A haplotype has a clear definition: it is simply a haploid genotype, representing the maternal and paternal contributions to variation at a single site extending the notion of an allele from one site to many. Although it is a widely used meaning of the term “haplotype block" is less clear. . In this section we contrast some of the more widely used definitions, and settle on one, which is based on branches in the underlying genealogy.

The simplest use of the term haplotype block is probably in reference to the correlational structure present in sets of DNA sequences. Specifically, samples of DNA sequence are often characterised by fewer haplotypes than we would expect if alleles combine randomly. This strong haplotype structure, which can be measured in terms of linkage disequilibrium amongst polymorphic sites, can be used to impute missing genotypes, and 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 observed SNP. 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 based on 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. 1). 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, such as studies of experimental evolution, we can now sequence the founders, and thus directly observe blocks defined in this way. Moreover, the evolutionary processes subsequent to the founding of the population are entirely described by the block structure (if we disregard mutation).

However, when we deal with natural populations, there is no obvious reference population, and the block structure will depend arbitrarily on our choice of reference (Fig. 2). Therefore, we will base our definition on the full ancestry of the sampled genomes, namely, on the ancestral recombination graph (ARG) (box x). This 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. We emphasise that there 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. However, spatial and genetic structure can also be included

We could define a haplotype blockk 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.

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 SNP in that class.

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