Understanding haplotypes

## Outline/Paragraph Plan

#### Introduction

“Haplotype”” and “Haplotype block” are widely used terms, and have increased in importance for several reasons:

1. Emergence of the new methods for obtaining reliable haplotype information: molecular phasing using linked reads, haplotagging.
2. Limitations of widely used site-based statistics are recognized, need for alternative approaches arises
3. Rapid development of methodology for genealogy inference (tsinfer, Relate)
4. Phasing and imputation depend on assumptions about haplotype structure
5. Haplotypes are increasingly used in inference, of selective sweeps, introgression, and population structure.

#### Definition & Theory

* Original and simple meaning of “haplotype”, other synonyms
* Definition through identity by descent (IBD)

“Haplotype blocks” can be defined through identity by descent from a reference population. In some cases, there is an obvious reference population (eg in a selection experiment), but in general, it is arbitrary.

* Definition through ancestral recombination graph (ARG)
  1. We should think in terms of the ARG; then, we see that haplotype blocks correspond to sets of genomes that depend from particular branches.

Using this definition, in which blocks descend from some branch, they can be detected through carrying mutations with a certain configuration;

The number of such mutations corresponds to the length of the branch.

Close relation with Konrad’s blockwise SFS, which also looks at the numbers of each mutational configuration in a window.

That extra information comes from knowing the block length (and how exactly do we define that?)

* How length of the haplotype block is defined? What information does it carry? What definitions are used in the literature?
* Importance of haplotype length information for inference

#### Length of the haplotype block: simulation example

1. What type of analysis can be improved while using haplotype length and frequency as a statistics? What is the new information we gain? What kind of analysis will gain extra power?
2. Simulation example:

* false positive - a region with (by chance) an unusually short genealogy
* neutral
* hard sweep
* island model
* balancing selection

#### Practical definition of the haplotype block and application of haplotype block length

1. Approaches to infer haplotype blocks: phasing, haplotype block inference.
   1. Phasing software underlying assumptions about haplotype structure:

Sequential Markov coalescent (SMC) is assumed as an approximation, whenever a Hidden Markov Model is used.

* 1. “Blockers”: HaploBlocker, HaploView
  2. How accurate are the algorithms in inferring the “true” blocks?

(hard to determine without a lot of simulation, but maybe one can make some comment)

* 1. Genealogy-aware algorithms (tsinfer, Relate):

Ancestral Haplotype reconstruction serves as a basis for block length definition

1. Using haplotype block length in inference of selection
   1. EHH and derivatives: practical definition of the haplotype block
2. iHS (Integrated Haplotype Score), xpEHH, w
   1. Is using Relate output is better approach for inferring selection?
3. Using haplotype block in population structure inference
   * 1. Definition of the haplotype block in inference of population structure, LD
     2. Relation to ADMIXTURE methods
     3. fineSTRUCTURE - haplotype-based inference approach (special case: RAD data)
4. How can we do better in the inference above?

## Summary and conclusion