Definition & Simulations

One of the first way to represent haplotypes or rather haplotype blocks comes from thinking about genomes in terms of identity by descent (IBD) (as for example in Fisher 1952) was in terms of IBD with a reference population. This depends on when one chooses the reference, and so we go on to develop a definition based on the ARG.

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). In general, 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. 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).

* Definition through ancestral recombination graph (ARG)

The *ancestral recombination graph* (ARG) traces genomes back to their various ancestors, through a series of recombination and coalescence events; in the absence of selection, it is generated by a simple coalescent process. 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. 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. 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 corresponds to the length of the branch. The number of SNPs 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. 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.

**Conclusion:** haplotype block is a segment of the genome inherited from a common ancestor, with no recombination occurring in the lineages of interest. The genealogy of the sample and branch length contains information about all blocks regardless of the time point where the common ancestor(s) is defined.

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.