REVIEW

The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics

Alexander L. Lewanski[1,2,3,4]*, **Michael C. Grundler**[4], **Gideon S. Bradburd**[2,4]

1 Department of Integrative Biology, Michigan State University, East Lansing, Michigan, United States of America, **2** W.K. Kellogg Biological Station, Michigan State University, Hickory Corners, Michigan, United States of America, **3** Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, United States of America, **4** Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America

* allewanski@gmail.com

OPEN ACCESS

Citation: Lewanski AL, Grundler MC, Bradburd GS (2024) The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. PLoS Genet 20(1): e1011110. https://doi.org/10.1371/journal.pgen.1011110

Editor: Bret Payseur, University of Wisconsin–Madison, UNITED STATES

Published: January 18, 2024

Funding: This work was supported by a University Distinguished Fellowship from Michigan State University (awarded to ALL), the National Defense Science & Engineering Graduate (NDSEG) Fellowship from the Department of Defense (awarded to ALL), and by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM137919 (awarded to GSB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In the presence of recombination, the evolutionary relationships between a set of sampled genomes cannot be described by a single genealogical tree. Instead, the genomes are related by a complex, interwoven collection of genealogies formalized in a structure called an *ancestral recombination graph* (ARG). An ARG extensively encodes the ancestry of the genome(s) and thus is replete with valuable information for addressing diverse questions in evolutionary biology. Despite its potential utility, technological and methodological limitations, along with a lack of approachable literature, have severely restricted awareness and application of ARGs in evolution research. Excitingly, recent progress in ARG reconstruction and simulation have made ARG-based approaches feasible for many questions and systems. In this review, we provide an accessible introduction and exploration of ARGs, survey recent methodological breakthroughs, and describe the potential for ARGs to further existing goals and open avenues of inquiry that were previously inaccessible in evolutionary genomics. Through this discussion, we aim to more widely disseminate the promise of ARGs in evolutionary genomics and encourage the broader development and adoption of ARG-based inference.

Introduction

Many of the principal pursuits in evolutionary genomics can be recast as questions about the transmission of genetic material from ancestors to descendants. For example, in the study of speciation and hybridization, we may be interested in identifying which sections of a hybrid genome were derived from which parental species [1,2]. As another example, we often want to know about the nature of selection on a genetic variant (e.g., [3–6]), which is, in essence, asking whether the variant has displayed a particular pattern of transmission. For example, a positively selected variant confers a fitness advantage and thus would be preferentially transmitted between generations. In applied settings, we may want to understand whether a human-made structure such as a road or dam (e.g., [7,8]) reduces connectivity between populations, which is implicitly asking how often ancestor–descendant relationships span the potential barrier (e.g., [9]). So far, direct knowledge of how genetic material is transmitted from ancestors to descendants is extremely limited in nearly all systems, save those with extensive pedigree and genomic information (e.g., Florida scrub-jays [10–12], economically important livestock like dairy cattle [13,14]). However, access to this information could revolutionize the study of numerous topics across evolutionary genomics.

In population genetics, the fundamental structure that describes how genetic material is passed from ancestors to descendants is called an *ancestral recombination graph* (ARG). Building on earlier developments in coalescent theory [15–19], ARGs were conceptualized in the 1990s by R.C. Griffiths and P. Marjoram [20–22] to describe ancestry in the presence of coalescence and recombination. ARGs have subsequently featured prominently in the theoretical and statistical realms of population genetics where they have been extensively studied for their biological, mathematical, and computational properties and utility.

Nonetheless, ARGs remain much less known and appreciated in the broader field of evolutionary genomics. This inattention can at least partially be ascribed to pragmatism—until recently, ARGs have been purely theoretical constructs, impractical to reconstruct in empirical systems or even simulate at biologically realistic scales. Additionally, although an expansive literature already exists on ARGs, much of this content is targeted at an audience with extensive theoretical or statistical expertise in population genetics, and thus may be unapproachable for biologists lacking this background. Excitingly, recent methodological advances in reconstructing (Box 1) and simulating (Box 2) ARGs together with concurrent progress in genome sequencing and increasingly available high-performance computation means that obtaining ARGs is becoming conceivable in many situations.

Box 1: ARG reconstruction

A growing arsenal of methods is available to infer ARGs from genomic data. `ARGweaver`, which was introduced in 2014 by [44], represents a seminal achievement in ARG inference. `ARGweaver` and its extension (`ARGweaver-D`; [88]) leverage approximations of the coalescent (SMC or SMC' [43,89]) and time discretization to simplify the space from which to sample candidate ARGs using Markov Chain Monte Carlo. These methods, along with other recent Bayesian approaches like `Arbores` [90] and `ARGinfer` [91], enable the rigorous treatment of uncertainty via the incorporation of an ARG's posterior distribution into downstream analyses. One general limitation of these methods is that, due to computational requirements, they can only handle fairly modest sample sizes. For example, `ARGweaver` can consider between 2 to about 100 samples [45].

Motivated by the extensive sequencing efforts in human genomics, several methods have been devised to accommodate large and complicated genomic datasets. For example, `ARG-Needle` [92], `tsinfer+tsdate` [33,41], and `Relate` [86] can infer genomic genealogies for tens of thousands (`Relate`) to hundreds of thousands (`tsinfer+tsdate, ARG-Needle`) of human samples. `Relate` and `tsinfer+tsdate` can additionally incorporate samples from different time periods and have been used to reconstruct unified genomic genealogies for modern humans and ancient samples of humans, Neanderthals, and Denisovans [32,33]. This scalability is facilitated by various statistical simplifications, which result in several limitations in the inferences of these approaches. For example, `Relate` and `tsinfer+tsdate` infer less information about recombination than methods like `ARGweaver`, which attempts to identify the specific recombination events associated with every breakpoint [25]. Additionally, they only provide point estimates for the tree topologies, which precludes comprehensive assessments of uncertainty in ARG structure.

So far, most ARG inference development has focused on human and other eukaryotic genomes. However, there are also active efforts to create methods tailored to other types of genomes and systems. For example, [93] developed a Bayesian approach dubbed Bacter, which is designed to infer ARGs for bacteria based on the ClonalOrigin model [94]. Spurred by the COVID-19 pandemic, [95] recently introduced a method (`sc2ts`) for ARG reconstruction that can involve millions or more of SARS-CoV-2 genomes. `sc2ts` is designed to construct and repeatedly update an ARG through time with new samples, which is relevant to ongoing surveillance during pandemics wherein pathogen samples are collected and sequenced in real time.

In summary, there is a burgeoning assortment of methods that enable ARG reconstruction across a range of dataset and system characteristics including data types, sample sizes, and sampling regimes. ARG reconstruction remains a formidable statistical and computational challenge, and many improvements in the robustness and flexibility of ARG reconstruction are still needed [74,96,97]. However, ARG inference has emerged as a nexus of methodological development in statistical population genetics, and ongoing efforts exist to address the limitations and combine the strengths of existing methods (e.g., [98]). Readers should be prepared for continued innovation in this area.

Box 2: Simulation

Concurrent with improvements in ARG reconstruction, revolutionary progress in population genomic simulation has occurred over the past decade. One of the most significant developments was `msprime` [40,47], which can simulate genomes and ancestry backward in time under a variety of population genetic models including several models of the coalescent. With the coalescent, only the ancestors of the samples (and not entire populations) must be tracked. This approach is highly efficient but generally entails an assumption of neutral evolution [although it is possible for coalescent theory and simulation to incorporate selection (e.g., [47,99–103])]. The notable innovation of `msprime` relative to previous coalescent programs is the speed at which it can perform simulations at biologically realistic scales under a variety of models and with recombination. For example, `msprime` has been used to simulate realistic whole genome sequences based on genealogical information for approximately 1.4 million people inhabiting Quebec, Canada [104].

Another noteworthy development in population genomic simulation over the past decade is `SLiM` [105]. In contrast to coalescent simulators, `SLiM` simulates forward in time using either Wright–Fisher or non-Wright–Fisher models [106]. The forward-in-time nature of `SLiM` means that all individuals in each generation (including historical individuals that are not genetic ancestors to the contemporary population) must be tracked in the simulation. This elevates the computational burden compared to coalescent simulation. However, it enables substantially more flexibility in the scenarios that can be simulated including complex selection and ecological interactions across multiple species [107]. Relevant to this review, both `SLiM` and `msprime` can record ARGs during simulation [47,108]. These and other simulation programs (e.g., `discoal` [103]) can be used for a variety of purposes in ARG-based research including exploration of biological phenomena, statistical and machine learning inference (e.g., [37,109,110]), and methods evaluation [74].

The field of evolutionary genomics thus finds itself at an intriguing threshold. Methodological breakthroughs have led to substantial and growing prospects for obtaining ARGs in a variety of circumstances. However, several factors are combining to prevent the field from fully exploiting this newfound capacity to learn about evolutionary phenomena from ARGs. First, a restricted awareness of ARGs and their value means that there are currently few empirical researchers inferring ARGs in their systems. This paucity of empirical ARGs is limiting momentum to develop statistical methods that make inferences (e.g., of selection or demography) from ARGs (although see the "ARGs in evolutionary genomics" section for some recent examples). And, completing the feedback loop, the lack of methods that leverage ARGs to make inferences disincentivizes empirical researchers from generating ARGs in their systems. Fully capitalizing on ARGs in evolutionary genomics will thus require addressing these interrelated deficiencies—fostering an understanding and appreciation of ARGs and developing methods that enable researchers to make inferences from them. As a step towards achieving these goals and moving the field towards the "era of the ARG," we view now as an opportune moment to provide a widely accessible resource for comprehending ARGs and their potential in evolutionary genomics.

We have 2 primary objectives for this paper. First, we provide a concise and gentle primer on ARGs, including an introduction to what an ARG is, what information can be encoded within it, and an exploration of some of its basic properties. Second, we discuss the potential for ARGs to benefit evolutionary genomics research. Our aim for the second objective is not to exhaustively review existing ARG-based research, but rather to articulate the promise of ARGs to advance diverse topics across evolutionary genomics. We supplement these 2 main objectives with an overview of recent methodological developments in reconstructing, simulating, and analyzing ARGs. Through this discussion, we hope to demonstrate the potential of ARG-based approaches for providing insights into many evolutionary genomics questions and galvanize further development in ARG-based inference.

An ARG primer

In the following section, we will incrementally develop an intuition for what ARGs are by starting with the fundamentals of sexual reproduction and genealogical relatedness, which will help clarify how ARGs emerge from these first principles of biology. To simplify our discussion, we will focus on the nuclear genome of sexual, diploid organisms and meiotic recombination throughout the paper. However, the ideas covered here are relevant to any organism across the tree of life as well as viruses whose genomes undergo any type of recombination (e.g., gene conversion, bacterial conjugation). For more technical treatments of ARGs, we direct interested readers to [22–25].

Background

In sexual, diploid organisms, haploid gametes are generated by the sampling of a single DNA copy of every position in the genome during meiosis. During reproduction, the parents' gametes fuse, which leads to a diploid offspring. The relationships between a set of individuals can be represented by a genealogical pedigree (Fig 1A; light gray portions), in which each individual has 2 parents, from each of whom it has inherited exactly half of its genome. The pedigree consists of nodes, which represent individual organisms, and edges, which connect a subset of the nodes and signify parent–offspring relationships.

By itself, the pedigree can provide coarse estimates of genetic ancestry, such as the expected genetic relatedness between individuals (e.g., 0.50 between full siblings; 0.125 between first cousins), or the expected proportion of the genome inherited from a particular genealogical ancestor. However, for any region of the genome, we are unable to ascertain from the pedigree alone whether it is the parent's maternal or paternal copy that has been transmitted. Thus, we are restricted to calculating expected quantities. We could therefore gain more in-depth knowledge of ancestry in the genome by explicitly tracking the transmission of DNA sequences down the pedigree from specific parental to offspring chromosomes.

This discussion of the pedigree highlights multiple key ideas in our build-up to ARGs. First, because each parent contributes only 1 DNA copy at a particular genomic position to its offspring, each copy experiences its own unique history of inheritance through the pedigree. Second, because a parent only contributes half of its genome to each offspring and not all individuals reproduce, only a subset of the genetic material possessed by historical individuals in the pedigree end up in contemporary individuals. As you travel further back in the pedigree, despite the geometric increase in the number of expected genealogical ancestors (a maximum of 2^n ancestors where n equals the number of generations back in time), an increasing proportion of these ancestors contributes no genetic material to their contemporary descendants [26,27].

If we concentrate on a particular position in an individual's genome, we see that each DNA copy traverses just one of the manifold possible paths (i.e., series of connected nodes and edges) in the pedigree. The specific pedigree paths through which copies at a particular position in contemporary individuals were transmitted from their ancestors represent the genetic genealogy at that position [28,29]. Similar to a pedigree, each edge in the genealogy represents a transmission event of genetic material from parent to offspring. However, in a pedigree, each node is a diploid individual, while in a genetic genealogy, each node represents 1 of 2 haploid sequences *within* a diploid individual—the specific genomic copy sampled to create a gamete that passes genetic material from a parent to the current individual. This genetic genealogy is embedded in the pedigree (Fig 1A; portions in dark gray and color). The sequence of relationships defined by the pedigree constrains the possible nodes and edges that can exist in the genealogy, but does not fully dictate the identities of these nodes and edges. The structure of a genetic genealogy is determined by both the pedigree structure and the outcome of the gametogenic genome sampling at each reproduction event in the pedigree.

The genetic perspective of relatedness is further complicated by another feature of meiosis: recombination. Meiotic recombination, the shuffling of genetic material in the genome during meiosis, occurs via 2 processes: (1) exchange of genetic material between homologous chromosomes via crossing over during prophase I; and (2) random assortment of homologous chromosomes during anaphase I. These recombinational processes can produce a mosaic of genetic ancestry across the haploid genome of the gamete so that a particular gametic genome potentially contains genetic material inherited from different parents both between non-homologous chromosomes and within chromosomes. Recombination therefore results in

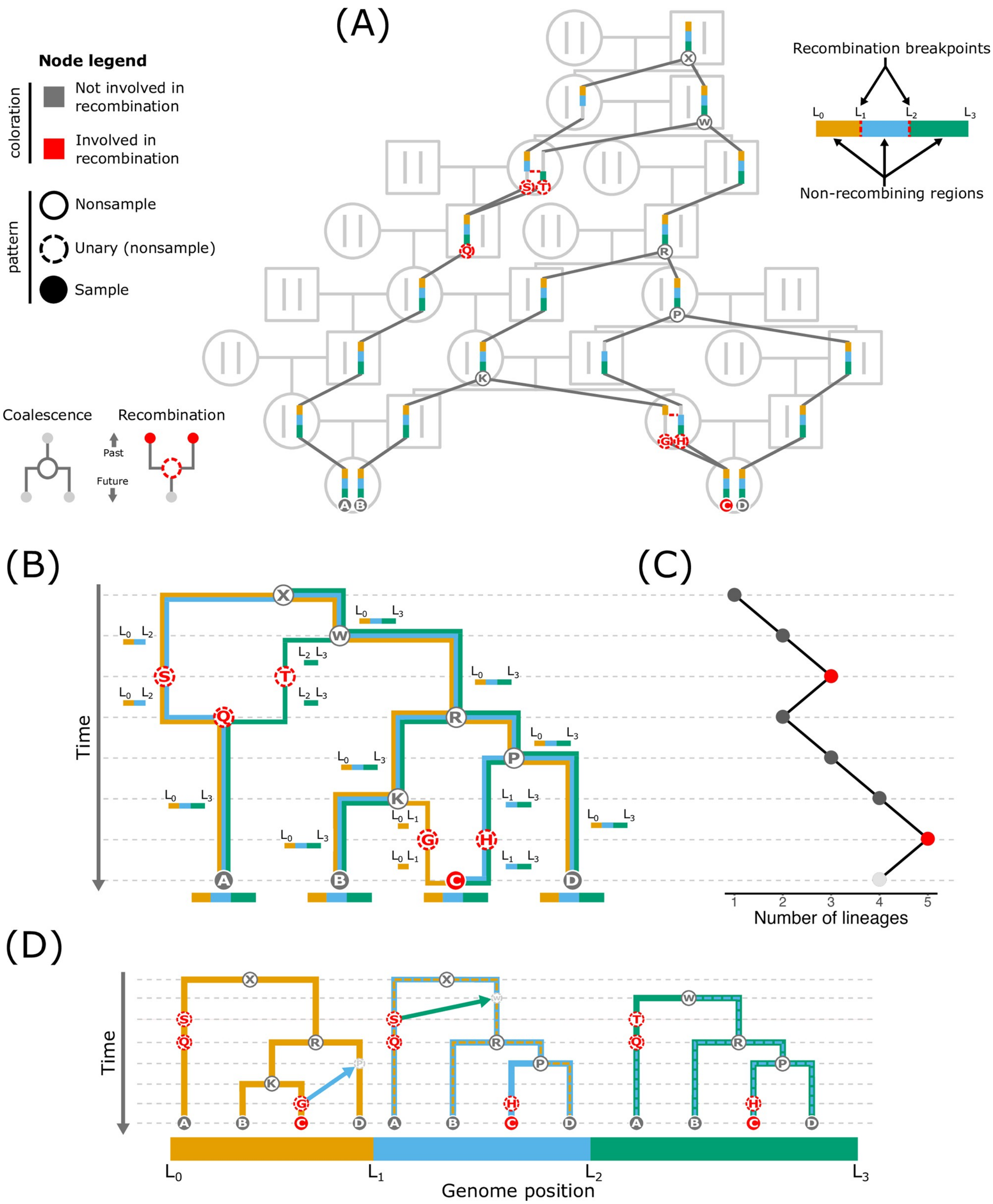


Fig 1. Overview of ARGs. In all ARG depictions (A, B, D), nodes are indicated by small circles, and each node represents a single set of one or more chromosomes (a haploid genome) of an individual. The node coloration indicates whether or not it is involved in recombination, and the specific pattern (shading and outline) of the node indicates its type: nonsample, unary (nonsample), sample. The genome is divided into 3 non-recombining regions (orange, blue, and green). (A) The relationships of multiple individuals can be organized into a pedigree (light gray portions). An ARG is embedded in a pedigree (portions in dark gray and color) and represents the set of pedigree paths through which genetic material is transmitted. (B)

The graphical representation of an ARG. Edges (the connections between nodes) are colored and annotated with the non-recombining region(s) that they transmit. (C) A plot recording the lineage count through time in the ARG. Backward in time, coalescent events, which occur at the dark gray points, merge lineages and thus reduce the lineage count. The red points highlight the times at which recombination occurs, which splits lineages backward in time and therefore increases the lineage count. (D) An ARG can be formulated as a series of local trees that share nodes and edges. Each non-recombining region possesses its own local tree. The regions are separated by a recombination event, which, when moving between regions, prunes a portion of the tree and regrafts it to another node. This action means that nearby trees are generally quite similar in structure. The arrows in the left 2 trees show how recombination relocates a branch in the tree (reconnecting to the small, light gray node) to form the tree of the region immediately to the right. The dashed lines in the second and third trees highlight each tree's shared structure with its leftward neighbor.

https://doi.org/10.1371/journal.pgen.1011110.g001

different histories of inheritance (and thus different genealogies) across the genome, with topological changes to the genealogy associated with recombination breakpoints and different chromosomes [30].

Ancestral recombination graphs

The complex web of genetic genealogies across the genome is recorded in a graphical structure known as an ARG, which provides extensive information regarding the history of inheritance for a set of sampled genomes. Each node in an ARG represents a haploid genome (a *haplotype*) in a real individual that exists now or in the past [25]. Each diploid individual therefore contains 2 haploid genomes and is represented by 2 nodes. We refer to nodes corresponding to sampled genomes (often, though not necessarily [31–33], sampled in the present) as *sample nodes* and all other nodes as *nonsample nodes*. If sample nodes have no sampled descendants, they constitute the tips of an ARG. Sample nodes are particularly salient because ARGs are generally specified in terms of the genetic ancestry of these genomes. Edges in an ARG indicate paths of inheritance between nodes. ARGs are technically described as "directed graphs" because genetic material flows unidirectionally from ancestors to descendants.

Assuming that sample nodes are sourced from contemporary individuals, the present time in an ARG (the bottom of the vertical axes in Fig 1B and 1D) contains a lineage (i.e., sets of one or more edges connected by nodes forming continuous paths of inheritance) for each sample. Tracing the lineages back in time, some nodes have 2 edges enter on the future-facing side but only a single outbound edge on the past-facing side (e.g., node Ⓡ in Fig 1B). These nodes represent haplotypes in which 2 lineages find common ancestry and thus merge into a single lineage, which reduces the lineage count by one (the dark gray points in Fig 1C). Common ancestry events additionally represent *coalescence* when (backward in time) the 2 merging edges contain the same portion of the genome (note that all nodes corresponding to common ancestry events in Fig 1 (Ⓚ, Ⓟ, Ⓡ, Ⓦ, and Ⓧ) also correspond to coalescence). From an organismal perspective, nodes corresponding to coalesence represent an instance in which a parent provides the same (portion of a) haploid genome to multiple offspring and thus splits a lineage into multiple lineages forward in time.

Conversely, other nodes have a single edge enter on the future-facing side but 2 edges exit the past-facing side (e.g., node Ⓠ in Fig 1B), which represents the outcome of recombination [2]. Backward in time, the node with 2 outbound edges on the past-facing side is the recombinant offspring node whose genome is inherited from 2 parental nodes (e.g., node Ⓒ in Fig 1). The 2 nodes that each receive one of the outbound edges are the parental nodes whose genomes are recombined in the offspring node. For example, in Fig 1, Ⓖ and Ⓗ are the parental nodes of Ⓒ. From an organismal perspective, these nodes occur when an offspring receives one of its haploid genomes from a parent and that haploid genome represents the outcome of recombination between the parent's 2 haploid genomes. Recombination splits the genome into separate lineages and thus each portion of the genome experiences a distinct history of inheritance between (traversing an ARG from present to past) the recombination event from which they split to the coalescence event in which they join back up. Consequently, each recombination event increases the number of lineages in an ARG by one (the red points in Fig 1C; [34]). From a forward-in-time perspective, recombination fuses portions of 2 parental genomes into a single haplotype (in the recombinant offspring), and thus unites separate lineages into a single lineage. Nodes through which genomic material that is eventually inherited by a sample node (hereafter *ancestral material*) is transmitted but are involved in neither common ancestry nor recombination for the ancestral material do not determine the topology of an ARG and thus are frequently omitted (we retain several of these nodes in Fig 1 to highlight the effects of recombination). More generally, nodes with only 1 descendant (*unary* nodes; e.g., node Ⓢ in Fig 1) do not directly influence genealogical relationships between the sample nodes. In simulations, unary nodes are often removed via a process called *simplification* [35].

ARGs generally record the timing of each node and the ancestral material that each edge transmits between ancestors and descendants. To trace the genealogy for a particular position in the genome, you follow the edges through the ARG that contain the focal position [22]. For example, in Fig 1B, to extract the genealogy for a position in the orange region (between positions L_0 and L_1) of sample node Ⓑ, you would follow the edges that transmit the orange region between nodes (i.e., Ⓑ → Ⓚ → Ⓡ → Ⓦ → Ⓧ).

The fact that each genomic region bracketed by recombination breakpoints (hereafter *non-recombining region*) possesses its own genealogy and that a non-recombining region in a single sample node traces only 1 path back to the MRCA of the entire sample suggests an alternative representation of an ARG: an ordered set of genealogical trees along the genome with labeled sample and nonsample nodes to specify how nodes are shared between trees (Fig 1D; [22]). Considering this representation of an ARG, which we refer to as the *tree representation*, is worthwhile because ARGs are often formulated (see Box 3) and operationalized in inference (e.g., [36–38]) based on this representation. In the tree representation, each non-recombining region has its own local tree that represents the region's evolutionary history. If each recombination breakpoint occurs at a unique position in the genome, as you shift from one local tree to the next (amounting to traversing one recombination breakpoint), the structure of the new tree is identical to its neighbor except for a single edge that is removed and then affixed to a (potentially) new node (Fig 1D). In computational parlance, this action is called a *subtree-prune-and-regraft* operation [39]. When all recombination events occur at unique locations and each event involves only 1 breakpoint, the total number of local trees will equal one more than the number of recombination events defining the evolutionary relationships in the genome. For example, in Fig 1 and 2, recombination events generate 3 trees. If recombination events occur at the same location (a breakpoint represents >1 recombination event), then moving between adjacent trees will involve a corresponding number of subtree-prune-and-regraft operations (one representing each recombination event), and the tree count will be less than one plus the number of recombination events.

Box 3: The succinct tree sequence

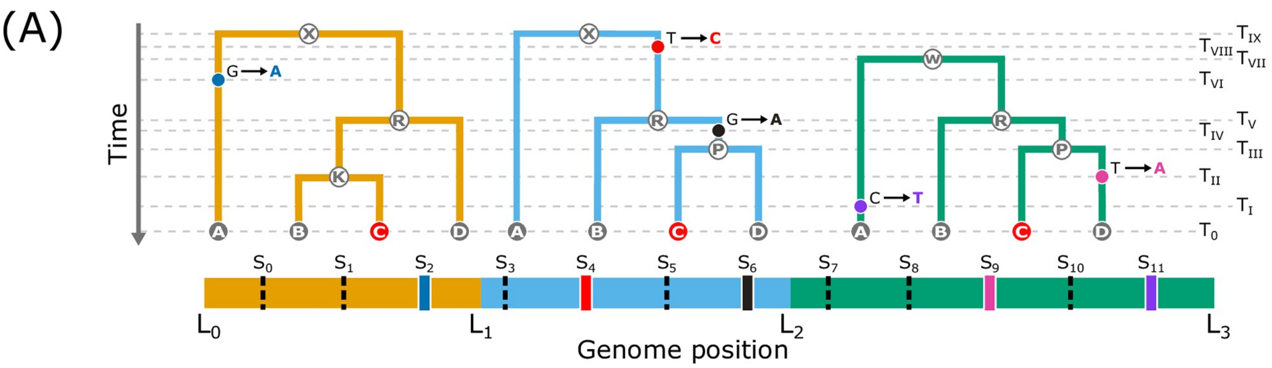
The correlated nature of an ARG's local trees can be exploited to compactly encode the trees in a data structure termed the *succinct tree sequence* or *tree sequence* for short (Fig 2A and 2B; [35,40]). The tree sequence defines the trees using 2 tables. The node table contains an identifier and the timing of each node (first table in Fig 2B). The edge table documents the edges shared between adjoining trees by recording the parent and offspring nodes of each edge and the contiguous extent of the genome that each edge covers (second table in Fig 2B). The key innovation here is that the data structure eliminates substantial redundancy. Instead of storing each tree independently, which would necessitate duplication of shared nodes and edges, the tree sequence records each shared component just once.

The basic tree sequence technically does not encode the full ARG, which includes all coalescent and recombination events. The basic tree sequence only explicitly contains information on the coalescent events and does not detail the timing and specific changes that differentiate adjacent trees; [41] explain this distinction as follows: the full ARG "encodes the events that occurred in the history of a sample" while the set of local trees recorded in the tree sequence "encodes the outcome of those events." Nonetheless, the tree sequence can be elaborated with recombination information to more exhaustively document genetic ancestry (e.g., [47,91]).

Several properties of the tree sequence have revolutionized ARG-based research. First, its concise nature means that an immensity of genealogical information can be stored in a highly compressed manner. The tree sequence is also a flexible format that can be augmented with additional tables to store other information such as location metadata and DNA data (e.g., third and fourth tables in Fig 2B; Fig 2A). Notably, relative to conventional genotype matrix formats (Fig 2C), DNA data can be represented much more efficiently using the tree sequence. For example, [41] estimated that the tree sequence format could store genetic variant data for 10 billion haploid human-like chromosomes in approximately 1 TB, which is many orders of magnitude smaller than the approximately 25 PB required to store these data in a VCF [67]. The efficiency of the tree sequence also permits significant speed-ups in computation (e.g., through the implementation of fast algorithms). These features have enabled advancements in the scale and scope of ARG-based analyses and are increasingly accessible given that the tree sequence underpins a growing ecosystem of methods and software including `tsinfer` [41], `sc2ts` [95], `ARGinfer` [91], `msprime` [47], and `tskit` [35] built to infer, simulate, and analyze ARGs. Further details on the tree sequence can be found in the papers introducing and expanding the tree sequence [35,40,91] and in the documentation of `tskit` [35].

With inclusion of all nodes involved in recombination and coalescence relevant to the sample nodes, it is straightforward to switch between the 2 ARG representations. As previously discussed, the local tree for a particular non-recombining region can be extracted from the graphical representation by starting at each sample node and tracing the lineages that transmit the region through the ARG until all lineages meet in the MRCA. Conversely, you can recover the graphical representation from the local trees by starting with the tree at one end of the set and then sequentially working across the trees, combining the shared nodes and edges, adding the nodes and edges that are not yet included in the graphical structure, and annotating each edge with the non-recombining region(s) that it transmits. As a brief illustration, in Fig 1D, the first 2 trees both contain nodes Ⓢ and Ⓠ with a connecting edge. In the graphical representation, these shared components would be merged and the edge would be annotated with the transmission of the regions between positions L_0 and L_2 (as shown in Fig 1B).

A recombination event can have several consequences for the structure of adjacent trees. First, it could alter the topology (i.e., the specific branching structure) if the new edge joins to a node on a different edge (e.g., the first and second trees in Fig 1D). However, if the new edge



(B) Succinct tree sequence

Nodes

Node ID	Age
A	T_0
B	T_0
C	T_0
D	T_0
K	T_{II}
P	T_{III}
R	T_{V}
W	T_{VII}
X	T_{IX}

Edges

Genome start	Genome end	Parent	Child
L_0	L_2	X	A
L_0	L_1	K	B
L_0	L_1	K	C
L_0	L_2	X	R
L_0	L_1	R	D
L_0	L_1	R	K
L_1	L_3	R	B
L_1	L_3	R	P
L_1	L_3	P	C
L_1	L_3	P	D
L_2	L_3	W	A
L_2	L_3	W	R

Sites

Site ID	Position	Ancestral
0	S_0	G
1	S_1	T
2	S_2	G
3	S_3	G
4	S_4	T
5	S_5	C
6	S_6	G
7	S_7	T
8	S_8	A
9	S_9	T
10	S_{10}	A
11	S_{11}	C

Mutations

Mutation ID	Site	Node	Derived	Age
0	2	A	A	T_{VI}
1	4	R	C	T_{VIII}
2	6	P	A	T_{IV}
3	9	D	A	T_{II}
4	11	A	T	T_{I}

(C) Genotype matrix

Site ID

Sample	0	1	2	3	4	5	6	7	8	9	10	11
A	G	T	A	G	T	C	G	T	A	T	A	T
B	G	T	G	G	C	C	G	T	A	T	A	C
C	G	T	G	G	C	C	A	T	A	T	A	C
D	G	T	G	G	C	C	A	T	A	A	A	C

Fig 2. The encoding of local trees and genotype data in the succinct tree sequence format. (A) Depiction of the local trees shown in Fig 1 with timing and location of mutation events mapped onto the branches and the location of each site shown on the genome. The black, dashed lines represent the invariant sites and the thicker, solid lines represent variant sites corresponding to each mutation. The trees are annotated with horizontal, dashed lines (labeled T_0–T_{IX}) that denote either the timing of coalescence or mutation events. (B) The trees and genotype data in the succinct tree sequence format. The trees are specified with the nodes and edges tables. The nodes table contains an ID and age for each node. The edges table contains the left (*Genome start*) and right (*Genome end*) positions of the genome over which each edge persists, while the *Parent* column contains the nodes that transmit material to the nodes in the *Child* column. The genotypic information is included in the sites [genomic position of each site (*Position*), ancestral state (*Ancestral*)] and mutations [derived state (*Derived*), mutation timing (*Age*)] tables. (C) The equivalent genotype data for the 4 sample nodes stored in a more conventional matrix format with the rows representing each sample node and the columns representing each genomic site. Note that with small amounts of genetic data such as this simple example, the tree sequence may require more storage space than a standard genotype matrix format. However, when considering realistic genomes, the tree sequence rapidly becomes much more efficient at storing genetic data with growing sample sizes [41].

https://doi.org/10.1371/journal.pgen.1011110.g002

joins to a different node on the same edge, the topology will remain unchanged, and only the edge lengths (i.e., coalescent times) will be modified (e.g., the second and third trees in Fig 1D). It is also possible for the lineage to coalesce back into the same node, which would result in no change to the tree structure. Each local tree contains every sample node because all samples possess the entire genome (and thus every non-recombining region represented by each tree). However, the collection of nonsample nodes can differ across trees. If an ARG includes all nodes (i.e., every nonsample node is retained), the absence of a node in a local tree signals that it does not represent a genetic ancestor for that region. If an ARG has been simplified (unary nodes removed), the absence of a node either means that it is not a genetic ancestor or that the node does not represent a genome in which coalescence occurred that involved the sample nodes.

There are several key characteristics of an ARG's tree representation. First, the subtree-prune-and-regraft operations that differentiate adjacent trees highlight that nearby trees are generally quite similar and frequently share many nodes and edges [28,30]. A series of shared nodes and edges between trees indicates that the corresponding non-recombining regions were found in the same lineage in that portion of the ARG. The correlated nature of the trees can be exploited for highly efficient tree storage and computation (Fig 2A and 2B; see Box 3 for further details; [35,40]). Second, although local trees can overlap in structure, a tree can contain components that are not universally found across the entire set of trees (e.g., in Fig 1D, node Ⓢ in the first tree is not found in the third tree). One feature that can frequently differ between trees is the node in which all sample nodes first find common ancestry (i.e., all lineages coalesce into a single lineage), which represents the region's *most recent common ancestor* (MRCA). When these local MRCAs exist at different times in the past, the trees will vary in height [28]. If all genomic regions trace their ancestry back to the same ancestor(s) in an ARG, the first node in which this occurs represents the *Grand MRCA* (GMRCA). It is possible for the same node to represent the GMRCA and one or more local MRCAs. For example, in Fig 1, node ⊗ is the GMRCA and the local MRCA for the first 2 non-recombining regions. However, this is not always the case. In fact, the GMRCA frequently predates any of the local MRCAs, which would result in it being absent from all of the (simplified) local trees.

Although the information contained in the graphical and tree representations of an ARG is the same, many readers, especially those with a background in phylogenetics, may prefer to think about ARGs via their tree representations. Unlike the graphical representation, each local tree is a familiar object: it is strictly bi- or multi-furcating, meaning that each node has exactly 1 ancestor and 2 or more descendants, and that therefore the tree contains no loops (i.e., it is non-reticulate), and is the desired result of a phylogenetic analysis run on a multiple sequence alignment of the DNA in the tree's non-recombining region. Building off this intuition, a phylogeneticist may draw on experience and imagine the set of local trees as analogous to a Bayesian posterior distribution of phylogenies. However, although this intuition may be initially useful, it is important to remember that each local tree is not independent of the others, both because each is generally separated from its neighbors by a small number of recombination events (so is therefore highly correlated), and because the same nodes and edges may appear across multiple local trees. The shared structure of trees imbues the nodes and edges with different properties relative to the analogous components in a standard phylogeny. For example, in a standard phylogeny, branches depict ancestor–descendant relationships through time and thus are one-dimensional. In contrast, edges in an ARG exist both through time and across the genome, and thus can be conceptualized as two-dimensional [42]. This two-dimensionality can be seen in Fig 1B where edges extend along the vertical, time dimension and also along different extents of the genome (edges contain different sets of genomic regions). Equivalently, the genome dimension of edges manifests in an ARG's tree representation (Fig 1D) through edges persisting across different sets of local trees. The overlapping nature of local trees (i.e., shared nodes and edges) underlies much of an ARG's utility and facilitates the power of ARG-based inference, which we discuss later in the review.

Modeling coalescence with recombination

In population genetics, ARGs are commonly generated by simulating under Hudson's [19] model of coalescent with recombination, which is closely connected to the original conception of ARGs [22]. Under this model, which assumes a Wright–Fisher population, a set of genomes exists in the present and the lineages describing each genome's ancestry are traced backward in time. Either coalescence or recombination can occur, which represent competing events with exponentially distributed waiting times. With coalescence, 2 lineages find common ancestry and merge into one. With recombination, a genomic position is selected uniformly as the breakpoint location. The offspring chromosome is inherited from one parental chromosome on one side of the breakpoint and the other parental chromosome on the other side. Recombination splits a lineage into 2 backwards in time. This process produces a series of genealogies across the genome that describes the ancestry of each genomic position. One question that may arise here is whether recombination could preclude the lineages from finding common ancestry because it increases the lineage count. However, backwards in time, the lineage count grows via recombination at a linear rate ($kR/2$ where k = lineage count and R = recombination rate), whereas lineages coalesce at a quadratic rate [$k(k–1)/2$], and thus finding common ancestry is guaranteed [22]. Later in the review, we will be simulating under this model to explore various features of ARGs.

ARGs in practice

In our introduction of ARGs, we mainly focus on the ancestors that are involved in common ancestry and recombination of ancestral material. However, when navigating the literature, it is important to recognize that the term *ancestral recombination graph* is frequently applied to structures that differ in various ways from each other and potentially from how we describe ARGs here. This variation stems from both terminological imprecision and inferential limitations.

The degree of completeness in which genetic inheritance from ancestors to descendants is documented can vary extensively. At the most comprehensive extreme, one could record all the genomic material that is passed between ancestors and descendants regardless of whether the material is ancestral or non-ancestral to the samples. Alternatively, one could render an ARG comprehensive to only the focal samples by only keeping track of the material that is ancestral to them (sometimes referred to as a *full ARG*). This structure could be further simplified in various ways such as removing nodes that are unary in one or more local trees. Although these descriptions of ancestry vary in the information that they include, they have all been referred to as ARGs in the literature [25].

Although ARGs may fully document genetic ancestry in theory, we rarely work with such a comprehensive structure in practice. First, in empirical settings, it is not possible to infer all of this information. The sample space of possible structures for a comprehensive ARG quickly becomes impractically vast with increasing genome and sample sizes. Hence, assumptions and shortcuts (e.g., the sequentially Markovian coalescent (SMC); [43]) are often employed [44], which sacrifices a capacity to infer a comprehensive and fully accurate ARG for the sake of computational tractability. There are also many components of ARGs that are largely unidentifiable and thus are necessarily omitted. For example, contemporary samples can provide only limited information on unary nodes, and certain features may be imperceptible in contemporary samples. An example of this is a "diamond" structure [44], where (going backward in time) recombination splits a lineage but then the lineages immediately coalesce again. Additionally, many sites in the genome are uninformative regarding the local tree topologies (e.g., invariant and singleton sites), which frequently precludes the identification of precise recombination breakpoint locations and other ARG features. More generally, patterns of shared variants represent the information from which ARGs are inferred, while recombination reduces the informative sites per genealogy by dividing the genome into smaller regions. ARG inference will therefore tend to decline in accuracy when the ratio of mutations to recombination is low [45]. This tension between mutation and recombination imposes a theoretical limit on ARG recoverability from sequencing data [46].

As a consequence of these obstacles, in practice, we are restricted in what we can infer about genetic ancestry from genomic data. For example, `tsinfer` [41] infers the collection of local trees and their shared structure (i.e., how nodes and edges overlap across trees) by first estimating ancestral haplotypes and then deducing the tree topologies by inferring how haplotypes relate to each other. This output can be thought of as representing the outcome of coalescence and recombination rather than completely encoding the events that generated the relationships [41]. That is, we are inferring the relationships across the genome produced by recombination and coalescence, but we lack detail on the recombination events that determine how these genealogies exactly knit together in an ARG. Importantly, even if we can acquire comprehensive information on genetic ancestry (e.g., in a simulation), many questions may only require certain subsets of this information, such as the structure of local trees. To accommodate both the existing terminological ambiguity and the reality of how well we can infer genetic ancestry, we permissively apply the term *ancestral recombination graph* to encompass structures that document genetic ancestry in the presence of recombination at varying levels of completeness.

Deepening ARG intuition with simulations

To further develop a foundational intuition for ARGs and reinforce content covered in the primer section, we implemented a series of simulations using `msprime` v1.2.0 [47] and the classical coalescent with recombination model. We completed post-simulation processing, analysis, and visualization using `tskit` [35], `numpy` [48], and `pandas` [49] in `Python 3.11.2` [50] and the following packages in `R 4.2.3` [51]: `TreeDist` [52], `ape` [53], `ggtree` [54], `dplyr` [55], `ggplot2` [56], `ggforce` [57], and `ggridges` [58]. We include all code on `GitHub` (https://github.com/AlexLewanski/arg_review).

First, to illustrate several general features of ARGs, we focus on a single simulation involving 1 population with an effective population size of 100 diploid individuals, a genome size of 10 kilobases (kb), a sample size of 12 diploid individuals, and a uniform recombination rate of 5×10^{-5} per base per generation. In the simulation, we recorded the full ARG, in which all nodes involved in common ancestry and recombination are retained. We then simplified the ARG structure, which involves removing unary nodes so that remaining nodes represent those that correspond to at least 1 coalescence event in the genome. Across the 593 local trees generated from this simulation, tree height (TMRCA of each non-recombining region) varied between 57.29 and 1,214.71 generations (non-integer generations are possible here because simulations involved a continuous time model) with a mean±standard deviation of 448.87 ±209.38 generations. The step-like pattern of tree height along the genome, in which height is constant for a stretch, then suddenly jumps to another value, appears because each tree (with a single height) applies to all sites in each non-recombining region (Fig 3A). As discussed in the primer section, another ubiquitous feature of ARGs is that nearby local trees are often highly similar. As a simple illustration of this, we quantified the dissimilarity of all pairwise combinations of local trees using the (approximate) subtree-prune-and-regraft (SPR) distance [59,60], which is the minimum number of subtree moves required to convert one tree to another only based on tip identities (ignoring identities of internal nodes). The topologies of nearby trees

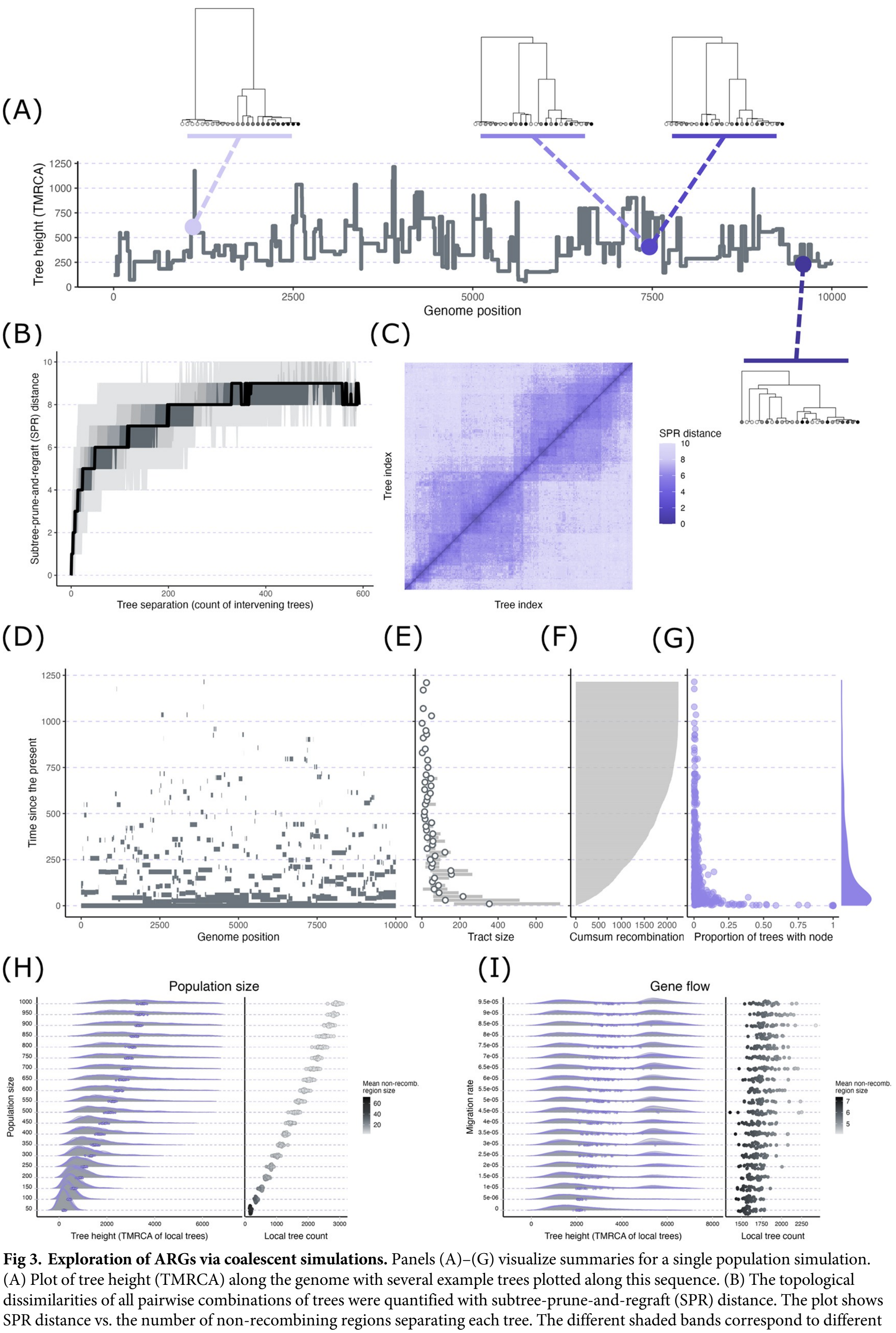


Fig 3. Exploration of ARGs via coalescent simulations. Panels (A)–(G) visualize summaries for a single population simulation. (A) Plot of tree height (TMRCA) along the genome with several example trees plotted along this sequence. (B) The topological dissimilarities of all pairwise combinations of trees were quantified with subtree-prune-and-regraft (SPR) distance. The plot shows SPR distance vs. the number of non-recombining regions separating each tree. The different shaded bands correspond to different percentiles of SPR distance values at each tree separation count: 0–100 (lightest gray), 10–90, 20–80, 30–70, 40–60, 50 (black line).

(C) Matrix of SPR distances for all combinations of trees organized by tree index (e.g., the 30th tree in the genome has an index of 30). (B) and (C) illustrate how nearby local trees are highly similar with similarity rapidly declining with growing number of breakpoints separating the trees. (D) Tracking the genomic material for 1 sample node back in time through its genetic ancestors recorded in the ARG. Continuous tracts of ancestral material get progressively smaller back in time as recombination repeatedly breaks the tracts into smaller pieces. (E) The sizes of tracts of ancestral material swiftly decline going back in time. The plot shows the mean (points) and 25th/75th percentiles of tract size (gray bars) for 20 generation bins. (F) The cumulative number of recombination events occurring backward in time. (G) The number of nodes and node sharing across local trees in an ARG quickly decline backward in time. The plot shows the location of each node in time (vertical axis) versus the proportion of local trees that contains each node (horizontal axis). The marginal density plot along the vertical axis shows the distribution of nodes through time. (H) A series of simulations with all conditions held constant except for population size. The left plot shows the distribution of tree height for each population size with the purple points representing the mean value per single simulation run. The right plot shows the mean tree count per simulation run with each point shaded with its mean non-recombining region size. (I) A series of simulations with all conditions held constant except for gene flow rate. The plots are identical to (H) except that they explore the effects of variation in gene flow rate instead of population size.

https://doi.org/10.1371/journal.pgen.1011110.g003

were highly similar, with similarity rapidly attenuating with increasing breakpoint separation between trees (Fig 3B). This can also be seen in the matrix of SPR distance values (Fig 3C), with lower values clustered around the diagonal (trees with similar indices and few intervening non-recombining regions) and values rapidly increasing away from this region. The attenuating similarity can also be qualitatively observed in the example trees included in Fig 3A, where the second and third trees, which are adjacent (the 437th and 438th trees, respectively), appear highly similar and are both clearly different in structure compared to the more distant first (45th) and fourth (576th) trees.

Next, using the same simulation, we tracked genetic material found in the sample nodes back in time through the samples' ancestors recorded in the ARG (ancestral material). Because we had already simplified the ARG, every non-sample node that we examined represents a common ancestor of at least 2 sample nodes. The simplified nature of the ARG also means that we were only tracking material back to the local MRCAs of each region. Fig 3D displays the location of ancestral material (horizontal axis) and the timing of the ancestor (the age of the node) carrying that material (vertical axis) for a single sample node from the simulation. At the contemporary time point (time = 0), the tract of ancestral material spans the entire genome because this represents the sample node that by definition possesses its entire genome as a single haplotype. Traveling back in time (up the vertical axis in Fig 3D), the tracts of ancestral material are broken up into small pieces. Consequently, the average tract length of ancestral material peaks in the contemporary time period and rapidly declines back in time (Fig 3E; this figure is based on ancestral material of all sample nodes). This pattern emerges because the cumulative number of recombination events that have occurred in the transmission of ancestral material grows through time (Fig 3F), resulting in the fragmentation of ancestral material into progressively smaller pieces.

This pattern can also be understood through the lens of node-sharing across the local trees. At the present, every node is shared across all trees because all regions of the genome are found in each sample node. However, moving back in time, the tracts of ancestral material become progressively smaller and thus span fewer non-recombining regions. This results in a decline in node-sharing across trees further back in time; any particular node is carrying ancestral material for a decreasing number of non-recombining regions. Fig 3G depicts this phenomenon. Nodes with the highest proportion of sharing between trees are exclusively located near the present, while nodes located further back in time (higher up the vertical axis) show low proportions of sharing. The reduced node-sharing through time corresponds to variation in how quickly the trees change at different time periods. Near the present, the high degree of node-sharing means that tree structures remain fairly stable. However, the more rapid turnover of nodes at deeper time points translates into faster changes as you move across the trees further back in time.

A variety of variables can systematically modify features of an ARG. As a brief illustration, we examined how effective population size and gene flow, which frequently vary across studies and systems, influence 3 fundamental features: tree height, the number of local trees, and the size of non-recombining regions in an ARG. For the population size demonstration, we completed a set of simulations that kept all variables constant (sequence length = 10 kb, recombination rate = 3 × 10^{-5} per base per generation, sample size = 10 diploid individuals) except for population size, which varied between 50 and 1,000 in increments of 50 (a total of 20 population sizes with 30 replicates per size). Tree height and local tree count both increased while mean region size decreased at greater population sizes (Fig 3H). The correlations between population size and the 3 variables emerge because, with higher effective population sizes, coalescent times will tend to increase because more individuals exist that act as possible ancestors and thus there is a lower probability of any 2 lineages finding common ancestry in a particular generation (see [61,62] for accessible introductions to this and other fundamental coalescent concepts). Because of the deeper coalescent times (which result in greater tree heights), more opportunities exist for recombination to occur, which results on average in more local trees and smaller non-recombining regions.

We generated another set of simulations for the gene flow demonstration where we kept all variables constant (sequence length = 10 kb, recombination rate = 3 × 10^{-5} per base per generation) except for migration. We simulated 2 populations of 500 individuals each that merged (backwards in time) after 5,000 generations. While the populations were separated, one of the populations (the *recipient population*) experienced continuous, unidirectional gene flow from the second population (the *donor population*) forward in time. We varied the migration rate between 0 and 9.5 × 10^{-5} in increments of 5 × 10^{-6} (a total of 20 different migration rates with 30 replicates per rate). We then sampled 10 diploid individuals from the recipient population. With increasing gene flow, trees tended to increase in height on average, which was associated with increasing bimodality in the distribution of tree heights. This bimodality phenomenon emerges because the presence of 2 populations along with gene flow result in 2 distinct time periods during which lineages can coalesce [63,64]. The left mode of the distribution corresponds to non-recombining regions whose entire history postdating the population split occurred within the recipient population, and thus coalescence for that region could occur fairly rapidly (small TMRCA values). However, with gene flow, part of a non-recombining region's history can occur in the donor population. Consequently, a region whose ancestry involves the donor population must wait until the 2 populations merge in the ancestral population before finding its MRCA. This results in the second, later mode in tree heights. The slight trends of increasing tree count and decreasing region size at greater migration rates occur because the tree heights are increasing on average, which provides opportunities for more recombination events.

Note that the ARG summaries we have reported here—tree height, number of local trees, length of non-recombining regions, similarity and node-sharing between local trees—only represent a small glimpse into the innumerable ways that ARGs can be dissected and summarized. We chose this set to exemplify fundamental features of ARGs and illustrate how they reflect and can therefore be informative about demographic and evolutionary phenomena that are frequently of interest in evolutionary genomics.

ARGs in evolutionary genomics

From a practical perspective, 2 questions logically ensue from the ARG introduction: what is the utility of ARGs in evolutionary genomics, and what advantages do they impart relative to existing approaches? As with many methodological advances, ARGs can offer multiple benefits, including strengthening our ability to answer existing questions and opening up entirely new fields of inquiry.

To understand how ARGs facilitate empirical inferences that are equal or superior to existing approaches, it is helpful to consider 2 topics: (1) how ARGs are shaped by evolutionary phenomena; and (2) how ARGs juxtapose with the prevailing approaches to address questions in evolutionary genomics. A critical idea is that the genealogies underlying the genome are the ultimate record of evolutionary history. The structure of an ARG is governed by processes, including selection, drift, and gene flow, that regulate the fitness and relatedness of haplotypes. The genomic composition of individuals is precisely reflected in an ARG's structure because ARGs encode the ancestral source(s) of samples' genomes, including how new mutations are propagated through time and across individuals (Fig 2A). Consequently, the genomes of sampled individuals and any summary of their content represent derivatives of the underlying ARG, and many of these genomic summaries can be reinterpreted as explicit descriptions of the ARG [65,66].

Currently in evolutionary genomics, genomic data are typically stored as a genotype matrix (e.g., a VCF file [67]; Fig 2C). The data are distilled down to a variety of summaries such as principal components [68,69], *F*-statistics [70–72], or the site frequency spectrum (SFS) that each reflect particular attributes of the samples' genomes. From these measures, we attempt to infer past phenomena (e.g., selection, demographic changes) that gave rise to the observed data, under the premise that disparities in the generative process translate to corresponding differences in genomic summaries. Indeed, these summaries can often provide powerful and accurate insights into evolutionary processes, and the field of statistical population genetics has made extraordinary strides in divining evolutionary processes from summaries of genetic and genomic data in the 6 decades following the first empirical measurements of molecular genetic variation [73]. As previously discussed, each summary measure calculated from these data (e.g., the SFS, F_{ST}, π, θ, individual heterozygosity, identity-by-state, identity-by-descent) is a low-dimensional summary of an ARG, so, to the extent that our ARG reconstruction methods are accurate and statistically consistent [74], we can recover any of these quantities at least as accurately as they are estimated from the genomic data from which an ARG is inferred [66]. [See [65,66] for instructive discussions of the ways common summaries of genomic data (and many other quantities) can be calculated and interpreted with ARGs]. And, because ARGs can offer computational efficiencies over traditional methods of storing genomic data, in many cases these quantities can be calculated more easily, and with less computational overhead, from ARGs [66,75].

A growing assortment of methods is demonstrating the strengths of ARG-explicit approaches. For example, [75] devised a method to efficiently represent linkage disequilibrium (LD) based on genomic genealogies (*LD graphical models*). These LD graphical models enable orders-of-magnitude reductions in computation time and memory usage for LD matrix computations and facilitate better polygenic prediction compared to a similar method using the LD correlation matrix. As another example, [76] found that an expected genetic relatedness matrix calculated from an ARG in a given genomic region more accurately captures relationships than the empirical genetic relatedness matrix calculated in the same region. The higher accuracy of these approaches may seem counterintuitive; after all, empirical ARGs are estimated from genomic data, so how could statistical inferences conducted on an ARG be *more* accurate than those made directly from the genotype matrix? To see how this can occur, consider the structure of the genealogies that comprise an ARG. Each local tree is usually separated from that of the adjacent non-recombining region by a small number of recombination events, leading to high correlation in the genealogical relationships contained in nearby trees (e.g., Figs 1D, 3B, and 3C). Because of this correlation, the other trees contain information about relatedness between samples in a focal tree. The mutational process is intrinsically random, so that the true genealogical relationships between a set of samples may not be apparent in patterns of shared variation associated with any particular region. By leveraging the information about relationships between samples contained across the entire set of trees, we can, in principle, side-step some of the "noise" in the data that exists due to the randomness of the mutational process [66].

Reference [38] provides another illustration of how ARGs can improve inferences in evolutionary genomics. The authors introduced a new framework (`gLike`) for calculating the likelihood of a genealogy under a defined demographic history; this framework can be used for demographic inference from an ARG's local trees. Using simulations, they found that, compared to a popular SFS-based method (`fastsimcoal2` [77]), `gLike` yields more accurate inferences based on both the true genealogies and genealogies inferred from the simulated genomic data. The better performance of `gLike` arises because a local tree's structure (e.g., coalescent times, topology) is sensitive to the underlying demographic history, and thus explicit analysis of the trees and their attributes can lead to accurate inferences. In contrast, information is inherently discarded by working with summaries of an ARG like the SFS. The coarseness of this summary can result in multiple histories mapping to the same SFS in some situations, rendering precise and accurate recovery of the true history more difficult.

Beyond facilitating more efficient and accurate inferences, the increasing availability of empirical ARGs will foster entirely new fields of ARG-based inquiry. A useful analogy here is the way in which the field of phylogenetics opened up the associated field of phylogenetic comparative methods. For example, the question of whether diversification rates vary across a phylogeny [78,79] is impossible to pose, let alone answer, without a phylogeny. It is difficult to guess what form the "comparative methods" field of ARGs (i.e., not just asking existing questions better or faster, but entirely new questions that are predicated on ARGs) will take, especially as empirical ARG inference is still in its infancy. However, we can highlight one particularly exciting direction that has already begun to materialize: geographic inference with ARGs.

The recent advances in the reconstruction of ARGs have sparked a revolution in spatial population genetics. In particular, several recent approaches [33,80] have begun to explore the feasibility of inferring the locations of the genetic ancestors of sampled individuals across space and through time. Although similar geographic inference has been done using non-recombining gene regions (e.g., [81–83]) or a single phylogenetic tree ("phylogeography" [84]), it is only with an ARG in hand that it has become feasible to infer locations for *all* the genetic ancestors of a sample. This power, in turn, has facilitated massively more detailed and nuanced understanding of how organisms move across space and through time. For example, [80] inferred the mean effective dispersal distance of *Arabidopsis thaliana* and [33] recovered the broad strokes of human dispersal history over the last 800,000 years. In the future, this type of inference of ancestral locations could empower specific and biologically principled definitions of "admixture" (e.g., 12.5% of the genetic ancestors of a focal individual are estimated to have lived inside a particular geographic region at a particular slice of time) [85]. The exciting enterprise of geographic inference of ancestor locations (more precisely, of the geographic locations of nodes in an ARG) and of the concomitant historical patterns of dispersal and density described by a sample's georeferenced genealogy, is entirely predicated on the existence of an inferred ARG for a set of samples.

An important qualifier to this discussion is that, despite the evident promise of ARG-based inference, there is still uncertainty about the extent and contexts that this promise can be realized in empirical applications. One of the main reasons for this uncertainty is, despite some awareness of empirical limits on ARG reconstruction, much remains unknown regarding the degree of accuracy needed to make quality downstream inferences from ARGs. For example, do accurate inferences generally require high precision and accuracy in all features of ARGs? Or do some questions perhaps only require accuracy in specific features of ARGs, such as node heights *or* local topologies. Available evidence suggests that this latter statement is likely true. For example, for ARG-based reconstruction of demographic history, [38] found that inferences were more accurate using ARGs inferred with `tsinfer+tsdate` [33,41] than `Relate` [86], which they attribute to better estimates of recent coalescent times by `tsdate`. Conversely, calculations of genome-wide expected relationship matrices were more accurate with ARGs from `Relate` than `tsinfer+tsdate`, which may stem from `Relate`'s more accurate reconstruction of deeper portions of ARGs [38,87]. More exhaustive knowledge of the sensitivities and requirements of downstream inferences will help uncover the particular facet(s) of ARG reconstruction whose improvements would be most beneficial, and will also help delineate the limits that empirical ARG reconstruction will enforce on downstream inferences.

Conclusions

In this review, we aimed to introduce ARGs, articulate the capacity of ARGs to enhance the study of evolutionary genomics, and describe the current and/or forthcoming practicability of using ARGs in empirical- and simulation-based research. Indeed, ARGs have the potential to advance evolutionary genomics in both minor and profound ways ranging from improving implementation of existing approaches (e.g., faster calculation of traditional population genetics statistics) to inspiring novel and previously inaccessible avenues of study. The nature and extent to which ARGs will reshape the field remains unclear and will depend on fundamental limits regarding the information contained in empirical ARGs, the degree to which ARGs are integrated into the methods canon of evolutionary genomics, and our collective ingenuity.

How do we fully capitalize on ARGs? First, a broader suite of inference methods and tools based on ARGs must (continue to) be developed, evaluated, and made readily accessible to the broader community. Until now, most ARG-based methods development has concentrated on ARG reconstruction and simulation. Although these topics will benefit from additional progress, we are reaching a stage where empirical- and simulation-based ARGs can be realistically acquired in many situations and readily stored and manipulated with an increasingly mature and powerful software infrastructure (e.g., `tskit`). A more expansive body of methods built on ARGs will enable wider adoption of ARG-based inference. The incipient nature of ARG methods presents an opportunity for more extensive synthesis and synergy between evolutionary genomics and both phylogenetic comparative methods and phylogeography. These fields have developed a sizeable assortment of phylogenetic methods that could be co-opted and modified for tree-based inference in the context of ARGs. Fully capitalizing on our growing ARG capabilities will clearly require a receptivity to new genealogically explicit approaches and ideas that have so far only featured sparingly in empirical evolutionary genomics. However, with a concerted embrace of ARGs, we are confident that this "holy grail of statistical population genetics" [45] will further realize its potential for many questions in evolutionary biology.

Acknowledgments

We thank Peter Ralph, Sarah Fitzpatrick, Yan Wong, Jerome Kelleher, and members of the Fitzpatrick and Bradburd labs for helpful discussions and manuscript feedback.

Author Contributions

Conceptualization: Alexander L. Lewanski, Michael C. Grundler, Gideon S. Bradburd.

Project administration: Alexander L. Lewanski.

Supervision: Gideon S. Bradburd.

Visualization: Alexander L. Lewanski.

Writing – original draft: Alexander L. Lewanski, Gideon S. Bradburd.

Writing – review & editing: Alexander L. Lewanski, Michael C. Grundler, Gideon S. Bradburd.

References

1. Marques DA, Meier JI, Seehausen O. A Combinatorial View on Speciation and Adaptive Radiation. Trends Ecol Evol. 2019; 34(6):531–544. https://doi.org/10.1016/j.tree.2019.02.008 PMID: 30885412
2. Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. The genomic consequences of hybridization. Elife. 2021; 10:e69016. https://doi.org/10.7554/eLife.69016 PMID: 34346866
3. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020; 20:555–572. https://doi.org/10.1038/s41568-020-0290-x PMID: 32778778
4. Schluter D, Rieseberg LH. Three problems in the genetics of speciation by selection. Proc Natl Acad Sci U S A. 2022; 119(30):e2122153119. https://doi.org/10.1073/pnas.2122153119 PMID: 35858397
5. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. Nature. 2015; 16:333–343. https://doi.org/10.1038/nrg3931 PMID: 25963372
6. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008; 23 (1):38–44. https://doi.org/10.1016/j.tree.2007.09.008 PMID: 18006185
7. Epps CW, Palsbøll PJ, Wehausen JD, Roderick GK, Ramey RR, McCullough DR. Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. Ecol Lett. 2005; 8 (10):1029–1038. https://doi.org/10.1111/J.1461-0248.2005.00804.X
8. Machado CB, Braga-Silva A, Freitas PD, Galetti PM. Damming shapes genetic patterns and may affect the persistence of freshwater fish populations. Freshw Biol. 2022; 67(4):603–618. https://doi.org/10.1111/FWB.13866
9. Jasper M, Schmidt TL, Ahmad NW, Sinkins SP, Hoffmann AA. A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. Mol Ecol Resour. 2019; 19 (5):1254–1264. https://doi.org/10.1111/1755-0998.13043 PMID: 31125998
10. Chen N, Cosgrove E, Bowman R, Fitzpatrick J, Clark A. Genomic Consequences of Population Decline in the Endangered Florida Scrub-Jay. Curr Biol. 2016; 26(21):2974–2979. https://doi.org/10.1016/j.cub.2016.08.062 PMID: 27746026
11. Aguillon SM, Fitzpatrick JW, Bowman R, Schoech SJ, Clark AG, Coop G, et al. Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. PLoS Genet. 2017; 13(8). https://doi.org/10.1371/journal.pgen.1006911 PMID: 28771477
12. Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, et al. Allele frequency dynamics in a pedigreed natural population. Proc Natl Acad Sci U S A. 2019; 116(6):2158–2164. https://doi.org/10.1073/pnas.1813852116 PMID: 30598449
13. Larkin DM, Daetwyler HD, Hernandez AG, Wright CL, Hetrick LA, Boucek L, et al. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. Proc Natl Acad Sci U S A. 2012; 109(20):7693–7698. https://doi.org/10.1073/pnas.1114546109 PMID: 22529356
14. Ma L O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, et al. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. PLoS Genet. 2015; 11(11):1005387. https://doi.org/10.1371/journal.pgen.1005387 PMID: 26540184
15. Griffiths RC. Neutral two-locus multiple allele models with recombination. Theor Popul Biol. 1981; 19 (2):169–186. https://doi.org/10.1016/0040-5809(81)90016-2
16. Kingman JFC. The coalescent. Stoch Process their Appl. 1982; 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4
17. Kingman JFC. On the genealogy of large populations. J Appl Probab. 1982; 19(A):27–43. https://doi.org/10.2307/3213548

18. Tajima F. Evolutionary relationships of DNA sequences in finite populations. Genetics. 1983; 105 (2):437–460. https://doi.org/10.1093/GENETICS/105.2.437 PMID: 6628982

19. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 1983; 23(2):183–201. https://doi.org/10.1016/0040-5809(83)90013-8 PMID: 6612631

20. Griffiths RC. The Two-Locus Ancestral Graph. Lecture Notes-Monograph Series. 1991; 18:100–117.

21. Griffiths RC, Marjoram P. Ancestral Inference from Samples of DNA Sequences with Recombination. J Comput Biol. 1996; 3(4):479–502. https://doi.org/10.1089/cmb.1996.3.479 PMID: 9018600

22. Griffiths RC, Marjoram P. An ancestral recombination graph. In: Donelly P, Tavaré S, editors. Progress in Population Genetics and Human Evolution. New York: Springer-Verlag; 1997. p. 257–270.

23. Wiuf C, Hein J. Recombination as a Point Process along Sequences. Theor Popul Biol. 1999; 55 (3):248–259. https://doi.org/10.1006/tpbi.1998.1403 PMID: 10366550

24. Hein J, Schierup MH, Wiuf C. The coalescent with recombination. In: Gene genealogies, variation and evolution: a primer in coalescent theory; 2005. p. 127–171.

25. Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. A general and efficient representation of ancestral recombination graphs. bioRxiv. 2023; p. 2023.11.03.565466. https://doi.org/10.1101/2023.11.03.565466 PMID: 37961279

26. Donnelly KP. The probability that related individuals share some section of genome identical by descent. Theor Popul Biol. 1983; 23(1):34–63. https://doi.org/10.1016/0040-5809(83)90004-7 PMID: 6857549

27. Chang JT. Recent common ancestors of all present-day individuals. Adv Appl Probab. 1999; 31 (4):1002–1026. https://doi.org/10.1239/AAP/1029955256

28. Hudson RR. Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. vol. 7. Oxford: Oxford University Press; 1991. p. 1–44.

29. Mathieson I, Scally A. What is ancestry? PLoS Genet. 2020; 16(3):6–11. https://doi.org/10.1371/journal.pgen.1008624 PMID: 32150538

30. Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet. 2002; 3(5):380–390. https://doi.org/10.1038/nrg795 PMID: 11988763

31. Schaefer NK, Shapiro B, Green RE. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. Sci Adv. 2021; 7(29):776–792. https://doi.org/10.1126/sciadv.abc0776 PMID: 34272242

32. Speidel L, Cassidy L, Davies RW, Hellenthal G, Skoglund P, Myers SR. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. Mol Biol Evol. 2021; 38(9):3497–3511. https://doi.org/10.1093/molbev/msab174 PMID: 34129037

33. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified genealogy of modern and ancient genomes. Science. 2022; 375(6583). https://doi.org/10.1126/science.abi8264 PMID: 35201891

34. Nordborg M. Coalescent theory. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of Statistical Genomics. Chichester, UK: Wiley; 2001. p. 179–212.

35. Kelleher J, Thornton KR, Ashander J, Ralph PL. Efficient pedigree recording for fast population genetics simulation. PLoS Comput Biol. 2018; 14(11):e1006581. https://doi.org/10.1371/journal.pcbi.1006581 PMID: 30383757

36. Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. 2019; 15(9):e1008384. https://doi.org/10.1371/journal.pgen.1008384 PMID: 31518343

37. Hejase HA, Mo Z, Campagna L, Siepel A. A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. Mol Biol Evol. 2022; 39(1). https://doi.org/10.1093/molbev/msab332 PMID: 34888675

38. Fan C, Cahoon JL, Dinh BL, Vecchyo DOD, Huber C, Edge MD, et al. A likelihood-based framework for demographic inference from genealogical trees. bioRxiv. 2023. https://doi.org/10.1101/2023.10.10.561787 PMID: 37873208

39. Song YS. On the Combinatorics of Rooted Binary Phylogenetic Trees. Ann Comb. 2003; 7(3):365–379. https://doi.org/10.1007/S00026-003-0192-0

40. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016; 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842 PMID: 27145223

41. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nat Genet. 2019; 51(9):1330–1338. https://doi.org/10.1038/s41588-019-0483-y PMID: 31477934

42. Shipilina D, Pal A, Stankowski S, Chan YF, Barton NH. On the origin and structure of haplotype blocks. Mol Ecol. 2023; 32:1141–1457. https://doi.org/10.1111/mec.16793 PMID: 36433653

43. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci. 2005; 360(1459):1387–1393. https://doi.org/10.1098/rstb.2005.1673 PMID: 16048782

44. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. PLoS Genet. 2014; 10(5):e1004342. https://doi.org/10.1371/journal.pgen.1004342 PMID: 24831947

45. Hubisz M, Siepel A. Inference of ancestral recombination graphs using ARGweaver. Methods Mol Biol. 2020; 2090:231–266. https://doi.org/10.1007/978-1-0716-0199-0_10 PMID: 31975170

46. Hayman E, Ignatieva A, Hein J. Recoverability of ancestral recombination graph topologies. Theor Popul Biol. 2023; 154:27–39. https://doi.org/10.1016/j.tpb.2023.07.004 PMID: 37544486

47. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022; 220(3):iyab229. https://doi.org/10.1093/genetics/iyab229 PMID: 34897427

48. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020; 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2 PMID: 32939066

49. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010. p. 51–56.

50. Python Software Foundation. Python; 2023.

51. R Core Team. R: A Language and Environment for Statistical Computing. 2023. Available from: https://www.R-project.org/.

52. Smith MR. TreeDist: Calculate and Map Distances Between Phylogenetic Trees. 2023. Available from: https://CRAN.R-project.org/package=TreeDist.

53. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018; 35:526–528.

54. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017; 8(1):28–36. https://doi.org/10.1111/2041-210X.12628

55. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2019. Available from: https://cran.r-project.org/package=dplyr.

56. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al.. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics; 2023.

57. Pedersen TL. ggforce: Accelerating ggplot2. 2022.

58. Wilke CO. ggridges: Ridgeline Plots in ggplot2. 2022.

59. Hein J, Jiang TB, Wang LC, Zhang K. On the complexity of comparing evolutionary trees. Appl Math. 1996; 71:153–169.

60. de Oliveira ML, Leal E, Kishino H. Phylogenetic Detection of Recombination with a Bayesian Prior on the Distance between Trees. PLoS ONE. 2008; 3(7):e2651. https://doi.org/10.1371/journal.pone.0002651 PMID: 18612422

61. Coop G. Genetic Drift and Neutral Diversity. In: Population and Quantitative Genetics; 2020. p. 59–89.

62. Pritchard JK. More on genetic drift: The coalescent. In: An Owner's Guide to the Human Genome: An introduction to human population genetics, variation and disease. 2023. p. 85–103.

63. Maruyama T. Effective number of alleles in a subdivided population. Theor Popul Biol. 1970; 1(3):273–306. https://doi.org/10.1016/0040-5809(70)90047-x PMID: 5527634

64. Rosenberg NA, Feldman MW. The relationship between coalescence times and population divergence times. In: Slatkin M, Veuille M, editors. Modern Developments in Theoretical Population Genetics. Oxford: Oxford University Press; 2002. p. 130–164.

65. Ralph PL. An empirical approach to demographic inference with genomic data. Theor Popul Biol. 2019; 127:91–101. https://doi.org/10.1016/j.tpb.2019.03.005 PMID: 30978307

66. Ralph P, Thornton K, Kelleher J. Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. Genetics. 2020; 215(3):779–797. https://doi.org/10.1534/genetics.120.303253 PMID: 32357960

67. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

68. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic Maps of Human Gene Frequencies in Europeans. Science. 1978; 201(4358):786–792. https://doi.org/10.1126/science.356262 PMID: 356262

69. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009; 5(10). https://doi.org/10.1371/journal.pgen.1000686 PMID: 19834557

70. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009; 461(7263):489–494. https://doi.org/10.1038/nature08365 PMID: 19779445

71. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192(3):1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212

72. Peter BM. Admixture, population structure, and F-statistics. Genetics. 2016; 202(4):1485–1501. https://doi.org/10.1534/genetics.115.183913 PMID: 26857625

73. Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in nature populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics. 1966; 54 (2):577–594. https://doi.org/10.1093/GENETICS/54.2.577 PMID: 5968642

74. Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics. 2022; 221(1). https://doi.org/10.1093/GENETICS/IYAC044 PMID: 35333304

75. Nowbandegani PS, Wohns AW, Ballard JL, Lander ES, Bloemendal A, Neale BM, et al. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. Nat Genet. 2023; 55:1494–1502. https://doi.org/10.1038/s41588-023-01487-8 PMID: 37640881

76. Link V, Schraiber JG, Fan C, Dinh B, Mancuso N, Chiang CWK, et al. Tree-based QTL mapping with expected local genetic relatedness matrices. bioRxiv. 2023. https://doi.org/10.1101/2023.04.07.536093 PMID: 37066144

77. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 2013; 9(10). https://doi.org/10.1371/journal.pgen.1003905 PMID: 24204310

78. Ricklefs RE. Estimating diversification rates from phylogenetic information. Trends Ecol Evol. 2007; 22(11):601–610. https://doi.org/10.1016/j.tree.2007.06.013 PMID: 17963995

79. Rabosky DL. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. PLoS ONE. 2014; 9(2):e89543. https://doi.org/10.1371/journal.pone.0089543 PMID: 24586858

80. Osmond MM, Coop G. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. bioRxiv. 2021. https://doi.org/10.1101/2021.07.13.452277

81. Neigel JE, Avise JC. Application of a random walk model to geographic distributions of animal mitochondrial DNA variation. Genetics. 1993; 135(4):1209–1220. https://doi.org/10.1093/genetics/135.4.1209 PMID: 8307331

82. Barton NH, Wilson I. Genealogies and geography. Philos Trans R Soc Lond B Biol Sci. 1995; 349 (1327):49–59. https://doi.org/10.1098/rstb.1995.0090 PMID: 8748019

83. Avise JC. Phylogeography: retrospect and prospect. J Biogeogr. 2009; 36(1):3–15. https://doi.org/10.1111/J.1365-2699.2008.02032.X

84. Knowles LL. Statistical Phylogeography. Annu Rev Ecol Evol Syst. 2009; 40:593–612. https://doi.org/10.1146/ANNUREV.ECOLSYS.38.091206.095702

85. Bradburd GS, Ralph PL. Spatial Population Genetics: It's About Time. Annu Rev Ecol Evol Syst. 2019; 50:427–449. https://doi.org/10.1146/annurev-ecolsys-110316-022659

86. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nat Genet. 2019; 51:1321–1329. https://doi.org/10.1038/s41588-019-0484-x PMID: 31477933

87. Fan C, Mancuso N, Chiang CWK. A genealogical estimate of genetic relationships. Am J Hum Genet. 2022; 109(5):812–824. https://doi.org/10.1016/j.ajhg.2022.03.016 PMID: 35417677

88. Hubisz MJ, Williams AL, Siepel A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS Genet. 2020; 16(8):e1008895. https://doi.org/10.1371/journal.pgen.1008895 PMID: 32760067

89. Marjoram P, Wall JD. Fast "coalescent" simulation. BMC Genet. 2006; 7(1):1–9. https://doi.org/10.1186/1471-2156-7-16 PMID: 16539698

90. Heine K, Beskos A, Jasra A, Balding D, De Iorio M. Bridging trees for posterior inference on ancestral recombination graphs. Proc R Soc A. 2018; 474:20180568. https://doi.org/10.1098/rspa.2018.0568 PMID: 30602937

91. Mahmoudi A, Koskela J, Kelleher J, Yb C, Balding D. Bayesian inference of ancestral recombination graphs. PLoS Comput Biol. 2022; 18(3):e1009960. https://doi.org/10.1371/journal.pcbi.1009960 PMID: 35263345

92. Zhang BC, Biddanda A, Gunnarsson AF, Cooper F, Palamara PF. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat Genet. 2023; 55:768–776. https://doi.org/10.1038/s41588-023-01379-x PMID: 37127670

93. Vaughan TG, Welch D, Drummond AJ, Biggs PJ, George T, French NP. Inferring ancestral recombination graphs from bacterial genomic data. Genetics. 2017; 205(2):857–870. https://doi.org/10.1534/genetics.116.193425 PMID: 28007885

94. Didelot X, Lawson D, Darling A, Falush D. Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. Genetics. 2010; 186(4):1435–1449. https://doi.org/10.1534/genetics.110.120121 PMID: 20923983

95. Zhan SH, Ignatieva A, Wong Y, Eaton K, Jeffery B, Palmer DS, et al. Towards Pandemic-Scale Ancestral Recombination Graphs of SARS-CoV-2. bioRxiv. 2023. https://doi.org/10.1101/2023.06.08.544212

96. Deng Y, Song YS, Nielsen R. The distribution of waiting distances in ancestral recombination graphs. Theor Popul Biol. 2021; 141:34–43. https://doi.org/10.1016/j.tpb.2021.06.003 PMID: 34186053

97. Ignatieva A, Favero M, Koskela J, Sant J, Myers SR. The distribution of branch duration and detection of inversions in ancestral recombination graphs. bioRxiv. 2023. https://doi.org/10.1101/2023.07.11.548567

98. Rasmussen DA, Guo F. Espalier: Efficient tree reconciliation and ARG reconstruction using maximum agreement forests. Syst Biol. 2023; 72:1154–1170. https://doi.org/10.1093/SYSBIO/SYAD040 PMID: 37458753

99. Kaplan NL, Darden T, Hudson RR. The Coalescent Process in Models with Selection. Genetics. 1988; 120(3):819. https://doi.org/10.1093/genetics/120.3.819 PMID: 3066685

100. Hudson RR, Kaplan NL. The coalescent process in models with selection and recombination. Genetics. 1988; 120(3):831–840. https://doi.org/10.1093/genetics/120.3.831 PMID: 3147214

101. Walczak AM, Nicolaisen LE, Plotkin JB, Desai MM. The Structure of Genealogies in the Presence of Purifying Selection: A Fitness-Class Coalescent. Genetics. 2012; 190(2):753–779. https://doi.org/10.1534/genetics.111.134544 PMID: 22135349

102. Spencer CCA, Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics. 2004; 20(18):3673–3675. https://doi.org/10.1093/bioinformatics/bth417 PMID: 15271777

103. Kern AD, Schrider DR. Discoal: flexible coalescent simulations with selection. Bioinformatics. 2016; 32 (24):3839–3841. https://doi.org/10.1093/bioinformatics/btw556 PMID: 27559153

104. Anderson-Trocmé L, Nelson D, Zabad S, Diaz-Papkovich A, Kryukov I, Baya N, et al. On the genes, genealogies, and geographies of Quebec. Science. 2023; 380(6647):849–855. https://doi.org/10.1126/science.add5300 PMID: 37228217

105. Messer PW. SLiM: Simulating evolution with selection and linkage. Genetics. 2013; 194(4):1037–1039. https://doi.org/10.1534/genetics.113.152181 PMID: 23709637

106. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol Biol Evol. 2019; 36(3):632–637. https://doi.org/10.1093/molbev/msy228 PMID: 30517680

107. Haller BC, Messer PW. SLiM 4: Multispecies Eco-Evolutionary Modeling. Am Nat. 2023; 201(5): E127–E139. https://doi.org/10.1086/723601 PMID: 37130229

108. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol Ecol Resour. 2019; 19(2):552–566. https://doi.org/10.1111/1755-0998.12968 PMID: 30565882

109. Campagna L, Mo Z, Siepel A, Uy JAC. Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species. PLoS Genet. 2022; 18(11): e1010474. https://doi.org/10.1371/journal.pgen.1010474 PMID: 36318577

110. Korfmann K, Sellinger T, Freund F, Fumagalli M, Tellier A. Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent. bioRxiv. 2023. https://doi.org/10.1101/2022.09.28.508873