nature reviews genetics

https://doi.org/10.1038/s41576-024-00772-4

Review article

Check for updates

Inference and applications of ancestral recombination graphs

Rasmus Nielsen [1,2,3] ✉, Andrew H. Vaughn[3,4] & Yun Deng[3,4]

Abstract

Ancestral recombination graphs (ARGs) summarize the complex genealogical relationships between individuals represented in a sample of DNA sequences. Their use is currently revolutionizing the field of population genetics and is leading to the development of powerful new methods to elucidate individual and population genetic processes, including population size history, migration, admixture, recombination, mutation and selection. In this Review, we introduce the readers to the structure of ARGs and discuss how they relate to processes such as recombination and genetic drift. We explore differences and similarities between methods of estimating ARGs and provide concrete illustrative examples of how ARGs can be used to elucidate population-level processes.

Sections

[1]Department of Integrative Biology and Department of Statistics, UC Berkeley, Berkeley, CA, USA. [2]GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [3]Center for Computational Biology, UC Berkeley, Berkeley, CA, USA. [4]These authors contributed equally: Andrew H. Vaughn, Yun Deng. ✉e-mail: rasmus_nielsen@berkeley.edu

Introduction

Coalescent theory has formed the foundation for analyses of population genetic data since the invention of PCR and DNA sequencing technologies[1–5]. The central object of coalescent theory is the coalescence tree (Fig. 1), which summarizes information about the genetic relationships between sampled individuals. The leaves (tips) of the tree represent individual DNA sequences in the sample; the edges (alternatively referred to as lineages or branches) represent lines of descent; and the internal nodes represent coalescence events (also known as coalescent events), that is, the points in time at which the lineages of specific individuals, or groups of individuals, in the sample merge (or 'coalesce') into their most recent common ancestor (MRCA). The root of the tree represents the MRCA of all individuals represented in the sample.

Kingman[1] and Hudson[2,6] discovered in the early 1980s that the relationship between individuals in a population sample could be described by a binary tree, that is, a tree in which each node has either zero children (leaf nodes) or two children (internal nodes), and that population genetic models of genetic drift provide simple predictions regarding the structure of the tree. This observation quickly led to the emergence of many statistical methods for inferring population-level parameters such as population size changes[7] and migration rates[8–10], and it spurred the development of new methods to detect natural selection using population genetic data[6,11,12]. The central insight forming the basis of these methods is that all the information in DNA sequencing data regarding population-level processes is represented by the coalescence tree. If the coalescence tree can be inferred with 100% accuracy, no more information is obtainable about these demographic processes from the sequence data. This principle led to the development of full likelihood and Bayesian statistical methods that achieved their objectives by integrating over the set of possible coalescence trees using Markov chain Monte Carlo (MCMC) or other stochastic methods; that is, they took uncertainty regarding the structure of the coalescence trees into account by considering many trees weighted by their relative probabilities[9,10,13,14].

By the time the human genome had been sequenced in 2003, a mature set of tree-based statistical methods for analysis of population genetic data had emerged[9,10,13,14].There was only one problem: these methods assumed no recombination and were therefore useful only for mitochondrial DNA, chloroplast DNA, the Y chromosome, some viral sequencing data and a few other types of marker that are not subject to recombination. However, with the emergence of next-generation sequencing technologies, these methods were rapidly becoming outdated, as the focus now changed to genomic data with an abundance of recombination. For genomic data, there is not just one coalescence tree, but as many coalescence trees as there are recombination events in the history of the data. Each recombination event splits a segment of DNA into two separate segments, each of which will have a different coalescence process (also known as coalescent process), thus forming distinct trees on each side of the recombination break point. Some approaches[15,16] attempted to deal with this problem by assuming that the genome could be divided into short genomic regions with free recombination between regions but no recombination within regions. However, mutation rates and recombination rates in humans are fairly similar[17,18], which means that, on average, every time a new mutation occurs that might provide information about the tree structure, a new recombination event will also occur that changes the tree structure. Furthermore, many of the methods aimed at short non-recombining regions simply did not scale computationally to thousands of genomic regions. Consequently, the field moved slowly away from full likelihood and Bayesian methods that analysed coalescence trees. Instead, the focus shifted to composite likelihood methods based on treating SNPs as if they were independent[19–21], analysing the distribution of allele frequencies[22–24], using approximate Bayesian computation[25], focusing on pairs of sequences[26] or other methods that used only a small subset of the information in the data. This was the state of the field until 2014, when it was demonstrated that inferences of coalescence trees in models with recombination are in fact computationally feasible, even for genome-scale data[27]. Using clever sampling methods, it was possible to estimate (and sample from a posterior) the coalescence trees along the length of the genome in a joint structure called an ancestral recombination graph (ARG). This development opened up the possibility of extracting much more information from population genetic sequencing data and, perhaps, developing full likelihood and Bayesian methods for inference in population genetics using ARGs (Fig. 1A) in place of individual coalescence trees (Fig. 1B). Here, we review developments in the field of ARG inference and analysis, with a special focus on the prospect of using ARGs as a general framework for rigorous statistical inferences in population genetics.

Ancestral recombination graphs

ARGs and the coalescent with recombination

The coalescence process with recombination (CwR), first described in 1983 (ref. 2), is formulated as a stochastic process running from the present backwards in time, allowing recombination events and coalescence events to occur until a MRCA has been found for each locus. In later work[28], this process was developed further, and the term ARG was coined to describe the random graph structure arising as a product of the originally described CwR[2]. The basic concept of ARGs is that they represent a coalescence process in which ancestral lineages not only merge by coalescence but also split owing to recombination (Fig. 1A). From the ARG, the individual coalescence trees for each segment of the genome can be deduced (Fig. 1B). Recombination will often generate trees with different topologies, although sometimes they may just differ in having different branch lengths (Fig. 1). The ARG contains the information regarding all the coalescence trees in the genome, but it also contains information about recombination events and their location on the lineages of the coalescence trees. Clearly, for genomic data with many individuals and tens or hundreds of thousands of recombination events, the ARG can become very complex. As a result, it is common to make some simplifications. In particular, the time of a recombination event on a lineage cannot be inferred from data and is, therefore, often not represented in the ARG. Further simplifications and approximations are often used to enable inferences based on ARGs.

ARGs as sequences of trees

The stochastic process generating ARGs[2,28] tracks recombination and coalescence events running from the present backwards in time. Another way of looking at the process generating ARGs is to consider how coalescence trees change along the length of the genome. As the trees can be deduced from the ARG, we might consider a process that runs from one end of the chromosome to the other end, generating coalescence trees that consider the specific recombination and coalescence events that affect each genomic segment. This process was shown to be non-Markovian[29], meaning that the probability distribution of a coalescence tree in one position depends not only on the coalescence tree in the previous position but also on all other trees along the length of the chromosome. Hence, considering the sequential process of trees

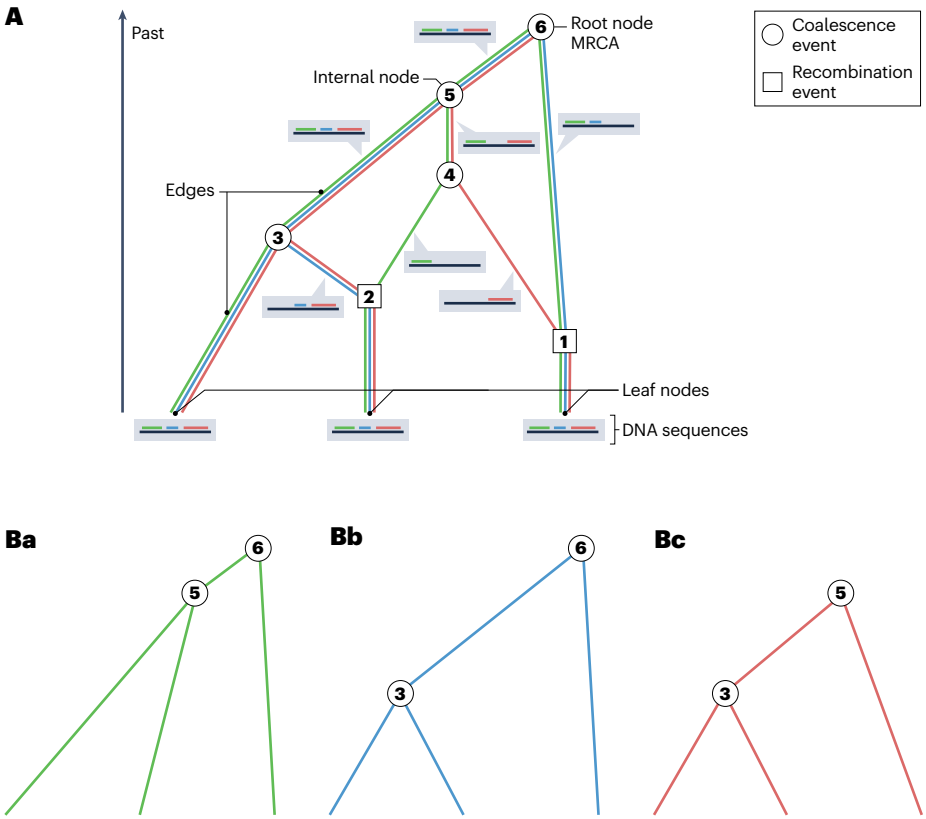

Fig. 1 | The ancestral recombination graph and derived coalescence trees. A, An example ancestral recombination graph (ARG) for three sample DNA sequences that traces the coalescence process back in time until a most recent common ancestor (MRCA) is found. Coalescence trees consist of a set of nodes connected by edges. The nodes that represent the observed DNA sequences are the leaves (leaf nodes) in the tree, and the nodes that represent coalescence events are internal nodes. The node at the top of the tree represents the MRCA of the three sequences. Only the internal nodes, and not the leaf nodes, are numbered and represented by circles in this depiction. Also, the trees are depicted with the root at the top and the leaves at the bottom, as is the tradition in computer science. In the ARG, circles represent coalescence events and squares represent recombination events. Each recombination event that occurs on a lineage breaks up the DNA sequence represented by that lineage into two segments, so the two recombination events in the history of these samples (1 and 2) separate the sample sequences into three segments (labelled with green, blue and red), each of which has a distinct coalescence tree. The coalescence events (3, 4, 5 and 6) merge lineages together into a common ancestor. Note that a lineage represents a line of descent that contains genetic material for some segment (or segments) of the genome that is ancestral to one or more individuals in the sample. The DNA segments containing ancestral genetic material for the three DNA sequences are depicted adjacent to each edge. For example, the lineage going from node 2 to node 4 contains only ancestral material for the first green segment of the DNA sequence. The rest of the DNA sequence on this lineage is not ancestral to any of the three individuals in the sample. **B**, Coalescence trees for three segments of DNA sequence (green, blue and red), respectively. All three coalescence trees are embedded in, that is, can be deduced from, the ARG (part **A**). Although the topologies are identical, the trees differ by having different branch lengths.

as they change along the length of the chromosome is computationally fairly complex because it is necessary to simultaneously keep track of all trees for purposes such as inference and simulation.

For a simpler and more computationally tractable process, a model was proposed that could approximate the tree-generating process along the length of the genome as a Markovian process, termed sequentially Markovian coalescent (SMC)[30]. The idea was to allow only coalescence events between lineages whose ancestral genetic material overlapped. For example, in Fig. 1A, a coalescence event occurs between two lineages that contain only ancestral genetic material for the green and the red genomic segments, respectively, forming node 4. Under the SMC process, such coalescence events are not allowed, as the coalescence tree at a particular position is allowed to depend only on the tree at the previous position, leading to a much-simplified process.

Whereas the SMC process allows only coalescence events between lineages that contain at least some shared ancestral genetic material, a refinement of the SMC process, called SMC′, also allows coalescence events between lineages with no shared genetic material, but with genetic material from adjacent loci[31]. This small extension of the SMC does not lead to much added computational complexity, but it drastically improves the accuracy of the approximation. Subsequent work[32] has shown that, by most measures, the SMC′ process is a very close approximation to the full CwR[28] and is today, arguably, the preferred approximation to the CwR used for population genetic inferences. The possible coalescence events allowed under the CwR, SMC or SMC′ model are detailed in Fig. 2.

Various assumptions about the process generating ARGs lead to different definitions of ARGs[33]. Some ARGs allowed in the full CwR are

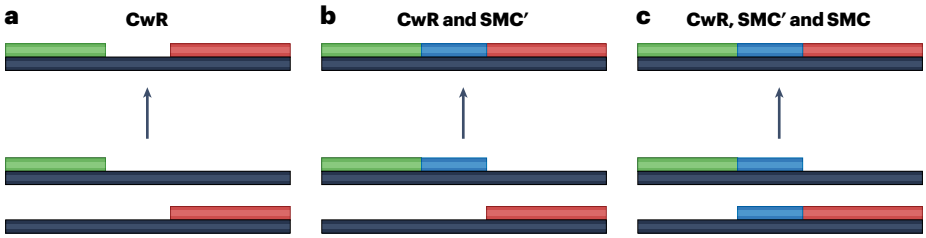

Fig. 2 | Examples of coalescence events permitted by different models of the CwR. a, A coalescence event between two lineages carrying non-overlapping and non-adjacent segments of DNA. Such coalescence events are allowed only under the full coalescent with recombination (CwR) and not by SMC′ or SMC. **b**, A coalescence event between two lineages carrying adjacent but non-overlapping segments of DNA. Such coalescence events are allowed under both the CwR and SMC′ processes but not under SMC. **c**, A coalescence event between overlapping segments of DNA. Such coalescence events are allowed under all three models. The limits that each of the SMC′ and SMC models place on the types of allowed coalescence restrict the state space of ancestral recombination graphs that can be simulated under these models, with SMC inducing a more severe restriction than SMC′. SMC, sequentially Markovian coalescent.

not supported by the SMC′ process, and further ARGs that are supported by the SMC′ process are not allowed by the SMC process. Sometimes, ARGs can be represented even more simply, as just a sequence of trees without explicit information about recombination and/or identification of which coalescence events are shared between trees[34,35]. Whether these sequences of trees are truly ARGs is then a matter of definition. However, irrespective of how ARGs are defined, they can be considered as products of a coalescence process governed by population genetic factors such as population sizes, migration rates and selection, and molecular forces such as mutation and recombination. The ARGs contain information about these processes, and methods based on ARGs that can extract this information have the potential to take on a central role in population genetics in the genomic era, similar to that of coalescence trees in the 1990s.

Inference of ARGs

Representations of ARGs and paradigms of ARG inference

Many different methods to infer ARGs have been developed, including Margarita[34], ARGweaver[27], RentPlus[36], Arbores[37], Relate[35], tsinfer + tsdate[38,39], ARGweaver-D[40], SARGE[41], KwARG[42], ARGinfer[43], ARG-Needle[44], SINGER[45], among others. We note that tsinfer and tsdate were tools developed separately to infer ARG topology and internal node age, respectively; we use ‘tsinfer + tsdate’ to refer to the combined use of these methods. These methods differ in many ways, including the way they represent ARGs, which is connected to implicit assumptions regarding the underlying generative process. Methods representing ARGs as a series of coalescence trees, without necessarily enforcing that the trees share nodes or branches with each other, include Relate[35] and RentPlus[36]. This representation is at times convenient but is not efficient in that many trees embedded in ARGs share the same nodes and edges (Fig. 1), which becomes particularly storage-inefficient with large sample size[46]. Furthermore, this representation destroys some of the naturally occurring correlation structure between trees. The concept of a correlated sequence of trees underlies simulation algorithms such as FastCoal[31], fastsimcoal/fastsimcoal2 (refs. 24,47) and ms[46] as well as inference programs such as tsinfer[38], ARG-Needle[44], SINGER[45] and ARGweaver[27]. It can also lead to a compact ‘succinct tree sequence’ data structure for the sequence data, as implemented in the widely used simulation software msprime[48–50], in which each node and edge only needs to be represented once, even though it may be a component of many trees. This ‘tree sequence’ representation of an ARG as a set of trees with shared nodes and edges[33,48–50] results in dramatically increased simulation speed and storage efficiency, and many current ARG inference methods provide the option to produce output in this compact format (for example, tsinfer, ARG-Needle and SINGER). msprime[48–50], which introduced this data format, is the most commonly used software for simulating ARGs and was the first software to fully take advantage of this data structure for fast ARG simulations.

The inference of ARGs is a computationally difficult problem because the number of possible graph topologies can be enormous. Additionally, the observed sequence data provide limited information about the structure of the local tree in any particular position of the genome. Some methods, such as ARGweaver/ARGweaver-D, ARGinfer and SINGER, address this problem by taking a Bayesian approach, where ARGs are sampled from a posterior distribution by MCMC. Instead of inferring a single ARG, many possible ARGs are sampled, which provides a representation of the uncertainty in ARG inference and facilitates rigorous downstream statistical analyses. Although Relate and tsinfer + tsdate infer a single fixed ARG topology, Relate additionally samples coalescence times for the inferred topology and tsdate outputs metadata about the marginal distribution of the age of each node. Most other methods currently provide only point estimates of both ARG topology and branch lengths. As the sampling of ARGs is computationally demanding, methods that estimate only a single ARG tend to be much faster than methods that sample multiple ARGs.

Scalability, accuracy, data requirements and output

ARG inference methods are in rapid development and are constantly improving. We provide an overview here and refer to recent comprehensive comparisons of various methods[51–53], which measure various aspects of evolutionary inference, such as recombination, coalescence times, allele frequencies and polygenic score histories.

Current ARG inference methods differ, for example, with respect to model assumptions, computational speed, inference accuracy, what type of inferences are made, and which data types are supported. Whereas some methods require phased modern data, other methods can use other types of data, including ancient DNA and genotype array data. ARGweaver and ARGweaver-D are the only methods that can use unphased data; tsinfer-sparse (a variation of tsinfer) and ARG-Needle are the only methods that support genotype array data. ARGweaver-D, Relate, tsinfer + tsdate and SARGE can support the use of ancient DNA, which must be computationally imputed and phased for all programs except ARGweaver-D. Relate and tsinfer

additionally require the polarization of ancestral states, which typically is inferred using outgroup species. The applicability of the different methods is summarized in Table 1.

In terms of computational speed, methods such as ARGweaver, Arbores, KwARG and ARGinfer can only handle tens of sequences, with KwARG and ARGinfer also being limited to relatively short sequences. Methods such as RentPlus, SARGE and SINGER can handle hundreds of whole-genome sequences, and Relate can handle thousands of sequences including the full 1000 Genomes Project data[54]. The most scalable methods now are tsinfer and ARG-Needle, which can handle several hundred thousand sequences including the full UK Biobank data[55] and other large genome-wide association study (GWAS) data.

Simulation studies have found that the fastest methods often, but not always, have the worst performance[51–53], which is not surprising as with the ARG inference problem, as in many other computational problems, there is a trade-off between computational speed and accuracy. However, methods with comparable speed sometimes have very different statistical properties, often with more recent methods outperforming older methods. We refer the reader to recent simulation comparisons for more specifics on the accuracy of the various methods[51–53].

As previously discussed, the methods also differ in whether they provide measurements of statistical uncertainty in the estimates, with ARGinfer, ARGweaver, ARGweaver-D and SINGER being the only methods providing full probabilistic modelling of statistical uncertainty. Finally, the methods differ in the representation of the ARG provided, with some methods, such as Relate, estimating trees in windows, whereas other methods, such as SINGER, ARGweaver and ARGweaver-D, provide graphs that represent a sequence of trees separated by individual recombination events.

Application of ARGs for population genetic inferences

Inference of demography and selection

ARGs can be used to make detailed and powerful inferences regarding population genetic processes, such as population size history, migration, natural selection, mutation rate and recombination rate. A major area where ARGs have been of use is in the inference of demography and the genealogical relationship between individuals. The rate of coalescence at a certain time point is inversely proportional to the effective population size (N_e) at that time. Therefore, one can use the temporal density of coalescence times of an ARG to estimate N_e through time (Fig. 3a).

ARGs have also been used to estimate the parameters of complex demographic models by extracting coalescence trees across the genome and considering the different possible migration histories of the lineages in each tree[56]. This idea of considering the possible paths taken by lineages through a demographic model has also been leveraged for local ancestry inference, in which each tree along the genome provides information about the path that tree took through the demographic model and, therefore, about the local ancestry at the segment spanned by this tree[57,58]. ARGs can also be used to infer natural selection. Inference methods to estimate the selection coefficient for a single SNP using ARGs have proceeded by extracting the local tree around the focal SNP and then examining the relative coalescent rate of lineages carrying the derived allele and those carrying the ancestral allele. For example, an allele under positive selection ($s > 0$) is expected to increase in frequency forwards in time more rapidly than a neutral allele ($s = 0$) (Fig. 3b). Therefore, looking backwards in time, the positively selected allele will seem to rapidly decrease in frequency, meaning the number of lineages carrying this allele will decline quickly. We would then expect a high density of coalescence events of lineages with the selected allele, analogous to the previous discussion of small population sizes generating high densities of coalescence events[59,60]. The usage of the coalescence density of lineages carrying different alleles to infer selection can either be done analytically by explicitly computing the likelihood and integrating over the derived allele frequency[61,62] or by applying machine learning techniques on the inferred ARGs themselves[63,64]. These methods can also be slightly modified to examine selection acting on multiple SNPs concurrently, that is, polygenic selection[65], or to reconstruct polygenic scores through time[52,66].

In addition to inferring demography and selection, there have been other exciting applications of ARGs for inferences, for example, the combination of inferred ARGs with a linear mixed model framework to refine association analyses of variants with complex traits[44].

Table 1 | The applicability of different ARG inference methods

	Samples branch lengths	Samples topologies	Scales to 1000 Genomes Project	Scales to UK Biobank	Supports ancient DNA	Supports genotyping array
ARGweaver	Yes	Yes	No	No	Yes	No
RentPlus	No	No	No	No	No	No
Arbores	Yes	Yes	No	No	No	No
Relate	Yes	No	Yes	No	Yes	No
tsinfer+tsdate	No	No	Yes	Yes	Yes	Yes
ARGweaver-D	Yes	Yes	No	No	Yes	No
SARGE	No	No	No	No	No	No
KwARG	No	No	No	No	No	No
ARGinfer	Yes	Yes	No	No	No	No
ARG-Needle	No	No	Yes	Yes	No	Yes
SINGER	Yes	Yes	No	No	No	No

The wording 'Samples branch lengths' and 'Samples topologies' refers to whether the methods just provide a single point estimate or sample multiple instances to provide measures of statistical uncertainty. ARG, ancestral recombination graph.

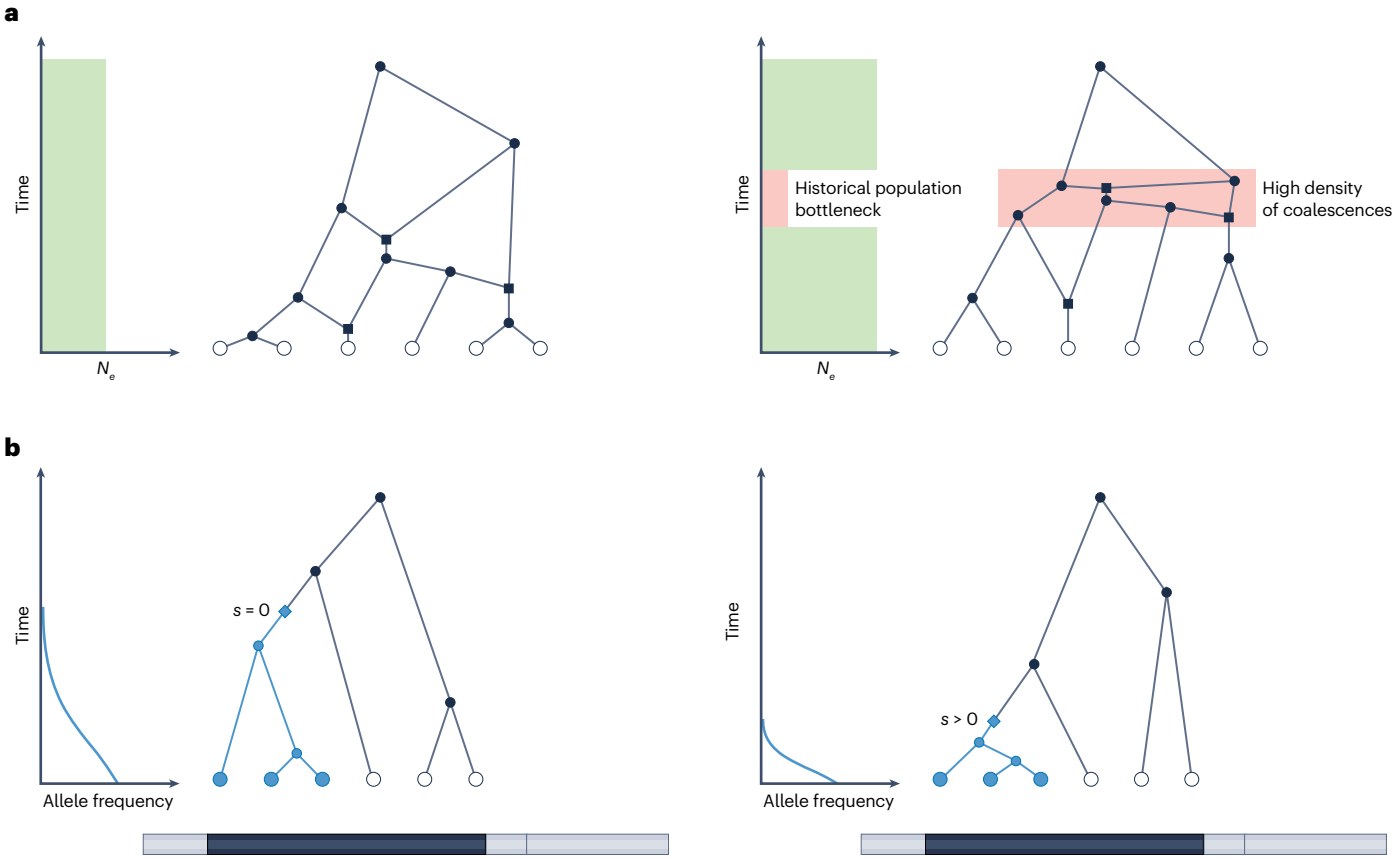

Fig. 3 | Different population genetic parameters are reflected in the ARG. **a**, Changes in historical population size are reflected in the ancestral recombination graph (ARG). The left panel shows an ARG in the presence of constant population size through time. The right panel shows an ARG in the presence of a historical population bottleneck. The smaller effective population size (reduced N_e) during this period results in a higher density of coalescences across the whole ARG compared with the constant population size history. Squares represent recombination events and circles represent coalescence events. **b**, Natural selection at a SNP results in a more rapid increase in the frequency of the selected allele, leading to a higher density of coalescences of lineages carrying the selected allele (shown in blue) at the tree spanning the SNP. Both panels show a coalescence tree for a particular region of the genome (marked in dark grey); of note, not the whole ARG is shown. A mutation on a lineage (shown as a blue diamond) causes all descendant lineages to inherit that allele. If this allele is neutral ($s = 0$) (left), no difference in coalescence density is expected. However, if this allele is advantageous ($s > 0$) (right), an increase in the coalescence density of blue lineages is expected.

A particular strength of their method was the ability to accurately detect large-effect associations of rare variants on sparse genotype array data. Another promising area of research is the use of ARGs to explicitly model the spatial location of ancestral lineages of a given sample[39,67–69]. Furthermore, ARGs have been used to study the evolution of the human mutation spectrum by examining the mappings of mutations to the branches of the ARG to study changes in mutation frequency over time[70]. A similar approach of implicitly considering the possible mappings of mutations to branches of coalescence trees was taken to estimate the ages of alleles[71], which in turn were used to infer historical generation times in different human populations[72]. Although controversies exist about the specific methodology used[73], the framework used nevertheless illustrates the exciting potential of ARGs to help understand complex molecular and demographic processes. Nascent ARG-based methods have also recently been developed to infer identity-by-descent segments[74], detect chromosomal inversions[75], compute time-stratified, tree-based *f*-statistics[76], simulate and analyse quantitative traits[77,78], model linkage disequilibrium[79] and simulate local ancestries[80]. In addition, developments in incorporating ARGs into forwards-in-time simulations, which enable ARGs to be generated from a much wider set of molecular and population-level processes than is possible with backwards-in-time simulations, has greatly extended the set of inference problems to which ARGs can be applied[49,81].

Using ARGs for full probabilistic inferences

Many ARGs are possible with a given data set, but analyses that use ARGs often rely on a single ARG estimated from the data. This treatment of a single inferred graph as the true graph greatly simplifies subsequent analyses, but it presents two distinct problems. First, it does not account for uncertainty in the underlying ARG, inducing false certainty in the results of any analysis that uses a single ARG as the true ARG. Second, any ARG is inferred under certain implicit or explicit assumptions (for example, no selection, constant population

size, no population structure) and will therefore be biased, as the true population genetic parameters may differ from the assumed scenario. Hence, the unobserved nature of the ARG necessitates both an integration over uncertainty in the ARG and a bias correction to achieve full probabilistic inference.

Future methods can leverage ARGs for full probabilistic inferences in one of two ways. The first way is through a statistical technique known as importance sampling[82–84]. In this version of importance sampling, a large set of ARGs is sampled under a specific model with specific assumptions about parameters, such as effective population sizes or selection coefficients (Fig. 4). These ARGs are then assigned weights that are inversely proportional to their probabilities under the sampling model. By reweighting the samples appropriately, one can approximate the likelihood function for a population genetic parameter of interest evaluated at any parameter value and thereby provide estimates of the parameters[61,62]. ARGs that are more likely under the original set of parameters used to sample from the posterior are given low weights, whereas ARGs that are less likely are given high weights. Finally, the likelihood of any parameter values can be approximated as a weighted average of the ARG probabilities for the sampled ARGs. This approximation becomes exact as more and more samples of ARGs are used.

The alternative approach to importance sampling is to incorporate parameter estimation directly into the MCMC algorithms used to sample ARGs. For example, in addition to proposing changes in ARG topologies and branch lengths, an ARG sampling algorithm could also propose changes in historical population size or mutation rate. There has been previous work on inferring ARGs under user-defined demographic models[40], and Relate[35] has the capability to alternate between the estimation of coalescence times and the estimation of historical population sizes, but a method that can jointly sample ARGs and parameters is yet to be created. Frameworks using this concept of joint sampling have been developed[19,85–88], but they have focused on small sample sizes or small numbers of total recombination events. This

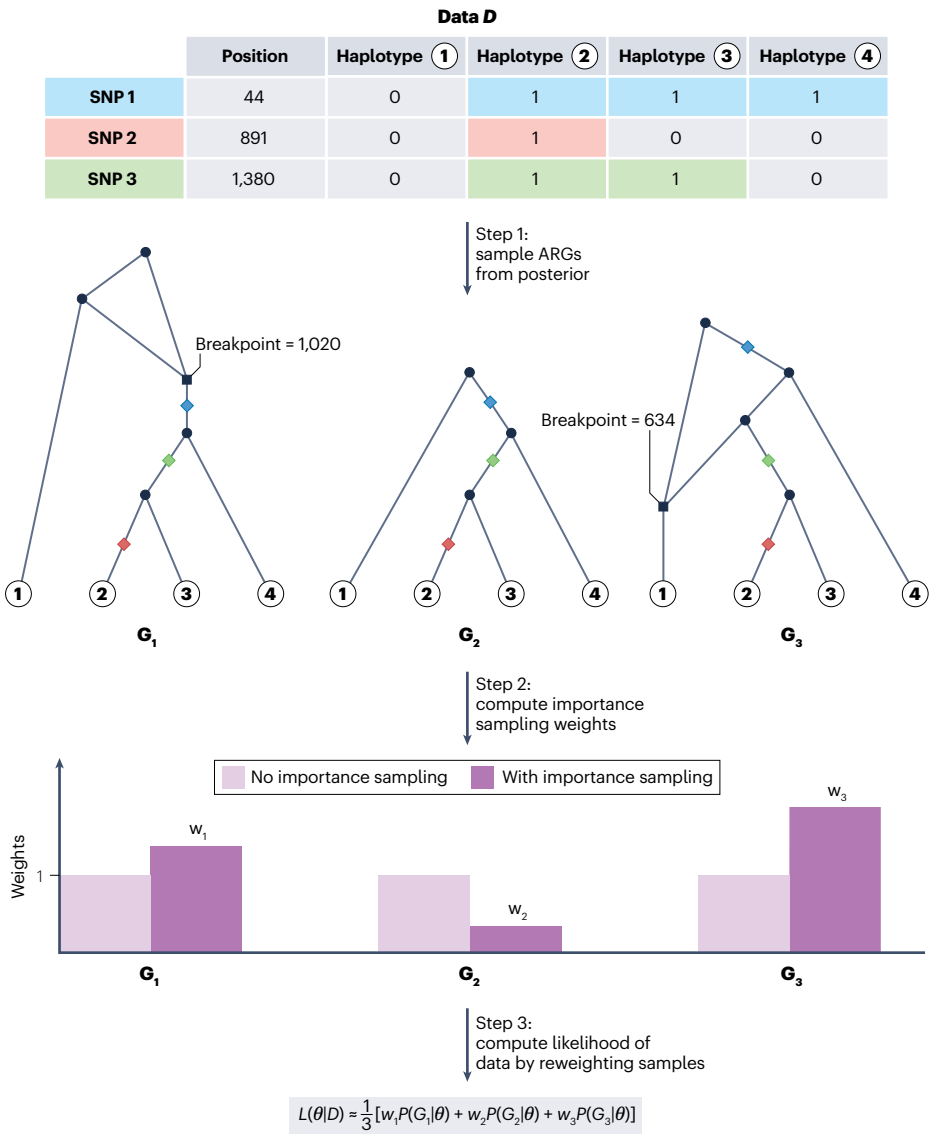

Fig. 4 | Estimating the likelihood function of population genetic parameters when using importance sampling with ARG inference methods. In step 1, many ancestral recombination graphs (ARGs) are sampled under one model from a certain posterior. For simplicity of presentation, only three sampled graphs are shown. As in Fig. 3, mutations are shown as coloured diamonds on the ARG, squares represent recombination events and circles represent coalescence events. In step 2, the importance sampling weight of each graph is computed. In step 3, we compute the likelihood of a parameter as the weighted average of the likelihood of each sampled graph. Here, D represents the sequence data, (G_1–G_3) represent different sampled ARGs and (w_1–w_3) represent the corresponding importance sampling weights for these ARGs. θ represents any population genetic parameter of interest, such as population size or selection coefficient.

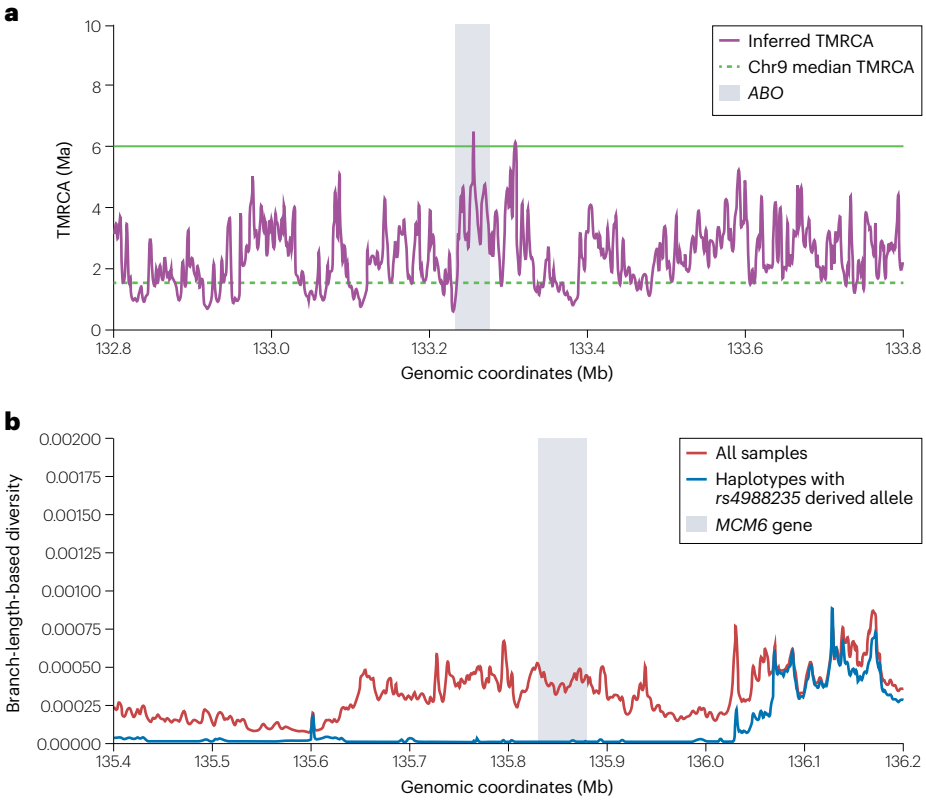

Fig. 5 | Using inferred ARGs to learn about the balancing selection at *ABO* and the selective sweep at *MCM6*. **a**, The purple lines show the inferred 1 kb time to most recent common ancestor (TMRCA) for 50 sequences from the CEU population from the 1000 Genomes Project[54,95], and the human–chimpanzee speciation time (around 6 Myr ago (Ma)) is shown as a green solid line. The *ABO* gene (shaded area) exhibits coalescence times older than the speciation time. **b**, The ancestral recombination graph (ARG)-inferred 1 kb branch-length-based diversity among all samples (red) versus that of carriers of the derived allele of *rs4988235* (blue) in the British (GBR) population from the 1000 Genomes Project[54,95]. The inferred ARG is available at: https://github.com/YunDeng98/ARG_review/tree/main/inferred_ts.

is likely due to the dramatically increased computational cost of doing this joint sampling in addition to the necessity of incorporating every parameter of interest into the ARG inference algorithm. By contrast, importance sampling can be carried out without changing the basic inference algorithm and, therefore, might provide more flexibility to infer a wider set of parameters. For example, one can imagine using ARGs to examine other processes, such as background selection, assortative mating, inference of admixture graphs. We therefore consider the use of importance sampling of ARGs for full probabilistic inference of population genetic parameters to be an important line of future research for the population genetics community. Other important future directions in this area include the development of fast, properly calibrated ARG-inference methods and finding efficient ways to compute the probability of an ARG or coalescence tree under a given set of population genetic parameters.

Illustration on human data

To illustrate the utility of inferred ARGs in understanding population genetic processes we provide three examples. We first applied SINGER to human sequencing data around the *ABO* and *MCM6* genes. Both are previously reported targets of natural selection[89–92], but here we show how to explore the selection signals using inferred ARGs.

The *ABO* gene, which determines an individual's blood type, is one of the most variable coding regions outside of the HLA region[89]. Comparative genomics analysis has shown evidence of trans-species polymorphism at this locus, and it has been hypothesized that balancing selection is maintaining its polymorphism in humans and other primates[89,93,94]. If this hypothesis is true, then we would expect to see unusually ancient times to the most recent common ancestor (TMRCA) in the *ABO* gene. Plotting the estimated 1 kb TMRCA around the *ABO* locus in the CEU population (Utah residents with northern and western European ancestries) from the 1000 Genomes Project[54,95], compared with the chromosome 9 median of 1.62 Myr ago (Ma), shows that an inferred TMRCA within this region occurred more than 6 Ma (roughly the speciation time between humans and chimpanzees)[96] (Fig. 5a). These results are compatible with the hypothesis of balancing selection maintaining trans-species polymorphism at this locus. Of note, as SINGER uses a standard coalescence prior, it likely underestimates the most ancient coalescence times, meaning the true TMRCA at this locus may be older than suggested here.

The derived allele at the *rs4988235* SNP in the *MCM6* gene has been identified as a causal variant for lactase persistence in Europeans and is believed to have been positively selected after the introduction of dairy farming[90–92]. Comparing the overall 1 kb branch-length-based diversity[45] in all British (GBR) samples from the 1000 Genomes Project[54,95] with the diversity in the haplotypes carrying the derived allele of *rs4988235* reveals that the within-carrier diversity is substantially lower than the overall diversity, and the depletion of diversity between carriers spans a very long region (Fig. 5b). This observation is consistent with a strong, recent selective sweep, as previously proposed[90,91].

We also demonstrate the ability of ARGs to infer the demographic history of different human populations (Fig. 6). We analysed an ARG inferred by tsinfer + tsdate on a large set of modern and ancient genomes[38,39]. We considered four of the populations on which this ARG was inferred: the Han Chinese (CHB), British (GBR) and Yoruba (YRI) populations from the 1000 Genomes Project[54], and the Quechua

population, an Indigenous population of South America, from the Simons Genome Diversity Project[97]. Calculating the genome-wide distribution of within-population pairwise TMRCAs for each population revealed that the Yoruba population shows the highest density of older coalescent times, as the ancestral lineages of African individuals were not subjected to the Out of Africa bottleneck that affected all non-African populations. The Han Chinese and British populations, whose ancestral lineages were affected by this bottleneck, exhibit a reduction in the number of very old pairwise TMRCAs relative to the Yoruba and a corresponding increase in the number of more recent coalescences (Fig. 6a). The Quechua population shows an even more extreme skew towards recent TMRCAs (Fig. 6a), resulting from Quechua lineages being affected by both the Out of Africa bottleneck and the bottleneck that occurred when a population of humans migrated across the Bering Land Bridge during the Last Glacial Maximum to people the Americas[98].

We also ran Relate on chromosome 1 for a sample of four populations from the 1000 Genomes Project[54]: Han Chinese, British, Yoruba and Finnish (FIN). Using the *EstimatePopulationSize* functionality, which infers historical population sizes in different epochs based on the density of inferred coalescence events, we reconstructed the demographic history of each population. We observe that at very ancient times, all populations seem to have identical population sizes, which is what would be expected if these populations had not yet diverged but were instead all part of the same ancestral population (Fig. 6b). In the period 10,000–200,000 years ago, we observe a significant decrease in the population sizes of all non-African populations relative to the Yoruba population, coinciding with the Out of Africa bottleneck. In very recent time periods, we observe a decrease in the Finnish population size relative to the British population size, which corresponds to the known small founding population of modern-day Finland when compared with other European populations[99].

Conclusions

Full probabilistic population genetic inferences on genomic data might have seemed impossible a few years ago. However, new computational methods have facilitated the use of ARGs to address several problems

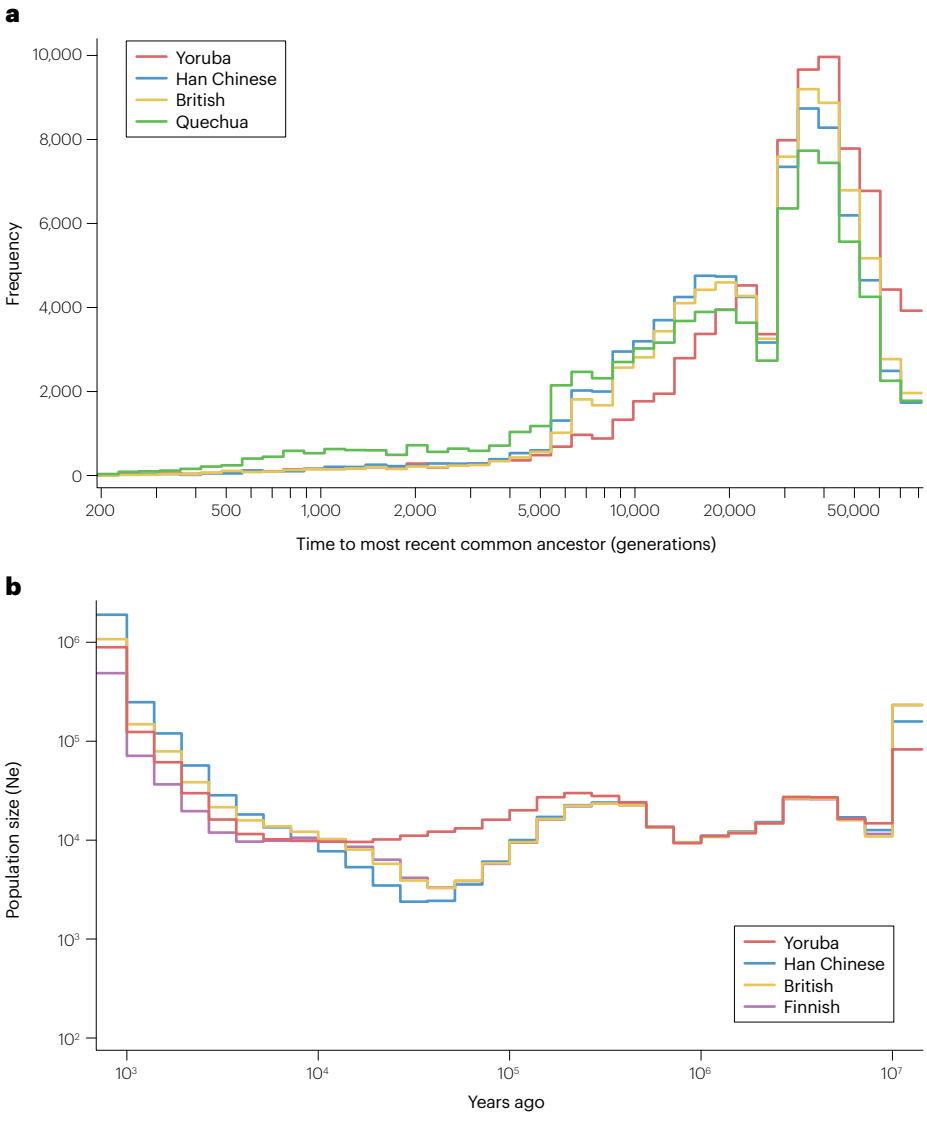

Fig. 6 | Using ARGs to reconstruct human demographic history. a, Genome-wide distribution of within-population pairwise times to recent common ancestors (TMRCA) in four populations: the Yoruba, Han Chinese and British populations from the 1000 Genomes Project[54] and the Quechua population from the Simons Genome Diversity Project[97]. For each population, three individuals were considered, and the pairwise TMRCA was calculated for each pair of lineages among these individuals at 100 kb intervals across the genome. This is inspired by ref. 39, which compares TMRCAs for African and non-African populations for the same ancestral recombination graph (ARG). The inferred ARG is available at https://zenodo.org/records/5495535. **b**, Inferred historical effective population sizes of the Yoruba, Han Chinese, British and Finnish populations from the 1000 Genomes Project[54]. Obtained by running Relate on chromosome 1. This is inspired by ref. 35.

Glossary

Ancestral recombination graph
(ARG). A graph describing genetic relationships between individuals in a sample in the presence of recombination where nodes represent either coalescence events or recombination events and edges represent lines of descent.

Approximate Bayesian computation
A simulation-based technique to approximate a posterior distribution using summary statistics instead of using the full likelihood function.

Coalescence tree
Also known as coalescent tree. A tree that describes the genetic relationships among individuals in a sample in a single locus, where nodes represent coalescence events and edges represent lines of descent.

Coalescent theory
Also known as coalescence theory. The population genetic theory that describes the stochastic processes that generate coalescence trees and ancestral recombination graphs.

Importance sampling
A statistical technique that allows approximation of a target distribution using simulations under another approximating distribution that subsequently is corrected.

Markov chain Monte Carlo
(MCMC). A statistical simulation technique, often used in Bayesian inference, that allows the approximation of a distribution that only is known up to a constant.

Most recent common ancestor
(MRCA). The MRCA of a group of individuals is the most recently living individual that is an ancestor to all those individuals. The MRCA can be represented by a node in a coalescence tree or an ancestral recombination graph.

Selection coefficient
The selection coefficient associated with an allele is the difference in fitness between individuals who carry one copy of the allele and individuals who do not carry the allele (in an additive model).

Sequentially Markov coalescent
(SMC). A stochastic process that approximates the full coalescence process with recombination by ignoring coalescence events that occur between lineages that do not share genetic material from the same genomic segments.

in population genetics such as inferring ages of mutations, identifying fine-scaled relationships between individuals, detecting natural selection acting on traits and individual alleles. Although still in their infancy, methods using either MCMC or importance sampling on ARGs promise to provide probabilistic frameworks that can take advantage of the rich amount of information in full genomic data. Even without such methods, ARG inference provides a crucial tool for visualizing genetic variation and detailed genetic relationships. Not all methods for inferring ARGs perform equally well. There is generally a trade-off between accuracy and computational complexity. Also, methods that provide measures of statistical accuracy (for example, full Bayesian methods such as ARGweaver and SINGER) are generally much slower than methods that do not provide measures of statistical uncertainty. Currently, only methods such as ARG-Needle and tsinfer are applicable to GWAS-sized data sets. We expect numerous developments over the next few years that will improve the computational and statistical aspects of ARG inference and expand the applications of ARGs to fully move computational population genetics into the genomic era.

Code availability

The results presented in Figs. 5 and 6 can be reproduced by code in the following GitHub repository: https://github.com/YunDeng98/ARG_review.git.

Published online: 30 September 2024

References

1. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Probab.* **19**, 27–43 (1982).
This paper rigorously derives the standard coalescence process, now known as the Kingman coalescent, and shows that the stochastic process of lines of descent of a population genetic sample converges to a strictly binary tree with exponentially distributed waiting times between coalescence events.
2. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
This paper describes the CwR and the resulting genealogical structure of ARGs (although it does not use that term).
3. Fu, Y. X. & Li, W. H. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10 (1999).
4. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**, 380–390 (2002).
5. Wakeley, J. Developments in coalescent theory from single loci to chromosomes. *Theor. Popul. Biol.* **133**, 56–64 (2020).
6. Hudson, R. R. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217 (1983).
7. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
8. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
9. Beerli, P. & Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl Acad. Sci. USA* **98**, 4563–4568 (2001).
10. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
11. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The "hitchhiking effect" revisited. *Genetics* **123**, 887–899 (1989).
This paper derives coalescence models for neutral loci linked to a locus under selection.
12. Nielsen, R. et al. Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
13. Griffiths, R. C. & Tavaré, S. Ancestral inference in population genetics. *Stat. Sci.* **9**, 307–319 (1994).
14. Wilson, I. J. & Balding, D. J. Genealogical inference from microsatellite data. *Genetics* **150**, 499–510 (1998).
15. Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.* **27**, 921–933 (2010).
16. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
17. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
18. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
19. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
20. Adams, A. M. & Hudson, R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712 (2004).
21. Garrigan, D. Composite likelihood estimation of demographic parameters. *BMC Genet.* **10**, 72 (2009).
22. Nielsen, R. et al. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* **19**, 838–849 (2009).
23. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
24. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
25. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
26. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
27. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
This paper describes the first method for full probabilistic inferences of ARGs (ARGweaver).
28. Griffiths, R. C. & Marjoram, P. An ancestral recombination graph. *Inst. Math. Appl.* **87**, 257 (1997).
This paper coins the term ARG and provides a rigorous derivation of the CwR.

29. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).
30. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B* **360**, 1387–1393 (2005).
31. Marjoram, P. & Wall, J. D. Fast "coalescent" simulation. *BMC Genet.* **7**, 16 (2006).
32. Wilton, P. R., Carmi, S. & Hobolth, A. The SMC′ is a highly accurate approximation to the ancestral recombination graph. *Genetics* **200**, 343–355 (2015).
33. Wong, Y. et al. A general and efficient representation of ancestral recombination graphs. *Genetics* **228**, iyae100 (2024).
34. Minichiello, M. J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
35. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
This paper presents the popular ARG inference method Relate.
36. Mirzaei, S. & Wu, Y. RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics* **33**, 1021–1030 (2017).
37. Heine, K., Beskos, A., Jasra, A., Balding, D. & De Iorio, M. Bridging trees for posterior inference on ancestral recombination graphs. *Proc. R. Soc. A.* **474**, 20180568 (2018).
38. Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
This paper presents the popular ARG inference method tsinfer, which is applicable to biobank-scale data.
39. Wohns, A. W. et al. A unified genealogy of modern and ancient genomes. *Science* **375**, eabi8264 (2022).
40. Hubisz, M. J., Williams, A. L. & Siepel, A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* **16**, e1008895 (2020).
41. Schaefer, N. K., Shapiro, B. & Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Sci. Adv.* **7**, eabc0776 (2021).
42. Ignatieva, A., Lyngsø, R. B., Jenkins, P. A. & Hein, J. KwARG: parsimonious reconstruction of ancestral recombination graphs with recurrent mutation. *Bioinformatics* **37**, 3277–3284 (2021).
43. Mahmoudi, A., Koskela, J., Kelleher, J., Chan, Y.-B. & Balding, D. Bayesian inference of ancestral recombination graphs. *PLoS Comput. Biol.* **18**, e1009960 (2022).
44. Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F. & Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet.* **55**, 768–776 (2023).
45. Deng, Y., Nielsen, R. & Song, Y. S. Robust and accurate bayesian inference of genome-wide genealogies for large samples. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.16.585351 (2024).
46. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
47. Excoffier, L. & Foll, M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
48. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
49. Kelleher, J., Thornton, K. R., Ashander, J. & Ralph, P. L. Efficient pedigree recording for fast population genetics simulation. *PLoS Comput. Biol.* **14**, e1006581 (2018).
50. Baumdicker, F. et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).
51. Y. C. Brandt, D., Wei, X., Deng, Y., Vaughn, A. H. & Nielsen, R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics* **221**, iyac044 (2022).
52. Peng, D., Mulder, O. J. & Edge, M. D. Evaluating ARG-estimation methods in the context of estimating population-mean polygenic score histories. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.24.595829 (2024).
53. Deng, Y., Song, Y. S. & Nielsen, R. The distribution of waiting distances in ancestral recombination graphs. *Theor. Popul. Biol.* **141**, 34–43 (2021).
54. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
55. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
56. Fan, C. et al. A likelihood-based framework for demographic inference from genealogical trees. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.10.561787 (2023).
57. Pearson, A. & Durbin, R. Local ancestry inference for complex population histories. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.06.529121 (2023).
58. Irving-Pease, E. K. et al. The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**, 312–320 (2024).
59. Coop, G. & Griffiths, R. C. Ancestral inference on gene trees under selection. *Theor. Popul. Biol.* **66**, 219–232 (2004).
60. Hejase, H. A., Dukler, N. & Siepel, A. From summary statistics to gene trees: methods for inferring positive selection. *Trends Genet.* **36**, 243–258 (2020).
61. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* **15**, e1008384 (2019).
This paper demonstrates how ARGs can be used to infer selection.
62. Vaughn, A. H. & Nielsen, R. Fast and accurate estimation of selection coefficients and allele histories from ancient and modern DNA. *Mol. Biol. Evol.* **41**, msae156 (2024).
63. Hejase, H. A., Mo, Z., Campagna, L. & Siepel, A. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Mol. Biol. Evol.* **39**, msab332 (2022).
64. Mo, Z. & Siepel, A. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLoS Genet.* **19**, e1011032 (2023).
65. Stern, A. J., Speidel, L., Zaitlen, N. A. & Nielsen, R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* **108**, 219–239 (2021).
66. Edge, M. D. & Coop, G. Reconstructing the history of polygenic scores using coalescent trees. *Genetics* **211**, 235–262 (2019).
67. Osmond, M. M. & Coop, G. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.13.452277 (2021).
68. Grundler, M. C., Terhorst, J. & Bradburd, G. S. A geographic history of human genetic ancestry. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.27.586858 (2024).
69. Deraje, P., Kitchens, J., Coop, G. & Osmond, M. M. Inferring the geographic history of recombinant lineages using the full ancestral recombination graph. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.10.588900 (2024).
70. Gao, Z., Zhang, Y., Cramer, N., Przeworski, M. & Moorjani, P. Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. *eLife* **12**, e81188 (2023).
71. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).
72. Wang, R. J., Al-Saffar, S. I., Rogers, J. & Hahn, M. W. Human generation times across the past 250,000 years. *Sci. Adv.* **9**, eabm7047 (2023).
73. Ragsdale, A. P. & Thornton, K. R. Multiple sources of uncertainty confound inference of historical human generation times. *Mol. Biol. Evol.* **40**, msad160 (2023).
74. Huang, Z., Kelleher, J., Chan, Y.-B. & Balding, D. J. Estimating evolutionary and demographic parameters via ARG-derived IBD. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.07.583855 (2024).
75. Ignatieva, A., Favero, M., Koskela, J., Sant, J. & Myers, S. R. The distribution of branch duration and detection of inversions in ancestral recombination graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.11.548567 (2023).
76. Speidel, L. et al. High-resolution genomic ancestry reveals mobility in early medieval Europe. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.15.585102 (2024).
77. Tagami, D., Bisschop, G. & Kelleher, J. tstrait: a quantitative trait simulator for ancestral recombination graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.13.584790 (2024).
78. Link, V. et al. Tree-based QTL mapping with expected local genetic relatedness matrices. *Am. J. Hum. Genet.* **110**, 2077–2091 (2023).
79. Salehi Nowbandegani, P. et al. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. *Nat. Genet.* **55**, 1494–1502 (2023).
80. Tsambos, G., Kelleher, J., Ralph, P., Leslie, S. & Vukcevic, D. link-ancestors: fast simulation of local ancestry with tree sequence software. *Bioinform. Adv.* **3**, vbad163 (2023).
81. Haller, B. C. & Messer, P. W. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
82. Tokdar, S. T. & Kass, R. E. Importance sampling: a review. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 54–60 (2010).
83. Hammersley, J. M. & Morton, K. W. Poor man's Monte Carlo. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **16**, 23–38 (1954).
84. Rosenbluth, M. N. & Rosenbluth, A. W. Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.* **23**, 356–359 (1955).
85. Kuhner, M. K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768–770 (2006).
86. Wang, Y. & Rannala, B. Bayesian inference of fine-scale recombination rates using population genomic data. *Phil. Trans. R. Soc. B* **363**, 3921–3930 (2008).
87. Wang, Y. & Rannala, B. Population genomic inference of recombination rates and hotspots. *Proc. Natl Acad. Sci. USA* **106**, 6215–6219 (2009).
88. Vaughan, T. G. et al. Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* **205**, 857–870 (2017).
89. Ségurel, L. et al. The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
90. Enattah, N. S. et al. Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
91. Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
92. Chin, E. L. et al. Association of lactase persistence genotypes (rs4988235) and ethnicity with dairy intake in a healthy U.S. population. *Nutrients* **11**, 1860 (2019).
93. Fortier, A. L. & Pritchard, J. K. Ancient trans-species polymorphism at the major histocompatibility complex in primates. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.28.497781 (2022).
94. Azevedo, L., Serrano, C., Amorim, A. & Cooper, D. N. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Hum. Genomics* **9**, 21 (2015).

95. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
96. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
97. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
98. Amos, W. & Hoffman, J. I. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci.* **277**, 131–137 (2010).
99. Kittles, R. A. et al. Dual origins of Finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* **62**, 1171–1179 (1998).

Acknowledgements
The authors' research was supported by NIH grants R01GM138634 and R35 GM153400-01 to R.N., NIH grant R56-HG013117 to Y.D. and NSF Graduate Research Fellowship 2146752 to A.H.V.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Reviews Genetics* thanks David Balding; Jerome Kelleher; and Adam Siepel, who co-reviewed with Melissa J. Hubisz, Nurdan Kuru and Armin Scheben, for their contribution to the peer review of this work.

Related links
Arbores: https://github.com/heinekmp/Arbores
ARGinfer: https://github.com/alimahmoudi29/arginfer
ARG-Needle: https://palamaralab.github.io/software/argneedle/
ARGweaver: https://github.com/mdrasmus/argweaver
ARGweaver-D: https://github.com/CshlSiepelLab/argweaver-d-analysis
KwARG: https://github.com/a-ignatieva/kwarg
Relate: https://github.com/MyersGroup/relate
RentPlus: https://github.com/SajadMirzaei/RentPlus
SARGE: https://github.com/nkschaefer/sarge
SINGER: https://github.com/popgenmethods/SINGER
tsdate: https://github.com/tskit-dev/tsdate
tsinfer: https://github.com/tskit-dev/tsinfer