OXFORD GENETICS

GENETICS, 2026, 232(1), iyaf103
https://doi.org/10.1093/genetics/iyaf103
Advance Access Publication Date: 29 May 2025
Investigation

Likelihoods for a general class of ARGs under the SMC

Gertjan Bisschop [1,*] Jerome Kelleher [1] Peter Ralph [2]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7LF, UK
[2]Institute of Ecology and Evolution, University of Oregon, 101C McArthur Court, Eugene, OR 97403, USA

*Corresponding author: Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7LF, UK.
Email: gertjanbisschop@gmail.com

Ancestral recombination graphs (ARGs) are the focus of much ongoing research interest. Recent progress in inference has made ARG-based approaches feasible across of range of applications, and many new methods using inferred ARGs as input have appeared. This progress on the long-standing problem of ARG inference has proceeded in two distinct directions. First, the Bayesian inference of ARGs under the Sequentially Markov Coalescent (SMC), is now practical for tens-to-hundreds of samples. Second, approximate models and heuristics can now scale to sample sizes two to three orders of magnitude larger. Although these heuristic methods are reasonably accurate under many metrics, one significant drawback is that the ARGs they estimate do not have the topological properties required to compute a likelihood under models such as the SMC under present-day formulations. In particular, heuristic inference methods typically do not estimate precise details about recombination events, which are currently required to compute a likelihood. In this article, we present a backwards-time formulation of the SMC (conventionally regarded as an along-the-genome process) and derive a straightforward definition of the likelihood of a general class of ARG under this model. We show that this formulation does not require precise details of recombination events to be estimated, and is robust to the presence of polytomies. We discuss the possibilities for ARG inference that this new formulation opens.

Keywords: ancestral recombination graphs; coalescent; SMC

Introduction

Following recent breakthroughs in simulation and inference methods, there is now strong interest in applying Ancestral Recombination Graphs (ARGs) to a variety of questions in genetics (Brandt DY *et al.* 2024 ; Lewanski *et al.* 2024; Nielsen *et al.* 2025). ARGs describe the intricately linked paths of genetic inheritance resulting from recombination (Hudson 1983; Griffiths and Marjoram 1996; Wong *et al.* 2024), and contain rich detail about ancestral processes. Numerous applications taking advantage of this detailed history in inferred ARGs are now appearing (Stern *et al.* 2019; Fan *et al.* 2022; Guo *et al.* 2022; Hejase *et al.* 2022; Ignatieva *et al.* 2022; Wang and Coop 2022; Ignatieva *et al.* 2023; Nowbandegani *et al.* 2023; Zhang *et al.* 2023; Deraje *et al.* 2024; Korfmann *et al.* 2024; Osmond and Coop 2024; Fan *et al.* 2025; Grundler *et al.* 2025; Huang *et al.* 2025; Speidel *et al.* 2025) and it seems likely that many more will follow (Harris 2019, 2023). Although there are many different approaches to ARG inference (Wong *et al.* 2024), two broad classes of method have been the focus of most recent interest.

The first class of inference methods are probabilistic approaches that sample ARGs under a population genetics model such as the coalescent with recombination (Hudson 1983), and its approximation, the Sequentially Markovian Coalescent (McVean and Cardin 2005; Marjoram and Wall 2006). ARGweaver (Rasmussen MD *et al.* 2014; Hubisz *et al.* 2020) has been the most widely used and studied, and until recently consistently out-performed other methods in terms of accuracy in a variety of benchmarks (Brandt D *et al.* 2022), and has successfully been applied in several contexts (e.g. De Manuel *et al.* 2016; Shriner and Rotimi 2018; Hejase *et al.* 2020; Stankowski *et al.* 2024). The recently-introduced method SINGER takes a similar approach, and by using an improved Monte Carlo algorithm promises to be even more accurate and substantially more scalable (Deng *et al.* 2024). The second class of inference methods that has been of recent interest are based on heuristics and approximate models. Relate (Speidel *et al.* 2019), tsinfer (Kelleher *et al.* 2019), ARG-Needle (Zhang *et al.* 2023) and Threads (Gunnarsson *et al.* 2024) work on quite different principles, but share some common properties. First, they are all heuristic and approximation driven, favoring computational scalability over a basis in a probabilistic model. Second, they all produce a single ARG as output, inferring a deterministic point estimate. Third, they regard the precise timing of nodes as a separate problem, focusing only on the relative ordering of nodes when producing the initial ARG. General purpose ARG dating methods are appearing (Wohns *et al.* 2022; Deng *et al.* 2025) and can also improve dating performance in ARGs sampled from the SMC (Deng *et al.* 2025). Fourth (and most importantly for the purposes of this article), the ARGs that heuristic methods produce lack some of the topological information present in the output of ARGweaver or SINGER. This occurs, roughly speaking, because of a basic difference in their approach to uncertainty.

There is considerable uncertainty in ARG inference, and often a fundamental lack of information in the sequence data to distinguish different possibilities. Consider, for example, a case in

Received on 24 February 2025; accepted on 19 May 2025

which we have three ancestral lineages with identical sequences. We know that there must have been two coalescence events, but there is no mutational information to distinguish their relative ordering. Sampling methods overcome this problem by randomly choosing an ordering, with the uncertainty communicated by the order changing in the different ARGs sampled from the posterior. The other approach (used by methods such as tsinfer) is to communicate the uncertainty structurally by means of a polytomy: we have no information about the intermediate coalescent event, and so we omit it. Similarly, there is often a fundamental lack of information about the details of recombination events, and heuristic methods omit these details in different ways. The significant progress made in scalability by these heuristic methods is in part because they don't attempt to precisely reconstruct recombination events.

There are substantial advantages to the statistically rigorous sampling approaches of ARGweaver and SINGER. In particular, they tend to be more accurate (at the scale at which they can be compared), and sampling ARGs from a posterior distribution provides a means of quantifying the uncertainty around point estimates. On the other hand, heuristic methods have the potential to use the greater amount of information in large datasets, capturing fine-scale details about the recent past. Although SINGER is a major step forward in terms of scalability over ARGweaver and other methods, inference is still only feasible for hundreds of samples. Meanwhile, tsinfer, ARG-Needle and Threads have been applied to datasets of hundreds of *thousands* of samples. With the explosive growth in dataset sizes in recent years (Caulfield *et al.* 2017; Bycroft *et al.* 2018; Turnbull *et al.* 2018; Backman *et al.* 2021; Halldorsson *et al.* 2022; UK Biobank Whole-Genome Sequencing Consortium 2023; All of Us Research Program Genomics Investigators 2024; Stark *et al.* 2024; Cook *et al.* 2025) there is a pressing need for more statistical rigor in large-scale ARG inference.

A key problem facing the field is that the ARGs estimated by heuristic methods lack a concrete connection to population genetic theory because we cannot compute their likelihood under the SMC or similar models. Computing the likelihood of an ARG under a model is a fundamental element of probabilistic inference. The classical Kuhner–Yamato–Felsenstein (2000) formulation (henceforth: KYF) works by assigning a probability to each recombination and common ancestor event, and to the inter-event waiting times. Although the KYF formulation is a natural and elegant way to describe the likelihood of an ARG, the high level of detail about the underlying events that is required is not present in many estimated ARGs (Wong *et al.* 2024) or even in most simulations. Simulators such as the classical ms program (Hudson 2002) output ARGs in a tree-by-tree format, which does not contain sufficient detail to reconstruct the "full ARG" required for likelihood computation (Wong *et al.* 2024). The msprime simulator (Kelleher *et al.* 2016; Baumdicker *et al.* 2022) by default outputs a more complete and efficient representation of the simulated ancestry, but still does not provide the exhaustive detail required. Indeed, providing the information required to support likelihood calculations originally motivated the addition of the `record_full_arg` option to msprime (Baumdicker *et al.* 2022). Forwards-time simulations that output ARGs (Haller *et al.* 2018; Kelleher *et al.* 2018) also do not typically contain the level of detail required to compute a likelihood under the KYF approach.

In this article, we provide a new formulation of ARG likelihood which has much more lenient input requirements than KYF, and gracefully handles incomplete and imprecise ARGs. To do this, we give a backwards-in-time description of the SMC which naturally leads to a straightforward method for computing the likelihood of any ARG, and describe an efficient algorithm to do so. We show that explicitly identifying recombination events via nodes in the graph is not necessary (from the perspective of computing the likelihood) as long as the "locally unary" (Fritze *et al.* 2024; Wong *et al.* 2024) sections of common ancestor nodes are retained. We also show via an empirical example that the likelihood is robust to the presence of polytomies and other imprecise information, and illustrate how it can be decomposed into separately evaluated time slices. Finally, we discuss how these results may be applied and extended.

Results

Ancestral recombination graphs

Although the term ARG has been used to mean several things, in the generic sense, it is a labelled graph structure that records the inheritance of genetic material (Wong *et al.* 2024). An ARG of the sort we consider describes the history of inheritance of the genomic material of a focal set of individuals, or *samples*. Concretely, an ARG is equivalent to a collection of (non-contradictory) statements of the form "c inherited from p on the genomic segment from x to y," where c and p are (haploid) genomes, and x and y are genetic coordinates. We record each such statement in an "edge," often summarized as the tuple (p, c, x, y). By "describes the history," we mean that each node in the ARG represents a specific genome, and that the ARG specifies the genealogical tree describing how the sampled genomes are related at each base pair.

An ARG can therefore be thought of either as a collection of inheritance relationships between sampled and ancestral haplotypes (haploid genomes), or as a sequence of genealogical trees that have common node labels. Figure 1 shows a small ARG that describes relationships between four samples on a 100 bp piece of genome, represented as both a graph (top) and a series of local trees (bottom). Both structures represent the same information: for instance, the fact that the ancestral genome at A inherited the left half of its genome from ancestor g and the right half from ancestor f is seen in the fact that the first two trees have A below g, and the remaining have A below f. The blue dotted line from A in the second tree to f in the third tree draws our attention to the fact that the path traced by A's ancestry switches at that point. In fact, A is in blue because it is a *recombinant* node—an ancestor that has a segment of genome inherited by a sample, but that segment was inherited in two pieces from their parent. Other nodes (labelled by black letters a-j) are either sample nodes or *common ancestry* nodes—ancestors that are the most recent common ancestor for two or more samples at some place along the genome.

Crucial for the equivalence between the two representations is the presence of nodes that only have a single child along one or more local trees—i.e. are *locally unary* (Wong *et al.* 2024). Recombinant nodes (in blue capital letters) have a single child in all trees. However, common ancestry nodes can be locally unary as well: for instance, sample d inherited a longer stretch of genome from ancestor f than sample c. In the top plot, this is visible because the pink bar below node f extends further to the left than does the blue bar, while in the bottom plot, this is visible because node f is parent to d in the second tree, but not to node c, and so is unary in the second tree.

These locally unary nodes enable us to uniquely identify all lineages that were hit by a recombination event, even if we simplify the ARG by removing all recombinant nodes (A, B, and C in Fig. 1). In the absence of these recombinant nodes, a

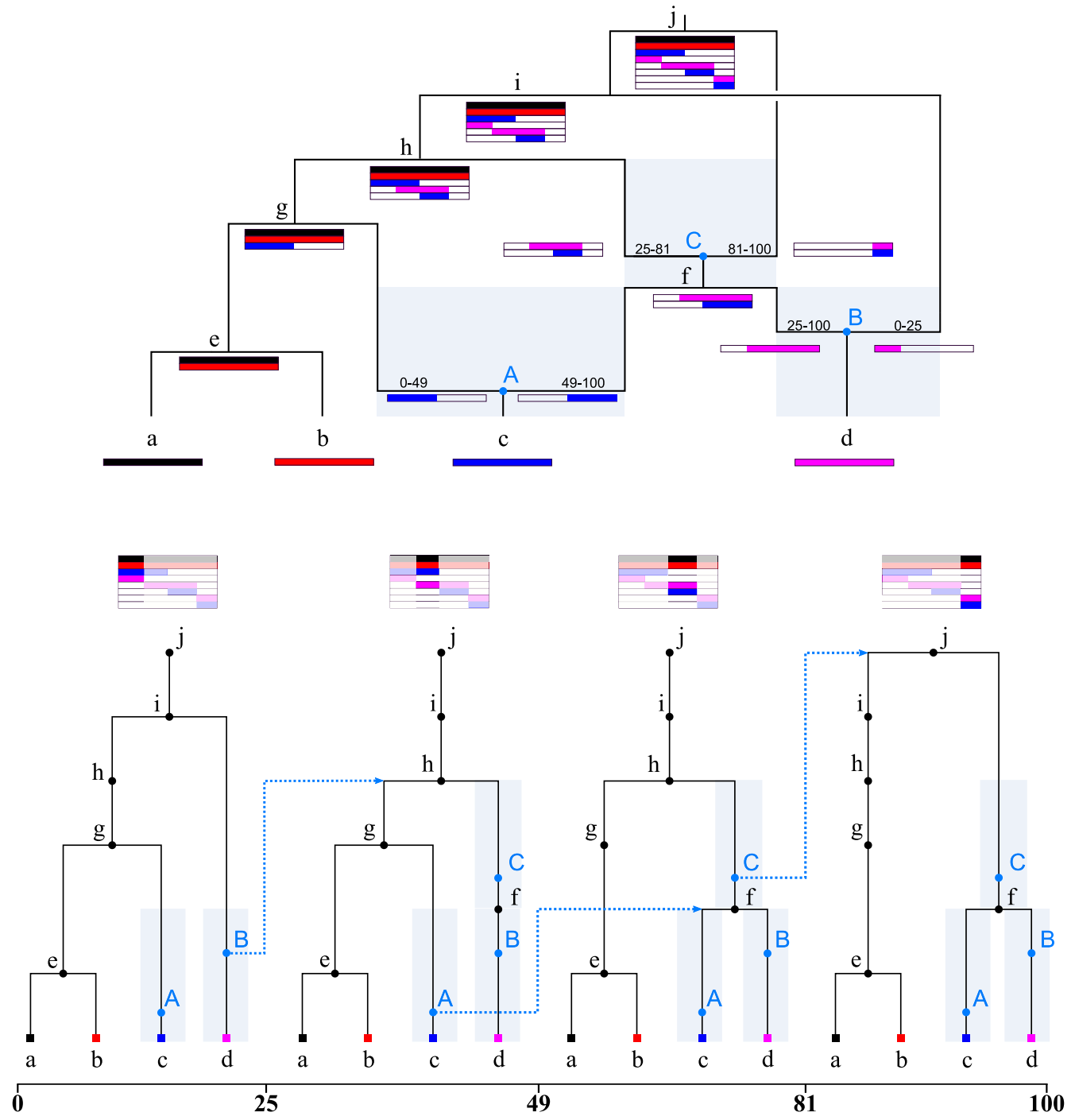

Fig. 1. An ARG represented as a graph (top) and as a series of local trees (bottom). To ensure the one-to-one correspondence between both representations a node is recorded for each local tree whenever a node is encountered along its ancestry path in the graph representation. This results in nodes that (locally) have only a single child, and are therefore (locally) unary. Nodes A, B, and C are recombinants, i.e. have two parents in the graph and are unary in the local trees. The shaded regions indicate the range of possible times associated with these recombinants. The dotted arrows (bottom) show the standard left-to-right logic of the SMC whereby the floating lineage coalesces randomly with the remaining portion of the local tree following recombination.

recombination can be observed easily enough as a change in parent going from one genomic region to the next. Without the recombinant nodes we no longer know the precise time of each recombination event, but we do still have constraints on it; for instance, we know that C happened sometime between the times of node f and node h. More generally, a recombinant node must happen in the time interval between the child node and the most recent of the parent nodes. This is again easier to visualize using the graph representation (shaded area, Fig. 1). Although these unary nodes are useful and to some degree clearly inferrable, they are not always represented by simulators or ARG inference methods. However, to our understanding this is because the field is used to thinking in terms of the trees rather than in terms of relationships between haplotypes, not

because of any question of knowability. See Fritze et al. (2024) for more discussion on unary nodes and haplotype aware ARG comparison metrics.

SMC backwards-in-time

The coalescent process describes the ancestry of a set of sampled genomes. The original algorithm to simulate the coalescent with recombination was formulated backwards-in-time by Hudson (1983). Wiuf and Hein later described a tree-by-tree method to simulate the same stochastic process (Wiuf and Hein 1999a). This along-the-genome approach is considerably more complex as the next tree to be simulated depends on all previous trees. This idea however provided the basis of the SMC (McVean and Cardin 2005), restricting the set of all state space transitions possible under the Hudson coalescent to obtain a process that is Markovian both backwards-in-time, as well as along-the-genome. In its backwards-in-time formulation, the SMC only requires a simple modification to the coalescent with recombination. Instead of allowing any pair of lineages to coalesce, the SMC is the process in which only lineages with *overlapping* genomic intervals may coalesce. The SMC' is a refinement of the SMC which also allows lineages with adjoining genomic intervals to coalesce (Marjoram and Wall 2006). Although McVean and Cardin (2005) verbally described the SMC using this backwards-in-time description, their analysis was almost entirely in left-to-right terms, and subsequent papers followed this trend (Li H and Durbin 2011; Paul et al. 2011; Rasmussen MD et al. 2014; Schiffels and Durbin 2014).

Linking the resulting tree-valued process along the genome back to the relationship between the "graph" and "local trees" representation of ARGs, the Markovian property can be formulated as follows. The distribution of the next tree within the series of local trees can be generated only knowing the current tree, provided that we remove recombinant nodes entirely, and erase the unary portions of coalescent nodes. To see that the locally unary nodes make the process non-Markov, consider that a node in a tree might be unary either because it's *about to be* (in the left-to-right sense) a coalescent node or it *already has been*: looking at the second tree in Fig. 1, we know node f will be coalescent further along the sequence because it was not present in the first tree. So, the first tree gives us information about subsequent trees that the second tree alone could not. Further note that by limiting common ancestry events to overlapping pairs of lineages, we are guaranteed that all nodes in the ARG only carry ancestral material, that is, genetic material that is inherited by either one of a set of samples (Wiuf and Hein 1999b). We therefore require a single half-open interval to describe the ancestral material associated with each node in the ARG.

We now define the process more formally, by reintroducing the notation of McVean and Cardin (2005) required to describe the likelihood computation. This section does not deviate from the original description of the SMC apart from how coalescable pairs are counted. This subtle nuance greatly simplifies the ability to keep track of coalescable pairs going backwards-in-time. At any point in time, the state of the process is the set of labelled lineages extant at time t, $L(t) = \{X_j(t)\}$. A lineage $X_j(t) = [x_j,\ y_j)$ is labelled by an integer j and defines the half-open genomic interval of its ancestral material.

Because we look backwards-in-time, $t = 0$ is today, and $L(0)$ consists of n sampled lineages, labelled 0 to $n-1$, each represented by a single interval spanning the entire genome. Then, the process evolves by a succession of common ancestor and recombination events until each segment of ancestral material is only present in one lineage. The waiting time to the next event is determined by these two competing processes with exponentially distributed waiting times as outlined below.

Recombination is described by a Poisson process of rate r per unit of genomic length and time: so, if $T(t) = \sum_j |X_j(t)|$ is the total length of ancestral material carried by lineages at time t, then recombination occurs at instantaneous rate $rT(t)$. A recombination to the left of x (i.e. between base pairs x and $x-1$) with $x_j < x < y_j$ splits lineage j into two new lineages, given new, unique labels. We then remove lineage j and add the two new lineages to the state $L(t)$. This operation keeps the total amount of ancestral material unchanged.

Assuming a randomly mating, diploid population of constant size N_e, common ancestry occurs between two *overlapping* lineages at rate $\lambda = 1/(2N_e)$. Here, we deviate from the McVean and Cardin formulation, and define a strict total order on all lineages in $L(t)$: for each lineage $X_j(t) = [x_j,\ y_j)$ we define its order by the left coordinate x_j and label j (to break ties). The instantaneous rate of coalescence then equals λ multiplied by the number of overlapping pairs, i.e. $\lambda \sum_{j \neq k} I_{jk}$, where

$$I_{jk} = \begin{cases} 1 & X_j > X_k \text{ and } x_j < y_k \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We say that lineage k can coalesce *into* lineage j if $I_{jk} = 1$. Note that since $X_j > X_k$ implies that $x_j \geq x_k$, this condition implies that the left endpoint x_j of X_j falls in the interval defined by X_k. From the definition of the strict total order it follows that if $I_{jk} = 1$ then $I_{kj} = 0$. The newly formed lineage acquires the union of both intervals, the original lineages are removed, and $L(t)$ is updated accordingly.

For later use, also define $I_j(t)$ to be the total rate at which lineages are coalescing into lineage j, and $C_j(t)$ to be the set of lineages that can coalesce into lineage j. In other words, $C_j(t) = \{X_k \in L(t) | I_{jk}(t) = 1\}$ and $|C_j(t)| = I_j(t) = \sum_k I_{jk}(t)$, and the total rate of coalescence at time t is $\lambda \sum_j |C_j(t)|$. Because I_{jk} is defined in terms of a total order, the sets $C_j(t)$ are disjoint, which will simplify likelihood computations later. In particular, although recombination affects *which* lineages can coalesce into j, it does not change the *number* of such lineages: in other words, a recombination event changes $C_j(t)$ but not $I_j(t)$.

ARG likelihood

We have defined the SMC process in terms of rates of various types of events. In general, the likelihood for a continuous-time Markov process specified in this way is $\exp(-\Lambda) \prod_i \lambda_i$, where λ_i is the rate of the ith realised event, and Λ is the sum of all rates of all possible events integrated over the entire process (also called the "total hazard"). Each λ_i would be either equal to λ (if the event is a coalescence) or $r|X_j(t)|$ for some j and t (if it is a recombination), and since the total rates at time t of recombination events is $rT(t)$ and of coalescent events is $\lambda \sum_j |C_j(t)|$, the total hazard up to time T is $\Lambda = \int_0^T rT(t) + \lambda \sum_j |C_j(t)| dt$.

However, as discussed in the Introduction, in many cases these events are not what is estimated by inference methods, which produce, more fundamentally, a collection of genetic inheritance relationships (edges) between ancestral haplotypes (nodes). Our goal, therefore, is to decompose the overall likelihood into the per-node and edge contributions, allowing us to compute likelihoods for the incomplete ancestries estimated by current methods. Using the out- and in-degree of the nodes we'll first reason about the number of common ancestor events and recombination

events that could have resulted in the observed ARG. Next the information contained in the edges is leveraged to compute the total hazard associated with both types of events.

Let $\deg_o(u) = |\{e \in E : p_e = u\}|$ be the out-degree of a node u, i.e. the number of edges having u as a parent. Because mergers are binary in the SMC, this must have been the result of $\deg_o(u) - 1$ common ancestry events, and so this contributes a factor of $\lambda^{\deg_o(u)-1}$ to the likelihood for nodes with any children. Note however that this description is general and holds for the case of more-than-binary mergers. By treating such polytomies as if there were zero-length edges between a sequence of mergers, the order of these unresolved mergers does not affect the likelihood. Under the SMC likelihood of a zero-length edge is zero. This is not special: since edge lengths are continuously distributed, the likelihood of any particular value is also zero.

Similarly, let $\deg_i(u) = |\{e \in E : c_e = u\}|$ be the in-degree of a node u, i.e. the number of edges having u as a child. Each additional edge that is ancestral to a given node implies one additional ancestral recombination event, and thus a factor of $r^{\deg_i(u)-1}$ in the likelihood, if the node has any parents. Put together, these two contribute

$$\prod_{u \in N} \lambda^{\deg_o(u)-1} r^{\deg_i(u)-1} = r^{|E|-|N|+n_r} \lambda^{|E|-|N|+n_s} \tag{2}$$

to the likelihood, where n_r is the number of nodes with in-degree 0 (roots), and n_s is the number of nodes with out-degree 0 (usually, samples). (The simple form follows because each edge contributes exactly one to some node's out-degree and some node's in-degree.)

Now consider the "hazard" associated with recombination. Begin with an edge $e = (p, c, x, y)$. This implies that there was a sequence of lineages from the time of the child c (call this time t_c) back to the time of the parent p (again, t_p), that contained the segment (x, y), and the lineage that directly inherited from p had the segment (x, y) (see Fig. 2). We therefore know there was no recombination between x and y on any of those lineages over this time span, the probability of which is $\exp(-r\mathcal{A}_e)$, where r is the recombination rate and $\mathcal{A}_e$ is the total "area" of the edge. The area of an edge is its length in time, $t_p - t_c$, multiplied by the number of "eligible links" along its span. The number of eligible links is the number of adjacent base pairs in the edge on which recombination did not occur, which is either $y - x$ (if this edge is the leftmost of all edges above node c) or $y - x - 1$ (otherwise). So, the first contribution to the likelihood from this edge e is

$$A_e(r) = \exp(-r(y - x - 1_{x>x_c})(t_p - t_c)), \tag{3}$$

where as for lineages, x_c is the leftmost position of the edges whose children are c.

Finally, we turn to the hazard associated with common ancestry events. To do this, we need only consider what happens along the left side of each edge. Following what we did for lineages, we define a total order on edges by ordering by left endpoint and breaking ties by child (i.e. $(p, c, x, y) < (p', c', x', y')$ if $x < x'$, or if $x = x'$ and $c < c'$). The instantaneous rate of coalescence of a lineage whose left endpoint is at x at a given time is equal to λ multiplied by the number of earlier lineages that overlap x at that time. Using this total order, we can define $I_e(t)$ to be the total number of earlier edges that overlap edge e and are present at time t. Then, the cumulative hazard for common ancestry events of edge e between time s and t is given by λ multiplied by $F(e, s, t) = \int_s^t I_e(z)dz$. If the edge $e = (p, c, x, y)$ is the leftmost segment of material inherited by node c (e.g. the left hand segment ancestral to node c in Fig. 2), then no recombination occurred along this edge and this edge simply contributes $\exp(-\lambda F(e, t_c, t_p))$ to the likelihood. However, if the edge is *not* the leftmost segment (as shown in Fig. 2), then it was initiated by a new recombination breakpoint, and we need to integrate over possible times the breakpoint occurred. If this is the case, there is another edge (p', c, x', x) with the same child node, immediately adjacent to the focal edge (p, c, x, y). The recombination event that split the two must have happened between t_c and the smaller of t_p and $t_{p'}$. Taking all this into consideration, the contribution to the likelihood here is

$$B_e(r, \lambda) = \begin{cases} e^{-\lambda F(e, t_c, t_p)} & \text{if } x = x_c \\ \int_{t_c}^{t_p \wedge t_{p'}} e^{-rs} e^{-\lambda F(e, s, t_p)} ds & \text{if } x > x_c. \end{cases} \tag{4}$$

The overall likelihood of an ARG $\mathcal{G}$ defined by the set of edges E and nodes N given parameters θ is then given by

$$\mathcal{L}(\mathcal{G}|r, \lambda) = (r)^{|E|-|N|+n_r} (\lambda)^{|E|-|N|+n_s} \prod_{e \in E} A_e(r) B_e(r, \lambda). \tag{5}$$

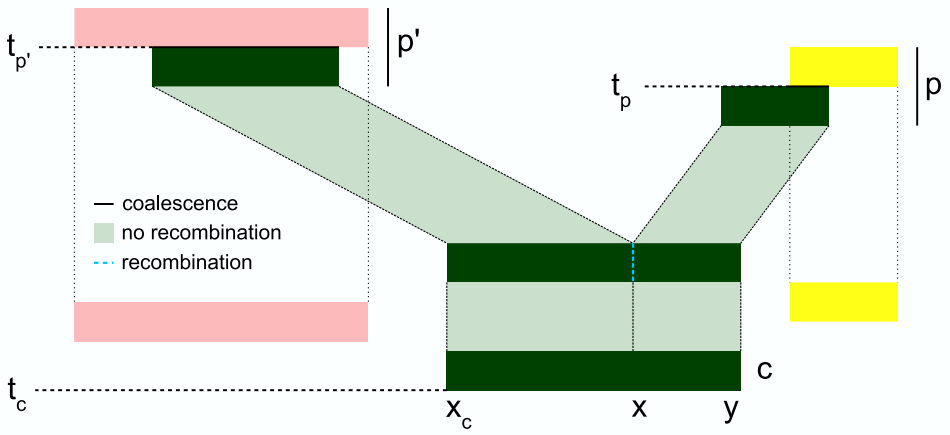

Fig. 2. Graphical representation of the events associated with the edge (p, c, x, y). The lineage carried by c spans (x_c, y) until a recombination event (dotted line at position x) splits it into two segments. Left and right hand segments coalesce (with the pink and yellow lineages, respectively) at times $t_{p'}$ and t_p, respectively, bounding the time to recombination by $\min(t_p, t_{p'})$.

One significant advantage of this formulation is that the computation time is linear in the number of edges, and an efficient algorithm exists to keep track of $I_e(t)$ (see Appendix and Fig. A1). Note that we have only considered the topology and branch length information here and not considered mutational information. Assuming the infinite sites model, computing the likelihood of an ARG given a mutation rate μ and set of mutations is straightforward, as the overall likelihood can be directly decomposed into the per-edge contributions (Baumdicker *et al.* 2022; Mahmoudi *et al.* 2022).

Robustness to partial and incorrect information

The ARG likelihood formulation in Eq. (5) allows us to compute a likelihood for any collection of nodes (haplotypes that existed at some time t) and edges (a record that child node c inherited the genomic interval $[x, y)$ from parent node p) under the SMC. Other than some basic consistency constraints (e.g. parents must be older than children) there are no requirements on the topology described, and in particular, no requirement that the events from the model (the SMC) should be explicitly identified in the graph. Our formulation reasons about the events that *could have resulted in* the observed nodes and edges, and therefore frees us from the requirement of having to explicitly write those down. In this section we illustrate the robustness of this approach to incomplete and potentially incorrect information about recombination and coalescent events via a proof-of-concept example.

In Fig. 3, we begin with a sample from the SMC model containing complete information about every simulated event (Full ARG; 9,378 nodes; 13,221 edges; 2,933 trees). We then plot the likelihood computed by Eq. (5) as we vary the recombination rate parameter r, as a ratio of the likelihood at the true value of r. Reassuringly, we can see that the minimum of the likelihood-ratio curve when using the full simulated ARG is close to the true value of the parameter. (The likelihood -ratio curve obtained when computing the likelihood under the full coalescent with recombination model using the KYF formula is identical; not shown.) We next compute the likelihood-ratio for the ARG obtained when we simplify (Kelleher *et al.* 2018; Wong *et al.* 2024) out the recombination nodes from the full ARG ("RE nodes removed"; 3,332 nodes; 6,901 edges; 2,933 trees; note that msprime uses distinct nodes for the left and right parent of each recombination). The curve we obtain is indistinguishable from that of the full ARG, demonstrating that the locally unary portions of coalescent nodes contain all the information that we need about recombination events for this application. When we fully simplify the simulated ARG to remove all locally unary portions of nodes from trees ("Fully simplified"; 3,332 nodes; 10,674 edges, 2,933 trees) we can see there is a significant loss of information, and the minimum of the likelihood-ratio surface is biased away from its true value, as expected, because removing the unary portions of nodes has removed some of the area of the ARG in which recombination did not occur. (See Wong *et al.* 2024 for more discussion on varying degrees of ARG simplification.)

Figure 3 shows that Eq. (5) is also remarkably robust to the presences of polytomies in this example. For each of the three ARGs discussed in the previous paragraph, we also evaluated the likelihood for ARGs in which 5% or 25% of the internal ARG nodes are removed and edges adjusted accordingly. This introduces

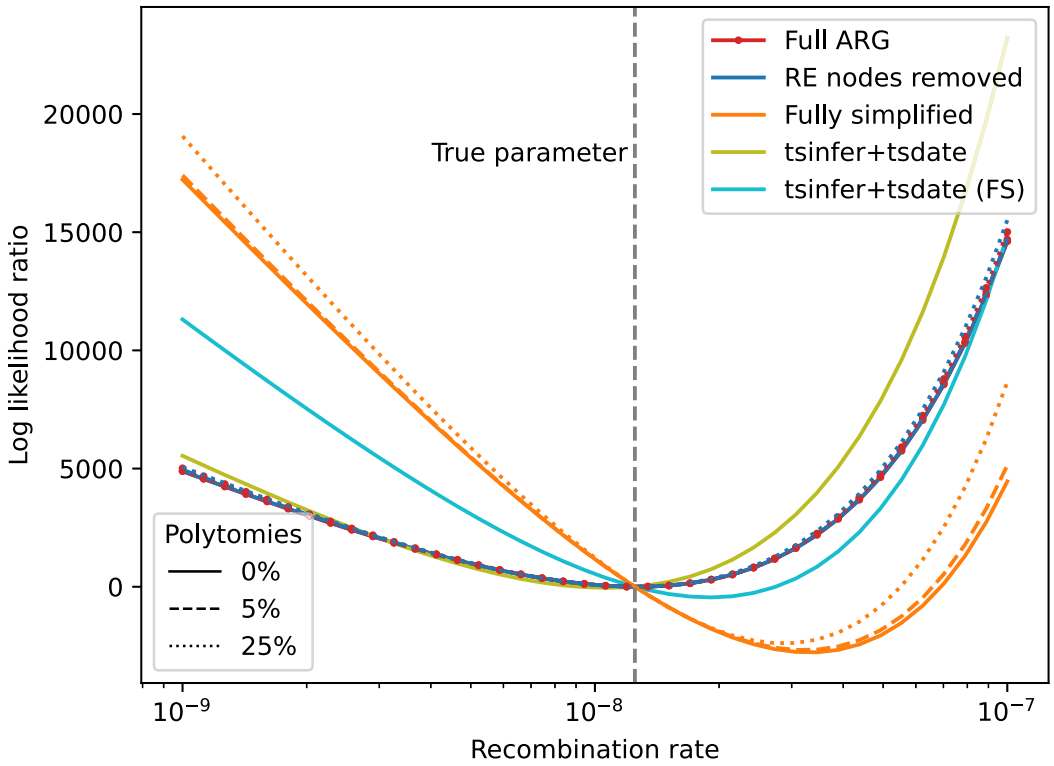

Fig. 3. Log likelihood-ratio curves for ARGs with various properties. The "Full ARG" here is the result of an msprime simulation of the SMC (100 diploid samples, genome length 1 Mb, $r = \mu = 1.25 \times 10^{-8}$, $N_e = 1 \times 10^4$). The "RE nodes removed" ARG is the result of simplifying out all of the recombination nodes, and the "Fully simplified" ARG has all locally unary nodes removed. Each of these simulated ARGs additionally has 5% and 25% of the internal nodes removed to create polytomies. We also show the results for ARGs inferred from the original simulation data using tsinfer and tsdate (see text for details), in which we use the default output of tsinfer including unary nodes (tsinfer+tsdate), and in which we fully simplify before passing to tsdate (tsinfer +tsdate FS). See the text for details on how the likelihood ratio values are computed. The "Full ARG" and "RE nodes removed" lines coincide.

polytomies at varying degrees in the different ARGs because deleting a node in a tree (and connecting its children directly to its parent) will only create a polytomy if the parent is not unary. Thus, while deleting 25% of the internal nodes in the fully simplified ARG created an average of 33 polytomies over the 2,933 trees along the sequence (each tree having 200 leaf nodes), it resulted in an average of 21 polytomies per tree in the ARG where recombination nodes have been removed, and only 10 in the full ARG (which has an average of 1,026 unary nodes per tree). Given this disparity in the number of polytomies introduced, the effects on the likelihood across the three examples are not entirely comparable. For the ARGs containing unary nodes, removing 25% of the internal nodes has negligible effect on the likelihood curve in Fig. 3, and even on the fully simplified ARG has a very minor effect on the location of the minimum. Supplementary Fig. S1 explores more extreme examples in which we remove 25%, 50% and 75% of the internal nodes in these ARGs, and demonstrates a remarkable robustness to the loss of the majority of the internal nodes in these examples. Interestingly, the loss of 75% of the internal nodes in the fully simplified ARG (1,841 of 2,455, leading to an average of 27 polytomies per tree) appears to improve performance, substantially reducing the bias away from the true parameter value.

Figure 3 also shows results for ARGs inferred from the simulated data. To do this we first simulated mutations on the full ARG at rate 1.25×10^{-8} using msprime, resulting in an ARG with 2,989 mutations. We then used tsinfer and tsdate (Wohns *et al.* 2022) to infer a dated ARG, using default parameters for tsinfer and the true values of mutation rate and N_e for tsdate. In the first example, we use the default output of tsinfer (2,148 nodes; 5,589 edges; 1,307 trees; 3,039 mutations). This ARG has an average of 192 unary nodes, 120 binary nodes, and 32 nodes with three or more children per tree. Tsinfer does not directly estimate recombination events; the breakpoints between trees are the consequence of switches between parental haplotypes in the Li and Stephens (2003) copying process. Information about recombination events in the returned ARG is therefore quite imprecise, and likely to contain incorrect and contradictory details if one were to attempt to reconstruct the exact subtree-prune-and-regraft moves (e.g. Rasmussen and Guo 2023). It is encouraging, therefore, that in this toy example our likelihood function is robust to these omitted and incorrect details in the output of tsinfer, with the minimum close to the true parameter value. We also show the likelihood curve obtained when we fully simplify the output of tsinfer (1,865 nodes; 7,511 edges; 1,252 trees; 3,039 mutations) before dating. We can see that, as with the true data, removing all the unary nodes results in a bias away from the true value of the recombination rate parameter.

Piecewise decomposition over time

Our final example explores decomposition of the likelihood into separate components for different segments of time. By explicitly associating a single likelihood with each edge, we can compute the likelihood for any slice of the ARG taken in time and/or along the genome. In the case of a fully precise ARG, in which each recombination event has an exact time, this is true as long as the slicing operation does not erase the information on whether an edge is associated with the left or right hand segment generated by a

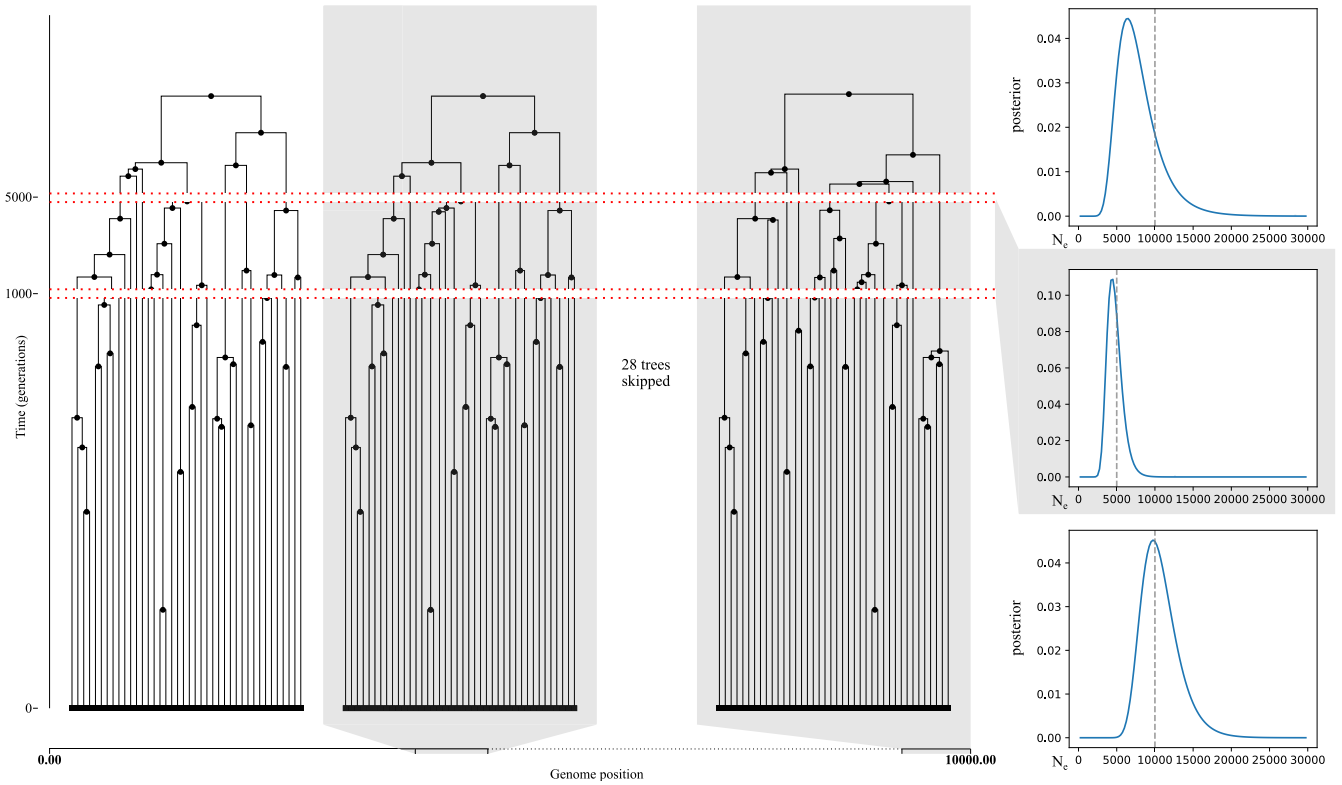

Fig. 4. Simulated ARG under the Hudson coalescent (left). N_e is constant for each of the 3 time slices: from 0 to 1,000 generations ago N_e is 10,000; from 1,000 to 5,000 generations ago it is 5,000; and past that it is 10,000. Remaining parameters: genome length is 10 kb, 20 diploid samples, and $r = 1.25 \times 10^{-8}$. There are 148 edges in the resulting ARG. A single N_e value was estimated for each of the three slices of the ARG separated by the true population size change time points. The inferred posterior distribution (right) using the SMC-likelihood was obtained using a uniform prior. The number of edges per time slice (from most recent to oldest) was: 68, 57, and 23. Locally unary nodes were omitted for clarity.

recombination event. In the absence of the exact time to recombination, the integral in Eq. (4) means that the likelihood for the entire ARG does not factorise into the likelihoods of these individual slices. More precisely, the likelihood associated with any edge that is initiated by a recombination breakpoint cannot be factorized into the product of two integrals of the same form. However, as long as the individual slices of the ARG contain sufficient information we can approximate the true integral as the product of the integral of the slices. Figure 4 demonstrates this idea on a toy example. We simulated an ARG under human-like parameters with two effective population size changes, and subsequently computed the posterior distribution on the three effective population sizes using a uniform prior on each slice independently. The result is reasonably accurate, despite only using 10 kb of sequence.

Discussion

The key contribution of this article is to rederive a classical ARG likelihood function in a way that supports the sort of partial and imprecise information provided by some modern ARG inference tools, using a novel backwards-time derivation of the SMC. The underlying process that we reason about probabilistically is still composed of events affecting lineages, but, because we do not necessarily observe or estimate these events, we integrate over the possible timings implied by the information in the input ARG. Our method can compute a likelihood under the SMC for a very general class of ARG (Wong *et al.* 2024), a significant increase in flexibility over the strict input requirements for the classical KYF formulation. Locally unary nodes play a key role, and although not currently a common focus of ARG research (Wong *et al.* 2024), they are estimated by tsinfer and may be imputed to a high degree of accuracy from fully simplified ARGs (Fritze *et al.* 2024). We have shown via some illustrative examples that the formulation is robust to incomplete and imprecise information, tolerating the loss of a significant fraction of the internal nodes in the ARG to minimal effect. We have also shown that the likelihood function performs well (at least in our toy example) on ARGs inferred by tsinfer+tsdate, successfully capturing information about latent recombination events from imprecise and noisy data. However, there are many more aspects of the likelihood function that could be examined, beyond coarse parameters like recombination rate or effective populations size. Theoretical explanations of this apparent robustness, and fully characterizing the properties of the likelihood on the output of tsinfer and other methods is an important avenue for future work. Similarly, while there are many potential applications to concrete inferential problems, the details are nontrivial and must be consigned to future work.

Our results, while preliminary, suggest that "full ARGs," in which each recombination and common ancestor event is precisely estimated, contain a significant degree of redundancy in terms of capturing information about the generative process. It is possible, therefore, that explicitly estimating *all* such events is not necessary to sample ARGs under the SMC (as ARGweaver and SINGER do). A combined approach, in which the uncertainty around the ordering of events that are weakly informed by the data is encoded structurally in the graph and other forms of uncertainty captured by sampling from the posterior, may yield further improvements in scalability for statistically rigorous inference. Another exciting prospect for ARG inference is the development of "hybrid" inference methods, where heuristic methods such as tsinfer or ARG-Needle are used to infer the details of the recent past in huge datasets such as UK Biobank, passing over to a rigorous probabilistic approach for the ancient past. This is similar in spirit to the widely-used approach of simulating complex dynamics of the recent past using detailed but slow models and using faster coalescent simulations in the distant past when the assumptions required are more justified (Bhaskar *et al.* 2014; Haller *et al.* 2018; Kelleher *et al.* 2018; Nelson *et al.* 2020). Just as forward and backward simulations are useful in different contexts and can be combined to great effect, large-scale heuristic and probabilistic inference methods have their applications, and there may be synergy in their combination. The practical details of "handing over" between methods will require cooperation and precise communication; the tskit library, with its well-defined data model and flexible metadata support, is the ideal basis for this.

Data availability

All scripts and data used for this manuscript are available at https://github.com/gertjanbisschop/smc-bit-paper, and an implementation of the algorithm, named "Recombination (time) UNaware SMC" (runsmc), is available at https://github.com/gertjanbisschop/runsmc.

Supplemental material available at GENETICS online.

Funding

JK acknowledges EPSRC (research grant EP/X024881/1) and the Robertson Foundation. JK and PR acknowledge support from NIH (research grants HG011395 and HG012473).

Conflicts of interest

The author(s) declare no conflicts of interest.

Literature cited

All of Us Research Program Genomics Investigators. 2024. Genomic data in the all of us research program. Nature. 627(8003):340. doi:10.1038/s41586-023-06957-x.

Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, *et al.* 2021. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 599(7886):628–634. doi:10.1038/s41586-021-04103-z.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, *et al.* 2022. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 220(3):iyab229. doi:10.1093/genetics/iyab229.

Bhaskar A, Clark AG, Song YS. 2014. Distortion of genealogical properties when the sample is very large. Proc Natl Acad Sci USA. 111(6):2385–2390. doi:10.1073/pnas.1322709111.

Brandt DY, Huber CD, Chiang CW, Ortega-Del Vecchyo D. 2024. The promise of inferring the past using the ancestral recombination graph. Genome Biol Evol. 16(2). doi:10.1093/gbe/evae005.

Brandt D, Wei X, Deng Y, Vaughn AH, Nielsen R. 2022. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics. 221(1):iyac044. doi:10.1093/genetics/iyac044.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, *et al.* 2018. The UK

Biobank resource with deep phenotyping and genomic data. Nature. 562(7726):203–209. doi:10.1038/s41586-018-0579-z.
Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, Hubbard T, Jostins L, Maltby N, Mahon-Pearson J, *et al.* 2017. National Genomic Research Library v5.1, Genomics England.
Cook MB, Sanderson SC, Deanfield JE, Reddington F, Roddam A, Hunter DJ, Ali R. 2025. Our future health: A unique global resource for discovery and translational research. Nat Med. 31: 728–730. doi:10.1038/s41591-024-03438-0.
De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, *et al.* 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 354(6311):477–481. doi: 10.1126/science.aag2602.
Deng Y, Nielsen R, Song YS. 2024. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. bioRxiv 585351. https://doi.org/10.1101/2024.03.16.585351, pp. 2024–03, preprint: not peer reviewed.
Deng Y, Song YS, Nielsen R. 2025. A general framework for branch length estimation in ancestral recombination graphs. bioRxiv 638385. https://doi.org/10.1101/2025.02.14.638385, pp. 2025–02, preprint: not peer reviewed.
Deraje P, Kitchens J, Coop G, Osmond MM. 2024. The promise and challenge of spatial inference with the full ancestral recombination graph under Brownian motion. bioRxiv 588900. https://doi.org/10.1101/2024.04.10.588900, preprint: not peer reviewed.
Fan C, Cahoon JL, Dinh BL, Ortega-Del Vecchyo D, Huber CD, Edge MD, Mancuso N, Chiang CW. 2025. A likelihood-based framework for demographic inference from genealogical trees. Nat Genet. 57:865–874. doi:10.1038/s41588-025-02129-x.
Fan C, Mancuso N, Chiang CW. 2022. A genealogical estimate of genetic relationships. Am J Hum Genet. 109(5):812–824. doi:10.1016/j.ajhg.2022.03.016.
Fritze H, Pope N, Kelleher J, Ralph PL. 2024. A forest is more than its trees: Haplotypes and inferred ARGs. bioRxiv 626138. https://doi.org/10.1101/2024.11.30.626138, preprint: not peer reviewed.
Griffiths R, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol. 3(4): 479–502. doi:10.1089/cmb.1996.3.479.
Grundler MC, Terhorst J, Bradburd GS. 2025. A geographic history of human genetic ancestry. Science. 387(6741):1391–1397. doi:10.1126/science.adp4642.
Gunnarsson "F, Zhu J, Zhang BC, Tsangalidou Z, Allmont A, Palamara PF. 2024. A scalable approach for genome-wide inference of ancestral recombination graphs. bioRxiv 610248. https://doi.org/10.1101/2024.08.31.610248, pp. 2024–08, preprint: not peer reviewed.
Guo F, Carbone I, Rasmussen DA. 2022. Recombination-aware phylogeographic inference using the structured coalescent with ancestral recombination. PLoS Comput Biol. 18(8):e1010422. doi:10.1371/journal.pcbi.1010422.
Halldorsson BV, Eggertsson HP, Moore KH, Hauswedell H, Eiriksson O, Ulfarsson MO, Palsson G, Hardarson MT, Oddsson A, Jensson BO, *et al.* 2022. The sequences of 150,119 genomes in the UK Biobank. Nature. 607(7920):732–740. doi:10.1038/s41586-022-04965-x.
Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2018. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol Ecol Resour. 19:552–566. doi:10.1111/1755-0998.12968.
Harris K. 2019. From a database of genomes to a forest of evolutionary trees. Nat Genet. 51(9):1306–1307. doi:10.1038/s41588-019-0492-x.
Harris K. 2023. Using enormous genealogies to map causal variants in space and time. Nat Genet. 55:730–731. doi:10.1038/s41588-023-01389-9.
Hejase HA, Mo Z, Campagna L, Siepel A. 2022. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. Mol Biol Evol. 39(1):msab332. doi:10.1093/molbev/msab332.
Hejase HA, Salman-Minkov A, Campagna L, Hubisz MJ, Lovette IJ, Gronau I, Siepel A. 2020. Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. Proc Natl Acad Sci USA. 117(48):30554–30565. doi:10.1073/pnas.2015987117.
Huang Z, Kelleher J, Chan YB, Balding D. 2025. Estimating evolutionary and demographic parameters via ARG-derived IBD. PLoS Genet. 21(1):e1011537. doi:10.1371/journal.pgen.1011537.
Hubisz MJ, Williams AL, Siepel A. 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS Genet. 16(8):e1008895. doi: 10.1371/journal.pgen.1008895.
Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 23(2):183–201. doi:10.1016/0040-5809(83)90013-8.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18(2):337–338. doi:10.1093/bioinformatics/18.2.337.
Ignatieva A, Favero M, Koskela J, Sant J, Myers SR. 2023. The distribution of branch duration and detection of inversions in ancestral recombination graphs. bioRxiv 548567. https://doi.org/10.1101/2023.07.11.548567, preprint: not peer reviewed.
Ignatieva A, Hein J, Jenkins PA. 2022. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. Mol Biol Evol. 39(2):msac028. doi:10.1093/molbev/msac028.
Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 12(5):e1004842. doi:10.1371/journal.pcbi.1004842.
Kelleher J, Thornton KR, Ashander J, Ralph PL. 2018. Efficient pedigree recording for fast population genetics simulation. PLoS Comput Biol. 14(11):1–21. doi:10.1371/journal.pcbi.1006581.
Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019. Inferring whole-genome histories in large population datasets. Nat Genet. 51(9):1330–1338. doi:10.1038/s41588-019-0483-y.
Korfmann K, Sellinger TPP, Freund F, Fumagalli M, Tellier A. 2024. Simultaneous inference of past demography and selection from the ancestral recombination graph under the beta coalescent. Peer Community J. 4(11). doi:10.24072/pcjournal.397.
Kuhner MK, Yamato J, Felsenstein J. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics. 156(3):1393–1401. doi:10.1093/genetics/156.3.1393.
Lewanski AL, Grundler MC, Bradburd GS. 2024. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. PLoS Genet. 20(1): e1011110. doi:10.1371/journal.pgen.1011110.
UK Biobank Whole-Genome Sequencing Consortium, Li S, Carss KJ, Halldorsson BV, Cortes A. 2023. Whole-genome sequencing of half-a-million UK Biobank participants. medRxiv 23299426. https://doi.org/10.1101/2023.12.06.23299426, pp. 2023–12, preprint: not peer reviewed.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature. 475(7357):493–496. doi:10.1038/nature10231.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 165(4):2213–2233. doi:10.1093/genetics/165.4.2213.

Mahmoudi A, Koskela J, Kelleher J, Chan Y, Balding D. 2022. Bayesian inference of ancestral recombination graphs. PLoS Comput Biol. 18(3):e1009960. doi:10.1371/journal.pcbi.1009960.

Marjoram P, Wall JD. 2006. Fast "coalescent" simulation. BMC Genet. 7(1):1–9. doi:10.1186/1471-2156-7-16.

McVean G, Cardin NJ. 2005. Approximating the coalescent with recombination. Philos Trans R Soc B Biol Sci. 360(1459):1387–1393. doi:10.1098/rstb.2005.1673.

Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. 2020. Accounting for long-range correlations in genome-wide simulations of large cohorts. PLoS Genet. 16(5):e1008619. doi:10.1371/journal.pgen.1008619.

Nielsen R, Vaughn AH, Deng Y. 2025. Inference and applications of ancestral recombination graphs. Nat Rev Genet. 26(1):47–58. doi:10.1038/s41576-024-00772-4.

Nowbandegani PS, Wohns AW, Ballard JL, Lander ES, Bloemendal A, Neale BM, O'Connor LJ. 2023. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. Nat Genet. 55:1494–1502. doi:10.1038/s41588-023-01487-8.

Osmond M, Coop G. 2024. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. Elife. 13: e72177. doi:10.7554/eLife.72177.

Paul JS, Steinrucken M, Song YS. 2011. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics. 187(4):1115–1128. doi:10.1534/genetics.110.125534.

Ralph P, Thornton K, Kelleher J. 2020. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. Genetics. 215(3):779–797. doi:10.1534/genetics.120.303253.

Rasmussen DA, Guo F. 2023. Espalier: Efficient tree reconciliation and ancestral recombination graphs reconstruction using maximum agreement forests. Syst Biol. 72(5):1154–1170. doi:10.1093/sysbio/syad040.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10(5). doi:10.1371/journal.pgen.1004342.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 46(8):919–925. doi:10.1038/ng.3015.

Shriner D, Rotimi CN. 2018. Whole-genome-sequence-based haplotypes reveal single origin of the sickle allele during the holocene wet phase. Am J Hum Genet. 102(4):547–556. doi:10.1016/j.ajhg.2018.02.003.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nat Genet. 51(9): 1321–1329. doi:10.1038/s41588-019-0484-x.

Speidel L, Silva M, Booth T, Raffield B, Anastasiadou K, Barrington C, Götherström A, Heather P, Skoglund P. 2025. High-resolution genomic history of early medieval Europe. Nature. 637(8044): 118–126. doi:10.1038/s41586-024-08275-2.

Stankowski S, Zagrodzka ZB, Garlovsky MD, Pal A, Shipilina D, Castillo DG, Lifchitz H, Le Moan A, Leder E, Reeve J, *et al.* 2024. The genetic basis of a recent transition to live-bearing in marine snails. Science. 383(6678):114–119. doi:10.1126/science.adi2982.

Stark Z, Glazer D, Hofmann O, Rendon A, Marshall CR, Ginsburg GS, Lunt C, Allen N, Effingham M, Hastings Ward J, *et al.* 2024. A call to action to scale up research and clinical genomic data sharing. Nat Rev Genet. 26:141–147. 10.1038/s41576-024-00776-0.

Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from dna sequence data. PLoS Genet. 15(9):e1008384. doi:10.1371/journal.pgen.1008384.

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Baple E, Craig C, Hamblin A, Henderson S, *et al.* 2018. The 100,000 genomes project: Bringing whole-genome sequencing to the NHS. BMJ. 361: k1687. doi:10.1136/bmj.k1687.

Wang S, Coop G. 2022. A complex evolutionary history of genetic barriers to gene flow in hybridizing warblers. bioRxiv 516535. https://doi.org/10.1101/2022.11.14.516535, pp. 2022–11, preprint: not peer reviewed.

Wiuf C, Hein J. 1999a. Recombination as a point process along sequences. Theor Popul Biol. 55(3):248–259. doi:10.1006/tpbi.1998.1403.

Wiuf C, Hein J. 1999b. The ancestry of a sample of sequences subject to recombination. Genetics. 151(3):1217–1228. doi:10.1093/genetics/151.3.1217.

Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, Patterson N, Reich D, Kelleher J, McVean G. 2022. A unified genealogy of modern and ancient genomes. Science. 375(6583). doi:10.1126/science.abi8264.

Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. 2024. A general and efficient representation of ancestral recombination graphs. Genetics. 228(1):iyae100. doi:10.1093/genetics/iyae100.

Zhang BC, Biddanda A, Gunnarsson AF, Cooper F, Palamara PF. 2023. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat Genet. 55(5): 768–776. doi:10.1038/s41588-023-01379-x.

Appendix

Algorithm

In this section we describe how to efficiently compute the likelihood under the SMC from an ARG stored using the `tskit` tree sequence encoding. The tree-sequence format encodes relationships using *edges*, along with additional information that allows us to sequentially recover local trees (and reason about their differences) efficiently (Kelleher *et al.* 2016; Ralph *et al.* 2020). The order of the edge deletions and insertions required to move from one local tree T_1 to the next, T_2 imposes the strict total ordering on edges (see definition (1)) that is needed to efficiently maintain the state of $I_e(t)$ for each focal edge e. Indeed, following the removal of edges that do not persist in T_2, we simply declare that each subsequently added edge can only coalesce with the already-visited edges that make up the (partially) reconstructed local tree T_2.

This allows us to compute a likelihood in a single pass over all edges. More concretely, the likelihood is computed by maintaining a single vector I that tracks the number of extant lineages in each internode interval (see Fig. A1). When moving from T_1 to T_2, the vector I is first decremented along the internode intervals spanned by the time of the child and the parent of all edges that do not persist in T_2. For each subsequently inserted edge, we first compute the likelihood contribution of that edge by computing the integral in Eq. (5) using I, and then update the counts in I. We further keep track of the last parent a child node was connected to to detect recombination events. The pseudocode for the algorithm is included in the Supplementary Material (Algorithm 2). Note that Supplementary Algorithm 1 details how to compute the integrals contained in Eq. (4).

Editor: M. Sillanpää

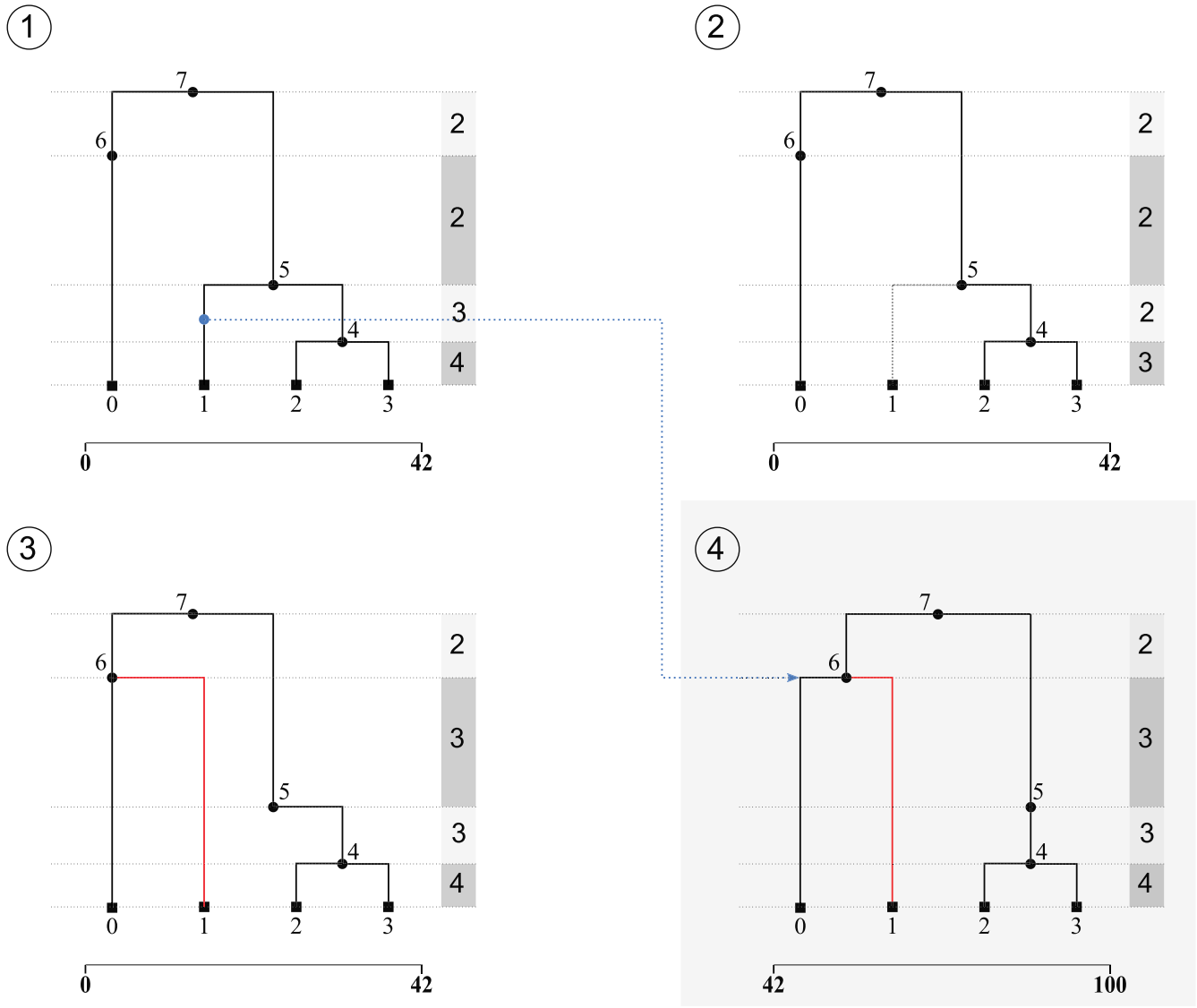

Fig. A1. Likelihood computation algorithm. Moving along the genome we transition from local tree T_1 to the next, T_2 in three steps. Edges that do not persist in T_2 are removed first. Edges starting at the left coordinate of T_2 are inserted. After deletion or insertion of an edge the count vector I is updated along the corresponding internode intervals. Following an insertion, and prior to updating I, the likelihood for that edge is computed as I then represents the total number of lineages the lineage associated with this new edge can coalesce with. By moving along the genome in this way, each edge gets visited exactly once.