**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. Our understanding of this term must depend on the processes of coalescence and recombination that generate haplotype structure. Thus informed, we contrast alternative definitions, and settle on one, which is based on branches in the underlying genealogy.

Samples of DNA sequence often appear to be structured into blocks, with a much smaller number of haplotypes than expected if alleles combine randomly (e.g. Fig. 1)*.* In other words, there is often strong haplotype structure, reflecting linkage disequilibrium amongst polymorphic sites; this is often described pragmatically in terms of "haplotype blocks". 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.

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, if there are only a few common haplotypes in the population, one resolution may become the most likely, so that “statistical phasing” is possible (Browning ref). A region of genome that is consistent with a particular common haplotype may then be referred to as a “block”. *Is this correct?*

Haplotype blocks can be defined in a more concrete way via the classical concept of identity by descent (IBD). Imagine an initial population, where each founder genome is labelled by a different colour. At some later time, each region of genome must derive from one or other founder, and so it is natural 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. Moreover, if we disregard mutation, 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. 3). 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 (Box1). We emphasise that these are real events: coalescence occurs when an 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 gives the structure of the ARG at one genetic locus.

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. 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 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 very small (especially with large samples) and will usually differ trivially (Fig. 3). Moreover, as samples get larger, blocks defined in this way will get ever smaller.

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 may carry derived SNP alleles that are shared by the specific set of genomes that were brought together in the focal branch by the corresponding coalescence event. With enough SNPs, the haplotype block is revealed directly by SNPs in that class (Fig. 4).

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 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 the same set of lineages by chance, so that we can usually identify unique coalescence events as corresponding to unique sets of lineages.