On the origin and structure of haplotype blocks

Daria Shipilina*[1,2,3$], Sean Stankowski*[2], Arka Pal[2], Yingguang Frank Chan4, Nicholas H. Barton[2]

[1]Evolutionary Biology Program, Department of Ecology and Genetics (IEG), 75236 Uppsala University, Sweden; [2]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria, 3Swedish Collegium for Advanced Study, 75236, Uppsala, Sweden,

Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany

Deleted: [4]

*Both authors contributed equally to this work

[$]Corresponding author

ORCID IDs:

DS: 0000-0002-1145-9226

SS: 0000-0003-0472-9299

AP: 0000-0002-4530-8469

YFC: 0000-0001-6292-9681

NHB: 0000-0002-8548-5240

Abstract

The term "haplotype block" is commonly used in the developing field of haplotype-based inference methods. We argue that the term should be defined based on the structure of the Ancestral Recombination Graph (ARG), which contains complete information on the ancestry of a sample. We use simulated examples to demonstrate key features of the relation between haplotype blocks and ancestral structure, emphasising the stochasticity of the processes that generate them. Even the simplest cases of neutrality or of a "hard" selective sweep produce a rich structure, which is missed by commonly used statistics. We highlight a number of novel methods for inferring haplotype structure as full ARG, or as a sequence of trees. While some of these new methods are computationally efficient, they still lack features to aid exploration of the haplotype blocks, as we define them, thus calling for the development of new methods. Understanding and applying the concept of the haplotype block will be essential to fully exploit long and linked-read sequencing technologies.

Keywords

haplotype block, ancestral recombination graph, haplotype-based methods, coalescent

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). 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 organisms across the tree of life. In light of this 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 across many disciplines (Delaneau et al., 2019; International HapMap Consortium, 2005; Leitwein et al., 2020). An important, but often overlooked fact, is that populations 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; Clark, 2004). Thus, disentangling the evolutionary history underlying genomic patterns can be challenging using solely site-based statistics. For example, while whole-genome scans for signatures of selection can reveal loci that affect fitness (Poelstra et al., 2014; Tavares et al., 2018), it is hard to determine the actual causes of these signals (Burri, 2017; Grossman et al., 2010; Ravinet et al., 2017; Rockman, 2012; Stankowski et al., 2019; Tavares et al., 2018; Wolf & Ellegren, 2017). For example, shifts in polygenic scores from genome-wide association studies (GWAS) can be misinterpreted as signals of selection instead of being artifacts of population structure (Berg et al., 2019; Novembre & Barton, 2018; Sella & Barton, 2019). Similarly, methods for estimating population density and gene flow struggle to distinguish among a virtually infinite number of possible population structures (Richardson et al., 2016; Sousa et al., 2011; Whitlock & Mccauley, 1999).

Commented [1]: Nick, Dasha (this is a pretty strange comment!), how to address?

- I found the paragraph starting at L43 a bit confusing. The authors list several examples of challenging analyses where haplotype structure is important (selection scans, polygenic selection, polygenic scores, gene flow), but the implied connection with haplotype blocks is a bit vague.

Deleted: to be

By accounting for haplotype structure, it should be possible to make inferences more accurate and more efficient. Haplotypes carry information not only from *mutation* but also from *recombination*, which provides 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), human and medical genetics (Crawford & Nickerson, 2005), and animal and plant breeding (Bhat et al., 2021; Mészáros et al., 2021). As a result, there is often little consensus on how haplotype blocks are defined. More practically, this disparity complicates comparison of results, and may preclude insights that may otherwise arise by combining 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 (ARG). Using simulations of simple but general scenarios, we 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 practical inference methods and their applications in large-scale population studies. We consider how different methods make use of haplotype information and infer haplotype blocks, their underlying assumptions and respective limitations.

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 (Al Bkhetan et al., 2019; Clark, 2004; International HapMap Consortium, 2005; Schwartz et al., 2003; Taliun et al., 2014; Zhang et al., 2002). 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 2^n possible pairs of haplotypes - more than a million with just n = 20. However, there are usually just a few haplotypes, due to 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 (S. R. Browning & Browning, 2011). Looking across individuals in larger genotype panels, the more frequent haplotypes often appear as stretches of shared, "banded" blocks of SNPs (Fig. 1A) (Patil et al., 2001). This can be especially striking when different haplotypes become fixed across populations, which can produce block-like patterns in data even when individual haplotypes cannot be observed (Fig. 1B); in some cases, the outstanding regions have been referred to as 'haploblocks' (Todesco et al., 2020).

Whilst a blocklike structure may be apparent within empirical genetic 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 have also been defined in a more concrete way via the classical concept of identity by descent (Carmi et al., 2013; Hartl et al., 1997; Thompson, 2013). Imagine an initial population, where each founder genome is labelled 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 colours, each corresponding to their ancestors. This naturally defines blocks that descend from a given set of founders (Fig. 2). Fisher (Fisher, 1954) 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 (Lundberg et al., 2017; Otte & Schlötterer, 2021; Wallberg et al., 2017). Moreover, if we disregard new mutations, the evolutionary processes subsequent to the founding of the population are entirely described by the block structure.

Deleted: can be

Identity-by-descent is usually defined with respect to a specific ancestral reference population (but note that coalescent definitions of IBD also exist (Wakeley & Wilton, 2016)). However, when we deal with natural populations, there is no obvious reference population, so the block structure will vary depending on our arbitrary choice of founders at an arbitrary time point (Figure 2). To eliminate this subjectivity, we will base our definition of 'haplotype block' on the full ancestry of the sampled genomes, namely, on the ancestral recombination graph (ARG) (R. R. 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 events that join lineages and of recombination events that split lineages (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 (Hudson 1983, Griffiths, Marjoram 1997). 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, i.e. a local gene tree. 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 A in Fig. 3 and Fig. B1A). Moreover, as samples get larger, blocks defined 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 simulations (neutral scenario and a selective sweep; supplement 1). The simulation generates the ancestral recombination graph, by tracking the ancestors of a sample of genomes back through time, until all ancestral genomes are ancestors to the whole sample. The simulation can be conditional on a selective sweep, which is defined by the numbers of copies of the favourable allele in the population. A Wright-Fisher model with a constant population size 2N haploid genomes is assumed. A region of genome of map length R is

Commented [2]: Maybe mention it later?

Commented [3R2]: I think its ok like this, but I pointed out above that there are 2 diff scenarios followed, with the selected locus at the leftmost point (i.e., at 0). For simplicity, we allow at most one crossover, with probability R; we simulate R <<1, so this is close to the case with no interference between crossovers. Once the ARG is constructed, genealogies along the genome can be constructed, and their branches identified. Neutral SNPs can be added, assuming infinite-sites mutation; each SNP is associated with a branch in the ARG.

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 (Supplement 1). SNPs were generated by infinite-sites mutation with mutation at twice the rate of recombination. This simulation is general, because time and map distance both scale with population size (Richard R. Hudson, 1990). 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, where N is the effective size. Thus rescaled, this simulation shows a generic pattern, independent of population size.

The central panel of Fig. 3 (middle panel, ‘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). Recombination events have divided the genome 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; top panel, ‘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 labelled a - o). For illustration, we show one pair of genealogies that have the same topology, but differ in depth (k_1 and k_2), B in Fig. 3).

Deleted: The genome

Deleted: is divided

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 arise. 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 genomic lengths 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 in

Fig. 3). 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. 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'). Haplotype blocks will also overlap in the genome, but not in time, when branches are nested in the genealogy, giving rise to nested haplotype blocks. For example, branch *ii* (orange), which is ancestral to genomes 4 and 8 descends in the middle part of the genome from branch *i* (blue), which is ancestral to genomes 4, 7, 8 and 10. Thus, haplotype block *i* is nested above block *ii* in Fig. 3 (see also F for another example of nested haplotype blocks).

Deleted: (

Deleted:)

Deleted: branch

Deleted: branch

Deleted: ing

If we start at a particular site of the chromosome, 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 incidence 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, 'Blocks'), where branches consist of segments

Deleted: point on the map

Deleted: rate of

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 fact, the distributions of areas of blocks is highly skewed, and so most SNPs are on a few deep branches (Supplement 1).

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 SNPs 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 (Wakeley, 2009). Selective sweeps have a somewhat similar effect. In the classic case (Maynard Smith & 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 (Supplement 2). The sweep greatly reduces diversity around the selected locus, because all lineages must trace back to the successful mutation (Fig. 4B, 1); this region of complete coalescence is shown in red; note that it still contains some diversity, due to mutation subsequent to the sweep. As we move away from the selected locus, lineages recombine out onto the ancestral background, and coalesce with the rest of the genealogy much further back (Fig. 4B). This process can be seen in the time to the MRCA (Fig. 4D), which jumps from a low value at the selected locus, through successive recombination events, back to a time that fluctuates around $4N_e=800$ generations, under the standard coalescent. However, the replicates in the lower panel show that there is considerable variation in this process, which sets a fundamental limit on our power to detect a sweep and estimate its properties.

At the selected locus, all lineages coalesce in the favoured mutation. Successive recombinations each free one or a few lineages from the new background, so that the exceptionally large and recent cluster gradually diminishes in size, until the genealogies follow a close to neutral

Commented [4]: NICK:
Line 260-261: "In simulation, deep branches tend to be wider than expected…" – it is not clear exactly what the expectation was, or how much the simulated data differed from the expectation.

Also:
- L260: the authors observe that deep branches are longer (Supp 1). I believe this is also likely due to re-coalescence events, which are less likely on short branches at the bottom (see difference between SMC and SMC' models by McVean & Cardin and Marjoram & Wall).

Commented [5R4]: I looked at this more closely and expanded the SI ("Inverse relation..."). It is not an easy question - please look at this section in the SI and see whether what I say here is reasonable.

distribution. Thus, branches with large numbers of descendants are associated with the sweep, and can be distinguished by the characteristic sets of SNPs that they carry; nine such branches are illustrated in Fig. 4C.

We close this section by commenting on possible connections between our description of the ARG, and practical inference. Stern et al. (2019) propose a method that infers the allele frequency trajectory from the genealogy at the selected locus, which is assumed to be known. The extent of the focal branch along the genetic map gives additional information, with a predicted constant rate of recombination out into the ancestral background, at a rate equal to the frequency of the ancestral allele. Additional branches give more information: in particular, several lineages may coalescence early in the sweep, but then recombine out (e.g. the second genealogy in Fig. 4B). This generates multiple long branches, whose distribution depends on $4N_es$ (Barton, 1998). There is considerable scope for using the extent of branches along the genome, as well as the genealogy at specific loci.

Nevertheless, we make two cautionary comments. First, there is considerable variability between different realisations, given the same trajectory (e.g. Fig. 4D), and there is yet more variability in the trajectory generated by given selection. Moreover, if the locus is identified from a genome-wide scan, ascertainment bias will distort the ARG: indeed, sequence variation around a neutral locus that experiences a sweep *by chance* may be indistinguishable from a genuinely selected locus. This poses fundamental limits to our ability to estimate selection at a particular locus. Second, sophisticated methods based on simple scenarios will be confounded by deviations from the model. For example, the extent of reduced diversity along the genome is the inverse of the time taken to reach high frequency - but that may be greatly increased by population structure. The visualisations that we develop here may have greatest value in allowing us to check whether the fine structure of a candidate region is actually consistent with some simple model. It remains to be seen how far the rich information contained in the structure of such branches will help us improve our inferences.

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 make 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 (B. L. Browning & Browning, 2009, 2013; Davies et al., 2016; B. Howie et al., 2011; Marchini et al., 2007). 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. Given that our definition is rooted in the features of the ARG, 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 and highlight where they could be extended in light of our proposed haplotype block definition. Separately, in Box 2, we also outline classes of simpler methods that use fixed genomic windows or genomic segments as a proxy of the haplotype block.

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, ARGweaver (Rasmussen et al., 2014) and its extension ARGweaver-D (Hubisz et al., 2020) 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; extended by Marjoram and Wall 2006(McVean & Cardin, 2005) to sample from an ARG distribution. While making inference more tractable, SMC precludes the inference of disjunct blocks, because only one immediately prior state is considered as one moves along the genome. Besides these limitations, inference of the "full" ARG is still computationally expensive. As such, ARGweaver is most suitable for mid-sized datasets on the order of fifty sequences.

Two recent methods*,* tsinfer and 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 of the ARG. 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 and Stephens' hidden Markov model (N. Li & Stephens, 2003) see Box 2 for further details) to infer local pairwise distances (Relate) or ancestral haplotypes (tsinfer). As an added advantage, which doubles as an efficient, lossless compression algorithm by indexing population genomic variation as SNPs-on-trees as opposed to the traditional (and highly redundant) SNP-by-individual matrix (implemented as a tskit library (Kelleher et al., 2019). 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 encapsulates a number of population genetics summary statistics (Kelleher et al., 2019; Ralph et al., 2020). These developments may prove essential, as sequencing of entire national populations increasingly becomes routine.

Deleted: hidden Markov model (

Among practical methods, tsinfer and Relate represent the state-of-the-art in representing large populations. All three approaches, including ARGweaver, approximate some aspects of the ARG well, and give accurate coalescence time estimates under simulation of the standard coalescent (Brandt et al., 2021). For our purposes, they are also useful approximations of ARG that highlight some of the key advantages we wish to emphasise 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 alone in tsinfer. Having dated branches allows, 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.

We note that efforts are already underway to bridge across methods and address their limitations. 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 have noted that that they may not be accurately captured under the Li–Stephens models in tsinfer and Relate, in a way that may bias the inferred ARG. However, this is something of open question, so more work is needed to understand how different methods performance across a range of parameters relevant to non-model organisms.

Deleted:

Deleted: they

Deleted: can be

Deleted: rather poorly

In summary, there has been a recent spurt in innovation in genealogy/ARG-based methods. Among these, ARGweaver arguably comes closest to inferring the full ARG, but at considerable computational cost. 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.

Deleted: Taken together

Commented [6]: DASAH, ARKA, FRANK
Line 397: "…but infer haplotype blocks only as an incidental output." In my understanding, the output of Relate tells which branches existed over which genomic intervals and maps specific mutations to specific branches. The shortcoming of this type of output data for identifying haplotype blocks, as defined in this manuscript, is unclear.

Assuming that a method becomes available for inferring blocks as we have defined them, there are still practical considerations that we will 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 (for example) hard to define the extent of haplotype blocks in any simple way, especially since they may be disjunct. Should this be their maximum length, or should it rather be weighted by the depth? It is not clear which description would be better for inference and this may even depend upon the specific process that we wish to infer. These kinds of issues could be investigated by estimating parameters under a variety of specific models in which case we can evaluate the strength and weaknesses of different descriptions of haplotype structure in characterizing different processes.

Conclusions and future directions

In this article, 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. In our view, haplotypes and haplotype blocks should be the core concepts through which we understand population genetic processes. Under 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, analysis software, and visualization tools that are designed with sequencing and population datasets in mind (Davies et al., 2021; Meier et al., 2021).

Deleted: and

Our simulations show that haplotype blocks contain rich information about the demographic and selective history of the locus. Making the most of this information will require a fundamental rethink of our linear, reference-based genome assemblies, and a move towards a graph-based assembly standard (Eggertsson et al., 2017; Hickey et al., 2020). We will also need new concepts and vocabulary to describe features in these graphs (e.g., super-graphs and “bubbles”; (Cheng et al., 2021; Turner et al., 2018; Weisenfeld et al., 2017)) informed by a robust understanding of the generative process discussed above, and we need to align our mental models with inference schemes and their encoding (as in, e.g., tsinfer). For that reason, we hope our discussion here can focus our effort towards this new standard, as haplotype-resolved sequencing becomes routine.

Acknowledgements

We thank the Barton group for useful discussion and feedback during the writing of this article. Funding was provided by SCAS (Natural Sciences Programme, Knut and Alice Wallenberg Foundation), an FWF Wittgenstein grant (PT1001Z211), an FWF standalone grant (grant P 32166), and an ERC advanced grant. YFC was supported by the Max Planck Society.

Deleted:

Deleted: and

Commented [7]: mention funding from ERC?

Deleted: .

References

Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K., & Goudey, B. (2019). Exploring effective approaches for haplotype block phasing. *BMC Bioinformatics*, *20*(1), 540.

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K., & Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. *ELife*, *8*. https://doi.org/10.7554/eLife.39725

Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., & Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. *Communications Biology*, *4*(1), 1–12.

Brandt, D. Y. C., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2021). Evaluation of methods for the inference of ancestral recombination graphs. In *bioRxiv* (p. 2021.11.15.468686). https://doi.org/10.1101/2021.11.15.468686

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223.

Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*(2), 459–471.

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews. Genetics*, *12*(10), 703–714.

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*. https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.14

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., & Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, *43*(10), 956–963.

Carmi, S., Palamara, P. F., Vacic, V., Lencz, T., Darvasi, A., & Pe'er, I. (2013). The Variance of Identity-by-Descent Sharing in the Wright–Fisher Model. *Genetics*, *193*(3), 911–928.

Castro, J. P. L., Yancoskie, M. N., Marchini, M., Belohlavy, S., Hiramatsu, L., Kučka, M., Beluch, W. H., Naumann, R., Skuplik, I., Cobb, J., Barton, N. H., Rolian, C., & Chan, Y. F. (2019). An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *ELife*, *8*, e42014.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology*, *27*(4), 321–333.

Crawford, D. C., & Nickerson, D. A. (2005). Definition and clinical importance of haplotypes. *Annual Review of Medicine*, 56, 303–320.

Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965–969.

Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C. M., Chan, Y. F., & Myers, S. (2021). Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, *53*(7), 1104–1111.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1), 5436.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49(11), 1654–1660.

Fisher, R. A. (1954). A fuller theory of “Junctions” in inbreeding. *Heredity*, *8*(2), 187–197.

Griffiths, R. C., & Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and Its Applications*, *87*, 257.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2010). A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, *327*(5967), 883–886.

Hartl, D. L., Clark, A. G., & Clark, A. G. (1997). *Principles of population genetics* (Vol. 116). Sinauer associates Sunderland, MA.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, *21*(1), 35.

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3* , *1*(6), 457–470.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529.

Hubisz, M. J., Williams, A. L., & Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genetics*, *16*(8), e1008895.

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

Hudson, Richard R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, *7*(1), 44.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61.

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

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*(3), 235–248.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*(1), e1002453.

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P.-A., & Bernatchez, L. (2020). Using Haplotype Information for Conservation Genomics. *Trends in Ecology & Evolution*, *35*(3), 245–258.

Li, H., & Ralph, P. (2019). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics*, *211*(1), 289–304.

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

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834.

Lohse, K., Chmelik, M., Martin, S. H., & Barton, N. H. (2016). Efficient Strategies for Calculating Blockwise Likelihoods Under the Coalescent. *Genetics*, *202*(2), 775–786.

Lundberg, M., Liedvogel, M., Larson, K., Sigeman, H., Grahn, M., Wright, A., Åkesson, S., & Bensch, S. (2017). Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. Evolution Letters, *1*(3), 155–168.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913.

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics*, *206*(1), 429–438.

Maynard Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35.

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

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N. J., Bridle, J. R., Rolian, C., Barton, N. H., McMillan, W. O., Jiggins, C. D., & Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25). https://doi.org/10.1073/pnas.2015005118

Mészáros, G., Milanesi, M., Ajmone-Marsan, P., & Utsunomiya, Y. T. (2021). Editorial: Haplotype analysis applied to livestock genomics. *Frontiers in Genetics*, *12*, 660478.

Novembre, J., & Barton, N. H. (2018). Tread Lightly Interpreting Polygenic Tests of Selection. *Genetics*, *208*(4), 1351–1355.

Otte, K. A., & Schlötterer, C. (2021). Detecting selected haplotype blocks in evolve and resequence experiments. *Molecular Ecology Resources*, *21*(1), 93–109.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., … Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, *294*(5547), 1719–1723.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M. G., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., & Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, *5*(6), e1000519.

Ralph, P., Thornton, K., & Kelleher, J. (2020). Efficiently Summarizing Relationships in Large

Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics*, *215*(3), 779–797.

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

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477.

Richardson, J. L., Brady, S. P., Wang, I. J., & Spear, S. F. (2016). Navigating the pitfalls and promise of landscape genetics. *Molecular Ecology*, *25*(4), 849–863.

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution; International Journal of Organic Evolution*, *66*(1), 1–17.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918.

Schwartz, R., Halldórsson, B. V., Bafna, V., Clark, A. G., & Istrail, S. (2003). Robustness of inference of haplotype block structure. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *10*(1), 13–19.

Sella, G., & Barton, N. H. (2019). Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics*, *20*, 461–493.

Sousa, V. C., Grelaud, A., & Hey, J. (2011). On the nonidentifiability of migration time estimates in isolation with migration models. *Molecular Ecology*, *20*(19), 3956–3962.

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

Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biology*, *17*(7), e3000391.

Stankowski, S., & Streisfeld, M. A. (2015). Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proceedings. Biological Sciences / The Royal Society*, *282*(1814). https://doi.org/10.1098/rspb.2015.1666

Steinrücken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 17115–17120.

Steinrücken, M., Spence, J. P., Kamm, J. A., Wieczorek, E., & Song, Y. S. (2018). Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, *27*(19), 3873–3888.

Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, *76*(3), 449–462.

Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, *18*(4), 676–682.

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827.

Taliun, D., Gamper, J., & Pattaro, C. (2014). Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics*, *15*, 10.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., Elleouet, J., Burrus, M., Andalo, C., Li, M., Li, Q., Xue, Y., Rebocho, A. B., Barton, N. H., & Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(43), 11006–11011.

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–326.

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L., Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), 602–607.

Turner, I., Garimella, K. V., Iqbal, Z., & McVean, G. (2018). Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics* , *34*(15), 2556–2565.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), e72.

Wakeley, J. (2009). *Coalescent theory :an introduction /* (575:519.2 WAK). sidalc.net. http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=FCL.xis&method=post&formato=2&cantidad=1&expresion=mfn=010195

Wakeley, J., & Wilton, P. R. (2016). *Coalescent and models of identity by descent*.

Wallberg, A., Schöning, C., Webster, M. T., & Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLoS Genetics*, *13*(5), e1006792.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767.

Whitlock, M. C., & Mccauley, D. E. (1999). Indirect measures of gene flow and migration: FST≠1/(4Nm+1). *Heredity*, *82*(2), 117–125.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2021). A unified genealogy of modern and ancient genomes. In *bioRxiv* (p. 2021.02.16.431497). https://doi.org/10.1101/2021.02.16.431497

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews. Genetics*, *18*(2), 87–100.

Zhang, K., Calabrese, P., Nordborg, M., & Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *American Journal of Human Genetics*, *71*(6), 1386–1394.

Data Accessibility Statement

The Mathematica code used to generate the simulations presented in the paper are provided in the supplementary materials. The simulated data are available on Dryad under accession number xxxxxx.

Benefit-Sharing Statement

Benefits Generated: Benefits from this research accrue from the sharing of our simulation code as described above.

Author Contributions

All authors conceived the ideas and contributed to the writing of the manuscript. NHB conducted the simulations.

Figure 1. Block-like patterns in empirical data. (A) Block-like patterns in phased DNA sequences from *Mimulus auranticus* within the gene *MaMyb2* (Stankowski & Streisfeld, 2015). Rows show 24 individual haplotypes. Each column is a site with yellow and blue squares representing ancestral and derived sites, respectively. (B) An F_{st} Scan across *Helilconius* chromosome 2 reveals a large plateau of differentiation on chromosome 2 between races of *H. erato* (Meier et al., 2021). This large block-like pattern coincides with a chromosomal inversion, the boundaries of which are illustrated by the dashed line.

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 on either side of a recombination event (indicated by the vertical black line); the light grey branch in both trees shows how recombination changes the structure of the genealogy on either side. 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, T_a and T_b, yielding different haplotype structure.

Deleted: the effect of

Deleted: in the

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 labelled a - o). The 8 branches that we focus on in this example are coloured and labelled i – vii. (A) Two neighbouring topologies that differ only slightly due to recombination. (B) An example of two trees (k_1 and k_2) that have the same topologies but different branch lengths. The central panel shows 10 haploid genomes (labelled 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 grey. 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 centre 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 not revealed by mutations arising on the corresponding branch. (E & F) Examples of nested haplotype blocks, where the ancestral block is highlighted with a coloured outline.

Figure 4. The effects of a recent selective sweep on linked genealogies. (A) A mutation with advantage 10% arose in a population of 400 haploid individuals, and swept to fixation in 110 generations, at which time 20 genomes were sampled; 20cM of the genome is followed back in time, with the selected locus at the left.; dashed lines (T_1 - T_4) show times when the favoured allele was in 1 copy, at 10%, at 50%, and at 90% (110, 53, 38, 22 generations back). (B) shows genealogies at positions 0, 1.3cM, 5.3cM, and 20 cM, branches are coloured in red when on the fitter background, and black when on the ancestral background. Thus, changes in colour show recombination events that change the genomic background. Note that such events are unlikely when the allele is near fixation (i.e., at the base of the tree, below the lower dashed line), and conversely, become common whilst the allele is rare, simply because it will almost always meet with the opposite background. Before the mutation occurs (i.e., above the upper dashed line) lineages must either trace back to that mutation (top left) or recombine out into the ancestral background; thus, all lineages must appear black above the upper dashed line (110 generations back). Note that the disjunct branches in trees 2 - 4 all coalesce further back in time, but only 200

generations are shown for visibility. (C) shows SNPs along the 20 sampled genomes. The 9 most substantial branches are shown. (These have more than 8 descendants, formed by coalescence more recently than the sweeping mutation, and have areas >0.5). The red block at the left shows the region linked to the selected locus, which coalesces in a single common ancestor 69 generations back, just after the sweeping mutation arose. Grey dots show those SNPs that are not on these 9 highlighted branches. (D) shows the time back to the most recent common ancestry (TMRCA) along the genome, on a log scale. The bold line shows the example simulated above, whilst the three grey lines show replicates, generated conditional on the same sweep; the break in the line shows an area where the TMRCA extends further back than the extent of the y-axis. The dashed line across the plot corresponds to T_1 in panel A.

Commented [8]: Nick, is this correct? i.e., these are the only branches that meet the criteria in parentheses?

Deleted: of the

Figure B1. Relationship between Genealogies and the ARG. (A) Genealogical trees along the genome, corresponding to the ARG - each tree describes the ancestral relationship for each of the 4 non-recombined regions. $c_1, c_2, \ldots, c_6$ denote time points for each coalescence event. Trees can either change, 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. $c_1, c_2, \ldots, c_6$ denotes time points for each coalescence event. r_1, r_2, r_3 denotes time points for each recombination event.

Figure B2. Relationship between Genealogies and the ARG. (A) Genealogical trees along the genome, corresponding to the ARG - each tree describes the ancestral relationship for each of the 4 non-recombined regions. $c_1, c_2, \ldots, c_6$ denote time points for each coalescence event. Trees can either change, 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. $c_1, c_2, \ldots, c_6$ denotes time points for each coalescence event. r_1, r_2, r_3 denotes time points for each recombination event.

Figure B3. Schematic representation of Li and Stephens hidden Markov model. A new haplotype can be sampled as an imperfect copy of *n* reference haplotypes (hidden states). To find the most likely path taken through the hidden states, the LS model works along the genome (*k-1*, *k*, *k+1*, …), calculating the probabilities of changes in the attributed haplotype. The transition probability to continue or switch the attributed haplotype is a function of the recombination rate *(r)* between adjacent sites, whilst the emission probability to copy the attributed allele with or without error is a function of the mutation rate (*p*). Moving along the genome, the LS model compares the probability of every possible copying path and infers the most likely one.

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 (Griffiths & Marjoram, 1997; R. R. Hudson, 1983). 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).

Deleted: (Griffiths & Marjoram, 1997; Hudson, 1983)

The full ARG (Fig B1B) is a graph structure that depicts individuals (both ancestral and extant), lineage relationships in time. Each node in the ARG 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, an 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 t_2 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-recombining 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, F_{st}, 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.

Box 2: Methods for haplotype detection

Many methods for inferring evolutionary processes make use of haplotype structure. These can be roughly grouped into three types based on their underlying paradigm: window-based methods, segment-based methods and tree-based methods. These methods vary in complexity from simple heuristics to full statistical treatments. Here we discuss window-based and segment-based methods, but we reserve our discussion of tree-based methods to the main text.

Of the three classes, window-based methods tend to be the simplest, and primarily operate *across* sets of individuals. In the simplest form, haplotypes are operationally defined as the set of alleles observed at the segregating sites within a predefined window of an arbitrary length, say, 50 SNPs or 100 kilobase. Ideally, window sizes should be short enough to minimize spanning recombination breakpoints. One example is H_{12}, 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 (H_1 and H_2) to dominate the population. The H_{12} 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 (H. Li & Ralph, 2019; Todesco et al., 2020). While window-based methods do not explicitly infer or use information of haplotype block length, they sometimes do take the genealogical structure into account, e.g., *Twisst* (Lohse et al., 2016; Martin & Van Belleghem, 2017). Often, the simplicity of window-based methods is also their main appeal in the era of SNP genotyping.

Segment-based methods are more sophisticated. They 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 et al., 2002) is an excellent example of such segment-based statistics for inferring selection. Along with its derivatives, such as integrated haplotype score (iHS) (Szpiech & Hernandez, 2014; Voight et al., 2006) and cross-population EHH (XP-EHH) (Sabeti et al., 2007), they have been widely used to detect selection in many systems (Cao et al., 2011; International HapMap Consortium, 2005). 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 are especially helpful in uncovering admixture and introgression (e.g., fineSTRUCTURE (Lawson et al., 2012). This allows for the visualization of the haplotype-specific ancestry and improved fine-scale analysis of population structure that is not obvious from unlinked markers.

Box 3: Application and limits of Li and Stephens Model

Li and Stephens (2003) (LS) proposed a hidden Markov model (HMM) framework that underpins a large number 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 (S. R. Browning & Browning, 2007; B. N. Howie et al., 2009; Y. Li et al., 2010; Marchini et al., 2007; Stephens & Scheet, 2005), inference of population structure and demographic history (Hellenthal et al., 2014; Lawson et al., 2012; Steinrücken et al., 2019, 2018), characterisation of local admixture (Price et al., 2009; Sundquist et al., 2008), inference of local genealogies (Kelleher et al., 2019; Rasmussen et al., 2014; 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.

Deleted: (Browning & Browning, 2007; Howie et al., 2009; Y. Li et al., 2010; Marchini et al., 2007; Stephens & Scheet, 2005)

The LS algorithm requires a reference sample of haplotypes, or if presented in a sequence, previously observed 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 LS 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, which yields disjunct haplotype blocks.