G}_{E',V',D',M',H}$ in (21) is defined as the set of position vectors G' in $\mathbb{G}_{E',V',D',M'}$ such that for any pair $(i, j) \in H$, the position $g'_{j,r}$ corresponding to the parent edge $e'_{j,r} \in \mathbf{e}'_j$ of v'_j is different from $\sum_{x<i}(d_x + m_x) + r$.

Given $V = (v_1, \ldots, v_p)$, E, M, K, and G of size p, the consensus tree $C_{\mathcal{T}_V}(E)$ is constructed by grafting the trees $\mathbf{S}(T_{v_i}/\bar{\mathbf{e}}_i)$ at their roots. Edges in $C_{\mathcal{T}_V}(E)$ are named by the positioning vector G. The shape of consensus tree $C_{\mathcal{T}_V}(E) \setminus\setminus K$ is obtained by removing k_i tips directly connected to v_i in $C_{\mathcal{T}_V}(E)$ for all i. Then, $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G')$ in (21) is 1 if the following conditions are satisfied, 0 otherwise:

1. $C_{\mathcal{T}_V}(E) \setminus\setminus K = C_{\mathcal{T}'_{V'}}(E') \setminus\setminus K'$,
2. $k_i = (1 - m_i)|F_{v_i}|$ and $k'_j = (1 - m'_j)|F'_{v'_j}|$.

5.3 Recursive Equations for the Algorithm

We present here the key equations for the recursive derivation of $R(V, V', D, D', K, K', M, M', H)$, which is used to calculate $\gamma_S(S', d)$ through (22). The first theorems initialize the R values, while Theorems 6, 7, and 8 enable the decomposition of R values during the recursion through the tree. More specifically, we start with $V = (\nu_0)$ and $V' = (\nu'_0)$ as in (22). We first use Theorem 6 to compute R through augmented V as replacing ν_0 by its children. Theorem 6 is repeatedly applied to the leftmost node in V satisfying the conditions in Theorem 6 until any $m_i = 1$. We then move on Theorem 7 to augment V' as replacing ν'_0 by its children. Similarly, Theorem 7 is repeatedly applied to the leftmost node in V' satisfying the conditions in Theorem 7 until any newly introduced $m'_j = 1$. Then, Theorem 8 is applied to factorize R. This process is repeated until the value of R is obtained by Theorems 2-5. All proofs are found in Appendix D, which can be found in the online supplemental material.

Theorem 2. *If $k_i = |T_{v_i}|$ and $k'_j = |T'_{v'_j}|$ for all i, j, then*

$$R(V, V', \mathbf{0}, \mathbf{0}, K, K', \mathbf{0}, \mathbf{0}, H) = \prod_j N\big(T'_{v'_j} \backslash_m \emptyset\big),$$

otherwise $R(V, V', \mathbf{0}, \mathbf{0}, K, K', \mathbf{0}, \mathbf{0}, H) = 0$.

Theorem 3. *$R = R(V, V', D, D', K, K', M, M', H) = 0$ if $\sum_i(d_i + m_i) \neq \sum_j(d'_j + m'_j)$, and if V and V' do not contain both children of the pseudoroot. Generally, $R = 0$ if $\Delta \neq \Delta'$, where $\Delta = \sum_i(d_i + m_i)$ if V does not contain both children of the pseudoroot, $\Delta = d_1 + d_2 + m_1$ otherwise. Δ' is defined similarly.*

Theorem 4. *$R = 0$ if there exists an index i satisfying at least one of the following conditions:*

1. *v_i is a cherry, $d_i > 0$;*
2. *$d_i > 0, k_i > |T_{v_i}| - 2$;*
3. *$d_i = m_i = 0, k_i \neq |T_{v_i}|$;*
4. *$m_i = 0$, $k_i + 1$ or more tips are directly connected to v_i; or*
5. *$d_i > |T_{v_i}| - 2$.*

Similarly, if there exists an index j satisfying at least one of the analogous conditions in terms of V', D', K', M', then $R = 0$.

Theorem 5. *Consider trees T_ν and $T'_{\nu'}$ on the same number of taxa with d internal nodes. If they have the same shape, then $R((\nu), (\nu'), (d), (d), (k), (k), (0), (0), \emptyset) = 1$ if k is the number of tips directly connected to ν; 0 otherwise. If T_ν and $T'_{\nu'}$ have different shapes, then $R((\nu), (\nu'), (d), (d'), (k), (k'), (0), (0), \emptyset) = 0$ for all k and k'.*

Theorems 6 and 7 decompose R into a sum of R values, where one node in V or V' is replaced by its children.

Theorem 6 (Formula dismantling a node in T). *Consider $v_x \in V$ such that $m_x = 0$, $d_x \geq 1$, and v_x has $r(\leq 3)$ internal nodes and k_0 tips as children. Let $w_1, \ldots, w_r$ be the r internal node children of v_x. We define the following sets:*

$$\mathcal{C} = \Big\{ \tilde{\mathbf{d}}, \tilde{\mathbf{m}} \mid \tilde{\mathbf{d}} = (\tilde{d}_x, \ldots, \tilde{d}_{x+r-1}), \tilde{\mathbf{m}} = (\tilde{m}_x, \ldots, \tilde{m}_{x+r-1}), \sum_{i=x}^{x+r-1} (\tilde{d}_i + \tilde{m}_i) = d_x; \tilde{m}_2 = 1 \text{ and } \tilde{d}_1 + \tilde{m}_1 + \tilde{d}_2 = d_x \text{ if } \nu_x \text{ is the pseudoroot and } \tilde{m}_1 = 1 \Big\},$$

$$\mathcal{K}_{\tilde{\mathbf{m}}} = \Big\{ \tilde{\mathbf{k}} \mid \tilde{\mathbf{k}} = \{\tilde{k}_x, \ldots, \tilde{k}_{x+r-1}\}, \sum_{i=x}^{x+r-1} \tilde{k}_i^{(1-\tilde{m}_i)} = k_x - k_0;\ \tilde{k}_i = 0 \text{ if } \tilde{m}_i = 1 \Big\}.$$

Then,

$$R(V, V', D, D', K, K', M, M', H) = \sum_{\tilde{\mathbf{d}}, \tilde{\mathbf{m}} \in \mathcal{C}} \sum_{\tilde{\mathbf{k}} \in \mathcal{K}_{\tilde{\mathbf{m}}}} R(\tilde{V}, V', \tilde{D}, D', \tilde{K}, K', \tilde{M}, M', \tilde{H}),$$

where $\tilde{V}$ is similar to V except that v_x is replaced by its children. More specifically, $\tilde{v}_i = v_i$, for $i \leq x-1$; w_{i-x+1}, for $x \leq i \leq x+r-1$; v_{i-r+1}, for $i \geq x+r$. $\tilde{D}'$, $\tilde{K}'$ and $\tilde{M}'$ are defined similarly. By definition, $\tilde{H}$ contains (i,j) if $(i,j) \in H$, and $i \leq x-1$; $(i+r-1, j)$ if $(i,j) \in H$ and $i \geq x+1$. Note that $|H| = |\tilde{H}|$.

Theorem 7 (Formula dismantling a node in T'). *Consider $v'_x \in V'$ such that $m'_x = 0$, $d'_x \geq 1$ and v'_x has $r(\leq 3)$ internal nodes and k'_0 tips as children. Let $w'_1, \ldots, w'_r$ ($r \leq 3$) be the r internal node children of v'_x. We define the following sets:*

$$\mathcal{C} = \Big\{ \tilde{\mathbf{d}}', \tilde{\mathbf{m}}' \mid \tilde{\mathbf{d}}' = (\tilde{d}'_x, \ldots, \tilde{d}'_{x+r-1}), \tilde{\mathbf{m}}' = (\tilde{m}'_x, \ldots, \tilde{m}'_{x+r-1}), \sum_{i=x}^{x+r-1} (\tilde{d}'_i + \tilde{m}'_i) = d'_x; \tilde{d}'_j + \tilde{m}'_j = 0 \text{ if } \tilde{d}'_{j-1} + \tilde{m}'_{j-1} = 0, \text{ and if } \tilde{w}'_{j-1} \text{ and } \tilde{w}'_j \text{ are symmetric}; \tilde{m}'_2 = 1 \text{ and } \tilde{d}'_1 + \tilde{m}'_1 + \tilde{d}'_2 = d'_x \text{ if } v'_x \text{ is the pseudoroot and if } \tilde{m}'_1 = 1 \Big\},$$

$$\mathcal{K}_{\tilde{\mathbf{m}}'} = \Big\{ \tilde{\mathbf{k}}' \mid \tilde{\mathbf{k}}' = \{\tilde{k}'_x, \ldots, \tilde{k}'_{x+r-1}\}, \sum_{i=x}^{x+r-1} \tilde{k}'_i{}^{(1-\tilde{m}'_i)} = k'_x - k'_0;\ \tilde{k}'_i = 0 \text{ if } \tilde{m}'_i = 1 \Big\},$$

$$\mathrm{sym}_{F_{v'_x}}(v'_x) = \begin{cases} 1 & \text{if none of } \mathbf{S}\ (F_{w'_i}) \text{ are the same,} \\ 2 & \text{if exactly 2 of } \mathbf{S}\ (F_{w'_i}) \text{ are same,} \\ 3 & \text{if } r = 3 \text{ and all 3 } \mathbf{S}\ (F_{w'_i}) \text{ are same.} \end{cases}$$

Then,

$$R(V, V', D, D', K, K', M, M', H) = \sum_{\tilde{\mathbf{d}}', \tilde{\mathbf{m}}' \in \mathcal{C}} \sum_{\tilde{\mathbf{k}}' \in \mathcal{K}_{\tilde{\mathbf{m}}'}} \Big\{ R(V, \tilde{V}', D, \tilde{D}', K, \tilde{K}', M, \tilde{M}', \tilde{H}) \times k'_x! \Big/ \Big(\mathrm{sym}_{F_{v'_x}}(v'_x)! \prod_{i=x}^{x+r-1} (\tilde{k}'_i!)^{(1-\tilde{m}'_i)} \Big) \Big\},$$

where $\tilde{V}'$ is similar to V' except that v'_x is replaced by its children, as defined by $\tilde{v}'_i = v'_i$, for $i \leq x-1$; w'_{i-x+1}, for $x \leq i \leq x+r-1$; v'_{i-r+1}, for $i \geq x+r$. $\tilde{D}'$, $\tilde{K}'$ and $\tilde{M}'$ are defined similarly. By definition, $\tilde{H}$ contains (i,j) if $(i,j) \in H$ and $j \leq x-1$; $(i, j+r-1)$ if $(i,j) \in H$ and $j \geq x+1$. Note that $|H| = |\tilde{H}|$.

Theorem 8 (Factorization formula). *Consider $v_x \in V$ such that $m_x = 1$, and assume that the partial sum $\sum_{i=1}^{x-1}(d_i + m_i) = 0$. Define $\mathcal{Z}$ as the index set of nodes v'_j in V' that can be paired with v_x to define the same bipartition, as specified below. If V and V' contain all internal node children of roots ν_0 and ν'_0, $\mathcal{Z} = \{j \mid (x,j) \notin H, m'_j = 1, k'_j = 0, d_x = d'_j, |T_{v_x}| = |T'_{v'_j}|, v'_j$ has no symmetric sibling in $(v'_1, \ldots, v'_{j-1})\}$. More generally,*

$$\mathcal{Z} = \big\{ j \mid (x,j) \notin H, m'_j = 1, k'_j = 0, d_x = d'_j, |T_{v_x}| = |T'_{v'_x}|; \forall \nu' \leq v'_j, \text{ symmetric sibling of either } v'_j \text{ or its ancestor and } \forall W \text{ maximal antichain in } T'_{\nu'},\ W \not\subseteq V' \big\}.$$

Let H^* be the augmented constraint set $H^* = H \cup \{(x,j) : j \in \mathcal{Z}\}$. Then,

$$R(V, V', D, D', K, K', M, M', H) = R(V, V', D, D', K, K', M, M', H^*) + \sum_{j \in \mathcal{Z}} \Bigg[\sum_{k=0}^{|T_{v_x}|} R\big((v_x), (v'_j), (d_x), (d_x), (k), (k), (0), (0), \emptyset\big) \times R(V_{-x}, V'_{-j}, D_{-x}, D'_{-j}, K_{-x}, K'_{-j}, M_{-x}, M'_{-j}, \tilde{H}) \Bigg],$$

where V_{-x} contains all elements in V except for v_x, and we similarly define V'_{-j} and so on. We also define $\tilde{H} = \{(\tilde{i}, \tilde{j})$: $(i,l) \in H$, where $i = \tilde{i}$ if $\tilde{i} < x$; $\tilde{i}+1$ if $\tilde{i} \geq x$, and $l = \tilde{j}$ if $\tilde{j} < j$; $\tilde{j}+1$ if $\tilde{j} \geq j\}$.

6 Approximations to the Normalizing Function

Although we can calculate the exact values of the normalizing function $Z_L(\beta)$ through (14), (15), and the algorithm outlined in Section 5, its computation is usually too expensive to be repeated at each iteration of an MCMC algorithm. Therefore, we propose two approximations to this normalizing function.

6.1 Large-L Normal Approximation

Recall that L denotes the number of segments and N denotes the number of taxa. We can write

$$Z_L(\beta) = Z_1 + \zeta_L(1)e^{-\beta} + \sum_{x=2}^{D_L} \zeta_L(x)e^{-\beta x}, \tag{23}$$

where $D_L = (L-1)(N-3)$, $Z_1 = (2N-5)!!$ was defined previously,

$$\zeta_L(x) = \#\left\{(T_1, \ldots, T_L) : \sum_{l=1}^{L-1} d(T_l, T_{l+1}) = x\right\},$$

and $\zeta_L(1)$ is easily shown to be $\zeta_L(1) = (L-1)\,2(N-3)Z_1$. The sum in (23) is approximated using the following central limit theorem.

Theorem 9. *Consider independent, uniformly distributed unrooted N-taxon trees $(T_i)_{i\geq 1}$. Let $S_L = \sum_{l=1}^{L-1} d(T_l, T_{l+1})$. Then, $P(S_L \leq 1)$ goes to 0 as L goes to infinity and both $(S_L - \mu_L)/\sigma_L$ and*

$$(S_L - \mu_L)/\sigma_L \, \mathbb{I}(S_L \geq 2) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad L \to \infty,$$

where $\mu_L = (L-1)E(d(T_1, T_2))$ and

$$\sigma_L^2 = (L-1)\{\mathrm{var}(d(T_1,T_2) + \mathrm{cov}(d(T_1,T_2), d(T_2,T_3))\}.$$

The proof (Appendix F.1, which can be found in the online supplemental material) rests on the weak dependence of the sequence $(d(T_i, T_{i+1}))_{i\geq 1}$. The second part results in a normal approximation for the sum in (23), from which we obtain the *normal approximation* $Z_L(\beta) \approx \hat{Z}_{(1)}$:

$$\begin{aligned}\hat{Z}_{(1)} = {} & Z_1 + (L-1)\zeta_2(1)e^{-\beta} + \left\{Z_1^L - Z_1 - (L-1)\zeta_2(1)\right\} \\ & \times \left[\Phi\left(D_L + .5; \mu_L - \beta\sigma_L^2, \sigma_L^2\right) - \Phi\left(2 - .5; \mu_L - \beta\sigma_L^2, \sigma_L^2\right)\right] \\ & \times \exp\left[-\beta\mu_L + \frac{\beta^2\sigma_L^2}{2}\right],\end{aligned} \tag{24}$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the cumulative distribution function of the normal distribution with mean μ and variance σ^2.

6.2 Independence Approximation

Our second approximation is simply obtained by ignoring the dependence between distances $d(T_{l-1}, T_l)$ and $d(T_l, T_{l+1})$, for $l = 1, \ldots, L-1$. We can write

$$\begin{aligned} Z_L(\beta) &= Z_1^L E\left(e^{-\beta\sum_{l=1}^{L-1} d(T_l, T_{l+1})}\right) \approx \hat{Z}_{(2)} \\ \hat{Z}_{(2)} &= Z_1^L E\left(e^{-\beta d(T_1,T_2)}\right)^{L-1}. \end{aligned} \tag{25}$$

Note that $d(T_{l-1}, T_l)$ are indeed independent when there is only one possible tree shape, i.e., when $N \leq 5$. We prove in the appendix, which can be found in the online supplemental material, that for all N, L, and all β,

$$Z_{(2)}(\beta) \leq Z_L(\beta).$$

6.3 Accuracy of Approximations

The proposed approximations (24)-(25) to the normalizing function are compared with the true value $Z_L(\beta)$ for various values of β, on trees with 5 taxa and 10 taxa, and when the length of the alignment varies from 10 to 1,000 (see Fig. 5). The normalizing function $Z_L(\beta)$ quickly drops to Z_1 as β grows. The extent of the decline is more profound with more segments or more taxa. Since distances between tree topologies $\{d(T_i, T_{i+1}) : i \geq 1\}$ are independent when there is only one tree shape, the independence approximation $\hat{Z}_{(2)}$ in (25) is exact for $N \leq 5$. The large-L normal approximation $\hat{Z}_{(1)}$ in (24) is a good approximation except for $\beta \in (1.5, 5)$ approximately. Note that the distribution of the sum of tree distances S_L is skewed left because its mean μ_L is approximately $(L-1)(N-3-1/8)$ [25], which is very close to its maximum value $(L-1)(N-3)$. The symmetric normal approximation to the distribution of S_L is thus expected to underestimate the true probabilities at small values. These small values of x are given more weight by the exponential term in (23), so $\hat{Z}_{(1)}$ is expected to underestimate the true Z_L. This is indeed what we observe in Fig. 5. The proposed approximations showed similar accuracy on 10 taxa. In particular, the independence approximation $Z_{(2)}$ is still very close to the true normalizing function.

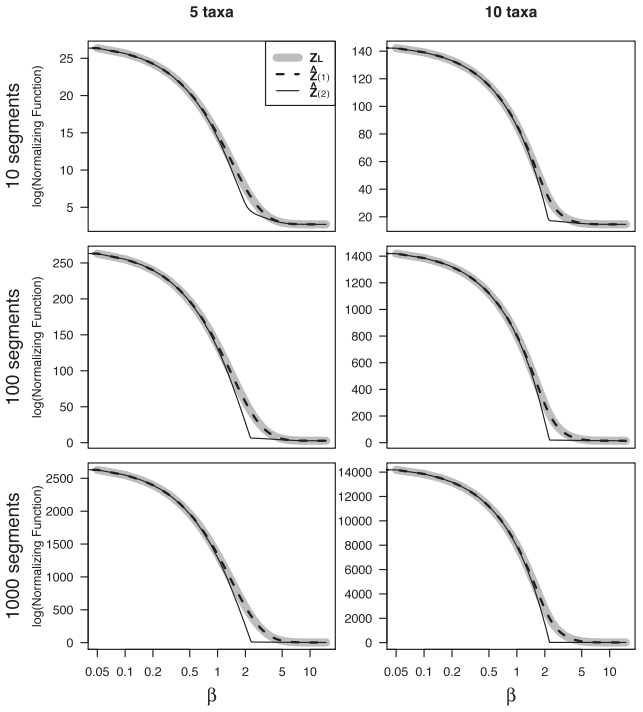

Fig. 5. Accuracy of approximations to the normalizing function $Z_L(\beta)$ on 5 taxa and 10 taxa, when the number of segments is $L = 10$, $L = 100$, or $L = 1{,}000$. The true normalizing function $Z_L(\beta)$ in the thick gray line is compared with two approximations: the normal approximation $\hat{Z}_{(1)}$ (—) and the independence approximation $\hat{Z}_{(2)}$ (- -).

Fig. 6 shows the impact of using $\hat{Z}_{(2)}$ instead of the true normalizing function in terms of hyperprior densities. Although the hyperprior actually used on β has a slightly higher density than the assumed hyperprior on small β values when $\lambda = 0.01$ (see Fig. 6d), the difference is small enough to be ignored. Overall, the hyperprior actually used is very close to the assumed hyperprior.

7 Discussion

In this work, we first show empirical evidence that the phylogenetic trees of neighboring genomic regions are correlated, in the sense that they are more similar than

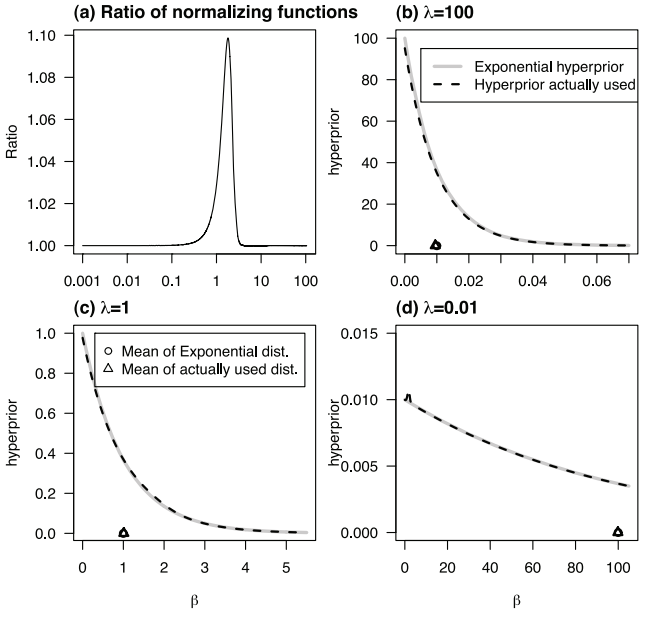

Fig. 6. Impact of using the independence approximation on 10 taxa with $L = 10$ segments. (a) Ratio of the true normalizing function to independence approximation $Z_L/\hat{Z}_{(2)}$. (b)-(d) The thick gray line indicates the exponential distribution $\mathcal{E}(\lambda)$, with mean indicated by a circle (○). The hyperprior density actually used is indicated with a dotted line (- -) with mean indicated by a triangle (Δ) when $\lambda = 100$, 1, and 0.01. Note that the axis for β in (a) is on the log scale.

expected by chance. In *Escherichia* and *Shigella* genomes, the correlation between neighboring trees was shown to span across distances of about 2 kb. This is in support of methods that go beyond detecting gene tree discordance, toward the analysis of the dissimilarity of discordant gene trees. Leigh et al. [26] take this approach to cluster predefined genes based on the similarity of their gene trees. We focus here on long alignments for which recombination-free loci are not predefined. We consider a Bayesian approach to simultaneously detect recombination breakpoints and phylogenetic trees based on a Gibbs prior distribution, to account for the correlation between phylogenetic trees at neighboring loci. The behavior of the Gibbs distribution is controlled by a parameter β which scales with the inverse recombination rate per segment. The dissimilarity between tree topologies is measured by the RF distance. We show how to calculate the normalizing function of the Gibbs distribution exactly, and propose fast and accurate approximations. We, thus, provide the mathematical foundation for the future implementation of Gibbs-distribution-based methods to simultaneously infer recombination breakpoints and the phylogenetic history of individual recombination blocks.

The RF distance is not the ideal dissimilarity measure to quantify gene tree discordance due to recombination, because one recombination event is expected to cause the trees on the left and right side of the breakpoint to disagree by one SPR rearrangement [27]. Therefore, we use the RF distance here to measure the presence of recombination and detect breakpoints, but *not* as a measure of the amount of recombination. Computing the SPR distance between two trees is computationally heavy unfortunately [28], requiring approximations like in biomc2. On the other hand, computing the RF distance is fast. Additionally, there is a wide lack of tools to study the normalizing function of the Gibbs distribution based on the SPR distance. For instance, the distribution of the SPR distance between a random tree and a fixed tree, as a function of the shape of the fixed tree, is unknown. The diameter of the SPR metric space is bounded above by $N - 3$ and below by $N/2 - o(N)$, where N is the number of taxa [29].

The core of the present work is an algorithm to calculate the joint distribution of the shape of a random tree and its RF distance to another fixed tree (code available upon request). This joint distribution completely determines the Gibbs distribution for the trees at two neighboring segments. It is then used to recursively calculate the normalizing function of the Gibbs distribution on any number of segments. The core algorithm to calculate the joint distribution of tree shape and RF distance builds on Bryant and Steel [16], who provide the distribution of the RF distance only, based on the shape of the fixed tree. Their algorithm recursively calculates a quantity analogous to $R(v, d, k)$, where v is the root of a subtree and d relates to the RF distance between two subtrees. To also track the second tree shape, our algorithm needs to condition the R value on many other variables, making the algorithm much more complicated. We had to add arguments such as v', d', and k' for the other tree. To specify the shared bipartitions between two trees, additional arguments m, m', and H were introduced to avoid matching some pairs of edges multiple times.

When both trees are fixed, the complexity of the algorithm calculating $\zeta_S(S', d)$ for all RF distance values (d) depends on the shapes S and S' of the trees. If both are caterpillar trees whose shape is the most asymmetric shape a tree can have [30], then the algorithm runs in a polynomial time. If both trees are fully symmetric, then the algorithm has an exponential time complexity (see Appendix E, which can be found in the online supplemental material).

Two approximations to the normalizing function were proposed, and our "independence" approximation showed excellent performance. Both approximations require the marginal distribution of the RF distance between a random tree and a fixed tree, whose shape is known but arbitrary. This can be calculated in polynomial time [16]. These practical considerations are important, because the normalizing function needs to be evaluated each time a new prior inverse recombination rate β is proposed during Bayesian inference with Markov Chain Monte Carlo. Bryant and Steel [16] also provide two approximations to their normalizing function (5) when β is either small or large. Their approximations cut down computing time substantially, as they do not require the distribution of the RF distance. Our attempts to use their small β and large β approximations to speed up our independence approximation resulted in large errors unfortunately, and increasingly more so as more segments were considered. Instead, our independence approximation provides a substantial computing time reduction without misleading the MCMC results.

ACKNOWLEDGMENTS

The authors would like to thank Guy Plunkett III for providing alignments of 33 *Escherichia* genomes and eight *Shigella* genomes. The authors would also like to thank Aaron Darling for technical assistance with the alignments. This work was funded in part by US National Science Foundation awards 0936214 and 0949121.

REFERENCES

[1] J. Hein, "A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination," *J. Molecular Evolution,* vol. 36, no. 4, pp. 396-405, 1993.

[2] C. Ané, "Detecting Phylogenetic Breakpoints and Discordance from Genome-Wide Alignments for Species Tree Reconstruction," *Genome Biology and Evolution,* vol. 3, pp. 246-258, Jan. 2011.

[3] N. Grassly and E. Holmes, "A Likelihood Method for the Detection of Selection and Recombination Using Nucleotide Sequences," *Molecular Biology and Evolution,* vol. 14, no. 3, pp. 239-247, 1997.

[4] X. Didelot, D. Lawson, A. Darling, and D. Falush, "Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences," *Genetics,* vol. 186, no. 4, pp. 1435-1449, 2010.

[5] D. Husmeier and G. McGuire, "Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo," *Molecular Biology and Evolution,* vol. 20, no. 3, pp. 315-337, 2003.

[6] W.P. Lehrach and D. Husmeier, "Segmenting Bacterial and Viral DNA Sequence Alignments with a Trans-Dimensional Phylogenetic Factorial Hidden Markov Model," *J. Royal Statistical Soc. Series C,* vol. 58, no. 3, pp. 307-327, 2009.

[7] A. Webb, J.M. Hancock, and C.C. Holmes, "Phylogenetic Inference Under Recombination Using Bayesian Stochastic Topology Selection," *Bioinformatics,* vol. 25, no. 2, pp. 197-203, 2009.

[8] B. Boussau, L. Guéguen, and M. Gouy, "A Mixture Model and a Hidden Markov Model to Simultaneously Detect Recombination Breakpoints and Reconstruct Phylogenies," *Evolutionary Bioinformatics,* vol. 5, pp. 67-79, 2009.

[9] V.N. Minin, K.S. Dorman, F. Fang, and M.A. Suchard, "Dual Multiple Change-Point Model Leads to More Accurate Recombination Detection," *Bioinformatics,* vol. 21, no. 13, pp. 3034-3042, 2005.

[10] M.A. Suchard, R.E. Weiss, K.S. Dorman, and J.S. Sinsheimer, "Oh Brother, Where Art Thou? A Bayes Factor Test for Recombination with Uncertain Heritage," *Systematic Biology,* vol. 51, no. 5, pp. 715-728, 2002.

[11] F. Fang, J. Ding, V.N. Minin, M.A. Suchard, and K.S. Dorman, "cBrother: Relaxing Parental Tree Assumptions for Bayesian Recombination Detection," *Bioinformatics,* vol. 23, no. 4, pp. 507-508, 2007.

[12] E.W. Bloomquist, K.S. Dorman, and M.A. Suchard, "StepBrothers: Inferring Partially Shared Ancestries among Recombinant Viral Sequences," *Biostatistics,* vol. 10, no. 1, pp. 106-120, 2009.

[13] L. de Oliveira Martins, E. Leal, and H. Kishino, "Phylogenetic Detection of Recombination with a Bayesian Prior on the Distance between Trees," *PLoS ONE,* vol. 3, no. 7, article e2651, 2008.

[14] M. Steel and A. Rodrigo, "Maximum Likelihood Supertrees," *Systematic Biology,* vol. 57, no. 2, pp. 243-250, 2008.

[15] D.F. Robinson and L.R. Foulds, "Comparison of Phylogenetic Trees," *Math. Biosciences,* vol. 53, nos. 1/2, pp. 131-147, Feb. 1981.

[16] D. Bryant and M. Steel, "Computing the Distribution of a Tree Metric," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 6, no. 3, pp. 420-426, July-Sept. 2009.

[17] B.A. Cipra, "An Introduction to the Ising Model," *Am. Math. Monthly,* vol. 94, pp. 937-959, Dec. 1987.

[18] A.E. Darling, B. Mau, and N.T. Perna, "ProgressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement," *PLoS ONE,* vol. 5, no. 6, article e11147, 2010.

[19] J.P. Huelsenbeck and F. Ronquist, "MRBAYES: Bayesian Inference of Phylogenetic Trees," *Bioinformatics,* vol. 17, no. 8, pp. 754-755, 2001.

[20] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA," *J. Molecular Evolution,* vol. 22, no. 2, pp. 160-174, 1985.

[21] L. de Oliveira Martins and H. Kishino, "Distribution of Distances between Topologies and Its Effect on Detection of Phylogenetic Recombination," *Annals of the Inst. of Statistical Math.,* vol. 62, pp. 145-159, 2010.

[22] C.J. Preston, "Generalized Gibbs States and Markov Random Fields," *Advances in Applied Probability,* vol. 5, no. 2, pp. 242-261, 1973.

[23] I.P. Goulden and D.M. Jackson, *Combinatorial Enumeration.* Dover Publications, 2004.

[24] G. Furnas, "The Generation of Random, Binary Unordered Trees," *J. Classification,* vol. 1, no. 1, pp. 187-233, Dec. 1984.

[25] M.A. Steel and D. Penny, "Distributions of Tree Comparison Metrics—Some New Results," *Systematic Biology,* vol. 42, no. 2, pp. 126-141, 1993.

[26] J.W. Leigh, K. Schliep, P. Lopez, and E. Bapteste, "Let Them Fall Where They May: Congruence Analysis in Massive Phylogenetically Messy Data Sets," *Molecular Biology and Evolution,* vol. 28, no. 10, pp. 2773-2785, 2011.

[27] Y. Song and J. Hein, "Constructing Minimal Ancestral Recombination Graphs," *J. Computational Biology,* vol. 12, no. 2, pp. 147-69, 2005.

[28] M. Bordewich and C. Semple, "On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance," *Annals of Combinatorics.,* vol. 8, no. 4, pp. 409-423, Jan. 2005.

[29] B.L. Allen and S. Mike, "Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees," *Annals of Combinatorics,* vol. 5, pp. 1-15, 2001.

[30] C. Semple and M. Steel, *Phylogenetics.* Oxford Univ. Press, 2003.

Yujin Chung received the PhD degree in statistics from the University of Wisconsin-Madison in 2012. She is currently working as a postdoctoral associate in the Center for Computational Genetics and Genomics (CCGG), Department of Biology at Temple University.

Nicole T. Perna received the PhD degree in genetics from the University of New Hampshire, Durham, in 1996. She is currently a professor of genetics at the University of Wisconsin-Madison. Her research interests include microbial genome evolution and molecular evolution of complex traits like virulence and host range.

Cécile Ané received the PhD degree in probability from the University of Toulouse, France, in 2000. She is currently an associate professor in both the Department of Statistics and the Department of Botany at the University of Wisconsin-Madison. Her research interests include statistical methods for molecular evolution and trait evolution.