On the origin and structure of haplotype blocks

**Task list 15/12/21:**

**To do:**

* Abstract
* Conclusion
* Add and cite references
* Sweep figures and text
* Refs
* Selective sweep update
* Read full text
* Mathematica notebook prepare for submission (GitHub?)

Meeting zoom link: <https://uu-se.zoom.us/j/9672768660>

### Abstract

…

### Introduction

One of the breakthroughs of long and linked-read sequencing technologies is the emergence of new methods for obtaining reliable haplotype information for large data sets [(Meier et al. 2021; ​​Lutgen et al., 2020)](https://paperpile.com/c/8P42ec/qEPZ). Although most studies of genome-wide variation still focus on SNP data, we are approaching the stage where population-scale haplotype information will be widely available for non-model organisms from across the tree of life. In light of this upcoming shift from site-based to haplotype-based inference, this article considers one of the fundamental concepts for haplotype-based inference—the definition of the haplotype block.

“Haplotype” and “Haplotype block” are widely used terms in evolutionary genetics, and have increased in importance for several reasons [(International HapMap Consortium 2005, Delaneau et al. 2019, Leitwein et al. 2020)](https://paperpile.com/c/8P42ec/qLoF+wHqb+zAV4). Most fundamentally, populations actually evolve through changing frequencies of blocks of the genome, rather than of individual sites. Therefore, we should be most interested in understanding the trajectories of the underlying haplotypes, yet these are not fully reflected by the SNPs that we see [(Castro et al. 2019)](https://paperpile.com/c/8P42ec/n0J1). Thus, disentangling the evolutionary history underlying genomic patterns can be challenging if based only on site-based statistics. For example, while whole-genome scans for signatures of selection can reveal regions that affect fitness [(Tavares et al. 2018; Poelstra et al. 2014)](https://paperpile.com/c/8P42ec/G3l1+Zh1w), it is hard to determine the actual causes of these signals [(Tavares et al. 2018; Ravinet et al. 2017; Rockman 2012; Stankowski et al. 2019; Wolf and Ellegren 2017; Burri 2017)](https://paperpile.com/c/8P42ec/G3l1+FvVp+fejK+wB8W+YfyE+iG6r). For example, shifts in polygenic scores from genome-wide association studies (GWAS) can be misinterpreted to be signals of selection instead of being artifacts of population structure [(Novembre and Barton 2018; Sella and Barton 2019; Berg et al. 2019)](https://paperpile.com/c/8P42ec/bzy7+Rv9L+okXB). Similarly, methods for estimating population density and gene flow struggle to distinguish among a virtually infinite number of possible population structures [(Whitlock and Mccauley 1999; Sousa, Grelaud, and Hey 2011; Richardson et al. 2016)](https://paperpile.com/c/8P42ec/PNU7+zolH+PSOI).

By accounting for haplotype structure, it should be possible to make inferences more accurate and more efficient. Haplotypes carry information from *mutation* as well as *recombination*, providing an additional ‘clock’ that can help to reveal past events. Primarily for these reasons, there has been a steady increase in analytic methods that aim to infer haplotype structure from sequence data, or that exploit haplotype structure to make inferences about selection, gene flow, and population structure.

Although there has been significant progress toward the broader use of haplotype information in empirical studies, much of this work is fragmented across many subfields, including evolutionary and conservation genetics [(Leitwein et al. 2020)](https://paperpile.com/c/8P42ec/qLoF), human and medical genetics [(Crawford and Nickerson 2005)](https://paperpile.com/c/8P42ec/0ygb), and animal and plant breeding [(Mészáros et al. 2021; Bhat et al. 2021)](https://paperpile.com/c/8P42ec/Cd8l+znGn). As a result, there is often little consensus on how haplotype blocks defined. More practically, this disparity complicates comparison of results, and may preclude insights that may otherwise arise by different perspectives.

The main goal of this paper is to critically examine the fundamental definition of the haplotype block. Specifically, we propose a definition of haplotype block based on the full genealogy, represented by the Ancestral Recombination Graph. Using simulations of simple but general scenarios, we briefly explore how the characteristics of haplotype blocks relate to the origin of the samples and segregating SNP variation. We then discuss how the proposed definition relates to the various ways that haplotype block sare defined in practice. We consider how different methods make use of haplotype information and construct haplotype blocks, discuss limitations and the assumptions that they make.

### Defining haplotype blocks

A haplotype has a clear definition: it is simply a haploid genotype (for example, the genotype of the sperm or egg). In contrast, the term “haplotype block" is used widely, but in many different ways (Zhang et al. 2002, Schwarz et al. 2004, The International HapMap Consortium, 2005; Taliun et al, 2014, Bkhetan et al, 2019). Our understanding of this term must depend on the processes of coalescence and recombination that generate haplotype structure. With this in mind, we contrast alternative definitions, and settle on one, which is based on branches in the underlying genealogy.

In sequence data, we usually observe the diploid genotypes; resolving them into the two haploid genotypes is termed "phasing". With *n* heterozygous sites, there are 2n possible pairs of haplotypes - more than a million with just n=20. However, there are usually just a few of them. Linkage disequilibrium (LD) across polymorphic sites, which often produces strong haplotype structure. This allows “statistical phasing”, through which one reconciles diploid genotypes into the underlying haplotype pair (Browning, Browning 2011). Looking across individuals in larger genotype panels, the more frequent haplotypes often appear as stretches of shared, “banded” blocks of SNPs. This is especially striking when different haplotypes become fixed across populations (see example in Fig. 1).

Whilst a blocklike structure may be apparent within data, we argue here that there should be a more fundamental definition of haplotype block that is based on the true ancestry of the sequences, independent of the mutations that generated observable SNPs. Thus, we separate the *definition* of haplotype blocks from the *estimation* of these blocks from actual data.

Haplotype blocks can be defined in a more concrete way via the classical concept of identity by descent (IBD; Hartl, Clark 1997, Carmi et al. 2013, Thompson 2013). Imagine an initial population, where each founder genome is labeled by a different colour. At some later time, each region of the genome must derive from one or other founder, and so will appear as a mosaic of blocks of different colors, each corresponding to their ancestors. This naturally defines blocks that descend from a given set of founders (Fig. 2). Fisher (1952) showed that the junctions between IBD blocks segregate like Mendelian variants, and used this idea to understand the distribution of runs of homozygosity. In artificial populations, we can now sequence the founders, and thus directly observe blocks defined in this way (eg. Gabriel 2002, Wallberg 2017, Lundberg 2017, Otte 2021). Moreover, if we disregard new mutations, the evolutionary processes subsequent to the founding of the population are entirely described by the block structure.

Identity by descent is defined with respect to a specific ancestral reference population. However, when we deal with natural populations, there is no obvious reference population, so the block structure will depend on our arbitrary choice of founders at an arbitrary time point. As Figure 2 shows, these arbitrary choices can complicate analysis, and it is not clear how to resolve the definition issue. Therefore, we will base our definition on the full ancestry of the sampled genomes, namely, on the ancestral recombination graph (ARG) (Hudson, 1983). The ARG consists of the segments of past genomes that are ancestral to our sample; looking back in time, it is generated by a series of coalescence and of recombination events (Box 1). We emphasise that these are real events: coalescence occurs when an actual individual leaves two or more offspring that are each ancestral to our sample, and recombination occurs between the two haploid parent genomes during meiosis in an ancestral individual. Together, these processes are embedded in the ARG (Fig. B1).

In large populations, and over long timescales, the ARG is approximated by the coalescent with recombination; in the simplest case, the rate of coalescence is the inverse of the effective (haploid) population size, and the rate of recombination is just the rate of crossover (). Importantly, the coalescent does not describe the entire genealogical relationship of the whole population. Rather, it summarises how the subset of sampled individuals are related to each other. Spatial and genetic structure can also be included: ancestral lineages carry a particular set of selected alleles (i.e., a particular genetic background), and are at a particular spatial location. Tracing back in time, lineages move between backgrounds by recombination, and between locations by migration.

Informed by the ARG, we could define a haplotype block as a contiguous region of the genome in which all sites share the same genealogy. However, adjacent genealogies differ by a single recombination event, and so blocks defined in this way will be vanishingly small (especially with large samples) and will usually differ trivially (see blocks A and on the Fig. B1). Moreover, as samples get larger, blocks defined in this way will become so small as to be impractical.

Instead, we define a haplotype block as the set of genomic regions that descend from a particular branch in the ARG; this branch is defined by a unique coalescence event. Crucially, such regions should carry a shared set of derived SNP alleles that arise on the focal branch that just precedes the coalescence event. With enough SNPs, the haplotype block is revealed directly by these shared SNPs.

### Implications of the definition

We next elaborate on the definition and illustrate the relationships between genealogies, SNPs and haplotype blocks using example simulation. Figure 3 shows a neutral example capturing the ancestry of 10 genomes, sampled from a population of 100 haploid individuals, across 10cM of the genetic map. SNPs were generated by infinite-sites mutation with mutation at twice the rate of recombination. This simulation is general (Hudson, 1990 ref), because time and map distance both scale with population size. Thus, the 268 generations taken for every part of the simulated genome to coalesce in a single common ancestor scales to 2.68N, and the simulated map length scales to 10/N. Thus rescaled, this simulation shows a generic pattern, independent of population size.

The central panel of Fig. 3 (SNPs) shows the distribution of SNPs on the ten sampled genomes, coloured according to the branch on which they arose (we illustrate 8 branches with four or more SNPs each, out of 55 unique branches). The genome is divided into 34 non-recombining intervals, but it contains only 24 different genealogies, because some longer genealogies were split into multiple intervals by intervening recombination events (Fig. 3; trees). This illustrates how recombination interacts with the coalescent (also see Fig.B1 for schematic representation of the process). If we disregard branch lengths, the trees can be further simplified into 15 distinct topologies shown in the top panel Fig. 3 (trees and corresponding regions on the genome labeled a - o). For illustration, we show one pair of genealogies that have the same topology, but differ in depth (k1 and k2), B in Fig. 3) .

The coloured blocks shown in the lower panel of Fig. 3 (Blocks) illustrate the extent of each branch along the genome, and through time. The mutations arising on each branch are projected onto the block at the time and genomic position that they arose. The number of SNPs arising on each branch is Poisson distributed, with the expected number proportional to the area of the block; this area is the sum of the lengths of the genome that each ancestor carries, and that is ancestral to the coalescence event that defines the branch. We emphasise that this is a random process, so some regions may not carry any informative SNPs. For example, though branch i (light blue) is relatively well covered by 9 SNPs, none of them fall in the shallow region to the left (C). Similarly, branch ii has only 6 SNPs, none of which happen to fall in the rightmost region (D). Ultimately, the distribution of SNPs sets a limit on what can be inferred from sequence data; branches without mutations will be invisible to us, and our ability to infer the length of a block depends entirely on where mutations happen to fall.

Each branch coincides with a specific coalescence event that brings together a specific set of lineages: in other words branches are defined by both the coalescence event *and* the set of lineages. A single coalescence, i.e. a single ancestor, may generate multiple branches: the two genomes that come together in that event may carry a mosaic of ancestral material, in several combinations. A single coalescence event may even generate a branch that carries disjunct segments of the genome, ancestral to the same set of descendants (see the schematic representation on the Fig. B1). This did not occur for any of the focal branches in the example of Fig. 3, but is not unlikely, especially in a selective sweep (see simulation). Conversely, two different coalescence events may happen to bring together the same sets of lineages; their branches would be hard to distinguish.

Because each branch is generated by a single coalescence, it begins at the same time across its whole extent (so, branches are bounded by a horizontal line at their base in the lower panel of Fig. 3). Recombination events split distal segments, thus limiting the span of the block. Tracing back in time, branches must end in coalescence events that combine them with yet more descendants. These may occur at different times if there have been recombination events.

Haplotype blocks overlap in their genomic extent, since multiple lineages exist at any time after the MRCA; this is shown by the overlapping 3-D blocks in Fig. 3 (Blocks). Because they correspond to branches in a genealogy, blocks can show a nested structure. For example, branch i, which is ancestral to {4, 8} descends in the middle part of the genome from branch i (blue), which brings together {4, 7, 8, 10}. Thus, branch i is nested above branch ii in Fig. 3 (see also F for another example of nesting).

If we start at a particular point on the map, and work along the genome, at some point a branch will be split by a recombination event; the new lineage will trace back and eventually coalesce, most likely ending the branch. The rate of recombination is proportional to the branch length, and so we expect that if a branch traces back deep into time, it will span a short region of the genome. Conversely, shallow branches will extend over a longer genomic span. This pattern is seen clearly in Fig. 3 (lower panel), where branches consist of segments that are either deep and narrow, or shallow and wide. However, this relationship is not *precisely* inverse; if it were, blocks would tend to have the same area, whether they were deep or shallow, and hence would carry similar numbers of SNPs. In simulation, deep branches tend to be wider than expected from the naive argument given here, and so most SNPs are on a few deep branches.

Note that under the coalescent process, large numbers of sampled lineages rapidly coalesce down to a few, which are then likely to trace back deep into the genealogy. Thus, in a given region of the genome a substantial fraction of SNP will fall on long, deep, branches, whereas the tips of the genealogy will be hard to resolve. Moreover, in a large sample, it is unlikely that different coalescence events will bring together exactly the *same* set of lineages by chance, so that we can usually identify unique coalescence events as corresponding to unique sets of lineages.

Figure 3 illustrates the simplest case of the standard coalescent with recombination. In reality, population structure and selection complicate genealogies. For example, in the island model, lineages either coalesce quickly within a deme, or escape to coalesce much further back in time. This exaggerates the tendency for genealogies to be dominated by a few long branches. Selective sweeps have a somewhat similar effect. In the classic case (Maynard Smith and Haigh, 1974), all lineages at the selected locus coalesce in the individual that carries the favoured mutation. Moving out from this locus, recombination frees lineages to coalesce much further back.

Figure 4 illustrates such a selective sweep. On average, the sweep reduces diversity around the selected locus. We observe longer haplotypes blocks around the causal locus, however one should note, that overlap between them is still available. As we move further from the recombining region, selected branches start to show deeper coalescent times. One way of looking at the inference of selection can be looking at the average time to MRCA, which is available under our definition. In Figure 4 (E). We see shallow genealogies around the focal locus. However, in any one realisation, its effect is highly random, making it relatively easy to identify the causal *branch*, but impossible in principle to precisely locate the causal mutation/SNP (see example genealogies 1 and 2 on figure 4B).

### The definition in practice

Having defined haplotype blocks conceptually, we next consider the problem of inferring haplotype blocks from empirical datasets. Current sequencing and genotyping technology makes it straight-forward to call SNPs or indel variants, but it remains non-trivial to connect these to the haplotypes in which they are embedded. For that reason, sophisticated algorithms have been developed for phasing, genotype imputation and inference of genealogies (Marchini, 2007, Kong, 2008, Browning, Browning 2009, Howie, 2011, Browning, Browning 2013, Davies 2017). These tasks all engage different facets of the same problem, and rely to some extent on haplotype structure. However, these methods tend to focus on phasing and most stop short of inferring haplotype blocks as we define them. In view of how central genealogy (and the ARG) is in shaping haplotype blocks, we will focus our discussion around a selection of methods that make active use of the ARG and its approximations. We will discuss the underlying assumptions of these genealogy-based methods, but also where they could be extended in light of our proposed haplotype block definition. Separately, in Box 3, we also outline families of simpler methods that use fixed genomic windows or genomic segments (and their limitations).

The full ARG contains information about branches of the genealogy and, in theory, encodes all the information needed for applying the haplotype block definition to empirical datasets. Therefore, a direct (but impractical) way to define and analyse haplotype blocks in a dataset would be to infer the full ARG from the sample of sequences. Nevertheless, as we will soon see below, the state space of every possible ancestral history of a sample of genomes is effectively infinite, so inferring the ARG in its full form is intractable. Instead, most practical methods rely on various trade-offs to simplify the problem.

For direct inference of ARG (ref), ARGweaver (Rasmussen et al, 2014) and its extension ARGweaver-D are among the most powerful, and widely used. ARGweaver solves the infinite state space issue by discretizing time, effectively making the ARG space finite by limiting recombination and coalescence events within discrete time points. Further, ARGweaver uses a coalescent-with-recombination model (Sequentially Markov Coalescent, or SMC; McVean & Cardin 2005; Marjoram and Wall 2006) to sample from an ARG distribution. While making inference more tractable, SMC precludes overlapping blocks as we have proposed here. Besides the limitations due to binning and non-overlapping blocks, inference of the “full” ARG is still computationally expensive, thus ARGweaver is most suitable for mid-sized datasets on the order of fifty sequences.

Two recent methods*,* tsinferand Relate (Kelleher et al, 2019, Speidel et al, 2019), have attempted to approximate the ARG in much larger populations with thousands of samples by focusing on topology (or ‘succinct tree sequences’), rather than a full inference. They do so by representing genomes as a series of tree topologies: Relate as distinct trees; tsinfer as ‘tree sequences’ connected via ancestral haplotypes. Both achieve this remarkable speed-up by relying on the Li-Stephens model (2003; see Box 2 for further details) to infer local pairwise distances (Relate) or ancestral haplotypes (tsinfer). As an added advantage, tsinfer doubles as an efficient, lossless compression algorithm by indexing population genomic variation SNPs-on-trees as opposed to the highly-redundant, traditional SNP-by-individual matrix. Put in another way, the tree sequence encoding can fully capture the variation data in entire populations, for a fraction of the storage space. Such a representation also effectively syntheses a number of population genetics summary statistics. These developments can be very interesting, as national sequencing projects become routine.

Among practical methods, tsinfer and Relate represent the state-of-the-art in representing large populations. All three approaches, including ARGweaver, approximate ARG well, and give accurate coalescence time estimates under simulation (Brandt et al., 2021). For our purposes, they are also useful approximations of ARG that highlight key advantages we wish to emphasize in our haplotype block definition. For example, Relate presents a selection statistics suite that goes beyond SNP information. One advantage of Relate is that branches are dated, as opposed to a strict encoding of topology only in tsinfer. Having dated branches allow, among other things, the possibility of estimating temporal changes in mutation rates. Another useful feature, in our view, is tsinfer’s placement of SNPs onto branches, which is the key feature that distinguishes haplotype blocks from each other under our definition.

Already, we note that efforts are underway to address limitations of the methods. For instance, tsdate now adds coalescence times estimates and branch lengths from tsinfer’s output (Wohns et al., 2021). In the context of our exploration of haplotype blocks and their overlapping structure (Fig. 3C, D), we note that they can be captured rather poorly under the Li–Stephens models in tsinfer and Relate, in a way that may bias the inferred ARG.

In contrast to the methods discussed above, HaploBlocker is an inference method that does allow for overlapping haplotype blocks (Pook et al., 2019). HaploBlocker internally defines a large set of overlapping blocks and then selects the most relevant ones based on the ratio of block length and haplotype frequency. The authors emphasize that the inferred haplotypes are subgroup-specific, which is analogous to selecting branches of the related haplotypes. The main weakness of this method is that genealogical information is not considered.

Taken together, there has been a recent spurt in innovation in genealogy/ARG-based methods. Among these, ARGweaver arguably comes closest to inferring the full ARG, at great computational costs. Both tsinfer and Relate are robust and scalable to thousands of samples with minimal, reasonable tradeoffs, but infer haplotype blocks only as an incidental output. Ultimately, we hope our discussion here will encourage development of new methods to infer haplotype blocks as we define them.

Assuming that a method becomes available for inferring blocks as we have defined them, there are still practical considerations that we need to face. For example, we see from Fig. 3 that haplotype blocks, defined via branches in the genealogy, have a complex structure, tracing back in time for a number of generations that varies along their span (e.g., blocks ii and iii). This makes it hard to define the haplotype blocks in any simple way. Should this be their maximum length, or should it rather be weighted by the depth? It is not clear which definition would be better for inference. These kinds of issues could be investigated with more detailed simulations of more diverse processes than were examined here.

### Conclusions and outstanding questions

Thus far, we have outlined a definition of the haplotype block, explored the implications of the definition with simple simulations, and considered how current methods can infer such blocks from empirical data. Indeed, it is our view that haplotypes and haplotype blocks should be the core unit through which we understand population genetic processes. If we adopt this view, it follows that ideally, genomic datasets should come directly as resolved haplotypes, rather than diploid genotypes that require phasing and further processing. We therefore welcome new developments in linked- and long-read sequencing techniques and softwares that are designed with sequencing and population datasets in mind (Meier et al., 2021; Davies et al., 2021).

Our simulations show that haplotype blocks contain rich information about the demographic and selection history of the locus. To make the most of this information, it will require a fundamental rethink of our linear, reference-based genome assemblies, and move towards a graph-based assembly standard. We will also need new concepts and vocabulary to describe features in these graphs (e.g., super-graphs, “bubbles”, etc.) informed by a robust understanding of the generative process discussed above, and align our mental models with inference schemes and their encoding, e.g., tsinfer. For that reason, we hope our discussion here can focus our effort towards making sense of this new standard, as haplotype-resolved sequencing becomes a matter of routine.

### Acknowledgements

FWF Wittgenstein, SCAS, Barton Group

### Figures and captions:

#### 

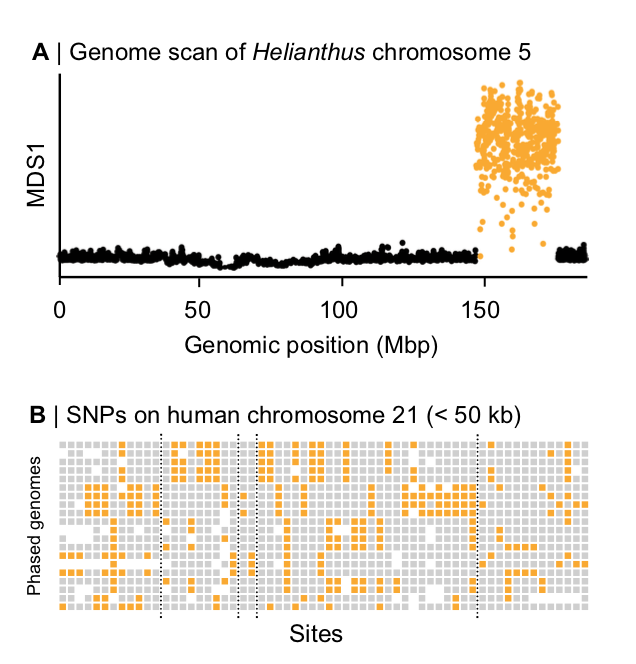


Figure 1. Block-like patterns in empirical data. (A) A local PCA analysis conducted on SNP data from wild sunflowers reveals a large ‘haploblock’ on chromosome 5 (Todesco et al., 2020). Each circle is the MDS1 score for each window, illustrating variation in the PC results for each window (See Li and Ralph (2019) for more details on the local PCA method). The large block of distinct MDS1 scores implies that large, non-recombining haplotypes reside in this region, without the need to observe the haplotypes directly. B) Haplotype blocks on human chromosome 21 are visible in phased SNP data (Patel et al. 2001). Rows show 20 individual haplotypes. Each column is a site with gray and yellow squares representing ancestral and derived sites, respectively. White areas are missing data, the dashed lines demarcate ‘blocks’ that contain multiple distinct haplotypes.

#### 

#### Figure 2. IBD definition

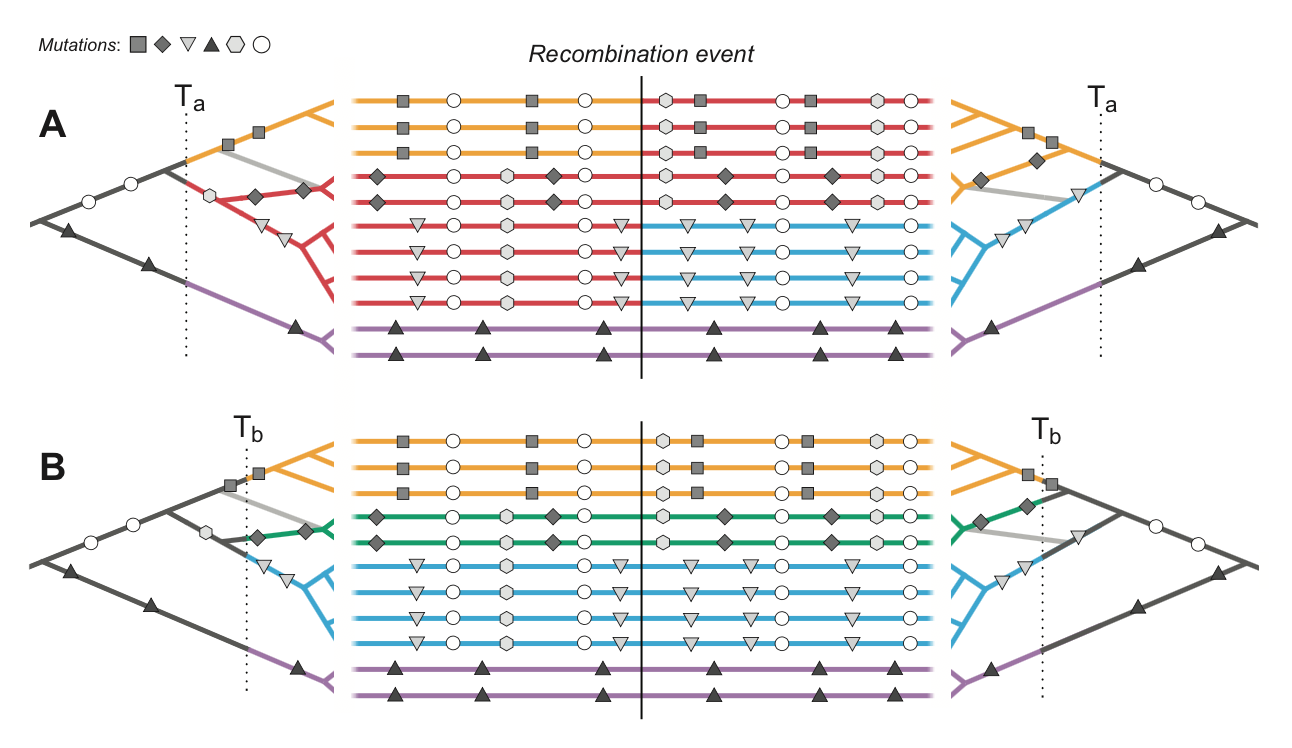


Figure 2. Haplotype blocks defined through identity by descent (IBD). Panels A and B show the same 11 hypothetical DNA sequences depicted as horizontal lines. The trees on the left and right sides show the genealogy for the set of sequences either side of a recombination event (indicated by the vertical black line); the light gray branch in both trees shows the effect of recombination in the genealogy. Mutations are shown as symbols that correspond to the branches upon which they arose. Under the IBD definition, haplotype blocks can be defined based on DNA segments that derive from a given set of ancestors, shown here by the coloured sections of branch and DNA sequence. The only difference between panels A and B is that these ancestors are defined at two different arbitrary time points, Ta and Tb, yielding different haplotype structure.

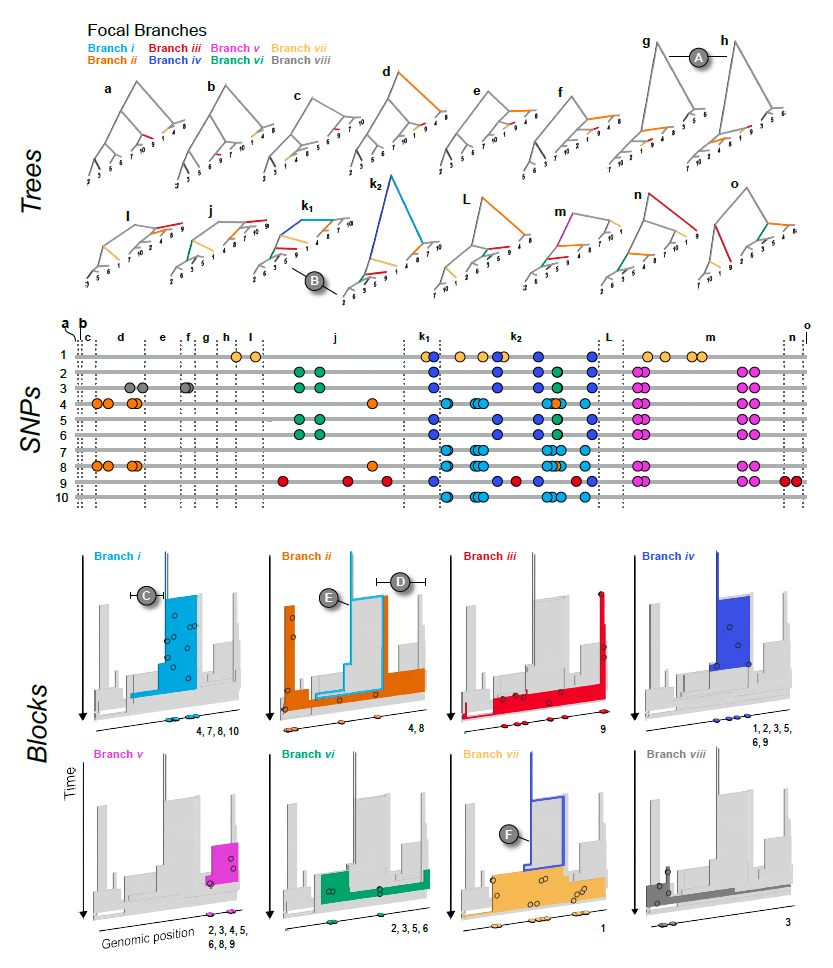
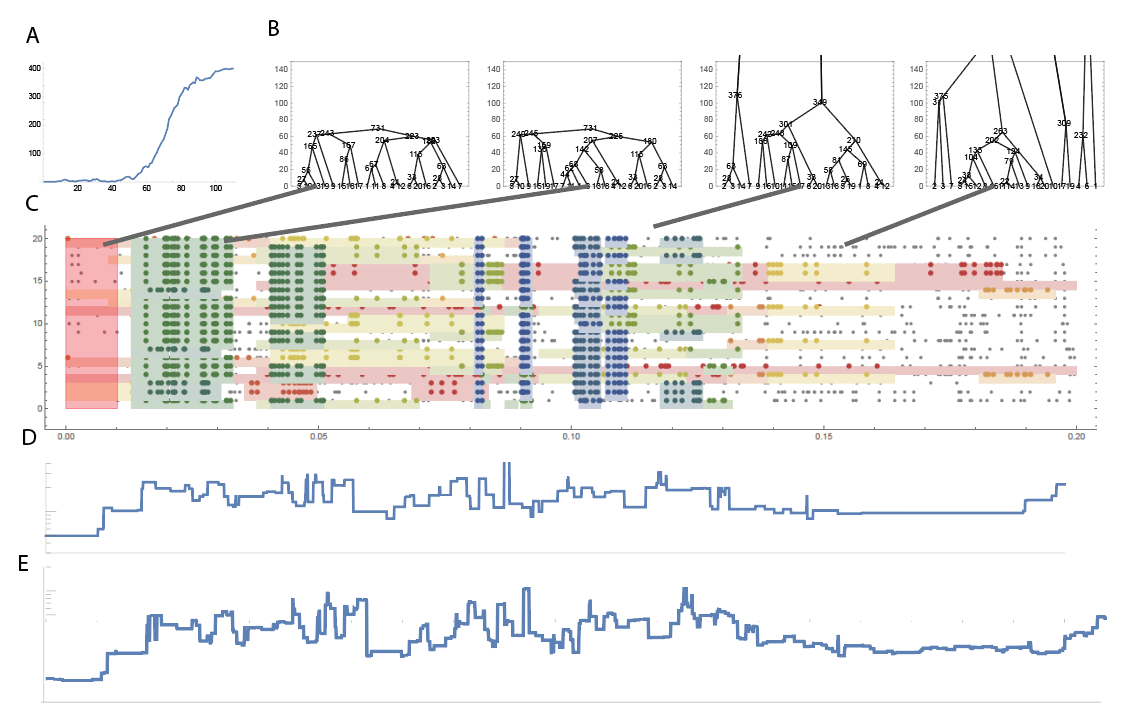


Figure 3. The relationship between trees (top), SNPs (middle), and haplotype blocks (bottom) in the neutral simulation (see main text for simulation details). The trees (a - o) show all of the unique topologies that coincide with the genomic spans shown in the central panel (also labeled a - o). The 8 branches that we focus on in this example are coloured and labeled i – vii. (A) Two neighboring topologies that differ only slightly due to recombination. (B) An example of two trees (k1 and k2) that have the same topologies but different branch lengths. The central panel shows 10 haploid genomes (labeled 1 – 10, top to bottom, coinciding with the tips of the trees). The SNPs that arose on the 8 focal branches are indicated by the coloured circles. The lower panel (Blocks) shows the haplotype blocks for each focal branch. The coloured block in each panel is the focal branch, with the other 7 blocks shown in gray. The mutations shown in the central panel are projected onto each block (black circles) at the genomic location and time that they arose. They are also plotted onto the genomic position axis to make the connection with the center panel mode explicit. Similarly, the numbers at the bottom right corner indicate which DNA sequences the mutations are associated with. (C & D) Examples of regions of blocks that, by chance, are revealed by mutations arising on the corresponding branch. (E & F) Examples of nested blocks, where the ancestral block is highlighted with a coloured outline.

#### Figure 4. Sweep simulation



Caption

### Boxes:

#### BOX 1: Ancestral Recombination Graph (ARG)

The ARG describes the complete ancestry of a sample of genomes through a series of real coalescence and recombination events (Hudson 1983, Griffiths, Marjoram 1997). At any given site on the genome, the relationship can be described through a genealogy (Kingman, 1982); all contemporary samples coalesce and eventually trace back to one single ancestor. Moving along the genome, the relationship inevitably changes due to recombination. This leads to a series of observable genealogies along the genome (Fig B1A), which are embedded in a single structure - the ARG (Fig B1B).

The full ARG (Fig B1B) is a graphical representation with nodes as ancestral or contemporary genomes and branches as ancestral lineages. Branches can either split due to recombination or merge due to coalescence, so that the number of ancestral lineages fluctuates. Each node represents a real coalescence or recombination event, whilst branches represent the ancestry of a particular genomic segment, along a genetic lineage (depicted by coloured/grey segment for inherited/non-inherited genetic material in Fig B1B). Altogether, the full ARG describes the entire ancestral history - each recombination and each coalescence event, which imply the genealogy for each non-recombined genomic block. Crucially, the ARG describes ancestry but not allelic state, so is independent of all the mutations that lead to the observed polymorphism in the present sample.

It is important to note that the full ARG (Fig B1B) contains more information than the series of tree sequences along the genome (Fig B1A). First, a series of tree sequences lack information on the timing of recombination events, unless these are separately stored. Second, while some recombination events lead to observable changes in genealogical trees, others might not. Figure B1A depicts such cases - some recombination events might not change the tree topologies at all (trees *ii* and *iv* are exactly the same), whereas others might only lead to temporal changes in coalescence nodes (tree *i* differs from trees *ii* and *iv* by 1 node position, but all have the same topology). Therefore, while there are 4 non-recombining genomic regions, there are only 2 unique tree topologies (trees *i*, *ii* and *iv* have the same topology) and 3 distinct trees (trees *ii* and *iv* are exactly the same). Some coalescence events can also be entirely invisible and not be represented in any of the individual trees – coalescence at *t2* in Fig B1B is not represented in the series of trees in Fig B1A. Furthermore, two disjunct blocks of the genome can be inherited from the same ancestor, so that a unique coalescence event (e.g. marked by \* in Fig B1A) can generate disjunct blocks of ancestry. It should also be noted that although Fig B1 shows the inevitable coalescence of the whole genome into a single common ancestor, this typically takes an astronomically long time: each non-recombined region of the genome coalesces at various time points, and the single lineages ancestral to each region then take an extremely long time to coalesce in one common ancestor, in a process which is in principle unobservable.

Since the ARG contains full information about the genealogy of the sample, it is in theory sufficient to infer any evolutionary process: the ARG necessarily gives more information than commonly used statistics like SFS, Fst, EHH, which are low-dimensional summaries of the ARG (Ralph et al, 2020). Therefore, the ARG should serve as the foundation for developing new methodologies. However, we note that whilst the ARG is a sufficient statistic, it remains an open question how much the extra information it gives can improve inference: the intrinsic variability of the evolutionary process sets a bound on the accuracy of our inferences.

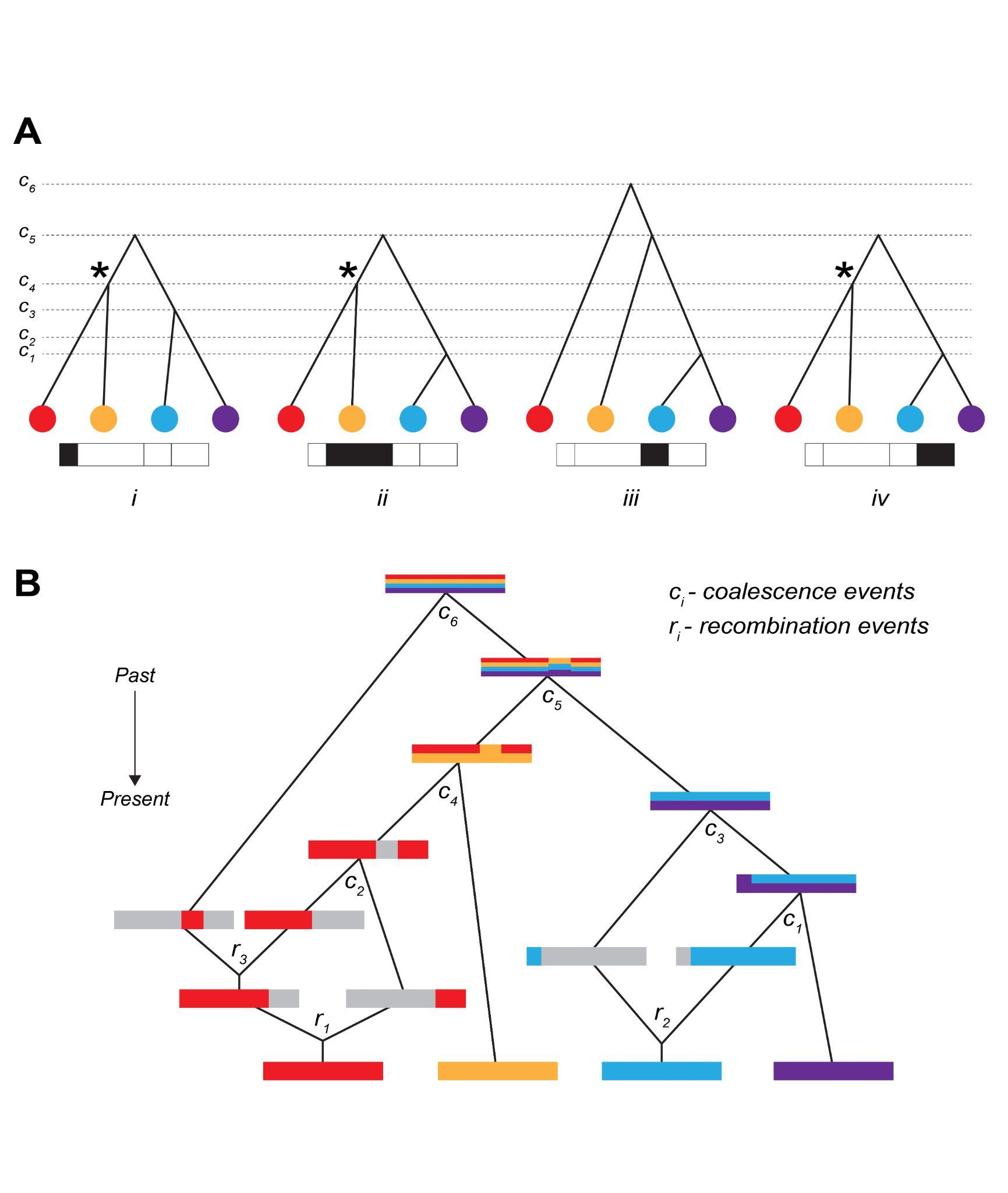


Figure B1. Caption

(A) Genealogical trees along the genome, corresponding to the above ARG - each tree describes the ancestral relationship for each of the 4 non-recombined regions. *t1,t2,...,t6* denotes time points for each coalescence event. Trees can either have the same topology, or marginally differ by only temporal positions of coalescence nodes. Asterisk (\*) denotes a unique coalescence event that is ancestral to disjunct genomic regions. (B) Full representation of Ancestral Recombination Graph (ARG) - Tracing back ancestry of four genomes, there is either recombination splitting lineages or coalescence merging lineages. Inherited ancestral genomic regions are coloured corresponding to the contemporary genomes. Recombination is represented by splitting the genome into two; where grey denotes non-ancestral genomic region. Coalescence is represented by two genomes merging, with inherited genomic regions denoted by mixed colours. There are 3 recombination and 6 coalescence events in the full ancestral history of the four genomes. *t1,t2,...,t6* denotes time points for each coalescence event. *r1, r2, r3* denotes time points for each recombination event.

#### Box 2: Application and limits of Li and Stephens Algorithm

Li and Stephens (2003) (LS) proposed a hidden Markov model (HMM) framework that underpins a large proportion of existing inference methods. Originally developed to model patterns of linkage disequilibrium, it has since been widely applied to develop analytical tools and address empirical problems, such as, phasing and imputation of genomic data (Browning and Browning 2007, 2015; Stephens and Scheet 2005; Marchini et al. 2007; Howie et al. 2009; Li et al. 2010), inference of population structure and demographic history (Lawson et al. 2012; Steinrücken et al. 2018, 2019, Hellenthal et al. 2014), characterisation of local admixture (Sundquist et al. 2008; Price et al. 2009), inference of local genealogies (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019), and many more. The LS HMM framework is highly tractable and efficient. However, underlying assumptions make it incompatible with the haplotype definition we propose.

The LS algorithm requires a reference sample of haplotypes. It gives a framework to decide whether some focal haplotype represents a) an entirely new haplotype or b) a mosaic of previously encountered haplotypes, and determines the breakpoints and transitions in this mosaic. Whilst the Li Stephens model captures genetic relatedness among chromosomes through recombination, it assumes that the reference haplotypes are known. This would be valid in a selection experiment, if we know the founder genomes; in this case, blocks are defined by IBD to this reference population. However, if we only have contemporary genomes, the reference panel is an approximation. Secondly, the model assumes that genomic states depend solely on the immediately preceding site. This is also an approximation, since in the true ARG, recombinant lineages can coalesce back to any lineage that existed in the preceding genome.

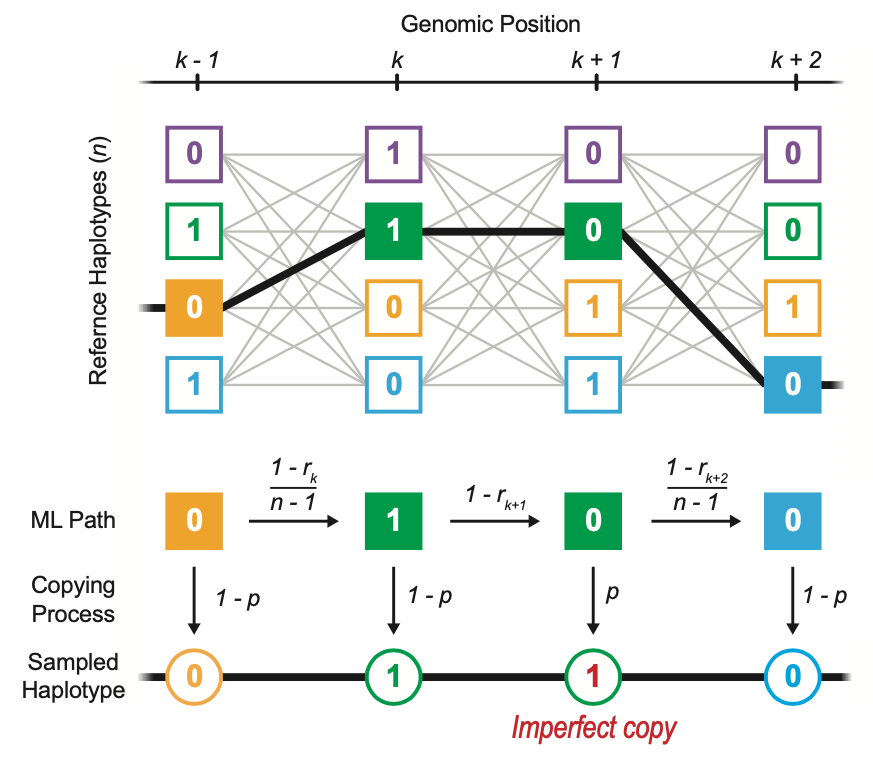


Figure B2. Caption

Schematic representation of Li and Stephens hidden Markov model. A new haplotype as an imperfect copy of *n* previously existing reference haplotypes at each genomic position (*k-1*, *k*, *k+1*, …). The more likely path taken through the reference haplotypes is computed by the transition and emission probability. The transition probability is a function of *rk* (recombination rate between adjacent sites *k* and *k-1*) and decides whether to continue or switch the reference haplotype that the immediately previous genomic site was copied from. Then, each site is copied either perfectly or with error from the chosen haplotype with a copying probability, that is a function of the mutation rate (*p*). A perfect copy denotes no mutation; imperfect copy reflecting mutations.

It is based on a hidden Markov model, where the mosaic of reference haplotypes are the hidden states. To find the most likely path taken through a set of hidden states, it works along the genome, calculating the probabilities of changes in the attributed haplotype. The transition probability is a function of the recombination rate (*r*, see Fig. B2) between adjacent sites, whilst the emission probability is a function of the rate of mutation (*p*) (Fig B2). Moving along the genome, the LS model compares the probability of every possible copying path and infers the most likely one.

#### Box 3: Haplotype-based methods

In practice, three main paradigms are commonly used to operationally define haplotypes. They focus on the entire population (windows), individual haplotypes (segments), or the underlying generative process (trees) and based on windows, segments and trees. Within each of these groups complexity of the data treatment can vary from simple heuristics to full statistical treatments. We consider tree (ARG) based methods in the main text, but review other methods, their application and connection to the proposed haplotype block definition in this box. Defining the haplotype block by selecting arbitrary window has obvious limitation as it ignores haplotypes of variable length and does not allow overlaps. “Segment” type of methods attempts to reconstruct more realistic haplotype blocks, often referred to in the literature as IBD-blocks. Despite the limitations, those methods proved to be practical for several applications.

Across three classes of methods, window-based methods tend to be the simplest, and primarily operate *across* individuals. In the simplest form, haplotypes are operationally defined as the set of alleles observed at the segregating sites within a window of arbitrary length, say, 20 to 100 SNPs or kilobase. They typically rely on pre-defined windows of arbitrary length, say, 20 to 100 SNPs, or ten kilobase. Ideally, window sizes should be short enough to avoid spanning recombination breakpoints. One example is H12, which detects selective sweeps (Garud et al, 2015). In this test, for any given window, haplotypes are rank-ordered by their frequencies; in the case of a selective sweep at a given locus, we expect the two most common haplotypes (H1 and H2) to dominate the population. The H12 test features enhanced power to detect selection, especially under competing sweeps between recurring mutations. However, the test does not attempt to capture the real haplotype block length and is rather heuristic. Other fixed window-based applications include ones exploiting local genomic structures, especially ones showing geographical structure or associated with local adaptation (data-driven clustering/DDC in Jones et al., 2012, see also (Todesco et al., 2020, doi:10.1038/s41586-020-2467-6; Li and Peter, Genetics, doi: 10.1534/genetics.118.301747}). While window-based methods do not explicitly infer or use information of haplotype block length, they sometimes do take into account genealogy, e.g., *Twisst* (Martin et al, 2017, Lohse et al 2016). Often, the simplicity of window-based methods are also their main appeal to in the era of SNP genotyping.

The second family of methods operate primarily on individual sequences, with the aim to represent haplotypes as a mosaic of segments from a haplotype panel, often under some version of Li and Stephens’ algorithm (Box 2). These segments offer a more realistic model of recombination breakpoints and confer superior power to capture signatures due to linkage. Extended haplotypes homozygosity (EHH) (Sabeti, 2002) is an excellent example of such segment-based statistics for inferring selection. Along with its derivatives, such as integrated haplotype score (iHS) (Voight et al., 2007; Szpiech and Hernandez, 2014) and cross-population EHH (XP-EHH) (Sabeti et al. 2007), they have been widely used to detect selection in many systems (Cao 2011, International HapMap Consortium, [Meier et al. 2021](https://paperpile.com/c/8P42ec/qEPZ)). These methods typically seek to capture the decay of a signal, say, in the extent of haplotype sharing, from an *a priori* defined core SNP. More sophisticated methods based on hidden Markov models to infer the haplotype structure and especially helpful in uncovering admixture and introgression: treemix, (Lawson et al. 2012). This allows for visualizing the haplotype-specific ancestry and improved fine-scale analysis of population structure not obvious from unlinked markers. This approach is another example of IBD definition of the halpotype block, it’s based on set of ancestors.

### 