The length of haplotype blocks and signals of structural variation in reconstructed genealogies

Anastasia Ignatieva[1,2†], Martina Favero[3*], Jere Koskela[4,5*], Jaromir Sant[1*], and Simon R. Myers[1,6]

[1]Department of Statistics, University of Oxford
[2]School of Mathematics and Statistics, University of Glasgow
[3]Department of Mathematics, University of Stockholm
[4]Department of Statistics, University of Warwick
[5]School of Mathematics, Statistics and Physics, Newcastle University
[6]Centre for Human Genetics, University of Oxford
[*]Contributed equally, ordered alphabetically
[†]Correspondence: anastasia.ignatieva@stats.ox.ac.uk

February 4, 2025

Abstract

Recent breakthroughs have enabled the accurate inference of large-scale genealogies. Through modelling the impact of recombination on the correlation structure between genealogical local trees, we evaluate how this structure is reconstructed by leading approaches. Despite identifying pervasive biases, we show that applying a simple correction recovers the desired distributions for one algorithm, Relate. We develop a statistical test to identify clades spanning unexpectedly long genomic regions, likely reflecting regional suppression of recombination in some individuals. Our approach allows a systematic scan for inter-individual recombination rate variation at an intermediate scale, between genome-wide differences and individual hotspots. Using genealogies reconstructed with Relate for 2 504 human genomes, we identify 50 regions possessing clades with unexpectedly long genomic spans ($p < 1 \cdot 10^{-12}$). The strongest signal corresponds to a known inversion on chromosome 17. The second strongest uncovers a novel 760kb inversion on chromosome 10, common (21%) in S. Asians and correlated with GWAS hits for a range of phenotypes. Other regions indicate additional genomic rearrangements: inversions (8), copy number changes (2), or other variants (12). The remaining regions appear to reflect recombination suppression by previously unevidenced mechanisms. They are enriched for precisely spanning single genes ($p = 5 \cdot 10^{-10}$), specifically those expressed in male gametogenesis, and for eQTLs ($p = 2 \cdot 10^{-3}$). This suggests an extension of previously hypothesised crossover suppression within meiotic genes, towards a model of suppression varying across individuals with different expression levels. Our methods can be readily applied to other species, showing that genealogies offer previously untapped potential to study structural variation and other phenomena impacting evolution.

1 Introduction

In the presence of recombination, the genealogical history of a sample can be fully captured in the form of an ancestral recombination graph (ARG). This can be represented as a sequence of local trees describing the sample genealogy at each locus, connected by recombination events that reshape these trees along the genome. ARGs can, in principle, capture the effects of all the evolutionary forces that have shaped the observed genetic diversity of a sample of sequences, while providing a much more efficient representation of genomes than multiple sequence alignments (Kelleher et al., 2019). Thus, statistical inference methods which take ARGs as inputs have the potential to provide very powerful insights into evolutionary events and parameters.

However, the true underlying genealogy is usually not observable in practice, and the ARG must be reconstructed from the data, which typically comprises a set of genetic sequences sampled at the present time. This is a notoriously difficult problem, due to the computational cost of traversing the huge search space of plausible ARGs. This has been the main bottleneck for the widespread development of genealogy-based inference, but has seen impressive recent breakthroughs,

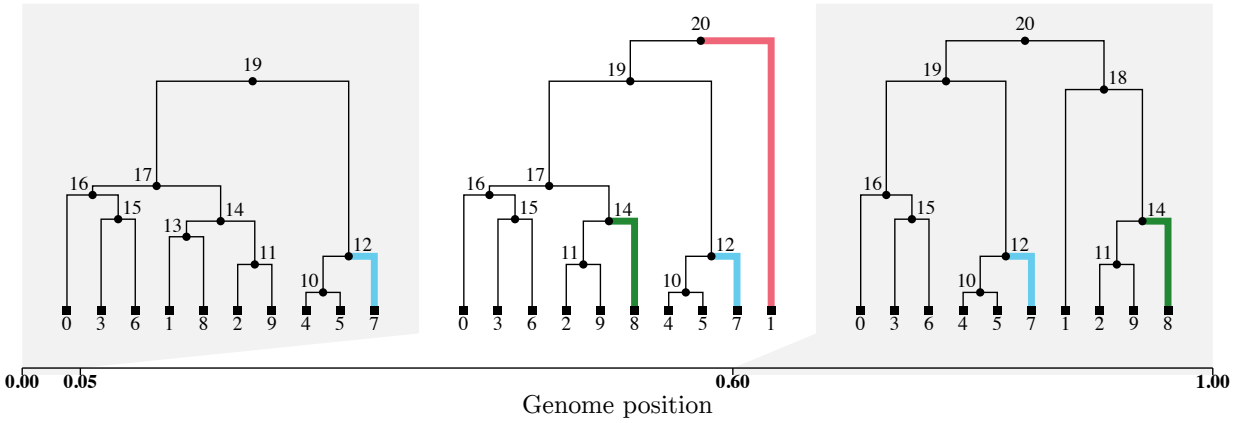

Figure 1: An ARG with $n = 10$ sequences, represented as a sequence of the corresponding local trees along the genome. The *time-length* of an edge is given by the time of its parent node less the time of its child node; *older* edges are those closer to the root of a local tree. Edge highlighted in red ($1 \to 20$) spans one local tree and has span 0.55; edge highlighted in green ($8 \to 14$) spans two local trees and has span 0.95; edge highlighted in blue ($7 \to 12$) spans all three local trees and has span 1. ARG simulated using msprime and visualised using tskit (Baumdicker et al., 2022; Kelleher et al., 2016).

with methods now capable of reconstructing and efficiently storing ARGs for tens or even hundreds of thousands of samples (Speidel et al., 2019; Wohns et al., 2022; Zhang et al., 2023; Zhan et al., 2023). Inference using ARGs reconstructed from large-scale human sequencing datasets is in its infancy, but has already produced novel scientific insights, for instance in untangling the history of human demography (Speidel et al., 2019; Wohns et al., 2022) and understanding the phenotypic effects of genetic variants (Zhang et al., 2023).

This is a rapidly developing field, however progress has been hampered by the fact that methods which work very well on simulated ARGs often lose power and accuracy when applied to reconstructed genealogies, for reasons that are, in general, poorly understood. Methods for quantifying the quality of ARG reconstruction have generally been limited to comparing simulated to reconstructed ARGs, to quantify how well local tree topology (Speidel et al., 2019; Kelleher et al., 2019) and pairwise coalescence times (Brandt et al., 2022) are recovered. Deng et al. (2021) and McKenzie and Eaton (2022) took the approach of deriving the distribution of the genomic distance between consecutive local trees (under a given model), and comparing this to the empirical distributions calculated from reconstructed ARGs. All of these studies have broadly demonstrated that different tools have somewhat different strengths, but since they commonly output strikingly different ARGs for the same dataset (Wong et al., 2024), more in-depth exploration is needed to understand (and correct) the underlying causes.

Moreover, all currently available methods of recording and inferring ARGs typically only account for mutations in the form of single base substitutions, and ignore the presence of genomic structural variants (SVs), such as duplications and inversions. SVs are a ubiquitous and evolutionarily important type of mutation, playing a key role in speciation and local adaptation (Kirkpatrick and Barton, 2006), and within human populations through altering the structure and expression of genes (Chiang et al., 2017; Abel et al., 2020). Thus, identifying and analysing the evolution of structural variants at all scales is an incredibly important goal, but so far no studies have attempted to leverage ARGs for this purpose, beyond simulation (e.g. Peischl et al., 2013).

Within an ARG, each edge has a well-defined genomic span during which it is present in the local trees (Figure 1). We analytically derive the theoretical distribution of this quantity using the well-known SMC' model (Marjoram and Wall, 2006), which is an excellent approximation to the coalescent with recombination (Griffiths and Marjoram, 1997). Through calculating the empirical distribution of edge span in ARGs reconstructed by ARGweaver, Relate, tsinfer and tsdate, and ARG-Needle in simulation studies, we find that these tools recover the correct distribution to varying

degrees of success.

We then derive the distribution of the length of a haplotype block within an ARG, defined as the genomic span of a given clade of samples (as by Shipilina et al., 2023). Recombination is suppressed in individuals heterozygous for an inversion, which manifests in the ARG as a clade of samples that persists for a longer stretch of the genome than would otherwise be expected: we use this idea to construct a computational tool for detecting localised (between-clade) suppression of recombination (DoLoReS: Detection of Localised Recombination Suppression). This tool is tailored for use with Relate ARGs, implementing suitable adjustments to correct for possible sequencing and phasing errors in the data, and the particularities of the ARG reconstruction method. We demonstrate the power and accuracy of this tool for both simulated and reconstructed ARGs.

Finally, we apply DoLoReS to an ARG for the 1000 Genomes Project (1KGP; 1000 Genomes Project Consortium, 2015) reconstructed using Relate (Speidel et al., 2019). We detect several known inversions: for instance, one of the top significant hits is the 17q21.31 inversion polymorphism common in European populations (Stefansson et al., 2005). The second strongest signal corresponds to a previously unknown 760kb inversion on 10q22.3, which we validate using data from the Human Pangenome Reference Consortium (HPRC; Liao et al., 2023); this inversion is common in S. Asian populations (with a frequency of 21%), spans a number of genes associated with lung function, and correlates with a number of GWAS hits for hematological and immunological traits. We find several other new SVs, and show that our method also detects distinguishable signals of other structural variants, particularly copy number variants (CNVs) and complex rearrangements. This demonstrates that, while Relate only uses SNP data and does not explicitly infer or account for SVs, the reconstructed ARGs still capture the signal of SV presence.

Code implementing the methods and used to produce the figures is publicly available at `github.com/a-ignatieva/dolores`.

2 Results

2.1 The probability that an edge is broken up by the next recombination event along the genome is biased in reconstructed ARGs, particularly for old edges

For a given edge of the ARG, its span can be defined as the genomic positions where it is present in the corresponding local trees (as illustrated in Figure 1); this is determined by recombination events that change local tree topologies and coalescence times. Intuitively, the longer (resp. shorter) the *time-length* of an edge, the more (resp. less) likely it is to be broken up by recombination as we move along the genome, so its *span* along the genome should be shorter (resp. longer). This is a fundamental property of the ARG which has several implications. For instance, in the absence of recombination, the genealogy is a single tree with all edge spans equal to the length of the genome, and the probability that a given edge carries at least one mutation is proportional to its time-length (assuming mutations occur as a Poisson process along the edges). On the other hand, in the presence of recombination, the mutation rate per edge is less variable, because of the interplay between edge time-length and span.

The SMC' is an approximation to the coalescent with recombination (CwR) model characterising the distribution over sample genealogies. Under this model, an ARG can be simulated by starting with an initial binary tree $\mathcal{T}_1$, representing the sample genealogy at the leftmost point of the chromosome. The distance until the next recombination breakpoint along the genome is then drawn from an exponential distribution, with rate $\mathcal{L}_{\mathcal{T}_1} \cdot \rho/2$, where ρ is the population-scaled recombination rate, and $\mathcal{L}_{\mathcal{T}_1}$ is the total branch length of $\mathcal{T}_1$. Then a recombination point is chosen uniformly at random along the edges of $\mathcal{T}_1$, the subtree underneath this point is allowed to coalesce at a new location (with the usual coalescent dynamics), and the result is the next local tree $\mathcal{T}_2$. This process is repeated until the end of the chromosome is reached (see SI, Sections S1.1 and S1.2, for full details).

We first calculate analytically, under the SMC', the probability $\mathbb{P}_{\mathcal{T}}(b \text{ disrupted})$ that a given edge b in a local tree $\mathcal{T}$ is disrupted by the next recombination event along the genome (Methods,

Section 4.1). This probability does not have a simple form, but can be calculated exactly for a given edge and local tree. The left panel of Figure 2 shows $\mathbb{P}_{\mathcal{T}}(b \text{ disrupted})$ calculated for each edge in an ARG simulated under the SMC', against the normalised age of the edge. Middle and right panels show the same quantities, but calculated for each edge of an ARG reconstructed from the simulated data: using Relate (middle), and tsinfer/tsdate (right).

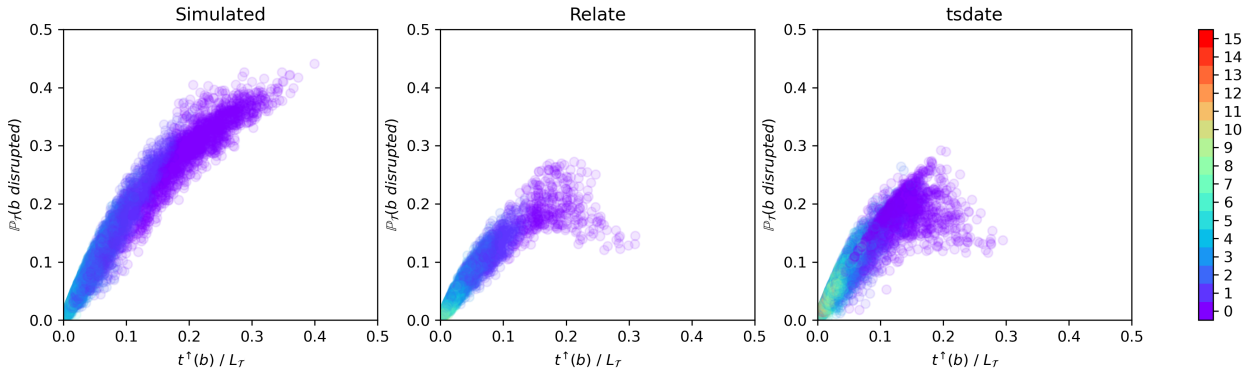

Figure 2: Value of $\mathbb{P}_{\mathcal{T}}(b \text{ disrupted})$ calculated for each edge b, against the time at its upper endpoint $t^{\uparrow}(b)$ divided by the total branch length of the tree, $\mathcal{L}_{\mathcal{T}}$. Left panel: for a uniform random sample of 10 000 edges in one simulated ARG with $n = 100$ samples (dataset 1 parameters given in Methods, Section 4.6.1). Middle and right: same quantity calculated for each edge of the ARG reconstructed from the simulated data using Relate (middle) and tsinfer/tsdate (right). Colour shows number of edges between the top end of the edge and the root of $\mathcal{T}$ (purple dots correspond to edges extending from the MRCA node).

For the simulated ARG, the probability that a given edge is disrupted by a recombination event is higher for older edges: this is as expected, as edges at the top of a local tree tend to have greater time-length and co-exist with fewer other lineages, so are more likely to be quickly broken up by recombination. However, old edges are generally difficult to accurately reconstruct using the sequences at the leaves due to lack of signal, unless they are strongly supported by mutations (which implies longer edge span along the genome). As a result, reconstructed ARGs have fewer old edges (with fewer points lying towards the right of the plots), and the old edges that *are* present tend to have lower probability of disruption than those of a similar age in the simulated ARG. This bias is seen for both Relate and tsinfer/tsdate. This demonstrates explicitly the difficulty with faithfully reconstructing the ARG topology and event times in the deep past.

2.2 The distribution of edge span is recovered with varying accuracy by different ARG reconstruction methods

We next derive an approximation to the distribution of edge span along the genome. If an edge b first appears at a position of the genome where the corresponding local tree is $\mathcal{T}$, we show that the waiting distance along the genome until is it broken up by a recombination event is approximately exponentially distributed as

$$\mathrm{Exp}\left(\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}} \cdot \rho/2\right), \tag{2.1}$$

we also derive an equivalent approximation for the case where the recombination rate varies along the genome (Methods, Section 4.2). Given a simulated or reconstructed ARG, we can thus calculate a corresponding p-value for each edge under this model (assuming edges have independent exponentially distributed spans), and check if the p-values follow the expected (uniform) distribution using a Q-Q plot (SI, Section S1.8).

We apply this to ARGs reconstructed from simulated data using Relate, tsinfer/tsdate, ARG-Needle, and ARGweaver. The resulting Q-Q plots are shown in Figure 3 (the corresponding histograms are also shown in SI, Figure S10). For ARGs reconstructed using Relate and tsinfer/tsdate, we adjust for the fact that these tools do not detect recombination events that affect only edge lengths (SI, Section S1.8.1). Moreover, we adjust for the fact that Relate does not attempt to infer

edge span when the edge is not supported by at least one mutation (SI, Section S1.8.2), in which case the local topology is resampled from one local tree to the next (unlike tsinfer, which more directly captures the span of edges in the reconstructed ARG based on shared ancestry).

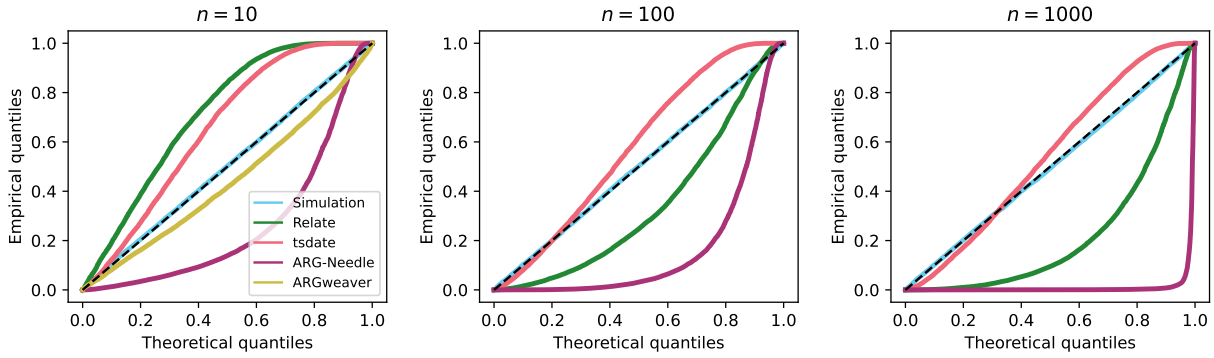

Figure 3: Q-Q plots for ARGs simulated with parameters of dataset 1 (dataset 2 for ARGweaver) given in Methods, Section 4.6.1, with $n = 10$ (left panel), $n = 100$ (middle panel), $n = 1\,000$ (right panel). Dashed line: diagonal from $(0, 0)$ to $(1, 1)$. ARGweaver results only shown for $n = 10$ due to excessive runtimes for larger sample sizes. Calculated for a random sample of 10 000 edges for each ARG. An overabundance of edges with short (resp. long) spans would drag the points below (resp. above) the diagonal.

ARGweaver accurately captures the edge span distribution, with the small deviation from the diagonal due to the difference between the SMC and SMC' models (SI, Section S1.12), although these results were only calculated for $n = 10$ due to excessive runtimes for larger sample sizes. Note that while it is possible to use the SMC' model with the latest version of ARGweaver, we found that the resulting ARGs contained cycles, which prevented us from calculating the required probabilities.

ARGs reconstructed by tsinfer/tsdate consistently have an excess of edges with long spans, while for Relate this depends on the sample size; both tools however produce skewed distributions of edge span. This is likely to be, in part, due to the waiting distances between trees being generally skewed, as shown by Deng et al. (2021). Moreover, the ARGs produced by tsinfer contain polytomies (nodes with more than two children), which is also not accounted for under the SMC' model. This results in the total branch length of a reconstructed local tree to often be greater than it would be if the polytomies were resolved to make the tree binary. In addition, some recombination events that would disrupt the edge under the SMC' (such as recombination points located on child edges), may not do so when the tree contains polytomies.

Relate first reconstructs the sequence of (correlated) local trees along the entire genome and then calculates edge spans, using an argument based on the similarity of clades subtended by the edge in successive trees. Thus, edge spans are calculated approximately, rather than being explicitly inferred, which we do not account for.

The threading procedure used by ARG-Needle to reconstruct the ARG uses the ASMC model (Palamara et al., 2018) to estimate coalescence times between the added sequence and the closely-related samples already in the ARG, with the lowest possible resulting coalescence time dictating which edge the sequence is threaded to and over what genomic length. This procedure (which relies on a combination of maximum a posteriori and posterior mean estimates for the coalescence times) is optimised for metrics having a direct effect on downstream analyses, and appears to lead to ARG-Needle consistently underestimating edge spans. We note that this is somewhat affected by the choice of time discretisation used within the ASMC (particularly for small sample sizes), but our overall findings do not change significantly when toggling this parameter.

Considering the unconditional distributions of the edge spans in the simulated and reconstructed ARGs (i.e. calculating the observed span of each edge and plotting the overall histogram) results in similar conclusions (SI, Figure S12). Further, histograms of the expected number of mutations per edge (being the product of the observed edge span, observed time-length, and the mutation rate) demonstrate large deviations between the distributions for simulated and reconstructed ARGs (SI,

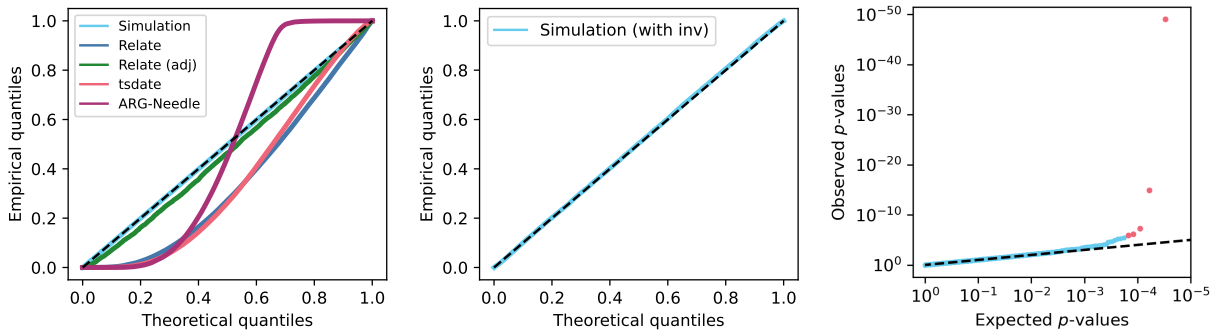

Figure 4: Left: Q-Q plot for clade p-values in simulated and reconstructed ARGs with $n = 100$ and parameters as for dataset 1 given in Section 4.6.1. An overabundance of clades with long spans would drag the points below the diagonal. Middle and right: Q-Q plot and p-value plot for ARG simulated using SLiM with one inversion under balancing selection (red points correspond to clades with p-value below the Bonferroni-corrected significance threshold).

Figure S13).

2.3 The distribution of the length of a haplotype block is recovered well by Relate after a suitable correction

A clade G of an ARG can be defined through the set of samples it contains, by writing $G = (g_1, g_2, \ldots, g_n)$, where $g_i = 1$ if sample i is in G, and 0 otherwise. The genomic span of G is defined as the interval $[a, b]$, where a is the leftmost position at which the corresponding local tree has a branch subtending exactly the samples in G, and b is the rightmost such position. For instance, in Figure 1, the clade $G = (1, 0, 1, 1, 0, 0, 1, 0, 1, 1)$ (containing samples 0, 2, 3, 6, 8, 9) has a genomic span of 0.55, since G first appears at position 0.05 and is broken up by the following recombination event at position 0.60.

Using the SMC' model, we derive the distribution of the genomic span of a given clade G, conditional on the local tree $\mathcal{T}$ in which it first appears (Methods, Section 4.3). This is approximately exponentially distributed as

$$\mathrm{Exp}\left(\mathbb{P}_{\mathcal{T}}(G \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}} \cdot \rho/2\right),$$

where $\mathbb{P}_{\mathcal{T}}(G \text{ disrupted})$ is the probability that G either gains or loses at least one sample following the next recombination event along the genome. We also derive an equivalent result when the recombination rate varies along the genome. Since our definition of a clade is equivalent to that of a haplotype block (Shipilina et al., 2023), this distribution is that of the length of a haplotype block. It is important to emphasise that since we condition on G and the local tree in which it first appears, this distribution is conditional on the frequency and age of G (as well as the local tree topology and event times, and the recombination map), rather than averaged over all clades.

For a given (simulated or reconstructed) ARG, we can thus calculate a corresponding p-value for each observed clade under this model, and again check if these follow the expected uniform distribution using a Q-Q plot. Figure 4 (left panel) shows that the approximation provides an excellent fit for an ARG simulated under the SMC' (blue points). Clade spans in ARGs reconstructed using tsinfer/tsdate (red points) tend to be over-estimated in general (possibly due to the presence of polytomies), while ARG-Needle (purple points) both under- and over-estimates this quantity. Relate (dark blue points) also tends to over-estimate clade spans; however, through analysing and correcting the causes of this, we propose a correction which effectively removes this bias (green points). Firstly, a clade might only be supported by mutations intermittently along its span, and between these regions Relate does not attempt to keep the clade intact. This causes some clade spans to be too short, and we correct this by extending the calculated span of a clade, if the clade disappears and subsequently reappears within a given distance. Secondly, since it is difficult to reconstruct the endpoints of clades exactly, which may results in genomic span being overestimated,

we use the leftmost and rightmost mutations that support a given clade to calculate its genomic span. See SI, Section S1.13.1 for full details.

2.4 Simulated and reconstructed ARGs capture signals of structural variation

Chromosomal inversions are a type of structural variant, whereby as a result of recombination, the genome breaks at two points and the segment between the breakpoints is reinserted in the opposite orientation. While recombination can proceed normally in individuals homozygous for the inversion, in heterozygotes recombination is substantially suppressed in the region containing the inversion, since a crossover recombination within the region is likely to result in the production of unbalanced gametes (Kirkpatrick, 2010). Detecting inversions computationally from sequencing data typically relies on paired-end mapping (detecting reads mapping in the opposite orientation to the reference), split-read methods (detecting reads that map onto the reference with gaps), and *de novo* assembly (directly reconstructing the sequenced genome to look for differences with the reference) (Tattini et al., 2015). For instance, DELLY (Rausch et al., 2012) implements a combination of these approaches and was used to identify hundreds of inversions in 1KGP data (Sudmant et al., 2015). In general, however, such methods suffer from high false positive rates and poor sensitivity, with their performance depending on the size of the inverted region, particularly for short-read sequencing data (Lucas Lledó and Cáceres, 2013); the detection of structural variants from long-read data is challenging due to high sequencing error rates (Sedlazeck et al., 2018). Inversions can also be detected by looking for disrupted patterns of linkage disequilibrium (LD) using population data (Li and Ralph, 2019; Bansal et al., 2007), however this is sensitive to noise in the LD patterns and cannot be used to reliably detect inversions that are not large and high-frequency.

Suppose that in a given ARG, an inversion happens on an edge g which subtends a clade G. Suppression of recombination in heterozygotes implies that if a lineage within G undergoes a recombination, it will coalesce with lineages in G with high probability; likewise, if a recombination event happens on an edge not carrying the inversion, with high probability it will coalesce with edges outside G (SI, Figure S3). This implies that inversions can be detected by looking for clades that last for "too long" along the genome due to this local suppression of between-clade recombination. Note that we are imposing the simplifying assumption that recombination in heterozygotes is suppressed completely in the inverted region (so the clade G remains completely intact). In reality, recombination can occur in heterozygotes: multiple breakpoints occurring in the inverted region would enable this, however such events have relatively low probability unless the inversion region is very large or the recombination rate is very high.

Test 1: Phrasing the above as a hypothesis test, for each clade we calculate its genomic span $[a, b]$, and calculate a p-value as the probability of this clade having a span greater than $b - a$ (Methods, Section 4.3.1). Simulation studies confirm that for ARGs simulated under the SMC' model without inversions, these p-values are approximately uniformly distributed (Figure 4, left panel, blue points), as expected.

Test 2: Alternatively, we can estimate the number of recombination events R that occurred within the genomic interval $[a, b]$, and calculate a p-value as the probability that G stays intact after at least R recombination events (Methods, Section 4.3.2). This is exactly equivalent to Test 1 when the ground truth ARG and recombination map are known. For reconstructed ARGs, we apply both of these tests since they are susceptible to false positives in different practical scenarios: Test 1 is sensitive to the choice of recombination map and presence of sequencing gaps, while Test 2 can result in false positives if there is a high level of recurrent mutation (which can cause reconstruction errors).

Both of these tests are implemented in DoLoReS, which outputs calculated genomic spans and other characteristics of each clade within an input ARG (in tskit format) and the corresponding p-values.

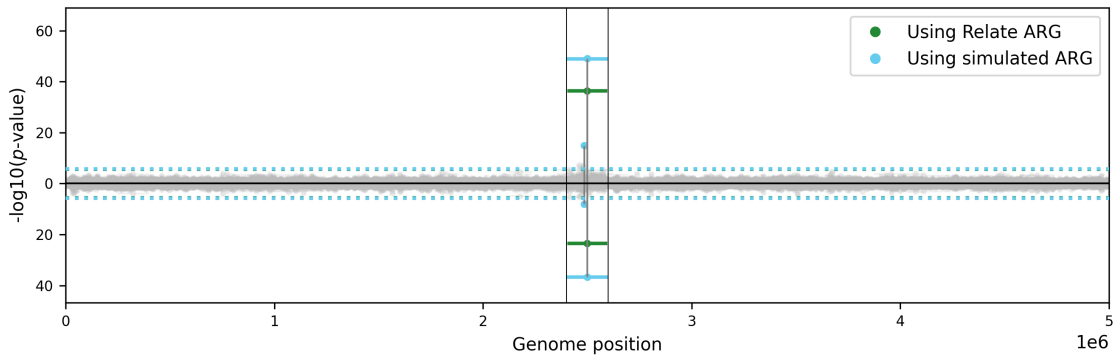

Figure 5: p-values for clades in simulated ARG and ARG reconstructed using Relate. Horizontal lines show the genomic span of each clade (with midpoint marked by circle); those with non-significant p-values are shown in grey, otherwise in blue and green for the simulated and Relate ARG, respectively. Corresponding p-values using Test 1 and Test 2 are shown on the y-axis above and below 0, respectively. Vertical black lines delineate the region of the inversion; dotted blue (resp. green) horizontal line shows Bonferroni-corrected significance threshold for calculations using the simulated (resp. Relate) ARG (note these overlap very closely).

2.4.1 Simulated ARGs

To demonstrate the power of these tests in detecting inversions, we used SLiM (Haller and Messer, 2019; Haller et al., 2019) to simulate an ARG with one 200kb inversion under balancing selection (Methods, Section 4.6.2). Figure 4 shows the corresponding Q-Q plot (middle panel) and p-values for each clade (right panel) using Test 1. The spans of most clades are unaffected by the inversion, so most of the points on the Q-Q plot adhere tightly to the diagonal. However, there are outliers in the tail (right panel, shown in red) with significant p-values (after Bonferroni correction for multiple testing). Figure 5 shows the p-values for each clade using Test 1 (above the 0 line) and Test 2 (below the 0 line). The clades with significant p-values (using a Bonferroni-corrected significance threshold of $2 \cdot 10^{-6}$) overlap the location of the inverted region, and the clade subtended by the edge carrying the inversion is a significant outlier with the lowest p-value. Repeating the simulation using SLiM with no inversion (and otherwise the same parameters), the p-values are approximately uniformly distributed and there are no clades with significant p-values (SI, Figure S14).

2.4.2 Reconstructed ARGs

We next applied DoLoReS to an ARG reconstructed using Relate for the data simulated using SLiM as described above (applying the correction described in Section 2.3). We detect one clade which has significant p-values using both tests, shown in green in Figure 5, which is exactly the clade carrying the simulated inversion. The corresponding Q-Q and p-value plots for Test 1 are shown in the top row of SI, Figure S15, showing that the Q-Q plot very close to the diagonal. The equivalent plots for a simulation with no inversion are shown in the bottom row, demonstrating that the p-values are approximately uniformly distributed and there are no false positives.

2.5 Structural variants can be detected using ARGs reconstructed from 1KGP data

We applied DoLoReS to an ARG reconstructed using Relate for the 1KGP (Phase 3) data by Speidel et al. (2019), splitting the ARG into the five super-populations to avoid confounding by population structure, accounting for varying population size through time, and applying several filters to correct for possible sequencing errors and artefacts (Methods, Section 4.4). The resulting p-values are shown in Figure 6 (and SI, Figures S16 and S17). There are a total of 125 significant clades, clustering into 50 localised regions.

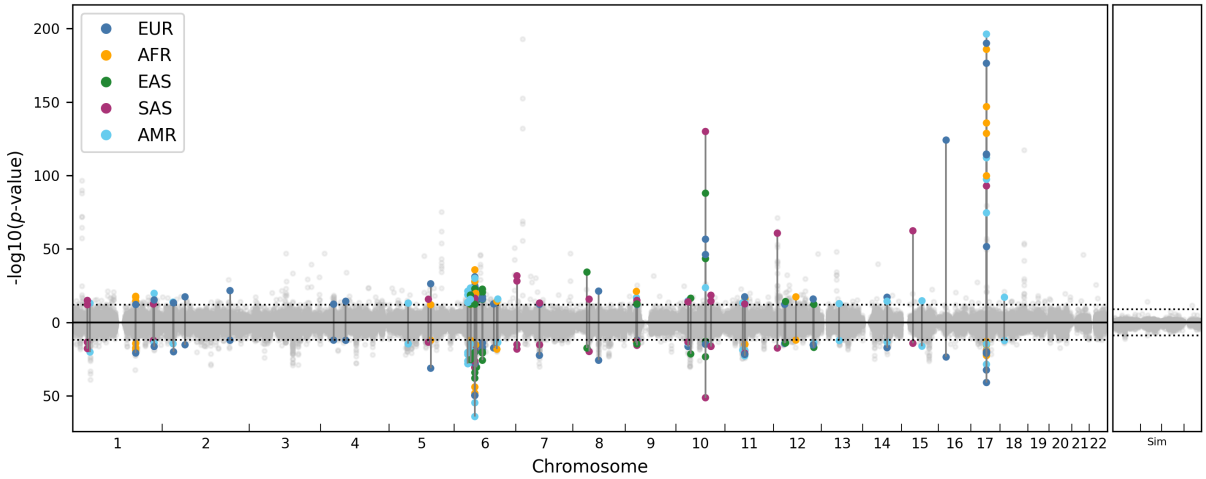

Figure 6: Left panel: p-values for clades in 1KGP ARG reconstructed using Relate. Each p-value shown as a point, with position on x-axis being the midpoint of the corresponding clade span; p-values using Test 1 and Test 2 are shown on the y-axis above and below 0, respectively. Points for clades with significant p-values shown in colour corresponding to the super-population (legend); corresponding p-values for Test 1 and Test 2 for each significant clade are connected by solid vertical lines. Dotted black horizontal lines show Bonferroni-corrected significance threshold of $1 \cdot 10^{-12}$. Right panel: results for reconstructed ARG using simulated data, dotted line shows Bonferroni-corrected significance threshold of $1 \cdot 10^{-9}$ (Methods, Section 4.6.3).

2.5.1 17q21.31 inversion

One of the detected regions with the lowest p-values, on chromosome 17, corresponds to the known 900kb inversion common in European populations, with two distinct haplotypes H1 and H2 (Stefansson et al., 2005). Since we separately test each clade within each of the population-specific ARGs, multiple clades from the same population can have significant p-values (creating the vertical stack of points seen in Figure 6). This is because if the samples are perfectly split into two clades A and B (of carriers and non-carriers of the inversion, respectively), both will show detectable signal of between-clade recombination suppression. Moreover, sub-clades of A or B can also have longer genomic spans due to the effects of locally suppressed recombination.

We detect strong signal of this inversion in all populations apart from E. Asian, estimating its average frequency at approximately 24% in European, 15% in American, 6% in S. Asian, and 2% in African populations, which aligns well with prior estimates (Donnelly et al., 2010). The ARG also allows for the estimation of the time of the inversion, through identifying the predicted clade carrying the inversion in each local tree, and extracting the time at the lower and upper end of the branch subtending this clade (Figure 7); this gives an average age of between 8 090 and 122 688 generations. This aligns with the estimates of 3m years of Stefansson et al. (2005) and 2.3m years of Steinberg et al. (2012), but highlights a large amount of uncertainty; the inversion has also been estimated to be much younger at around 100k years by Donnelly et al. (2010). We detect a region where the inversion appears relatively much more recent (highlighted in red on Figure 7), overlapping the 5'UTR region of the *CRHR1* gene (SI, Figure S18). This is the same as the region of very low sequence divergence between the H1 and H2 haplotypes found by Steinberg et al. (2012). The topology of the ARG, with H1 and H2 forming two disjoint clades with a very ancient MRCA time (SI, Figure S18K), is consistent with the inversion being very old. From the ARG we infer a change of local tree topology and H1/H2 MRCA time within the highlighted region in Figure 7, which is suggestive of a historic double crossover event between the two haplotypes (with the two breakpoints separated by around 40kb), as posited by Steinberg et al. (2012) (Methods, Section 4.5.3).

We also identify signal of a CNV at around 44.3Mb (SI, Figure S18G), through identifying instances where for a large number of individuals (between 6 and 19 depending on genomic position), their chromosomes perfectly segregate into the same two clades (Methods, Section 4.5.2). The clades

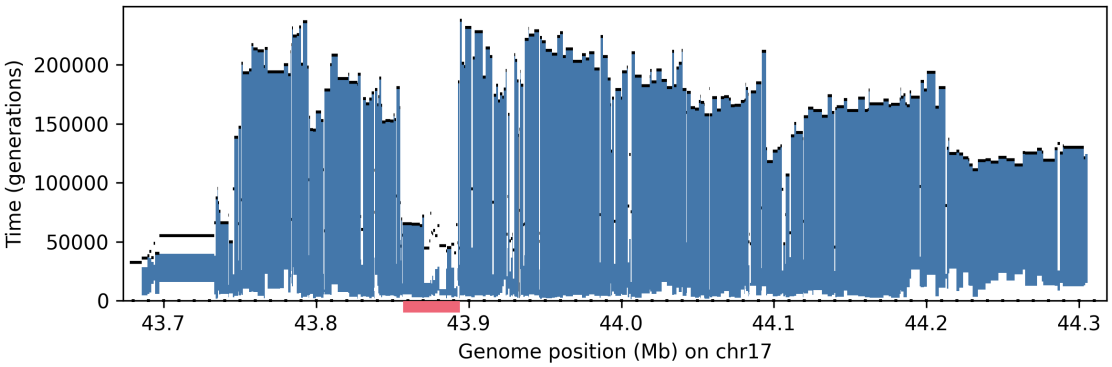

Figure 7: Age estimate for 17q21.31 inversion (the H2 haplotype) using the 1KGP ARG subsetted to European populations. Vertical lines at each genomic position are drawn between the time at the lower and upper end of the branch in the corresponding local tree which subtends the predicted carriers of the inversion. Black lines show MRCA time in full (all-population) ARG. Time is measured in generations, genome positions given in GRCh37 coordinates. Red bar highlights region where inversion appears much more recent.

we identify correspond to individuals who are homozygous for three copies of a known 25kb CNV at this position (which is in LD with the inversion). This demonstrates that CNVs are also detectable from reconstructed ARGs, through the signal they leave in the data that causes errors in ARG topology reconstruction.

Some significant regions correspond to other known inversions, including on 4q13.2 (Korbel et al., 2007), 11p11.12 (Porubsky et al., 2022), and a possible pericentromeric inversion on chromosome 6 (Martínez-Fundichely et al., 2014).

2.5.2 16p12.2 complex structural polymorphism

The significant clade on chromosome 16 corresponds to a known 1.1Mb structural polymorphism, which was posited to be an inversion in a number of studies (Tuzun et al., 2005; Bansal et al., 2007). This was subsequently shown to be the result of mis-assembly of the reference genome due to complex structural variation in this region, with two common haplotypes that differ by a large (333kb) segmental duplication (Antonacci et al., 2010). Our method thus captures signal of local recombination suppression in individuals heterozygous for this polymorphism.

Other significant hits include known regions of complex structural variation on 11q11 (Korbel et al., 2007) and 15q13.3 (Antonacci et al., 2010).

2.5.3 Structural variation on chromosome 6

There are a total of 43 significant clades on chromosome 6, clustering into 10 distinct regions (SI, Figure S16). Details of the top significant hit on 6p11.2 are shown in SI, Figure S21. The detected clades have a genomic span covering around 340kb, but (since we merge together the genomic span of a clade if it disappears and reappears within a given distance) there are relatively large stretches where the clades are not in fact present in the local trees (compare trees 2 and 4 to tree 3 in SI, Figure S21K). This appears to be due to the presence of multiple CNVs, which are not in LD with the detected clades, but are likely to cause phasing errors in the data, leading to artefacts in the reconstructed ARG. Poor reconstruction in these regions is also indicated by the relatively high number of mutations not uniquely mapping to a branch in the local trees in these regions (Methods, Section 4.5.1). This highlights that while SVs are ignored by the reconstruction method, their presence significantly distorts the resulting genealogies.

2.5.4 10q22.3 inversion

We detect a 760kb region of locally suppressed recombination on chromosome 10; details of the genomic spans of the significant clades are shown in Figure 9. This indicates an inversion with an

average frequency of approximately 9% (21% in S. Asian, 15% in American, 7% in European and 2% in E. Asian populations), with an age of between 2 625 and 13 392 generations (based on the lower and upper time of the edge subtending the inversion, averaged across its span). Analysis of the genealogies in this region indicates that the non-inverted orientation is ancestral.

The presence of segmental duplications (blocks of DNA larger than 1kb and with > 90% sequence similarity, occurring multiple times along the genome) can enable non-allelic homologous recombination (NAHR), a potential mechanism through which large structural polymorphisms arise (Lupski and Stankiewicz, 2005). NAHR between two segmental duplications which appear in opposite orientations can, specifically, lead to an inversion of the sequence that they flank. The endpoints of the significant clades (Figure 9A) align exactly with the positions of inverted segmental duplications of length 50kb (Figure 9F), supporting the possibility of an inversion in this region. We confirmed that for predicted carriers of the inversion there is an enrichment of discordantly mapped paired-end reads between these breakpoints.

To further validate this finding, we used complete long-read sequencing data from 47 individuals generated by the HPRC (Liao et al., 2023) and the T2T-CHM13 reference (Nurk et al., 2022). The HPRC sequenced a sample of children in parent-child 1KGP trios, whereas the ARG includes the parents; we thus used tagging SNPs to determine the predicted status of each HPRC sequence. We identify five sequences carrying an inversion within the predicted region, corresponding to one homozygous (HG01258) and three heterozygous (HG01123, HG01978, HG02257) individuals (Figure 8 and SI, Figure S19).

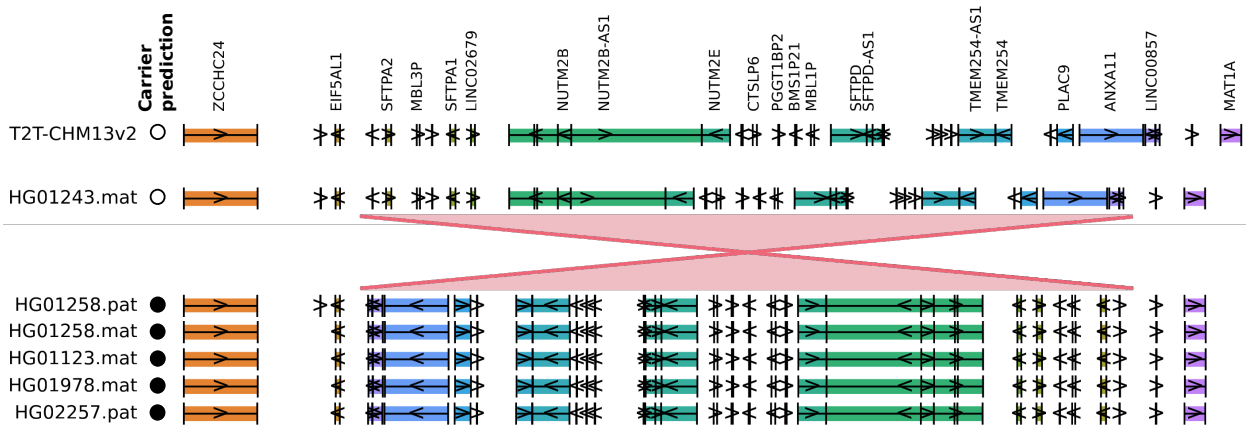

Figure 8: Validation of 10q22.3 inversion using HPRC data for 47 diploid individuals and the T2T-CHM13 reference. Five sequences (out of 95) display an inversion in the predicted region (shown in red), as indicated by a reversal of gene ordering compared to non-carriers (HG01243 chosen as a representative example; see SI, Figure S19 for a comparison against all sequences). Circles show predicted status of carrying the inversion (white = non-carrier, filled = carrier).

There are two relatively large CNVs within the span of the inversion: CNV1 at 81 474 561bp (30kb with 0 or 2 copies, overall allele frequency = 0.47) and CNV2 at 81 505 304bp (100kb with 0 or 2 copies, overall allele frequency = 0.02). CNV2 is not in LD with the inversion, but appears to distort the reconstructed genealogy around this region similarly to that on 6p11.2 (resulting in regions where the detected significant clades are broken up temporarily within the ARG, Figure 9A). However, the inverted haplotype has a deletion of CNV1, with all individuals homozygous for the inversion also homozygous for 0 copies of the CNV (using the 1KGP SV call set and analysing read depth as shown in SI, Figure S20; the deletion occurs within the *NUTM2B* gene and is also visible in Figure 8).

The region contains a number of genes associated with lung function (pulmonary-surfactant associated proteins *SFTPA1*, *SFTPA2*, *SFTPD*, and *DYDC2*) and immunity (*ANXA11*); SNPs supporting the significant clades in these regions are significantly associated with their expression (Supplementary Table S1). Searching for significant GWAS hits in LD with the predicted inversion carriers (Methods, Section 4.5.4) identified highly correlated variants associated with blood levels of SFTPD (rs2146192: $r^2 = 0.67$) and Cystatin C (rs55855057: $r^2 = 0.64$), decreased haemoglobin

(rs61859980: $r^2 = 0.98$, rs61863508: $r^2 = 1.0$), decreased haematocrit (rs61859980: $r^2 = 0.98$), and increased levels of blood urea (rs36073865: $r^2 = 1.0$, rs55838345: $r^2 = 0.62$, rs17678338: $r^2 = 0.6$, rs17678338: $r^2 = 0.6$).

2.5.5 Other SVs

We scanned all of the identified significant regions for those with relatively high frequency, large genomic spans, (direct or inverted) segmental duplications near the identified breakpoints, and evidence from analysis of reads pointing to possible structural variation. This identified a total of 10 inversions (5 novel), 1 known deletion, 1 novel possible CNV or complex rearrangement, 7 complex rearrangements or other variants (3 novel), and 5 regions with strong indications of structural variation but no clear classification (2 novel). For 12 of these 24 variants, we were able to confirm the presence of structural variants in these regions in the HPRC data. The remaining 26 regions show no clear evidence of structural variation. Full details are presented in Supplementary Table S1. The identified variants include:

- a 550kb region on 11q11 (average carrier frequency of 23%), flanked by inverted segmental duplications of length 30kb, and overlapping a CNV from the 1KGP call set which is in LD with the predicted carriers (SI, Figure S22). This corresponds to a known complex rearrangement identified by Korbel et al. (2007); we verified the presence of structural variation in the region in the HPRC sample. The variant overlaps a number of genes and correlates with a large number of GWAS hits for cardiovascular traits.

- a 375kb region on 11q12.1 (200kb away from the region described above), with an average frequency of 4%, flanked by directly oriented segmental duplications (SI, Figure S23). Analysis of reads in this region indicates the presence of a small deletion and inversion within the predicted span correlated with the predicted carriers. This variant is in LD with a variant significantly associated with adolescent idiopathic scoliosis (rs17500359, $r^2 = 1.0$).

- a 450kb region on 7q11.21, with an average frequency of 13%, flanked by a large number of long inverted segmental duplications (SI, Figure S24) with large number of discordant paired-end reads correlated with predicted carriers. Although we do not find any predicted carriers in the HPRC sample, there is evidence of a 130kb duplication within the region corresponding to a tandem duplication of *RABGEF1* (confirmed as present in 1KGP samples HG01356, HG01351 and HG01140 using analysis of reads), suggesting frequent copy number changes and rearrangements in this region.

3 Discussion

We have found that the distribution of edge span is very accurately captured by our approximation based on the SMC'. The differences between the distribution of edge spans in ARGs simulated under the SMC' model and that in ARGs produced using reconstruction tools are due to both model misspecification and the particularities of each algorithm. Our corrections for Relate result in almost complete recovery of the theoretical distributions. We suggest that the bias seen in ARGs reconstructed using tsinfer stems from the presence of polytomies, which lead to an excess of deep edges with long spans. It is nontrivial to adjust for their presence within our model, and it is not currently possible to break polytomies at random without re-sampling edges in each tree independently (which would prevent a proper calculation of their span). In general, deep edges can arise either through true demographic events or due to inaccuracies in ARG reconstruction; our method can be used to detect deep edges that are likely to be artefactual. We note that apart from ARGweaver, none of the ARG reconstruction methods we consider are explicitly optimised to recover the distribution of edge span, focusing instead on other aspects such as node times, local tree topologies, and patterns of LD (which *are* recovered well in our simulation studies). Our results can potentially be used to improve the estimates of edge span produced by tsinfer and Relate during the topology reconstruction step, and hence also improve the downstream inference of node times.

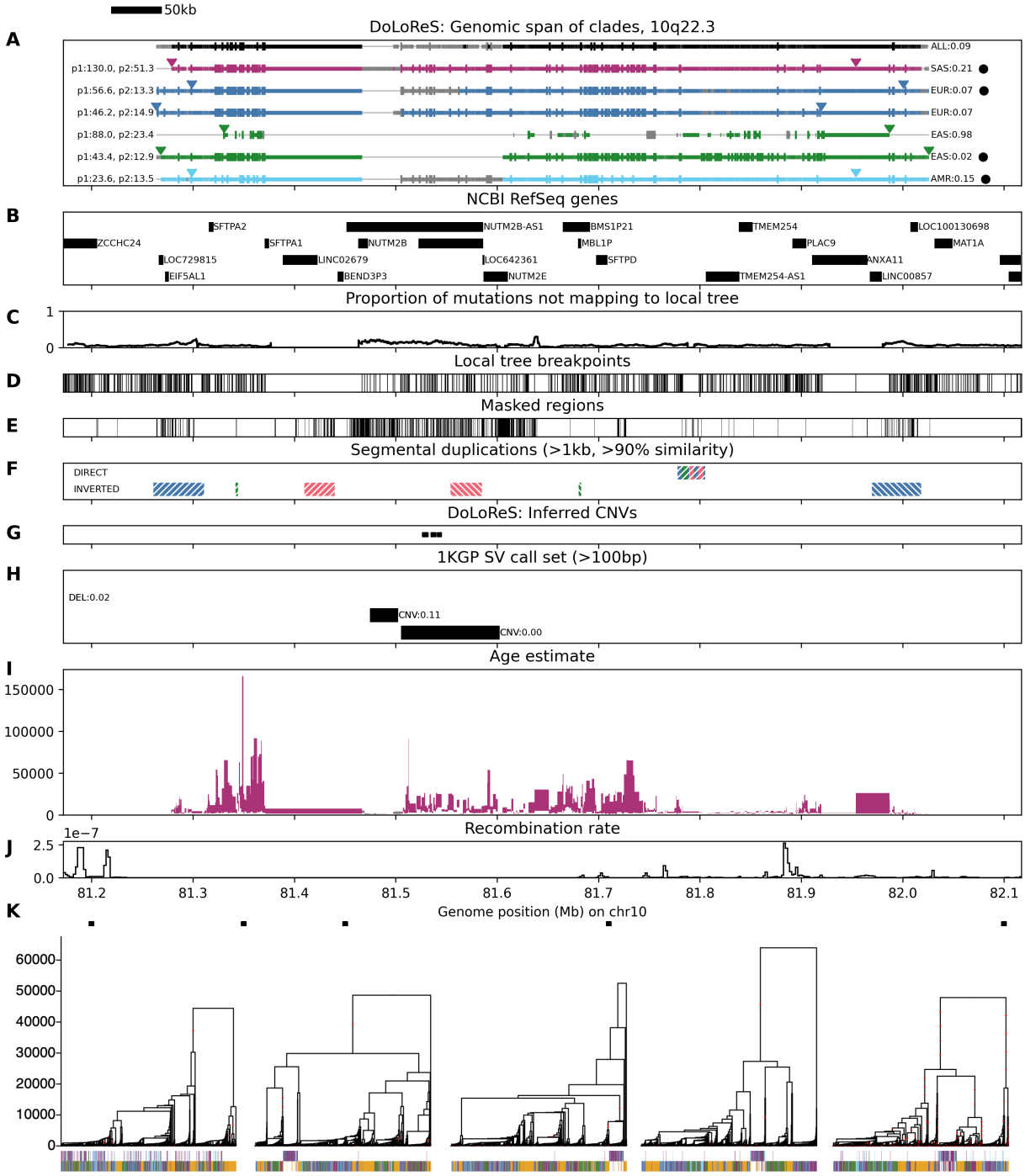

Figure 9: Details of 10q22.3 region. **A**: span of each significant clade shown as horizontal line (grey: clade is not present at that position exactly, but another highly correlated clade is). Vertical lines show positions of SNPs supporting the clade. Left label: p-values for Tests 1 and 2. Right label: population frequency. Black circles indicate predicted carriers. See Methods, Section 4.5 for details of this and other panels. **B**: positions of genes (UCSC Genome Browser NCBI RefSeq track). **C**: moving average of the proportion of SNPs not mapping onto the corresponding local tree. **D**: positions of breakpoints between local trees. **E**: regions masked during ARG reconstruction. **F**: segmental duplications. **G**: predicted positions of CNVs based on ARG clade analysis (Methods, Section 4.5.2). **H**: SVs in the 1KGP call set (labels show correlation with predicted inversion carriers). **I**: age of predicted inversion estimated using the ARG subsetted to S. Asian populations (Methods, Section 4.5.3). **J**: recombination rate (using HapMapII recombination map; averaged in bins of 2kbp). **K**: local trees at the positions indicated by squares; y-axis is time measured in generations; vertical lines drawn below each sample with colour corresponding to population (those belonging to a significant clade drawn in top row).

The SMC'-based approximation we construct for the distribution of the length of a haplotype block (the span of a clade of samples) also provides a very close fit, based on simulations. The corresponding tool we develop for detecting regions of locally suppressed recombination has excellent performance on simulated ARGs, as well as (with appropriate adjustments) ARGs reconstructed using Relate. Since the method detects long clade span after adjusting for the age of the clade and the local tree topology, and hence specifically detects localised (between-clade) suppression of recombination, it has the power to discriminate inversions from other genealogy-distorting events, such as point mutations under balancing selection. Our method can be used with arbitrary models of varying population size, and (based on simulation studies) appears to be generally robust to misspecification of demographic history. An inherent limitation of the method is that deep, ancient, population structure can result in a similar signal of localised recombination suppression as SVs, so cannot be easily distinguished. The method also cannot identify the specific types of genomic variants that might be causing suppression of recombination in heterozygotes, without utilising other sources of information (such as direct or inverted segmental duplications and other genomic features near the identified regions, or additional analysis of other types of data). However, it provides an alternative line of evidence to methods based on the analysis of paired-end reads, which can miss the presence of complex structural variants or those occurring in regions with poor read mapping.

Applying DoLoReS to the 1KGP ARG reconstructed using Relate identifies a number of regions with both known and novel SVs. Using the ARG allows for the genealogy-based analysis of the age and population frequencies of an SV, and potentially identification of recurrent inversion events. Our tool identifies a large and relatively common inversion on chromosome 10, which has remained unobserved using previous methods due to a lack of clear signal in this region from paired-end read mapping. This demonstrates the power of our method to pick up signals of localised recombination suppression. We limited our detailed examination to regions of size at least 50kb which are well-supported by SNPs, but note that there is a large number of other smaller regions with significant p-values (SI, Figure S17), which are more difficult to confidently validate.

Roughly half of the detected regions show no clear signals of structural variation based on our analysis. We suggest that other, non-structural reasons may explain allele-specific suppressed recombination in these regions. Previous evidence suggests that recombination crossovers are suppressed within the boundaries of genes expressed in meiosis (McVicker and Green, 2010). This suggests that expression quantitative trait loci (eQTLs) altering meiotic gene expression might impact crossover rates by suppressing recombination crossovers on carriers of one particular allele and in particular in heterozygous individuals, similar to structural drivers. To test whether the identified regions and carriers supported this possibility, we first tested (as detailed in Section 4.5.6) whether our regions are enriched for closely matching gene boundaries. We observe strong enrichment (observed 14 single-gene regions, expected 0.67, OR = 31.2, $p = 5 \cdot 10^{-10}$), with most of the observed overlaps (10) among those not showing structural evidence. Moreover, the 14 corresponding genes are significantly enriched for being highly expressed during male gametogenesis (9 genes; OR=3.2; p = 0.047). Secondly, we tested whether SNPs defining carriers of recombination-suppressed alleles are enriched for known eQTLs. Again, we see evidence of enrichment ($p = 2 \cdot 10^{-3}$), although whether these function in meiotic tisues is unknown. Specific example eQTL regions include *SCMH1* on 1p34.2, *SPATA6* on 1p33, and *ZFAND3* on 6p21.2, all essential for normal spermatogenesis (Takada et al., 2007; Yuan et al., 2015; de Luis et al., 1999). This enrichment of our regions for almost perfect overlap with genes, and eQTLs altering their expression, supports the hypothesis of meiotic allele-dependent suppression of recombination within genes. In contrast, although many (30) of our regions are in LD with GWAS hits, this overlap is not significantly higher than expected by chance ($p = 0.14$). Neither do we observe significant excess overlap with regions under selection identified by Akbari et al. (2024) ($p = 0.60$), although several regions contain one or more significant selection SNPs: 12q24.11 (110.7–111.1Mb), 9p21.1 (31.9–32.1Mb), 2p23.1 (31.8–32.4Mb), 17q21.31 (43.6–44.4Mb), with only the last possessing clear evidence of structural variation. Overall, the data suggest mainly mechanistic drivers of suppressed inter-allelic recombination crossover regions, rather than, for instance, local epistatic selection among trait-influencing variants.

It is clear that even though Relate only uses SNP data and does not explicitly model the presence of SVs, the reconstructed ARG faithfully captures some of these signals. While this allows for their detection, this also means that the genealogies can be distorted by the presence of SVs, most obviously through the phasing errors that they induce in the data. This highlights the value of considering structural variation as an important source of information when developing future ARG reconstruction and analysis methods.

4 Materials and methods

4.1 Probability that an edge is disrupted by a recombination event

Let $\mathcal{T}$ be a fixed local tree, and consider a particular edge b within this tree. We would like to calculate, under the SMC' and conditional on $\mathcal{T}$, the probability $\mathbb{P}_{\mathcal{T}}(b \text{ disrupted})$ that the next recombination event arriving along the genome disrupts b: that is, it changes the time-length of b, or the topology of the tree around b (we refer to the latter as b being *topologically disrupted*). This can happen because either (1) the recombination point is on b and the coalescence point is not on b, or (2) the recombination point is not on b but the coalescence point is (SI, Figure S1). The possible events that do and do not disrupt b are illustrated in SI, Figure S2. By integrating over the possible positions of the recombination point and new coalescence event, we calculate the probability of each illustrated event as detailed in SI, Section S1.3. The resulting expression is in closed form and can be calculated for any given tree and branch. We derive an equivalent probability $\mathbb{P}_{\mathcal{T}}(b \text{ topologically disrupted})$, which only takes into account topology-changing recombination events, as detailed in SI, Section S1.4.

4.2 Distribution of edge span

Conditional on the local tree $\mathcal{T}$, for a particular branch b, we are interested in the distribution of the genomic span before b is disrupted by a recombination event. We approximate this distribution under the SMC' by making the assumption that as recombination events arrive along the genome, if they do not disrupt b then they also do not otherwise change the rest of the local tree (so $\mathcal{T}$ stays fixed along the genome). Then since recombination events arrive along the genome as a Poisson process with rate $\mathcal{L}_{\mathcal{T}} \cdot \rho/2$ (where $\mathcal{L}_{\mathcal{T}}$ is the total branch length of $\mathcal{T}$), by thinning, recombination events that disrupt b arrive at rate

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}} \cdot \rho/2. \tag{4.1}$$

Thus the waiting time until b is disrupted by a recombination event is exponential with this rate. We derive an equivalent result when the recombination rate is not constant along the genome (SI, Section S1.8), when considering only topology-changing events (SI, Section S1.8.1), and when we condition on the branch having at least one mutation event (SI, Section S1.8.2).

The assumption that recombination events do not change $\mathcal{T}$ within the span of b is very strong. However, through quantifying the effect of recombination on the height (SI, Section S1.7) and total branch length (SI, Section S1.6) of local trees, we show that the averaged effect of recombination on the rest of the tree does not appear to significantly affect the probability that the given edge is disrupted, making this an excellent approximation (SI, Section S1.11).

4.3 Distribution of clade span

For a given local tree $\mathcal{T}$, we define each clade G through the samples it contains as in Section 2.3. We calculate the probability $\mathbb{P}_{\mathcal{T}}(G \text{ disrupted})$ that, under the SMC' and conditional on $\mathcal{T}$, G is disrupted by the next recombination event along the genome: that is, that the membership of sample nodes in the clade changes in the next local tree (allowing events that disrupt edges within the clade without changing the group of subtended samples). This can happen when the recombination point is on a edge within the clade and the coalescence point is on an edge outside the clade, or if the recombination point is on an edge outside the clade and the coalescence point is

on an edge within the clade, as illustrated in SI, Figure S3. We calculate the probability of these events as detailed in SI, Section S1.5.

The distribution of the genomic span of a clade G can then be similarly approximated by making the assumption that recombination events that do not disrupt G do not change $\mathcal{T}$; then the waiting time until G is disrupted by a recombination event is exponentially distributed with rate

$$\mathbb{P}_{\mathcal{T}}(G \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}} \cdot \rho/2.$$

We derive an equivalent result when the recombination rate changes along the genome (SI, Section S1.9).

4.3.1 Test 1

Given an ARG, for each clade G_i we calculate its genomic span $[a, b]$, and use the approximation (4.1) to compute a corresponding one-sided p-value to test whether G_i has a significantly longer span than under the null hypothesis of no local (between-clade) recombination suppression:

$$p_i = \exp\left(-\mathbb{P}_{\mathcal{T}}(G \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}} \cdot \int_a^b \rho(w)/2dw\right),$$

where $\rho(w)$ is the recombination rate at position w (SI, Section S1.9). Simulation studies show that the test has excellent performance for simulated ARGs even for small inversions, and maintains good sensitivity for detecting inversions of over 50kb for Relate ARGs, while accurately pinpointing their position (SI, Section S1.13.2).

4.3.2 Test 2

Under the model described above, an equivalent test can be constructed using the number of recombination events R occurring within the genomic span of G, which has a geometric distribution with rate $\mathbb{P}_{\mathcal{T}}(G \text{ disrupted})$ (SI, Section S1.14). For each clade G_i within a given ARG, we thus calculate a corresponding (one-sided) p-value

$$p_i = [1 - \mathbb{P}_{\mathcal{T}}(G \text{ disrupted})]^{R-1}.$$

In practice, R is unknown, so we instead calculate the number of breakpoints between local trees within $[a, b]$ (this is conservative as it strongly underestimates the number of recombination events in this interval).

4.4 1KGP ARG

We use tskit (Kelleher et al., 2016) to split the ARGs into the five super-populations (EUR: European, AFR: African, SAS: S. Asian, EAS: E. Asian and AMR: American), and analyse them separately. We adjusted for varying population size (as estimated by Relate) as detailed in SI, Section S1.9.1, applied the corrections detailed in SI, Section S1.13.1, and used the 1KGP genomic mask to set the recombination rate in regions marked as 'not passing' to 0 (to avoid false positives for Test 1). Additionally, to be robust to the presence of phasing switch errors in the data, instead of defining a clade by the samples it contains as described in Section 2.3, we instead count how many samples from each individual it contains. That is, we assign each clade G a "genotype" ID $(G_1, \ldots, G_N)$, where G_i is the number of sequences (0, 1, or 2) within G from individual i, and consider the clade present in a given local tree if there is a clade with the same genotype ID. We filter out clades supported by fewer than 10 mutations, spanning less than 50kb or fewer than 10 local trees, and those having fewer than 10 or more than $N - 10$ samples (where N is the total number of samples in the ARG). Tests 1 and 2 are then applied for each remaining clade independently (using the HapMapII GRCh37 recombination map for Test 1, and counting the number of local tree breakpoints to estimate the number of recombination events R for Test 2). We require both p-values to be below the Bonferroni-corrected threshold of $1 \cdot 10^{-12}$ (being 0.05 divided by the total number of clades in the ARGs).

4.5 Analysis of results

For each significant clade G, in each local tree within its genomic span, we identify the clade $\widetilde{G}$ with which it is most highly correlated (as measured by the correlation coefficient between their genotype IDs). In Figure 9A, the span of each significant clade G is plotted as a solid horizontal line (in colour corresponding to the super-population), if within the local tree at the given genomic position, there is a clade $\widetilde{G}$ with a correlation coefficient of at least 0.95 (and in grey if the correlation coefficient is between 0.9 and 0.95). The leftmost and rightmost positions at which the clade appears in the ARG exactly are indicated by triangles.

We also compute the combined genotype ID for the corresponding predicted clade of samples in the full (all-population) ARG, and calculate its genomic span (shown in black in Figure 9A). Existence of this "superclade" provides additional evidence that the clades identified independently in each population ARG are not false positives.

4.5.1 Genomic features and measures of ARG reconstruction quality

For each identified region, we extract the positions of nearby genes, using the UCSC Genome Browser NCBI RefSeq track (Pruitt et al., 2005). To check for known genomic features that tend to co-occur with SVs, we also extract the positions of segmental duplications of length greater than 1kb and $> 90\%$ sequence similarity falling within this region (Bailey et al., 2001, 2002), and the positions of SVs in the 1KGP call set (Sudmant et al., 2015). To detect issues caused by ARG reconstruction artefacts, we extract the positions of breakpoints between adjacent local trees, and the regions masked during ARG reconstruction (labelled as "not passing" in the 1KGP pilot mask); we also calculate a moving average (in 10kb windows) of the proportion of SNPs in the 1KGP data that cannot be uniquely mapped onto a branch in the local tree at the corresponding position (as a measure of ARG reconstruction error). To detect instances where poor estimation of event times may inflate the probability that the clade is disrupted by recombination, we also checked (for each significant clade) the proportion of mutations that fall within vs. outside the clade, and compared this to the average proportion of tree total branch length within vs. outside the clade (expecting these quantities to be similar if times are well estimated).

4.5.2 Phasing errors

Using the individual-based definition of a clade means that it is possible for two clades in the same local tree to have identical IDs, if one clade contains one chromosome from each of k individuals, and the other clade contains the other chromosome for each of these k individuals. This has negligible probability to arise by random chance unless k is very small (4 or less, based on simulations with 1KGP-like parameters). This can, however, arise for larger k as the result of phasing errors due to structural variation, in particular due to mis-alignment of CNVs which results in ARG reconstructed artefacts. We thus record all instances where, in a local tree, two clades of size at least 6 have identical IDs, to look for this signal.

4.5.3 Inversion age

A lower and upper bound on the age of an inversion can be obtained by identifying the most highly correlated clade $\widetilde{G}$ in each local tree within the genomic span of G, and (if the correlation coefficient is at least 0.9) obtaining the times at the bottom and top end of the branch that subtends $\widetilde{G}$ (call these times s and t, respectively).

While a neutral mutation can arise at any time uniformly distributed along this branch, an inversion prevents certain types of recombination events, so the changes in ARG topology and in s and t along the genome are informative of its age. Suppose that the inversion is old. Then Figures 10A-B imply s should be relatively recent (being approximately distributed as the coalescence time of k samples), while Figures 10C,D,H imply that both t and the MRCA time will be large, since they are constrained to be larger than the age of the inversion. Moreover, if the carriers and non-carriers form two (disjoint) clades in the ARG, the only type of event that can change this topology is that

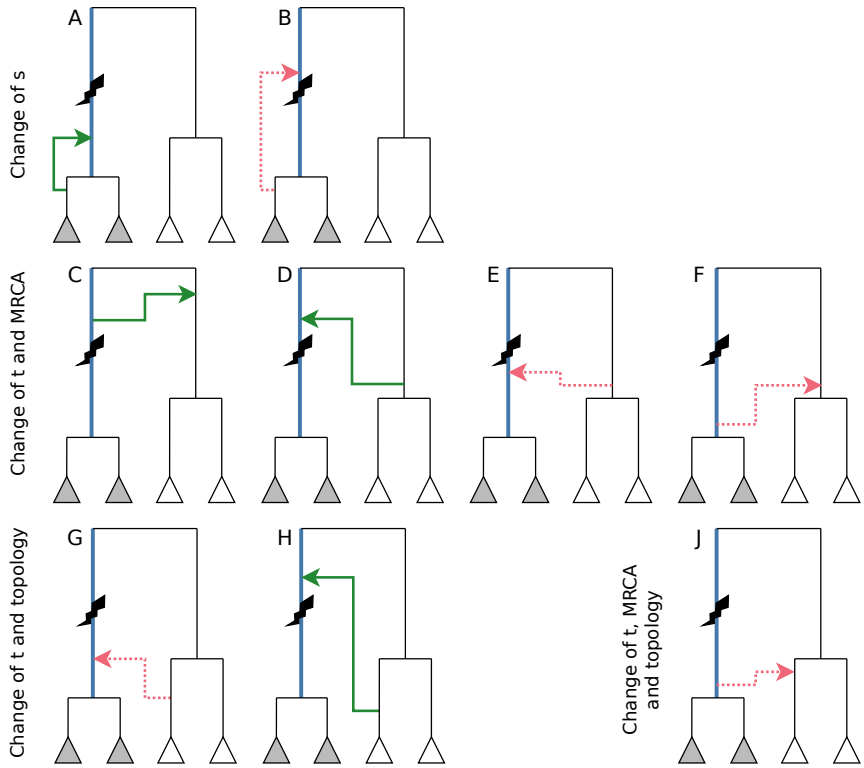

Figure 10: Possible recombination events that change s (time of the bottom end of the branch subtending the samples carrying the inversion, shown in blue), t (the of the top end of the branch), the MRCA of the carriers and non-carriers, and/or the order of coalescence of the clades. Inversion shown as lightning bolt. Recombination events shown as arrows, where the start of the arrow shows the time and location of the recombination event and the arrowhead shows that of the subsequent coalescence; green solid (resp. red dotted) arrows show feasible (resp. infeasible) events.

shown in Figure 10H, which has a very low probability under the SMC' assumptions. However, a double crossover within the inverted region can allow any of the events shown in red, locally changing s, t, the MRCA time and/or the topology, in the region between the two recombination breakpoints.

This description aligns with the observed ARG topology and branch time estimates for the 17q21.31 inversion (SI, Figure S18): the carriers and non-carriers form two disjoint clades with a very large MRCA time for most of the inverted region, apart from the region highlighted in red in Figure 7. In this region the corresponding local tree topologies change in a way consistent with an event of type F around position 43.86Mb (which changes t and the MRCA time), followed by a number of recombination events of type J between 43.87-43.90Mb (which change t, the MRCA time, and the tree topology as shown in SI, Figure S18K).

4.5.4 Significant eQTLs and GWAS hits in LD with identified variants

For each identified significant clade, we searched the Open Targets Genetics catalog (Ghoussaini et al., 2021; Mountjoy et al., 2021) for genome-wide significant GWAS hits ($p < 5 \cdot 10^{-8}$) in LD ($r^2 > 0.6$) with SNPs supporting the clade. The full details of the identified SNPs are presented in Supplementary Table S2. We also checked SNPs supporting the clade for significant associations with gene expression using the QTL catalog.

4.5.5 Analysis of HPRC data

For each identified significant region, we predicted the carrier status of each sequence in the HPRC dataset by checking whether it carries SNPs that occur on the branches that subtend the significant clades. For each sequence, we then calculated k-mer counts (setting $k = 20$) in (and

around) the region, and counted the number of k-mers deleted, duplicated, or inverted when compared to the T2T-CHM13v2 reference. We then calculated the correlation between these k-mer counts and predicted carrier status. To look for changes in ordering indicating rearrangements or inversions, we also (1) plotted the ordering of genes in the region as annotated for each sequence, and (2) selected a subsample of equally spaced k-mers on the T2T-CHM13v2 reference, and plotted the relative positions of these k-mers on each sequence.

4.5.6 Testing for enrichment

We tested for enrichment of our identified regions for overlapping with genes, genes involved in meiosis, SNPs under selection, GWAS hits and eQTLs. For each significant region we selected a tagging SNP in the middle of the region, constructed a list of 10 best-matched SNPs on the same chromosome (matching on the overall frequency, frequency in Europeans, and average recombination rate within the surrounding 1Mb region), and defined a region around the matched SNPs of the same physical (and approximately the same genetic) size. We then checked each matched region for overlapping SNPs under selection from Akbari et al. (2024), genes (using the Genome Browser NCBI RefSeq track), genes involved in meiosis (taking clusters 2-4 from Xia et al. (2020, Table S3) and filtering for genes within the top 25% by expression level in testis bulk sequencing data (obtained from the GTEx portal on 19/01/2025, file `gene_reads_v10_testis.gct.gz`), and GWAS hits and eQTLs as described in Section 4.5.4 (filtering the Open Targets eQTLs for those with an association score of at least 0.6). We then calculated bootstrapped p-values by resampling using these matched regions.

4.6 Simulation parameters

4.6.1 Neutral simulations

Using stdpopsim, we simulated two ARGs under the SMC' with $n = 100$ (haploid samples) and the following two sets of parameters:

- dataset 1: Chr21 with HapMapII_GRCh38 recombination map, mutation rate $1.29 \cdot 10^{-8}$ per site per generation, $N_e = 10\,000$ diploids, constant population size model;
- dataset 2: 5Mb of Chr21 with flat recombination map, recombination rate $1.2 \cdot 10^{-8}$ per site per generation, mutation rate $1.29 \cdot 10^{-8}$ per site per generation, $N_e = 10\,000$ diploids, constant population size model.

For each simulated dataset, we used ARGweaver, Relate (v1.1.9), tsinfer/tsdate (v0.3.0 and v0.1.5), and ARG-Needle to reconstruct an ARG, using the true simulation parameters as inputs for each tool (and for ARGweaver, a time discretisation grid with 100 points and selecting the MAP ARG from 1000 posterior samples, for ARG-Needle using 50 time discretisation points). We sense-checked the output of the ARG reconstruction methods using a number of metrics (comparing the MRCA times, local tree topologies, and LD decay, against the simulated ARGs).

4.6.2 Simulations with inversion

We used SLiM (Haller and Messer, 2019; Haller et al., 2019) to simulate an ARG with one inversion under balancing selection ($n = 100$, 5Mb with recombination rate $1 \cdot 10^{-8}$ per bp per generation, constant population size 10 000, inverted segment length 200kb, neutral mutations at rate $1 \cdot 10^{-8}$ per bp per generation added using msprime).

4.6.3 1KGP simulation

To simulate data using parameters similar to the 1KGP data, we used stdpopsim with the AmericanAdmixture_4B11 demographic model, simulating the same number of African, European, Asian and Admixed samples as in the 1KGP, for chromosomes 18-22 (using the HapMapII GRCh37

recombination map). We then applied the 1KGP genomic mask and reconstructed an ARG using Relate, and applied the methods described in Section 4.4 to calculate p-values, using a Bonferroni-corrected significance threshold of $1 \cdot 10^{-9}$. The results are shown in Figure 6 (right panel).

5 Data availability statement

Data sharing is not applicable to this article as no new data were generated or analysed.

6 Acknowledgements

This work was initiated at the *Stochastic modelling in the life sciences* Junior Trimester Programme held at the Hausdorff Research Institute for Mathematics, University of Bonn (funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy EXC-2047/1-390685813). SRM and AI are supported by the Wellcome Trust (Investigator Award 212284/Z/18/Z), MF by the Knut and Alice Wallenberg Foundation (Program for Mathematics, grant 2020.072, hosted by the Department of Statistics, University of Warwick), JK by the EPSRC (research grant EP/V049208/1), and JS by the ERC (Starting Grant ARGPHENO 850869). We thank Yan Wong and Pier Palamara for helpful discussions, and Sebastian Quintanilla Terminel and Kari Heine for useful comments.

References

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

Abel, H., Larson, D., Regier, A., Chiang, C., Das, I., Kanchi, K., Layer, R., Neale, B., Salerno, W., Reeves, C., et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, **583**(7814): 83–89, 2020.

Akbari, A., Barton, A. R., Gazal, S., Li, Z., Kariminejad, M., Perry, A., Zeng, Y., Mittnik, A., Patterson, N., Mah, M., et al. Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. *bioRxiv*, 2024.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Teague, B., Ventura, M., Girirajan, S., Alkan, C., Campbell, C. D., Vives, L., Malig, M., et al. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature Genetics*, **42**(9): 745–750, 2010.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. Recent segmental duplications in the human genome. *Science*, **297**(5583): 1003–1007, 2002.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research*, **11**(6): 1005–1017, 2001.

Bansal, V., Bashir, A., and Bafna, V. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research*, **17**(2): 219–230, 2007.

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

Brandt, D. Y. C., Wei, X., Deng, Y., Vaughn, A. H., and Nielsen, R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, **221**(1): iyac044, 2022.

Chiang, C., Scott, A., Davis, J., Tsang, E., Li, X., Kim, Y., Hadzic, T., Damani, F., Ganel, L., Consortium, G., et al. The impact of structural variation on human gene expression. *Nature Genetics*, **49**(5): 692–699, 2017.

de Luis, O., López-Fernández, L. A., and del Mazo, J. Tex27, a gene containing a zinc-finger domain, is up-regulated during the haploid stages of spermatogenesis. *Experimental Cell Research*, **249**(2): 320–326, 1999.

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

Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S. Q., Kajuna, S. L., Barta, C., Kungulilo, S., Karoma, N., Lu, R.-B., et al. The distribution and most recent common ancestor of the 17q21 inversion in humans. *American Journal of Human Genetics*, **86**(2): 161–171, 2010.

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, **49**(D1): D1311–D1320, 2021.

Griffiths, R. C. and Marjoram, P. An ancestral recombination graph. In P. Donnelly and S. Tavaré, eds., *Progress in population genetics and human evolution*, 257–270. Springer, New York, 1997.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**(2): 552–566, 2019.

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

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

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

Kirkpatrick, M. How and why chromosome inversions evolve. *PLOS Biology*, **8**(9): e1000501, 2010.

Kirkpatrick, M. and Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics*, **173**(1): 419–434, 2006.

Korbel, J. O., Urban, A. E., Grubert, F., Du, J., Royce, T. E., Starr, P., Zhong, G., Emanuel, B. S., Weissman, S. M., Snyder, M., et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *PNAS*, **104**(24): 10110–10115, 2007.

Li, H. and Ralph, P. Local pca shows how the effect of population structure differs along the genome. *Genetics*, **211**(1): 289–304, 2019.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., et al. A draft human pangenome reference. *Nature*, **617**(7960): 312–324, 2023.

Lucas Lledó, J. I. and Cáceres, M. On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLOS One*, **8**(4): e61292, 2013.

Lupski, J. R. and Stankiewicz, P. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLOS Genetics*, **1**(6): e49, 2005.

Marjoram, P. and Wall, J. D. Fast coalescent simulation. *BMC Genetics*, **7**(1): 1–9, 2006.

Martínez-Fundichely, A., Casillas, S., Egea, R., Ramia, M., Barbadilla, A., Pantano, L., Puig, M., and Caceres, M. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research*, **42**(D1): D1027–D1032, 2014.

McKenzie, P. F. and Eaton, D. A. R. Estimating waiting distances between genealogy changes under a multi-species extension of the sequentially Markov coalescent. *bioRxiv*, 2022.

McVicker, G. and Green, P. Genomic signatures of germline gene expression. *Genome Research*, **20**(11): 1503–1511, 2010.

Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M. A., et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, **53**(11): 1527–1533, 2021.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A., Mikheenko, A., Vollger, M., Altemose, N., Uralsky, L., Gershman, A., et al. The complete sequence of a human genome. *Science*, **376**(6588): 44–53, 2022.

Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, **50**(9): 1311–1317, 2018.

Peischl, S., Koch, E., Guerrero, R., and Kirkpatrick, M. A sequential coalescent algorithm for chromosomal inversions. *Heredity*, **111**(3): 200–209, 2013.

Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maggiolini, F. A. M., Harvey, W. T., et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, **185**(11): 1986–2005, 2022.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **33**: D501–D504, 2005.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18): i333–i339, 2012.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M. C. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, **15**(6): 461–468, 2018.

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

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

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., et al. A common inversion under selection in Europeans. *Nature Genetics*, **37**(2): 129–137, 2005.

Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, **44**(8): 872–880, 2012.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571): 75–81, 2015.

Takada, Y., Isono, K.-i., Shinga, J., Turner, J. M., Kitamura, H., Ohara, O., Watanabe, G., Singh, P. B., Kamijo, T., Jenuwein, T., et al. Mammalian Polycomb Scmh1 mediates exclusion of Polycomb complexes from the XY body in the pachytene spermatocytes. *Development*, **134**(3): 579–590, 2007.

Tattini, L., D'Aurizio, R., and Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, **3**: 92, 2015.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. Fine-scale structural variation of the human genome. *Nature Genetics*, **37**(7): 727–732, 2005.

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

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

Xia, B., Yan, Y., Baron, M., Wagner, F., Barkley, D., Chiodin, M., Kim, S. Y., Keefe, D. L., Alukal, J. P., Boeke, J. D., et al. Widespread transcriptional scanning in the testis modulates gene evolution rates. *Cell*, **180**(2): 248–262, 2020.

Yuan, S., Stratton, C. J., Bao, J., Zheng, H., Bhetwal, B. P., Yanagimachi, R., and Yan, W. Spata6 is required for normal assembly of the sperm connecting piece and tight head–tail conjunction. *PNAS*, **112**(5): E430–E439, 2015.

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

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

Supplementary Information

S1 Supplementary Methods **1**

S1.1 The SMC' model and ARG reconstruction . . . 1
S1.2 Notation and background . . . 2
S1.3 Probability that an edge is disrupted by a recombination event . . . 3
S1.4 Probability that an edge is topologically disrupted by a recombination event . . 6
S1.5 Probability that a clade is disrupted by a recombination event . . . 7
S1.6 Change in total branch length of tree following a recombination event . . . 8
S1.7 Change in tree height following a recombination event . . . 9
S1.8 Distribution of edge span . . . 9
S1.9 Distribution of clade span . . . 12
S1.10 Quality of approximation to the distribution of edge span . . . 13
S1.11 Effects of recombination on a local tree . . . 13
S1.12 Comparison of simulation models . . . 15
S1.13 Detection of local recombination suppression: Test 1 . . . 16
S1.14 Detection of local recombination suppression: Test 2 . . . 18

S2 Proofs **20**

S3 Supplementary Figures **32**

S1 Supplementary Methods

S1.1 The SMC' model and ARG reconstruction

ARGs were first described by Griffiths and Marjoram (1997) as realisations of the coalescent with recombination (CwR), a stochastic process operating backwards in time, generating a genealogy through a sequence of coalescence and recombination events (Hudson, 1983). Wiuf and Hein (1999) reframed the CwR as a stochastic process operating spatially along the genome: starting with the leftmost endpoint, local trees are generated sequentially moving to the right, reshaped by recombination events. While calculating the properties of the ARG under both frameworks is generally intractable, this seminal work spurred on a suite of simplifying approximations, enabling applications to large-scale genomic data.

The sequentially Markovian coalescent (SMC) model, proposed by McVean and Cardin (2005), imposed the assumption that the process along the genome is Markovian, which, in essence, prohibits recombination events in genetic material not ancestral to the sample. This was followed by the SMC' extension (Marjoram and Wall, 2006), which was shown to be an excellent approximation to the CwR, based on the joint distribution of pairwise coalescent times and a quantification of bias in population size estimates (Wilton et al., 2015), and the distribution of the next local tree conditional on the current one in a two-locus model (Hobolth and Jensen, 2014). Thus, for a small trade-off in accuracy, the SMC' model offered a substantially more tractable way of calculating analytic approximations to various quantities of interest, such as the correlation between coalescence time and linkage probability for a randomly sampled pair of sequences (Eriksson et al., 2009), and identity-by-descent tract length distributions and related quantities (Harris and Nielsen, 2013; Carmi et al., 2014). It also enabled the development of powerful new inference methods: for instance, by considering the genealogy of a single pair of sequences, Li and Durbin (2011) developed a HMM-based approach (the pairwise SMC, or PSMC) for inferring the history of human population sizes, which was subsequently extended by Schiffels and Durbin (2014) to multiple samples.

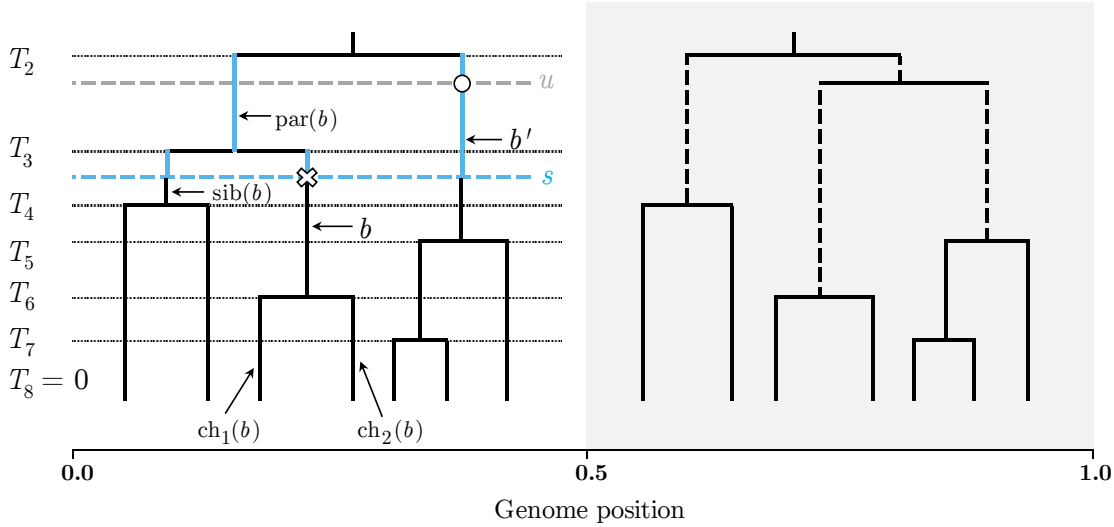

Figure S1: Illustration of the notation used throughout. The ARG has two marginal trees, where the tree on the left is $\mathcal{T}$, with $n = 7$. Coalescent event times are shown as black dotted lines. For the edge labelled b, $t^{\uparrow}(b) = T_3$, $t^{\downarrow}(b) = T_6$, $d^{\leftarrow}(b) = 0$, $d^{\rightarrow}(b) = 0.5$ (so the span of the edge is 0.5); $\mathcal{A}(b) = \{b, \mathrm{sib}(b), \mathrm{par}(b)\}$ and $\mathcal{B}(b) = \{b, \mathrm{sib}(b), \mathrm{ch}_1(b), \mathrm{ch}_2(b)\}$. The recombination event occurs at genomic position 0.5; the recombination point $\mathcal{R} = (b, s)$ is shown as a cross; the coalescence point $\mathcal{C} = (b', u)$ is shown as a circle; $n(s) = 3$ and $L_{\mathcal{T}}(s)$ gives the total length of the edges shown in blue. The tree on the right $\mathcal{T}'$ is obtained by pruning the subtree below $\mathcal{R}$ and reattaching at $\mathcal{C}$; solid vertical lines show edges that have not been affected by the recombination event.

Meanwhile, the definition of the ARG has become decoupled from its initial description as the realisation of a stochastic process, to more broadly denote a genealogical network that captures genetic inheritance. Using this looser definition, the problem of explicitly reconstructing plausible ARGs from sequencing data has seen significant recent progress driven by the use of heuristic methods and principled approximations to the CwR. ARGweaver (Rasmussen et al., 2014) implements an MCMC scheme based on a time-discretised version of the SMC (or SMC') to obtain posterior samples of ARGs compatible with a given dataset. Relate (Speidel et al., 2019) and tsinfer/tsdate (Kelleher et al., 2019; Wohns et al., 2022) reconstruct a single ARG from data, by using methods based on the Li and Stephens (2003) framework to first reconstruct the topologies and then estimating the edge lengths using Bayesian approaches with coalescent-based priors. ARG-Needle (Zhang et al., 2023) reconstructs a single ARG by sequentially threading in each sample, by first identifying the most closely related samples already in the ARG via genotype hashing, and subsequently estimating coalescence times under the Ascertained Sequentially Markovian Coalescent (ASMC) model (a coalescent-based HMM). These methods scale to thousands of human genome-length samples and have already been applied to many large-scale datasets, resulting in powerful inference of evolutionary events and parameters, such as the history of human demography (Wohns et al., 2022), past population sizes (Speidel et al., 2019), signals of selection (Hejase et al., 2022), and genetic associations for complex traits (Zhang et al., 2023).

S1.2 Notation and background

The notation is illustrated in Figure S1. Let $\mathcal{T}$ be a fixed local tree with n leaves. Denote by T_i (for $i \in \{2, \cdots, n\}$) the population-scaled time at which the number of lineages in $\mathcal{T}$ jumps from i to $i-1$, with T_2 being the time of MRCA and setting $T_{n+1} := 0$. Let $n(t)$ be the number of lineages at time t, so $n(0) = n$, with $n(T_j) = j - 1$ and $n(t) = 1$ for $t \geq T_2$.

For an edge $b \in \mathcal{T}$, denote the lower end time by $t^{\downarrow}(b) \geq 0$ and the upper end time by $t^{\uparrow}(b) \leq T_2$, with the *time-length* of the edge given by $\bar{t}(b) = t^{\uparrow}(b) - t^{\downarrow}(b)$. Let $d^{\leftarrow}(b)$ and $d^{\rightarrow}(b)$ be the leftmost and rightmost endpoints of the genomic span of edge b, respectively, with its *span* given by $d^{\rightarrow}(b) - d^{\leftarrow}(b)$.

Let $\mathrm{par}(b), \mathrm{sib}(b), \mathrm{ch}_1(b)$ and $\mathrm{ch}_2(b)$ denote the parent, sibling, left child and right child edge of b respectively, such that we have the following relations:

$$\begin{aligned}
&t^{\downarrow}(\mathrm{par}(b)) = t^{\uparrow}(b) \\
&t^{\uparrow}(\mathrm{sib}(b)) = t^{\uparrow}(b), \\
&t^{\uparrow}(\mathrm{ch}_1(b)) = t^{\uparrow}(\mathrm{ch}_2(b)) = t^{\downarrow}(b), \\
&\mathrm{ch}_1(b) = \mathrm{ch}_2(b) = \emptyset \text{ if } t^{\downarrow}(b) = 0 \text{ (b extends from a leaf node).}
\end{aligned}$$

Define the sets of edges $\mathcal{A}(b) := \{b, \mathrm{sib}(b), \mathrm{par}(b)\}$ and $\mathcal{B}(b) := \{b, \mathrm{sib}(b), \mathrm{ch}_1(b), \mathrm{ch}_2(b)\}$, and denote by b_r the root lineage extending past the MRCA node.

Let $L_{\mathcal{T}}(t)$ be the sum of edge lengths in $\mathcal{T}$ above time t and up to T_2:

$$L_{\mathcal{T}}(t) = \sum_{b \in \mathcal{T}: t^{\uparrow}(b) > t} \left[t^{\uparrow}(b) - \max(t, t^{\downarrow}(b)) \right],$$

so that $L_{\mathcal{T}}(0)$ is the total branch length of $\mathcal{T}$ (condensed as $\mathcal{L}_{\mathcal{T}} := L_{\mathcal{T}}(0)$ in the main text). Denote by $\mathcal{T}_x$ the local tree at position x along the genome.

Under the SMC', moving along the genome, a recombination event happens after an exponentially distributed waiting time with rate $\mathcal{L}_{\mathcal{T}}(0) \cdot \rho/2$; when this event happens, a location $\mathcal{R}$ is selected uniformly at random along the edges of $\mathcal{T}$, say on edge b at time s, which we denote as $\mathcal{R} \in b$ or $\mathcal{R} = (b, s)$. A new coalescence point $\mathcal{C}$ is selected by allowing the recombining lineage to coalesce at rate 1 with all the lineages present above time s (including b_r). We denote a coalescence point on edge b' at time u as $\mathcal{C} \in b'$ or $\mathcal{C} = (b', u)$. The next tree along the genome is then formed by pruning the subtree below the recombination point $\mathcal{R}$ and reattaching it at the chosen coalescence point $\mathcal{C}$. We write $\mathcal{R} \in \mathcal{B}(b)$ to mean that the recombination point is on one of the edges in $\mathcal{B}(b)$.

The difference with the spatial formulation of the CwR is that the coalescence point is restricted to be on the local tree $\mathcal{T}$, whereas under the CwR it could be placed on any edges of the ARG corresponding to the full sequence of trees to the left of the recombination position. In essence, the SMC' approximation disallows any recombination events that occur in non-ancestral material, making the process Markovian along the genome.

The difference between the SMC and SMC' models is that under the SMC', the coalescence point can be chosen on the same edge b containing the recombination point, above time s (so that recombinations can occur that do not change the tree topology or edge lengths), whereas events of this type are disallowed under the SMC.

S1.3 Probability that an edge is disrupted by a recombination event

Considering a fixed edge $b \in \mathcal{T}$, when a recombination event occurs, we would like to know the probability that b is affected by this recombination event. This includes both changes in the time-length of b and events that change the topology of the clade around b (we will say in these cases that b is *topologically* disrupted by the recombination). This can happen because either (1) the recombination point is on $b' \in \mathcal{B}(b)$ and the coalescence point is not on b', or (2) the recombination point is not on $\mathcal{B}(b)$ and the coalescence point is on b. The possible scenarios that do and do not disrupt b are illustrated in Figure S2.

We have, for a given tree $\mathcal{T}$,

$$\begin{aligned}
1 &= \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in \mathcal{B}(b)) + \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)) \\
&= \sum_{b' \in \mathcal{B}(b)} \left[\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b' | \mathcal{R} \in b') + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b' | \mathcal{R} \in b') \right] \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\
&\quad + \left[\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b | \mathcal{R} \notin \mathcal{B}(b)) \right] \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)).
\end{aligned}$$

The probability that b is disrupted by the recombination event is thus

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) = \sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b' \text{ and } \mathcal{C} \notin b') + \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b) \text{ and } \mathcal{C} \in b)$$

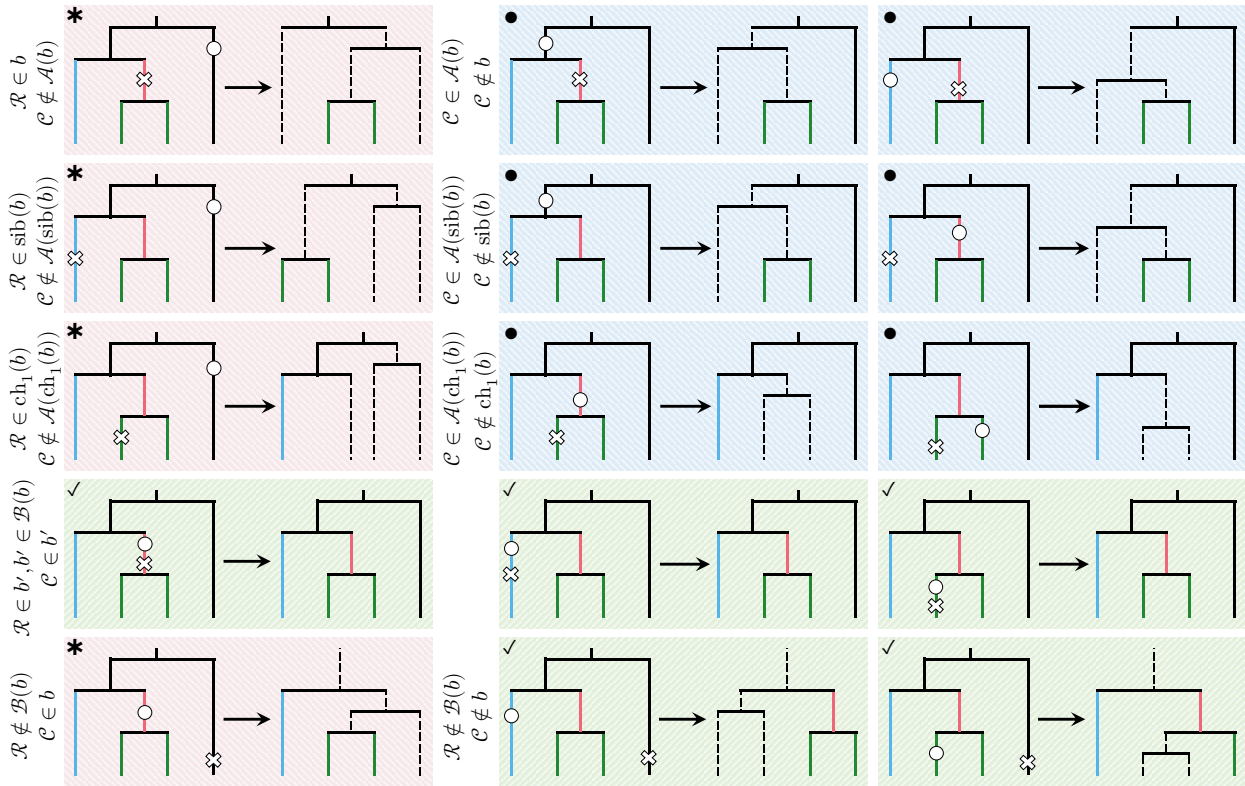

Figure S2: Possible events that do and do not disrupt edge b (shown in red); $\mathrm{sib}(b)$ is shown in blue, $\mathrm{ch}_1(b)$ and $\mathrm{ch}_2(b)$ in green. Recombination points are shown as crosses; coalescence points as circles. edges that are disrupted (or newly added) are shown as dashed lines. Events highlighted in green (marked with ticks) do not disrupt b. Events highlighted in red (marked with stars) disrupt the edge in terms of both edge length and topology (b is topologically disrupted); those highlighted in blue (marked with dots) disrupt b via changing only its time-length.

$$= \sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin b' | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') + \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)). \tag{S1}$$

We now calculate each of these probabilities in turn for an arbitrary edge $\beta \in \mathcal{T}$ under the SMC'.

S1.3.1 Probability recombination point is on edge β

Under the SMC', the recombination point is chosen uniformly at random along the edges of the tree. Thus, the probability that the recombination event happens on edge β is the ratio of the edge length to the total branch length of $\mathcal{T}$, so

$$\mathbb{P}_{\mathcal{T}}(\mathcal{R} \in \beta) = \frac{\bar{t}(\beta)}{L_{\mathcal{T}}(0)}, \tag{S2}$$

and

$$\mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(\beta)) = 1 - \sum_{b' \in \mathcal{B}(\beta)} \frac{\bar{t}(b')}{L_{\mathcal{T}}(0)}. \tag{S3}$$

S1.3.2 Probability coalescence point is not on β given recombination point is on β

The probability that, conditional on the recombination event happening on edge β at time s, the coalescence point is on β has been derived by Deng et al. (2021), which in our notation is as follows.

Proposition S1.1 (Deng et al. (2021), Proposition 1)**.** *Letting $k = n(s)$, so that T_k is the first coalescence time just above time s,*

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} = (\beta, s)) = \frac{1}{k} + e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj}, \tag{S4}$$

where

$$Q_{kk} := -\frac{1}{k} e^{-kT_k}, \tag{S5}$$

and

$$Q_{kj} = e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \frac{1}{j} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_j)} \right). \tag{S6}$$

Marginalising out the recombination time s, hence summing over $k = n(s)$ in (S4), gives the following.

Proposition S1.2 (Deng et al. (2021), Proposition 2)**.**

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \beta | \mathcal{R} \in \beta) = 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right), \tag{S7}$$

where

$$\widetilde{Q}^1(k) := \frac{1}{k} \left(T_k - T_{k+1} \right), \tag{S8}$$

and for $x, y, A, B \in \mathbb{Z}$, $x \geq k$, $2 \leq y \leq x$,

$$\widetilde{Q}^2(k, x, y, A, B) := \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=y}^{x} (Aj + B) \cdot Q_{kj}. \tag{S9}$$

The proofs, translated into our notation, are given in Sections S2.1 and S2.2.

S1.3.3 Probability coalescence point is on β given recombination point is not on $\mathcal{B}(\beta)$

We start by conditioning on the recombination time s to obtain the following.

Proposition S1.3. *Conditional on the recombination point $\mathcal{R}$ being at time s and on an edge outside the set $\mathcal{B}(\beta)$, with $k = n(s)$, the probability that β is disrupted is*

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \begin{cases} e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj} & s < t^{\downarrow}(\beta) \\ \frac{1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj} & t^{\downarrow}(\beta) \leq s < t^{\uparrow}(\beta) \\ 0 & \textit{otherwise.} \end{cases} \tag{S10}$$

with Q_{kk} and Q_{kj} as defined in (S5) *and* (S6)*, respectively.*

The proof is given in Section S2.3.

Let t_1, t_2, t_3, t_4 denote the event times $t^{\downarrow}(\mathrm{ch}_1(\beta)), t^{\downarrow}(\mathrm{ch}_2(\beta)), t^{\downarrow}(\mathrm{sib}(\beta)), t^{\downarrow}(\beta)$ sorted in increasing order, and define $t_0 := 0$ and $t_5 := t^{\uparrow}(\beta)$. Integrating out the recombination time in (S10), we have the following.

Proposition S1.4. *Conditional on the recombination point being on an edge outside the set $\mathcal{B}(\beta)$, the probability that β is disrupted is*

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta)) = \frac{1}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} \left\{ \sum_{k=n(t_1)+1}^{n} k \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \right.$$

$$
\begin{aligned}
&+ \sum_{k=n(t_2)+1}^{n(t_1)} (k-1)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&+ \sum_{k=n(t_3)+1}^{n(t_2)} (k-2)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&+ \sum_{k=n(t_4)+1}^{n(t_3)} \Big[\mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) < t^{\downarrow}(\beta))(k-3)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) \geq t^{\downarrow}(\beta))(k-1) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Big] \\
&+ \sum_{k=n(t_5)+1}^{n(t_4)} (k-2) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Bigg\}, \qquad \text{(S11)}
\end{aligned}
$$

with $\widetilde{Q}^1$ *and* $\widetilde{Q}^2$ *as defined in* (S8) *and* (S9)*, respectively.*

The proof is given in Section S2.4.

Substituting the expressions (S2), (S3), (S7) and (S11) into (S1) gives the desired probability that edge b is disrupted by the next recombination event.

S1.4 Probability that an edge is topologically disrupted by a recombination event

Most ARG reconstruction algorithms focus on identifying the presence of recombination events through finding patterns of mutations not consistent with tree-like evolution. This, in general, does not allow for the detection of recombination events that only change edge lengths (panels highlighted in blue in Figure S2). We therefore also calculate the probability that an edge b is topologically disrupted (corresponding to panels highlighted in red in Figure S2).

Theorem S1.1. *The probability that an edge b is topologically disrupted by a recombination event is given by*

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(b \textit{ topologically disrupted}) = &\sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(b') | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\
&+ \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)), \qquad \text{(S12)}
\end{aligned}
$$

with

$$
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta) | \mathcal{R} \in \beta) = 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} G_\beta(k),
$$

where, for $k \leq n(t^{\downarrow}(\mathrm{sib}(\beta)))$*,*

$$
G_\beta(k) := 2 \cdot \widetilde{Q}^1(k) + 2 \cdot \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) + \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\mathrm{par}(\beta))) + 1, 0, 1),
$$

and for $k > n(t^{\downarrow}(\mathrm{sib}(\beta)))$

$$
\begin{aligned}
G_\beta(k) := \widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\downarrow}(\mathrm{sib}(\beta))) + 1, 0, 1) + 2 \cdot \widetilde{Q}^2(k, n(t^{\downarrow}(\mathrm{sib}(\beta))), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
+ \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\mathrm{par}(\beta))) + 1, 0, 1),
\end{aligned}
$$

and $\widetilde{Q}^1$ *and* $\widetilde{Q}^2$ *are as defined in* (S8) *and* (S9)*, respectively.*

The proof is given in Section S2.5.

S1.5 Probability that a clade is disrupted by a recombination event

We now calculate the probability that a particular *clade* of edges is disrupted by the next recombination event, i.e. that the membership of sample nodes in the clade changes from one local tree to the next (but allowing for events that disrupt edges within the clade without changing the group of subtended samples). This can happen when a lineage within the clade recombines and coalesces outside the clade or its root edge, or if a lineage from outside the clade recombines and coalesces into the clade, as illustrated in Figure S3.

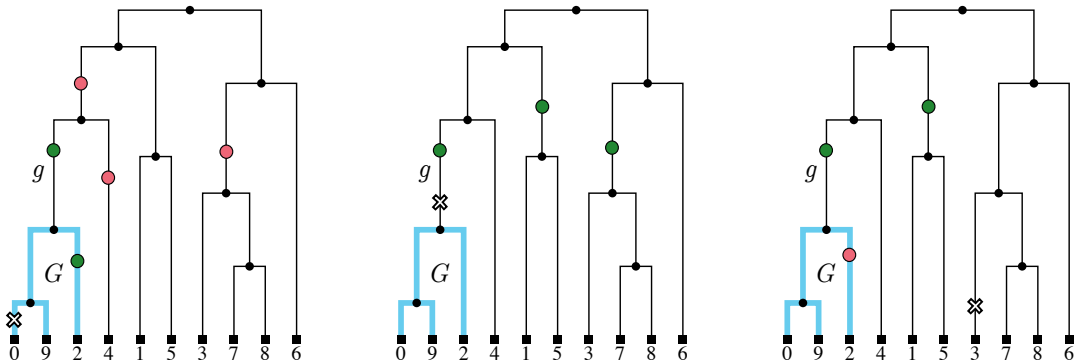

Figure S3: Recombination events that do and do not disrupt a clade. Clade G, subtended by edge g, contains samples $\{0, 2, 9\}$; edges belonging to G are shown in blue. In each tree, for the given recombination point (marked by a cross), red (resp. green) circles show examples of coalescence points that would (resp. would not) result in G being disrupted.

Let G be the set of edges subtended by an edge g, with clade MRCA time $t^{\downarrow}(g) = T_m$, $n_G(t)$ the number of lineages in clade G at time t, abusing notation to write $G \cup g := G \cup \{g\}$. Let $n_{G\cup g}(t)$ be the number of lineages in $G \cup g$ at time t and $L_G(t)$ the total branch length within the clade above time t. Then

$$\begin{aligned} 1 = \mathbb{P}(\mathcal{R} \in G) \cdot (\mathbb{P}(\mathcal{C} \notin G \cup g | \mathcal{R} \in G) + \mathbb{P}(\mathcal{C} \in G \cup g | \mathcal{R} \in G)) + \\ + \mathbb{P}(\mathcal{R} \in g) \\ + \mathbb{P}(\mathcal{R} \notin G \cup g) \cdot (\mathbb{P}(C \notin G | \mathcal{R} \notin G \cup g) + \mathbb{P}(C \in G | \mathcal{R} \notin G \cup g)) \,. \end{aligned}$$

Considering only the events that disrupt the clade, we obtain

$$\mathbb{P}(G \text{ disrupted}) = \mathbb{P}(\mathcal{R} \in G) \cdot \mathbb{P}(\mathcal{C} \notin G \cup g | \mathcal{R} \in G) + \mathbb{P}(\mathcal{R} \notin G \cup g) \cdot \mathbb{P}(C \in G | \mathcal{R} \notin G \cup g). \quad \text{(S13)}$$

Theorem S1.2. *The probability that a clade G is disrupted by a recombination event is*

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(G \textit{ disrupted}) = \frac{L_G(0)}{L_{\mathcal{T}}(0)} - \frac{1}{L_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g)+1)}^{n} \left[n_{G\cup g}(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G \cup g) \right] n_G(T_{k+1}) \\ + \frac{1}{L_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} (k - n_{G\cup g}(T_{k+1})) \left[n_G(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G) \right], \end{aligned} \quad \text{(S14)}$$

where

$$\widetilde{Q}^4(k, A) = \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=n(t^{\uparrow}(A))+1}^{k} n_A(T_{j+1}) Q_{kj},$$

taking $t^{\uparrow}(G \cup g) = t^{\uparrow}(g)$ and $t^{\uparrow}(G) = t^{\downarrow}(g)$.

The proof is given in Section S2.10.

S1.6 Change in total branch length of tree following a recombination event

Conditioning on the tree $\mathcal{T}$, we now consider the distribution of

$$C = L_{\mathcal{T}'}(0) - L_{\mathcal{T}}(0),$$

the change in total branch length following a single recombination event. First, considering the sign of the change, we have the following.

Proposition S1.5. *The probability of C being negative, zero, or positive is given by*

$$\mathbb{P}_{\mathcal{T}}(C<0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left((k-1)\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,n(t^{\uparrow}(b))+1,1,-1)\right) \tag{S15}$$

$$\mathbb{P}_{\mathcal{T}}(C=0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,n(t^{\uparrow}(b))+1,0,1)\right) \tag{S16}$$

$$\mathbb{P}_{\mathcal{T}}(C>0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^2(k,n(t^{\uparrow}(b)),2,1,0) + \widetilde{Q}^3(k)\right), \tag{S17}$$

respectively, where

$$\widetilde{Q}^3(k) = \frac{1}{k}\left(e^{-L_{\mathcal{T}}(T_k)} - e^{-L_{\mathcal{T}}(T_{k+1})}\right). \tag{S18}$$

The proof is given in Section S2.6.

To explore the distribution of the magnitude of the change in edge length (when this is non-zero), we derive an approximation of its density.

Proposition S1.6. *Conditional on the change in total branch length being non-zero, the density of C is given approximately by*

$$p^C_{\mathcal{T}}(c|C\neq 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b\in\mathcal{T}} \xi_b(c), \tag{S19}$$

where $\xi_b(c)$ is given by

$$\begin{cases} (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))}\cdot \sum_{k=l}^{n(t^{\downarrow}(b))} \widetilde{P}^1_{kl} - \frac{e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1}{l-1}\right) & t^{\uparrow}(b)\neq T_2,\ -\bar{t}(b)\leq c<0 \\ (n(c+2T_2-T_3)-1)\left(e^{-(l-1)(c+2T_2-T_3)}\cdot \sum_{k=l}^{n(t^{\downarrow}(b))} \widetilde{P}^1_{kl} - \frac{e^{(l-1)(T_l-c-2T_2+T_3)}-1}{l-1}\right) & t^{\uparrow}(b)=T_2, \\ & t^{\downarrow}(b)+T_3-2T_2\leq c<2(T_3-T_2) \\ \frac{1}{2}(n(c/2+T_2)-1)\left(e^{-(c/2+T_2)}\cdot \sum_{k=l}^{n(t^{\downarrow}(b))} \widetilde{P}^1_{kl} - e^{T_l-c/2-T_2}+1\right) & t^{\uparrow}(b)=T_2,\ 2(T_3-T_2)\leq c<0 \\ n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\cdot \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{P}^2_k & 0<c\leq T_2-t^{\uparrow}(b) \\ \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\cdot \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{P}^3_k & c>T_2-t^{\uparrow}(b) \\ 0 & \textit{otherwise.} \end{cases}$$

and

$$\widetilde{P}^1_{kl} := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(l\cdot T_l - k\cdot T_k + L_{\mathcal{T}}(T_l) - L_{\mathcal{T}}(T_k)\right)$$

$$\widetilde{P}^2_k := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(t^{\uparrow}(b) - k\cdot T_k - L_{\mathcal{T}}(T_k) + L_{\mathcal{T}}(t^{\uparrow}(b))\right)$$

$$\widetilde{P}^3_k := \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)\exp\left(t^{\uparrow}(b) - k\cdot T_k - L_{\mathcal{T}}(T_k)\right).$$

The proof is given in Section S2.7. This is an approximation rather than an exact result under the SMC', since it assumes that after conditioning on the coalescence point not being on the same branch as the recombination point, the coalescent dynamics follow the SMC model (we find this to give a very close approximation, and simplifies our calculations).

S1.7 Change in tree height following a recombination event

The height of the tree, $H(\mathcal{T}) = T_2$, can change following a recombination event if (1) the coalescence point is above T_2, in which case a new root is formed and the tree height increases, or (2) the recombination happens on one of the two lineages descending from the MRCA, then the tree height can either increase or decrease. Let the set $\mathcal{M} := \{\mathrm{ch}_1(b_r), \mathrm{ch}_2(b_r)\}$ contain the two edges descending from the MRCA, and let $H = H(\mathcal{T}') - H(\mathcal{T})$ be the magnitude of the change in height. Then we have the following.

Proposition S1.7. *Conditional on $\mathcal{T}$, the probability of the change in tree height being negative, zero, or positive is given by*

$$\mathbb{P}_{\mathcal{T}}(H < 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left\{ (k-1)\widetilde{Q}^1(k) - \widetilde{Q}^2(k,k,2,0,1) - \widetilde{Q}^3(k) \right\} \tag{S20}$$

$$\mathbb{P}_{\mathcal{T}}(H = 0) = \frac{1}{L_{\mathcal{T}}(0)} \left\{ \sum_{b \in \mathcal{M}} \sum_{k=2}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k,k,2,0,1) \right) + \sum_{b \notin \mathcal{M}} \left(\bar{t}(b) - \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k) \right) \right\} \tag{S21}$$

$$\mathbb{P}_{\mathcal{T}}(H > 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k), \tag{S22}$$

where $\widetilde{Q}^1$, $\widetilde{Q}^2$ and $\widetilde{Q}^3$ are as defined in (S8), (S9) *and* (S18), *respectively.*

The proof is given in Section S2.9.

S1.8 Distribution of edge span

Under the SMC', for a given edge b, the distribution of its span can be characterised by considering the rate at which edge-disrupting recombination events arrive as we move left-to-right along the genome. However, the instantaneous rate at which b is disrupted at position τ may not be the same as that at position $\tau' > \tau$, due to the effect of other recombination events that might occur between τ and τ'. Thus, the span of an edge is the waiting time to the next edge-disrupting recombination event, but the rate at which this happens is inhomogeneous along the genome (and is, in fact, itself random).

Similarly to Deng et al. (2021), however, we find that if a recombination event between adjacent trees $\mathcal{T}$ and $\mathcal{T}'$ does not disrupt b, then

$$P_{\mathcal{T}}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}}(0) \approx P_{\mathcal{T}'}(b \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}'}(0),$$

where $P_{\mathcal{T}}(b \text{ disrupted})$ is given by (S1), based on simulation results (Section S1.11). Thus, an approximation to the distribution of edge span can be constructed by assuming that the rate at which edge-disrupting recombination events arrive *is* homogeneous along the genome, which is equivalent to assuming that recombination events that do not disrupt the edge b also do not change the local tree: so if $\mathcal{T}$ is the local tree at position $d^{\leftarrow}(b)$, after each recombination event the newly formed local tree is $\mathcal{T}' = \mathcal{T}$. Recombination events occur as a Poisson process along the genome with rate $\mathcal{L}_{\mathcal{T}}(0) \cdot \rho/2$, allowing us to thin the process by multiplying this rate by the probability that the event is edge-disrupting, thereby offering a tractable approximation for the arrival of edge-disrupting recombination events. Then conditional on $d^{\leftarrow}(b)$, the edge span $d^{\rightarrow}(b) - d^{\leftarrow}(b)$ is distributed as the waiting time to the first event in a Poisson process with rate

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2}.$$

That is,

$$d^{\rightarrow}(b) - d^{\leftarrow}(b) \mid d^{\leftarrow}(b) \sim \mathrm{Exp}\left(\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2} \right),$$

or, by rescaling,

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho}{2}(d^{\rightarrow}(b) - d^{\leftarrow}(b)) \mid d^{\leftarrow}(b) \sim \text{Exp}(1).$$

Analogously, if the recombination rate is not constant along the genome, with the population-scaled recombination rate at position w given by $\rho(w)/2$, then the intensity of the process at position w is instead given by

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(w)}{2},$$

and we have

$$\mathbb{P}_{\mathcal{T}}(b \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b)}^{d^{\rightarrow}(b)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(b) \sim \text{Exp}\,(1)\,. \tag{S23}$$

The quality of this approximation can be verified by simulation, using the probability integral transform as follows. For the i-th edge $b_i \in \{b_1, \ldots, b_m\}$ of a simulated ARG, take $\mathcal{T}$ to be the local tree at position $d^{\leftarrow}(b_i)$, compute q_i as

$$q_i := \mathbb{P}_{\mathcal{T}}(b_i \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b_i)}^{d^{\rightarrow}(b_i)} \frac{\rho(w)}{2} dw,$$

and let $p_i = 1 - e^{-q_i}$. Then a Q-Q plot can be constructed by plotting the ordered quantities $p_{(1)} \leq \ldots \leq p_{(m)}$ against the corresponding quantiles of the uniform distribution $\frac{1}{1+m}, \ldots, \frac{m}{1+m}$. If the approximation fits well, the points should lie on the diagonal. A Kolmogorov–Smirnov (K–S) goodness of fit test can be used to test the null hypothesis that the computed p_i values are uniformly distributed.

S1.8.1 Considering only topology-disrupting events

If we were to consider only events that topologically disrupt the edge, we instead have the approximation

$$\mathbb{P}_{\mathcal{T}}(b \text{ topologically disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \int_{d^{\leftarrow}(b)}^{d^{\rightarrow}(b)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(b) \sim \text{Exp}\,(1)\,. \tag{S24}$$

A similar procedure to that described above can be used to check goodness of fit.

S1.8.2 Conditioning on edge having at least one mutation

ARG reconstruction algorithms utilise mutations to infer changes in local tree topologies due to recombination, so it may be of interest to consider only edges in reconstructed ARGs that are supported by at least one mutation. Suppose that mutations occur as a Poisson process along the edges with constant rate θ, and the recombination rate at position x is $\rho(x)/2$. Conditional on the left endpoint of the given edge $d^{\leftarrow}$, let D be its right endpoint, which using (S24) has the density

$$p_{\mathcal{T}}^{D}(\delta | d^{\leftarrow}) = \lambda(d^{\leftarrow}) \exp\left(- \int_{d^{\leftarrow}}^{\delta} \lambda(x) dx \right),$$

where

$$\lambda(x) = \mathbb{P}(b \text{ topologically disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(x)}{2}.$$

Let $\bar{t}$ be the time-length and M the number of mutations on the edge. Then the conditional distribution of D is given by

$$P_{\mathcal{T}}^{D}(d^{\rightarrow} | M > 0, d^{\leftarrow}) = \frac{\int_{d^{\leftarrow}}^{d^{\rightarrow}} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p_{\mathcal{T}}^{D}(\delta | d^{\leftarrow}) d\delta}{\int_{d^{\leftarrow}}^{\infty} \mathbb{P}(M > 0 | \delta, d^{\leftarrow}) p_{\mathcal{T}}^{D}(\delta | d^{\leftarrow}) d\delta},$$

with

$$\mathbb{P}(M > 0|\delta, d^{\leftarrow}) = 1 - \exp(-\theta\bar{t}(\delta - d^{\leftarrow})).$$

We have

$$\mathbb{P}(M > 0|\delta, d^{\leftarrow})p^D_{\mathcal{T}}(\delta|d^{\leftarrow}) = \lambda(d^{\leftarrow})\exp\left(\int_{d^{\leftarrow}}^{\delta}\lambda(x)dx\right) - \lambda(d^{\leftarrow})\exp\left(-\theta\bar{t}(\delta - d^{\leftarrow}) - \int_{d^{\leftarrow}}^{\delta}\lambda(x)dx\right).$$

Assuming that the recombination map is piecewise constant, we split the part of the genome to the right of $d^{\leftarrow}$ into portions where the recombination rate is constant between the (ordered) breakpoints $d^{\leftarrow} =: w_0 < w_1 < w_2 < \ldots < w_k < \ldots$, adding an extra breakpoint $w_k := d^{\rightarrow}$. Then we can write

$$\begin{aligned}
\int_{d^{\leftarrow}}^{\infty}\mathbb{P}(M > 0|\delta, d^{\leftarrow})p^D_{\mathcal{T}}(\delta|d^{\leftarrow})d\delta &= 1 - \lambda(d^{\leftarrow})\int_{d^{\leftarrow}}^{\infty}\exp\left(-\theta\bar{t}(\delta - d^{\leftarrow}) - \int_{d^{\leftarrow}}^{\delta}\lambda(x)dx\right)d\delta \\
&= 1 - \lambda(d^{\leftarrow})\sum_{i=0}^{\infty}\exp\left(-\int_{d^{\leftarrow}}^{w_i}\lambda(x)dx\right)\int_{w_i}^{w_{i+1}}\exp\left(-\theta\bar{t}(\delta - d^{\leftarrow}) - \int_{w_i}^{\delta}\lambda(x)dx\right)d\delta \\
&= 1 - \sum_{i=0}^{\infty}S_i,
\end{aligned}$$

where, by integrating,

$$S_i = \frac{\lambda(d^{\leftarrow})\exp\left(-\int_{d^{\leftarrow}}^{w_i}\lambda(x)dx\right)\left[\exp(-\theta\bar{t}(w_i - d^{\leftarrow})) - \exp(-\theta\bar{t}(w_{i+1} - d^{\leftarrow}) - \lambda(w_i)(w_{i+1} - w_i))\right]}{\theta\bar{t} + \lambda(w_i)}.$$

Similarly,

$$\int_{d^{\leftarrow}}^{d^{\rightarrow}}\mathbb{P}(M > 0|\delta, d^{\leftarrow})p^D_{\mathcal{T}}(\delta|d^{\leftarrow})d\delta = 1 - \exp\left(-\Lambda(b)\right) - \sum_{i=0}^{k-1}S_i,$$

where

$$\Lambda(b) = \int_{d^{\leftarrow}}^{d^{\rightarrow}}\lambda(x)dx.$$

Combining, we have

$$P^D_{\mathcal{T}}(d^{\rightarrow}|M > 0, d^{\leftarrow}) = \frac{1 - \exp(-\Lambda(b)) - \sum_{j=0}^{k-1}S_j}{1 - \sum_{j=0}^{\infty}S_j}. \tag{S25}$$

Note that in the limit $\theta \to \infty$, this reduces to $1 - \exp\left(-\Lambda(b)\right)$, as expected (since this effectively removes the conditioning).

For the case where the recombination rate is constant along the genome, with $\rho(x) = \rho$ and $\lambda(x) = \lambda$, we obtain

$$P^D_{\mathcal{T}}(d^{\rightarrow}|M > 0, d^{\leftarrow}) = 1 - \frac{\theta\bar{t} + \lambda}{\theta\bar{t}}\exp(-\lambda(d^{\rightarrow} - d^{\leftarrow})) + \frac{\lambda}{\theta\bar{t}}\exp(-(\lambda + \theta\bar{t})(d^{\rightarrow} - d^{\leftarrow})).$$

The p-values for each edge of a simulated ARG can now be computed by evaluating the cdf (S25) for the given values of $d^{\leftarrow}$ and $d^{\rightarrow}$, and the Q-Q plot can again be constructed by plotting these against the corresponding quantiles of the uniform distribution. However, this potentially requires summing over a large number of increments of the recombination map. Instead, we propose to approximate (S25) by

$$\widetilde{P}^D_{\mathcal{T}}(d^{\rightarrow}|M > 0, d^{\leftarrow}) := 1 - \frac{\theta\bar{t} + \Lambda(b)}{\theta\bar{t}}\exp(-\Lambda(b)) + \frac{\Lambda(b)}{\theta\bar{t}}\exp(-(\Lambda(b) + \theta\bar{t}d)).$$

We find this to be a very close match to the exact distribution based on simulations with human-like parameters, while being very fast to compute.

S1.9 Distribution of clade span

Similar approximations to those employed when investigating the distribution of edge span along the genome can be used for the distribution of the waiting time until a clade G is broken up by a recombination event (that is, when the clade either gains or loses one or more samples as a consequence of recombination). Consider an inhomogeneous Poisson process with intensity at position w given by

$$\mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G)}}(G \text{ disrupted}) \cdot L_{\mathcal{T}}(0) \cdot \frac{\rho(w)}{2}, \tag{S26}$$

conditional on the local tree at $d^{\leftarrow}(G)$ (defined as the leftmost position along the genome where the clade arises). Again through rescaling time, we have

$$\mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G)}}(G \text{ disrupted}) \cdot L_{\mathcal{T}_{d^{\leftarrow}(G)}}(0) \cdot \int_{d^{\leftarrow}(G)}^{d^{\rightarrow}(G)} \frac{\rho(w)}{2} dw \mid d^{\leftarrow}(G) \sim \text{Exp}\,(1)\,. \tag{S27}$$

Thus, for each clade G_i in a simulated ARG, we can calculate its left and right endpoints $d^{\leftarrow}(G_i)$ and $d^{\rightarrow}(G_i)$, respectively. Letting

$$q_i := \mathbb{P}_{\mathcal{T}_{d^{\leftarrow}(G_i)}}(G_i \text{ disrupted}) \cdot L_{\mathcal{T}_{d^{\leftarrow}(G_i)}}(0) \cdot \int_{d^{\leftarrow}(G_i)}^{d^{\rightarrow}(G_i)} \frac{\rho(w)}{2} dw, \tag{S28}$$

the quality of the approximation can again be checked using a Q-Q plot as described above.

S1.9.1 Adjusting for varying population size

Given a population size function $N(t)$, $t \geq 0$, let

$$\Lambda(t) = \int_0^t \frac{1}{N(t)} dt.$$

Let $\widetilde{L}_G(0)$ be the total length of branches in G measured in generations, and $\widetilde{L}_{\widetilde{\mathcal{T}}}(0)$ the total branch length of $\mathcal{T}$ measured in generations. Then we have

$$\mathbb{P}(\mathcal{R} \in G) = \frac{\widetilde{L}_G(0)}{\widetilde{L}_{\mathcal{T}}(0)}, \quad \mathbb{P}(\mathcal{R} \notin G \cup g) = \frac{\widetilde{L}_T(0) - \widetilde{L}_G(0) - \Lambda^{-1}(t^{\uparrow}(g)) + \Lambda^{-1}(t^{\downarrow}(g))}{\widetilde{L}_{\mathcal{T}}(0)}.$$

Conditional on the recombination point being on a branch within G, the density of the recombination time is

$$p_S(s|\mathcal{R} \in G) = \begin{cases} \frac{n_G(s) N(\Lambda^{-1}(s))}{\widetilde{L}_G(0)} & s \leq T_m \\ 0 & \text{otherwise,} \end{cases}$$

and similarly,

$$p_S(s|\mathcal{R} \notin G \cup g) = \begin{cases} \frac{(n(s) - n_{G \cup g}(s)) N(\Lambda^{-1}(s))}{\widetilde{L}_T(0) - \widetilde{L}_G(0) - \Lambda^{-1}(t^{\uparrow}(g)) + \Lambda^{-1}(t^{\downarrow}(g))} & s \leq T_2 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose now that $N(t)$ is piecewise constant, and for each $2 \leq k \leq n$, write

$$[T_{k+1}, T_k] = [\tau_k^0, \tau_k^1] \cup \ldots \cup [\tau_k^{\omega(k)-1}, \tau_k^{\omega(k)}],$$

where $\tau_0 := T_{k+1}$, $\tau_k^{\omega(k)} := T_k$, and $N(\tau) = N_k^v$ if $\tau \in [\tau_k^{v-1}, \tau_k^v]$. That is, $\omega(k)$ is the minimal number of (disjoint) intervals where the population size is piecewise constant, while there are k lineages in the tree. Then following similar calculations as in the proof of Theorem S1.2, we have

$$\mathbb{P}(\mathcal{C} \notin G \cup g | \mathcal{R} \in G) \mathbb{P}(\mathcal{R} \in G) = \frac{\widetilde{L}_G(0)}{\widetilde{L}_{\mathcal{T}}(0)} - \frac{1}{\widetilde{L}_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} n_G(T_{k+1}) \cdot$$

$$\cdot \left[n_{G\cup g}(T_{k+1})\widetilde{Q}^1(k)\frac{\widetilde{T}_k - \widetilde{T}_{k+1}}{T_k - T_{k+1}} + \frac{1}{k}\left(\sum_{j=n(t^{\uparrow}(g))+1}^{k} n_{G\cup g}(T_{j+1})Q_{kj} \right) \sum_{i=1}^{\omega(k)} N_k^i \left(e^{k\tau_k^i} - e^{k\tau_k^{i-1}} \right) \right],$$

where $\widetilde{T}_j = \Lambda^{-1}(T_j)$, and

$$\mathbb{P}(\mathcal{C} \in G | \mathcal{R} \notin G \cup g)\mathbb{P}(\mathcal{R} \notin G \cup g) = \frac{1}{\widetilde{L}_{\mathcal{T}}(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} [k - n_{G\cup g}(T_{k+1})] \cdot$$
$$\cdot \left[n_G(T_{k+1})\widetilde{Q}^1(k)\frac{\widetilde{T}_k - \widetilde{T}_{k+1}}{T_k - T_{k+1}} + \frac{1}{k}\left(\sum_{j=n(t^{\downarrow}(g))+1}^{k} n_G(T_{j+1})Q_{kj} \right) \sum_{i=1}^{\omega(k)} N_k^i \left(e^{k\tau_k^i} - e^{k\tau_k^{i-1}} \right) \right],$$

which can be substituted into (S13), and in turn into the expressions in Section S1.9, to give the corresponding approximation to the genomic span of G under an arbitrary piecewise constant population size model. This can also be applied to an arbitrary population size model, through averaging the population size over a suitable time grid and thus approximating it with a piecewise constant function.

S1.10 Quality of approximation to the distribution of edge span

We first assess the quality of the approximation derived in Section S1.8 by simulating an ARG under the SMC' and checking if the simulated edge spans follow (S23), by using the procedure described in Section S1.8. The simulation parameters are given in Section 4.6.1 (main text), and we sampled 10 000 edges from each ARG (uniformly at random) for testing, to speed up computation. The corresponding Q-Q plots are shown in Figure S4 (blue points). The points adhere very closely to the diagonal, demonstrating that the approximation provides an excellent fit. The K–S p-values of 0.31 (left panel) and 0.75 (right panel) also suggest good agreement. Grouping edges by their depth (the number of edges on the way to the MRCA) or clade size (number of samples subtended by the edge) in the tree and constructing Q-Q plots for each group also did not reveal significant deviation from the diagonal (Figure S11), suggesting that the approximation holds for all edges.

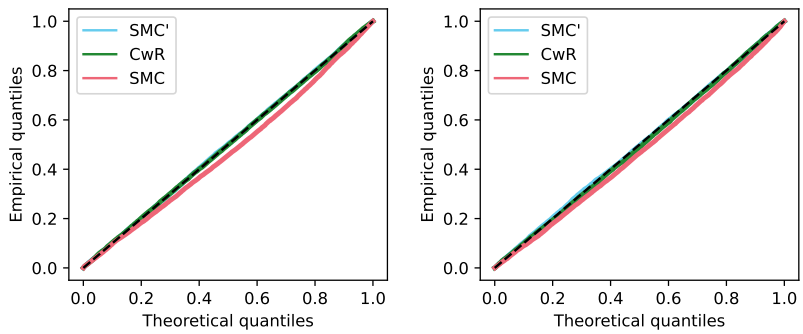

Figure S4: Q-Q plots using (S23) computed from ARGs simulated using the SMC' (blue), CwR (green), and SMC (red) models with $n = 100$. Note the blue and green points closely overlap, and overlay the diagonal. Left panel: dataset 1 parameters; right panel: dataset 2 parameters. Dashed line: diagonal from $(0, 0)$ to $(1, 1)$.

S1.11 Effects of recombination on a local tree

The very good quality of the approximation above can be understood by considering the effects of a recombination event on properties of the local tree.

For a given edge b_i that exists in local trees $\mathcal{T}_{(1)}, \ldots, \mathcal{T}_{(k)}$, let

$$f(b_i, l) := P_{\mathcal{T}_{(l)}}(b_i \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}_{(l)}}(0),$$

for $1 \leq l \leq k$. Let

$$H_1^l(b_i) := f(b_i, l+1)/f(b_i, l),$$

for $1 \leq l \leq k-1$, and

$$H_2(b_i) = \max_l f(b_i, l)/\min_l f(b_i, l).$$

We calculate these quantities for a uniform random sample of 1 000 edges from an ARG simulated using dataset 1 parameters in Section 4.6.1 (Main Text); the corresponding histograms are shown in Figure S5. These suggest that the quantity $P_{\mathcal{T}}(b_i \text{ disrupted}) \cdot \mathcal{L}_{\mathcal{T}}(0)$ stays relatively conserved following each recombination event, even if individually $P_{\mathcal{T}}(b_i \text{ disrupted})$ and $\mathcal{L}_{\mathcal{T}}(0)$ may vary more significantly. This is similar to the findings of Deng et al. (2021, Figure 5).

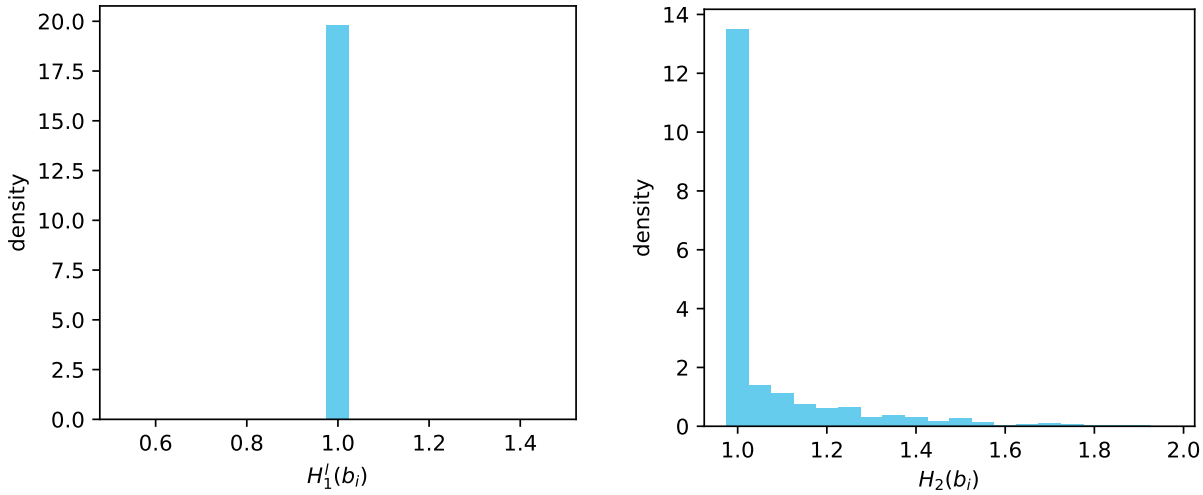

Figure S5: Histograms of $H_1^l(b_i)$ and $H_2(b_i)$ for a simulated ARG.

To calculate a Monte Carlo estimate of the marginal probability that the change in total branch length is negative, zero, or positive, we average over local trees simulated using msprime under the SMC' model. The results are shown in Figure S6. Recombination has a stabilising effect on the total branch length of local trees: when the total branch length is small (resp. large), the probability that the recombination event will increase the total branch length increases (resp. decreases).

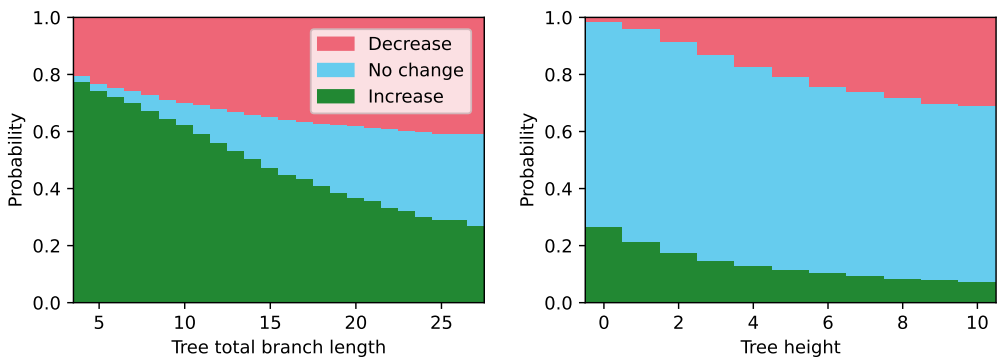

Figure S6: Mean probability of change in total branch length (left) and tree height (right) being negative (red, top stack), zero (blue, middle stack) or positive (green, bottom stack). Trees were simulated and binned according to total branch length (left) or height (right), with 100 trees simulated per bin, and sample size $n = 100$. For each tree, probabilities were calculated using (S15)-(S17), the stacked bar plot shows the mean probabilities for each bin.

Similarly, we can average over trees simulated under the SMC' model to estimate the marginal probability that the change in tree height following a recombination event is negative, zero, or positive. The results are shown in Figure S6 (right panel). As with total branch length, recombination

tends to increase (resp. decrease) the tree height with higher probability when the tree height is small (resp. large).

Figure S7 shows the density (S19) for three simulated trees with varying total branch lengths $L_{\mathcal{T}}(0)$, for $n = 10$ and $n = 100$. In both cases, the density is concentrated around zero and skewed to the left (resp. right) when the total branch length is small (resp. large); it is roughly symmetric about zero for middling values of $L_{\mathcal{T}}(0)$.

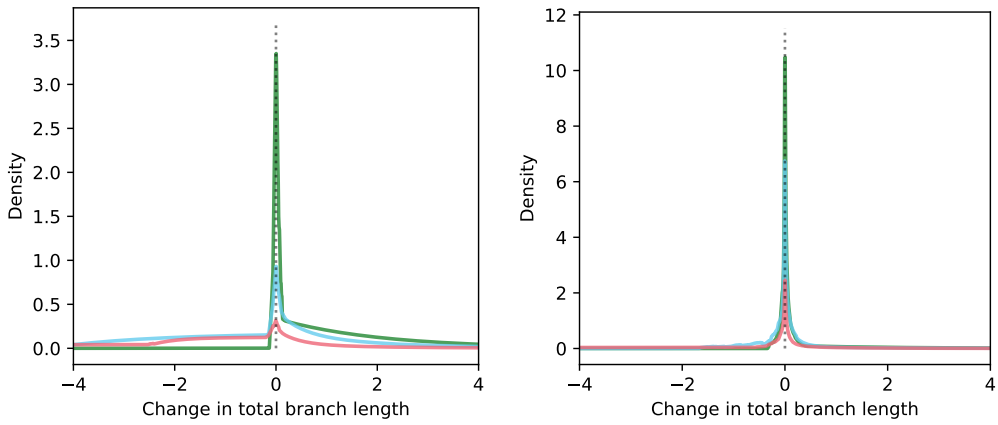

Figure S7: Density of change in total branch length (S19) for three simulated trees, with $n = 10$ (left panel) $n = 100$ (right panel). Left: trees have total branch length 1 (green), 6 (blue) and 24 (red). Right: trees have total branch length 5 (green), 10 (blue) and 28 (red).

These results shed light on why, despite the strong assumption that recombination events that do not disrupt the given edge also do not change the rest of the local tree, our approximation to the distribution of edge span gives an extraordinarily close fit for data simulated under the SMC' model. For an edge that is close to the leaves, the probability that the edge is disrupted by the next recombination event is very small (Figure 2). Thus, many recombination events will occur before this edge is disrupted. As can be seen in Figure S6, recombination has the effect of stabilising the total branch length, with events causing an increase (decrease) in total branch length being more likely if the current total branch length is relatively small (large). Thus, it seems the fluctuations in total branch length average out and do not significantly affect the overall rate of edge-disrupting recombination events. On the other hand, for an edge that is close to the root of the tree, per Figure 2 the probability of the edge being disrupted by the next recombination is relatively high. Thus, a relatively small number of recombination events are likely to occur before they affect the given edge. As illustrated in Figure S7, when a recombination changes the total branch length of the tree, the magnitude of this change is concentrated around 0. Thus, the effect of recombination on the rest of the tree does not appear to significantly affect the probability that the edge is disrupted.

S1.12 Comparison of simulation models

We next simulate ARGs under the CwR and under the SMC, with the same two parameter settings given in Section 4.6.1 (main text) and again compare the resulting edge spans to (S23). For the CwR, the span of an edge is taken to be the sum of all the genomic intervals where that edge appears in the local tree (to account for the presence of recombination events that occur in non-ancestral material). Figure S4 shows that the approximation is an excellent fit to the CwR (green points), with K–S p-values of 0.06 (left panel) and 0.94 (right panel). This suggests that the distribution of edge spans under the SMC' and that under the CwR are remarkably close.

Under the SMC (red points), edge spans are shorter in general, with points falling below the diagonal (with both K–S p-values < 0.001). This is due to the model disallowing recombination events that do not change the local tree, so edges are more frequently disrupted by recombination.

S1.13 Detection of local recombination suppression: Test 1

Given an ARG, for each clade G_i, we calculate its left and right endpoints $d^{\leftarrow}(G_i)$ and $d^{\rightarrow}(G_i)$, and would like to estimate the probability of observing a clade span greater than $d^{\rightarrow}(G_i) - d^{\leftarrow}(G_i)$ via $p_i = e^{-q_i}$, with q_i as defined in (S28). This is a one-sided p-value, and we test whether G_i has a significantly longer span than otherwise expected by comparing this against a Bonferroni-corrected significance threshold (0.05 divided by the total number of tested clades). Simulation studies confirm that for ARGs simulated under the SMC' model without inversions, these p-values are approximately uniformly distributed (Figure 4, left panel, blue points), as expected.

S1.13.1 Reconstructed ARGs

As can be seen from Figure 4 (left panel), similarly to edge span, the distribution of clade span in ARGs reconstructed using Relate, tsinfer/tsdate and ARG-Needle is skewed, with clade span generally overestimated by these methods. Thus, directly applying the test as described above will lead to a high false positive rate. We now describe a correction which can be applied to counteract two main problematic features of reconstructed ARGs, focusing particularly on Relate (due to the presence of polytomies for tsinfer/tsdate being difficult to correct for, and the large bias seen with ARG-Needle which both under- and overestimates clade span).

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ be a list of all clades in the reconstructed ARG. The first issue is that due to a lack of mutations around the leftmost and rightmost endpoints of a clade, Relate may overestimate its span, causing false positives. To correct for this, we proceed as follows. Suppose that the root edges of a clade G_i in trees $\mathcal{T}_{d^{\leftarrow}(G_i)}, \ldots, \mathcal{T}_{d^{\rightarrow}(G_i)}$ have mutations at positions $m_1^i < m_2^i < \ldots < m_K^i$. We (1) remove from $\mathcal{G}$ all clades with fewer than three mutations in total and fewer than M mutations per kb on average, and (2) measure an adjusted clade span $\widetilde{d}$ using the positions of the leftmost and rightmost mutations that support the given clade. That is, we define

$$\widetilde{d}^{\leftarrow}(G_i) = m_1^i,\ \widetilde{d}^{\rightarrow}(G_i) = m_K^i.$$

The second issue is that the clade carrying the inversion may not be supported by mutations uniformly along the inverted region, causing it to appear and disappear multiple times in quick succession in the reconstructed ARG, which can cause false negatives. We correct for this by "merging" pairs of clades that are nearby on the genome (in terms of genetic distance, to allow for varying recombination rates). For two clades $G_i, G_j \in \mathcal{G}$ that have identical sets of sample descendants and are less than L cM apart, that is

$$100 \cdot \int_{d^{\rightarrow}(G_i)}^{d^{\leftarrow}(G_j)} \rho(w) dw < L,$$

form $G_{i,j} := G_i$ but setting $d^{\leftarrow}(G_{i,j}) := d^{\leftarrow}(G_i)$, $d^{\rightarrow}(G_{i,j}) := d^{\rightarrow}(G_j)$, $\widetilde{d}^{\leftarrow}(G_{i,j}) := \widetilde{d}^{\leftarrow}(G_i)$, $\widetilde{d}^{\rightarrow}(G_{i,j}) := \widetilde{d}^{\rightarrow}(G_j)$, and update $\mathcal{G}$ as

$$\mathcal{G} = \{G_{i,j}\} \cup \mathcal{G} \backslash \{G_i, G_j\}.$$

We apply this to all pairs of clades in $\mathcal{G}$ iteratively, until no more clades can be merged together.

For each clade in the reduced list $G_i \in \mathcal{G}$, we thus calculate an adjusted version of (S28):

$$\widetilde{q}_i := \mathbb{P}_{\mathcal{T}_{\widetilde{d}^{\leftarrow}(G_i)}}(G_i \text{ disrupted}) \cdot L_{\mathcal{T}_{\widetilde{d}^{\leftarrow}(G_i)}}(0) \cdot \int_{\widetilde{d}^{\leftarrow}(G_i)}^{\widetilde{d}^{\rightarrow}(G_i)} \frac{\rho(w)}{2} dw, \tag{S29}$$

again taking $\widetilde{p}_i = e^{-\widetilde{q}_i}$. Note that with the above definitions, this can be computed even though the clades in the reduced list will now not necessarily exist (with the re-defined spans) in the ARG itself. We thus obtain adjusted p-values for each clade, applying a significance threshold of $0.05/N$, where N is the original number of clades in the reconstructed ARG.

We apply this to the ARG reconstructed using Relate for the data simulated using SLiM (as described in Section 4.6.2), setting $L = 0.01$ and $M = 0.05$. The resulting Q-Q and p-value plots are

shown in the top row of Figure S15, showing that the correction brings the points on the Q-Q plot very close to the diagonal, and there are three significant clades (all of which overlap the inverted region, and the clade spanning the entire inverted region remains a significant outlier with the lowest p-value). The equivalent plots for a simulation with no inversion are shown in the bottom row of Figure S15, demonstrating that the p-values are approximately uniformly distributed and there are no false positives.

S1.13.2 Test error rates

We examined the performance of the test by applying it to ARGs simulated using SLiM with the parameters given in Section 2.4.1, varying the length of the inverted region from 0 to 200kb (100 ARGs in each case). Defining positive detection as there being at least one clade within the inverted region with a significant p-value, the resulting ROC curve is shown in Figure S8 (left panel). Fixing the false positive rate at 5% (corresponding roughly to one false positive per 100Mb) gives the confusion matrix shown in Table 1a, demonstrating very high sensitivity. For each inversion length, out of all the clades with significant p-values across the simulations, a high percentage lie within the inverted region.

Reconstructing an ARG using Relate for each simulated dataset and applying the adjustments described in Section 2.4.2 gives the ROC curve shown in Figure S8 (right panel); the results in Table 1b show high sensitivity is maintained for inversions longer than around 100kb. These results demonstrate very good performance in detecting the presence of inversions, as well as pinpointing the candidate clade and its position along the genome.

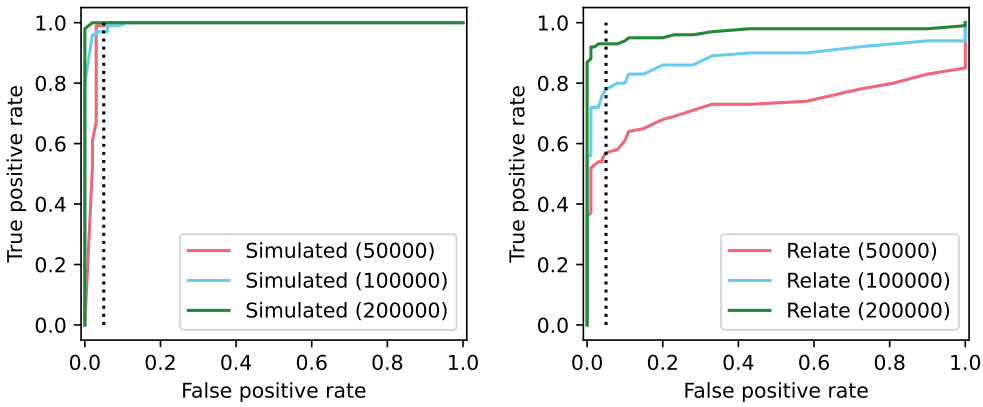

Figure S8: ROC curve for inversion detection test, based on 100 simulations for each given length of the inverted region. Left: using the simulated ARGs; right: ARGs reconstructed using Relate. Dotted line corresponds to a false positive rate of 5% (false positive being defined as an ARG simulated with no inversion but having at least one significant clade). Colours correspond to the different inversion lengths.

(a)

		Inversion length (kb)			
		200	100	50	0
Inv. detected	+	100	97	99	5
	−	0	3	1	95
Clades tested		3.3m	3.4m	3.4m	3.4m
Significant clades		201	144	127	5
Within inv. region		96%	92%	97%	0%

(b)

		Inversion length (kb)			
		200	100	50	0
Inv. detected	+	93	78	57	5
	−	7	22	43	95
Clades tested		116k	116k	117k	122k
Significant clades		227	123	79	5
Within inv. region		96%	97%	96%	0%

Table 1: Confusion matrices and results summaries for inversion detection test, based on 100 simulations for each given length of the inverted region: using (a) the simulated ARGs, (b) ARGs reconstructed using Relate.

S1.14 Detection of local recombination suppression: Test 2

Under our approximation, recombination events arrive as an inhomogeneous Poisson process along the genome with rate $\lambda(w)$ given by (S26). For a particular clade G, call recombination events which do not change the membership of G "Type 1", and other events "Type 2" (our key assumption is that Type 1 events also don't change the local trees). We thus have Type 1 events arriving at rate $z\lambda(w)$, and Type 2 events at rate $(1-z)\lambda(w)$, where $z := \mathbb{P}_{\mathcal{T}_{d\leftarrow(G)}}(G \text{ disrupted})$. Let D be the number of Type 1 events before the first Type 2 event. Then it is easy to show that the marginal distribution of D is geometric with parameter z.

Thus, if G is first disrupted by the R-th recombination event, we can calculate a corresponding p-value as

$$p_i = \left[1 - \mathbb{P}_{\mathcal{T}_{d\leftarrow(G)}}(G \text{ disrupted})\right]^{R-1}.$$

If R is known exactly, this is equivalent to Test 1.

References

Carmi, S., Wilton, P. R., Wakeley, J., and Pe'er, I. A renewal theory approach to IBD sharing. *Theoretical Population Biology*, **97**: 35–48, 2014.

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

Eriksson, A., Mahjani, B., and Mehlig, B. Sequential Markov coalescent algorithms for population models with demographic structure. *Theoretical Population Biology*, **76**(2): 84–91, 2009.

Griffiths, R. C. and Marjoram, P. An ancestral recombination graph. In P. Donnelly and S. Tavaré, eds., *Progress in population genetics and human evolution*, 257–270. Springer, New York, 1997.

Harris, K. and Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLOS Genetics*, **9**(6): e1003521, 2013.

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

Hobolth, A. and Jensen, J. L. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, **98**: 48–58, 2014.

Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**(2): 183–201, 1983.

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

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

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

Marjoram, P. and Wall, J. D. Fast coalescent simulation. *BMC Genetics*, **7**(1): 1–9, 2006.

McVean, G. A. and Cardin, N. J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459): 1387–1393, 2005.

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

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

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

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

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

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

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

S2 Proofs

S2.1 Proof of Proposition S1.1

Conditioning on the recombination happening on edge β, the density of the recombination event time is

$$p^S_{\mathcal{T}}(s|\mathcal{R}\in\beta) = \begin{cases} \frac{1}{t(\beta)} & t^{\downarrow}(\beta)\le s\le t^{\uparrow}(\beta) \\ 0 & \text{otherwise,} \end{cases}$$

as the recombination time is chosen uniformly at random along the length of the edge. The conditional density of the coalescence time is

$$p^U_{\mathcal{T}}(u|\mathcal{R}=(\beta,s)) = p^U_{\mathcal{T}}(u|\mathcal{R}=(\cdot,s)) = n(u)\exp\left(-\int_s^u n(t)dt\right), \tag{S1}$$

for $u>s$ and 0 otherwise. Conditional on the coalescence time u, the probability that the coalescence point is on β is

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C}\in\beta|\mathcal{C}=(\cdot,u),\mathcal{R}=(\beta,s)) = \begin{cases} \frac{1}{n(u)} & s<u<t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}$$

Letting $k=n(s)$, so that T_k is the first coalescence time just above time s,

$$\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C}\in\beta|\mathcal{R}=(\beta,s)) &= \int_0^\infty \mathbb{P}(\mathcal{C}\in\beta|\mathcal{C}=(\cdot,u),\mathcal{R}=(\beta,s))\cdot p^U_{\mathcal{T}}(u|\mathcal{R}=(\cdot,s))du \\
&= \int_s^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t)dt\right)du \\
&= \int_s^{T_k} \exp\left(-\int_s^u k\,dt\right)du + \sum_{j=n(t^{\uparrow}(\beta))+1}^{k-1}\int_{T_{j+1}}^{T_j}\exp\left(-\int_s^u n(t)\,dt\right)du,
\end{aligned}$$

note that $T_{n(t^{\uparrow}(b))+1} = t^{\uparrow}(b)$. The first term is

$$\begin{aligned}
\int_s^{T_k}\exp\left(-\int_s^u k\,dt\right)du &= \int_s^{T_k}\exp\left(-k(u-s)\right)du \\
&= \left[-\frac{1}{k}e^{-k(u-s)}\right]_s^{T_k} \\
&= \frac{1}{k}-\frac{1}{k}e^{-kT_k}e^{ks},
\end{aligned} \tag{S2}$$

and the summands of the second term are

$$\begin{aligned}
\int_{T_{j+1}}^{T_j}\exp\left(-\int_s^u n(t)\,dt\right)du &= \int_{T_{j+1}}^{T_j}\exp\left(-\int_s^{T_{j+1}} n(t)\,dt - \int_{T_{j+1}}^u n(t)\,dt\right)du \\
&= \exp\left(-\int_s^{T_{j+1}} n(t)\,dt\right)\int_{T_{j+1}}^{T_j}\exp\left(-\int_{T_{j+1}}^u j\,dt\right)du \\
&= \exp\left(-k(T_k-s)-\sum_{i=j+1}^{k-1} i(T_i-T_{i+1})\right)\left[-\frac{1}{j}e^{-j(u-T_{j+1})}\right]_{T_{j+1}}^{T_j} \\
&= e^{ks}\exp\left(-kT_k+L_{\mathcal{T}}(T_{j+1})-L_{\mathcal{T}}(T_k)\right)\frac{1}{j}\left(1-e^{-j(T_j-T_{j+1})}\right).
\end{aligned}$$

Thus,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C}\in\beta|\mathcal{R}=(\beta,s)) = \frac{1}{k}+e^{ks}\cdot\sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj},$$

where

$$Q_{kk} := -\frac{1}{k} e^{-kT_k},$$

and

$$\begin{aligned} Q_{kj} &:= \frac{1}{j}\left(1 - e^{-j(T_j - T_{j+1})}\right) \exp\left(-kT_k + L_{\mathcal{T}}(T_{j+1}) - L_{\mathcal{T}}(T_k)\right) \\ &= \frac{1}{j} e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_{j+1}) - j(T_j - T_{j+1})}\right) \\ &= e^{-kT_k} e^{-L_{\mathcal{T}}(T_k)} \frac{1}{j} \left(e^{L_{\mathcal{T}}(T_{j+1})} - e^{L_{\mathcal{T}}(T_j)}\right). \end{aligned}$$

S2.2 Proof of Proposition S1.2

Marginalising out the recombination time in (S4),

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \beta | \mathcal{R} \in \beta) &= 1 - \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} = (\beta, s)) p_{\mathcal{T}}^{S}(s | \mathcal{R} \in \beta)\, ds \\ &= 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \int_{T_{k+1}}^{T_k} \left(\frac{1}{k} + e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj}\right) ds \\ &= 1 - \frac{1}{\bar{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1)\right), \end{aligned}$$

where

$$\widetilde{Q}^1(k) := \frac{1}{k}\left(T_k - T_{k+1}\right),$$

and for $x, y, A, B \in \mathbb{Z}$, $x \geq k$, $2 \leq y \leq x$,

$$\widetilde{Q}^2(k, x, y, A, B) := \frac{1}{k}\left(e^{kT_k} - e^{kT_{k+1}}\right) \sum_{j=y}^{x} (Aj + B) \cdot Q_{kj}.$$

S2.3 Proof of Proposition S1.3

We have

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{C} = (\cdot, u), \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \begin{cases} \frac{1}{n(u)} & \max(s, t^{\downarrow}(\beta)) < u < t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}$$

For $s < t^{\downarrow}(\beta)$,

$$\begin{aligned} \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t)dt\right) du &= \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \int_{T_{j+1}}^{T_j} \exp\left(-\int_s^u n(t)dt\right) du \\ &= \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \exp\left(-\int_s^{T_{j+1}} n(t)\, dt\right) \int_{T_{j+1}}^{T_j} \exp\left(-\int_{T_{j+1}}^u j\, dt\right) du \\ &= e^{ks} \cdot \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj}, \end{aligned}$$

and the case $s \geq t^{\downarrow}(\beta)$ is given by (S2). Thus,

$$\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) = \int_{\max(s, t^{\downarrow}(\beta))}^{t^{\uparrow}(\beta)} \exp\left(-\int_s^u n(t)dt\right) du$$

$$
= \begin{cases} e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} Q_{kj} & s < t^{\downarrow}(\beta) \\ \frac{1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(\beta))+1}^{k} Q_{kj} & t^{\downarrow}(\beta) \leq s < t^{\uparrow}(\beta) \\ 0 & \text{otherwise.} \end{cases}
$$

S2.4 Proof of Proposition S1.4

Consider all of the possible orderings of the event times $t_1, \ldots, t_4$, as illustrated in Figure S9.

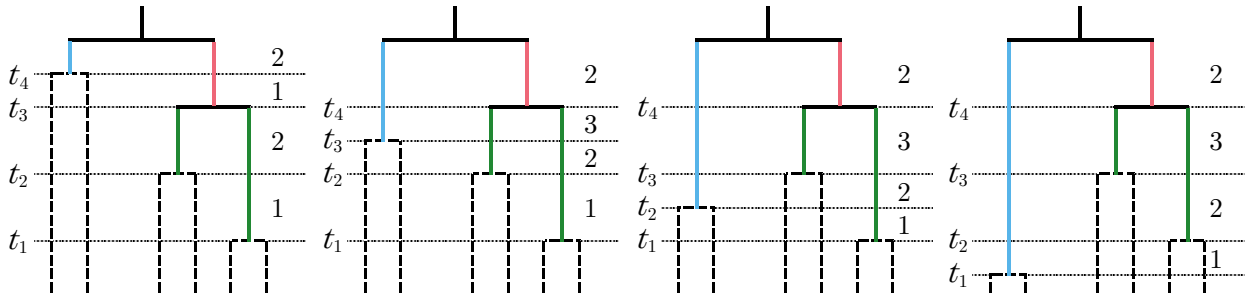

Figure S9: All possible orderings of $t^{\downarrow}(\text{ch}_1(\beta)), t^{\downarrow}(\text{ch}_2(\beta)), t^{\downarrow}(\text{sib}(\beta)), t^{\downarrow}(\beta)$. The edge β is shown in red, $\text{sib}(\beta)$ in blue, $\text{ch}_1(\beta)$ and $\text{ch}_2(\beta)$ in green. Numbers to the right of each tree show the number of lineages in $\mathcal{B}(\beta)$ in each time interval. For instance, in the leftmost tree, $t_1 = t^{\downarrow}(\text{ch}_2(\beta))$, $t_2 = t^{\downarrow}(\text{ch}_1(\beta))$, $t_3 = t^{\downarrow}(\beta)$ and $t_4 = t^{\downarrow}(\text{sib}(\beta))$.

The number of lineages in set $\mathcal{B}(\beta)$ at time s can be written as

$$
\begin{aligned}
n_{\mathcal{B}(\beta)}(s) &:= \sum_{b' \in \mathcal{B}(\beta)} \mathbb{1}(s \in [t^{\downarrow}(b'), t^{\uparrow}(b')]) \\
&= \begin{cases} 1 & t_1 \leq s < t_2 \\ 2 & t_2 \leq s < t_3 \\ 1 + 2 \cdot \mathbb{1}(t^{\downarrow}(\text{sib}(\beta)) < t^{\downarrow}(\beta)) & t_3 \leq s < t_4 \\ 2 & t_4 \leq s < t_5 = t^{\uparrow}(\beta) \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

where $\mathbb{1}(\cdot)$ is the indicator function. Then conditional on the recombination point not being on an edge in the set $\mathcal{B}(\beta)$, the density of the recombination event time is

$$
p^S_{\mathcal{T}}(s|\mathcal{R} \notin \mathcal{B}(\beta)) = \begin{cases} \frac{n(s) - n_{\mathcal{B}(\beta)}(s)}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} & 0 \leq s \leq T_2 \\ 0 & \text{otherwise.} \end{cases}
$$

Marginalising out the recombination time in (S10),

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta)) &= \int_0^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \beta | \mathcal{R} \notin \mathcal{B}(\beta), \mathcal{R} = (\cdot, s)) \cdot p^S_{\mathcal{T}}(s|\mathcal{R} \notin \mathcal{B}(\beta))\, ds. \\
&= \frac{1}{L_{\mathcal{T}}(0) - \sum_{b' \in \mathcal{B}(\beta)} \bar{t}(b')} \Bigg\{ \sum_{k=n(t_1)+1}^{n} k \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \sum_{k=n(t_2)+1}^{n(t_1)} (k-1) \widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{k=n(t_3)+1}^{n(t_2)} (k-2)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&+ \sum_{k=n(t_4)+1}^{n(t_3)} \Big[\mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) < t^{\downarrow}(\beta))(k-3)\widetilde{Q}^2(k, n(t^{\downarrow}(\beta)), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\
&\quad + \mathbb{1}(t^{\downarrow}(\mathrm{sib}(\beta)) \geq t^{\downarrow}(\beta))(k-1) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Big] \\
&+ \sum_{k=n(t_5)+1}^{n(t_4)} (k-2) \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) \right) \Bigg\}.
\end{aligned}
$$

S2.5 Proof of Theorem S1.1

If we now say that an edge is disrupted only if there is a change in topology (but not edge length), we can allow the events shown in Figure 2 in blue (marked with dots), i.e. those where the recombination point is on edge $\beta \in \mathcal{B}(\beta)$, and the coalescence point is on one of the edges in $\mathcal{A}(\beta)$. Thus,

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(b \text{ topologically disrupted}) = &\sum_{b' \in \mathcal{B}(b)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(b') | \mathcal{R} \in b') \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \in b') \\
&+ \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in b | \mathcal{R} \notin \mathcal{B}(b)) \cdot \mathbb{P}_{\mathcal{T}}(\mathcal{R} \notin \mathcal{B}(b)).
\end{aligned}
$$

To calculate the probability $\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta) | \mathcal{R} \in \beta)$, we follow the same approach as the proofs of Propositions S1.1 and S1.2, first conditioning on the recombination point $\mathcal{R} = (\beta, s)$. Let $k = n(s)$, so that T_k is the first coalescence time just above time s. Let

$$
r(u) := \exp\left(-\int_s^u n(t) dt \right).
$$

Then if $t^{\downarrow}(\mathrm{sib}(b)) < s$,

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{R} = (\beta, s)) &= \int_0^{\infty} \mathbb{P}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{C} = (\cdot, u), \mathcal{R} = (\beta, s)) \cdot p_{\mathcal{T}}^U(u | \mathcal{R} = (\cdot, s)) du \\
&= \int_s^{t^{\uparrow}(\beta)} 2 \cdot r(u)\, du + \int_{t^{\uparrow}(\beta)}^{t^{\uparrow}(\mathrm{par}(\beta))} r(u)\, du \\
&= \frac{2}{k} + e^{ks} \cdot \left(\sum_{j=n(t^{\uparrow}(\beta))+1}^{k} 2 \cdot Q_{kj} + \sum_{j=n(t^{\uparrow}(\mathrm{par}(\beta)))+1}^{n(t^{\uparrow}(\beta))} Q_{kj} \right),
\end{aligned}
$$

and if $t^{\downarrow}(\mathrm{sib}(b)) \geq s$,

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{R} = (\beta, s)) &= \int_s^{t^{\downarrow}(\mathrm{sib}(\beta))} r(u)\, du + \int_{t^{\downarrow}(\mathrm{sib}(\beta))}^{t^{\uparrow}(\beta)} 2 \cdot r(u)\, du + \int_{t^{\uparrow}(\beta)}^{t^{\uparrow}(\mathrm{par}(\beta))} r(u)\, du \\
&= \frac{1}{k} + e^{ks} \cdot \left(\sum_{j=n(t^{\downarrow}(\mathrm{sib}(\beta)))+1}^{k} Q_{kj} + \sum_{j=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\mathrm{sib}(\beta)))} 2 \cdot Q_{kj} + \sum_{j=n(t^{\uparrow}(\mathrm{par}(\beta)))+1}^{n(t^{\uparrow}(\beta))} Q_{kj} \right).
\end{aligned}
$$

Marginalising out the recombination time,

$$
\begin{aligned}
\mathbb{P}_{\mathcal{T}}(\mathcal{C} \notin \mathcal{A}(\beta) | \mathcal{R} \in \beta) &= 1 - \int_{t^{\downarrow}(\beta)}^{t^{\uparrow}(\beta)} \mathbb{P}_{\mathcal{T}}(\mathcal{C} \in \mathcal{A}(\beta) | \mathcal{R} = (\beta, s)) p_{\mathcal{T}}^S(s | \mathcal{R} \in \beta)\, ds \\
&= 1 - \frac{1}{\overline{t}(\beta)} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} G_{\beta}(k),
\end{aligned}
$$

where for $k \leq n(t^{\downarrow}(\text{sib}(\beta)))$,

$$G_\beta(k) \;=\; 2 \cdot \widetilde{Q}^1(k) + 2 \cdot \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1) + \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\text{par}(\beta))) + 1, 0, 1),$$

and for $k > n(t^{\downarrow}(\text{sib}(\beta)))$

$$\begin{aligned} G_\beta(k) = \widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\downarrow}(\text{sib}(\beta))) + 1, 0, 1) + 2 \cdot \widetilde{Q}^2(k, n(t^{\downarrow}(\text{sib}(\beta))), n(t^{\uparrow}(\beta)) + 1, 0, 1) \\ + \widetilde{Q}^2(k, n(t^{\uparrow}(\beta)), n(t^{\uparrow}(\text{par}(\beta))) + 1, 0, 1). \end{aligned}$$

S2.6 Proof of Proposition S1.5

S2.6.1 Probability of no change in total branch length

The probability that the recombination event does not result in a change in total branch length is

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(C = 0) &= \sum_{b \in \mathcal{T}} \frac{\bar{t}(b)}{L_{\mathcal{T}}(0)} \left(1 - P_{\mathcal{T}}(\mathcal{C} \notin b | \mathcal{R} \in b)\right) \\ &= \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(\beta))+1}^{n(t^{\downarrow}(\beta))} \left(\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 0, 1)\right), \end{aligned}$$

using (S7). This is equal to the probability derived by Deng et al. (2021, Theorem 1).

S2.6.2 Probability that change in total branch length is negative

We have

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(C < 0 | \mathcal{R} = (b, s)) &= \int_s^{t^{\uparrow}(b)} \frac{n(u) - 1}{n(u)} \cdot n(u) \exp\left(-\int_s^u n(t)\, dt\right) du \\ &= \frac{k-1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(b))+1}^{k} (j-1) Q_{kj}. \end{aligned}$$

Integrating over the recombination time,

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(C < 0 | \mathcal{R} \in b) &= \frac{1}{\bar{t}(b)} \int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)} \left(\frac{k-1}{k} + e^{ks} \sum_{j=n(t^{\uparrow}(b))+1}^{k} (j-1) Q_{kj}\right) ds \\ &= \frac{1}{\bar{t}(b)} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left((k-1)\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 1, -1)\right). \end{aligned}$$

Then

$$\mathbb{P}_{\mathcal{T}}(C < 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left((k-1)\widetilde{Q}^1(k) + \widetilde{Q}^2(k, k, n(t^{\uparrow}(\beta)) + 1, 1, -1)\right).$$

S2.6.3 Probability that change in total branch length is positive

Similarly, the probability that the change in total branch length is positive is given by

$$\begin{aligned} \mathbb{P}_{\mathcal{T}}(C > 0 | \mathcal{R} = (b, s)) &= \int_{t^{\uparrow}(b)}^{\infty} n(u) \exp\left(-\int_s^u n(t)\, dt\right) du \\ &= \int_{t^{\uparrow}(b)}^{T_2} n(u) \exp\left(-\int_s^u n(t)\, dt\right) du + \int_{T_2}^{\infty} \exp\left(-\int_s^u n(t)\, dt\right) du \end{aligned}$$

$$= e^{ks} \sum_{j=2}^{n(t^{\uparrow}(b))} jQ_{kj} + e^{-L_{\mathcal{T}}(s)},$$

giving

$$\mathbb{P}_{\mathcal{T}}(C > 0 | \mathcal{R} \in b) = \frac{1}{\bar{t}(b)} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^2(k, n(t^{\uparrow}(b)), 2, 1, 0) + \widetilde{Q}^3(k) \right),$$

where

$$\begin{aligned}
\widetilde{Q}^3(k) &= \int_{T_{k+1}}^{T_k} e^{-L_{\mathcal{T}}(s)}\, ds \\
&= \int_{T_{k+1}}^{T_k} e^{ks} \exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right)\, ds \\
&= \frac{1}{k}\left(e^{kT_k} - e^{kT_{k+1}}\right) \exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right) \\
&= \frac{1}{k}\left(e^{-L_{\mathcal{T}}(T_k)} - e^{-L_{\mathcal{T}}(T_{k+1})}\right).
\end{aligned}$$

Thus,

$$\mathbb{P}_{\mathcal{T}}(C > 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \left(\widetilde{Q}^2(k, n(t^{\uparrow}(b)), 2, 1, 0) + \widetilde{Q}^3(k) \right).$$

S2.7 Proof of Proposition S1.6

We first calculate the density of change in total branch length conditional on the recombination point $\mathcal{R} = (b, s)$, then marginalise out the recombination time and edge.

S2.8 Density of change in total branch length conditional on recombination point

Suppose that the recombination point is on edge b at time s. Then given the coalescence time u, the change in total branch length $C > t^{\downarrow}(b) + T_3 - 2T_2$ given $C \neq 0$ is

$$C = \begin{cases} u + T_3 - 2T_2 & t^{\uparrow}(b) = T_2 \text{ and } u \leq T_3 \\ 2u - t^{\uparrow}(b) - T_2 & t^{\uparrow}(b) = T_2 \text{ and } u > T_3, \text{ or } u > T_2 \\ u - t^{\uparrow}(b) & \text{otherwise.} \end{cases}$$

Let $l = n(u)$, so T_l is the time of the first coalescence event just above time u. We need to condition on the coalescence point not being on edge b, and as a simplification we take

$$p^U_{\mathcal{T}}(u | \mathcal{R} = (\cdot, s), \mathcal{C} \notin b) = \begin{cases} n(u) \exp\left(-\int_s^u n(t) dt\right) & u \geq t^{\uparrow}(b) \\ (n(u) - 1) \exp\left(-\int_s^u (n(t) - 1) dt\right) & s \leq u < t^{\uparrow}(b) \\ 0 & \text{otherwise.} \end{cases}$$

This essentially assumes SMC rather than SMC' dynamics, since the two models differ only in that the latter allows the coalescence event to occur on the same branch as the recombination event, so this is a very close match for the conditional distribution (and simplifies our calculations). Thus, through a change of variable in (S1), for $t^{\uparrow}(b) \neq T_2$ and $s - t^{\uparrow}(b) < c < 0$,

$$\begin{aligned}
p^C_{\mathcal{T}}(c | C \neq 0, \mathcal{R} = (b, s), t^{\uparrow}(b) \neq T_2) &= (n(u) - 1) \exp\left(- \int_s^u (n(t) - 1)\, dt \right) \\
&= (n(u) - 1) \exp\left(\left[- \int_s^{T_k} - \int_{T_k}^{T_l} + \int_u^{T_l} \right] (n(t) - 1)\, dt \right)
\end{aligned}$$

$$
\begin{aligned}
&= e^{(k-1)s} P^1_{kl}(n(u)-1)e^{-(l-1)u} \\
&= e^{(k-1)s} P^1_{kl}(n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))},
\end{aligned}
$$

where

$$
\begin{aligned}
P^1_{kl} &:= \exp\left(-(k-1)T_k - \sum_{i=l}^{k-1}(i-1)(T_i - T_{i+1}) + (l-1)T_l\right) \\
&= \exp\left(l \cdot T_l - k \cdot T_k + L_{\mathcal{T}}(T_l) - L_{\mathcal{T}}(T_k)\right).
\end{aligned}
$$

For $t^{\uparrow}(b) = T_2$ and $t^{\downarrow}(b) + T_3 - 2T_2 \le c \le 2(T_3 - T_2)$, similarly,

$$
p^C_{\mathcal{T}}(c|C \neq 0, \mathcal{R} = (b,s), t^{\uparrow}(b) = T_2) = e^{(k-1)s} P^1_{kl}(n(c+2T_2-T_3)-1)e^{-(l-1)(c+2T_2-T_3)},
$$

and for $t^{\uparrow}(b) = T_2$ and $2(T_3 - T_2) < c < 0$, since $l = 2$,

$$
p^C_{\mathcal{T}}(c|C \neq 0, \mathcal{R} = (b,s), t^{\uparrow}(b) = T_2) = \frac{1}{2} e^{(k-1)s} P^1_{kl}\left(n\left(c/2 + T_2\right) - 1\right) e^{-(c/2+T_2)}.
$$

For $0 < c \le T_2 - t^{\uparrow}(b)$,

$$
\begin{aligned}
p^C_{\mathcal{T}}(c|C \neq 0, \mathcal{R} = (b,s)) &= n(u) \exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt - \int_{t^{\uparrow}(b)}^{u} n(t)\,dt\right) \\
&= \exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt\right) n(u) \exp\left(-\int_{t^{\uparrow}(b)}^{u} n(t)\,dt\right) \\
&= e^{(k-1)s} P^2_k n(c+t^{\uparrow}(b)) \exp\left(-\int_{t^{\uparrow}(b)}^{c+t^{\uparrow}(b)} n(t)\,dt\right) \\
&= e^{(k-1)s} P^2_k n(c+t^{\uparrow}(b)) \exp\left(-\left[L_{\mathcal{T}}(t^{\uparrow}(b)) - L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]\right),
\end{aligned}
$$

where

$$
\begin{aligned}
P^2_k &:= \exp\left(-(k-1)T_k - \sum_{i=n(t^{\uparrow}(b))+1}^{k-1} (i-1)(T_i - T_{i+1})\right) \\
&= \exp\left(t^{\uparrow}(b) - k \cdot T_k - L_{\mathcal{T}}(T_k) + L_{\mathcal{T}}(t^{\uparrow}(b))\right).
\end{aligned}
$$

Finally, for $c > T_2 - t^{\uparrow}(b)$,

$$
\begin{aligned}
p^C_{\mathcal{T}}(c|C \neq 0, \mathcal{R} = (b,s)) &= \exp\left(-\int_s^{t^{\uparrow}(b)} (n(t)-1)\,dt - \int_{t^{\uparrow}(b)}^{T_2} n(t)\,dt - \int_{T_2}^{u} 1\,dt\right) \\
&= \exp\left(-\int_s^{T_2} n(t)\,dt + \int_s^{t^{\uparrow}(b)} 1\,dt - \int_{T_2}^{u} 1\,dt\right) \\
&= e^{(k-1)s} P^3_k e^{-(u-T_2)} \\
&= \frac{1}{2} e^{(k-1)s} P^3_k e^{-(c+t^{\uparrow}(b)-T_2)/2},
\end{aligned}
$$

where

$$
\begin{aligned}
P^3_k &:= \exp\left(-(k-1)T_k - \sum_{i=2}^{k-1} i(T_i - T_{i+1}) + \sum_{i=n(t^{\uparrow}(b))+1}^{k-1} (T_i - T_{i+1})\right) \\
&= \exp\left(t^{\uparrow}(b) - k \cdot T_k - L_{\mathcal{T}}(T_k)\right).
\end{aligned}
$$

Note that

$$e^{(k-1)s}P_k^3 = \exp\left(-\left[L_{\mathcal{T}}(s) - (t^{\uparrow}(b) - s)\right]\right)$$

gives the probability that the coalescence event happens above T_2. The conditional density $p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))$ is thus

$$\begin{cases} e^{(k-1)s}P_{kl}^1(n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))} & t^{\uparrow}(b) \neq T_2,\ s - t^{\uparrow}(b) \leq c < 0 \\ e^{(k-1)s}P_{kl}^1(n(c+2T_2-T_3)-1)e^{-(l-1)(c+2T_2-T_3)} & t^{\uparrow}(b) = T_2,\ t^{\downarrow}(b) + T_3 - 2T_2 \leq c < 2(T_3 - T_2) \\ \frac{1}{2}e^{(k-1)s}P_{kl}^1\left(n\left(\frac{c}{2}+T_2\right)-1\right)e^{-(c/2+T_2)} & t^{\uparrow}(b) = T_2,\ 2(T_3 - T_2) \leq c < 0 \\ e^{(k-1)s}P_k^2 n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b)) - L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]} & 0 < c \leq T_2 - t^{\uparrow}(b) \\ \frac{1}{2}e^{(k-1)s}P_k^3 e^{-(c+t^{\uparrow}(b)-T_2)/2} & c > T_2 - t^{\uparrow}(b) \\ 0 & \text{otherwise.} \end{cases}$$

S2.8.1 Density of change in total branch length

Marginalising out the position and time of the recombination point, we have

$$\begin{aligned} p_{\mathcal{T}}^C(c|C \neq 0) &= \sum_{b \in \mathcal{T}} \mathbb{P}(\mathcal{R} \in b) \int_0^{\infty} p_{\mathcal{T}}^S(s|\mathcal{R} \in b)\, p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds \\ &= \frac{1}{L_{\mathcal{T}}(0)} \sum_{\substack{b \in \mathcal{T}: \\ c \geq -\bar{t}(b), \\ t^{\uparrow}(b) \neq T_2}} \int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)+\min(0,c)} p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds \\ &+ \frac{1}{L_{\mathcal{T}}(0)} \sum_{\substack{b \in \mathcal{T}: \\ t^{\downarrow}(b)+T_3-2T_2 \leq c < 2(T_3-T_2), \\ t^{\uparrow}(b) = T_2}} \int_{t^{\downarrow}(b)}^{c+2T_2-T_3} p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds \\ &+ \frac{1}{L_{\mathcal{T}}(0)} \sum_{\substack{b \in \mathcal{T}: \\ 2(T_3-T_2) \leq c < 0, \\ t^{\uparrow}(b) = T_2}} \int_{t^{\downarrow}(b)}^{c/2+T_2} p_{\mathcal{T}}^C(c|C \neq 0, \mathcal{R} = (b,s))\, ds. \end{aligned}$$

Let

$$\begin{aligned} \widetilde{P}_{kl}^1 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s}P_{kl}^1\, ds = \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)P_{kl}^1, \\ \widetilde{P}_k^2 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s}P_k^2\, ds = \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)P_k^2, \\ \widetilde{P}_k^3 &:= \int_{T_{k+1}}^{T_k} e^{(k-1)s}P_k^3\, ds = \frac{1}{k-1}\left(e^{(k-1)T_k} - e^{(k-1)T_{k+1}}\right)P_k^3. \end{aligned}$$

Then for $t^{\uparrow}(b) \neq T_2$ and $-\bar{t} \leq c < 0$,

$$\begin{aligned} &\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)+c} e^{(k-1)s}P_{kl}^1(n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))}\, ds \\ &\quad = (n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))}\left(\sum_{k=l}^{n(t^{\downarrow}(b))} \int_{T_{k+1}}^{T_k} e^{(k-1)s}P_{kl}^1\, ds - \int_{c+t^{\uparrow}(b)}^{T_l} e^{(l-1)s}\, ds\right) \end{aligned}$$

$$= (n(c+t^{\uparrow}(b))-1)e^{-(l-1)(c+t^{\uparrow}(b))}\left(\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl} - \frac{1}{l-1}\left(e^{(l-1)T_l} - e^{(l-1)(c+t^{\uparrow}(b))}\right)\right)$$

$$= (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl} - \frac{1}{l-1}\left(e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1\right)\right).$$

Similarly, for $t^{\uparrow}(b)=T_2$ and $t^{\downarrow}(b)+T_3-2T_2 \le c < 2(T_3-T_2)$,

$$\int_{t^{\downarrow}(b)}^{c+2T_2-T_3} e^{(k-1)s}P^1_{kl}(n(c+2T_2-T_3)-1)e^{-(l-1)(c+2T_2-T_3)}\,ds$$

$$= (n(c+2T_2-T_3)-1)\left(e^{-(l-1)(c+2T_2-T_3)}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl} - \frac{1}{l-1}\left(e^{(l-1)(T_l-c-2T_2+T_3)}-1\right)\right),$$

and for $t^{\uparrow}(b)=T_2$ and $2(T_3-T_2)\le c<0$,

$$\int_{t^{\downarrow}(b)}^{c/2+T_2}\frac{1}{2}e^{(k-1)s}P^1_{kl}(n(c/2+T_2)-1)e^{-(l-1)(c/2+T_2)}\,ds$$

$$= \frac{1}{2}(n(c/2+T_2)-1)\left(e^{-(l-1)(c/2+T_2)}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl} - e^{T_l-c/2-T_2}+1\right).$$

For $0<c\le T_2-t^{\uparrow}(b)$,

$$\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)} e^{(k-1)s}P^2_k n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\,ds$$

$$= n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\int_{T_{k+1}}^{T_k} e^{(k-1)s}P^2_k\,ds$$

$$= n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}^2_k.$$

For $c>T_2-t^{\uparrow}(b)$,

$$\int_{t^{\downarrow}(b)}^{t^{\uparrow}(b)}\frac{1}{2}e^{(k-1)s}P^3_k e^{-(c+t^{\uparrow}(b)-T_2)/2}\,ds$$

$$= \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\int_{T_{k+1}}^{T_k} e^{(k-1)s}P^3_k\,ds$$

$$= \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}^3_k.$$

Thus,

$$p^C_{\mathcal{T}}(c|C\neq 0) = \frac{1}{L_{\mathcal{T}}(0)}\sum_{b\in\mathcal{T}}\xi_b(c),$$

where $\xi_b(c)$ is given by

$$\begin{cases} (n(c+t^{\uparrow}(b))-1)\left(e^{-(l-1)(c+t^{\uparrow}(b))}\cdot\sum\limits_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl}-\frac{e^{(l-1)(T_l-t^{\uparrow}(b)-c)}-1}{l-1}\right) & t^{\uparrow}(b)\neq T_2,\ -\bar{t}(b)\leq c<0 \\ (n(c+2T_2-T_3)-1)\left(e^{-(l-1)(c+2T_2-T_3)}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl}-\frac{e^{(l-1)(T_l-c-2T_2+T_3)}-1}{l-1}\right) & t^{\uparrow}(b)=T_2, \\ & t^{\downarrow}(b)+T_3-2T_2\leq c<2(T_3-T_2) \\ \frac{1}{2}(n(c/2+T_2)-1)\left(e^{-(c/2+T_2)}\cdot\sum_{k=l}^{n(t^{\downarrow}(b))}\widetilde{P}^1_{kl}-e^{T_l-c/2-T_2}+1\right) & t^{\uparrow}(b)=T_2,\ 2(T_3-T_2)\leq c<0 \\ n(c+t^{\uparrow}(b))e^{-\left[L_{\mathcal{T}}(t^{\uparrow}(b))-L_{\mathcal{T}}(c+t^{\uparrow}(b))\right]}\cdot\sum\limits_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}^2_k & 0<c\leq T_2-t^{\uparrow}(b) \\ \frac{1}{2}e^{-(c+t^{\uparrow}(b)-T_2)/2}\cdot\sum\limits_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{P}^3_k & c>T_2-t^{\uparrow}(b) \\ 0 & \text{otherwise.} \end{cases}$$

S2.9 Proof of Proposition S1.7

The probability of a negative change in tree height, conditional on the location and time of the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H<0|\mathcal{R}=(b,s))=\begin{cases}\frac{k-1}{k}-e^{ks}\left(\exp\left(-kT_k-L_{\mathcal{T}}(T_k)\right)+\sum_{j=2}^{k}Q_{kj}\right) & b\in\mathcal{M} \\ 0 & b\notin\mathcal{M}.\end{cases}$$

Integrating over the recombination time gives

$$\mathbb{P}_{\mathcal{T}}(H<0|\mathcal{R}\in b)=\begin{cases}\frac{1}{\bar{t}(b)}\sum\limits_{k=2}^{n(t^{\downarrow}(b))}\left\{(k-1)\widetilde{Q}^1(k)-\widetilde{Q}^2(k,k,2,0,1)-\widetilde{Q}^3(k)\right\} & b\in\mathcal{M} \\ 0 & b\notin\mathcal{M}.\end{cases}$$

Summing over the edges and multiplying by the corresponding probability, the unconditional probability that the change in tree height is negative is thus

$$\mathbb{P}_{\mathcal{T}}(H<0)=\frac{1}{L_{\mathcal{T}}(0)}\sum_{b\in\mathcal{M}}\sum_{k=2}^{n(t^{\downarrow}(b))}\left\{(k-1)\widetilde{Q}^1(k)-\widetilde{Q}^2(k,k,2,0,1)-\widetilde{Q}^3(k)\right\}.$$

The probability of no change in tree height, conditional on the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H=0|\mathcal{R}=(b,s))=\begin{cases}\frac{1}{k}+e^{ks}\sum_{j=2}^{k}Q_{kj} & b\in\mathcal{M} \\ 1-e^{ks}\exp\left(-kT_k-L_{\mathcal{T}}(T_k)\right) & b\notin\mathcal{M},\end{cases}$$

and

$$\mathbb{P}_{\mathcal{T}}(H=0|\mathcal{R}\in b)=\begin{cases}\frac{1}{\bar{t}(b)}\sum\limits_{k=2}^{n(t^{\downarrow}(b))}\left(\widetilde{Q}^1(k)+\widetilde{Q}^2(k,k,2,0,1)\right) & b\in\mathcal{M} \\ 1-\frac{1}{\bar{t}(b)}\sum\limits_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{Q}^3(k) & b\notin\mathcal{M}.\end{cases}$$

The probability of no change in tree height is thus

$$\mathbb{P}_{\mathcal{T}}(H=0)=\frac{1}{L_{\mathcal{T}}(0)}\left\{\sum_{b\in\mathcal{M}}\sum_{k=2}^{n(t^{\downarrow}(b))}\left(\widetilde{Q}^1(k)+\widetilde{Q}^2(k,k,2,0,1)\right)+\sum_{b\notin\mathcal{M}}\left(\bar{t}(b)-\sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))}\widetilde{Q}^3(k)\right)\right\}.$$

Finally, the probability of a positive change in height, conditional on the recombination point, is

$$\mathbb{P}_{\mathcal{T}}(H > 0|\mathcal{R} = (b, s)) = e^{ks} \exp\left(-kT_k - L_{\mathcal{T}}(T_k)\right),$$

and

$$\mathbb{P}_{\mathcal{T}}(H > 0|\mathcal{R} \in b) = \frac{1}{\bar{t}(b)} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k).$$

The probability that the change in tree height is positive is

$$\mathbb{P}_{\mathcal{T}}(H > 0) = \frac{1}{L_{\mathcal{T}}(0)} \sum_{b \in \mathcal{T}} \sum_{k=n(t^{\uparrow}(b))+1}^{n(t^{\downarrow}(b))} \widetilde{Q}^3(k).$$

S2.10 Proof of Theorem S1.2

Conditional on $\mathcal{T}$, the recombination point is chosen uniformly along the edges, so

$$\mathbb{P}(\mathcal{R} \in G) = \frac{L_G(0)}{L_{\mathcal{T}}(0)}, \quad \mathbb{P}(\mathcal{R} \notin G \cup g) = \frac{L_{\mathcal{T}} - L_G(0) - \bar{t}(g)}{L_{\mathcal{T}}(0)}.$$

Conditional on the recombination point being in G and letting the clade MRCA time be $t^{\downarrow}(g) = T_m$, the density of the recombination time is

$$p_S(s|\mathcal{R} \in G) = \begin{cases} \frac{n_G(s)}{L_G(0)} & \text{for } s \leq T_m \\ 0 & \text{otherwise,} \end{cases}$$

and similarly

$$p_S(s|\mathcal{R} \notin G \cup g) = \begin{cases} \frac{n(s) - n_{G \cup g}(s)}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} & \text{for } s \leq T_2 \\ 0 & \text{otherwise.} \end{cases}$$

First conditioning on the recombination time,

$$\begin{aligned}
\mathbb{P}(\mathcal{C} \in G \cup g|\mathcal{R} = (G, s)) &= \int_s^{t^{\uparrow}(g)} \frac{n_{G \cup g}(u)}{n(u)} n(u) \exp\left(-\int_s^u n(t)dt\right) du \\
&= \int_s^{T_k} n_{G \cup g}(T_{k+1}) \exp\left(-\int_s^u k dt\right) du \\
&+ \sum_{j=n(t^{\uparrow}(g))+1}^{k-1} \int_{T_{j+1}}^{T_j} n_{G \cup g}(T_{j+1}) \exp\left(-\int_s^u n(t)dt\right) du \\
&= \frac{1}{k} n_{G \cup g}(T_{k+1}) + e^{ks} \sum_{j=n(t^{\uparrow}(g))+1}^{k} n_{G \cup g}(T_{j+1}) Q_{kj},
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{P}(\mathcal{C} \in G \cup g|\mathcal{R} \in G) &= \frac{1}{L_G(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \int_{T_{k+1}}^{T_k} n_G(T_{k+1}) \mathcal{P}(\mathcal{C} \in G \cup g|\mathcal{R} \in G, \mathcal{R} = (\cdot, s)) ds \\
&= \frac{1}{L_G(0)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \left[n_{G \cup g}(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G \cup g) \right] n_G(T_{k+1}),
\end{aligned}$$

with

$$\widetilde{Q}^4(k, A) = \frac{1}{k} \left(e^{kT_k} - e^{kT_{k+1}} \right) \sum_{j=n(t^{\uparrow}(A))+1}^{k} n_A(T_{j+1}) Q_{kj}.$$

Similarly,

$$\begin{aligned}\mathbb{P}(\mathcal{C} \in G | \mathcal{R} \notin G \cup g, \mathcal{R} = (\cdot, s)) &= \int_s^{t^{\downarrow}(g)} \frac{n_G(u)}{n(u)} n(u) r(u) du \\ &= \frac{1}{k} n_G(T_{k+1}) + \sum_{j=n(t^{\downarrow}(g))+1}^{k} e^{ks} n_G(T_{j+1}) Q_{kj},\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\mathcal{C} \in G | \mathcal{R} \notin G \cup g) &= \frac{1}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} \int_{T_{k+1}}^{T_k} (k - n_{G \cup g}(T_{k+1})) \\ &\quad \cdot \mathbb{P}(\mathcal{C} \in G | \mathcal{R} \notin G \cup g, \mathcal{R} = (\cdot, s)) ds \\ &= \frac{1}{L_{\mathcal{T}}(0) - L_G(0) - \bar{t}(g)} \sum_{k=n(t^{\downarrow}(g))+1}^{n} (k - n_{G \cup g}(T_{k+1})) \left[n_G(T_{k+1}) \widetilde{Q}^1(k) + \widetilde{Q}^4(k, G) \right].\end{aligned}$$

S3 Supplementary Figures

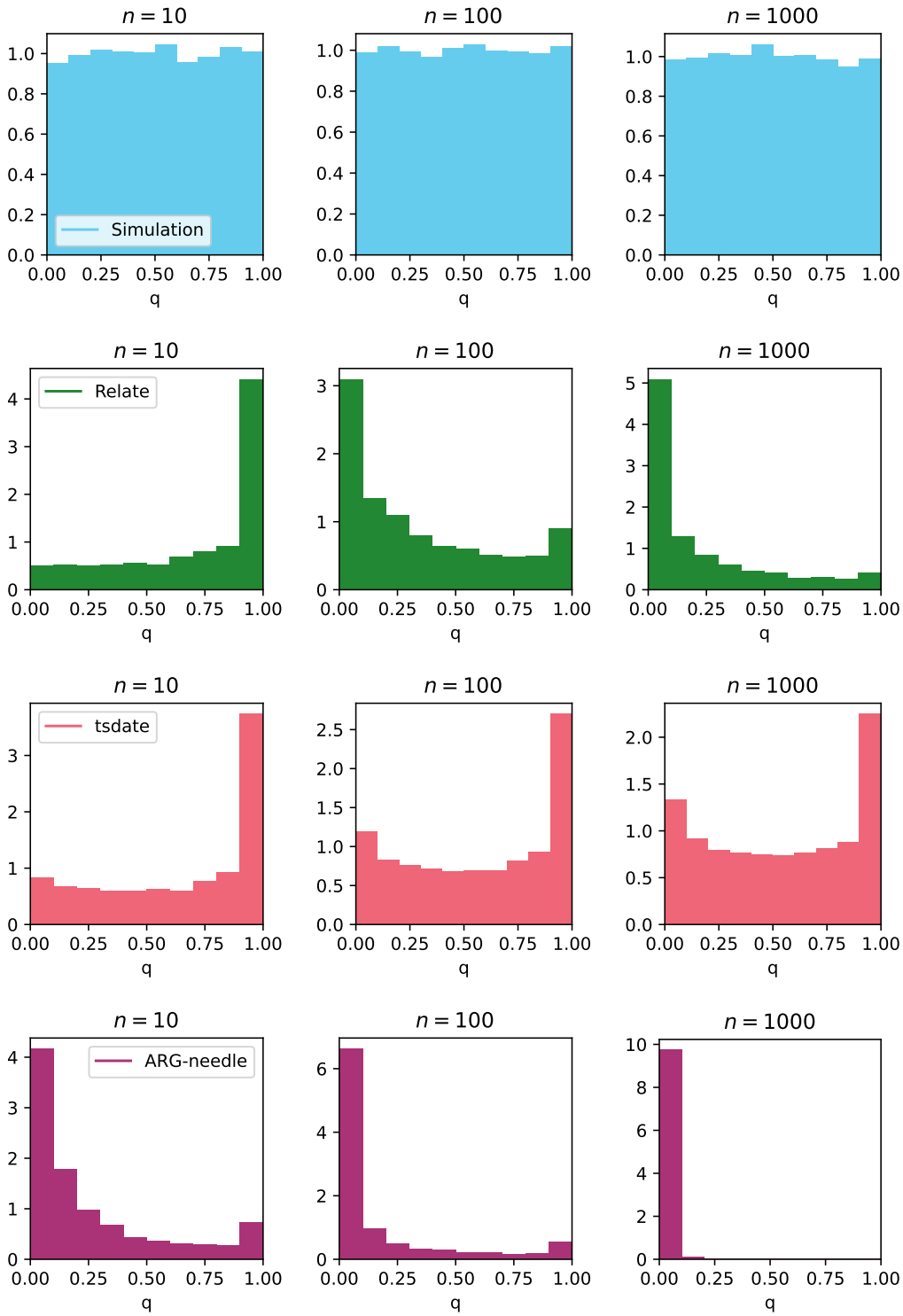

Figure S10: Histograms corresponding to Q-Q plots in Figure 3.

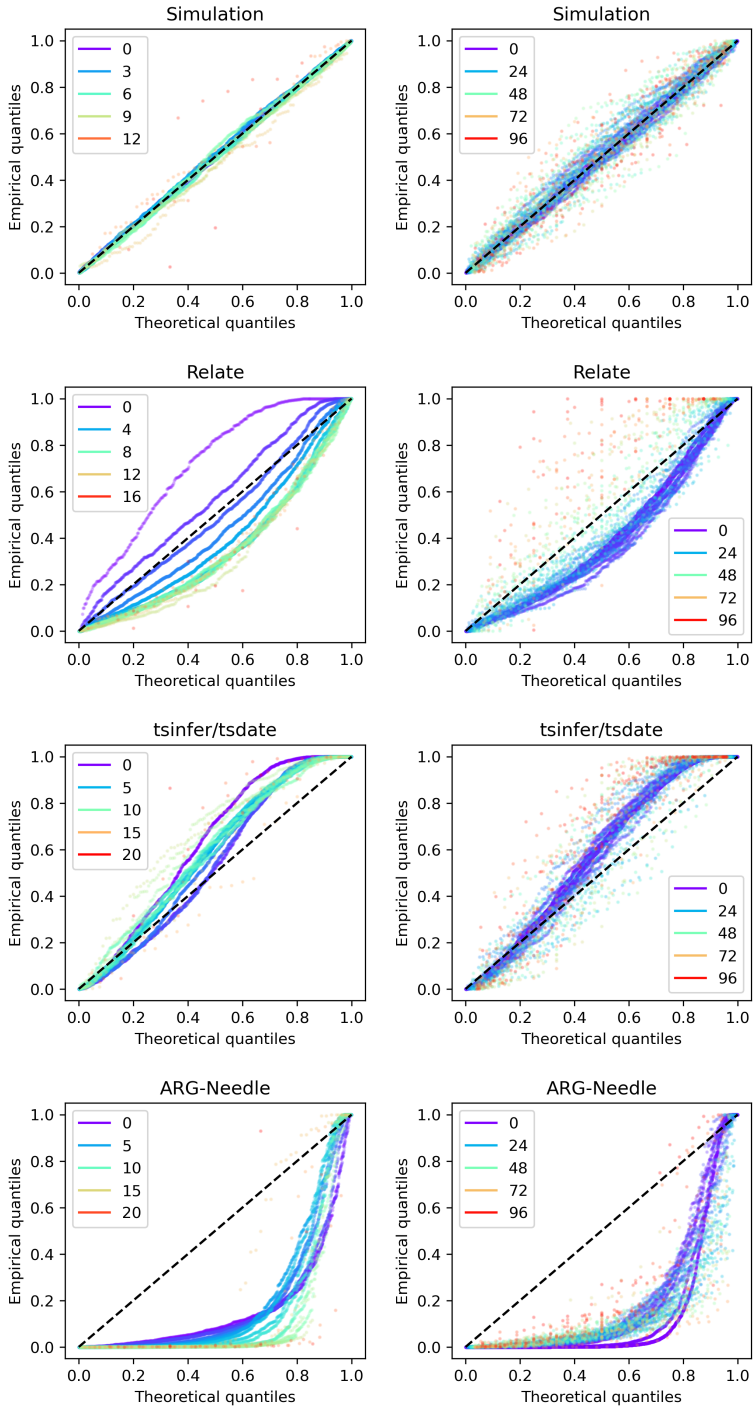

Figure S11: Q-Q plots split by clade size and depth. Q-Q plots using (S23) computed from an ARG simulated using dataset 1 parameters under the SMC' (with $n = 100$). Left panels: split by depth of the edge (defined as the number of edges to the root of the tree); right panels: split by clade size (defined as the number of samples subtended by the edge). Dashed line: diagonal from $(0, 0)$ to $(1, 1)$. For the simulated ARGs, none of the corresponding K–S p-values for each group are significant using a 0.05 significance threshold.

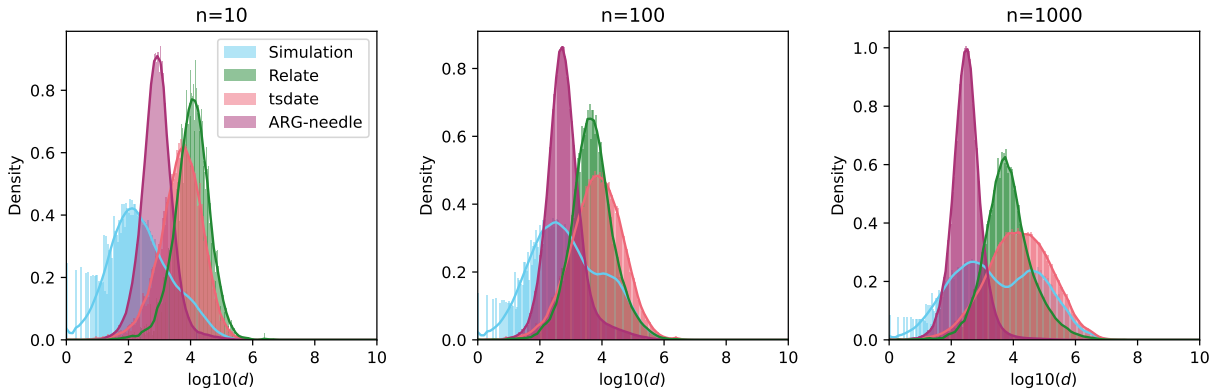

Figure S12: Histograms of observed edge span for simulated and reconstructed ARGs. Histograms of (observed) edge span, calculated as $d(b) = d^{\rightarrow}(b) - d^{\leftarrow}(b)$ for each edge b in simulated and reconstructed ARGs (same as those in Figure 3). Note log scale on the x-axis.

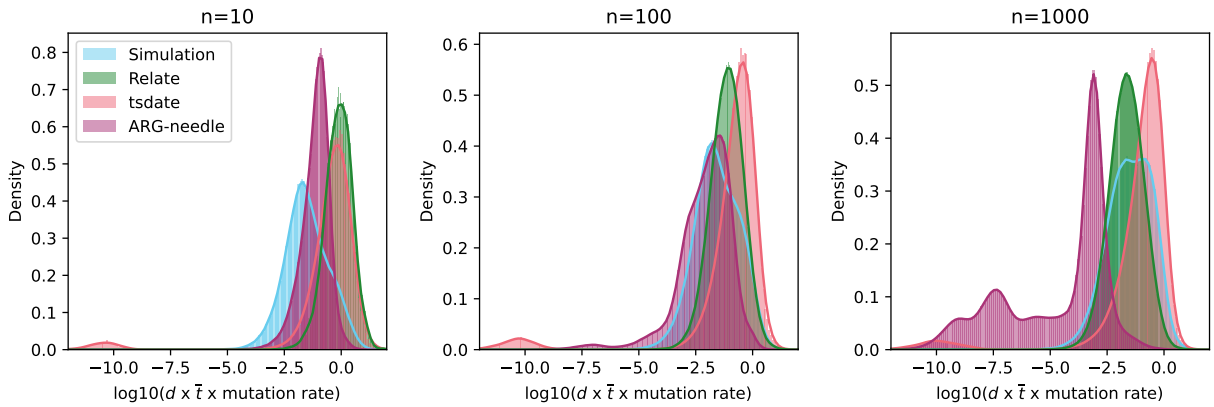

Figure S13: Histograms of expected number of mutations per edge for simulated and reconstructed ARGs. Histograms of (observed) expected number of mutations per edge, calculated as $(d^{\rightarrow}(b) - d^{\leftarrow}(b)) \cdot \bar{t}(b) \cdot \mu$ for each edge b in simulated and reconstructed ARGs (same as those in Figure 3). Note log scale on the x-axis.

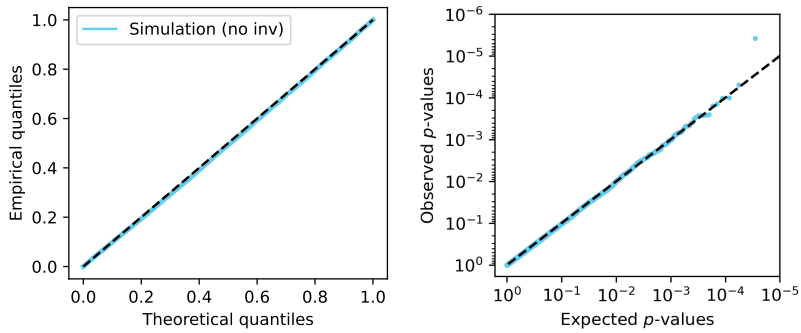

Figure S14: Q-Q and p-value plots (simulated ARG without inversion). Q-Q plot (left panel) and p-value plot (right panel) for ARG simulated using SLiM (without inversions and otherwise same parameters as in Section 4.6.2, main text). No clades have p-values below the Bonferroni-corrected significance threshold.

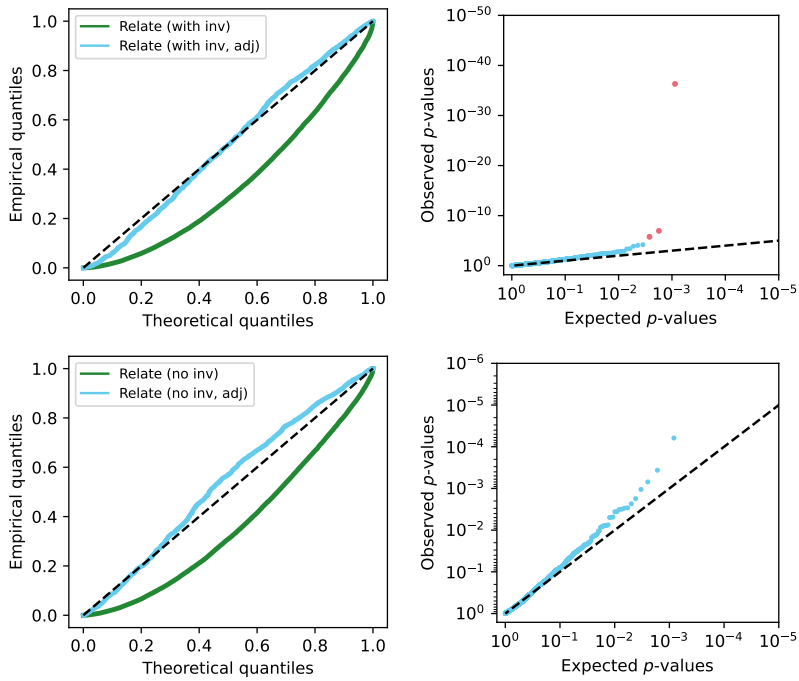

Figure S15: Q-Q and p-value plots (ARGs reconstructed using Relate). Top row: Q-Q plot (left panel) and p-value plot (right panel) for ARG reconstructed using Relate from data simulated using SLiM with one inversion under balancing selection (as described in Section 2.4.1). Blue (resp. green) points show values calculated after (resp. before) applying the adjustments described in Section S1.13.1; red points correspond to clades with p-value below the Bonferroni-corrected significance threshold. Bottom row: same using SLiM simulation without inversions (and otherwise same parameters).

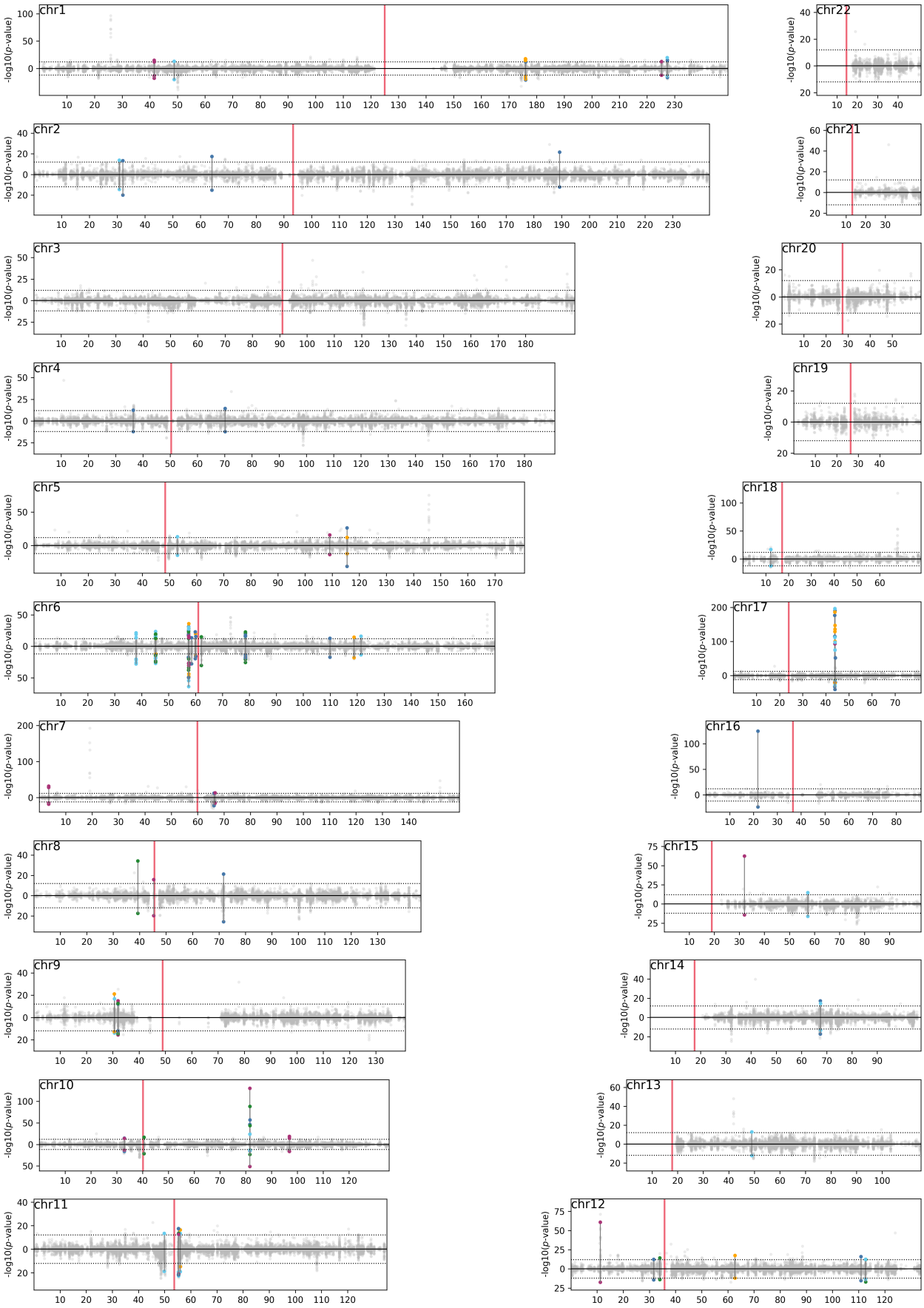

Figure S16: DoLoReS *p*-values for 1KGP ARG. Red vertical lines indicate positions of centromeres. See caption of Figure 6 (main text).

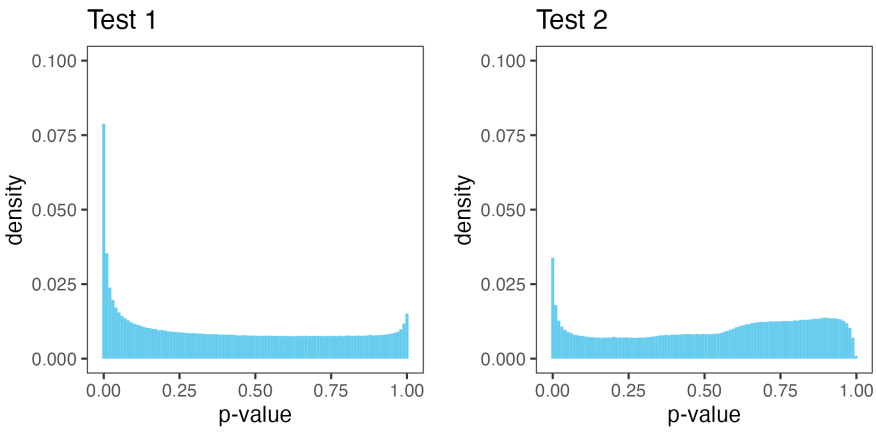

Figure S17: Histograms of p-values for Tests 1 and 2 for the 1KGP ARG (all populations combined, all clades with more than 2 mutations, at least 10 and at most $n-10$ samples, spanning at least 2 local trees).

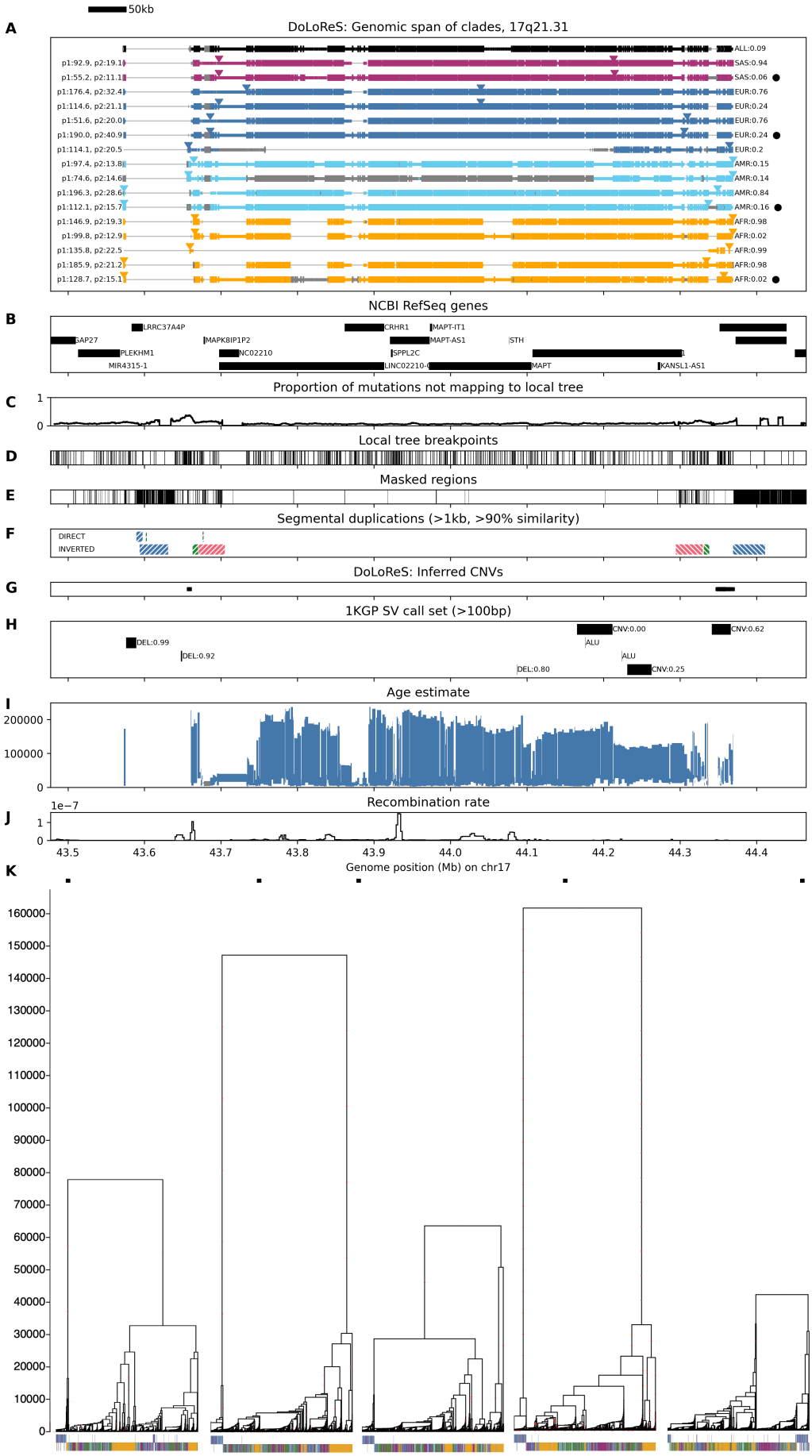

Figure S18: See caption of Figure 9 (main text). Age is estimated using the ARG subsetted to European populations.

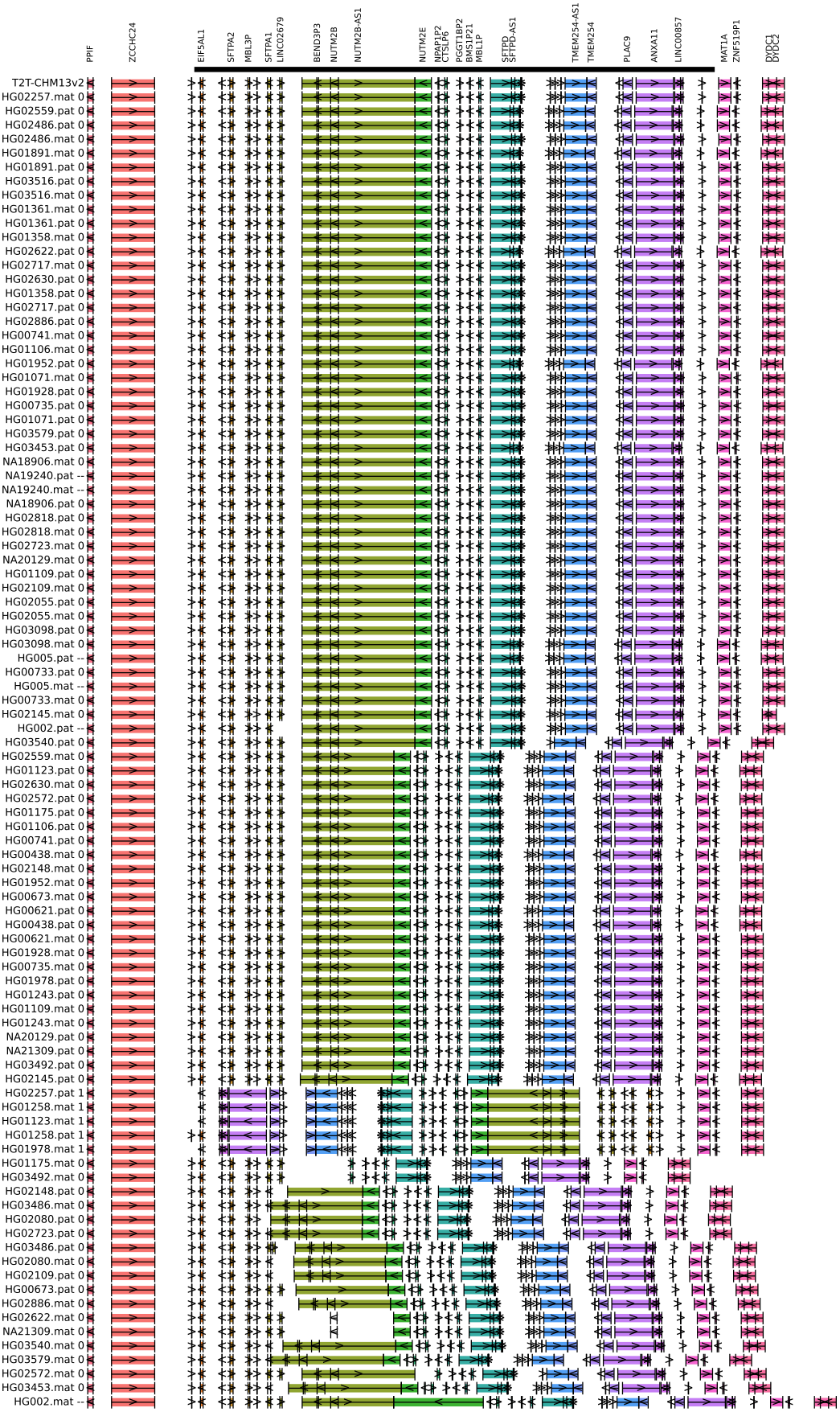

Figure S19: 10q22.3 inversion region for HPRC data and T2T-CHM13 reference. Each row is a sequence, labelled by the individual ID, whether it corresponds to the maternal (mat) or paternal (pat) haplotype, and the predicted inversion status (0 for non-carrier, 1 for carrier). Genes are coloured uniquely, from orange to purple left-to-right for the un-inverted (ancestral) orientation. Black bar shows predicted span of inversion.

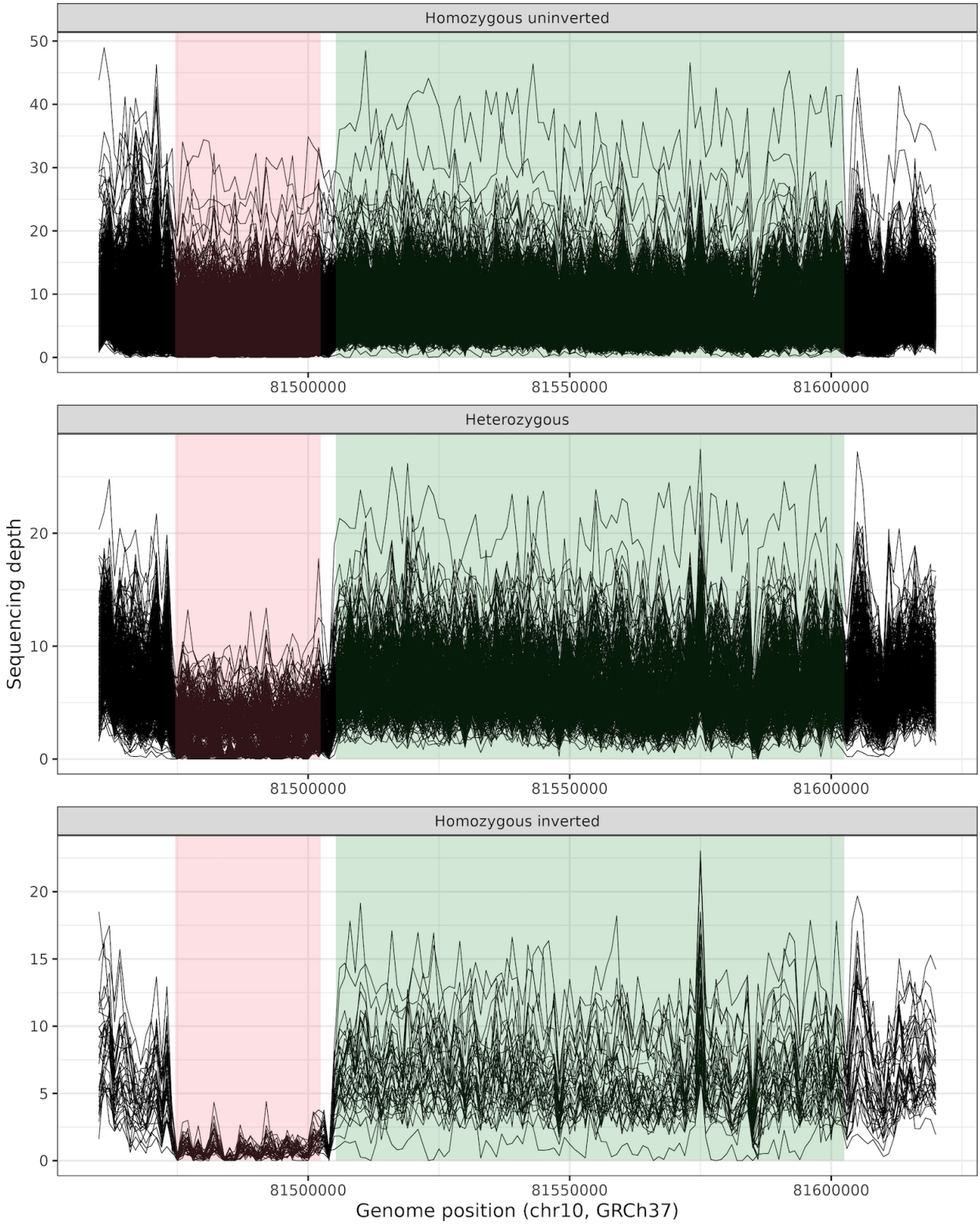

Figure S20: Sequencing read coverage on 10q22.3 around CNV1 (positions shown in red) and CNV2 (positions shown in green). Average coverage calculated in windows of 1000kb (each line corresponds to one individual) using 1KGP (Phase 3) low-coverage WGS GRCh37 data.

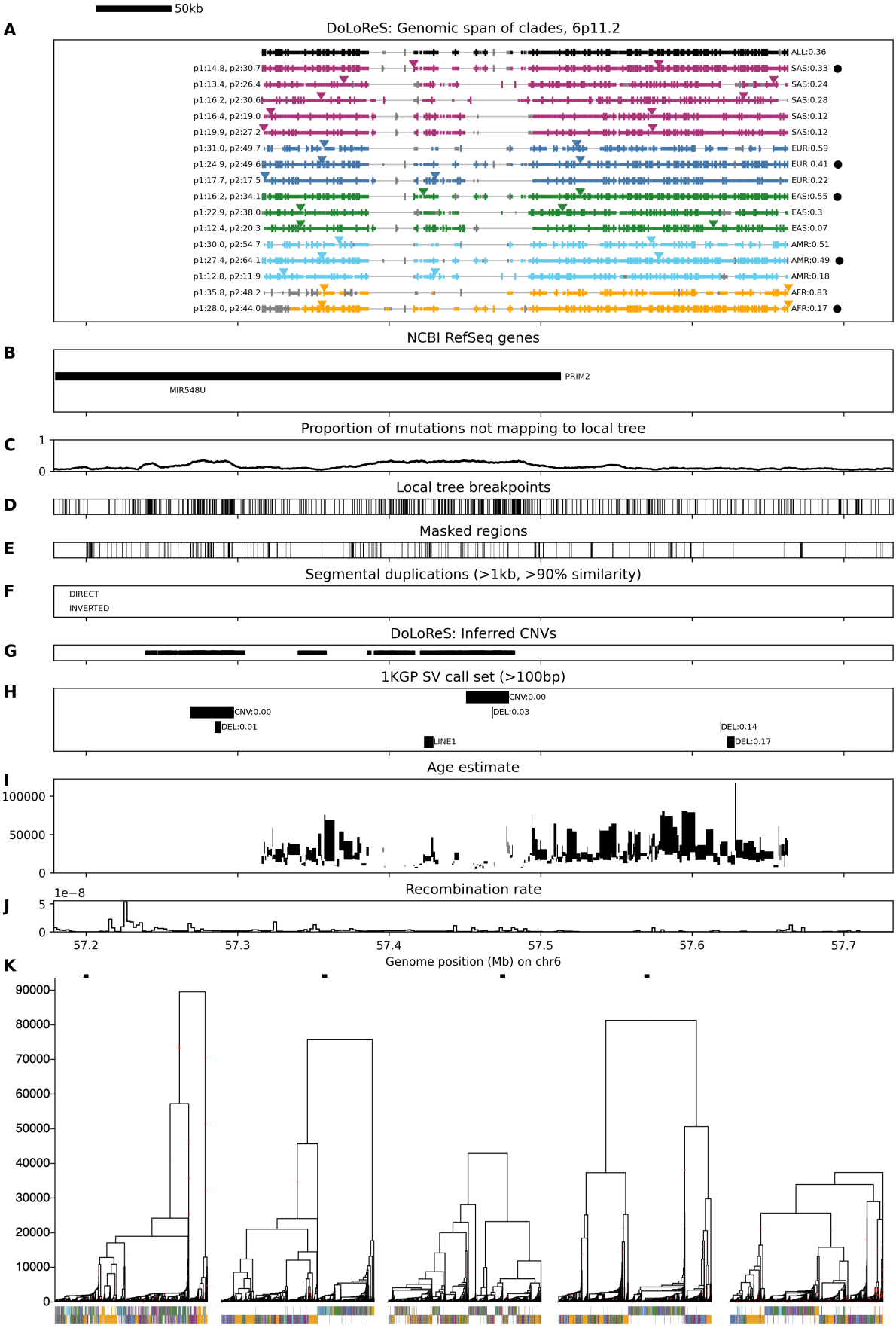

Figure S21: See caption of Figure 9 (main text). Age is estimated using the ARG for all populations.

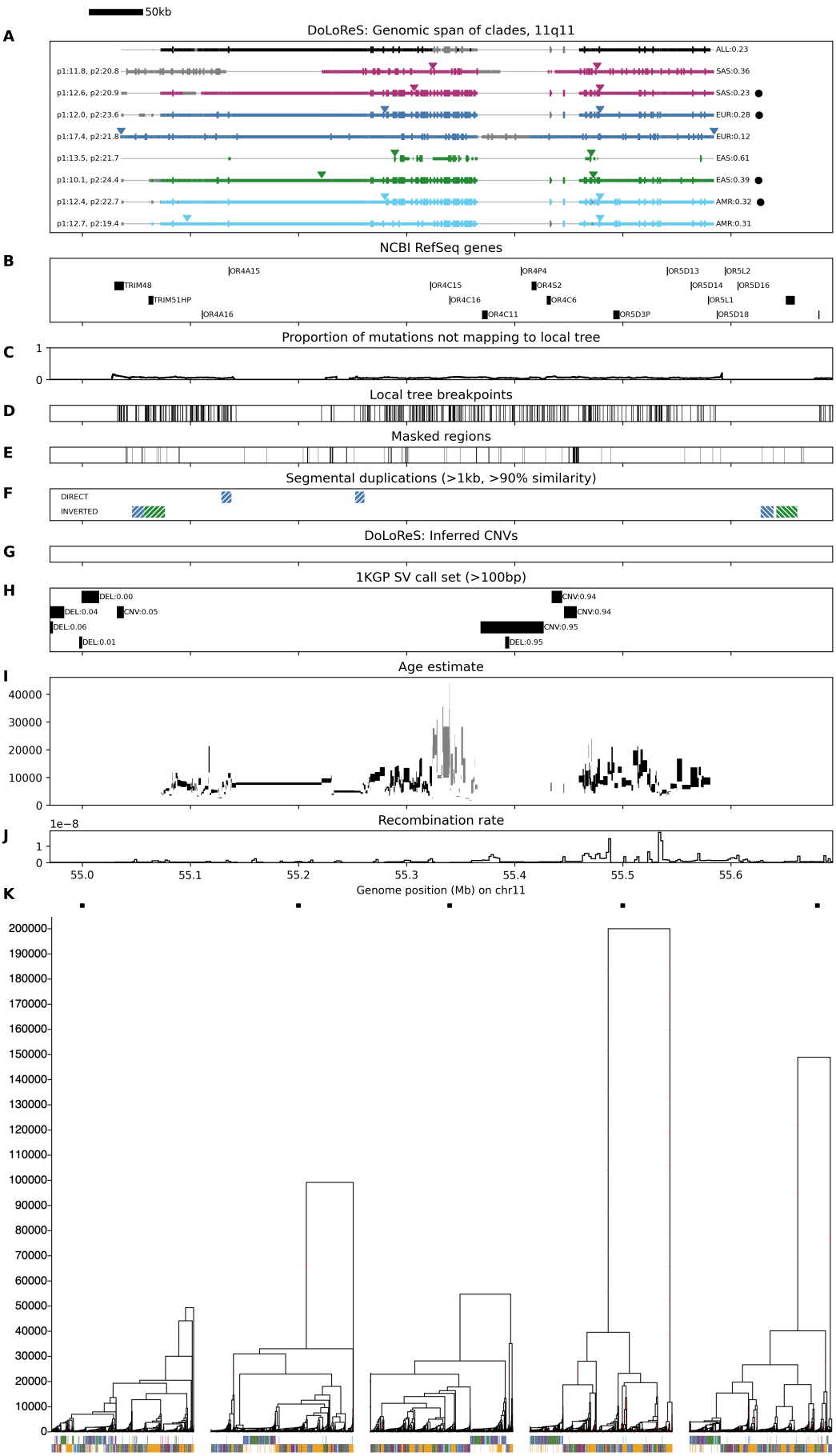

Figure S22: See caption of Figure 9 (main text). Age is estimated using the ARG for all populations.

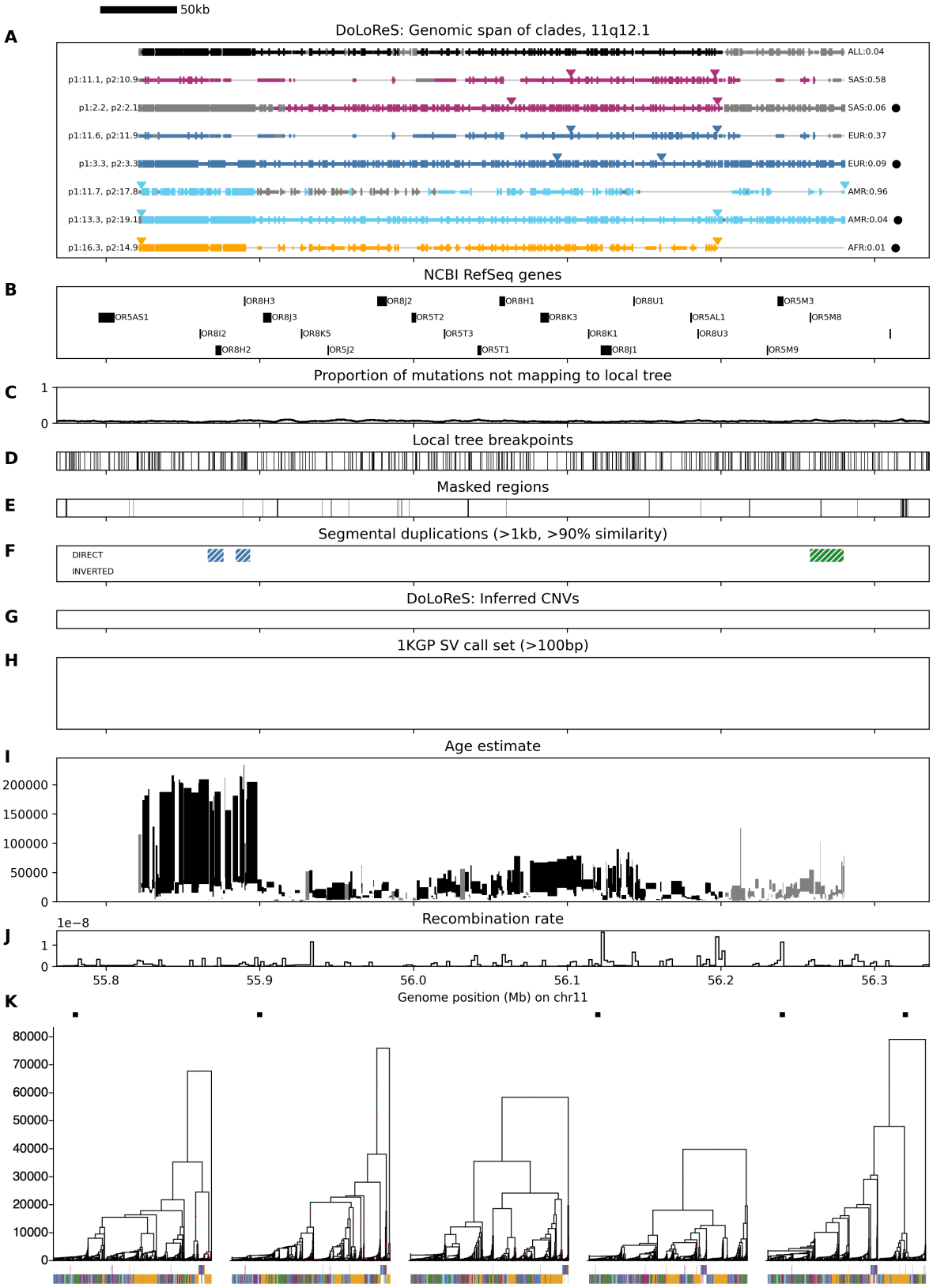

Figure S23: See caption of Figure 9 (main text). Age is estimated using the ARG for all populations.

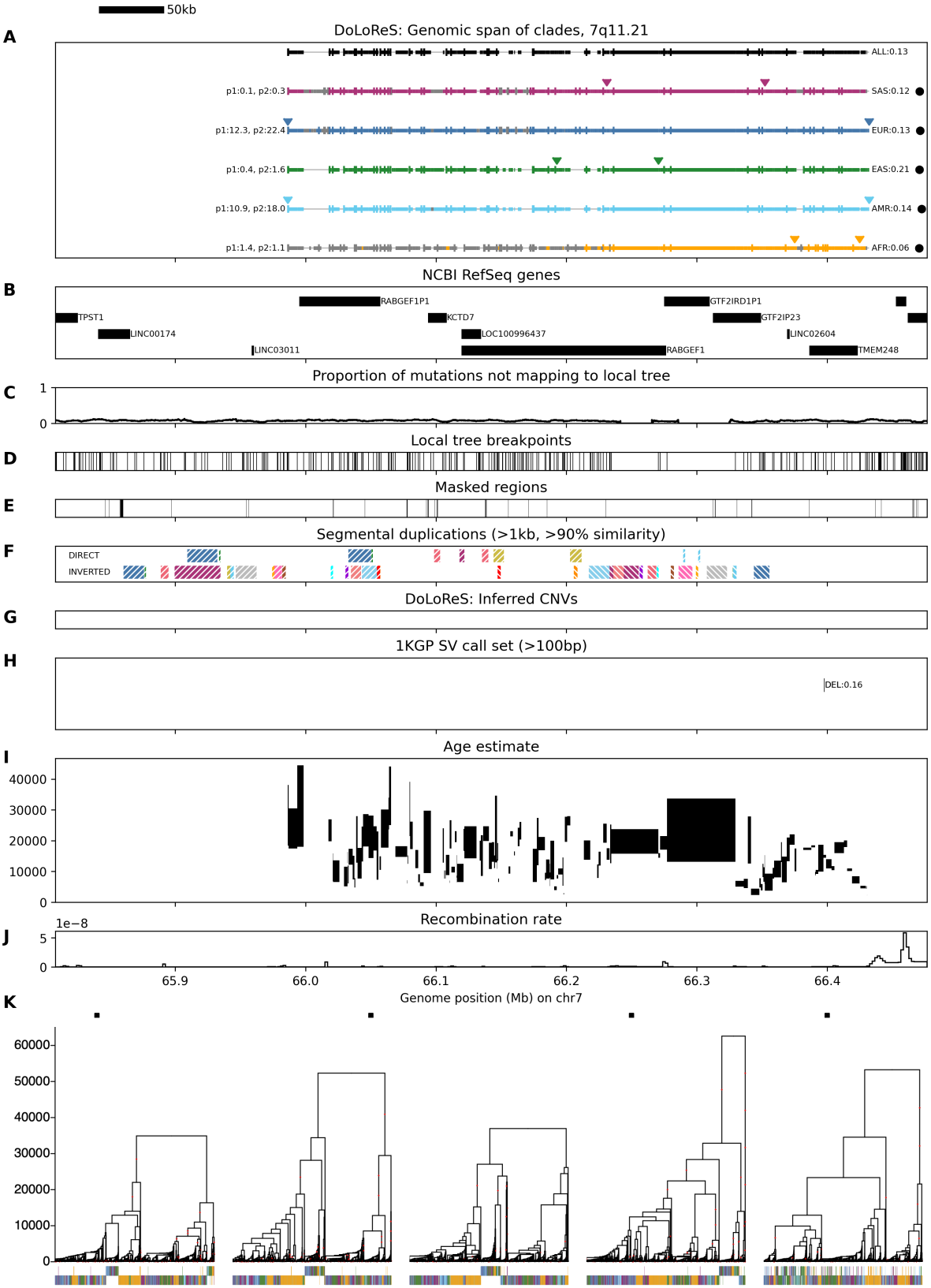

Figure S24: See caption of Figure 9 (main text). Age is estimated using the ARG for all populations.