Inferring Local Genealogies on Closely Related Genomes

Ryan A. Leo Elworth and Luay Nakhleh(✉)

Department of Computer Science, Rice University,
6100 Main Street, Houston, TX 77005, USA
{ral8,nakhleh}@rice.edu

Abstract. The relationship between the evolution of a set of genomes and of individual loci therein could be very complex. For example, in eukaryotic species, meiotic recombination combined with effects of random genetic drift result in loci whose genealogies differ from each other as well as from the phylogeny of the species or populations—a phenomenon known as incomplete lineage sorting, or ILS. The most common practice for inferring local genealogies of individual loci is to slide a fixed-width window across an alignment of the genomes, and infer a phylogenetic tree from the sequence alignment of each window. However, at the evolutionary scale where ILS is extensive, it is often the case that the phylogenetic signal within each window is too low to infer an accurate local genealogy. In this paper, we propose a hidden Markov model (HMM) based method for inferring local genealogies conditional on a known species tree. The method borrows ideas from the work on coalescent HMMs, yet approximates the model parameterization to focus on computationally efficient inference of local genealogies, rather than on obtaining detailed model parameters. We also show how the method is extended to cases that involve hybridization in addition to recombination and ILS. We demonstrate the performance of our method on synthetic data and one empirical data set, and compare it to the sliding-window approach that is, arguably, the most commonly used technique for inferring local genealogies.

1 Introduction

The coalescent is a stochastic process that considers a sample of alleles from a single population and its genealogical history under certain assumptions such as those made by the Wright-Fisher (W-F) model [21]. Under the coalescent, the genealogy of a sample takes the shape of a tree that links all extant samples to their most recent common ancestor (MRCA); Fig. 1(a). This model was extended to model multiple populations related by a tree structure, giving rise to the multispecies coalescent [6].

Part of this research was conducted while RALE was funded by a training fellowship from the National Library of Medicine (Award T15LM007093; PD Lydia E. Kavraki). Furthermore, research was funded in part by NSF grants CCF-1541979 and DMS-1547433.

J. Meidanis and L. Nakhleh (Eds.): RECOMB CG 2017, LNBI 10562, pp. 213–231, 2017.
DOI: 10.1007/978-3-319-67979-2_12

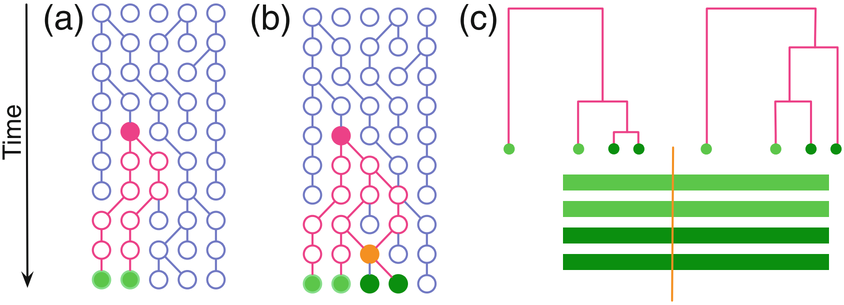

Fig. 1. The coalescent and recombination. (a) A single population of five individuals over 10 generations. The MRCA (top solid circle) of a sample of two extant individuals (two solid circles at the bottom) is shown. (b) The coalescent with recombination. The genealogy of a sample of four extant individuals (four solid circles at the bottom) from their MRCA (top solid circle) is shown. The recombination node (solid circle in the second layer from the bottom) results in an ancestral recombination graph (ARG) for the genealogy. (c) The multispecies coalescent with recombination viewed as a process along the genomes: The local genealogy changes as a recombination breakpoint (the vertical line) is encountered along the multiple sequence alignment (the four horizontal bars).

When recombination occurs, the genealogy takes the shape of a rooted, directed, acyclic graph known as an ancestral recombination graph, or ARG; Fig. 1(b). In this case, "walking" across the genome alignment, the local genealogy observed at each site could change as a recombination breakpoint is encountered; Fig. 1(c). Therefore, while the genomes evolve within the branches of a species tree (barring reticulation), individual genomic regions, or loci, across the genomes could have local genealogies that differ in shape from each other and from that of the species tree. Elucidating the species tree itself as well as the individual local genealogies is of great interest [13]. Sliding a fixed-size window across the genome alignment and building a tree on the sequence alignment pertaining to each window is the most common practice. For example, this is the approach employed in analyzing the genomes of *Staphylococcus aureus* [38], butterflies [42], and mosquitos [12], to demonstrate the extent of ILS. However, for genomes that are evolutionarily close enough for incomplete lineage sorting (ILS) to occur, the signal within a window could be problematic: for small window sizes, the signal could be weak to infer a tree with well supported branches, and for large window sizes, recombination would be extensive within the window rendering the assumption of a treelike genealogy incorrect. For example, in the work of [12], the window size was set to 50 kb, and, in a more recent study of sparrow genomes [10], the window size was even set to 100 kb. These are window sizes large enough to contain extensive recombination breakpoints within each window. There has already been debate in the community as to the size of genomic regions that would be recombination-free (or almost recombination-free) and could truly have a single underlying evolutionary tree [9,36]. In particular, it is claimed that a recombination-free locus is about 12bp or shorter.

In this work, we present a hidden Markov model (HMM) based method for inferring local genealogies from genomic data conditional on the knowledge of a species phylogeny. Wiuf and Hein [39] observed that to sample an ARG and its parameters from genomic sequences, a process can be followed that moves along the sequences and modifies the local genealogy as recombination breakpoints are encountered; This observation allowed for introducing the *sequentially Markov coalescent* [25] and *coalescent hidden Markov model* (coalescent HMMs) [16]. These models are very detailed and their main utility is to obtain accurate estimates of evolutionary parameters, such as ancestral population sizes and recombination rates (in most applications of these models, the local genealogies are treated as nuisance parameters). Here, we make use of the observation of [39] but depart from the detailed and computationally intensive work of [16,25], as we focus on inferring local genealogies, rather than on estimating evolutionary parameters. This departure allows us to make the computation faster for obtaining good estimates of the local genealogies. In particular, we "integrate out" the branch lengths of local genealogies in computing emission probabilities and use the Normalized Rooted Branch Score, or NRBS, distance [15] to approximate the transition probabilities. Running standard, efficient algorithms on the resulting HMM and the genomic sequence data provides estimates of local genealogies along with their posterior support. We demonstrate the utility of our model on both simulated and biological data sets and discuss its limitations, in addition to comparing its performance against the traditional window-sliding approach. Furthermore, we show how our model can be extended, with a slight modification, to account for cases of gene flow and show results of corresponding simulations with reticulate evolution.

Other earlier methods exist for elucidating local genealogies; e.g., [2,4,22, 26,29,34,40]. All this work notwithstanding, high profile analyses continue to employ the simple sliding-window approach [10,12,42]. The focus of our work here is to highlight the performance of a sliding-window approach on data sets that consist of or include closely related genomes, and demonstrate the gains one can obtain by incorporating spatial dependencies along the genome in improving the accuracy of obtained local genealogies.

2 Model and Methods

Let $\mathcal{S}$ be a set of aligned genomes on taxa $\mathcal{X}$ and $\mathcal{S}_i$ be the i^{th} site in the alignment. Given a species tree Ψ on $\mathcal{X}$, we denote by $H(\Psi)$ the set of all possible coalescent histories [5]; Fig. 2. Every $\mathcal{S}_i$ has evolved down a local coalescent, or gene, history in $H(\Psi)$. We define a hidden Markov model (HMM) $\mathcal{M}$ whose states are $q_1, q_2, \ldots, q_{|H(\Psi)|}$, such that state q_i corresponds to coalescent history $g_i \in H(\Psi)$; see Fig. 2. Each internal node of the species tree Ψ has a divergence time (in years) associated with it, such that τ_v is the divergence time of node v (all the leaves are assumed to be at time 0). Furthermore, with every branch

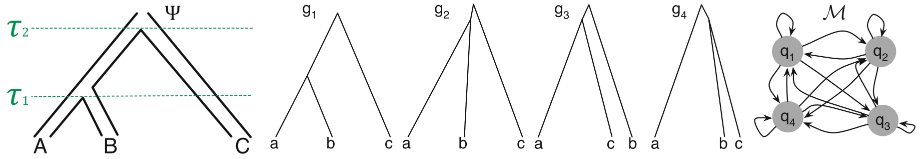

Fig. 2. A species tree Ψ on three taxa A, B, and C, and the four possible coalescent (or, gene) histories on three taxa g_1, g_2, g_3, and g_4. The HMM $\mathcal{M}$ has four states that correspond to the four coalescent histories. Divergence time τ_1 is when species A and B diverged from their MRCA and divergence time τ_2 is when species C and the MRCA of A and B diverged from their MRCA. While the coalescence times of both nodes in each of g_2, g_3, and g_4 must be larger than τ_2, the coalescence time of a and b in g_1 must be between τ_1 and τ_2.

e of the species tree we associate two quantities, N_e and g_e, which correspond to the population size and generation time for the population corresponding to that branch.

The joint probability of a sequence H of hidden states and a sequence alignment $\mathcal{S}$, both of length n, is

$$P(\mathcal{S}, H) = P(H_1) \prod_{i=1}^{n} \left[P(\mathcal{S}_i|H_i) \cdot P(H_i|H_{i-1})\right]. \tag{1}$$

The likelihood of the model $\mathcal{M}$ is

$$L(\mathcal{M}) = \sum_{H} P(\mathcal{S}, H) \tag{2}$$

which can be efficiently computed by the forward (or backward) algorithm [8]. Furthermore, the local genealogy and its support for each site are efficiently computed using the Viterbi, forward, and backward algorithms [8] once the model is parameterized, e.g., to maximize the likelihood. For the initial probability, the method of [5] can be used to compute $P(H_1 = q_i) = P(g_i|\Psi)$. In our analyses below, we used a uniform distribution for the initial probability given the very small number of taxa we use.

We now describe how we derive the other two terms in the equation for $P(\mathcal{S}, H)$: the emission probability $P(\mathcal{S}_i|H_i)$ and the transition probability $P(H_i|H_{i-1})$.

2.1 Emission Probabilities

The HMM emits an observation $\mathcal{S}_k$ from state q_i with probability $p(\mathcal{S}_k|q_i, \Psi, \theta)$, where θ are the parameters of the assumed substitution model. In this work, we assume the Jukes-Cantor model [20] of sequence evolution, and θ consists only of μ, the per individual substitution rate per site per generation. The role that the species tree Ψ plays in determining the emission probabilities is in that

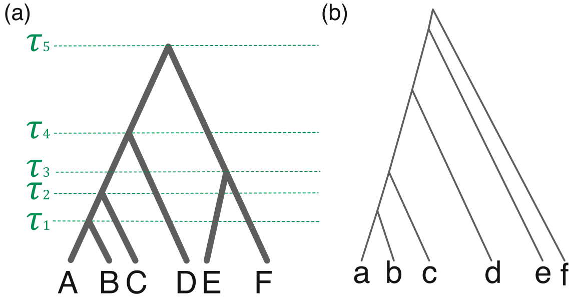

Fig. 3. (a) A species tree Ψ with times at its internal nodes in years. All the leaves are assumed to have time 0. (b) A coalescent history h, where ab coalesce in AB, abc coalesce in ABC, $abcd$ coalesce in $ABCD$, $abcde$ and $abcdef$ coalesce above the root.

it puts constraints on the ranges over which to integrate the branch lengths of the coalescent history. Let us consider species tree Ψ in Fig. 3(a) and coalescent history g in Fig. 3(b).

Given the times associated with the nodes of the tree Ψ, the lengths, in years, of the branches of history h satisfy:

- $\tau_1 \leq \ell_a, \ell_b \leq \tau_2$;
- $0 \leq \ell_{ab} \leq (\tau_4 - \tau_1)$;
- $\tau_2 \leq \ell_c \leq \tau_4$;
- $0 \leq \ell_{abc} \leq (\tau_4 - \tau_2)$;
- $\tau_4 \leq \ell_d \leq \tau_5$;
- $0 \leq \ell_{abcd} < \infty$;
- $\tau_5 \leq \ell_e < \infty$;
- $0 \leq \ell_{abcde} < \infty$; and,
- $\tau_5 \leq \ell_f < \infty$.

To turn these bounds into units of expected numbers of mutations, each bound for branch x (e.g., $x = abcd$ for the branch incoming into the MRCA of a, b, c, d in h) is multiplied by θ_x / g_x, where $\theta_x = 2N_x\mu$, where N_x is the population size associated with branch x, and g_x is the generation time (number of years per generation) for branch x.

Let ζ be a vector of upper and lower bounds on the lengths (in expected number of mutations per site per generation) of coalescent history g's branches. For every branch e in g, these bounds are derived from the constraints of the mapping between the nodes of g and the branches of the species tree as discussed above for the example in Fig. 3. Then, we have

$$p(\mathcal{S}_k|q_i, \Psi, \theta) = p(S_k|g_i, \Psi, \theta) = \int_{\mathbf{b}} p(S_k|g_i, \zeta, \theta) p(\mathbf{b}) d\mathbf{b} \tag{3}$$

where the definite integral is taken over all branch lengths of g_i, where for branch e, the integration is over $[\zeta_{lower}, \zeta_{upper}]$, or $[\zeta_{lower}, \infty)$ in the case of gene history nodes mapped to the root branch of the species tree, and $p(\mathbf{b})$ is a prior on the branch lengths.

The integrated likelihood of Eq. (3) can be computed efficiently with a modification to Felsenstein's "pruning" algorithm [11]. Recall that Felsenstein's algorithm computes the likelihood of a gene history with fixed branch lengths. It does so using a dynamic programming algorithm that operates on the tree in a bottom-up fashion, computing for a given site and for each node v, the quantity $C(x, v)$ which equals the probability of the subtree rooted at v given that v is labeled with state x (in the case of DNA sequences, $x \in \{A, C, T, G\}$).

In our case, the only difference from Felsenstein's algorithm is in computing $C(x, v)$ for internal nodes. If v is a node whose children are u and w, then we have

$$C(x,v) = \left(\int_{\zeta_{vu_{lower}}}^{\zeta_{vu_{upper}}} \left(\sum_y C(y,u) \cdot P_{x \to y}(b_{vu}) \right) p(b_{vu}) \mathrm{d}b_{vu} \right) \cdot \left(\int_{\zeta_{vw_{lower}}}^{\zeta_{vw_{upper}}} \left(\sum_y C(y,w) \cdot P_{x \to y}(b_{vw}) \right) p(b_{vw}) \mathrm{d}b_{vw} \right) \quad (4)$$

where $P_{x \to y}(b_{vu})$ is the probability of state x changing to state y over time that is given by the length of branch (v, u), b_{vu}. This is equal to

$$C(x,v) = \left(\sum_y C(y,u) \cdot \left(\int_{\zeta_{vu_{lower}}}^{\zeta_{vu_{upper}}} \left(P_{x \to y}(b_{vu}) \right) p(b_{vu}) \mathrm{d}b_{vu} \right) \right) \cdot \left(\sum_y C(y,w) \cdot \left(\int_{\zeta_{vw_{lower}}}^{\zeta_{vw_{upper}}} \left(P_{x \to y}(b_{vw}) \right) p(b_{vw}) \mathrm{d}b_{vw} \right) \right). \quad (5)$$

Assuming an $Exp(1)$ prior on branch lengths, we have $p(b_{vu}) = e^{-b_{vu}}$. As pointed out in [1], extremely long branch lengths are not realistic and a proper branch length prior puts decreasing density on higher branch lengths as is the case with our $Exp(1)$ prior. Each of the integrations in Eq. (5) can be computed analytically under the JC model as

$$\int_{\zeta_{e_{lower}}}^{\zeta_{e_{upper}}} \left(P_{x \to y}(b_e) \right) p(b_e) \mathrm{d}b_e = \int_{\zeta_{e_{lower}}}^{\zeta_{e_{upper}}} \frac{1}{4}(1 + 3e^{-4b_e/3}) e^{-b_e} \mathrm{d}b_e = -\frac{1}{4} e^{-\zeta_{e_{upper}}} + \frac{1}{4} e^{-\zeta_{e_{lower}}} - \frac{9}{28} e^{-7\zeta_{e_{upper}}/3} + \frac{9}{28} e^{-7\zeta_{e_{lower}}/3} \quad (6)$$

for the case of $x = y$ and

$$\int_{\zeta_{e_{lower}}}^{\zeta_{e_{upper}}} \left(P_{x \to y}(b_e) \right) p(b_e) \mathrm{d}b_e = \int_{\zeta_{e_{lower}}}^{\zeta_{e_{upper}}} \frac{1}{4}(1 - e^{-4b_e/3}) e^{-b_e} \mathrm{d}b_e = -\frac{1}{4} e^{-\zeta_{e_{upper}}} + \frac{1}{4} e^{-\zeta_{e_{lower}}} + \frac{3}{28} e^{-7\zeta_{e_{upper}}/3} - \frac{3}{28} e^{-7\zeta_{e_{lower}}/3} \quad (7)$$

for the case of $x \neq y$.

2.2 Transition Probabilities

The length of a non-recombining tract before encountering a recombination breakpoint is exponentially distributed with intensity equal to the total branch length of the gene history (in units of scaled recombination distance, or centiMorgans per megabase pair, cM/Mb) [17]. With this in mind, we approximate the transition probabilities by

$$p(q_i, q_{i'}) = \begin{cases} 1 - \frac{1}{L*+1} & i = i' \\ \frac{log(NRBS(g_i,g_{i'})+e)}{\sum_{i,i'} log(NRBS((g_i,g_{i'}+e)))} \cdot \frac{1}{L*+1} & i \neq i' \end{cases} \tag{8}$$

where L* is the expected tract length for the state we are currently in, or the reciprocal of the total branch length of the gene history in units of scaled recombination distance for the current state, i.e., $L* = \frac{1}{\sum_i b_i \cdot \lambda}$, where the sum is taken over all branch lengths of the genealogy. Here, we introduce the parameter λ to scale the branches of our trees to units of scaled recombination distance, which is to be learned from the data. Given that we are using integrated coalescent histories, our gene histories do not have specific branch lengths with which to derive $L*$. To obtain an estimate of L*, which requires specific branch lengths, we make another approximation by mapping our gene history nodes to the midpoint of the species tree branches that they coalesce inside of. This allows us to derive branch lengths which we convert to scaled recombination distance in order to finally derive L*, with gene history nodes coalescing above the species tree root being set to coalescing at the root node of the species tree.

For transitions between two different states, we make the simplifying assumption that when crossing a recombination breakpoint, the local genealogy changes to one that is very similar to the current one. We make use of the log scaled Normalized Rooted Branch Score, or NRBS, distance of [15]. The NRBS is the adaptation of the "branch score" given by [23] but for rooted trees. It gives a measure of tree similarity between two rooted trees with branch lengths. The transition probability between two different states is proportional to this distance and thus proportional to the similarity of the local genealogies in question.

2.3 Gene Flow: The Species Network Case

When the evolutionary history of the genomes under analysis involves gene flow (in addition to recombination), the species tree can be replaced by a phylogenetic network [24]. In this case, one can view every site in the genomic alignment as evolving down a local genealogy, but such a genealogy evolving within the branches of one of the parental trees inside the network [43].

One can view the phylogenetic network as a mixture of parental trees and each local genealogy evolves within one of the parental trees defined by the network [43].

Let $\Psi = \{\psi_1, \ldots, \psi_m\}$ and $H(\psi) = \{g_1, \ldots, g_r\}$ be the sets of all parental trees defined by the given phylogenetic network Ψ and all coalescent histories possible under a particular parental tree ψ, respectively. Then, the set of states of the HMM $\mathcal{M}$ in this case is $Q = \{q_{i,j} : 1 \leq i \leq m, 1 \leq j \leq r\}$, where state $q_{i,j}$ corresponds to parental tree ψ_i and coalescent history g_j. The emission probabilities in this case remain unchanged from the previous section (the constraints on the branch lengths of the coalescent history g_j in state $q_{i,j}$ come from the parental tree ψ_i). The initial probability is now calculated using the method of [41].

The transition probabilities, however, are modified slightly from the case of no gene flow. We assume that the HMM remains in a state with the same parental species tree with probability α and transitions to a state with a different parental species tree with probability $1-\alpha$ (normalized by the number of parental species trees minus 1). Multiplying α by the functions described in the no gene flow case is sufficient to handle the case of switching gene trees while remaining in the same parental tree:

$$p(q_{i,j}, q_{i',j'}) = \frac{h(q_{i,j}, q_{i',j'})}{\sum_{u,v} h(q_{i,j}, q_{u,v})}, \tag{9}$$

where u and v iterate over all parental trees and local genealogies, respectively, and

$$h(q_{i,j}, q_{i',j'}) = \begin{cases} 1 - \frac{1}{L*+1} & i = i', j = j' \\ \alpha \cdot \frac{log(NRBS(g_i, g_{i'})+e)}{\sum_{i,i'} log(NRBS((g_i, g_{i'}+e)))} \cdot \frac{1}{L*+1} & i = i', j \neq j' \\ \frac{1-\alpha}{m-1} \cdot \frac{log(NRBS(g_i, g_{i'})+e)}{\sum_{i,i'} log(NRBS((g_i, g_{i'}+e)))} \cdot \frac{1}{L*+1} & i \neq i' \end{cases}. \tag{10}$$

The value of α is to be learned during the training of the HMM. With this approximation in our model, the HMM switching between states of different parental trees denotes entering or exiting an introgressed region.

2.4 Learning the Model Parameters

To sum it up, under our formulation, the model parameters that need to be learned are the times τ associated with the species tree's or network's nodes, the population sizes N and generation times g associated with the species tree's branches, the DNA substitution model parameters θ (μ for this study as we only consider the Jukes-Cantor model of evolution), λ, and, in the case of a phylogenetic network, α. In this work, we assume known values for N and g instead of learning them from the data. For parameter inference, we rely on hill-climbing heuristics to obtain the parameter settings that maximize the likelihood of our model, $L(\mathcal{M})$. For the experiments conducted in this study, the parameters were learned using the BOBYQA multivariate optimizer based on [31], part of the Apache Commons Math library in Java.

When *a priori* knowledge of the model parameters exists, that knowledge could be used instead of learning the corresponding parameters. Furthermore, to the best of our knowledge, no work currently exists on investigating the identifiability of parameters in coalescent HMMs. Such an investigation is worth pursuing in future research.

3 Experimental Evaluation

For various simulated and biological data sets, we trained the model and analyzed the genomes using the Viterbi algorithm and/or the posterior decoding [32]. In the simulated data sets, the true topology of the gene history under which each site evolved is known. Therefore, in this case we are able to compare the true topology of the gene histories to the ones our model is confident about based on a combination of the forward and backward algorithms that compute the posterior decoding [32].

To obtain the labeling of sites with local genealogies using our method, we focused on the posterior decoding given by

$$P(H_i = g_k|\mathcal{S})$$

where H_i is the hidden state for site i in the alignment $\mathcal{S}$, and g_k is any of the possible gene histories on the three taxa. This quantity is efficiently computable using a combination of the forward and backward algorithms [32].

We also compare the results of our method against a commonly used method of obtaining local genealogies where a fixed window is passed across a genomic alignment and maximum likelihood or bayesian phylogenetic inference tools are used to infer a local genealogy for the window. In [3,12], maximum likelihood trees are built for every sequential non overlapping 50 kbp of their alignments. We do the same in our analyses, comparing our results to the local genealogies reported by RAxML [37] for sequential, non overlapping 50 kbp windows in our simulated alignments.

3.1 Results on Simulated Data with No Gene Flow

Three-taxon genome alignments were generated by first generating gene histories with recombination using `ms` [18] followed by the generation of nucleotide sequence alignments of length 500,000 using seq-gen [33] under the Jukes-Cantor model of evolution [20]. The specified species tree provides the population structure for `ms`, and the various parameters were set as we describe below. The program `ms` outputs a collection of gene genealogies, each corresponding to a genomic locus (specified by the start and end locations on the genome). The program seq-gen was then used to simulate sequence data down each genealogy (the length of the sequence alignment of each genealogy is specified by the genomic coordinates produced by `ms`). Finally, the sequence alignments of the loci are concatenated (while preserving the order produced by `ms`), and those are fed to the method for inferring local genealogies.

As in [16], our simulations consisted of three populations where the parameters used approximate the human/chimp/gorilla evolutionary scenario. Here, we label the human, chimp, and gorilla taxa as A, B, and C, respectively, and refer back to the parameters from Fig. 2. We set τ_1 to four million years and τ_2 to five and a half million years ago. We also use the same population parameters as [16] with N_A, N_B, and N_C being 30,000 versus the N_{AC} and N_{ABC} population sizes of 40,000. All generation times were set to be 25 (years per generation) and a recombination rate of r = 0.0075 was used, again following [16]. We set the mutation rate to $3.4 \cdot 10^{-8}$ from the estimates of human mutation rates in [27]. After generating the 500 kbp simulated data set, we heuristically obtained maximum likelihood estimates of the model parameters by using 160 runs of optimization allowing for up to 3,000 iterations per run.

Let K be a $1 \times 500,000$ vector indicating for each site in the alignment the true gene genealogy under which that site evolved (this is known in simulations). Let M^t be a $1 \times 500,000$ vector indicating for each site i in the alignment the gene genealogy g_s that satisfies $P(H_i = g_s|\mathcal{S}) > t$. Notice that when $t \leq 0.5$, more than one gene tree could satisfy the condition for site i. For each of the three possible gene trees g_s $(s = 1, 2, 3)$, we define the true positives rate (TPR) and false positives rate (FPR) as

$$TPR(s,t) = \frac{|\{i : 1 \leq i \leq 500,000, K[i] = M^t[i] = g_s\}|}{|\{i : 1 \leq i \leq 500,000, K[i] = g_s\}|}$$

and

$$FPR(s,t) = \frac{|\{i : 1 \leq i \leq 500,000, K[i] \neq g_s, M^t[i] = g_s\}|}{|\{i : 1 \leq i \leq 500,000, K[i] \neq g_s\}|}.$$

Finally, for each of the three gene trees, g_s $(s = 1, 2, 3)$, we plotted the receiver operating characteristic (ROC) curve, where t serves as the discrimination threshold. The three ROC curves for the three topologies for this simulation are shown in Fig. 4.

The results show a good performance of the method, despite its heavily simplifying approximations. To zoom in on the performance, we report on the TPR and FPR values for a few select thresholds in Table 1.

Clearly, as the discrimination threshold value increases, the true positive ratios decrease, but the false positives ratio also improves. Interestingly, the method performs consistently better in terms of correctly identifying the local genealogy g_1, whose topology is identical to that of the species tree, than either of the other two genealogies.

Alongside the analyses of our method, we split the alignment into non overlapping windows and generated maximum likelihood (ML) trees. For this, we added a fourth, outgroup taxon diverging 18 million years ago (corresponding to orangutans), also following [16], so as to enable inference of rooted three-taxon trees (the outgroup was used to root the ML trees). We used window sizes of 1 kbp, 10 kbp, and 50 kbp. In the case of window size 50 kbp, all windows for this no-gene flow case yielded the same ((a,b),c) topology in agreement with the divergence pattern of the overall population history and regardless of the true local genealogy. In the case of window size 10 kbp, only two of three possible topologies were estimated. The TPR and FPR of the window-sliding method for all three window sizes are shown in Table 1. Clearly, in these simulations, a window size of 1 kbp is the best choice, and is comparable in performance to our method with a discrimination threshold between 0.08 and 0.1. Nevertheless,

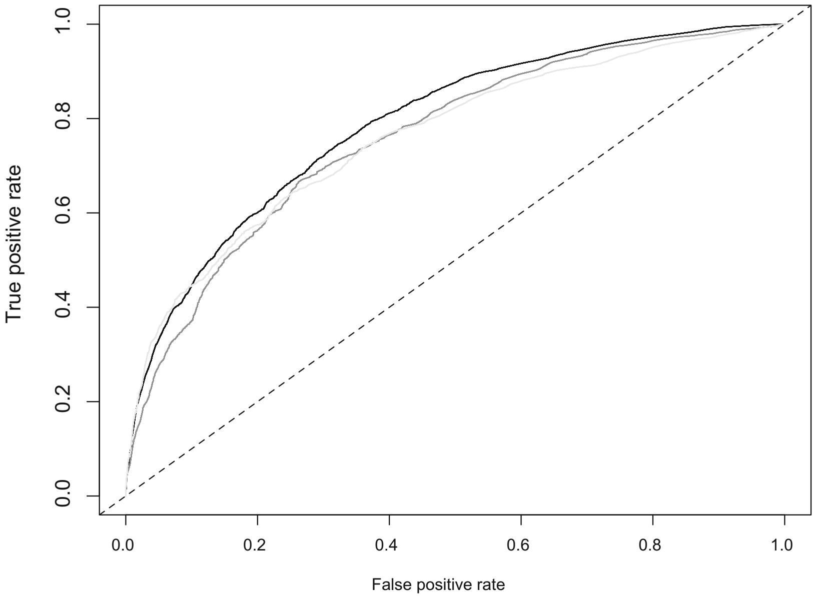

Fig. 4. ROC curves of the three gene topologies based on the posterior decoding support calculated by our method: $(((a, b), c)$ in black, $((a, c), b)$ in dark grey and $((b, c), a)$ in light grey. We used the ROC commands from the clinfun R package to generate this plot [35].

Table 1. True positive and false positive rates of our method for various posterior support thresholds, ARGweaver, and the window sliding method. The data was simulated with no gene flow, and with recombination rate $r = 0.0075$.

Discrimination threshold (t)	g_1 TPR	g_1 FPR	g_2 TPR	g_2 FPR	g_3 TPR	g_3 FPR
0.001	0.92	0.61	0.77	0.41	0.75	0.38
0.01	0.87	0.49	0.69	0.30	0.66	0.28
0.05	0.83	0.43	0.63	0.24	0.61	0.23
0.08	0.82	0.42	0.61	0.23	0.59	0.22
0.1	0.82	0.41	0.60	0.22	0.58	0.21
0.5	0.72	0.30	0.49	0.15	0.49	0.14
0.9	0.62	0.21	0.37	0.10	0.42	0.08
0.99	0.53	0.15	0.31	0.064	0.34	0.046
0.999	0.41	0.08	0.23	0.040	0.24	0.025
0.999999	0.14	0.011	0.078	0.0076	0.092	0.0067
N/A 1 kbp Windows	0.83	0.39	0.43	0.08	0.54	0.10
N/A 10 kbp Windows	0.97	0.92	0.083	0.034	0	0
N/A 50 kbp Windows	1	1	0	0	0	0
N/A ArgWeaver	0.84	0.33	0.47	0.10	0.53	0.09

it is important to make two points here. First, 10 kbp and 50 kbp window sizes are more commonly used in recent studies than 1 kbp or smaller window sizes. Second, in empirical data sets, we expect 1 kbp windows to have much less signal than we observe in simulations under our condition.

We also ran ARGweaver [34] (using its default parameter settings); the results are shown in Table 1. The method's performance is similar to that of our method with a discrimination threshold between 0.08 and 0.1 and to that of the 1 kbp window-sliding method. It is important to note here that ARGweaver is more powerful than our method and the window-sliding method in the sense that it computes many more quantities than the local genealogies.

We observed similar results of our method when varying the recombination rate as for the cases of $r = 0.0019$ and $r = 0.03$. With a recombination rate of $r = 0.03$, four times the rate used in [16], the method's accuracy drops. With extremely high rates of recombination the contiguous regions sharing an identical topology begin to become very short and recombination breakpoints start to become very frequent along the genome. As contiguous regions sharing the same topology become especially small, it is easy for there to not be enough signal for the learned HMM to switch states every few nucleotides, even with a higher learned λ value. Too high of a λ would also cause the areas that actually do have long contiguous regions to very easily switch states when there are alternative mutations by chance.

To better understand the behavior of methods for inference of local genealogies, particularly on closely related genomes, it is important to zoom in on the individual regions and assess the signal there. After all, the model local genealogy could be a fully resolved trees, but the true local genealogy would be completely unresolved if the the sequences have no substitutions. We zoomed in on a 1 kbp region of our simulated alignment from positions 14,501 to 15,500 in Fig. 5 and on a 1 kbp region from positions 58,001 to 59,000 in Fig. 6, where we see poor and excellent performance of the method, respectively.

In Fig. 5 we see three recombination breakpoints that alternate the true local genealogy, all of which are mislabeled by our method. The signal from the variable sites, however, is completely deceptive, and this figure highlights a pattern that would plague methods for inferring local genealogies. The first two regions (blue and green), are accompanied almost exclusively with variable sites which give signal for a gene history with topology ((b,c),a). We then see a very dense burst of variable sites supporting a gene history of ((a,c),b) whereas the true genealogy is the blue topology representing the ((a,b),c) gene history.

Conversely, in Fig. 6 we see our method working nearly perfectly in recovering the true gene history topologies as output by the simulation. We see two large blue and green regions recovered nearly exactly in both the Viterbi labeling and posterior decoding. Here, we see that the signal in the informative variable sites strongly supports the inferences as in the case of the extremely dense blue and green variable sites under the corresponding topology regions. Our method is even able to detect the recombination breakpoint occurring between these two regions almost exactly, within a few base pairs from the actual breakpoint.

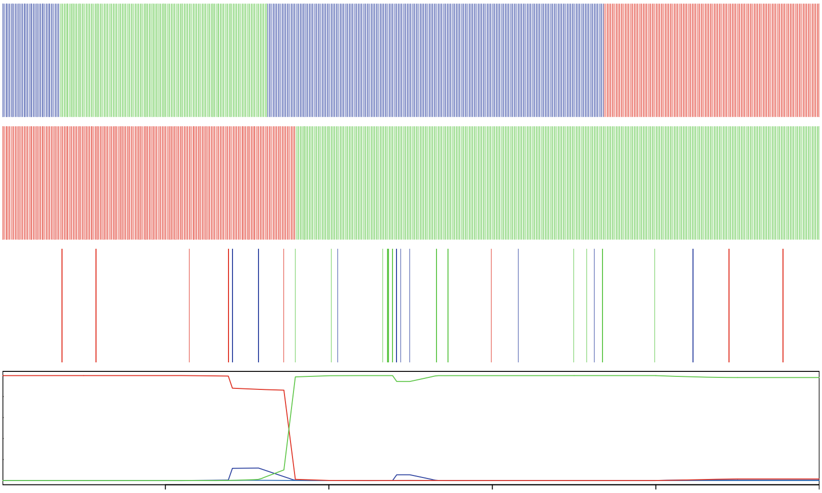

Fig. 5. Topology labels from alignment position 14,501 to 15,500 with recombination rate of r = 0.0075. (Top panel) True topologies as output by ms. (Second panel from top) Viterbi labelings from the trained HMM. (Third panel from top) Informative single nucleotide polymorphisms (sites where a and b have the same nucleotide and c is a different nucleotide shown in blue, etc. Invariant sites shown in white). (Bottom panel) Posterior support from the posterior decoding $((a, b), c)$ in blue, $((a, c), b)$ in green and $((b, c), a)$ in red. (Color figure online)

It has been long recognized that recombinations could be undetectable due to lack of signal [19,30]. This could arise, for example, if recombination results in the crossover between two identical sequences. This issue makes it hard, if not impossible, for methods to correctly delineate recombination breakpoints and, consequently, local genealogies.

3.2 Results on Simulated Data with Gene Flow

In this case, we simulated data as above, with recombination rates $r = 0.0075$ and a migration event with migration rate $M = 0.4$ (the species phylogeny was a network obtained by adding a horizontal edge from C to B to the species tree in Fig. 2).

When we assume the evolutionary history of our genome alignment is a network, we are interested not only in the accuracy of the local gene genealogies, but also of the areas that were gained across species boundaries through introgression (in our model, these would be regions labeled by a different parental tree). In analyzing the performance of our method, in addition to posterior support for the true gene tree topologies as shown before, we are able to look at whether the method is able to correctly label regions gained from migration or introgression. For our simulated data sets, we assessed this by analyzing cases where the true simulated gene genealogy's branch lengths allow B and C to coalesce before they would otherwise be allowed to. Unfortunately, the ms software

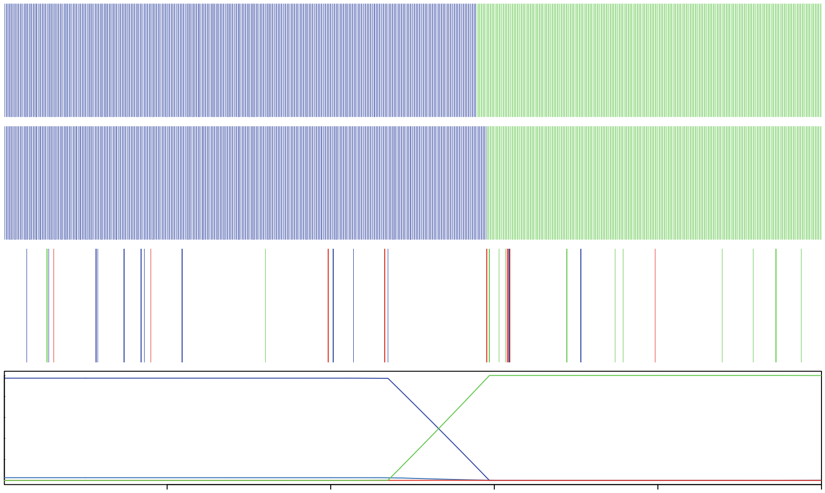

Fig. 6. Topology labels from alignment position 58,001 to 59,000 with recombination rate of r = 0.0075. (Top panel) True topologies as output by ms. (Second panel from top) Viterbi labels from the trained HMM. (Third panel from top) Informative single nucleotide polymorphisms (sites where a and b have the same nucleotide and c is a different nucleotide shown in blue, etc. Invariant sites shown in white). (Bottom panel) Posterior support from the posterior decoding $(((a, b), c)$ in blue, $((a, c), b)$ in green and $((b, c), a)$ in red. (Color figure online)

package does not support exact annotation of which gene trees are of introgressive origin, and modifying it to achieve this is not a simple task. Thus, our annotation of introgressed regions represents a lower bound on the true number of introgressed loci.

In this case, we focused on the performance of the method in terms of whether it elucidates that a site has been introgressed or not (that is, a binary classification problem). The ROC curve of the method is shown in Fig. 7. As above, the posterior decoding value was used as the discrimination threshold.

Once again, the method has good performance despite its simplistic approximations of the coalescent with recombination and gene flow. For example, it can achieve a TPR of about 0.8 and an FPR of about 0.3.

While the 50 kbp windows all yielded the same topology in the no gene flow case, we did see two windows with alternate topologies indicating introgressed tracts in this simulation.

3.3 Results on a Biological Data Set

In [16], regions of the human/chimp/gorilla genome alignment were analyzed and posterior support values were shown across the genome. We reevaluated a 32.6 kbp region with sufficient length and good patterns of large mixed gene genealogies to reanalyze with our new method. While in [16], the CoalHMM's

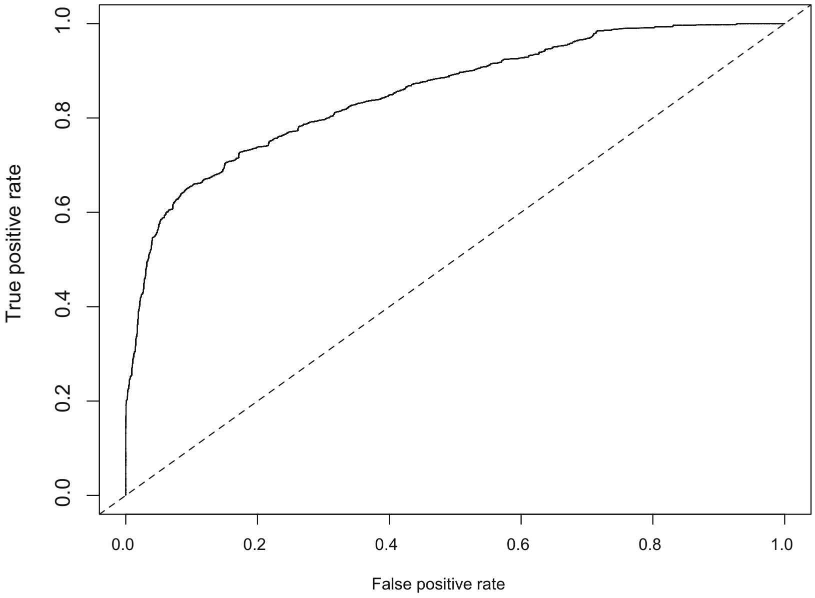

Fig. 7. ROC curve for labeling introgressed regions by our method based on varying posterior support as the discrimination threshold. We used the ROC commands from the clinfun R package to generate this plot [35].

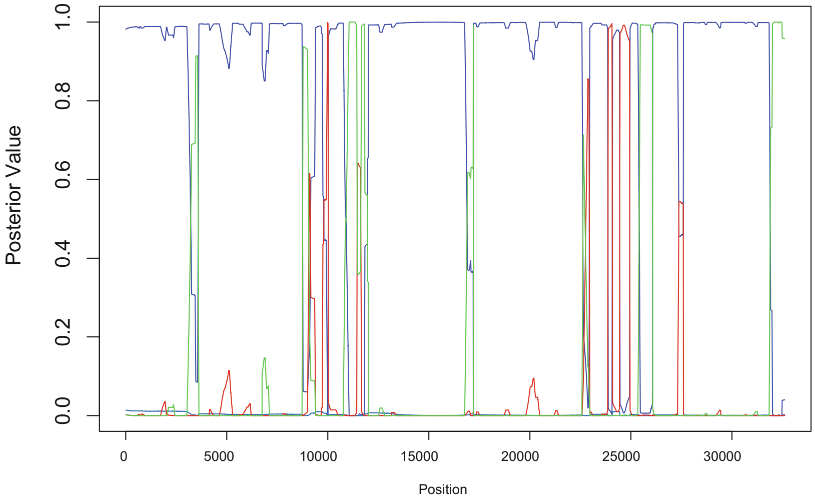

Fig. 8. Posterior support for a 32-kbp stretch of DNA obtained from human/chimp/gorilla genomes and analyzed in [16]. Posterior support for the three gene history topologies ((H,C),G), (H,(C,G)), and ((H,G),C) are in blue, green, and red, respectively. (Color figure online)

formulation is heavily reliant on *a priori* knowledge (to calibrate times, etc.), here the only *a priori* knowledge necessary was a given species phylogeny, population size, and generation time. We used a population size of 35,000 and generation time of 25 as we used in the inference step in our simulation analyses. With our generalized formulation, we found that our method was able to detect very similar regions of alternative gene topologies but with somewhat different support. Our results, shown in Fig. 8, directly mirror the results found in the posterior plot from Fig. S2 in [16] for the first large contiguous region stretching from ∼5 kbp to ∼35 kbp. We also used the window-based method to build trees based on 10 kbp sequential non overlapping windows. We found that all four 10 kbp windows gave the same topology of (h,(c,g)), once again disagreeing with both our results and the results from [16] indicating strong signal for alternative topologies in this region.

4 Discussion and Future Research

In this paper, we introduced an HMM-based method for inferring local genealogies on a genomic alignment in the presence of recombination and gene flow. The method is inspired by the work on coalescent HMMs [16,24], yet it strongly approximates the emission and transition probabilities since the focus here is on the local genealogies themselves, rather than accurate estimate of the evolutionary parameters. In our method, the gene tree branch lengths are integrated out when estimating emission probabilities, thus avoiding the uncertainty that comes with estimating branch lengths and fixing them for all sites (as done, for example, in [16]). Furthermore, transition probabilities are based on a rough measure of similarity between gene histories.

While we studied the performance of the method on small data sets in terms of the number of taxa, we showed that even for such small data sets, our method improves much over the common practice of sliding a window across the genomes, particularly for the commonly used window sizes of 10 kbp and 50 kbp. A direction for future research is to investigate the method's performance on larger data sets. This would give rise to a challenge that requires innovative solutions, namely how to ameliorate the "state explosion" problem. The number of gene, or coalescent, histories grows very fast with the number of taxa. Therefore, this approach becomes infeasible for large numbers of taxa. In particular, beyond a certain number of taxa, the number of possible gene histories far exceeds the number of sites in the genomes under study. This issue plagues not only our method, but all coalescent HMM methods. The window-sliding approach does not suffer from this problem, but its accuracy could suffer even more with more taxa. A marriage of the two approaches could provide one solution as follows. Sliding a window across the genomes yields a set of plausible local genealogies. However, this set could miss some of the potential local genealogies. Therefore, the set of trees identified by the window-sliding approach could be enriched. One way to enrich the set is to add, for example, all 1-NNI (nearest neighbor interchange) neighbors of all trees. Finally, this enriched set of local genealogies constitute the main states of the HMM. We will explore this aspect of the method.

As Hein *et al.* [14] noted, "local trees cannot be reduced to coalescent topologies or unrooted tree topologies, because trees with long branches will on average encounter a recombination breakpoint sooner than trees with short branches." Nevertheless, introducing states for every possible set of branch lengths is not possible. A strong approximation made by our method is that of the total length of a local genealogy (the b_i values in calculating the quantity $L*$). This approximation could be problematic in certain cases of larger data sets. A happy medium between integrating out branch lengths and using fixed ones must be found for all these approaches to scale. Still, an important message of this study is that sliding a window across the genomes to quantify incongruence is not the solution; after all, it could very well be the case that in cases of extensive incongruence the phylogenetic signal within a window is too small to infer accurate local genealogies. The Markov dependence on the neighboring genealogies in HMM-based approaches help ameliorate this problem as we demonstrated. However, it is important to acknowledge that at such low evolutionary scales, all these methods will run into problems of parameter (lack of) identifiability due to low signal.

In addition to our study being limited in the number of taxa, there are a number of simplifications to the evolutionary process that are assumed. Further research is required to expand the method to handle cases of gaps in genomic alignments, varied recombination and mutation rates across the genome (as in the case of recombination hotspots) and in expanding the mathematics of the integrated branch length emission probabilities to go beyond the Jukes-Cantor model. Additionally, augmenting the model to handle larger data sets in terms of the number of taxa will allow for more studies of existing empirical data sets.

Last but not least, we conducted a simple comparison to the ARGweaver method here. In future research, we will conduct a more through study of the various methods that have been devised for inferring local genealogies. In the case of gene flow, we will compare the performance of our method to others that identify introgression in sequence alignments, such as [7,28].

References

1. Alfaro, M.E., Holder, M.T.: The posterior and the prior in bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. **37**, 19–42 (2006)
2. Boussau, B., Guéguen, L., Gouy, M.: A mixture model and a hidden Markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. Evol. Bioinf. Online **5**, 67 (2009)
3. Heliconius Genome Consortium, et al.: Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**(7405), 94–98 (2012)
4. de Oliveira Martins, L., Leal, E., Kishino, H., Kishino, H.: Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS One **3**(7), e2651 (2008)

5. Degnan, J.H., Salter, L.A.: Gene tree distributions under the coalescent process. Evolution **59**(1), 24–37 (2005)
6. Degnan, J., Rosenberg, N.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. **24**(6), 332–340 (2009)
7. Durand, E.Y., Patterson, N., Reich, D., Slatkin, M.: Testing for ancient admixture between closely related populations. Mol. Biol. Evol. **28**(8), 2239–2252 (2011)
8. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
9. Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., Zhong, B., Wu, S., Lemmon, E.M., Lemmon, A.R., Leache, A.D., Liu, L., David, C.C.: Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. **94**, 447–462 (2016)
10. Elgvin, T.O., Trier, C.N., Tørresen, O.K., Hagen, I.J., Lien, S., Nederbragt, A.J., Ravinet, M., Jensen, H., Sætre, G.-P.: The genomic mosaicism of hybrid speciation. Sci. Adv. **3**(6), e1602996 (2017)
11. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**(6), 368–376 (1981)
12. Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K., Slotman, M.A., Emrich, S.J., Hahn, M.W., Besansky, N.J.: Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science **347**(6217), 1258524 (2015)
13. Hahn, M.W., Nakhleh, L.: Irrational exuberance for resolved species trees. Evolution **70**(1), 7–17 (2016)
14. Hein, J., Schierup, M.H., Wiuf, C.: Gene Genealogies, Variation and Evolution. Oxford University Press, Oxford (2005)
15. Heled, J., Drummond, A.J.: Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. **27**(3), 570–580 (2010)
16. Hobolth, A., Christensen, O., Mailund, T., Schierup, M.: Genomic relationships and speciation times of human, chimpanzee, and gorilla from a coalescent hidden Markov model. PLoS Genet. **3**(2), e7 (2007). doi:10.1371/journal.pgen.0030007
17. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7**(1), 44 (1990)
18. Hudson, R.R.: Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18**(2), 337–338 (2002)
19. Hudson, R.R., Kaplan, N.L.: Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**(1), 147–164 (1985)
20. Jukes, T., Cantor, C.: Evolution of protein molecules. In: Munro, H. (ed.) Mammalian Protein Metabolism, pp. 21–132. Academic Press, NY (1969)
21. Kingman, J.F.C.: The coalescent. Stochast. Processes Appl. **13**, 235–248 (1982)
22. Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D.: Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. **23**(10), 1891–1901 (2006)
23. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**, 459–468 (1994)
24. Liu, K., Dai, J., Truong, K., Song, Y., Kohn, M.H., Nakhleh, L.: An HMM-based comparative genomic framework for detecting introgression in eukaryotes. PLoS Comput. Biol. **10**(6), e1003649 (2014)

25. McVean, G.A., Cardin, N.J.: Approximating the coalescent with recombination. Philos. Trans. R. Soc. London B: Biol. Sci. **360**(1459), 1387–1393 (2005)
26. Minichiello, M.J., Durbin, R.: Mapping trait loci by use of inferred ancestral recombination graphs. Am. J. Hum. Genet. **79**, 910–922 (2006)
27. Nachman, M.W., Crowell, S.L.: Estimate of the mutation rate per nucleotide in humans. Genetics **156**(1), 297–304 (2000)
28. Pease, J.B., Hahn, M.W.: Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. **64**(4), 651–662 (2015)
29. Pond, S.L.K., Posada, D., Stawiski, E., Chappey, C., Poon, A.F., Hughes, G., Fearnhill, E., Gravenor, M.B., Brown, A.J.L., Frost, S.D.: An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput. Biol. **5**(11), e1000581 (2009)
30. Posada, D., Crandall, K., Holmes, E.: Recombination in evolutionary genomics. Annu. Rev. Genet. **36**, 75–97 (2002)
31. Powell, M.J.: The bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06. University of Cambridge, Cambridge (2009)
32. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **2**(2), 257–286 (1989)
33. Rambaut, A., Grass, N.C.: Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci.: CABIOS **13**(3), 235–238 (1997)
34. Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A.: Genome-wide inference of ancestral recombination graphs. PLoS Genet. **10**(5), e1004342 (2014)
35. Seshan, V.E.: clinfun: Clinical trial design and data analysis functions. R Package Version, **1**(6) (2014)
36. Springer, M.S., Gatesy, J.: The gene tree delusion. Mol. Phylogenet. Evol. **94**, 1–33 (2016)
37. Stamatakis, A.: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**(21), 2688–2690 (2006)
38. Takuno, S., Kado, T., Sugino, R., Nakhleh, L., Innan, H.: Population genomics in bacteria: a case study of staphylococcus aureus. Mol. Biol. Evol. **29**(2), 797–809 (2012)
39. Wiuf, C., Hein, J.: Recombination as a point process along sequences. Theor. Popul. Biol. **55**, 248–259 (1999)
40. Wu, Y.: New methods for inference of local tree topologies with recombinant snp sequences in populations. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **8**(1), 182–193 (2011)
41. Yu, Y., Degnan, J.H., Nakhleh, L.: The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. **8**(4), e1002660 (2012)
42. Zhang, W., Dasmahapatra, K.K., Mallet, J., Moreira, G.R., Kronforst, M.R.: Genome-wide introgression among distantly related heliconius butterfly species. Genome Biol. **17**, 25 (2016)
43. Zhu, J., Yu, Y., Nakhleh, L.: In the light of deep coalescence: Revisiting trees within networks. BMC Genom. **17**(14), 271 (2016)