A general and efficient representation of ancestral recombination graphs

Citation for published version:
Wong, Y, Ignatieva, A, Koskela, J, Gorjanc, G, Wohns, AW & Kelleher, J 2023 'A general and efficient representation of ancestral recombination graphs' bioRxiv, bioRxiv, pp. 1-32.
https://doi.org/10.1101/2023.11.03.565466

Digital Object Identifier (DOI):
10.1101/2023.11.03.565466

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

A general and efficient representation of ancestral recombination graphs

Yan Wong[1], Anastasia Ignatieva[2,3⋆], Jere Koskela[4,5⋆], Gregor Gorjanc[6], Anthony W. Wohns[7,8], and Jerome Kelleher[1†]

November 3, 2023

Abstract

As a result of recombination, adjacent nucleotides can have different paths of genetic inheritance and therefore the genealogical trees for a sample of DNA sequences vary along the genome. The structure capturing the details of these intricately interwoven paths of inheritance is referred to as an ancestral recombination graph (ARG). New developments have made it possible to infer ARGs at scale, enabling many new applications in population and statistical genetics. This rapid progress, however, has led to a substantial gap opening between theory and practice. Standard mathematical formalisms, based on exhaustively detailing the "events" that occur in the history of a sample, are insufficient to describe the outputs of current methods. Moreover, we argue that the underlying assumption that all events can be known and precisely estimated is fundamentally unsuited to the realities of modern, population-scale datasets. We propose an alternative mathematical formulation that encompasses the outputs of recent methods and can capture the full richness of modern large-scale datasets. By defining this ARG encoding in terms of specific genomes and their intervals of genetic inheritance, we avoid the need to exhaustively list (and estimate) *all* events. The effects of multiple events can be aggregated in different ways, providing a natural way to express many forms of approximate and partial knowledge about the recombinant ancestry of a sample.

Keywords: Ancestral recombination graphs

1 Introduction

Estimating the genetic genealogy of a set of DNA sequences under the influence of recombination, usually known as an Ancestral Recombination Graph (ARG), is a long-standing goal in genetics. Broadly speaking, an ARG describes the different paths of genetic inheritance caused by recombination, encapsulating the resulting complex web of genetic ancestry (see Lewanski et al. (2023) for a biologically oriented introduction). Recent breakthroughs in large-scale inference methods (Rasmussen et al., 2014; Kelleher et al., 2019b; Speidel et al., 2019; Schaefer et al., 2021; Wohns et al., 2022; Zhang et al., 2023; Zhan et al., 2023) have raised the realistic prospect of ARG-based analysis becoming a standard part of the population and statistical genetics toolkit (Hejase et al., 2020). Applications using inferred ARGs as input have begun to appear (Osmond and Coop, 2021; Fan et al., 2022; Hejase et al., 2022; Guo et al., 2022; Zhang et al., 2023; Nowbandegani et al., 2023; Ignatieva et al., 2023; Fan et al., 2023) and many more are sure to follow (Harris, 2019, 2023).

Although it is widely accepted that ARGs are important, there is some confusion about what, precisely, an ARG *is*. In its original form, developed by Griffiths and colleagues, the ARG is an alternative formulation of the coalescent with recombination (Hudson, 1983a), where the stochastic

[1] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK
[2] School of Mathematics and Statistics, University of Glasgow, UK
[3] Department of Statistics, University of Oxford, UK
[4] School of Mathematics, Statistics and Physics, Newcastle University, UK
[5] Department of Statistics, University of Warwick, UK
[6] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK
[7] Broad Institute of MIT and Harvard, Cambridge, USA
[8] Department of Genetics, Stanford University School of Medicine, Stanford, USA
[*] Joint second author, listed alphabetically
[†] Correspondence: jerome.kelleher@bdi.ox.ac.uk

process of coalescence and recombination among ancestral lineages is formalised as a graph (Griffiths, 1991; Ethier and Griffiths, 1990; Griffiths and Marjoram, 1996, 1997). Subsequently, an ARG has come to be thought of as a data structure (Minichiello and Durbin, 2006), i.e. describing a *realisation* of such a random process, or an inferred ancestry of a sample of genomes. The distinction between stochastic process and data structure is not clear cut, however, and subfields use the term differently (Appendix A). Other subtly different concepts that we must be careful to distinguish are the "true" ARG, describing the actual history of a sample of genomes, and a "population" ARG which is the true ARG of every individual in a population. True ARGs are the underlying real history of a sample, perfectly resolved into binary splits and mergers by the cellular processes of meiosis and mitosis, regardless of sampling density or population processes (Appendix D). Although population-scale true ARGs unquestionably exist, they can also never be entirely known, in part because of a fundamental lack of mutational information. Even if mutation rate were high enough to uniquely identify every recent branching point, such a high rate would saturate the genome with mutations and obscure deeper history.

Population ARGs may seem fanciful, but the scale of modern datasets makes it necessary for us to grapple with the idea. The UK Biobank (UKB), for example, has genotype data for around 500,000 humans (Bycroft et al., 2018), along with exome (Backman et al., 2021) and whole genome sequence (Halldorsson et al., 2022) data for large subsets of the cohort. UKB is just one of many such population-scale sequencing projects (e.g. Turnbull et al., 2018; Karczewski et al., 2020; Tanjo et al., 2021). Agricultural datasets are on a similar scale, and also include dense multi-generational sampling and near-perfect pedigree information (e.g. Hayes and Daetwyler, 2019; Ros-Freixedes et al., 2020). Recent advances have made it possible to actually *estimate* ARGs at population scale: ARGs have already been inferred for the 500,000 humans in UKB (Kelleher et al., 2019b; Zhang et al., 2023) and over a million SARS-CoV-2 genomes (Zhan et al., 2023). While this new population-scale reality presents many exciting opportunities, it also poses substantial challenges to existing methodologies.

A major problem currently facing the field is that classical mathematical formalisms and terminology cannot adequately describe these vast inferred ARGs. Fundamentally, these formalisms assume that an ARG is known in complete detail and are not suited to describing partial or approximate knowledge. As we are actively inferring ARGs at the population scale, and such ARGs can never be known in complete detail, there is currently a substantial gap between our theoretical frameworks and practical application. The breakthroughs in scale achieved by recent methods (e.g. Kelleher et al., 2019b; Speidel et al., 2019; Zhang et al., 2023) are all based, in different ways, on inferring approximate *structures* instead of a complete and fully detailed history. (Note that it is important to distinguish here between structures and models: whether an inference method is based on heuristics or a rigorous mathematical model is orthogonal to the level of detail provided in its estimate. One could heuristically estimate a fully precise ARG, or statistically sample a partial, approximate ARG under a model such as the coalescent.) Although the term "ARG" is now often used in a general sense (e.g. Mathieson and Scally, 2020; Hejase et al., 2020; Schaefer et al., 2021; Harris, 2023; Zhang et al., 2023; Fan et al., 2023), informally encompassing the varied approximate structures output by modern simulation and inference methods (Rasmussen et al., 2014; Palamara, 2016; Haller et al., 2018; Kelleher et al., 2019b; Speidel et al., 2019; Baumdicker et al., 2022; Zhang et al., 2023), there is no corresponding mathematical definition that is sufficiently general.

We address this problem by providing a simple formal definition of an ARG data structure, based on recording the intervals of genetic inheritance between specific genomes. We call this the "genome ARG", or gARG encoding. We contrast this with the classical formal definition of an ARG, based on recording common ancestor and recombination events, which we refer to as the "event ARG" or eARG encoding. We show that the new gARG encoding is a substantial generalisation of the classical eARG approach, providing much more flexibility in how genetic inheritance can be represented, and encompasses the outputs of modern methods. We show that the gARG approach can represent many different types of approximation, in particular allowing us to systematically describe uncertainty about the temporal ordering of multiple recombinations. It is important to note that throughout we are interested in the details of these competing mathematical formulations and their practical consequences.

We begin in Section 2 by providing a precise formal definition of a gARG, illustrated by an example ARG embedded in pedigree. We then provide a similar definition of the classical eARG approach in Section 3, and consider its limitations in the context of current datasets and research questions. Following this, we discuss the important concept of ancestral material in Section 4, and how it relates to the process of converting an eARG to a gARG. We continue in Section 5 by considering the relationship between an ARG and its local trees. Contrary to the prevailing view, we show that a suitably encoded sequence of local trees contains precisely as much information as the corresponding ARG. The gARG encoding opens a rich new set of details about ARGs, including the ideas of locally unary nodes (Section 6), the levels of detail that can be represented in an ARG (Section 7), and the degrees of precision about recombination that can be stored and we may seek to infer (Section 8). These ideas have important practical considerations, which we illustrate by examining the qualitative properties of ARGs inferred by four recent methods for a classical benchmark dataset in Section 9. We then discuss how the gARG framework can be efficiently implemented in Section 10, and finish with an assessment of the key challenges facing the field in the Discussion. Finally, the literature on ARGs is large and confusing, and we attempt to clarify some important aspects in appendices, including a brief history of ancestral graphs (Appendix A), a description of the Big and Little ARG stochastic processes (Appendix B), a survey of ARG inference methods (Appendix C), and a discussion of ARGs at an individual vs cell lineage level (Appendix D).

2 Genome ARGs

We define a genome as the complete set of genetic material that a child inherits from one parent. A diploid individual therefore carries two genomes, one inherited from each parent (we assume diploids here for clarity, but the definitions apply to organisms of arbitrary ploidy). We will also use the term "genome" in its more common sense of "the genome" of a species, and hope that the distinction will be clear from the context. We are not concerned here with mutational processes or observed sequences, but consider only processes of inheritance, following the standard practice in coalescent theory. We also do not consider structural variation, and assume that all samples and ancestors share the same genome coordinate space.

A genome ARG (gARG) is a directed acyclic graph in which nodes represent haploid genomes and edges represent genetic inheritance between an ancestor and a descendant. The topology of a gARG specifies that genetic inheritance occurred between particular ancestors and descendants, but the graph connectivity does not tell us which *parts* of their genomes were inherited. In order to capture the effects of recombination we "annotate" the edges with the genome coordinates over which inheritance occurred. This is sufficient to describe the effects of inheritance under any form of homologous recombination (such as multiple crossovers, gene conversion events, and many forms of bacterial and viral recombination).

We can define a gARG formally as follows. Let $N = \{1, \ldots, n\}$ be the set of nodes representing the genomes in the gARG, and $S \subseteq N$ be the set of sampled genomes. Then, E is the set of edges, where each element is a tuple (c, p, I) such that $c, p \in N$ are the child and parent nodes and I is the set of disjoint genomic intervals over which genome c inherits from p. Thus, each topological connection between a parent and child node in the graph is annotated with a set of inheritance intervals I. Here, the terms parent and child are used in the graph sense; these nodes respectively represent ancestor and descendant genomes, which can be separated by multiple generations. We will use these two sets of terms interchangeably.

How nodes are interpreted, exactly, is application dependent. Following Hudson (1983a), we can view nodes as representing gametes, or we can imagine them representing, for example, the genomes present in cells immediately before or after some instantaneous event (Appendix D). A node can represent any genome along a chain of cell divisions or can be interpreted as representing one of the genomes of a potentially long-lived individual. All these interpretations are potentially useful, and equally valid under the assumptions of the gARG encoding. In many settings, nodes are dated, i.e. each node $u \in N$ is associated with a time τ_u, and how we assign precise times will vary by application. The topological ordering defined by the directed graph structure and an arrow of time (telling us which direction is pastwards) is sufficient for many applications, however, and we assume node dates are not known here. In practical settings, we will wish to associate additional metadata with nodes such as sample identifiers or quality-control metrics. It is therefore best to think of the integers used here in the definition of a node as an *identifier*, with which arbitrary additional information can be associated.

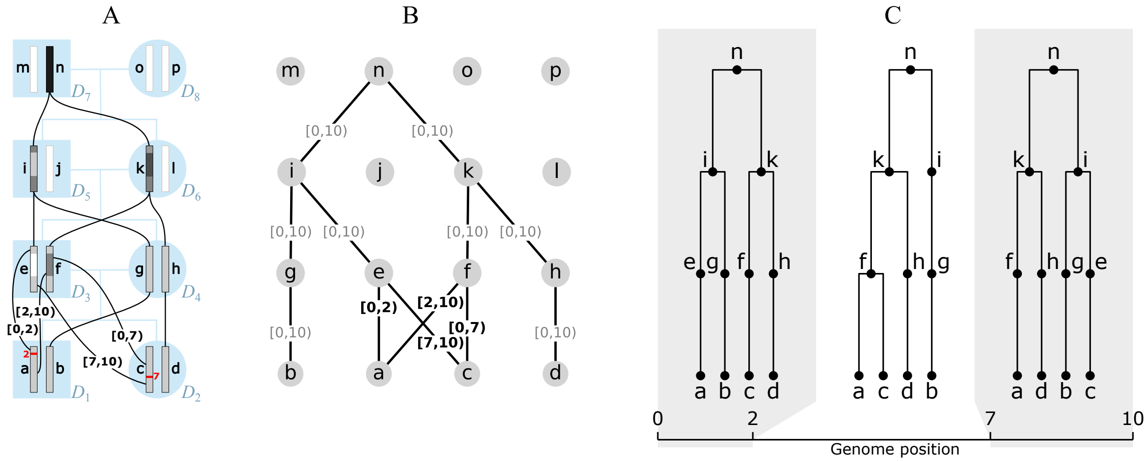

Figure 1: An example genome ARG (gARG) embedded in a pedigree. (A) Diploid individuals (blue), visualised in a highly inbred pedigree and labelled D_1 to D_8, contain both paternal and maternal genomes labelled a to p. Black lines show inheritance paths connecting genomes in the current generation (a to d) with their ancestors. Genomes a and c are the product of two independent meioses (recombination events, in red) between the paternal genomes e and f, and regions of genome inherited are shown with shaded colour. Genomes are shaded such that where, backwards in time, they merge into a common ancestor, the merged region is darker. (B) The corresponding gARG along with inheritance annotations on all edges (partial inheritance in bold). (C) The corresponding local trees.

As illustrated in Fig. 1, the gARG for a given set of individuals is embedded in their pedigree. The figure shows the pedigree of eight diploid individuals and their sixteen constituent genomes (each consisting of a single chromosome), along with paths of genetic inheritance. Here, and throughout, nodes are labelled with lowercase alphabetical letters rather than integer identifiers to avoid confusion with genomic intervals. Thus individual D_1 is composed of genomes a and b, which are inherited from its two parents D_3 and D_4. Each inherited genome may be the recombined product of the two genomes belonging to an individual parent. In this example, genome b was inherited directly from D_4's genome g without recombination, whereas genome a is the recombinant product of D_2's genomes e and f crossing over at position 2. Specifically, genome a inherited the (half-closed) interval $[0, 2)$ from genome e and $[2, 10)$ from genome f. These intervals are shown attached to the corresponding graph edges. The figure shows the annotated pedigree with realised inheritance of genomes between generations (A), the corresponding gARG (B), and finally the corresponding sequence of local trees along the genome (C). The local trees span the three genome regions delineated by the two recombination breakpoints that gave rise to these genomes; see Section 5 for details on how local trees are embedded in an ARG.

The genome ARG framework defined here is in many ways simply a clarification of existing treatments (e.g. Mathieson and Scally, 2020; Shipilina et al., 2023), adding concrete details to describe the differential inheritance of genetic material between genomes. It is important to note that here, and throughout, we are not questioning the form of the actual ancestral processes that occur in nature, but rather how we *represent* the outcomes of such processes in a practical manner. These practical details, as demonstrated in later sections, have important consequences not only for how methods exchange information about simulated and inferred ARGs, but more fundamentally in how we set our goals for inference and evaluate the success of results.

3 Event ARGs

In this section we define the classical view of an ARG data structure, and illustrate its limitations. We are interested in the details of how ARGs are described mathematically, and as a consequence, how they are represented in a practical sense as the output of inference programs. Where details of an ARG data structure (the encoding) are provided (e.g. Wiuf and Hein, 1999b; Gusfield, 2014;

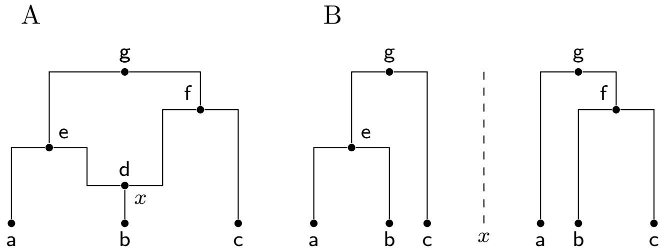

Figure 2: A classical event ARG (eARG). (A) Standard graph depiction with breakpoint x associated with the recombination node d. Nodes e, f and g are common ancestor events. (B) Corresponding local trees to the left and right of breakpoint x (note these are shown in the conventional form in which only coalescences within the local tree are included; see Section 5 for a discussion of this important point).

Hayman et al., 2023) they follow the approach described by Griffiths and colleagues (but see Parida et al. (2011) and Zhang et al. (2023) for notable exceptions), and a large number of ARG inference methods use it as an output format (e.g. Song and Hein, 2004; Song et al., 2005; Rasmussen et al., 2014; Heine et al., 2018; Ignatieva et al., 2021). In this Griffiths encoding we have two types of internal node in the graph, representing the common ancestor and recombination events in the history of a sample. At common ancestor nodes, the inbound lineages merge into a single ancestral lineage with one parent, and at recombination nodes a single lineage is split into two independent ancestral lineages. Recombination nodes are annotated with the corresponding crossover breakpoints, and these breakpoints are used to construct the local trees. This is done by tracing pastwards through the graph from the samples, making decisions about which outbound edge to follow through recombination nodes based on the breakpoint position (Griffiths and Marjoram, 1996). Because it is focused on recording events and their properties, we will refer to this Griffiths encoding as the "event ARG" or eARG encoding. Fig. 2 shows an example of a classical eARG with three sample genomes (a, b, and c), three common ancestor events (e, f, and g) and a single recombination event (node d) with a breakpoint at position x. Assigning a breakpoint to a recombination node is not sufficient to uniquely define the local trees, and either some additional ordering rules (e.g. Griffiths and Marjoram, 1996) or explicit information (e.g. Gusfield, 2014; Ignatieva et al., 2021) is required to distinguish the left and right parents. We assume in Fig. 2 that d inherits genetic material to the left of x from e and to the right of x from f.

While the Griffiths approach of annotating recombination nodes with a breakpoint in an eARG is a concise and elegant way of describing realisations of the coalescent, it is inherently limited when implemented literally. The eARG encoding explicitly models only two different types of event and thus anything that is not a single crossover recombination or common ancestor event, must be incorporated either in a roundabout way using these events, or by adding new types of event to the encoding. For example, gene conversion could be accommodated either by stipulating a third type of event (annotated by two breakpoints and corresponding traversal conventions for recovering the local trees) or by two recombination nodes joined by a zero-length edge. From the perspective of practical interchange of data between inference methods and downstream applications, both workarounds are problematic, and the gARG encoding described in the previous section offers a much simpler solution.

Aside from these obvious practical challenges arising from a literal implementation of the Griffiths approach, there is a deeper issue with the implicit strategy of basing an ARG data structure on recording events and their properties (e.g. the crossover breakpoint for a recombination event). The fundamental problem is that this approach assumes all events are *knowable*, and does not provide any obvious mechanism for either aggregating multiple events or expressing uncertainty about them. While this is not a problem when describing the results of simulations (where all details are perfectly known), it is a major issue when we wish to formally describe the output of inference methods, particularly as datasets approach the population scale. As discussed in the introduction, the precise details of all events in these vast ARGs can never be known, and a data structure that *enforces* complete precision is therefore an impediment to progress.

There is also a certain clarity gained by explicitly modelling nodes in the inheritance graph as genomes. Outside of the context of a mathematical model, an "event" is a slippery concept. For example, *which* genome along a chain of cell divisions should be regarded as the one where an event occurred, or whether multiple coalescences within a single individual should be regarded as one or multiple events are debatable points (Appendix D). From the perspective of a concrete data structure, ideally forming the basis of an ecosystem of interoperable inference and analysis methods, such debates are unproductive.

4 Ancestral material and sample resolution

Ancestral material (Wiuf and Hein, 1999a,b) is a key concept in understanding the overall inheritance structure of an ARG (here, and throughout, we use the general term "ARG" when the details of the specific encoding are not important). It denotes the genomic intervals ancestral to a set of samples on the edges of an ARG. For example, in Fig. 1 we have four sample genomes, a–d. As we trace their genetic ancestry into the previous generation (e–h), we can think of their ancestral material propagating through the graph backwards in time. In the region $[2, 7)$, there is a local coalescence where nodes a and c find a common ancestor in f. Thus, in this region, the total number of genome segments that are ancestral to the sample is reduced from four to three. Fig. 1A illustrates this by (shaded) ancestral material being present in only three nodes (f, g, and h) in this region, while node e is blank as it carries *non-ancestral* material. This process of local coalescence continues through the graph, until all samples reach their most recent common ancestor in node n.

The process of tracking local coalescences and updating segments of ancestral material is a core element of Hudson's seminal simulation algorithm (Hudson, 1983b; Kelleher et al., 2016), and the key distinguishing feature between the "Big" and "Little" ARG stochastic processes (see Appendix B). The ability to *store* resolved ancestral material is also a key distinction between the eARG and gARG encodings. Because an eARG stores only the graph topology and recombination breakpoints, there is no way to locally ascertain ancestral material without traversing the graph pastwards from the sample nodes, resolving the effects of recombination and common ancestor events.

Efficiently propagating and resolving ancestral material for a sample through a pre-existing graph is a well-studied problem, and central to recent advances in individual-based forward-time simulations (Kelleher et al., 2018; Haller et al., 2018). In contrast to the usual "retrospective" view of ARGs discussed so far, these methods record an ARG forwards in time in a "prospective" manner. Genetic inheritance relationships and mutations are recorded exhaustively, generation-by-generation, leading to a rapid build-up of information, much of which will not be relevant to the genetic ancestry of the current population. This redundancy is periodically removed using the "simplify" algorithm (Kelleher et al., 2018), which propagates and resolves ancestral material. Efficient simplification is the key enabling factor for this prospective-ARG based approach to forward-time simulation, which can be orders of magnitude faster than standard sequence-based methods (see Section 7 for other applications of ARG simplification). We refer to a gARG that has been simplified with respect to a set of samples, such that the inheritance annotations on its edges contain no non-ancestral material, as sample-resolved.

Any eARG can be converted to a sample-resolved gARG via a two-step process illustrated in Fig. 3. The first step is to take the input eARG (Fig. 3A), duplicate its graph topology, and then add inheritance annotations to each of the gARG's edges (Fig. 3B) as follows. If a given node is a common ancestor event, we annotate the single outbound edge with the interval $[0, L)$, for a genome of length L. If the node is a recombination event with a breakpoint x, we annotate the two outbound edges respectively with the intervals $[0, x)$ and $[x, L)$. These inheritance interval annotations are clearly in one-to-one correspondence with the information in the input eARG. They are also analogous to the inheritance intervals we get on the edges in a prospective gARG produced by a forward-time simulation, which are concerned with recording the direct genetic relationship between a parent and child genome and are not necessarily minimal in terms of the resolved ancestral material of a sample. Thus, the final step is to use the "simplify" algorithm to perform the required sample resolution (Fig. 3C).

The sample-resolved gARG of Fig. 3C differs in some important ways to the original eARG (Fig. 3A). Firstly, we can see that some nodes and edges have been removed entirely from the graph. The "grand MRCA" q is omitted from the sample-resolved gARG because all segments of the genome have fully coalesced in k and p before q is reached. Likewise, the edge between g and j is omitted because the recombination event at position 5 (represented by node g) fell in non-ancestral material.

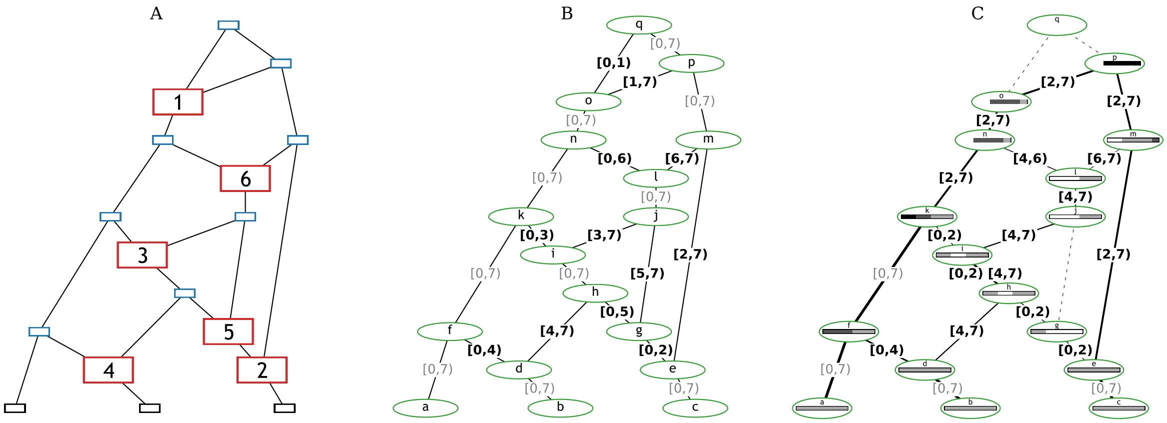

Figure 3: Converting the Wiuf and Hein (1999b, Fig. 1) example to a sample-resolved gARG. (A) The original eARG; square nodes represent sampling (black), common ancestor (blue), and recombination (red) events; the latter contain breakpoint positions. (B) The corresponding gARG with breakpoints directly converted to edges annotated with inheritance intervals. (C) The sample-resolved gARG resulting from simplifying with respect to the sample genomes, a, b, and c. Dashed lines show edges that are no longer present (in practice, nodes g, j, and q would also be removed). Coalescence with respect to the sample is indicated by shaded bars, as in Fig. 1A; nodes n, o, p, q have truncated bars showing that local ancestry of entirely coalesced regions is omitted. Line thickness is proportional to the genomic span of each edge. Nodes representing recombination events are retained for clarity, but could be removed by simplification if desired.

More generally, we can see that the sample resolved gARG of Fig. 3C allows for “local” inspection of an ARG in a way that is not possible in an eARG. Because the ancestral material is stored with each edge of a gARG, the cumulative effects of events over time can be reasoned about, without first “replaying” those events. Many computations that we wish to perform on an ARG will require resolving the ancestral material with respect to a sample. The gARG encoding allows us to perform this once and to store the result, whereas the eARG encoding requires us to repeat the process each time.

Note that the Wiuf and Hein (1999b) eARG in Fig. 3 is not particularly representative, because inference or simulation methods usually only generate ARGs containing nodes and edges ancestral to the sample (but see the discussion of the “Big ARG” stochastic process in Appendix B). Nonetheless, it is an instructive example from the literature which highlights several important properties of ARGs, and the general point about the need to resolve ancestral material “on the fly” for eARG traversals holds.

5 ARGs and local trees

The relationship between an ARG and its corresponding local trees is subtle and important. A fundamental property of genetics is that a given DNA nucleotide is inherited from exactly one parent genome, both at an organismal and cell-by-cell level (Appendix D). These paths of single-parent inheritance give rise, by definition, to a tree structure. As a result of recombination, adjacent nucleotides can have different paths of inheritance, and an ARG encodes the entire ensemble of local trees along the genome for a given set of sample nodes. Precisely defining the process by which local trees are extracted from an ARG is essential to our understanding of how ARGs and local trees are related, and we require a concrete mathematical structure to describe the local trees. It is important to note that although the following discussion is phrased in terms of the gARG encoding, the arguments apply equally to eARGs because any eARG can be converted to a gARG without loss of information (Section 4).

Oriented trees provide a convenient formalism to capture these parent-child relationships in a well-defined combinatorial object. Let $\pi_1 \dots \pi_n$ be a sequence of integers, such that π_u denotes the parent of node u, and $\pi_u = 0$ if u is a root (Knuth, 2011, p. 461). This encoding is particularly useful to describe evolutionary trees because parent-child relationships are important but the ordering of children at a node is not (Kelleher et al., 2013, 2014, 2016). Thus, for a given gARG with nodes $\{1,\dots,n\}$ and edges E (Section 2), we recover the local tree at position x as follows. We begin by setting $\pi_u = 0$ for each $1 \leq u \leq n$. Then, for each sample node in S we trace its path pastwards through the ARG for position x, and record this path in π. Specifically, at a given node u, we find an edge $(c, p, I) \in E$ such that $u = c$ and $x \in I$, and set $\pi_c \leftarrow p$. We then set $u \leftarrow p$, and repeat until either $\pi_u \neq 0$ (indicating we have traversed this section of the ARG already on the path from another sample) or there is no matching outbound edge (indicating we are at a root). Note that the local trees for an ARG are "sparse" (Kelleher et al., 2016), because many ancestral nodes will not be reachable from the samples at a given position (so their corresponding entries in π will be zero).

This combinatorial approach provides at least one novel insight, clarifying the fundamental relationship between ARGs and local trees. Suppose we are given a gARG defined by a set of nodes and edges. There is no requirement on the structure of this ARG beyond the basic definitions: it could correspond to an ARG in which every recombination event is exactly specified (e.g. Fig. 3) or one in which local trees are entirely disjoint (i.e. only the sample nodes are shared between them). If we are given the sequence of local trees for this gARG encoded as an oriented tree, along with the genome interval covered by each tree, we can recover the original gARG exactly. More formally, suppose we are given the local tree $\pi^x_1 \dots \pi^x_n$ for each nucleotide position $1 \leq x \leq L$ on a genome of length L. Then, the edges of the "local ARG" for this tree is given by $E^x = \{(u, \pi^x_u, \{x\}) \mid \pi^x_u \neq 0\}$. Because the ARG edges are defined by (c, p, I) tuples, where the set I defines the positions over which node c inherits from parent p, we can then simply combine the "local ARGs" for each position x to recover precisely the same set of edges as the original ARG. Thus, under this definition, there is a one-to-one correspondence between an ARG and the sequence of local trees that it encodes.

This is not the prevailing view, however. Kuhner and Yamato (2017) argue that the "interval-tree" representation of an ARG (the local trees and the genome intervals they cover) "does not contain all of the information in the underlying ARG: it lacks the number of recombinations occurring at each site, the times at which recombinations occurred, and the specific sequences involved as recombination partners." Shipilina et al. (2023) discuss the same ideas, and note that the "full ARG... contains more information than the series of tree sequences along the genome". These statements that an ARG contains more information than its local trees are true if we represent local trees in their conventional forms, but these forms discard important information that is available in an ARG.

There are two properties of how evolutionary trees are conventionally represented that lead to this disagreement about the relationship between local trees and an ARG. Firstly, the internal nodes of evolutionary trees are usually considered to be *unlabelled*, or equivalently, labelled by the leaves which they subtend. The same canonical labelling cannot be used for internal ARG nodes because the leaves they subtend will typically vary by genomic position. If we do not label the tree nodes in a way which is persistent across the sequence of local trees in the ARG, we lose the fact that the *same* ancestors sometimes persist across multiple trees. Defining ARG nodes as integers and using the oriented tree encoding explicitly labels internal nodes, and makes the relationship between tree and ARG nodes clear and precise.

The second property of how evolutionary trees are conventionally represented that is unhelpful in the context of ARGs is their focus on branching points (coalescences), i.e. nodes that have two or more children. As the introductory paragraph of this section emphasised, parent-child relationships are what fundamentally define a tree, and branching points can be seen as incidental. This is reflected by the oriented tree encoding which simply stores the local parent-child relationships, and does not, for example, directly tell us how many children a particular node has. The local tree at a given position records the *path* through the ARG; if this path omits nodes that are not branching points (such as `e` in Fig. 1), we lose information about the ARG. We return to this point in the following two sections, when we discuss "locally unary" nodes and the simplification process.

It is important that we make the distinction here between the local trees that we can derive from a known ARG (as just discussed), and an ARG that we can derive from a sequence of *estimated* local trees. The ARG inference method `Espalier` (Rasmussen and Guo, 2022) is illustrative in this context. It begins by splitting an input sequence alignment into segments that are assumed to be non-recombining. Within each segment, an initial local tree is estimated using standard phylogenetic methods. By necessity, these local trees will contain internal nodes that are unlabelled and consist only of branching points: there is no information shared between the independent tree estimation steps across segments. Part of the task of stitching these trees together into an ARG is then, essentially, to generate labels for the internal nodes, and decide which nodes persist across multiple local trees. `Espalier` approaches this task by identifying maximal subtrees that do not change between pairs of adjacent local trees and then heuristically exploring the space of possible rearrangements of these subtrees. To derive details about recombination events, `Espalier` then attempts to infer the precise subtree prune-and-regraft (SPR) operations (Hein, 1990; Song, 2003, 2006) induced by recombination between these partially reconciled local trees. Inferring the SPRs between leaf-labelled trees is NP-hard (Hein et al., 1996; Allen and Steel, 2001; Bordewich and Semple, 2005), but it is unclear what the complexity is when there is a degree of internal node sharing between trees. The combinatorial formulation of ARGs and local trees provided here may help clarify these fundamental questions.

6 Locally unary nodes

As discussed in the previous section, the local tree at a given position x is best seen as the path through the ARG at that position, defined by the oriented tree $\pi_1^x \dots \pi_n^x$. This path does not directly contain information about branching points, and defining a node's arity (number of child nodes) is therefore useful. The "local arity" of a node is the number of children it has in the local tree at position x, i.e., $a_u^x = |\{v : \pi_v^x = u\}|$ for each $1 \leq u \leq n$. The "ARG arity" of a node u is the number of children it has in the graph topology, i.e. $a_u = |\{v : (v, u, I) \in E\}|$. Thus, the local arity is less than or equal to the ARG arity (more precisely, $0 \leq a_u^x \leq a_u$), and the local arity of a node may change as we move along the genome.

This distinction between ARG and local arity is mainly of interest when we consider nodes that have a single child: those that are *unary*. Returning to the example in Fig. 1, nodes `g` and `h` are ARG-unary (Fig. 1B), and are consequently also unary in the local trees (Fig. 1C). On the other hand, node `f` has two children in the graph, but is binary only in the local tree covering the interval $[2, 7)$, representing the coalescence of samples `a` and `c` in this genome region. Over the interval $[0, 2)$ no coalescence occurs, but we still record the fact that genome `c` inherits from `f` in the local tree. Thus, node `f` has a single child in this interval: it is *locally unary*. In another example, `e` is binary in the graph, being a common ancestor of `a` and `c`, but is locally unary in all trees in which it is present. This is because no ancestral material coalesces in `e`: `a` inherits genetic material from the far left hand end of `e`, while `c` only inherits the (disjoint) right hand end.

By definition, ARG-unary nodes have one child but can have one or more parents. A node with one child and only one parent represents a "pass-through" node: these occur where we wish the record the passage of ancestral material through a known node. For example, in simulations it is sometimes useful to record the passage of ancestral material through known pedigree individuals regardless of whether common ancestry occurs. Nodes with one child and two parents arise when we model a recombination event using a single node in the classical manner (e.g. Fig. 3). It is also possible for sample nodes to be ARG-unary, for example in inferences from longitudinal datasets where genetic data is sampled at many timepoints and recombination is rare (e.g. SARS-CoV-2; see Discussion).

More generally, locally unary nodes, which can have one or more children in the graph, are a common and important feature of many different types of ARG. As discussed in the previous section, without these nodes marking the passage of ancestral material through specific ancestors, the local trees lack information about events other than local coalescence. For example, the local trees for the classical event ARG depicted in Fig. 2B follow the usual conventions and do not include any information about the recombination that occurred at node `d`. Given these two local trees in isolation we lack specific information about the recombination. Explicitly recording that node `d` lies on the branch joining `b` to `e` in the left hand tree, and `b` to `f` in the right hand tree resolves all ambiguity, and makes the collection of local trees exactly equivalent to the corresponding ARG (see previous section). Unary nodes are a vital link between ARGs and local trees, and we cannot fully reason about how a local tree is embedded in an ARG without them. As we see in the next two sections, both ARG and locally unary nodes occur in various scenarios, and are produced by a range of current inference methods.

7 Levels of simplification

ARG simplification is a powerful tool. In general, we can think of simplification as the process of removing nodes and re-writing edges (and their inheritance annotations) to remove various types of redundancy, much of which revolves around the presence of unary nodes (see previous section). We illustrate this successive removal of redundancy through a series of simplification steps in Fig. 4.

The ARG in Fig. 4A is the output of a backwards-time Wright-Fisher simulation for a sample of two diploid individuals (population size $N = 10$), and follows a similar process to the methods described by Nelson et al. (2020). As we proceed backwards in time, generation by generation, the extant lineages choose parents randomly. With a certain probability recombination occurs, and the ancestral material of a lineage is split between the two parental genomes. Local coalescence occurs when lineages with overlapping ancestral material choose the same parent genome. Note that in this simulation we do not explicitly model recombination *events* via an ARG node, but simply record the *outcome* of a recombination via edges to the parent's two genomes. Thus, a recombinant node such as g in Fig. 4 may also correspond to a coalescence. The distinction of using a single node to represent a recombination event, as is done in Fig. 3, or two to represent the parent genomes, as in Fig. 4, is usually not important. Either is possible in the gARG encoding, and the most convenient approach will vary by application (discussed e.g. in Appendix B). Note also that node k in Fig. 4 has three children. Polytomies like this are a natural feature of such a Wright-Fisher model (but see Appendix D).

The graph visualisations in Fig. 4 have three novel features which require some explanation. Firstly, edge weights (the thickness of the lines joining nodes) correspond to the length of the inheritance intervals they are annotated with. This allows us to distinguish edges that persist across many local trees from those that are less influential (contrast the edge (g, h) with (g, i) in Fig. 4A). Secondly, node colours denote the number of parents that they have in the graph, allowing us to easily see roots (those with zero parents), recombinants (those with two parents) and more complex situations arising from simplification (see below). Thirdly, the shading intensity of a node denotes the "coalescent span", the fraction of the node's span (the length of genome in which it is reachable from the samples in the local trees) over which it has more than one child. Nodes which are never locally unary therefore have a coalescent span of 100%, whereas nodes in which ancestral material never coalesces have a coalescent span of 0%.

Returning to the main topic of this section, Fig. 4A is the original simulation output, in which we retain all nodes involved in recombination or common ancestry events. This is the true history, and contains a very high level of detail, some of which may be considered redundant (or, from another perspective, unobservable). In Fig. 4A the local trees (right) contain many unary nodes, fewer as we successively simplify (Fig. 4B,C), until we reach Fig. 4D, where there are none.

The first level of simplification that we can perform is based only on the graph topology. An example of graph topology that we may consider redundant (or non-identifiable) is a "diamond" (Rasmussen et al., 2014) in which the two parent nodes of a recombination immediately join again into a common ancestor (e.g. j, l, m and n in Fig. 4A). Unless we are specifically interested in the recombination event or these ancestral genomes, the diamond can be replaced by a single edge without loss of information. More generally, any subgraph that is singly-connected in both the leafward and rootward direction (a "super-diamond") can be replaced by one edge. This definition includes the case of a node that has one inbound and one outbound edge, such as nodes f and h. Fig. 4B shows the result of this type of graph topology simplification.

Simplifying away diamonds will remove many unary nodes from the local trees, but there can still be nodes that are unary in all of the local trees. In particular, a node can represent a recombinant with multiple parents in the graph but only a single child (e.g. node n in Fig. 4B), or can represent a common ancestor with multiple children in the graph but in which no coalescence takes place in the local trees (node r in Fig. 4B). The distinction between the "common ancestry" of two or more genomes in an ancestral genome and the "coalescence" which may or may not occur in the local trees is important (Hudson, 1983b; Kelleher et al., 2016). Consider e in Fig. 4A, for example. We can see from the graph that it is a common ancestor of samples a and b, but it does not correspond to any coalescence in the local trees to the left of position 44, and is therefore unary in these three trees. Such nodes are not singly connected in the graph, but are nevertheless unary in all of the local trees. The operation to remove them therefore requires knowledge not just of the graph topology but also of

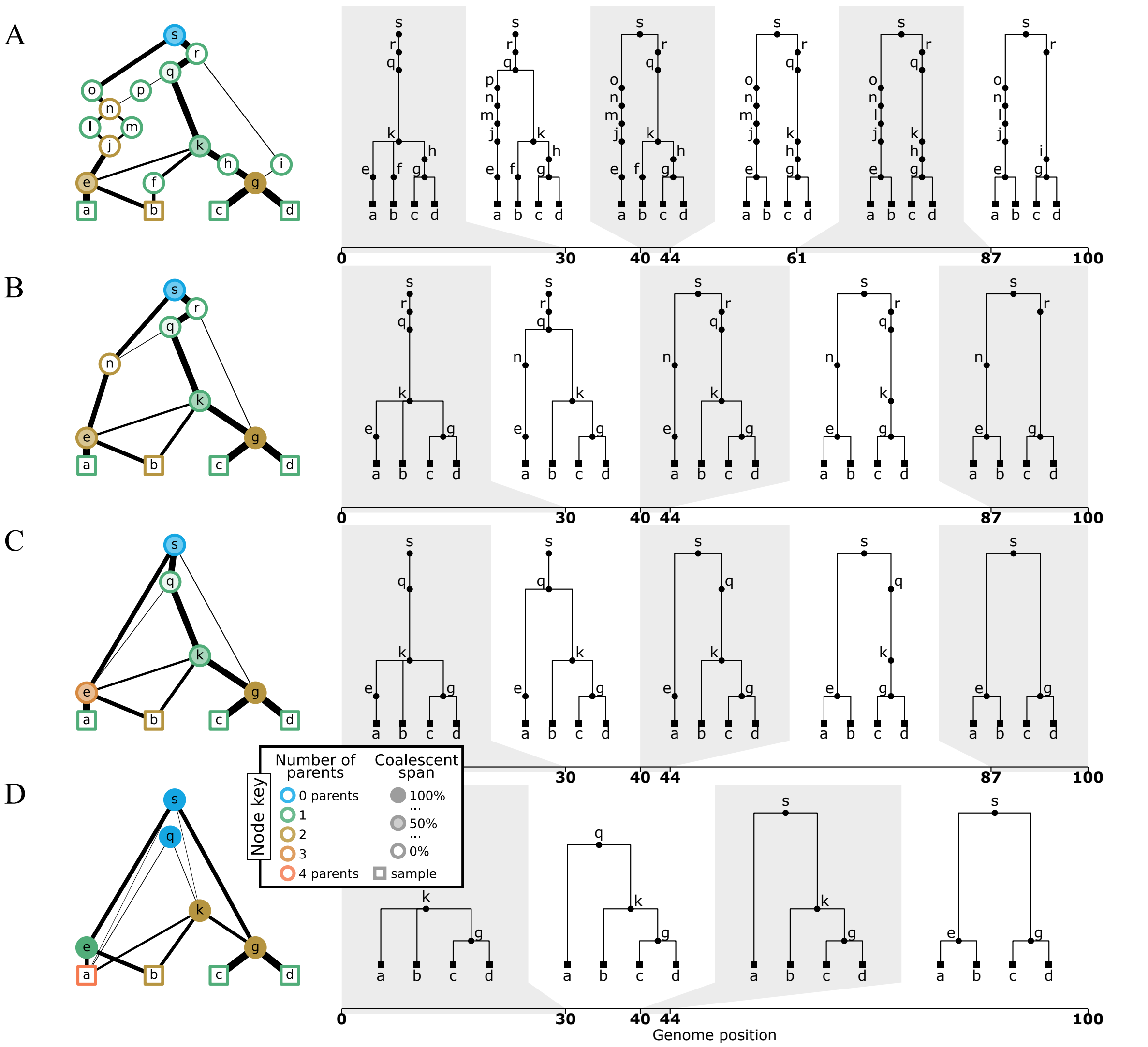

Figure 4: Levels of ARG simplification. (A) An example gARG simulated from a diploid Wright-Fisher model. (B) Remove all singly-connected graph components (e.g., diamonds such as jlnm). (C) Remove nodes that never represent coalescences, i.e. are unary everywhere (e.g. n and r). (D) Rewrite edges to bypass nodes in local trees in which they are unary. In each case, the graph is shown on the left and corresponding local trees on the right. In the interest of visual clarity, inheritance intervals are not shown on the graph edges; Supplementary Fig. S1 shows the graphs with these inheritance intervals included. Graph nodes are coloured by the number of parents and shaded according to the proportion of their span over which they are coalescent; see the text for more details.

the ancestral material associated with the edges. As we see in Fig. 4C, removal of recombinant nodes can produce graph nodes with more than two parents (e.g. node e); and likewise, removal of common ancestor but non-coalescent nodes can produce graph nodes with more than two children (e.g. node s). Both cases represent the merged *effects* of multiple evolutionary events in a single node (genome), and the ARG no longer contains the intermediate genomes corresponding to those events (see also Appendix D).

The remaining nodes are MRCAs of some subset of the samples at *some* positions along the genome. We still have some unary nodes in the local trees, but these nodes will correspond to a coalescence in at least one other local tree. For example, node k is unary in the fourth tree of Fig. 4C, but is either binary or ternary in all other local trees (recall this is a Wright-Fisher simulation). The final level of simplification is to alter the edge annotations such that, although no nodes are removed from the graph, all unary nodes disappear from the local trees (Fig. 4D). Note that although this last stage produces simpler local trees, by removing information about the exact paths taken by lineages through the graph, we lose potentially useful information about shared edges between trees. The msprime simulator, and the version of Hudson's algorithm described by Kelleher et al. (2016), produces ARGs that are fully simplified (i.e., contain no locally unary nodes). It is not difficult, however, to update these methods to record information about the passage of ancestral material through genomes under a range of conditions.

An important consequence of simplifying ARGs to remove unary nodes in local trees is that we lose some information about recombination events. This is related to the amount of *precision* about recombination events that we store and can hope to infer from sampled genomes, which is the topic of the next section.

8 Precision of recombination information

As illustrated in Fig. 4, successive levels of ARG simplification reduce the amount of information about the history of the sample that is stored. Some of the information lost, e.g. "diamond" removal (Fig. 4A), seems like a reasonable tradeoff for a simpler structure. The consequences of other simplifications, however, are more subtle and relate directly to what can be known about recombination events and the levels of precision that we should seek to infer about them.

The ARGs in Fig. 4 contain different numbers of local trees (6, 5, 5 and 4 respectively for A through D). When we move from A to B the local trees for the intervals $[44, 61)$ and $[61, 87)$ are merged because the only differences between them are their paths through nodes l and m. These nodes that participated in the diamond are removed from the ARG, and we have lost all information about the corresponding recombination at position 61. Other nodes (e.g. o and p) have also been removed but these represent the *parents* of recombinants. The recombinant nodes themselves (e.g. n) are still present, and represent precise information about the time, genomic location and lineages involved in the recombination event.

Fig. 4C has the same number of local trees as Fig. 4B, but has less precise information about recombination. Continuing the previous example, node n has been removed from the graph because it was unary in all of the local trees; its outbound edges to s and q have effectively been "pushed down" to e (which is retained because it is the coalescent parent of a and b over the interval $[44, 100)$). We have therefore lost precision about the *timing* of this recombination event, and know only that it must have occurred between the times of node e and q.

Fig. 4D removes all unary nodes from the local trees, and further reduces the precision of recombination information. Node e has not been removed from the graph because it is coalescent in the final tree, but we no longer know that the recombination event at position 30 was ancestral to it, or have any indication of its timings. Furthermore, trees for $[44, 87)$ and $[87, 100)$ were only distinguishable by the passage of the former tree through nodes e and q, and so the recombination on node g at position 87 has been lost entirely.

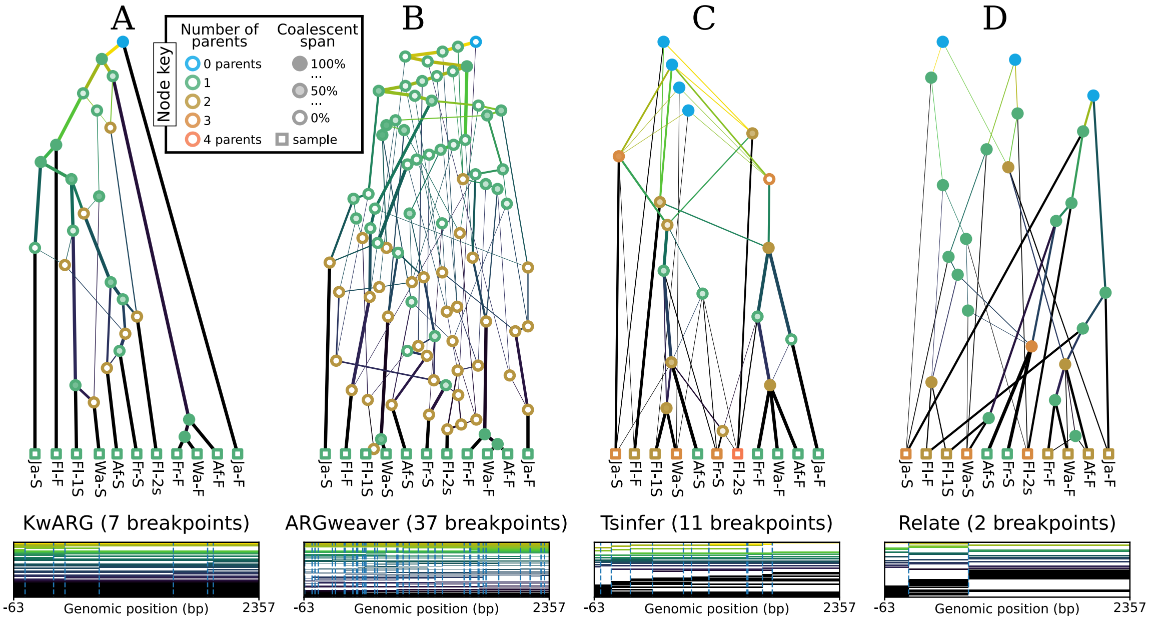

Figure 5: Inference of sample-resolved ARGs for 11 *Drosophila melanogaster* DNA sequences over a 2.4kb region of the ADH locus (Kreitman, 1983). Results for four different methods: (A) KwARG; (B) ARGweaver; (C) tsinfer; and (D) Relate. See the text for details of these methods. Edge colours indicate time of the edge's child node (lighter: older; darker: younger). Vertical and horizontal positions of graph nodes are arbitrary. Line width and node colour are as described in Fig. 4. Bottom row graphics show the genome positions, relative to the start of the ADH gene, for each graph edge from the corresponding ARG. Edge intervals are drawn as horizontal lines, stacked in time order (edges with youngest children at the bottom); vertical dashed lines denote breakpoints between local trees.

9 Example inferred ARGs

The scalability gains made by recent ARG inference methods such as Relate (Speidel et al., 2019) and tsinfer (Kelleher et al., 2019b) have been, in part, due to inferring lower levels of precision about recombination than classical methods. Neither method infers explicit recombination events, and therefore their outputs cannot be described using the classical eARG formalisms (Section 3). Nonetheless, both methods produce estimates in which nodes and edges persist across multiple trees, creating inheritance graphs which fit naturally into the gARG formulation. To illustrate the varying levels of information captured by current methods, and some qualitative differences between them, Fig. 5 shows graphical depictions of example ARGs produced by four tools using substantially different inference strategies.

The first two methods explicitly infer recombination events. KwARG (Ignatieva et al., 2021) is a parsimony based approach which searches the space of plausible ARGs, outputting minimal ones using heuristics. ARGweaver (Rasmussen et al., 2014) on the other hand is model-based, sampling from a discretised version of the SMC (McVean and Cardin, 2005; Marjoram and Wall, 2006). Note that both KwARG and ARGweaver produce many ARGs, and those shown in Fig. 5 are arbitrarily selected examples. While the second two methods both produce a single best-guess estimate and do not explicitly infer recombination events, they are based on quite different principles. Tsinfer works in a two-step process, first generating ancestral haplotypes via heuristics and then inferring inheritance relationships between them using the Li and Stephens model (Li and Stephens, 2003). Relate first reconstructs local tree topologies across the genome, using a variant of the Li and Stephens model to estimate the ordering of coalescence events in each tree, and then estimates branch lengths using MCMC with a coalescent-based prior. See Appendix C for more details on these and other inference methods.

Inferred ARGs are based on the Kreitman (1983) dataset, a standard benchmark in the classical

ARG literature. It consists of 43 biallelic SNPs spanning 2.4Kb of the *D. melanogaster* ADH locus on chromosome 2L. Where required for inference purposes we assume mutation and recombination rates of 5.49×10^{-9} and 2.40463×10^{-9} per site per generation (Schrider et al., 2013; Comeron et al., 2012) and a constant effective population size of 1,720,600 (Li and Stephan, 2006), as provided by the `stdpopsim` catalog (Adrion et al., 2020; Lauterbur et al., 2023). Software versions were `KwARG` v1.0, `ARGweaver-D` (2019), `tsinfer` v0.3.1, and `Relate` v1.1.9. Full details and code for generating these figures are available on GitHub (see Data Availability).

Considering Fig. 5, we can see that there is substantial variation in the number of recombination breakpoints inferred by different methods, with e.g. `ARGweaver` suggesting far more than the 7 required for this dataset under minimal parsimony assumptions (Song and Hein, 2003). A sense of the amount of recombination in each ARG is provided by the node colouring scheme, which shows the number of parents for each node. In Fig. 5A,B, each recombination event corresponds to a node with exactly two parents and one child. As these methods explicitly infer a recombination event for each breakpoint, the number of breakpoints equals the number of two-parent (brown) nodes. In contrast, Fig. 5C,D do not have this straightforward relationship between the number of nodes with multiple parents and number of breakpoints along the genome. In both ARGs the number of breakpoints is smaller than the number of multiple-parent ARG nodes, showing that several multiple-parent nodes must share breakpoint positions. There are also ARG nodes with multiple parents and multiple children, where one or more recombinations have been pushed down onto a more recent node. As a consequence, it may be difficult to condense each transition between trees in these ARGs into a set of SPR operations.

Shading within nodes in Fig. 5 indicates the fraction of the node's span over which it is coalescent (Section 6). For example, brown nodes in Fig. 5A,B are clear because there is no local coalescence at these recombination nodes (they are "ARG unary", and so local coalescence is impossible). The significant number of partially shaded nodes in Fig. 5A,B and C demonstrates that the `KwARG`, `ARGweaver` and `tsinfer` ARGs all contain locally unary nodes. Another difference between methods highlighted in this figure is the presence of polytomies, which only `tsinfer` creates. The most obvious example involves nodes Fr-F, Wa-F, and Af-F, which happen to have identical sequences. Because `KwARG`, `ARGweaver`, and `Relate` require bifurcating trees by design, each picks an arbitrary order of branching (hence Fig. 5A and B disagree in this order, and Fig. 5D even shows different orders in different trees).

The bottom row of Fig 5 shows the extent along the genome to which graph edges are shared between multiple trees. All four methods infer nodes and edges that are shared between multiple trees, to varying degrees. For example, all of the methods infer infer that Af-f, Fr-f, and Wa-f form a clade along the entire sequence. In particular, we can see both `tsinfer` and (to a lesser extent) `Relate` have edges that span multiple tree boundaries, indicating that they are not inferring a series of *unrelated* local trees. However, in comparison to `KwARG` and `ARGweaver` neither method results in extensive node sharing in the oldest time periods. Overall, Fig. 5 shows that `tsinfer` and `Relate` ARGs contain a level of detail that lies somewhere between a sequence of unrelated local trees on one extreme and an ARG with precisely specified recombination events on the other (Fig 5A,B).

10 Implementation and efficiency

The gARG encoding defined here leads to highly efficient storage and processing of ARG data, and has already been in use for several years. The succinct tree sequence data structure (usually known as a "tree sequence" for brevity) is a practical gARG implementation focused on efficiency. It was originally developed as part of the `msprime` simulator (Kelleher et al., 2016) and has subsequently been extended and applied to forward-time simulations (Kelleher et al., 2018; Haller et al., 2018), inference from data (Kelleher et al., 2019b; Wohns et al., 2022; Zhan et al., 2023), and calculation of population genetics statistics (Ralph et al., 2020). The succinct tree sequence encoding extends the basic definition of a gARG provided here by stipulating a simple tabular representation of nodes and edges, and also defining a concise representation of sequence variation using the "site" and "mutation" tables. The key property of the succinct tree sequence encoding that makes it an efficient substrate for defining analysis algorithms is that it allows us to sequentially recover the local trees along the genome very efficiently, and in a way that allows us to reason about the *differences* between those trees (Kelleher et al., 2016; Ralph et al., 2020).

The `tskit` library is a liberally licensed open source toolkit that provides a comprehensive suite of tools for working with gARGs (encoded as a succinct tree sequence). Based on core functionality written in C, it provides interfaces in C, Python and Rust. Tskit is mature software, widely used in population genetics, and has been incorporated into numerous downstream applications (e.g., Haller and Messer, 2019; Speidel et al., 2019; Adrion et al., 2020; Terasaki Hart et al., 2021; Baumdicker et al., 2022; Fan et al., 2022; Guo et al., 2022; Korfmann et al., 2023; Mahmoudi et al., 2022; Petr et al., 2022; Rasmussen and Guo, 2022; Zhang et al., 2023; Nowbandegani et al., 2023; Ignatieva et al., 2023; Fan et al., 2023). The technical details of `tskit`, and how it provides an efficient and portable platform for ARG-based analysis, are beyond the scope of this manuscript. In the interest of avoiding confusion, however, we list a few minor details in which the formal details of gARGs provided in Section 2 differ from their practical implementation in `tskit`. Firstly, "edges" in tree sequence terminology would perhaps be better described as "edge-intervals", as each describes a single contiguous interval of genome inheritance between a pair of nodes. This denormalisation of the gARG data model is for efficiency purposes. Secondly, zero- rather than one-based indexing is used for nodes in ARGs and oriented trees; consequently -1 is used to denote the presence of roots (rather than 0 as used here for notational simplicity).

11 Discussion

Recent breakthroughs have finally made large-scale ARG inference feasible in practice, leading to a surge of interest in inference methods, their evaluation, and their application to biological questions. The prospect of ARGs being used routinely within population and statistical genetics is tantalising, but in reality there is substantial work to be done to enable this. A necessary first step is a degree of terminological clarity. As discussed in Appendix A, the term "ancestral recombination graph" has several subtly different interpretations, depending on context. The trend to decouple ARGs from their original definition within the context of stochastic processes and instead use the term as a more general representation of any recombinant genetic ancestry seems useful; we have tried to clarify and systematise it here. Thus we can think of an ARG as any structure that encodes the reticulate genetic ancestry of a sample of colinear sequences under the influence of recombination. The "genome" ARG (gARG) encoding made explicit here is one way we can concretely define such recombinant ancestry, which we have shown is both flexible and efficient. The flexibility of the gARG encoding contrasts with the classical "event" ARG (eARG) encoding, which is more limited in what can be described. Importantly, gARGs do not require fully precise estimates of ancestral recombination events, and allow us to directly express important forms of temporal uncertainty.

Fully decoupling the general concept of an ARG from the coalescent with recombination (henceforth, "coalescent") is an important step. While the coalescent has proven to be a useful and robust model (Wakeley et al., 2012; Bhaskar et al., 2014; Nelson et al., 2020), many modern datasets have properties that grossly violate its assumptions. One key assumption is that sample size n is much less than the effective population size, N_e. Several human datasets now consist of hundreds of thousands of genomes (Turnbull et al., 2018; Bycroft et al., 2018; Karczewski et al., 2020; Tanjo et al., 2021; Halldorsson et al., 2022), and so sample size is an order of magnitude *larger* than the usually assumed N_e values. Agricultural datasets are an even more extreme departure from this assumption, with hundreds of thousands of samples embedded in multi-generational pedigrees (Hayes and Daetwyler, 2019; Ros-Freixedes et al., 2020) and effective population sizes of 100 and even less (MacLeod et al., 2013; Makanjuola et al., 2020; Hall, 2016; Pocrnic et al., 2016). A model assuming a single N_e would be a drastic over-simplification of course, but even if sufficiently complex demographic models (Gower et al., 2022) encompassing hundreds of populations, explosive growth rates and myriad interconnections of migration, were somehow estimated and provided as input, ARGs sampled from the coalescent cannot capture the complexities of family structure in these datasets (e.g. Turnbull et al., 2018; Ros-Freixedes et al., 2020). Another core assumption of the coalescent model is that the genome (or at least the region under study) is short enough that the number of extant lineages remains much smaller than N_e at all times. High-quality whole genome assemblies are now available for many species and projects are under way to obtain them for tens of thousands more (Darwin Tree of Life Project Consortium, 2022; Lewin et al., 2022), and so we can expect inferred ARGs to routinely span large fractions of a chromosome.

Recent large-scale methods have simplified the inference problem by making a single, deterministic best-guess at ARG inference (Kelleher et al., 2019b; Speidel et al., 2019; Zhang et al., 2023; Zhan et al., 2023). Even under strict parsimony conditions and for small sample sizes, the number of plausible ARGs compatible with a given dataset is vast. Thus, although it is clearly an oversimplification to arbitrarily choose *one* best guess, it is not clear that generating many guesses when sample sizes are large will achieve much. At the scale of millions of samples, we could only ever explore the tiniest corner of the incomprehensibly large space of plausible ARGs. Therefore, it is important to systematically describe and utilise uncertainty about ARG inference, and to incorporate uncertainty encountered during inference into the returned ARG. One approach, enabled by the gARG encoding described here, is to allow nodes to have more than two children (polytomies representing uncertainty over the ordering of coalescence events, Appendix D) or more than two parents (representing uncertainty over the ordering of multiple recombination events, Section 7). Development of other methods to capture, for example, uncertainty about node ages and recombination breakpoint positions, is an important aspect of future work. How this uncertainty can be utilised in downstream applications is an open question.

The timing, positions, and even the number of recombination events is generally not possible to infer precisely from genome sequencing data. Under coalescent-based models, the proportion of recombination events that change the ARG topology grows very slowly with sample size (Hein et al., 2004), and of those events only a small proportion are actually detectable from the data, assuming human-like mutation and recombination rates (Myers, 2002; Hayman et al., 2023). Even when a recombination event *is* detectable, its timing and breakpoint position can only be inferred approximately, depending on how much information can be elucidated from mutations in the surrounding genomic region. The fact that the eARG encoding *requires* precise information about recombination is therefore a fundamental limitation.

Besides the inherent limitations that exist on inferring fully precise ARGs from data, we should also consider the value that such exact estimates provide for downstream applications. Many applications work by examining local trees independently, making detailed information about recombination events superfluous. For example, the `Relate` selection test (Speidel et al., 2019) obtains p-values by computing clade size probabilities conditional on the timing of coalescence events in a given local tree. In their method for estimating dispersal rates and the locations of genetic ancestors, Osmond and Coop (2021) downsample trees along the genome so that they can be regarded as approximately independent. Similarly, Fan et al. (2023) compute the likelihood of an ARG under a particular demographic model as the product over a sample of widely-separated local trees, assumed to be independent. The SIA method for detecting selection (Hejase et al., 2022) encodes local trees as a set of lineage counts at discrete time intervals, and uses these as feature for a type of machine learning algorithm that takes "temporal" correlations into account. Thus, while SIA takes advantage of information about local tree correlation, it is in quite an indirect way, and clearly much of the detail about recombination events in an ARG is lost. The main application for fully precise ARGs thus far has been to compute a likelihood under the coalescent (e.g. Kuhner et al., 2000; Mahmoudi et al., 2022; Guo et al., 2022), which currently requires the details of all recombination events to be known.

The advantages of a model-agnostic representation that naturally incorporates uncertainty about the ordering of events in an ARG are well-illustrated by Zhan et al. (2023), who inferred ARGs using millions of SARS-CoV-2 sequences from the GISAID database (Shu and McCauley, 2017). In contrast to typical human sequencing datasets, the SARS-CoV-2 data is sampled continuously through time, sometimes with tens of thousands of sequences collected per day, with relatively little genetic diversity to distinguish them. The reconstructed ARGs thus contain polytomies and non-leaf sample nodes (sequences with descendants also present in the dataset) many of which only have a single child (i.e. are ARG-unary). Recombination is an important factor in the evolution of SARS-CoV-2 (VanInsberghe et al., 2021; Jackson et al., 2021; Ignatieva et al., 2022), and the inferred ARGs contain an unprecedented level of detail about the combined processes of viral mutation and recombination. Because parental sequences are generally never sampled themselves, and often a recombinant strain is the product of multiple recombination events, uncertainty around this is captured by recording the ancestry of each part of the recombinant sequence without arbitrarily assigning times or orderings for these events.

This view of ARGs, decoupled from generative models and without the hard requirement of complete precision on all historical events, may clarify inference goals and improve methods for evaluation.

In most cases, ARG inference is evaluated by simulating data from a known ground truth ARG, and comparing this to the inferred version via pairwise comparison of local trees along the genome using tree distance metrics (e.g. Robinson and Foulds, 1981; Kendall and Colijn, 2016), as described by Kuhner and Yamato (2015a). In comparing tree-by-tree along the genome, the effects of recombination are incorporated in a rather indirect manner through the correlations between the local trees, instead of directly taking into account the persistence of nodes and edges across multiple trees. The performance of tree distance metrics varies by application (Kuhner and Yamato, 2015b), and the correct approach to handling subtleties such as polytomies is an open question (Kelleher et al., 2019b; Zhang et al., 2023). Tree distance metrics often have $O(n^2)$ time complexity or worse and therefore cannot be applied to the very large sample sizes currently of interest. A recent trend has been to move away from such tree distance-based approaches and to examine more properties of the inferred ARGs, such as distributions of pairwise MRCA times (Brandt et al., 2022), waiting distances between local trees (Deng et al., 2021), and the genomic span of an edge or clade of samples (Ignatieva et al., 2023). In each case, simulation studies demonstrated substantial differences between these quantities in simulated and reconstructed ARGs that were not captured using tree-by-tree comparisons. Evaluations to-date have almost all been based on ground truth data from highly idealised simulations, with sample sizes limited to at most a few thousand (typically much fewer). Beyond the effects of very simplistic error models (e.g. Kelleher et al., 2019b), the effects of the richness of real data at biobank-scale on ARG inference are almost entirely unknown. The development of ARG evaluation metrics that take into account more of the global topology and can be applied to large ARGs would be a valuable and timely addition to the field. Using ARGs simulated from observed pedigree data (Anderson-Trocmé et al., 2023) as ground-truth would also add a valuable dimension to our understanding of how well methods perform when faced with realistic population and family structure.

Interest in ARG inference methods and downstream applications is burgeoning, with exciting developments arriving at ever-increasing pace. Without agreement on basic terminology and some standardisation on data formats, however, the ARG revolution may falter. For ARG-based methods to achieve mainstream status, we require a rich supporting software ecosystem. Ideally, this would comprise a wide range of inference methods specialised to different organisms, inference goals, and types and scales of data. If these diverse inference methods share a common, well-defined data format, their outputs could then be processed by many different downstream applications without the productivity-sapping problems of converting between partially incompatible formats (Excoffier and Heckel, 2006). Earlier efforts to standardise ARG interchange shared this vision, but did not succeed (Cardona et al., 2008; McGill et al., 2013). Current methods tend to tightly couple both ARG inference and downstream analysis within the same software package, which is ultimately not compatible with the widespread use of ARGs for routine data analysis, and a healthy and diverse software ecosystem. The gARG encoding described here is a significant generalisation of classical concepts, capable of describing even the bewildering complexity of contemporary datasets and encompassing a wide range of approximate ARG structures, and would be a reasonable basis for such a community interchange format.

Rigorously defining interchange formats (e.g. Kelleher et al., 2019a) is difficult and time-consuming, and no matter how precise the specification, in practise it is the *implementations* that determine how well methods interoperate. The BAM read alignment format (Li et al., 2009) is an instructive example. Originally developed as part of the 1000 Genomes project (1000 Genomes Project Consortium, 2015) to address the fragmented software ecosystem that existed at the time (Danecek et al., 2021), BAM has since become ubiquitous in bioinformatics pipelines. The excellent interoperability between methods exchanging alignment data is largely attributable to the success of `htslib` (Bonfield et al., 2021), the software library that *implements* BAM and several other foundational bioinformatics file formats. Today, there are thousands of software projects using `htslib` (Bonfield et al., 2021), and it this shared use of community software infrastructure that guarantees the smooth flow of data between applications. The emerging ARG software ecosystem could similarly benefit from the adoption of such shared community infrastructure to handle the mundane and time-consuming details of data interchange. The `tskit` library (Section 10) is a high-quality open-source gARG implementation, with proven efficiency and scalability (e.g. Anderson-Trocmé et al., 2023; Zhan et al., 2023), that is already in widespread use. Adopting it as a community standard may ease software implementation burden on researchers, freeing their time to address the many fascinating open questions and challenges that exist.

Acknowledgements

We are grateful to Nick Barton, Gideon Bradburd, Alex Lewanski, Peter Ralph and Andrew Vaughn for helpful discussions and comments on the manuscript, and to Iain Mathieson and Aylwyn Scally for interesting discussions that originally motivated this work. Gregor Gorjanc acknowledges support from the BBSRC ISP grant to The Roslin Institute (BBS/E/D/30002275, BBS/E/RL/230001A, and BBS/E/RL/230001C). Jerome Kelleher acknowledges support from the Robertson Foundation, NIH (research grants HG011395 and HG012473) and EPSRC (research grant EP/X024881/1). Jere Koskela acknowledges support from EPSRC research grant EP/V049208/1.

Data Availability

The public GitHub repository at `https://github.com/tskit-dev/what-is-an-arg-paper` can be used to reproduce all figures and tables in this paper. In particular this includes the ARG used in Fig. 3; the simulation code and functions used to generate Fig. 4; and for Fig. 5, the software versions, parameter settings, and (where necessary) functions to convert software outputs to the `tskit` gARG format.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, **526**(7571): 68, 2015.

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., et al. A community-maintained standard library of population genetic models. *eLife*, **9**: e54967, 2020.

Allen, B. L. and Steel, M. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**(1): 1–15, 2001.

Anderson-Trocmé, L., Nelson, D., Zabad, S., Diaz-Papkovich, A., Kryukov, I., Baya, N., Touvier, M., Jeffery, B., Dina, C., Vézina, H., et al. On the genes, genealogies, and geographies of Quebec. *Science*, **380**(6647): 849–855, 2023.

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, **599**(7886): 628–634, 2021.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, **220**(3), 2022. iyab229.

Bhaskar, A., Clark, A. G., and Song, Y. S. Distortion of genealogical properties when the sample is very large. *Proceedings of the National Academy of Sciences*, **111**(6): 2385–2390, 2014.

Birkner, M., Blath, J., and Eldon, B. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, **193**(1): 255–290, 2013.

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R. M. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, **10**(2): giab007, 2021.

Bordewich, M. and Semple, C. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, **8**(4): 409–423, 2005.

Brandt, D. Y., Wei, X., Deng, Y., Vaughn, A. H., and Nielsen, R. Evaluation of methods for the inference of ancestral recombination graphs. *Genetics*, **221**(1), 2022.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**: 203–209, 2018.

Cámara, P. G., Levine, A. J., and Rabadan, R. Inference of ancestral recombination graphs through topological data analysis. *PLOS Computational Biology*, **12**(8): e1005071, 2016.

Cardona, G., Rosselló, F., and Valiente, G. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, **9**: 532, 2008.

Comeron, J. M., Ratnappan, R., and Bailin, S. The many landscapes of recombination in *Drosophila melanogaster*. *PLOS Genetics*, **8**: 1–21, 2012.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., et al. Twelve years of SAMtools and BCFtools. *Gigascience*, **10**(2): giab008, 2021.

Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, **119**(4): e2115642118, 2022.

Deng, Y., Song, Y. S., and Nielsen, R. The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, **141**: 34–43, 2021.

Didelot, X., Lawson, D., Darling, A., and Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, **186**: 1435–1449, 2010.

Donnelly, P. and Kurtz, T. G. Genealogical processes for Fleming–Viot models with selection and recombination. *Annals of Applied Probability*, **9**(4): 1091–1148, 1999.

Etheridge, A. M. and Griffiths, R. C. A coalescent dual process in a Moran model with genic selection. *Theoretical Population Biology*, **75**(4): 320–330, 2009.

Ethier, S. N. and Griffiths, R. C. On the two-locus sampling distribution. *Journal of Mathematical Biology*, **29**(2): 131–159, 1990.

Excoffier, L. and Heckel, G. Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, **7**(10): 745–758, 2006.

Fan, C., Cahoon, J. L., Dinh, B. L., Vecchyo, D. O.-D., Huber, C. D., Edge, M. D., Mancuso, N., and Chiang, C. W. A likelihood-based framework for demographic inference from genealogical trees. *bioRxiv*, 2023.

Fan, C., Mancuso, N., and Chiang, C. W. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, **109**(5): 812–824, 2022.

Fearnhead, P. Perfect simulation from population genetic models with selection. *Theoretical Population Biology*, **59**(4): 263–279, 2001.

Fearnhead, P. Ancestral processes for non-neutral models of complex diseases. *Theoretical Population Biology*, **63**(2): 115–130, 2003.

Fearnhead, P. and Donnelly, P. Estimating recombination rates from population genetic data. *Genetics*, **159**(3): 1299–1318, 2001.

González Casanova, A. and Spanò, D. Duality and fixation in Ξ-Wright–Fisher processes with frequency-dependent selection. *Annals of Applied Probability*, **28**(1): 250–284, 2018.

Gower, G., Ragsdale, A. P., Bisschop, G., Gutenkunst, R. N., Hartfield, M., Noskova, E., Schiffels, S., Struck, T. J., Kelleher, J., and Thornton, K. R. Demes: a standard format for demographic models. *Genetics*, **222**(3): iyac131, 2022.

Griffiths, R. C. The two-locus ancestral graph. *Lecture Notes-Monograph Series*, **18**: 100–117, 1991.

Griffiths, R. C., Jenkins, P. A., and Lessard, S. A coalescent dual process for a Wright–Fisher diffusion with recombination and its application to haplotype partitioning. *Theoretical Population Biology*, **112**: 126–138, 2016.

Griffiths, R. C. and Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**(4): 479–502, 1996.

Griffiths, R. C. and Marjoram, P. An ancestral recombination graph. In P. Donnelly and S. Tavaré, eds., *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, vol. 87, 257–270. Springer-Verlag, Berlin, 1997.

Guo, F., Carbone, I., and Rasmussen, D. A. Recombination-aware phylogeographic inference using the structured coalescent with ancestral recombination. *PLOS Computational Biology*, **18**(8): e1010422, 2022.

Gusfield, D. *ReCombinatorics: the Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT press, 2014.

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

Hall, S. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal*, **10**(11): 1778–1785, 2016.

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*, **607**(7920): 732–740, 2022.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 2018.

Haller, B. C. and Messer, P. W. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, **36**(3): 632–637, 2019.

Harris, K. From a database of genomes to a forest of evolutionary trees. *Nature Genetics*, **51**(9): 1306–1307, 2019.

Harris, K. Using enormous genealogies to map causal variants in space and time. *Nature Genetics*, 1–2, 2023.

Hayes, B. J. and Daetwyler, H. D. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual Review of Animal Biosciences*, **7**: 89–102, 2019.

Hayman, E., Ignatieva, A., and Hein, J. Recoverability of ancestral recombination graph topologies. *Theoretical Population Biology*, **154**: 27–39, 2023.

Hein, J. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, **98**(2): 185–200, 1990.

Hein, J. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, **36**: 396–405, 1993.

Hein, J., Jiang, T., Wang, L., and Zhang, K. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, **71**(1-3): 153–169, 1996.

Hein, J., Schierup, M., and Wiuf, C. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*. Oxford University Press, USA, 2004.

Heine, K., Beskos, A., Jasra, A., Balding, D., and De Iorio, M. Bridging trees for posterior inference on ancestral recombination graphs. *Proc R Soc A*, **474**: 20180568, 2018.

Hejase, H. A., Dukler, N., and Siepel, A. From summary statistics to gene trees: methods for inferring positive selection. *Trends in Genetics*, **36**(4): 243–258, 2020.

Hejase, H. A., Mo, Z., Campagna, L., and Siepel, A. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular Biology and Evolution*, **39**(1): msab332, 2022.

Hubisz, M. and Siepel, A. Inference of ancestral recombination graphs using ARGweaver. In *Statistical Population Genomics*, 231–266. Humana, New York, NY, 2020.

Hubisz, M. J., Williams, A. L., and Siepel, A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLOS Genetics*, **16**(8): e1008895, 2020.

Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**: 183–201, 1983a.

Hudson, R. R. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, **37**(1): 203–217, 1983b.

Hudson, R. R. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**: 1–44, 1990.

Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2): 337–338, 2002.

Huson, D. H., Rupp, R., and Scornavacca, C. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.

Ignatieva, A., Favero, M., Koskela, J., Sant, J., and Myers, S. R. The distribution of branch duration and detection of inversions in ancestral recombination graphs. *bioRxiv*, 2023–07, 2023.

Ignatieva, A., Hein, J., and Jenkins, P. A. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. *Molecular Biology and Evolution*, **39**(2): msac028, 2022.

Ignatieva, A., Lyngsø, R. B., Jenkins, P. A., and Hein, J. KwARG: parsimonious reconstruction of ancestral recombination graphs with recurrent mutation. *Bioinformatics*, **37**: 3277–3284, 2021.

Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, **184**(20): 5179–5188, 2021.

Jenkins, P. A. and Griffiths, R. C. Inference from samples of DNA sequences using a two-locus model. *Journal of Computational Biology*, **18**: 109–127, 2011.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**(7809): 434–443, 2020.

Kelleher, J., Barton, N. H., and Etheridge, A. M. Coalescent simulation in continuous space. *Bioinformatics*, **29**(7): 955–956, 2013.

Kelleher, J., Etheridge, A. M., and Barton, N. H. Coalescent simulation in continuous space: Algorithms for large neighbourhood size. *Theoretical Population Biology*, **95**: 13–23, 2014.

Kelleher, J., Etheridge, A. M., and McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, **12**(5): e1004842, 2016.

Kelleher, J., Lin, M., Albach, C. H., Birney, E., Davies, R., Gourtovaia, M., Glazer, D., Gonzalez, C. Y., Jackson, D. K., Kemp, A., et al. htsget: a protocol for securely streaming genomic data. *Bioinformatics*, **35**(1): 119–121, 2019a.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, **14**(11): 1–21, 2018.

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

Kendall, M. and Colijn, C. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*, **33**(10): 2735–2743, 2016.

Kingman, J. F. C. The coalescent. *Stochastic processes and their applications*, **13**(3): 235–248, 1982a.

Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability*, **19**(A): 27–43, 1982b.

Knuth, D. E. *Combinatorial Algorithms, Part 1*, vol. 4A of *The Art of Computer Programming*. Addison-Wesley, Upper Saddle River, New Jersey, 2011.

Korfmann, K., Abu Awad, D., and Tellier, A. Weak seed banks influence the signature and detectability of selective sweeps. *Journal of Evolutionary Biology*, **36**(9): 1282–1294, 2023.

Koskela, J. and Wilke Berenguer, M. Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences*, **311**: 1–12, 2019.

Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, **304**(5925): 412–417, 1983.

Krone, S. M. and Neuhauser, C. Ancestral processes with selection. *Theoretical Population Biology*, **51**(3): 210–237, 1997.

Kuhner, M. K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**(6): 768–770, 2006.

Kuhner, M. K. and Yamato, J. Assessing differences between ancestral recombination graphs. *Journal of Molecular Evolution*, **80**(5): 258–264, 2015a.

Kuhner, M. K. and Yamato, J. Practical performance of tree comparison metrics. *Systematic Biology*, **64**(2): 205–214, 2015b.

Kuhner, M. K. and Yamato, J. A consensus method for ancestral recombination graphs. *Journal of Molecular Evolution*, **84**(2): 129–138, 2017.

Kuhner, M. K., Yamato, J., and Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**(3): 1393–1401, 2000.

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., et al. Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife*, **12**: RP84874, 2023.

Lewanski, A. L., Grundler, M. C., and Bradburd, G. S. The era of the ARG: an empiricist's guide to ancestral recombination graphs. *arXiv preprint arXiv:2310.12070*, 2023.

Lewin, H. A., Richards, S., Aiden, E. L., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., et al. The Earth BioGenome Project 2020: Starting the clock. 2022.

Li, H. and Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357): 493–496, 2011.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16): 2078–2079, 2009.

Li, H. and Stephan, W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLOS Genetics*, **2**(10): e166, 2006.

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

Lyngsø, R. B., Song, Y. S., and Hein, J. Minimum recombination histories by branch and bound. In *International Workshop on Algorithms in Bioinformatics*, 239–250. Springer, 2005.

MacLeod, I. M., Larkin, D. M., Lewin, H. A., Hayes, B. J., and Goddard, M. E. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Molecular Biology and Evolution*, **30**(9): 2209–2223, 2013.

Mahmoudi, A., Koskela, J., Kelleher, J., Chan, Y.-b., and Balding, D. Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, **18**(3): 1–15, 2022.

Makanjuola, B. O., Miglior, F., Abdalla, E. A., Maltecca, C., Schenkel, F. S., and Baes, C. F. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. *Journal of Dairy Science*, **103**(6): 5183–5199, 2020.

Marjoram, P. and Wall, J. D. Fast "coalescent" simulation. *BMC Genetics*, **7**: 16, 2006.

Mathieson, I. and Scally, A. What is ancestry? *PLOS Genetics*, **16**(3): e1008624, 2020.

McGill, J. R., Walkup, E. A., and Kuhner, M. K. GraphML specializations to codify ancestral recombinant graphs. *Frontiers in Genetics*, **4**: 146, 2013.

McVean, G. A. T. and Cardin, N. J. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, **360**: 1387–1393, 2005.

Medina-Aguayo, F. J., Didelot, X., and Everitt, R. G. Speeding up inference of homologous recombination in bacteria. *bioRxiv*, 2020.

Minichiello, M. J. and Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, **79**(5): 910–922, 2006.

Mirzaei, S. and Wu, Y. RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, **33**(7): 1021–1030, 2017.

Myers, S. R. *The detection of recombination events using DNA sequence data*. Ph.D. thesis, University of Oxford, 2002.

Myers, S. R. and Griffiths, R. C. Bounds on the minimum number of recombination events in a sample history. *Genetics*, **163**(1): 375–394, 2003.

Nelson, D., Kelleher, J., Ragsdale, A. P., Moreau, C., McVean, G., and Gravel, S. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*, **16**(5): e1008619, 2020.

Neuhauser, C. The ancestral graph and gene genealogy under frequency-dependent selection. *Theoretical Population Biology*, **56**(2): 203–214, 1999.

Neuhauser, C. and Krone, S. M. The genealogy of samples in models with selection. *Genetics*, **145**(2): 519–534, 1997.

Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**(2): 931–942, 2000.

Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*, **154**(2): 923–929, 2000.

Nowbandegani, P. S., Wohns, A. W., Ballard, J. L., Lander, E. S., Bloemendal, A., Neale, B. M., and O'Connor, L. J. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. *Nature Genetics*, 2023.

O'Fallon, B. D. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics*, **14**(1): 40, 2013.

Osmond, M. and Coop, G. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv*, 2021.

Palamara, P. F. ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, **32**(19): 3032–3034, 2016.

Parida, L., Melé, M., Calafell, F., Bertranpetit, J., and Consortium, G. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *Journal of Computational Biology*, **15**(9): 1133–1153, 2008.

Parida, L., Palamara, P. F., and Javed, A. A minimal descriptor of an ancestral recombinations graph. *BMC Bioinformatics*, **12**(1): 1–17, 2011.

Paul, J. S., Steinrucken, M., and Song, Y. S. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, **187**(4): 1115–1128, 2011.

Petr, M., Haller, B. C., Ralph, P. L., and Racimo, F. slendr: A framework for spatio-temporal population genomic simulations on geographic landscapes. *bioRxiv*, 2022.

Pocrnic, I., Lourenco, D. A., Masuda, Y., and Misztal, I. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. *Genetics Selection Evolution*, **48**(82), 2016.

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

Rasmussen, D. A. and Guo, F. Espalier: Efficient tree reconciliation and ARG reconstruction using maximum agreement forests. *bioRxiv*, 2022.

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

Robinson, D. F. and Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**(1-2): 131–147, 1981.

Ros-Freixedes, R., Whalen, A., Chen, C.-Y., Gorjanc, G., Herring, W. O., Mileham, A. J., and Hickey, J. M. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genetics Selection Evolution*, **52**(17), 2020.

Schaefer, N. K., Shapiro, B., and Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*, **7**(29): eabc0776, 2021.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**(8): 919–925, 2014.

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, **194**: 937–954, 2013.

Shipilina, D., Pal, A., Stankowski, S., Chan, Y. F., and Barton, N. H. On the origin and structure of haplotype blocks. *Molecular Ecology*, **32**(6): 1441–1457, 2023.

Shu, Y. and McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**(13): 30494, 2017.

Song, Y. S. On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, **7**(3): 365–379, 2003.

Song, Y. S. Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Annals of Combinatorics*, **10**(1): 147–163, 2006.

Song, Y. S. and Hein, J. Parsimonious reconstruction of sequence evolution and haplotype blocks. In *International Workshop on Algorithms in Bioinformatics*, 287–302. Springer, 2003.

Song, Y. S. and Hein, J. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology*, **48**(2): 160–186, 2004.

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

Song, Y. S., Wu, Y., and Gusfield, D. Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. *Bioinformatics*, **21**(suppl_1): i413–i422, 2005.

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

Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2): 437–460, 1983.

Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O., and Nagasaki, M. Practical guide for managing large-scale human genome data in research. *Journal of Human Genetics*, **66**(1): 39–52, 2021.

Terasaki Hart, D. E., Bishop, A. P., and Wang, I. J. Geonomics: Forward-time, spatially explicit, and arbitrarily complex landscape genomic simulations. *Molecular Biology and Evolution*, **38**(10): 4634–4646, 2021.

Thao, N. T. P. and Vinh, L. S. A hybrid approach to optimize the number of recombinations in ancestral recombination graphs. In *Proceedings of the 2019 9th International Conference on Bioscience, Biochemistry and Bioinformatics*, 36–42. Association for Computing Machinery, 2019.

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, **361**: k1687, 2018.

VanInsberghe, D., Neish, A. S., Lowen, A. C., and Koelle, K. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evolution*, **7**(2): veab059, 2021.

Vaughan, T. G., Welch, D., Drummond, A. J., Biggs, P. J., George, T., and French, N. P. Inferring ancestral recombination graphs from bacterial genomic data. *Genetics*, **205**(2): 857–870, 2017.

Wakeley, J. *Coalescent Theory: an Introduction*. Roberts and Company, Englewood, Colorado, 2008.

Wakeley, J., King, L., Low, B. S., and Ramachandran, S. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, **190**(4): 1433–1445, 2012.

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

Wang, Y. and Rannala, B. Bayesian inference of fine-scale recombination rates using population genomic data. *Philosophical Transactions of the Royal Society B*, **363**: 3921–3930, 2008.

Wang, Y. and Rannala, B. Population genomic inference of recombination rates and hotspots. *Proceedings of the National Academy of Sciences*, **106**(15): 6215–6219, 2009.

Wilton, P. R., Carmi, S., and Hobolth, A. The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, **200**(1): 343–355, 2015.

Wiuf, C. and Hein, J. The ancestry of a sample of sequences subject to recombination. *Genetics*, **151**(3): 1217–1228, 1999a.

Wiuf, C. and Hein, J. Recombination as a point process along sequences. *Theoretical Population Biology*, **55**(3): 248–259, 1999b.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. A unified genealogy of modern and ancient genomes. *Science*, **375**(6583): eabi8264, 2022.

Wu, Y. Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. *Journal of Computational Biology*, **15**(7): 667–684, 2008.

Wu, Y. New methods for inference of local tree topologies with recombinant SNP sequences in populations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**(1): 182–193, 2011.

Zhan, S. H., Ignatieva, A., Wong, Y., Eaton, K., Jeffery, B., Palmer, D. S., Murall, C. L., Otto, S., and Kelleher, J. Towards pandemic-scale ancestral recombination graphs of SARS-CoV-2. *bioRxiv*, 2023–06, 2023.

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., and Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, **55**: 768–776, 2023.

Appendix

A Ancestral graphs: a brief history

The coalescent (Kingman, 1982a,b; Hudson, 1983b; Tajima, 1983) models the ancestry of a sample of genomes under an idealised population model, and provides the theoretical underpinning for much of contemporary population genetics. It is a stochastic process, where each random realisation is a genealogical tree describing the genetic ancestry of the sample. Numerous extensions to the model have been proposed (Hudson, 1990; Hein et al., 2004; Wakeley, 2008), incorporating many evolutionary processes. Hudson (1983a) first incorporated recombination into the coalescent process, providing several fundamental analytical results and describing the basic simulation algorithm, still in widespread use (Hudson, 2002; Kelleher et al., 2016; Baumdicker et al., 2022). In the 1990s, Griffiths and colleagues revisited the coalescent with recombination from a different perspective, formulating it as a stochastic process where each realisation is encoded as a graph (Griffiths, 1991; Ethier and Griffiths, 1990; Griffiths and Marjoram, 1996, 1997). They referred to both the stochastic process and its random realisations as the Ancestral Recombination Graph (ARG). Although mathematically equivalent, it is important to note that the Griffiths and Hudson formulations of the coalescent with recombination are not identical; in particular, a direct implementation of the ARG process as originally described requires exponential time to simulate (see Appendix B for details). However, ARGs provided a way to reason about and infer recombinant ancestry as a single object, in a way that is not possible within Hudson's framework, which emphasised instead the collection of local trees along the genome resulting from recombination.

Subsequent work on ARGs proceeded in broadly three main directions: (1) exploring the mathematical properties of the coalescent with recombination and related stochastic processes; (2) inferring evolutionary parameters under (approximations to) this model, either with or without explicitly reconstructing the genealogy of the sample; and (3) treating the ARG as a discrete graph, ignoring the generating stochastic process, and studying its properties from a computational and algorithmic perspective.

An extensive body of work has been developed from studying the coalescent with recombination and other related graph-valued stochastic processes from a mathematical perspective. In particular, the Ancestral Selection Graph (ASG) (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997) uses a similar approach to model natural selection instead of recombination. Unlike the ARG process, the ASG imposes a hard distinction between the stochastic process, which constructs a random ARG-like graph, and an observable realisation, which is a single tree sampled from the graph in a non-uniform way to encode desired patterns of natural selection. Constructions of ASG-like stochastic processes encoding various forms of selection, often in parallel with recombination or other genetic forces, are an area of considerable and ongoing theoretical interest (e.g. Neuhauser, 1999; Donnelly and Kurtz, 1999; Fearnhead, 2001, 2003; Etheridge and Griffiths, 2009; González Casanova and Spanò, 2018; Koskela and Wilke Berenguer, 2019).

Early work on inference under the coalescent with recombination focused on the problem of inferring the parameters of the stochastic process, where the ancestry was regarded as a latent parameter to be averaged out (e.g. Griffiths and Marjoram, 1996; Kuhner et al., 2000; Nielsen, 2000; Fearnhead and Donnelly, 2001). These methods met with limited success because the state space of ARGs is overwhelmingly large, and lacks a simple geometry or neighbourhood structure for inference or sampling methods to exploit. Several breakthroughs in this direction were achieved through formulating simplified but more tractable approximations to the full model (McVean and Cardin, 2005; Marjoram and Wall, 2006; Li and Durbin, 2011; Paul et al., 2011; Schiffels and Durbin, 2014). The related problem of *sampling* genealogies compatible with a given dataset under the coalescent with recombination also proved notoriously difficult computationally; progress in explicitly inferring genealogies at scale has similarly been achieved through resorting to principled approximations (Rasmussen et al., 2014; Mahmoudi et al., 2022), or moving away from the coalescent with recombination altogether and seeking to infer a single plausible ARG (e.g. Minichiello and Durbin, 2006; Kelleher et al., 2019b; Speidel et al., 2019).

There has also been substantial interest in formulating and answering fundamental questions about properties of the ARG as a discrete graph structure, focusing on the ARG topology without considering either branch lengths or indeed the generating process. The first prominent problem was calculating (lower bounds on) the minimum number of recombinations required to reconstruct a valid genealogy for a given sample (Myers and Griffiths, 2003), and constructing the corresponding minimal (parsimonious) ARGs (Song and Hein, 2003; Song et al., 2005; Lyngsø et al., 2005). These problems are NP-hard in general (Wang et al., 2001), and progress has been achieved through studying various constrained special cases of ARGs (e.g. Gusfield et al., 2004) and other more general types of phylogenetic networks (Huson et al., 2010). The focus has been on algorithmic and combinatorial results (Gusfield, 2014) that are often not of direct relevance to the inference problems described above.

The goal of this historical overview is to illustrate that the meaning of the term "ARG" now strongly depends on the context in which it is used, and can mean both the stochastic process that generates genealogies in the presence of recombination (e.g. Nordborg, 2000; Birkner et al., 2013; Wilton et al., 2015; Griffiths et al., 2016), as well as, more commonly, the concrete realisation of ancestry from a process (e.g. Gusfield, 2014; Mathieson and Scally, 2020; Brandt et al., 2022).

B The Big and Little ARG

Here we review two important stochastic processes that construct ARGs: the "Big" ARG process of Griffiths and Marjoram (1997), and the "Little" ARG process of Hudson (1983a). The Big ARG process is mathematically simpler but is computationally intractable due to generating a vast number of ancestors which contribute no genetic material to the initial sample. The Little ARG process avoids non-genetic ancestors at the cost of more complex dynamics and state space. We also demonstrate that applications relying on the grouping of inheritance pathways into ancestral lineages, such as likelihood-based inference under the coalescent, requires that the gARG (or eARG) data structure be interpreted in a model-specific way.

A generic state of the Little ARG process consists of a finite collection of lineages L, each of which is a list of disjoint ancestry segments (ℓ, r, a), where $[\ell, r)$ is a half-closed genomic interval and a is an integer tracking the number of samples to which the lineage is ancestral over that interval. We also usually track the node associated with each segment, but that is not important for our purposes here so we omit it to lighten notation. The initial condition for a sample of n genomes of length m consists of n lineages of the form $\{(0, m, 1)\}$. The process traverses a series of common ancestor and recombination events backwards in time. Recombination events happen at rate $\rho\nu/(m-1)$, where $\rho \geq 0$ is a per-genome recombination rate and

$$\nu = \sum_{x \in L} \left(\max_{(\ell,r,a)\in x} r - \min_{(\ell,r,a)\in x} \ell - 1 \right)$$

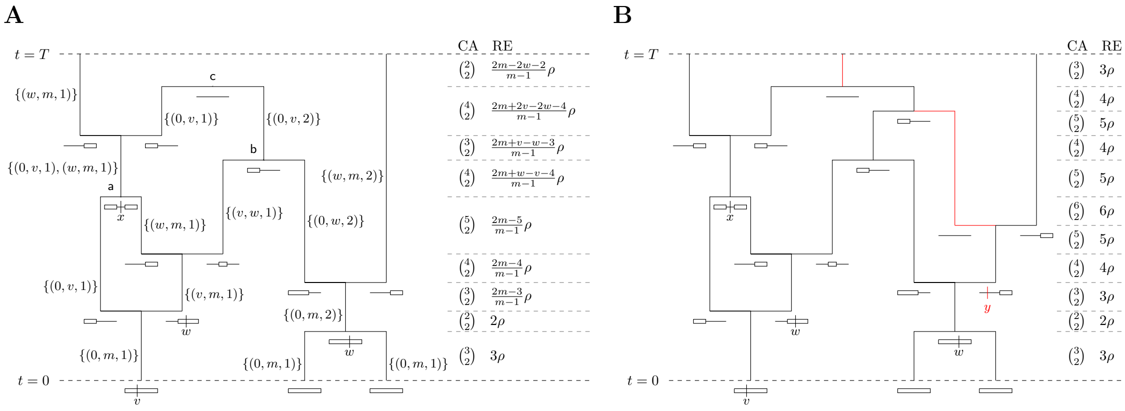

Figure A1: (A) A realisation of the graph traversed by Hudson's algorithm started from a sample of three chromosomes with m discrete sites each at time $t = 0$, and propagated until time T. The MRCA on the genetic interval $[v, w)$ is reached at event b, while that on $[0, v)$ is reached at event c. The non-ancestral segment $[v, w)$ above a contributes to the rate of effective recombinations because it is trapped between ancestral segments. The two columns titled CA and RE are the respective rates of mergers and recombinations when the recombination rate is ρ. (B) A corresponding realisation of a Big ARG, which augments Hudson's algorithm by tracking non-ancestral lineages. The result is a simpler state space and dynamics, at the cost of extra nodes and edges, highlighted in red, which do not affect the local tree at any site. Recombination positions are labelled alphabetically in time, and their ordering along the genome is $y < v < x < w$, of which the first only appears in panel B. There are two separate recombination events at link w.

is the number of available "links" surrounded by ancestral material. At a recombination event we choose one of these links uniformly and break it, replacing the original lineage in L with two new lineages containing the ancestral material to the left and right of the break point, respectively.

Common ancestor events occur at rate $\binom{|L|}{2}$. In a common ancestor event, two uniformly sampled lineages have their segments merged into a single ancestor lineage, which is added to L. If the lineages have overlapping intervals of ancestry, say, (ℓ, r, a_1) and (ℓ, r, a_2), a *coalescence* occurs. The result is a segment $(\ell, r, a_1 + a_2)$, and if $a_1 + a_2 < n$ it is included in the ancestor lineage. Otherwise, if $a_1 + a_2 = n$, we have found the most recent common ancestor of all samples in the interval $[\ell, r)$ and do not need to simulate its history any further. Non-overlapping intervals from the two lineages are included in the ancestor lineage without changes. Eventually, we find resultant lineages in which all segments have fully coalesced, and so the number of extant lineages gradually falls to zero.

In the Griffiths formulation (the Big ARG process), each edge in the graph corresponds to an extant lineage and nodes are events in the process. The n initial leaf nodes are sampling events. Common ancestor events occur at rate $\binom{|L|}{2}$. When a common ancestor event happens, two uniformly chosen lineages merge into a common ancestor lineage. Recombination events happen at rate $|L|\rho$. Here, we choose a lineage (i.e. edge) uniformly, and a breakpoint $0 < x < m$ uniformly on its genome. We terminate the edge at a node, record the breakpoint, and start two new edges from this node. The process then continues until there is only one lineage left (the Grand Most Recent Common Ancestor, GMRCA), which is guaranteed to happen in finite time because of the quadratic rate of coalescing vs. linear rate of branching.

The state-space of the Big ARG process is much simpler than that of the Little ARG process, which greatly facilitates mathematical reasoning. This simplicity comes at a substantial cost, however, if we wish to use it as a practical means of simulating recombinant ancestries. The number of events in the Big ARG all the way back to the GMRCA is $O(e^\rho)$ (Griffiths and Marjoram, 1997), whereas the number of events required to simulate the Little ARG is $O(\rho^2)$ (Hein et al., 2004; Baumdicker et al., 2022). This disparity arises because the majority of the events in the Big ARG are recombination events which occur outside of ancestral material, and these do not have any bearing on the ancestry of the initial sample. Because we don't keep track of the distribution of ancestral material during the process, we generate a vastly larger graph.

Figure A1 illustrates the more complex state space of the Little ARG process, as well as the extra events which occur in the Big ARG process. Moreover, it depicts the rates of common ancestors and recombination events in each interval of time of the realisations. In order to evaluate these rates, e.g. for likelihood-based inference (Baumdicker et al., 2022; Mahmoudi et al., 2022), it is necessary to know the number of lineages and number of extant links available for recombination in each time interval. Some representations may not provide this information. For example, in the gARG encoding depicted in Figure 3C, it is clear that a recombination takes place between nodes `i`, `k` and `j`. But the exact time of the recombination event is ambiguous: it could take place at any time between node `i` and its parents and produce the same gARG. Because a recombination increases the number of extant lineages by one (in the rootward direction of time), the number of lineages during the same time interval is ambiguous as well. In fact, this information cannot be recovered from the gARG encoding used in Figure 3C without an extrinsic convention. For the basic coalescent with recombination, it is sufficient to create *two* gARG nodes at the time of the recombination event, with the interpretation that the two rootward edges from node `i` in Figure 3C belong to the same lineage until the time of nodes `k` and `j`, and split into two separate lineages at that time point. Similarly, the trapped, non-ancestral links along that lineage remain available for effective recombination (i.e. one which splits up ancestral material) for the same time interval. This interpretation is highlighted in Figure A1 by drawing only one vertical edge between a recombinant child and its two parents.

C Survey of ARG inference methods

The problem of reconstructing ARGs for samples of recombining sequences has been of interest since the ARG was first defined. Early methods focused on finding *parsimonious* ARGs, i.e. those with a minimal number of recombination events (Hein, 1990). Two main approaches emerged: "backwards-in-time" (Lyngsø et al., 2005) and "along-the-genome" (Song and Hein, 2003, 2005). Backwards-in-time approaches start with a data matrix and reduce it to an empty matrix through row and column operations corresponding to coalescence, mutation, and recombination events, which construct an ARG from the bottom up (Song et al., 2005; Wu, 2008; Thao and Vinh, 2019; Ignatieva et al., 2021). Along-the-genome approaches begin from an initial local tree at a single focal site. Moving the focal site along the genome changes the local tree via a subtree prune and regraft operation whenever a recombination is encountered (Hein, 1993; Wu, 2011; Mirzaei and Wu, 2017). Rasmussen and Guo (2022) focus on parsimonious fusion of local trees into an ARG, while the method described by Cámara et al. (2016) is based on topological data analysis. Reconstructing a parsimonious ARG for a given data set is NP-hard (Wang et al., 2001), so parsimony-based methods resort to heuristics and are limited to analysing at most hundreds of sequences. Hence, a number of methods aim to balance computational efficiency with reconstruction of "reasonable", rather than parsimonious ARGs (Minichiello and Durbin, 2006; Parida et al., 2008; Kelleher et al., 2019b; Speidel et al., 2019; Schaefer et al., 2021; Zhang et al., 2023).

An alternative approach is to treat the ARG as a latent parameter to be averaged out by Monte Carlo methods, based either on importance sampling (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; Jenkins and Griffiths, 2011) or MCMC (Kuhner et al., 2000; Kuhner, 2006; Nielsen, 2000; Wang and Rannala, 2008, 2009; O'Fallon, 2013; Vaughan et al., 2017; Mahmoudi et al., 2022). These methods operate on representations of the "Little ARG" (see Appendix B), and are computationally expensive, being applicable to at most hundreds of samples consisting of tens or hundreds of kilobases with human-like parameters. State-of-the-art methods rely on cheaper, approximate models (Didelot et al., 2010; Heine et al., 2018; Hubisz et al., 2020; Hubisz and Siepel, 2020; Medina-Aguayo et al., 2020). The most scalable method, `ARGWeaver` (Rasmussen et al., 2014), can be applied to dozens of mammal-like genomes (Hubisz and Siepel, 2020).

Methods to sample ARGs generate a "cloud" of estimates, and Kuhner and Yamato (2017) provide an approach to generate a set of consensus breakpoints and local trees from such a cloud. The approach is based on examining the recombination breakpoints in all of the input ARGS, and including those that are in at least k of the input ARGs (with some additional filtering criteria) in the output. Within the resulting intervals, a consensus local tree is then generated using standard phylogenetic methods.

D Cell lineages and ARGs

In eukaryotes, ARGs are a result of the cellular processes of mitosis and meiosis. Mitosis leads to common ancestor events, and meiosis leads to recombination events (both crossover and gene conversion). Fig. 4 shows a schematic of the events and the genomes (chromosome icons) that occur in the cellular germline of a simplified, diploid multicellular hermaphrodite eukaryote with partially overlapping generations. Here, an event is not represented by a specific genome. Rather, genomes can be associated with, or "tag", events above (ancestral to) or below (descended from) them. For example, tagging the two genomes above a recombination event leads to the two node representation seen in Figs. 4 and A1, whereas tagging the genome below a recombination event leads to the more conventional graphs in Figs. 3 and 5A,B.

The schematic illustrates an important point about the biological reality of polytomies. Three lineages coalesce in the left-hand genome of individual D_{10}, but do so as the result of two successive bifurcations. This is *necessarily* so, because the only known method of reproducing DNA is by (semi-conservative) duplication. Whether this polytomy is resolvable depends on the available mutational data. Mutations can occur along any cell lineages. For example, a mutation in the first cell division of D_{10} could be shared between the two gametes produced by the cells in the left half of D_{10} but not shared by the right hand gamete. With enough mutations, each round of mitotic germline genome duplication within a single multicellular organism could in principle be distinguished.

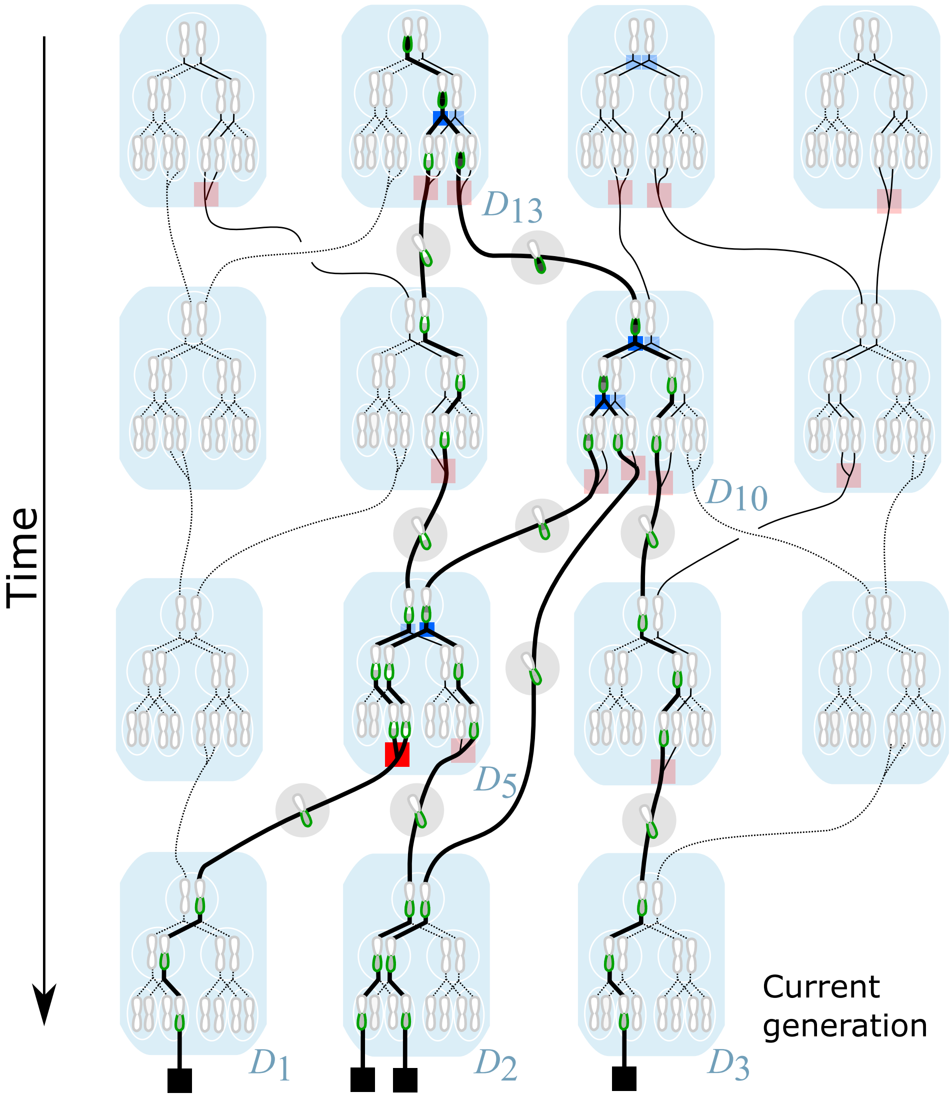

Figure A2: Cellular inheritance of a single chromosome in a diploid population. Individuals (blue) contain diploid cells (white circles enclosing a homologous pair of chromosomes). For clarity, only two rounds of mitotic germ-line cell division are shown per individual, and meiosis is not illustrated in detail. Lines show prospective inheritance paths for all chromosomes. Solid lines show all possible retrospective ancestry paths for four chosen chromosomes (indicated by square black "sampling events") sampled from 3 diploid individuals (D_1, D_2, D_3) in the current generation. Ancestral recombination events and coalescence events are shown as red and blue squares respectively. A realised ARG path for the lower arm of the sampled chromosomes is highlighted as a thick solid line, passing through a set of potential gARG nodes (green). This ARG involves a single recombination event and four coalescence events (highlighted as deep red and blue squares within individuals D_5, D_{10}, and D_{13}). ARG lineages also show gametic genomes, contained within shaded circles. As in Fig. 1A, inherited regions within the sampled chromosome arm are shaded by the number of descendant samples.

Supplementary Material

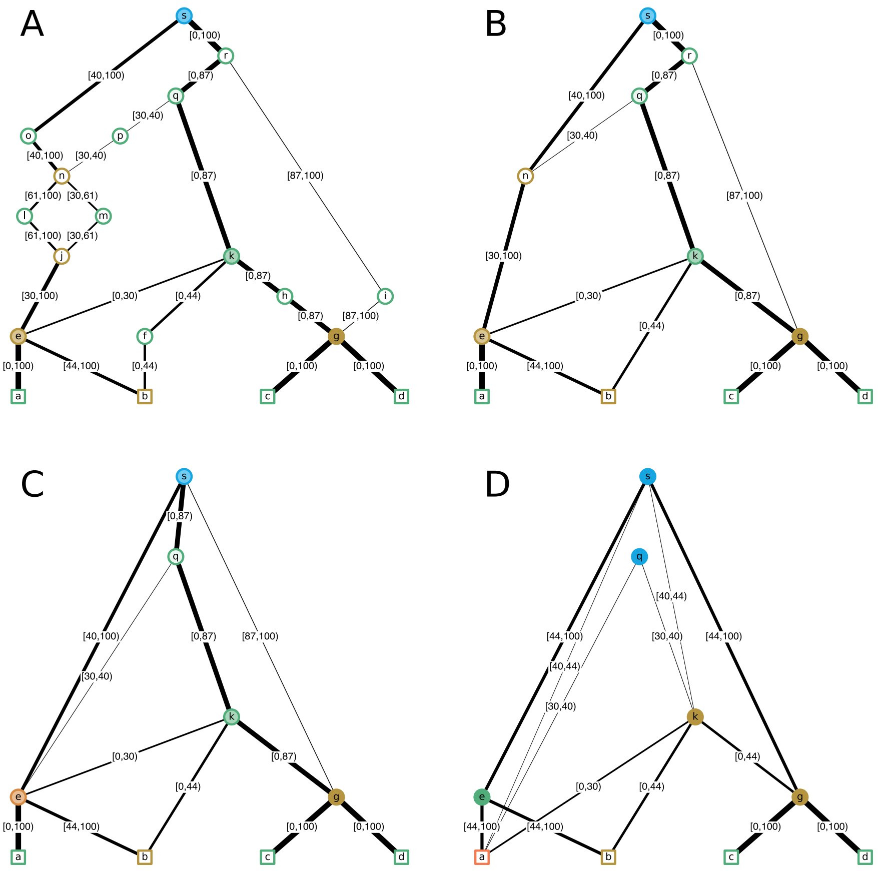

Figure S1: Example ARGs Fig. 4A–D, with edges annotated with inheritance intervals.