On the origin and structure of haplotype blocks

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

**To do:**

* Abstract
* Conclusion
* Add and cite references
* Read full text

Dasha:

* Monitor changes in the document
* Add Nick’s comments
* Read full text and send
* Sweep text
* Practical
* Conclusion
* Box 2
* Box 3
* Abstract

Sean:

* Sweep figures (when simulations are ready)
* Preferences

Arka:

* ARG box editing

Nick:

* Selective sweep update
* Mathematica notebook

Frank:

* Practical paragraph
* Box 3

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

[Abstract](#_gjuq5qamqwzm) **2**

[Introduction](#_5evphed5xuj9) **2**

[Defining haplotype blocks](#_is9lync626xi) **3**

[Implications of the definition](#_8xq0an65w9o7) **4**

[The definition in practice](#_p49jf7xe1b8e) **6**

[**Conclusions**](#_44sczekzt60) **9**

[Figures and captions:](#_px9lxhv4721o) **9**

[Figure 1. Inference of haplotype blocks from patterns of SNP variation in empirical data.](#_b3igbhcsffl5) 9

[Figure 2. IBD definition](#_hnrpck7q9mmr) 10

[Figure 3. Neutral simulation](#_pew6s6t95r89) 11

[Figure 4. Sweep simulation](#_328piexjga0t) 12

[Boxes:](#_nszd8661j7xj) **12**

[Box 1: Ancestral Recombination graph](#_qzujzsfc4z3d) 12

[Box 2: Li and Stephen’s Algorithm](#_z5d9ztzcqkfc) 13

[Box 3: Haplotype-based methods](#_33bdudw7bk0t) 14

### 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 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 information 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 in genetics, 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, haplotypes are often defined in different ways due to differences in terminology and focus. More practically, this disparity complicates comparison of results and interpretation, and may preclude insights that may otherwise arise through juxtaposing analyses and results across these 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 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 segregation SNP variation. We then discuss how the proposed definition relates to the various ways that haplotype blocks (or segments) are 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, 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 this 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, this is rarely the case. Instead, linkage—specifically linkage disequilibrium (LD)—across polymorphic sites often produces strong haplotype structure, such that we typically observe only a few common haplotypes in a given population. It therefore follows for “statistical phasing”, the process through which one reconciles diploid genotypes with 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 across SNPs. This is especially striking when different haplotypes become fixed across populations (see example in Fig. 1).

However, we aim for a more fundamental definition that is based on the true ancestry of the sequences, and that is independent of the mutations that generated the SNPs that we observe. 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, therefore will appear as a mosaic of blocks of different colors, corresponding to their ancestors. It is a natural way to define blocks that descend from these 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, which is done in practice in many studies (some recent examples: 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 (Fig. 2). As Figure 2 shows, these arbitrary choices can complicate analysis, but it is not always 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). 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, historical individual leaves two or more offspring that are each ancestral to our sample, and recombination occurs when, in meiosis, an ancestral genome inherits from two haploid parent genomes. What we can (in principle) observe is a sequence of genealogies along the genome, each of which shows how the genome-wide coalescence is refracted/inflected through local recombination at one genetic locus. Together, these sequential genealogies make up 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 only summarises how the 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 a particular spatial location. Tracing back in time, lineages move between backgrounds by recombination, and between locations by migration.

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 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 brought together by the focal branch that just precedes the coalescence event. With enough SNPs, the haplotype block is revealed directly by these shared SNPs (Fig. 4).

### Implications of the definition

We will now illustrate the relationships between genealogies, SNPs and haplotype blocks through simulation examples. 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 may not be intuitive, but illustrates how recombination interacts with the coalescent (also see Fig.B1 for schematic representation of the process). This can be further simplified into 15 distinct topologies, if we disregard branch lengths; these are 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 colored 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 these 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 resolution limit on what can be inferred from sequence data; branches without mutations will be invisible to us, and our ability to resolve 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 (see selective sweep simulation). Conversely, two different coalescence events may happen to bring together the same sets of lineages; these 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. Therefore, on Figure 3, the length of the blocks are bounded by recombination. Tracing back in time, branches must end in coalescence events that combine them with yet more (out-group) 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 (see SI and additional references).

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 carried 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. However, in any one realisation, its effect is highly random, making it relatively easy to identify the casual *branch*, but impossible in principle to precisely locate the causal mutation/SNP. *More to add here.*

### The definition in practice

Having defined the haplotype blocks conceptually, we next consider the problem of inferring haplotype blocks from empirical datasets. In general, it is now straight-forward to call SNPs or indel variants, but it is non-trivial to connect these to the haplotypes in which they are embedded. For that reason, 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, one needs to take one step further in order to be able to infer and operate haplotype blocks in a way we define them. Since we define the haplotype block through its manifestation in the ARG, in this section, we will review a selection of methods based on various approximations of ARG. We will further discuss their underlying assumptions, but also where they fall short in connection to the proposed haplotype block definition. Additionally, there is a family of methods, which uses a simpler definition of haplotype block and primarily operate on either fixed genomic windows or genomic segments. We outline methods belonging to the first two categories and their limitations in Box 3. Here we focus on genealogy-based methods, especially those based on the ARG, as these come closest to the way we have defined haplotype blocks here.

The full ARG contains information about branches of the genealogy and in theory encodes all the information needed for applying haplotype block definition to empirical datasets, therefore inferring ARG from the sample of sequences should be sufficient to practically define haplotype blocks. Several methods were proposed for direct inference of ARG (ref), of which ARGweaver (Rasmussen et al, 2014) and its extension ARGweaver-D are among the most powerful, and the most widely used. The main challenge of inferring the ARG, however, is that the state space of every possible ancestral history of a sample of genomes is effectively infinite, thus intractable in its full form. ARGweaver solves this issue by discretizing time to limit recombination and coalescence events within discrete time points, therefore making the ARG space finite. Further ARGweaver uses a hidden Markov model (more specifically Li-Stephens algorithm) to sample ARGs from a posterior distribution under the assumptions of sequential Markov coalescent. While making the inference more tractable, application of the hidden Markov model imposes significant limitations on defining the haplotype block in the proposed form by making inference of overlapping blocks not possible. As this approximation is used in all the methods described below, we discuss principle and imposed limitations in the box 2. In addition to limitation in terms of our definition, inference of “full” ARG is computationally expensive and tends to be sensitive towards user-defined parameters such as the mutation and recombination rate, and number of sampling iterations.

Recently, additional methods have been designed specifically to efficiently approximate ARG from much larger datasets by inferring succinct gene trees rather than full ARG. Of particular relevance here, the tree sequence contains information on tree topology and branches that is relevant to describing haplotype blocks as we defined them. Particularly notable among these are tsinfer+tsdate (Kelleher et al, 2019, Wohns 2020) and RELATE (Speidel et al, 2019). Rather than redundantly representing variation as a straight-forward SNP-by-individual matrix, tree encoding proposed by these algorithms maps variants onto nodes in a tree structure; and the genome as a sequence of such trees, hence ‘tree sequence’ (doi:[10.1534/genetics.120.303253](https://doi.org/10.1534/genetics.120.303253)). This encoding is highly efficient: it is effectively a population-level index that allows direct queries in extremely rapid ways. In data complexity terms, tree sequence encoding neatly remaps the data onto linear space in a way optimised for scalability. Notably, tree sequence encoding can be a form of extremely efficient storage for very large populations with relatively few singleton haplotypes. In fact, the underlying model in tsinfer and Relate uses ancestral haplotype reconstruction as the basis to assign individual sample haplotype segments. Methodology of tsinfer and Relate is slightly different, however like in ARGweaver they use a Li–Stephens (see Box 2) algorithm per haplotype and simply encode, rather than directly infer haplotype blocks as we have defined them. Therefore, the overlapping block structure (Fig.3 C,D) is not captured by these methods in comparison to our simulation (Fig. 3). The efficiency and accuracy of all the methods discussed above are compared in the recent preprint (Brandt et al 2021).

In contrast to the methods discussed above, HaploBlocker is an inference method that internally allows 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 (all some analogs: ) is that genealogical information is not considered.

<I think it would be important to have a paragraph here that summarizes our finding of this section regarding whether any of thr above methods defines haplotype blocks the way we say that they should. While each has features that do reflect different features of our definition, none of them really does the job. Maybe we could be explicit about this, i.e., argweaver does x but…tsinfer and relatte infer do x but don't allow for ... Ultimately this calls for development of a new method>

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, and ideally 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.

Recent methods which partially rely on our haplotype block definition already proved to be useful and efficient in various applications. However, it should be noted here that both methods are very recent and we expect to see novel applications and tests to emerge that can take advantage of the information from ARG and tree sequences. For example Relate allows to test for selection, estimate effective population size, and ancient introgression simultaneously. While applied to the haplotype data from the 1000 genomes project, Relate demonstrated an enrichment of evidence for selection at loci associated with such phenotypes as hair color. One of the interesting novelties is the possibility to estimate temporal changes in mutation rates. 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. In their paper, the authors have also presented tests for deep branches and other examples to demonstrate the added power to detection selection through features encoded in tree sequences (Speidel et al. 2019). However, it should be noted here that both methods are very recent and we expect to see novel applications and tests to emerge that can take advantage of the information from ARG and tree sequences.

### Acknowledgements

FWF Wittgenstein, SCAS Sweden, Barton Group

### Figures and captions:

#### 

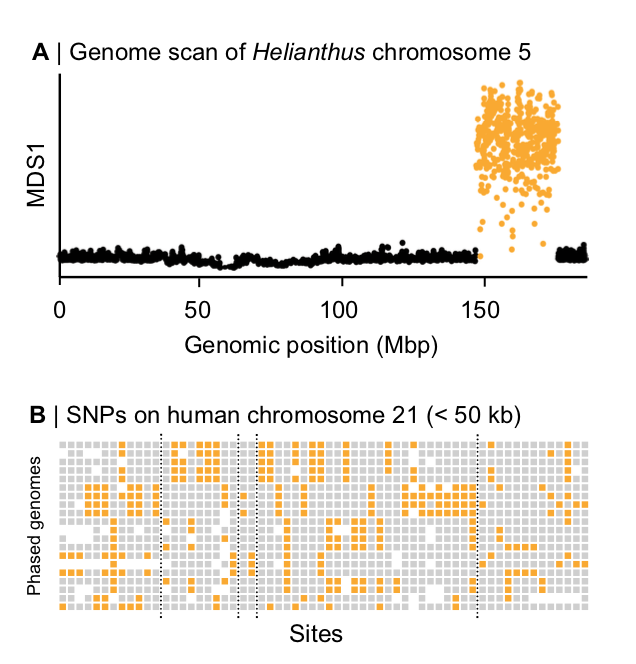


Figure 1. patterns in empirical data. (A) A local PCA analysis conducted on SNP data from wild sunflowers reveals a large ‘haploblock’ on chromosome 5 (Todesco et al., 2020). Each circle is the MDS1 score for each window, illustrating variation in the PC results for each window (See Li and Ralph (2019) 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

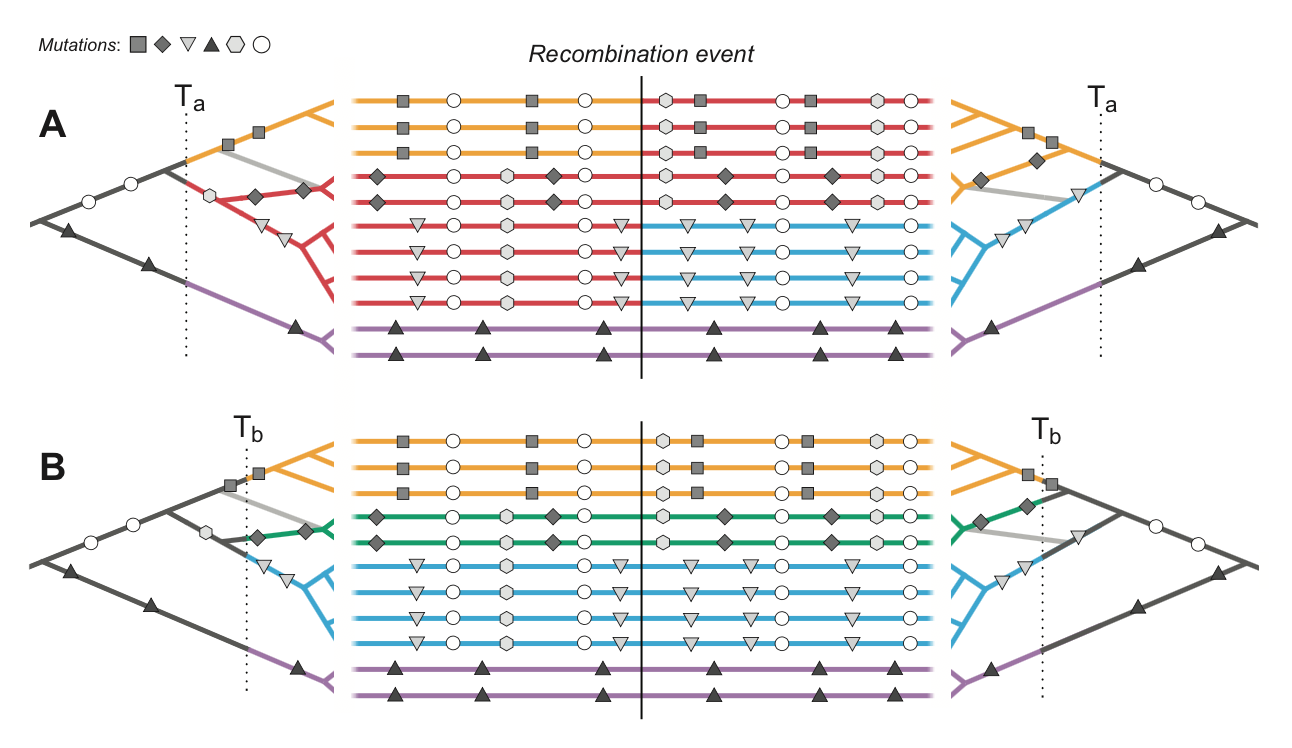


Figure 2. Haplotype blocks defined through identity by descent (IBD). Panels A and B show the same 11 hypothetical DNA sequences depicted as horizontal lines. The trees on the left and right sides show the genealogy for the set of sequences either side of a recombination event (indicated by the vertical 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.

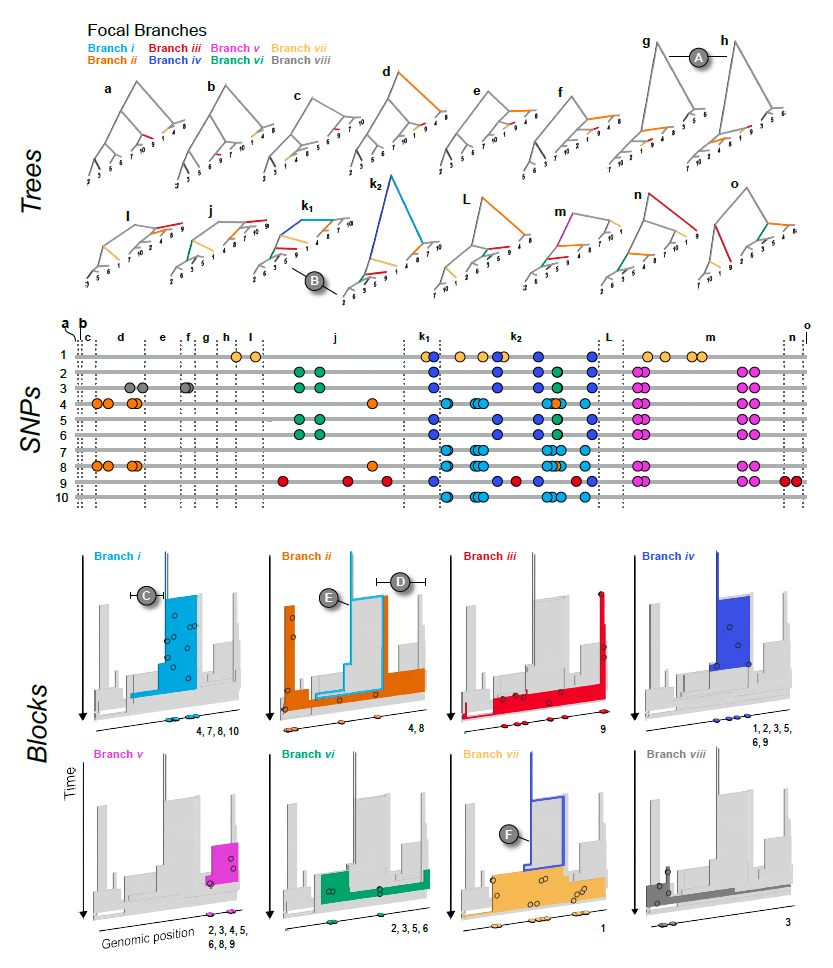
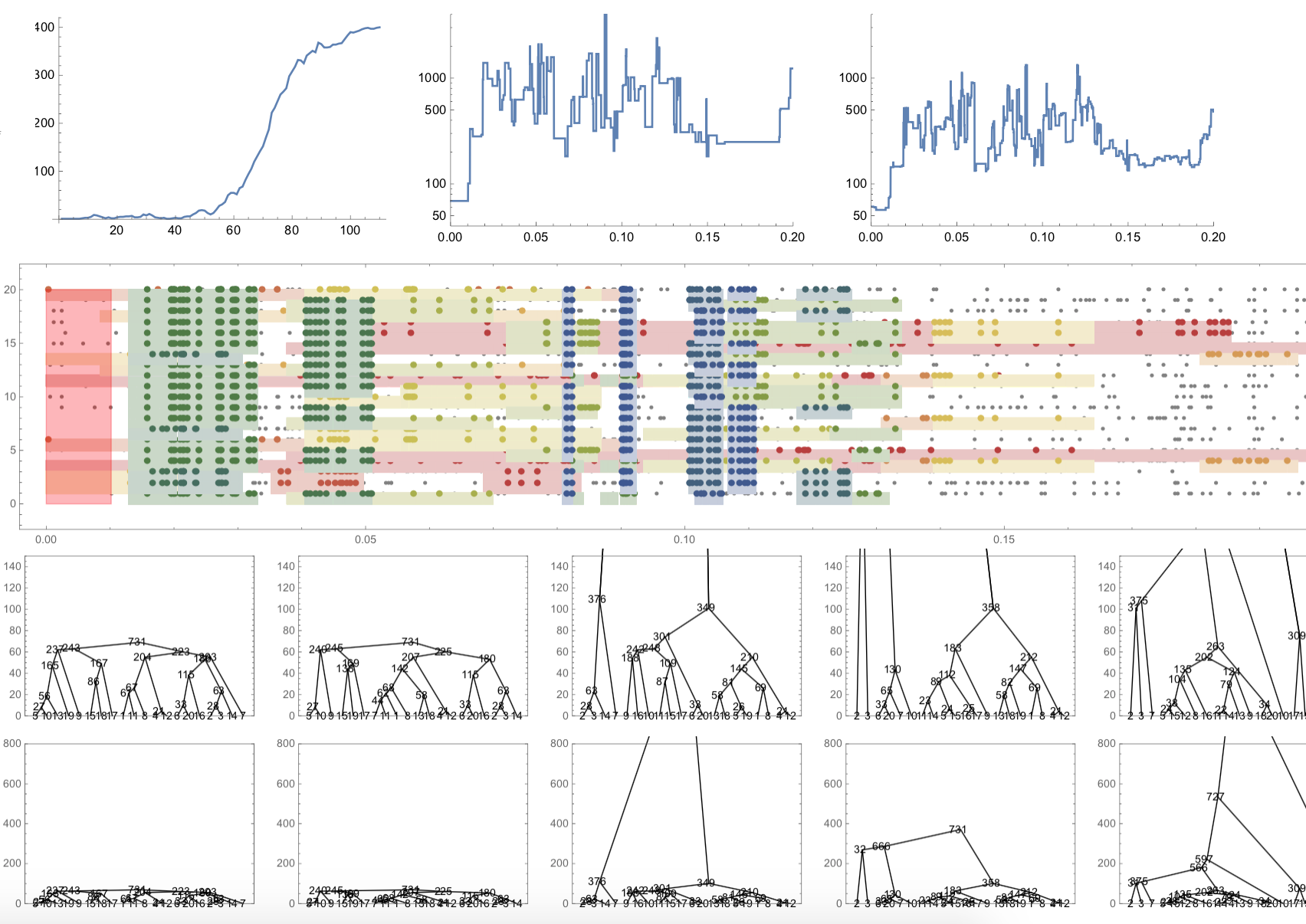


Figure 3. The relationship between trees (top), SNPs (middle), and haplotype blocks (bottom) in the neutral simulation (see main text for simulation details). The trees (a - o) show all of the unique topologies that coincide with the genomic spans shown in the central panel (also labeled a - o). The 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 genomic sequences through a series of real coalescence and recombination events (Hudson 1983, Griffits, Marjoram 1997). At any given site on the genome, the genealogical relationship between samples can be described through a genealogy (Fig. B1B), (Kingman, 1982); all chromosomes in the sample coalesce and eventually trace back to one single ancestor. Moving along the genome, the genealogy will inevitably change due to recombination. This leads to a series of observable genealogies along the genome (Fig B1, bottom panel), which are embedded in a single structure - the ARG (Fig B1, top panel). The ARG paints the complete ancestral picture of a sample - describing each recombination breakpoint and a genealogy for each non-recombined genomic block. In theory this includes trees which differ insignificantly, also termed marginal trees. Crucially, the ARG describes ancestry not allelic state, so is independent of all neutral mutations that lead to the observed polymorphism in the present sample.

<NB: this paragraph needs clarification>Another way to visualize the complex structure of ARG is by tracing back chromosomes of an organism (Fig. B1, top panel), following the coalescent process and trace the chromosomes back into the past. One generation back, it descends from a mosaic of segments, derived from the two homologous chromosomes of one or other parent via meiosis. In the next generation back, the chromosome is further broken into segments that descend from haploid chromosomes of two grandparents. Thus, by t generations into the past, meiotic recombination splits the ancestry of the present chromosome into up to 2t ancestral chromosomes. However, as the number of ancestral chromosomes increases, they coalesce into a smaller number of ancestors, leading to a stable equilibrium between the opposing effects of recombination and coalescence.

The ARG contains full information about the genealogy of the sample, which is in theory sufficient to infer any evolutionary process, such as selection, recombination and mutation rate estimation, inference of the effective population size and population structure. This information goes beyond 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.

#### 

Figure B1. Caption

(A) Ancestral Recombination graph representation - Tracing back ancestry of four chromosomes, there is either recombination when lineages split or coalescence when lineages merge. Genomic regions that are ancestral to the present chromosomes are coloured accordingly. Recombination is represented by splitting a chromosome into two; where grey denotes non-ancestral genomic background. Coalescence is represented by two chromosomes merging into one, with regions ancestral to both chromosomes denoted by mixed colours. There are 3 recombination and 6 coalescence events in the full ancestral history of the four chromosomes. (B) Genealogical trees along the genome, corresponding to the above ARG - each tree describes the ancestral relationship for each of the 4 non-recombined regions. Asterisk denotes a unique coalescence event that is ancestral to disjunct genomic regions. *t1,t2,...,t6* denotes time points for each coalescence events.

#### 

#### Box 2: Application and limits of Li and Stephen’s Algorithm haplotype inference

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.

At its core, the LS algorithm describes a framework to decide in a sample of haplotypes whether the focal haplotype represents *a)* an entirely new haplotype or *b)* a mosaic of previously encountered haplotypes, and if so, the breakpoints and transitions in this mosaic. This latter mosaic is achieved by finding the most likely path taken through a set of previously observed or reference haplotypes. Briefly, the model computes the conditional probability of choosing one of the reference haplotypes from which an allele can be copied. For the next allele, it can either continue copying from the same haplotype, or switch to a different one. The probability of switching is a function of the recombination rate *(r,* see Fig. B2*)* between adjacent sites. This scheme is further extended, such that alleles can be mis-copied from the reference with some probability, proportional to the reference sample size *(n)* and the rate of mutation *(p)* (Fig B2) Altogether, the LS model compares the probability of every possible copying path, and infers the most likely one. Compared to standard coalescence, the LS HMM framework is highly tractable and efficient. In terms of HMM, the referencehaplotypes are the hidden states. The transition probabilities specify the probability of either continuing or switching haplotypes to copy from - capturing recombination events. The emission probabilities specify the probability to copy with or without errors - capturing mutation events. Although the LS model captures two key features of genetic variation, such as genetic relatedness among chromosomes through recombination and newer haplotypes being increasingly similar to previously existing haplotypes, this model is non-genealogical and assumes the reference haplotypes to be actual ancestors. Moreover, the LS model assumes that genomic states depend solely on the immediately preceding site (i.e., a sequential Markov coalescent, SMC), while processes generating ARG makes no such assumption, since recombinant lineages can coalesce back to *any* lineage that existed in the preceding genome.

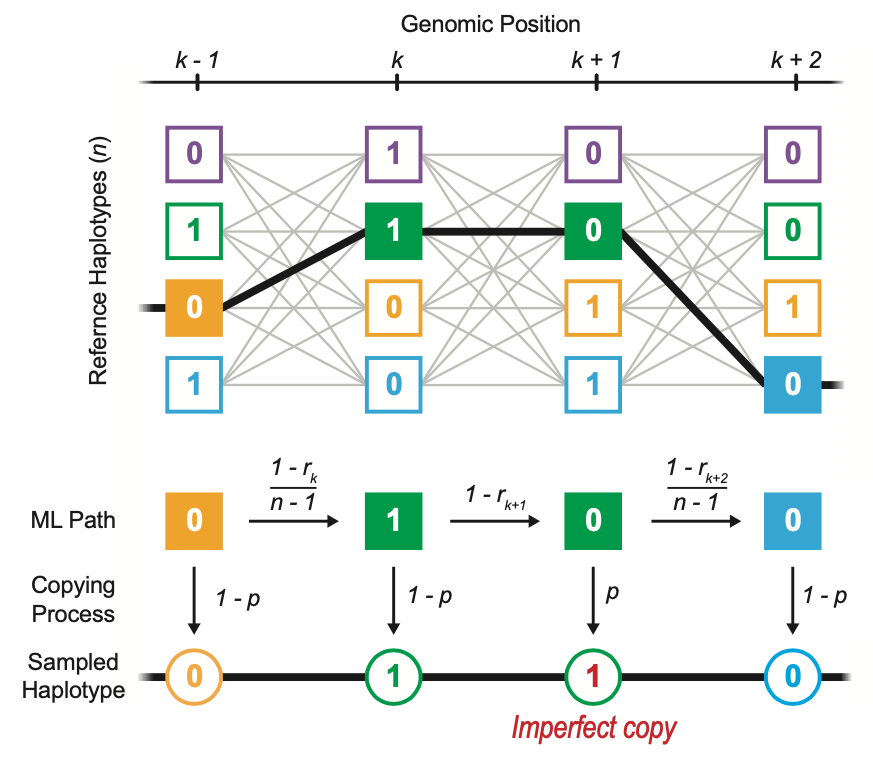


Figure B2. Caption

Schematic representation of Li and Stephen’s hidden markov model that samples a new haplotype as an imperfect copy of *n* previously existing reference haplotypes at each genomic position (*k-1*, *k*, *k+1*, …). The maximum likelihood (ML) path taken through the reference haplotypes is computed by the transition and copying 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. Switching haplotype capture recombination events. 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, whereas an imperfect copy captures mutations.

#### Box 3: Haplotype-based methods

In practice, three main paradigms based on windows, segments and trees are commonly used to operationally define haplotypes/ They focus on the entire population (windows), individual haplotypes (segments), or the underlying generative process (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 and their application in this box. <How does IBD is implemented?>

Across three classes of methods, window-based methods tend to be the simplest, and primarily operate *across* individuals. 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. In the simplest form, haplotypes are operationally defined as the set of alleles observed at the segregating sites within the window (refs). Next, summary statistics can be calculated for each window along the genome. 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 in the case of recurring mutations. However, the test does not attempt to capture the real haplotype block length, thus restricting its application to detecting selection. 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, Barton). Although, straight-forward, window-based methods are practical 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. These segments offer a more realistic model of recombination breakpoints and confer superior power to capture signatures due to linkage. Extended haplotypes homozygosity (EHH) is an excellent example of such segment-based statistics for inferring selection. Along with its derivatives, such as integrated haplotype score (iHS) and cross-population EHH (XP-EHH), they have been widely used to detect selection in many systems {REF; REF}. 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 can be also included into this group. One of many examples of applying this approach comes from analysis of population structure and admixture: fineSTRUCTURE. This allows for “painting” the chromosomes - assigning regions of the genome to a set of predefined ancestors and improved PCA-based analysis of population structure.

### 