RECOVERABILITY OF ANCESTRAL RECOMBINATION GRAPH TOPOLOGIES

ELIZABETH HAYMAN[1,*], ANASTASIA IGNATIEVA[2], AND JOTUN HEIN[3,4]

Abstract. Recombination is a powerful evolutionary process that shapes the genetic diversity observed in the populations of many species. Reconstructing genealogies in the presence of recombination from sequencing data is a very challenging problem, as this relies on mutations having occurred on the correct lineages in order to detect the recombination and resolve the placement of edges in the local trees. We investigate the probability of recovering the true topology of ancestral recombination graphs (ARGs) under the coalescent with recombination and gene conversion. We explore how sample size and mutation rate affect the inherent uncertainty in reconstructed ARGs; this sheds light on the theoretical limitations of ARG reconstruction methods. We illustrate our results using estimates of evolutionary rates for several biological organisms; in particular, we find that for parameter values that are realistic for SARS-CoV-2, the probability of reconstructing genealogies that are close to the truth is low.

Keywords. Recombination detection, ancestral recombination graph, coalescent, gene conversion.

1. Introduction

The reconstruction of genealogies from sequencing data in the presence of recombination has remained an important but challenging problem. Several algorithms have been developed recently for recovering the topology of local trees between recombination breakpoints, capable of tackling very large datasets using heuristic methods (e.g. Kelleher et al. 2019; Speidel et al. 2019). However, all methods that use sequencing data alone rely on mutations in the genealogical history in order to detect recombination and determine the ordering of coalescence events. Particularly when mutation rates are low, there may thus be significant uncertainty in the shape of the reconstructed local trees. Some tools (such as ARGweaver, Rasmussen et al. 2014) instead infer a distribution over genealogies, allowing inference methods to integrate over this uncertainty, although these are generally limited by computational power and can handle only moderate sample sizes.

In this article, we calculate the probability that the true topology of the genealogy can be recovered from the data, either in full or up to a specified number of ambiguous internal edges, under some simplifying assumptions. This sheds light on the performance of heuristic reconstruction methods, by quantifying how close to the truth they might get in the best case scenario, and the performance of methods exploring the distribution over compatible genealogies, by giving a sense of the size of the search space.

The coalescent with recombination is a widely used model for genealogies that extends coalescent trees to ancestral recombination graphs (ARGs) (Griffiths and Marjoram 1997). Under the commonly used *infinite sites* assumption, each mutation occurs at a new position of the genome. Recombination can then be detected using the *four gamete* test (Hudson and Kaplan 1985): denoting the ancestral allele by 0 and the derived allele by 1, if all four configurations 00, 01, 10 and 11 are observed at any two sites of a sample, then

[1] Department of Mathematics, University of Oxford, Andrew Wiles Building, OX2 6GG, UK
[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
[3] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK
[4] The Alan Turing Institute, British Library, London NW1 2DB, UK
[*] *E-mail*: elizabeth.hayman@keble.ox.ac.uk
Date: October 8, 2021.

the sample could not have been generated by mutation alone and at least one recombination must have occurred. For a recombination to generate such incompatible sites, the ARG topology must include a particular configuration of coalescence events preceding a recombination, and mutations must fall on the correct edges of the recombination cycle.

Under the coalescent with recombination, Myers (2003) derives the probability that conditional on a single recombination having occurred in the history of a sample, its effect is detectable from the sequencing data. This is achieved by constructing recursion equations for the probability of interest, by starting with the sample and considering each next event backwards in time. We utilise similar ideas and expand our consideration to the case of multiple recombination events, with the ARG topology constrained to be in the shape of a *galled tree*, i.e. an ARG where the recombination cycles do not interact with each other. This allows us to calculate the probability that, conditioning on R recombinations having occurred in the sample's history, these are all detectable, and the topology of each local tree can be recovered fully (or up to a fixed number of ambiguous internal edges). We also consider gene conversion—where a section of genetic material is taken from one parent genome, and the endpoints from another parent genome—and derive the probability that given one gene conversion event has occurred in the history of the sample, this is detectable from the sequencing data.

Where possible, we illustrate our findings using mutation and recombination rate parameters that are reasonable for biological organisms. Using published estimates of evolutionary rates for SARS-CoV-2, we take the population scaled mutation and recombination rates to be approximately $\theta = 100$ and $\rho = 0.1$ per genome, respectively (assuming a generation time of 7.5 days (Li et al. 2020), $N_e = 50$, mutation rate of $1 \cdot 10^{-3}$ per site per year (Duchene et al. 2020), recombination rate of $2 \cdot 10^{-6}$ per site per year (Müller et al. 2021)). We also consider *Drosophila melanogaster*, with $\theta = 8$ and $\rho = 21$ per kb, using estimates of Chan et al. (2012). For human populations, typical rates are $\theta = \rho = 0.1$ per kb, as used in previous analyses (Kelleher et al. 2019).

In Section 2, we first demonstrate our ideas in the simpler case that recombination is disallowed, i.e. when the genealogy is constrained to be a binary tree generated under the coalescent model. Then, in Section 3, we expand our results to include crossover recombination (under a two-locus model), and consider the probability of recovering the ARG topology when it is a galled tree. Further, in Section 4, we derive the probability that a gene conversion event is detectable from sequencing data. Discussion is presented in Section 5.

2. Recovering the topology of a tree

2.1. **Knowledge of full tree topology.** Disallowing recombination, we first consider the probability that the tree topology can be deduced fully from a sample of sequencing data. Without recombination, the coalescent history can be represented by a rooted tree, and a mutation on each internal edge is necessary and sufficient for unambiguously deducing the tree topology (Buneman 1971). In Figure 1, the first two trees are consistent with the same sequencing sample, so the sample is not sufficient to uniquely identify the coalescent tree topology. The third tree topology is uniquely associated with the sample, even though fewer mutations occur. Note that mutations on the terminal branches are not required in order to deduce the topology.

We note that some algorithms detect recombination by identifying changes to the marginal trees on either side of a breakpoint (Song and Hein 2005), so the detectability of recombination depends on how accurately the tree topologies at each locus can be reconstructed.

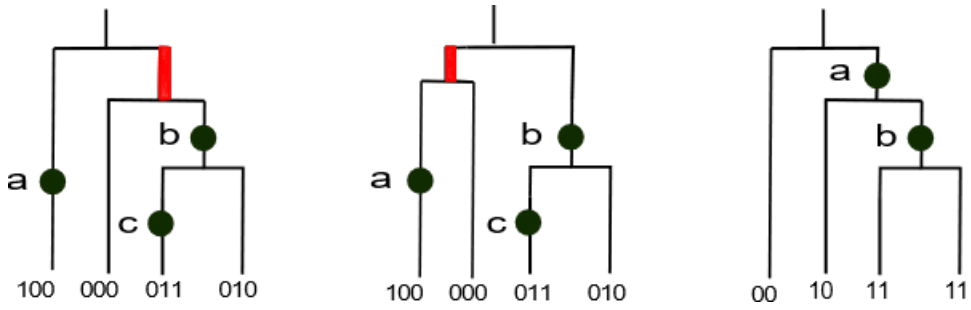


Figure 1. Examples of tree topologies that can and cannot be uniquely reconstructed from the data. Mutations are shown as dots; internal edges that do not carry a mutation are highlighted in red.

In order to derive the probability that the tree topology can be reconstructed fully from the data, we proceed by considering the genealogy backwards in time, and tracking whether at least one mutation has occurred on each internal lineage before it undergoes a coalescence event. At any point in time, each lineage can be in one of two states: if the lineage has not mutated since the last coalescence event, it is in State 1, and otherwise in State 2. Note that the terminal branches are taken to be in State 2.

Let n_l count the total number of lineages in the tree, and n_f count the number of lineages currently in State 2, that are therefore free to coalesce without losing knowledge of the tree topology. Define $p_{n_f}^{n_l}$ as the probability that the full tree topology is known given there are n_f lineages in State 2, while n_l lineages remain. Assigning all lineages to be in State 2 at the present time, p_n^n gives the probability the topology is fully recoverable starting with a sample size n.

We then construct recursion equations by considering the possible next event backwards in time. Letting $\lambda = \binom{n_l}{2} + n_l\,\theta/2$,

- with probability $\binom{n_l}{2}/\lambda$ the event is a coalescence of two lineages in State 2 (the number of lineages decreases by one, and the number of lineages in State 2 decreases by two);
- with probability $(n_l - n_f)\theta/2\lambda$ the event is a mutation of a lineage in State 1 (the number of lineages in State 2 increases by one, and consequently the number of lineages in State 1 drops by one);
- with probability $n_f\theta/2\lambda$, the event is a mutation of a lineage in State 2 (no change).

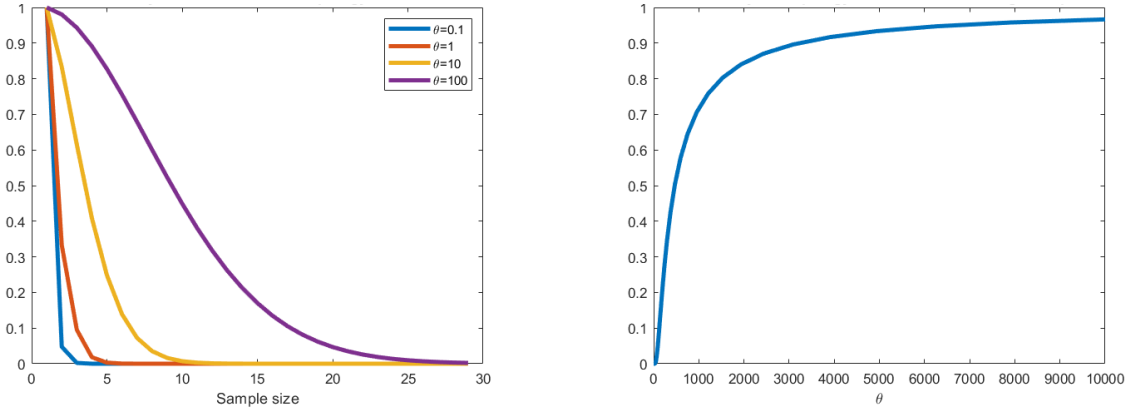
The recursion thus takes the form

$$\left(\binom{n_l}{2} + n_l\frac{\theta}{2}\right)p_{n_f}^{n_l} = \binom{n_f}{2}p_{n_f-2}^{n_l-1} + (n_l - n_f)\frac{\theta}{2}p_{n_f+1}^{n_l} + n_f\,\frac{\theta}{2}p_{n_f}^{n_l}, \quad 0 \le n_f \le n_l, \qquad (1)$$

with initial condition $p_0^1 = 1$. This is the simplest case of recursions we present, with a runtime of roughly n^2. Further on in the text, recursions become more complex, but can be solved efficiently via dynamic programming and shouldn't require any matrix inversion. We use MATLAB to solve the recursions.

Figure 2 illustrates the results for various values of n and θ. The left panel shows that the probability of knowing the full topology is monotonically decreasing in n for fixed θ; this is because larger values of n have shorter time periods between coalescence events, so it is less likely that mutations will occur on all the necessary edges. The right panel shows that the probability of detection for a sample of fixed size ($n = 20$) is increasing in θ, with a limit of 1 as $\theta \to \infty$. However, note that $\theta \approx 10^5$ is required for near certainty that the topology is recovered fully, which is unfeasibly large for biological samples (where typically

$\theta \leq 100$). For $\theta = 100$, which is typical for SARS-CoV-2 genomes, the probability of recovering the tree topology becomes very small for sample sizes over 25. For *Drosophila melanogaster*, with $\theta \approx 10$, the probability of recovering the true topology is minuscule for $n > 10$.



(a) Varying n for several fixed values of θ (colours).

(b) Fixed $n = 20$ and varying θ.

Figure 2. Probability that the full tree topology is known.

Fu and Li (1993) derive the total length of internal and external branches in a coalescent tree. Using our time scaling, the expected total length of internal branches is $2\left((\sum_{j=1}^{n-1} 1/j) - 1\right)$. Given that mutations occur as a Poisson process with rate $\theta/2$, the expected total number of mutations on interior branches is

$$\theta\left((\sum_{j=1}^{n-1} 1/j) - 1\right) \sim \theta \log(n).$$

This gives some intuitive understanding of the above graphs: for a sample of size n, there are $n-1$ coalescent events, and thus a minimum of $n-2$ mutations are required to have one on each interior branch. Therefore, even before the precise placement of individual mutations is considered, the total number of mutations needed to know the full topology increases like n, while the number of mutations expected to occur on the interior branches increases like $\log(n)$. Hence, the probability of knowing the full tree drops to 0 quickly.

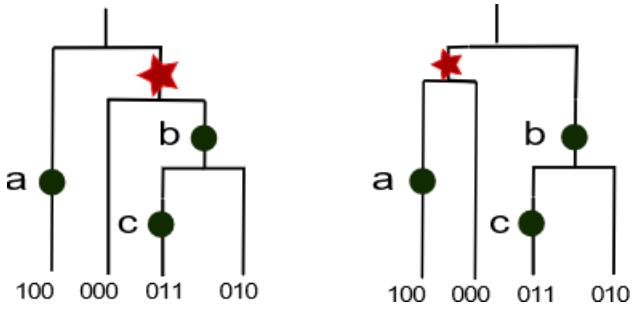


Figure 3. Two possible trees for the same sample data. Mutations are shown as dots on the branches. Starred internal edges carry no mutations.

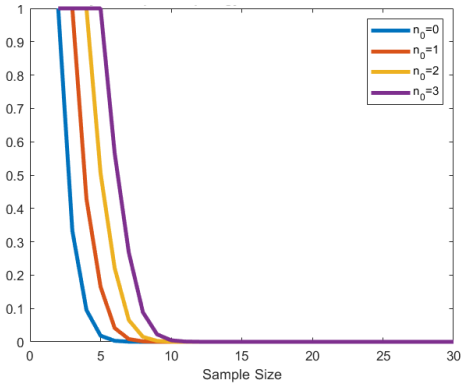
2.2. **Incomplete knowledge of tree topology.** The probability of recovering the full topology decreases rapidly as n increases, so we next consider the probability of partial knowledge of the tree. Here, the full topology is required except for the exact placement

of n_0 internal edges. As observed above, internal edges are fixed in the tree by having a mutation between each consecutive coalescence, so an "uncertainty" equates to one missed mutation. Figure 3 demonstrates two possible trees for $n_0 = 1$ and $n = 4$: the trees are consistent with the same sample data, and have the same topology apart from the interior branch joining the second sequence to the rest of the tree. A mutation at the starred position would fix the reconstructed genealogy to one of these two possibilities.

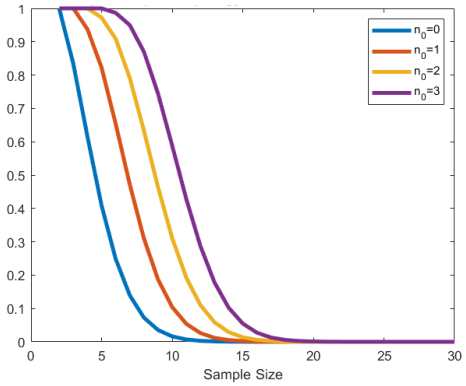
We extend the results of the previous section by including n_0 as a recursive index to track the number of internal branches that have not undergone at least one mutation. By considering the next event backwards in time when n_f out of n_l lineages are in State 2, the equivalent recursion to (1) is

$$\left(\binom{n_l}{2} + (n_l - n_f)\frac{\theta}{2}\right) p_{n_f}^{n_l,n_0} = \binom{n_f}{2} p_{n_f-2}^{n_l-1,n_0} + (n_l - n_f)\frac{\theta}{2} p_{n_f+1}^{n_l,n_0} + n_f(n_l - n_f) p_{n_f-1}^{n_l-1,n_0-1} + \binom{n_l - n_f}{2} p_{n_f}^{n_l-1,n_0-2}. \quad (2)$$

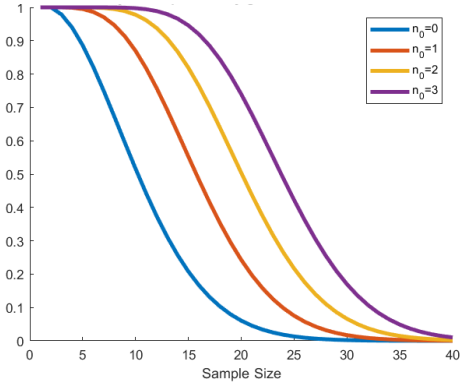
The initial conditions are $p_0^{1,n_0} = 1 = p_1^{1,n_0}$. Note that we are considering the probability of full knowledge of the tree topology except for the position of *up to* n_0 internal branches. The probability of missing *precisely* n_0 branches can be calculated as $p_{n_f}^{n,n_0} - p_{n_f}^{n,n_0-1}$.



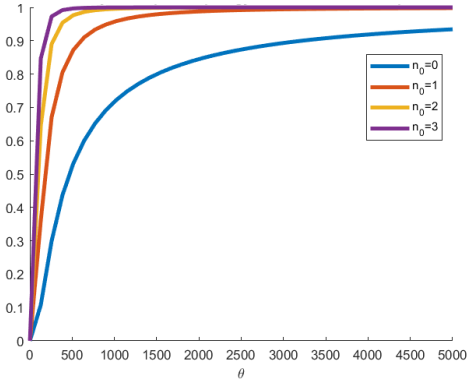
(a) $\theta = 1$.



(b) $\theta = 10$.



(c) $\theta = 100$.



(d) For a fixed sample $n = 20$ varying θ over biologically feasible values.

Figure 4. Probability of recovering the topology while allowing a small number n_0 of uncertain positions. Panels (a)-(c) show the probabilities for different values of the mutation rate against sample size, for a range of n_0 (colours). Panel (d) shows the probabilities for a fixed sample size $n = 20$ against θ, for a range of n_0 (colours).

Figure 4 shows how the probability of reconstructing the true topology up to n_0 internal branches varies with θ and n_0. Note that increasing n_0 in panels (a)-(c) appears to shift the probability curve to the right, for each value of θ, and that the magnitude of the shift increases with θ. Panel (d) highlights that allowing even a small number of unresolved edges results in a large increase in the probability of correctly reconstructing the rest of the tree.

2.3. **History of a specific lineage.** Even allowing for some uncertainty in recovering the placement of internal branches, the probability of knowing the tree topology still decreases rapidly with increasing sample size. We next focus on the probability of determining the history of a specific lineage. This is useful if there is particular interest in the history of a specific sequence: for instance, in the context of viral genealogies, our results quantify the probability that a particular viral strain can be accurately placed in the overall genealogy. Figure 5 shows an example of a tree topology where the history of the lineage highlighted in red can be recovered unambiguously, while allowing for uncertainty in the rest of the tree topology. This, again, requires that mutations occur on all of the internal branches highlighted in red.

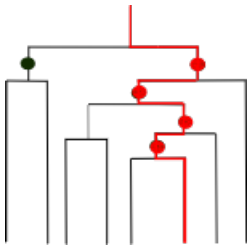


Figure 5. Example of a tree with a single lineage of interest (highlighted in red).

This simplifies the model considered in Section 2.1, as the dependence on n_l can be dropped, with the recursions only focussing on the state of the one lineage. Here, n_0 counts missed mutations that occur *only* on the single lineage of interest. The equivalent to (1) in this setting is

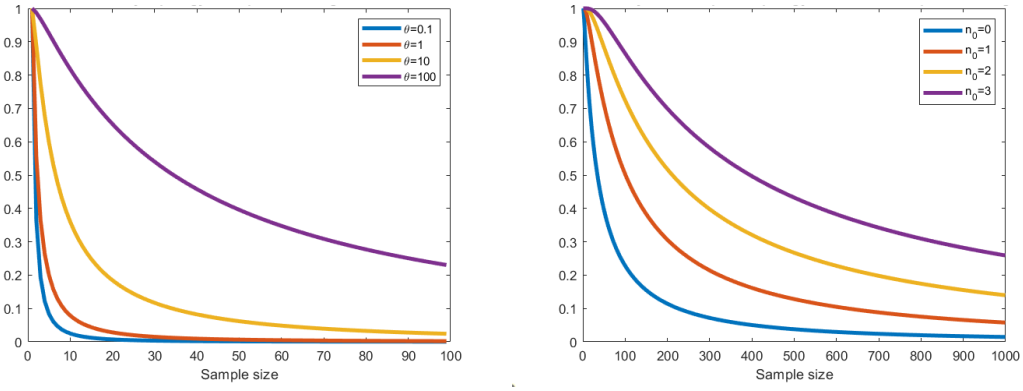
$$\begin{aligned}\left(\binom{n_l}{2}+\frac{\theta}{2}\right)p_1^{n_l,n_0} &= \binom{n_l-1}{2}p_1^{n_l-1,n_0}+(n_l-1)p_1^{n_l-1,n_0-1}+\frac{\theta}{2}p_2^{n_l,n_0},\\ \binom{n_l}{2}p_2^{n_l,n_0} &= \binom{n_l-1}{2}p_2^{n_l-1,n_0}+(n_l-1)p_1^{n_l-1,n_0}. \end{aligned} \tag{3}$$

The initial conditions are $p_1^{1,n_0}=1=p_2^{2,n_0}$.

Figure 6 shows plots of the resulting probabilities of knowing the history of a specific lineage. The probabilities at each point are consistently much greater for a single lineage, and are non-negligible for large sample sizes. Taking the SARS-CoV-2 value of $\theta\approx 100$, panel (a) shows that the genealogical history of a particular lineage from a sample size of 20 has 75% chance of being reconstructed accurately. Panel (b) considers partial knowledge of a single lineage, up to n_0 missed edges. As before, even a small degree of flexibility (up to three undetermined interior edges out of a sample of hundreds) leads to a significant improvement in recoverability.

3. Recovering the topology of an ARG

We now extend our results to include crossover recombination, through analysing the probability of recovering the (partial) ARG topology for a two-locus model. We constrain



(a) Varying n for several fixed values of θ (colours).

(b) $\theta = 100$, allowing up to n_0 undetermined internal edges (colours).

Figure 6. Solutions to the recursions focusing on a single lineage. Note the longer scale on the x-axis in panel (b).

the shape of the ARG to be a *galled tree*, disallowing interaction between the recombination loops. An example of ARGs which are and are not galled trees is shown in Figure 7.

First, we explore the probability that an ARG generated under the coalescent with recombination take the form of a galled tree, showing that our investigation is particularly relevant when the sample size is relatively small or the recombination rate is moderately low. Building on the work of Myers (2003), which presents detectability conditions in the case of a single recombination, we then derive recursion equations for the probability that the full (or partial) ARG topology is recoverable from the data, conditioning on R recombinations.

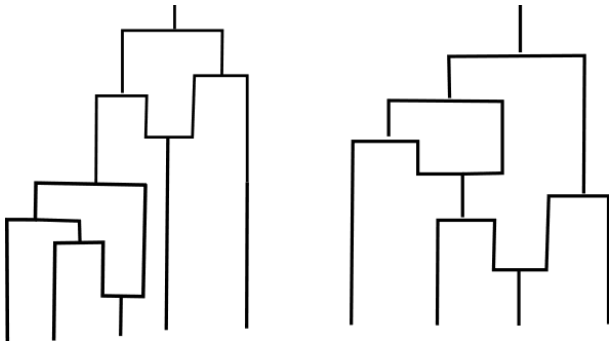


Figure 7. Left panel: ARG with two recombinations which is a galled tree (the recombination loops do not interact). Right panel: ARG that is not a galled tree, as the two recombinations loops are intertwined.

3.1. **Probability of an ARG being a galled tree.** Gusfield (2014, p.237) defines a galled recombination cycle in a phylogenetic graph as one which "shares no node with any other recombination cycle", and hence a *galled tree* as one where each recombination cycle satisfies this condition. Gusfield notes that ARGs are likely to be galled trees if the recombination rate is low, or if there is reason to believe that recombination has only occurred relatively close to the present. We derive an explicit expression for the probability that an ARG with n leaves and known recombination rate ρ contains only galled recombination cycles. Define an *open* recombination loop as one where the two

recombinant lineages have not yet coalesced back with each other, looking backwards in time.

First, we constrain the system by assuming that at most R recombinations have occurred in the history of the sample, with the intention of later removing this conditioning by taking $R \to \infty$. The relevant probabilities are $q^{n_l,r,R}$, being the probability that an ARG has at most R recombinations in the full history, given that there are currently n_l lineages present, with r out of R possible recombination loops currently open. By considering the genealogy backwards in time and conditioning on the next possible event, this is solved via the following recursion:

$$q^{n_l,R,R} = \frac{n_l - 1}{n_l - 1 + \rho} q^{n_l-1,R,R}, \tag{4}$$

$$q^{n_l,r,R} = \frac{n_l - 1}{n_l - 1 + \rho} q^{n_l-1,r,R} + \frac{\rho}{n_l - 1 + \rho} q^{n_l+1,r+1,R}, \quad r \leq R, \tag{5}$$

$$q^{n_l,R,R} = 0, \quad r > R.$$

The initial condition is $q^{1,r,R} = 1,\ r \leq R$. Then $q^{n,0,R}$ gives the probability starting from a sample of size n.

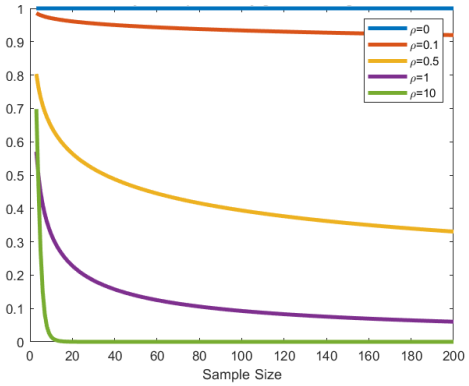
Now, let $p^{n_l,r,R}$ be the probability that an ARG has at most R *galled* recombinations in the history, conditional on n_l lineages currently present in the sample, with r out of R recombinations currently open. Any ARG with at most one recombination is trivially galled, so $p^{n_l,r,1} = q^{n_l,r,1}$. This gives the boundary conditions for the general case:

$$\begin{aligned}\frac{n_l}{2}(n_l - 1 + \rho) p^{n_l,R,R} &= R \cdot p^{n_l-1,R-1,R-1} \\ &\quad + \left(\frac{1}{2}(n_l - 2R)(n_l - 2R - 1) + 2R(n_l - 2R)\right) p^{n_l-1,R,R},\end{aligned} \tag{6}$$

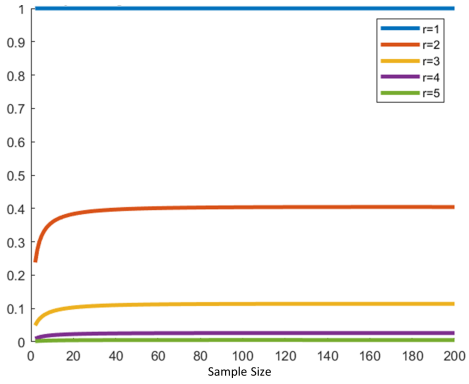
$$\begin{aligned}\frac{n_l}{2}(n_l - 1 + \rho) p^{n_l,r,R} &= r \cdot p^{n_l-1,r-1,R-1} \\ &\quad + \left(\frac{1}{2}(n_l - 2r)(n_l - 2r - 1) + 2r(n_l - 2r)\right) p^{n_l-1,r,R} \\ &\quad + \frac{\rho}{2}(n_l - 2r) p^{n_l+1,r+1,R}.\end{aligned} \tag{7}$$

Then $p^{n,0,R}/q^{n,0,R}$ gives the probability that an ARG with n leaves is a galled tree, conditional on at most R recombinations occurring. Taking $R \to \infty$ removes the conditioning on R (as a finite number of recombinations occurs in any history with probability 1).

Figure 8 illustrates the results for a range of recombination rates and values of R. The left panel demonstrates that when the recombination rate is low, ARGs are galled trees with high probability—this is both due to the ARGs being likely to contain at most one recombination node (and hence being trivially galled), or the recombinations being 'far apart' in the ARG so that the recombination loops are not likely to interact. The right panel shows that, conditioning on two recombinations and assuming a low recombination rate, the ARG is a galled tree with reasonably high probability, of around 0.4 when the sample size is moderate. This suggests that the galled tree restriction might be reasonable when analysing whole-genome SARS-CoV-2 data, for instance, and human or drosophila samples of relatively short genomic regions.



(a) Varying n for several fixed values of recombination rate (colours). To approximate the limit $R \to \infty$, a suitably high value of $R = 75$ was chosen.



(b) Varying n, conditioning on the number of recombinations (colours). To average over $\rho \in (0, 1)$, a uniform prior was used.

Figure 8. Probability that an ARG is a galled tree.

3.2. **Knowledge of full ARG topology.** Conditioning on one recombination having occurred in the history of a sample, Myers (2003) considers the probability of this recombination being detectable, i.e. the probability that the recombination changes the ARG topology, and mutations fall on the correct edges of the recombination loop to create incompatibilities in the data. Incompatibilities can then be detected by the four gamete test: under the infinite sites assumption and with the ancestral type known, at most two of the three non-ancestral allelic types generated by any pair of sites can be present in the data in the absence of recombination. The presence of three non-ancestral types therefore indicates a recombination event between the two sites (Hudson and Kaplan 1985). A topology changing recombination is one where the marginal trees to either side of the recombination give differing labelled graphs.

We first outline the results of Myers (2003) for the detectability of a single recombination. Then, through constructing similar recursion equations, we extend to the case of $R \geq 1$ recombinations having occurred, with the ARG in the shape of a galled tree; we calculate the probability that the ARG topology is recoverable, fully or with a specified number of unresolved internal edges.

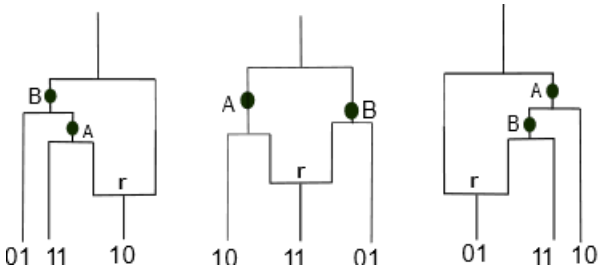


Figure 9. Positioning of mutations on the ARG with a single recombination that are required for the recombination to be detectable. Ancestral type is assumed to be 00.

3.2.1. *Detectability of a single recombination.* In this section, we summarise the results of Myers (2003, Section 4.3) calculating the probability that a recombination is detectable, when conditioning on a single recombination with breakpoint at $r \in [0, 1]$. The necessary conditions on the ARG topology and positions of mutations are illustrated in Figure 9;

one of the shown configurations of mutations inside the recombination loop must occur in order for the recombination to be detectable from the data. Denote A-type mutations as those occurring to the left of the recombination breakpoint, and B-type those to the right. Assuming that the ancestral type is known to be 00, all of the configurations shown in the Figure generate incompatible sites at A and B.

The probability of the recombination being detectable is calculated through constructing recursion equations, beginning with the sample and tracking the state of each lineage backwards in time. The possible states for the recombinant lineages emerging from the left-hand side of the recombination node are given in Table 1. The states for the right-hand lineage are equivalent, but with A and B reversed. Myers notes that for one recombination to be detectable, it is sufficient to have either lineage reach State 4, or for both lineages to simultaneously be in state ≥ 2.

Table 1. States described for the left recombinant edge denoted $\mathcal{E}$

State 0	No coalescence has occurred on edge $\mathcal{E}$ since the recombination.
State 1	There has been at least one coalescence since the recombination. No mutations have occurred since the last coalescence.
State 2	$\mathcal{E}$ has reached state 1 and a type A mutation has occurred since the last coalescence.
State 3	$\mathcal{E}$ has reached state 2 and undergone one further coalescence.
State 4	$\mathcal{E}$ has reached state 3 and a type B mutation has occurred since the last coalescence.

Let $p^n_{i,j}$ be the probability that while n lineages remain in the tree, $\mathcal{E}$ is in state i, and $\mathcal{F}$ is in state j. Condition on the recombination occurring while k lineages are present. The required recursion equations are then formulated by considering the next event back in time which changes the state of either recombinant lineage. A mutation moves a lineage in State 1 to State 2; a coalescence moves a lineage in State 2 to State 3 and reduces n by 1.

3.2.2. *Knowledge of full ARG topology.* We now calculate the probability of recovering the full ARG topology, conditioning on R recombinations having occurred. This requires at least one mutation between each coalescence event, and also mutations within the recombination loops occurring at positions certain to generate incompatible sites. This requires more states to track the recombinant lineages. As we consider only galled trees, the detection conditions stated above must hold in every gall.

A fixed number of recombinations R are allowed to occur in the history of a sample of size n. Suppose that at some point in time, the total number of remaining lineages is n_l, and the number of non-recombinant lineages which have undergone a mutation since the last coalescence event is n_f. The other indices are given in the third column of Table 2, tracking the number of left recombinant lineages in various states (with equivalent states for the right recombinant lineages). The indices $i, j, k_1, k_2, l_1, l_2, m_1, m_2$ (resp. $e, a, b_1, b_2, c_1, c_2, d_1, d_2$) count the number of left (resp. right) lineages in states 0, 1,..., 7. The index r tracks the number of recombination loops currently open. Note that we must have

$$r = i + j + k_1 + k_2 + l_1 + l_2 + m_1 + m_2 = e + a + b_1 + b_2 + c_1 + c_2 + d_1 + d_2.$$

Let

$$p^{n_l,n_f,r}_{i,j,k_1,k_2,l_1,l_2,m_1,m_2,e,a,b_1,b_2,c_1,c_2,d_1,d_2} \tag{8}$$

Table 2. States are described for the left recombinant edge denoted $\mathcal{E}$ (third column gives the index that counts the number of lineages in each state).

State 0	No coalescence has occurred on edge $\mathcal{E}$ since the recombination.	i
State 1	There has been at least one coalescence since the recombination. No mutations have occurred since the last coalescence.	j
State 2	$\mathcal{E}$ has reached state 1 and a type B mutation has occurred since the last coalescence.	k_2
State 3	$\mathcal{E}$ has reached state 1 and a type A mutation has occurred since the last coalescence.	k_1
State 4	$\mathcal{E}$ has reached state 3 and undergone one further coalescence.	l_1
State 5	$\mathcal{E}$ has reached state 4 and a type A mutation has occurred since the last coalescence.	m_2
State 6	$\mathcal{E}$ has reached state 4 and a type B mutation has occurred since the last coalescence.	m_1
State 7	$\mathcal{E}$ has reached state 6 and undergone one further coalescence.	l_2

be the probability that given r recombinations have occurred and n_l lineages remain, there are i recombination loops with left lineage in State 1, j in State 2, and so on. As we restrict to the case of galled trees, none of the recombinant lineages can interact with any other recombinant lineages except to close the recombination loop. The recursions for this system are described in full in the Appendix, Section A.1, and the recursion solved in MATLAB to find $p^{n,n,0}_{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}$.

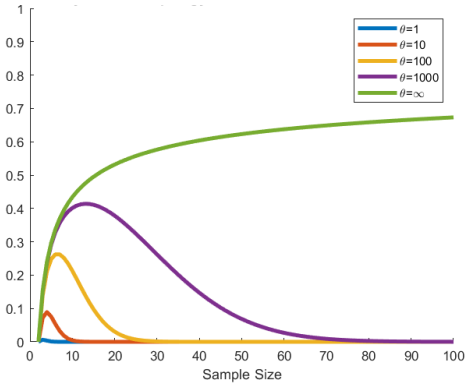
3.3. **Incomplete knowledge of ARG topology.** We also extend this system to consider the probability that the full topology is known apart from n_0 unresolved internal edges. This is done similarly to the case for binary trees described in Section 2, by including more states to track each lineage, so the full details of the resulting equations are presented in the Appendix, Section A.1.

3.4. **One recombination.** We first consider the results when conditioning on one recombination, setting $R = 1$; this is a realistic scenario when analysing sequencing data from species with a low recombination rate. We fix $\rho = 0.1$, being suitably small so that the assumption of a single recombination is valid (Myers 2003), noting that this is the estimated value of the recombination rate for SARS-CoV-2 genomes, and human samples of length 1kb.

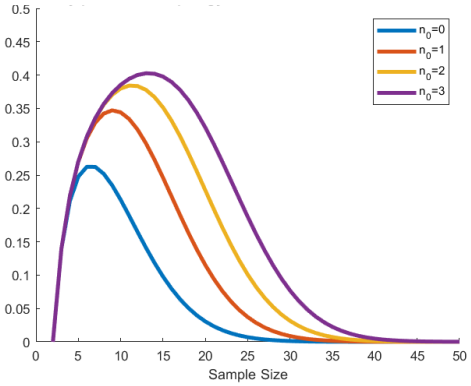
Figure 10 (a) shows solutions of the recursive system for various values of the parameters. Note that these curves are not monotonic: there must be a sufficient number of coalescences above the recombination to create incompatible sites in the sample, but increasing the number of lineages makes it unlikely that a mutation occurs between each coalescence (required to make the topology fully detectable). The results demonstrate that the probability of recovering the full ARG topology is very low for even moderate values of θ, increasing very slowly as $\theta \to \infty$.

Figure 10 (b) demonstrates that allowing just a small number of 'missed' internal edges substantially improves the probability of recovering the rest of the ARG topology correctly. For instance, with $n = 15$, the probability of recovering the ARG topology increases from around 0.1 to 0.4 if up to three unresolved internal edges are allowed.

Solutions to the recursive system while varying the breakpoint across the genome, z, show that taking a breakpoint closer to the centre of the genome gives slightly higher probabilities of detecting the full topology (see Figure 15 in the Appendix). For just



(a) Breakpoint fixed at $z = 0.5$, varying n for various values of θ (colours).

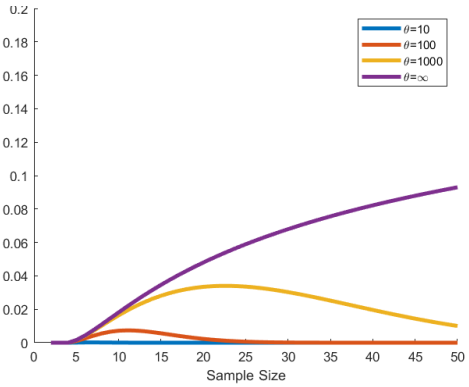


(b) Fixed $\theta = 100$, varying n, for various numbers n_0 of unresolved internal edges (colours).

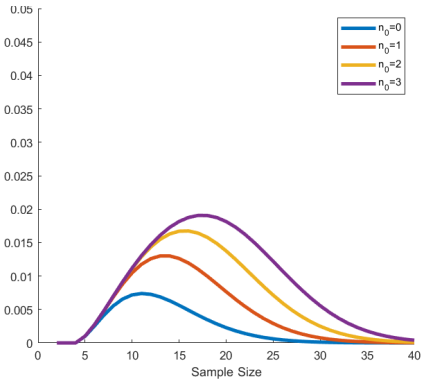
Figure 10. Probabilities of recovering the correct ARG topology conditional on one recombination.

detecting the recombination, a similar effect was noted by Myers (2003, p54), as mutations must occur on both sides of the breakpoint, and mutation rate varies linearly with genome length. By symmetry, having a breakpoint at z will result in identical probabilities to a breakpoint at $1 - z$.

3.5. **Two recombinations.** While ARGs containing only one recombination are trivially galled, Figure 8 shows that around 40 percent of trees containing 2 recombination nodes will be galled for $\rho = 0.1$. The probability of an ARG being a galled tree falls substantially when conditioning on more than two recombinations, so we do not analyse this case in further detail.



(a) Breakpoint fixed at $z = 0.5$. Varying n for various values of θ (colours).



(b) Fixed $\theta = 100$, varying n, for various numbers n_0 of unresolved internal edges (colours).

Figure 11. Probabilities of recovering the correct ARG topology conditional on two recombinations.

Figure 11 illustrates solutions of the recursion equations when conditioning on two recombinations. The probabilities of recovering the full ARG topology are significantly smaller, with less than a tenth of ARGs being fully recoverable even with an infinite mutation rate, for $n = 100$. Here, the conditioning restricts to two recombinations within the history. However, if instead we condition on two *galled* recombinations, Figure

8 suggests a maximum probability of $0.1/0.4 = 0.25$, meaning that a quarter of galled, two-recombination trees are fully detectable when the mutation rate is very large. In comparison, this probability is closer to 0.7 when conditioning on one recombination. Figure 11 (b) demonstrates again that allowing a small number of unresolved edges increases the probabilities substantially, but they still remain very low.

These results imply that even if an ARG reconstruction algorithm utilises all of the available sequencing data, there is still likely to be significant uncertainty in resolving the location of internal edges. This probability only decreases with increasing recombination rate, and improves very slowly with increasing mutation rate. This makes it very unlikely that reconstruction programs will successfully capture the full complexity of the ARG.

The parameter values $\theta \approx 100$ and $\rho \approx 0.1$, reasonable for SARS-CoV-2, might appear to be optimal for creating genealogies that are fully recoverable from the data: low recombination rates increase the probability of seeing a small number of galled cycles, and high mutation rates make it more likely that mutations will fall on all of the necessary edges. However, our results show that at most 30% of one-recombination, and 10% of two-recombination ARGS can be reconstructed fully.

4. Detectability of gene conversion

We have so far focused on crossover recombination events, where there is one breakpoint in the genome with genetic material on either side taken from the left or right parent. Gene conversion is also important in shaping genetic variation, although has been investigated less thoroughly (Song, Ding, et al. 2008). Figure 12 illustrates the key difference between crossover recombination (left panel) and gene conversion events (right panel). Genetic material ancestral to the orange section is taken from the right parent and material ancestral to the purple section from the left. In biological samples, the conversion tract (orange) is typically small compared to the length of the genome.

Figure 12. Two parent genomes undergoing crossover recombination with breakpoint at position z_1 (left panel), and gene conversion with conversion tract between positions z_1 and z_2 (right panel).

In this section, conditioning on a single gene conversion event in the history of the sample, we calculate the probability that this event is detectable (without the requirement that the rest of the ARG topology is recovered fully).

Let ρ be the population scaled rate of gene conversion. Similarly to the case of crossover recombination, a gene conversion is detectable if two pairs of sites spanning the two breakpoints are incompatible. Label the sections of the genome undergoing the gene conversion $[0, z_1)$, (z_1, z_2), $(z_2, 1]$ as A, B, C, respectively. Figure 13 demonstrates possible configurations of events inside the gene conversion loop. Following similar arguments to those of Myers (2003) for the case of a single recombination, if one of the three possibilities illustrated in Figure 13 appears as a subgraph of the ARG, the gene conversion is guaranteed to be detectable. Note that there is some flexibility in the arrangement of events, as the positions of A and C can be interchanged, and additional coalescence events can be added to the recombination loop. Note also that the sub-graphs corresponding to $[0, z_2)$ and $(z_1, 1]$ each must have one of the configurations given in Figure 9.

The recursion relations for this scenario take a similar form to those described in Section 3.2.1. They therefore condition on a single gene conversion in the history, occurring

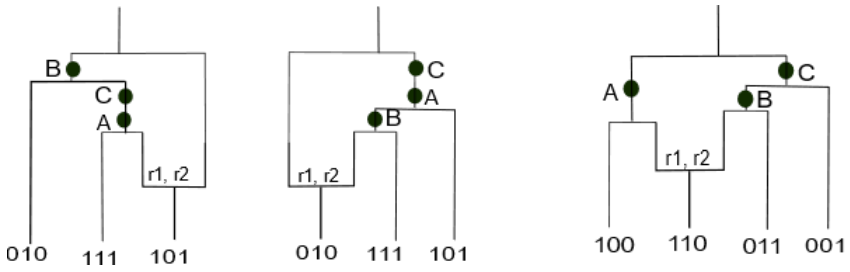


Figure 13. Conditions for a gene conversion to be detectable. Gene conversion nodes are labelled with the breakpoint positions. Each mutation is labelled by the section of the genome on which it must occur.

while k lineages are present. Let $p_{i,j}^{n_l}$ be the probability that the gene conversion will be detectable, conditional on n_l lineages currently in the ARG. Similarly to Myers (2003), the recursion only considers the events subsequent to the gene conversion (backwards in time), so n_l will include the two recombinant lineages. As we assume only one gene conversion event occurs in the history, it is more efficient to let index i track the state of the left recombinant lineage $\mathcal{E}$ as it undergoes mutations and coalescences, and likewise j for the right recombinant lineage $\mathcal{F}$. The required states are detailed in full in the Appendix, Tables 3 and 4. Unlike the case of crossover recombination, there is now a broken symmetry as two mutations on the outside (purple) parts of the genome are needed, and only one on the conversion tract (orange).

The full system of recursions is included in the Appendix, Section A.3. Each equation takes the form

$$\left(\binom{n_l}{2} + g(\theta) + \rho\frac{n_l}{2}\right)p_{i,j}^{n_l} = \binom{n_l-1}{2}p_{i,j}^{n_l-1} + (n_l-2)p_{i',j'}^{n_l-1} + \sum_{i',j'} g_{i',j'}(\theta)p_{i',j'}^{n_l}, \qquad (9)$$

where $g_{i,j}(\theta)$ are linear functions of θ which are different for each pair (i,j), and $g_{i,j}(\theta) = \sum_{i',j'} g_{i',j'}(\theta)$. These equations are formed by considering the next state that could be reached in the ARG, and applying the law of total probability. As some events will not change the state of the ARG, the recursive equations can be expressed as

$$\begin{aligned}&(\text{total rate of events that change the ARG}) \cdot p_{i,j}^{n_l} = \\ &\sum_{i',j'}\left([\text{rate of event that results in transition from states } (i,j) \to \text{states } (i',j')] \cdot p_{i',j'}^{n_l'}\right).\end{aligned}$$

Events which change the state of the ARG include coalescences and mutations (as the position of the gene conversion event is separately conditioned upon). Coalescence events always reduce n_l by one, and occur at rate $\binom{n_l}{2}$. These may involve only non-recombinant lineages, which will not alter the states i, j. If a recombinant lineage is involved, it may change state, from State $0 \to 1$ or from State $3 \to 4$ (see the Appendix for definitions of individual states); if the recombinant lineage is not in State 0 or 3, then the states i, j remain unchanged, but n_l decreases by 1. Mutation events may only alter the ARG's state if they occur on a recombinant lineage which is in State 1, 2, 4 or 5. The rate of an A-type mutation is $g_{i',j'}(\theta) = \theta_A/2 := z_1 \cdot \theta/2$, and likewise for B-type and C-type. A mutation event on a lineage in one of these states increases the state by 1.

The recursive system is solved to find $p_{0,0}^k$, and summing over k to remove the conditioning on when the gene conversion event occurs gives the desired full probability, $\mathcal{P}$, that a detectable gene conversion events occurs, conditional on precisely one such event

in the history. We have

$$\mathcal{P} = \frac{\sum_{k=2}^{n}\left(\prod_{l=k}^{n}\frac{l-1}{l-1+\rho}\right)\frac{\rho}{k-2+\rho}\cdot p_{0,0}^{k}}{\sum_{k=2}^{n}\left(\prod_{l=k}^{n}\frac{l-1}{l-1+\rho}\right)\frac{\rho}{k-2+\rho}\left(\prod_{l=2}^{k}\frac{l-1}{l-1+\rho}\right)} = \frac{\sum_{k=2}^{n}\left(\prod_{l=2}^{k-1}\frac{l-1}{l-1+\rho}\right)\frac{\rho}{k-2+\rho}\cdot p_{0,0}^{k}}{\sum_{k=2}^{n}\frac{\rho}{k-2+\rho}\frac{k-1}{k-1+\rho}}. \quad (10)$$

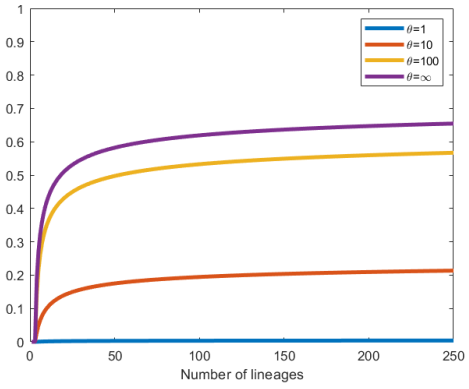
This is constructed as follows. Note that $p_{0,0}^{k}$ describes the state of an ARG directly after the gene conversion (looking backwards in time), which took the number of lineages from $k-1$ to k. Therefore, a detectable history with the gene conversion starting from a sample size n will have a sequence of $n-k$ coalescences, followed by the gene conversion event, followed by sufficient subsequent changes of state to allow the gene conversion to be detectable. These events, respectively, have probabilities

$$\prod_{l=k}^{n}\frac{l-1}{l-1+\rho}, \quad \frac{\rho}{k-2+\rho}, \quad p_{0,0}^{k}.$$

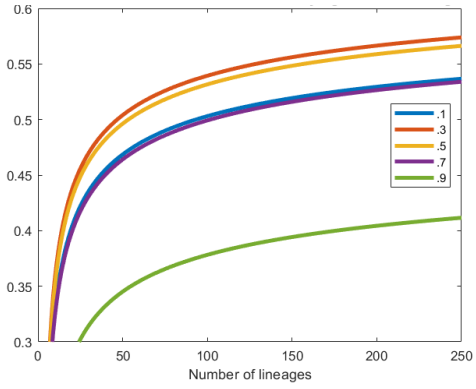
The denominator, conditioning on a single gene conversion, is constructed in a similar way, requiring first $n-k$ coalescent events followed by the gene conversion event. However, after the gene conversion event only a further k coalescences are required, with probability

$$\prod_{l=k}^{n}\frac{l-1}{l-1+\rho}.$$

Figure 14 (a) shows that detection probabilities for gene conversion events behave very similarly to those derived by Myers (2003, p.54) for detecting a single recombination, though are consistently slightly lower, as the gene conversion requires more mutation events to be detectable. In Figure 14 (b), the length of the conversion tract is varied; for scenarios where either the conversion tract, or its complement, is particularly short, the probability of detection decreases, as there is a lower probability of a mutation falling on the shorter section. As the mutation rate is assumed to be uniform across the genome, a conversion length of 1/3 gives the highest probabilities of detection. Note that the asymptotic probability as $\theta \to \infty$ tends to the probability that a single recombination changes the ARG topology (and hence agrees in the limit $\theta \to \infty$ with the probability given by Myers).



(a) Varying n for several fixed values of θ (colours), with breakpoints at $z = 0.33, 0.67$



(b) $\theta = 100$, varying the length of the converted section. Note the graph has been magnified for clarity. Mutation rate is uniform over the genome, and the conversion tract is centred about 0.5.

Figure 14. Probability of gene conversion being detectable.

5. Discussion

In this article, we have calculated the probability of recovering the tree or ARG topology under the coalescent with recombination, when the ARG topology is constrained to be in the shape of a galled tree. Galled trees have several attractive combinatorial and algorithmic properties that do not hold for general ARGs. For instance, there exists a polynomial time algorithm for reconstructing a parsimonious galled tree from sequencing data, if this is possible (Gusfield et al. 2004; Wang et al. 2001); there is a concise necessary and sufficient condition for the sample to be consistent with a galled tree (Song 2006); if a genealogy is in the shape of a galled tree, the sample can also be derived on a true tree (with no recombination) if at most one recurrent mutation per site is allowed (Gusfield 2014, Theorem 8.12.1). We have explicitly calculated the probability of an ARG being a galled tree, shedding light on how applicable these results might be in the analysis of real data. Our results indicate that genealogies in the form of galled trees are reasonably likely to be seen for $\rho < 1$ with moderate sample size.

Our results can also shed light on some theoretical properties of genealogical reconstruction algorithms. While some recently developed methods can handle impressive quantities of sequencing data, they are based on heuristic methods, making it difficult to obtain theoretical insights into their performance. In particular, while tsinfer retains polytomies (i.e. nodes with more than two child lineages) where the order of coalescence events cannot be resolved unambiguously, many other algorithms resolve the order of events randomly. Our results give a sense of how many such polytomies might be present in the history of a dataset, and how likely recombination events are to be detectable for a given value of evolutionary parameters. This provides an upper bound on how well genealogies can be reconstructed, even if the algorithm utilises all of the available sequencing data to the fullest extent.

In the absence of recombination, our results demonstrate that allowing a small number of unresolved internal edges can greatly improve the probability of reconstructing the rest of the tree correctly. This suggests that, for certain values of the parameters, there are likely to be a relatively small number of edges in the genealogy which are not supported by mutations and could be placed at many plausible positions.

For large sample sizes, the probability of recovering the tree or ARG topology with a high level of certainty is minuscule, for reasonable values of the mutation rate (such as those estimated for SARS-CoV-2). This strengthens the case for using Bayesian methods to integrate over the uncertainty of branch placements, or utilising additional data to resolve ambiguous event ordering. For instance, Ramazzotti et al. (2021) analysed variant frequencies using SARS-CoV-2 intra-host sequencing data, in order to resolve the ordering of transmission events in genealogies built using consensus sequences (i.e. at the level of one sequence per infected host).

For tractability, our analysis has focused on the particular case where the ARG topology is that of a galled tree, under the coalescent with recombination. A natural extension of this work would be to consider general ARGs and other models, with more complex scenarios that might include multiple loci or non-constant population size.

Acknowledgements

We thank Paul Jenkins for useful comments. This work was supported by the EPSRC and MRC OxWaSP Centre for Doctoral Training (EPSRC grant EP/L016710/1), and by the Alan Turing Institute (EPSRC grant EP/N510129/1).

References

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the archeological and historical sciences* (pp. 387–395).

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. *PLoS Genetics*, *8*(12), e1003090.

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution*, *6*(2), veaa061.

Fu, Y.-X., & Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*(3), 693–709.

Griffiths, R. C., & Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly & S. Tavare (Eds.), *Progress in population genetics and human evolution* (pp. 257–270). Springer.

Gusfield, D. (2014). *Recombinatorics: The algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT press.

Gusfield, D., Eddhu, S., & Langley, C. (2004). Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, *2*(01), 173–213.

Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, *111*(1), 147–164.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, *51*(9), 1330–1338.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y. et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, *382*, 1199–1207.

Müller, N. F., Kistler, K. E., & Bedford, T. (2021). Recombination patterns in coronaviruses. *bioRxiv*. https://doi.org/10.1101/2021.04.28.441806

Myers, S. (2003). *The detection of recombination events using DNA sequence data* (Doctoral dissertation). University of Oxford, Department of Statistics.

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., & Piazza, R. (2021). Verso: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. *Patterns*, *2*(3), 100212.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, *10*(5), e1004342.

Song, Y. S. (2006). A concise necessary and sufficient condition for the existence of a galled-tree. *IEEE/ACM transactions on computational biology and bioinformatics*, *3*(2), 186–191.

Song, Y. S., Ding, Z., Gusfield, D., Langley, C. H., & Wu, Y. (2008). Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of SNP sequences in populations. *Journal of Computational Biology*, *14*(10), 1273–86.

Song, Y. S., & Hein, J. (2005). Constructing minimal ancestral recombination graphs. *Journal of Computational Molecular Cell Biology*, *12*(2), 147–169.

Speidel, L., Forest, M., Shi, S., & Myers, S. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329.

Wang, L., Zhang, K., & Zhang, L. (2001). Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, *8*(1), 69–78.

Appendix A. Supplementary Material

A.1. **Model for full knowledge of the ARG.** This section presents the recursion equations for the probability of fully recovering the ARG topology. If a uniform mutation rate along the genome is assumed, with total mutation rate $\theta/2$ and a breakpoint at position $z \in [0,1]$; then type A mutations occur at rate $\theta_A/2 := z \cdot \theta/2$ and type B with rate $\theta_B/2 := (1-z)\cdot\theta/2$. It should be noted that the specific placement of the mutation within each locus does not matter.

For the sake of clarity, the subscript indices are contracted so that only those changing at each step are shown. For instance,

$$\tilde{p}^{n_0,n_l-1,n_f-1,r}_{i-1,j+1} = p^{n_0,n_l-1,n_f-1,r}_{i,j,k_1,k_2,l_1,l_2,m_1,m_2,e,a,b_1,b_2,c_1,c_2,d_1,d_2},$$

and $\tilde{p}^{n_0,n_l,n_f+1,r}$ indicates that none of the recombinant states have changed.

The index r tracks the number of open recombination loops, so there is a restriction

$$r = i + j + k_1 + k_2 + l_1 + l_2 + m_1 + m_2 = e + a + b_1 + b_2 + c_1 + c_2 + d_1 + d_2,$$

with $n \geq 2r$. Any values of $\tilde{p}$ out of these bounds are immediately set to 0.

For clarity, the main equation is broken up into several parts. As stated in the main text, the form of the equation is

$$\begin{aligned}&(\text{total rate of moves that change ARG state}) \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r} = \\ &\sum [\text{rate of event that results in transition from states } (i,j)] \cdot (p \text{ of resultant state}).\end{aligned}$$

The total rate of moves that change the state of the ARG is

$$\begin{aligned}Rate = \Bigg(&\binom{n_l}{2} + \frac{\rho n_l}{2} + \frac{\theta}{2}(n_l - n_f - 2r) \\ &+ \frac{\theta_A}{2}(j + k_2 + l_1 + a + c_1 + d_2) + \frac{\theta_B}{2}(j + l_1 + m_2 + a + b_2 + c_1)\Bigg)\end{aligned}$$

Moves that change the state of the ARG include the following event types.

(1) Coalescences of non-recombinant lineages, with rate

$$C_{NR} = \binom{n_f}{2}\tilde{p}^{n_0,n_l-1,n_f-2,r} + n_f(n-2r-n_f)\tilde{p}^{n_0-1,n_l-1,n_f-1,r} + \binom{n-2r-n_f}{2}\tilde{p}^{n_0-2,n_l-1,n_f,r}.$$

(2) The first mutation of a non-recombinant lineage since its last coalescence,

$$M_{NR} = (n - n_f - 2r)\frac{\theta}{2}\tilde{p}^{n_0,n_l,n_f+1,r}.$$

(3) A coalescence of one recombinant lineage, and one non-recombinant (taking care to distinguish whether the non-recombinant lineage has had a mutation since its

last coalescence),

$$
\begin{aligned}
C_R = {} & i \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{i-1,j+1} + i \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{i-1,j+1} + e \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{e-1,a+1} \\
& + e \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{e-1,a+1} + k_1 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{k_1-1,l+1} + k_1 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{k_1-1,l+1} \\
& + b_1 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{b_1-1,c+1} + b_1 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{b_1-1,c+1} + k_2 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{k_1+1,k_2-1} \\
& + k_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{k_1+1,k_2-1} + b_2 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{a+1,b_2-1} + b_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{a+1,b_2-1} \\
& + m_2 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{l_1+1,m_2-1} + m_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{l_1+1,m_2-1} + d_2 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{c_1+1,d_2-1} \\
& + d_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{c_1+1,d_2-1} + m_1 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{l_1+1,m_1-1} + m \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{l_1+1,m_1-1} \\
& + d_1 \cdot n_f \cdot \tilde{p}^{n_0,n_l-1,n_f-1,r}_{c_1+1,d_1-1} + d_1 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r}_{c_1+1,d_1-1} + j \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} \\
& + j \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r} + a \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} + a \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r} \\
& + l_1 \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} + l_1 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r} + l_2 \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} \\
& + l_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r} + c_1 \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} + c_1 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r} \\
& + c_2 \cdot n_f \cdot \tilde{p}^{n_0-1,n_l-1,n_f-1,r} + c_2 \cdot (n-2r-n_f) \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r}.
\end{aligned}
$$

(4) A mutation event that changes the state of a recombinant lineage,

$$
\begin{aligned}
M_R = {} & j \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{j-1,k_1+1} + a \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{a-1,b_1+1} + k_2 \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{k_1+1,k_2-1} + b_2 \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{b_1+1,b_2-1} + j \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{j-1,k_2+1} \\
& + a \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{a-1,b_2+1} + l_1 \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{l_1-1,m_1+1} + c_1 \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{c_1-1,d_1+1} + m_2 \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{m_1+1,m_2-1} + d_2 \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{d_1+1,d_2-1} \\
& + l_1 \cdot \frac{\theta_A}{2} \tilde{p}^{n_0,n_l,n_f,r}_{l_1-1,m_2+1} + c_1 \cdot \frac{\theta_B}{2} \tilde{p}^{n_0,n_l,n_f,r}_{c_1-1,d_2+1} + l_2 \cdot \frac{\theta}{2} \tilde{p}^{n_0,n_l,n_f,r}_{l_1-1,m_2+1} + c_2 \cdot \frac{\theta}{2} \tilde{p}^{n_0,n_l,n_f,r}_{c_1-1,d_2+1}.
\end{aligned}
$$

(5) The coalescence of two recombinant lineages, which for a galled tree must be the result of an open recombination loop closing. This requires a factor of $1/r$ in the probabilities, as each left recombinant lineage must choose to coalesce with its partner out of the r possible right recombinant lineages available,

$$
\begin{aligned}
C_{RR} = {} & \frac{l_2}{r} \cdot \Big(e \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,e-1} + a \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,a-1} + b_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,b_1-1} + b_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,b_2-1} \\
& + c_1 \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r-1}_{l_2-1,c_1-1} + c_2 \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r-1}_{l_2-1,c_2-1} + d_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,m_2-1} + d_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2-1,d_1-1} \Big) \\
& + \frac{m_1}{r} \cdot \Big(e \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,e-1} + a \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,a-1} + b_1 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,b_1-1} + b_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,b_2-1} \\
& + c_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{m_1-1,c_1-1} + c_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{m_1-1,c_2-1} + d_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,m_2-1} + d_1 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_1-1,d_1-1} \Big) \\
& + \frac{d_1}{r} \cdot \Big(i \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{i-1,d_1-1} + j \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{j-1,d_1-1} + k_1 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{k_1-1,d_1-1} + k_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{k_2-1,d_1-1} \\
& + m_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_2-1,d_1-1} + + l_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_1-1,d_1-1} \Big) + \frac{c_2}{r} \cdot \Big(i \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{i-1,c_2-1} + j \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{j-1,c_2-1} \\
& + k_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{k_1-1,c_2-1} + k_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{k_2-1,c_2-1} + m_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_2,m_2-1} \\
& + l_1 \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r-1}_{l-1,c_2-1} \Big) + \frac{k_1}{r} \cdot \Big(b_1 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{k_1-1,b_1-1} + d_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{k_1-1,m_2-1} + c_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{k_1-1,c_1-1} \Big) \\
& + \frac{m_2}{r} \cdot \Big(b \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_2-1,b_1-1} + d_2 \cdot \tilde{p}^{n_0,n_l-1,n_f,r-1}_{m_2-1,d_2-1} + c_1 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{m_2-1,c_1-1} \Big) \\
& + \frac{l_1}{r} \cdot \Big(b \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_1-1,b_1-1} + d_2 \cdot \tilde{p}^{n_0-1,n_l-1,n_f,r-1}_{l_1-1,m_2-1} + c_1 \cdot \tilde{p}^{n_0-2,n_l-1,n_f,r-1}_{l_1-1,c_1-1} \Big).
\end{aligned}
$$

(6) The opening of a new recombination loop,

$$R = \frac{\rho}{2}\Big((n - n_f - 2r)\cdot \tilde{p}_{i+1,e+1}^{n_0,n+1,n_f,r+1} + n_f \cdot \tilde{p}_{i+1,e+1}^{n_0,n+1,n_f-1,r+1}\Big). \tag{11}$$

Then the full equation can be expressed as

$$Rate \cdot \tilde{p}^{n_0,n_l,n_f,r} = C_{NR} + C_R + C_{RR} + M_{NR} + M_R + R. \tag{12}$$

Boundary data is given by $p_{0,0,0,0,0,0,0,0,0,0,0,0,0,0}^{2,n_f,0} = 1$ for $n_f = 0, 1, 2$ and 0 otherwise. This is solved by first setting $r = R$, $n = 2$ and iterating over the recombinant lineage state indices in reverse alphabetical order, then iterating forward in n until the sample size is reached, then backwards in r.

The computation time can be estimated using simple reasoning, despite some of the recursions looking quite complex. If a quantity is recursively defined using k integer arguments, and evaluation of the quantity for fixed values of the k arguments $(m_1, m_2, .., m_k)$ needs evaluation of some function $g(m_1, m_2, .., m_k)$, then two questions need to be considered: how many different arguments are there, and for each, what is g? Suppose we have a lower triangular matrix in k dimensions, and for each argument we need to evaluate all arguments smaller in each argument, then computation time will grow as k^2'th power. If we only need to refer to arguments smaller by a constant number, then it grows as k'th power.

The computation time needed to evaluate these recursions is of the order of $n_0 \cdot n^2 \cdot R^{17}$ where R is the total number of recombinations in the history. This quickly becomes unfeasibly large, but the restriction

$$r = i + j + k_1 + k_2 + l_1 + l_2 + m_1 + m_2 = e + a + b_1 + b_2 + c_1 + c_2 + d_1 + d_2,$$

can be exploited to significantly reduce the computation time. If $\mathcal{B}(r)$ is the number of tuples of eight non-negative integers that sum to r, then $\mathcal{B}(1, 2, 3, 4, 5) = (8, 36, 120, 330, 792)$, a major decrease from r^{17}. Exploiting this gives a reduced computational time of the order of $n_0 \cdot n^2 \cdot \mathcal{B}(R)^2$.

A.2. **Probability of topology known, varying breakpoint.**

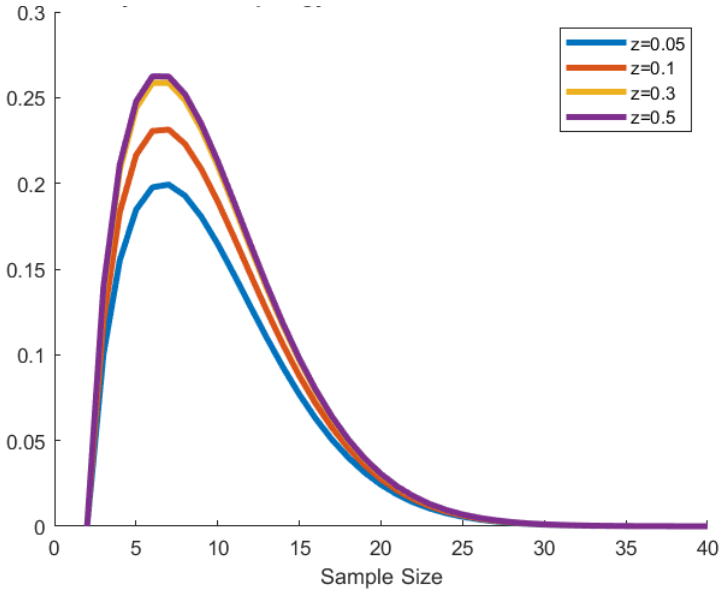


Figure 15. Fixed $\theta = 100$, varying breakpoint position z across $[0, 0.5]$.

A.3. **Detection of a gene conversion.** As observed in the main section, there is a certain amount of flexibility as to the order of the A- and C-type mutations in the history. Therefore, conditioning on the order of these mutations on both the $\mathcal{E}$ and $\mathcal{F}$ lineages is required. Under the uniform mutation rate assumption, with mutations occurring as competing Poisson processes, the probability of a type A mutation occurring before a type C is $\theta_A/(\theta_A+\theta_C)$, where as before $\theta_A = \theta \cdot \text{length}(A)$. Events on distinct lineages are independent.

Due to the breakdown of symmetry, the states for each lineage are given separately.

Table 3. States described for recombinant edge $\mathcal{E}$

State 0	No coalescence has occurred since the recombination.
State 1	There has been at least one coalescence since the recombination. No mutations have occurred since the last coalescence.
State 2	The first of the A/C-type mutations has occurred since the last coalescence.
State 3	The second of the A/C-type mutations has occurred since the last coalescence. This mutation must be different to the previous mutation in state 2.
State 4	$\mathcal{E}$ has reached state 3, and undergone one further coalescence.
State 5	Type B mutation has occurred since the last coalescence.

Table 4. States described for recombinant edge $\mathcal{F}$

State 0	No coalescence has occurred since the recombination.
State 1	There has been at least one coalescence since the recombination. No mutations have occurred since the last coalescence.
State 2	A B-type mutations has occurred since the last coalescence.
State 4	$\mathcal{F}$ has reached state 3, and undergone one further coalescence.
State 5	The first of the A/C-type mutations has occurred since the last coalescence.
State 6	The second of the A/C-type mutations has occurred since the last coalescence. Again this mutation must be different to the previous mutation in state 5.

Note that due to the choice of state labels, $\mathcal{F}$ does not have a State 3 equivalent. Again, we use the phrasing that the ARG being in state (i,j) means $\mathcal{E}$ is in state i and $\mathcal{F}$ is in state j.

The recombination will be detectable if $\mathcal{E}$ reaches state 5, or $\mathcal{F}$ reaches state 6, or $\mathcal{E}$ is in a state >2 and $\mathcal{F}$ in a state >1. If the ARG reaches one of these absorbing states, the subsequent probability of detection is given by q^n, the probability of no further gene conversion events in the sample. We have $q^n = \prod_{m=2}^{n}(m-1)/(m-1+\rho)$.

Denote the first of the A or C type mutations on lineage $\mathcal{E}$ (resp. $\mathcal{F}$) as l_1 (resp. r_1) and the second as l_2 (resp. r_2).

If $l_2 \neq r_2$, the ARG in state $(2,5)$ has the recombination detectable immediately, i.e. $p^n_{2,5} = q^n$. If $l_2 = r_2$, we have the relation

$$\left(\binom{n}{2} + \frac{\theta_{l_2}+\theta_{r_2}}{2} + \frac{\rho n}{2}\right)p^n_{2,5} = \left(\binom{n}{2} - 1\right)p^{n-1}_{2,5} + \frac{\theta_{r_2}}{2}q^n + \frac{\theta_{l_2}}{2}q^n,$$

and for every combination of l_i, r_i:

$$\left(\binom{n}{2} + \frac{\theta_{l_2} + \theta_{r_1}}{2} + \frac{\rho n}{2}\right) p^n_{2,4} = \left(\binom{n}{2} - 1\right) p^{n-1}_{2,4} + \frac{\theta_{r_1}}{2} p^n_{2,5} + \frac{\theta_{l_2}}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\theta_{l_2}}{2} + \frac{\rho n}{2}\right) p^n_{2,2} = \binom{n-1}{2} p^{n-1}_{2,2} + (n-2) p^{n-1}_{2,4} + \frac{\theta_{l_2}}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\theta_{l_1} + \theta_{r_2}}{2} + \frac{\rho n}{2}\right) p^n_{1,5} = \left(\binom{n}{2} - 1\right) p^{n-1}_{1,5} + \frac{\theta_{r_2}}{2} q^n + \frac{\theta_{l_1}}{2} p^n_{2,5}$$

$$\left(\binom{n}{2} + \theta_B + \frac{\rho n}{2}\right) p^n_{4,1} = \left(\binom{n}{2} - 1\right) p^{n-1}_{4,1} + \theta_B q^n$$

$$\left(\binom{n}{2} + \frac{\theta_{l_1} + \theta_{r_1}}{2} + \frac{\rho n}{2}\right) p^n_{1,4} = \left(\binom{n}{2} - 1\right) p^{n-1}_{1,4} + \frac{\theta_{r_1}}{2} p^n_{1,5} + \frac{\theta_{l_1}}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\theta_B}{2} + \frac{\rho n}{2}\right) p^n_{3,1} = \binom{n-1}{2} p^{n-1}_{3,1} + (n-2) p^{n-1}_{4,1} + \frac{\theta_B}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\theta_{l_1}}{2} + \frac{\rho n}{2}\right) p^n_{1,2} = \binom{n-1}{2} p^{n-1}_{1,2} + (n-2) p^{n-1}_{1,4} + \frac{\theta_{l_1}}{2} p^n_{2,2}$$

$$\left(\binom{n}{2} + \frac{\theta_B + \theta_{r_2}}{2} + \frac{\rho n}{2}\right) p^n_{2,1} = \left(\binom{n}{2} - 1\right) p^{n-1}_{2,1} + \frac{\theta_{r_2}}{2} p^n_{3,1} + \frac{\theta_B}{2} p^n_{2,2}$$

$$\left(\binom{n}{2} + \frac{\theta_B + \theta_{l_1}}{2} + \frac{\rho n}{2}\right) p^n_{1,1} = \left(\binom{n}{2} - 1\right) p^{n-1}_{1,1} + \frac{\theta_{l_1}}{2} p^n_{2,1} + \frac{\theta_B}{2} p^n_{1,2}$$

$$\left(\binom{n}{2} + \frac{\theta_{r_2}}{2} + \frac{\rho n}{2}\right) p^n_{0,5} = \binom{n-1}{2} p^{n-1}_{0,5} + (n-2) p^{n-1}_{1,5} + \frac{\theta_{r_2}}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\theta_{r_1}}{2} + \frac{\rho n}{2}\right) p^n_{0,4} = \binom{n-1}{2} p^{n-1}_{0,4} + (n-2) p^{n-1}_{1,4} + \frac{\theta_{r_1}}{2} p^n_{0,5}$$

$$\left(\binom{n}{2} + \frac{\theta_B}{2} + \frac{\rho n}{2}\right) p^n_{4,0} = \binom{n-1}{2} p^{n-1}_{4,0} + (n-2) p^{n-1}_{4,1} + \frac{\theta_B}{2} q^n$$

$$\left(\binom{n}{2} + \frac{\rho n}{2}\right) p^n_{3,0} = \binom{n-2}{2} p^{n-1}_{3,0} + (n-2)(p^{n-1}_{4,0} + p^{n-1}_{3,1})$$

$$\left(\binom{n}{2} + \frac{\rho n}{2}\right) p^n_{0,2} = \binom{n-2}{2} p^{n-1}_{0,2} + (n-2)(p^{n-1}_{0,4} + p^{n-1}_{1,2})$$

$$\left(\binom{n}{2} + \frac{\theta_{l_2}}{2} + \frac{\rho n}{2}\right) p^n_{2,0} = \binom{n-1}{2} p^{n-1}_{2,0} + (n-2) p^{n-1}_{2,1} + \frac{\theta_{l_2}}{2} p^n_{3,0}$$

$$\left(\binom{n}{2} + \frac{\theta_{l_1}}{2} + \frac{\rho n}{2}\right) p^n_{0,1} = \binom{n-1}{2} p^{n-1}_{0,1} + (n-2) p^{n-1}_{1,1} + \frac{\theta_{l_1}}{2} p^n_{0,2}$$

$$\left(\binom{n}{2} + \frac{\theta_B}{2} + \frac{\rho n}{2}\right) p^n_{1,0} = \binom{n-1}{2} p^{n-1}_{1,0} + (n-2) p^{n-1}_{1,1} + \frac{\theta_B}{2} p^n_{2,0}$$

$$\left(\binom{n}{2} + \frac{\rho n}{2}\right) p_{0,0}^{n} = \binom{n-2}{2} p_{0,0}^{n-1} + (n-2)(p_{0,1}^{n-1} + p_{1,0}^{n-1})$$