Definition & Simulations

In a simplest and earliest definition, a **haplotype** (or haploid genotype) is a combination of alleles in an organism that all have been inherited together from a single parent. In application to the selection experiment this definition needs to be extended to reflect joint inheritance of haplotype for a set of ancestors (Fisher 1952). This definition by itself doesn’t directly imply length of the haplotypes. We will refer to this definition as of identity by descent (IBD).

This definition is tailored to define blocks relative to **recent** common ancestors and in some cases, there is an obvious reference population (e.g. in a selection experiment). However, choice of common ancestor is arbitrary, since technically speaking each pair of the haplotypes has a common ancestor at each position of the genome, but this ancestor may have lived many millions of years ago. We illustrate IBD-based definition on the Fig. 1. While visualizing ancestry of the sample, one can notice, that choice of set of the ancestors or rather ancestor labels produce different number and assignment to haplotypes (shown in color) and doesn’t fully reflect genealogy of the sample.

Information of the genealogy of the sample is not reflected directly is the mutational state, however one can summarize it in the *ancestral recombination graph* (ARG). Ancestral recombination graph(ARG) is generated by a simple coalescent process and traces genomes back in time through a series of recombination and coalescence events, which leads to all possible ancestors. By including this full genealogical information, we are observing multiple branches leading to sets of haplotype blocks. For a sample of genes, or even, for the whole population, we could define a haplotype block as a region that shares the same genealogy. When defined by a specific coalescence event, haplotype block can be represented multiple times along the genome, and within each of the several contiguous segments, will trace back to different coalescent events (Fig. 2).

Using this definition, in which blocks descend from some branch, they can be detected through carrying mutations with a certain configuration and the number of such mutations is Poisson distributed with rate proportional to the area occupied by the block. Mutations on a branch will be shared by that set, and so we can (approximately) identify a block by the presence of such mutations.

Typically, genealogies are dominated by a few long branches, which will be distinguished by carrying alternative sets of mutations (Fig. 2). In a well-mixed population, the rate of coalescence is proportional to the number of pairs of lineages, and is therefore very fast for a large sample; the time to coalesce from many lineages down to two equals, on average, the time taken for those two to coalesce. In the simplest case of a single causal locus one can think of the focal genealogy plus associated material - arguably, all that matters is to know the focal genealogy, but we get information about it from junctions as well as mutations.

**Final definition:**

Leftover text:  
  
The most common practice defines haplotype block as a segment of the genome inherited from a single common ancestor, with no recombination occurring in the lineages of interest. As sequence evolves through time the common ancestor sequence is additionally disrupted by mutation, which can result in blocks identical by descent in fact looking different. Genotype error can be another source of uncertainty causing misinterpretation of the haplotype block.

In some methods, haplotype blocks are directly referred to as identity by descent (IBD) segments and length of the haplotype block equivalent to identification of their “endpoints” (Browning & Browning, 2020).

However, adjacent genealogies, which differ by a single pair of recombination and coalescence events, usually differ trivially, and undetectably.

We argue that the haplotype block should be defined in terms of the ARG; then, we see that haplotype blocks correspond to sets of genomes that descend from particular branches.