Haplotype is a long standing term, which originally referred to the entire genome that individual inherits from a single parent in sexual species. However in practice, term haplotypes is used not in necessarily in application to the whole genomes, but to some limited region or even a single site. One could generalize this 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. The key problem of this haplotype blocks is that it doesn’t take into account origin of this structure, which is ultimately presented in genealogy of the sample, in simple words we would like to know where haplotype structure is coming from. 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, which in our opinion presents the best way to approach the definition for natural populations.

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.~~

Originally haplotype blocks were observed empirically as characteristic set of haplotypes in the agricultural populations and raised the idea of identity by descent. Concept of identity by descent goes back to 1920-30th and proposed as an measure of degree of inbreeding in agricultural populations. Haplotype blocks can be defined in a more concrete way based on the classical concept of identity by descent (IBD). While talking about identity by descent we usually mean identity with respect to certain reference population. Imagine an initial population, where each founder genome is labelled by a different colour. At some later time, typically tens of generations, 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).

As we stated before, choice of the reference population is crucial in terms of this definition and imposes certain limitations on defining related haplotypes (in other words “chromosome painting”?). Figure 1 demonstrates effect of choice of ancestors/time point to define haplotype blocks. Earlier time point Ta leads to observing three different haplotypes, recombination event leads to reshuffling of ancestors, which is visible in change of block color. While ancestors defined at later (closer to present) time point all haplotypes are colored identically and ancestry information is lost. Note that mutation state, which corresponds to empirical observation, is identical between A and B.

However, when we deal with natural populations, there is no obvious reference population, so we have to think about the actual full ancestry of the sample, which is represented genealogies in the population. In theory each pair of individuals within certain species have some regions where they will share haplotype block (be identical by descent), but this genealogy may be determined by very long branches and this ancestral information may not have practical use. In contrast, above mentioned IBD definition will impose very recent ancestry in the sample. While working with the natural population we would like to have information about full ancestry and use branches of various length to define haplotype ancestry based on question of interest. Complete description of evolutionary relationships and ancestry within population is contained in specific structure: ancestral recombination graph (ARG).

Therefore, we will base our definition on the full ancestry of the sampled genomes - the ancestral recombination graph (ARG) (box x). As ARG is integral for our definition, we start from explaining it in order to get intuitive understanding first and then moving to more theoretical aspects. If we look at any particular point in the genome we will observe genealogical relationship between all individuals, which can be visualized as a tree like structure. All genes in the population will coalesce and eventually trace back to a one single ancestor. As we move along the genome at some point genealogy will change because the lineage tracing back undergone recombination event. Crossover during meiosis caused half of the recombination break come from one parental chromosome and another from other parental chromosome, therefore recombination event leads to splitting of the linages if we go back in time.

When we think of individual’s ancestry in many generations before, number of ancestors will be very large and therefore the genome traced back through these generations will be scattered among hundreds and hundreds of individuals in little blocks. Number of ancestors fluctuated in the random process due to drift in the population, coalescent will take into account situations when two different individuals produced by the same ancestors. ARG is interplay between processes of recombination, which increases amount of ancestors and process of coalescent, which reduces it.   
“The ARG naturally defines a set of recombination breakpoints, a set of haplotypes, and a genealogy for each non-recombining interval in the genome” (Hubisz, Siepel 2020).

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.

*Need to also explain that branches are present in regions of the ARG (Fig. 3c) and that they may be disjunct.*