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. Possibility for a haplotype statistics **box** here

#### Definition & Theory

Main point: haplotypes blocks contain both topology and mutation information

1. Original and simple meaning of “haplotype”, other synonyms
   1. What definitions are used in the literature?
   2. Brownings: “identity by descent (IBD) segments endpoints”
   3. Distinguish “haplotype” (simple) from “haplotype block” (subtle)
2. Definition through identity by descent (IBD)
3. Definition through ancestral recombination graph (ARG): blocks descend from the branch, can be detected through carrying mutations with a certain configuration

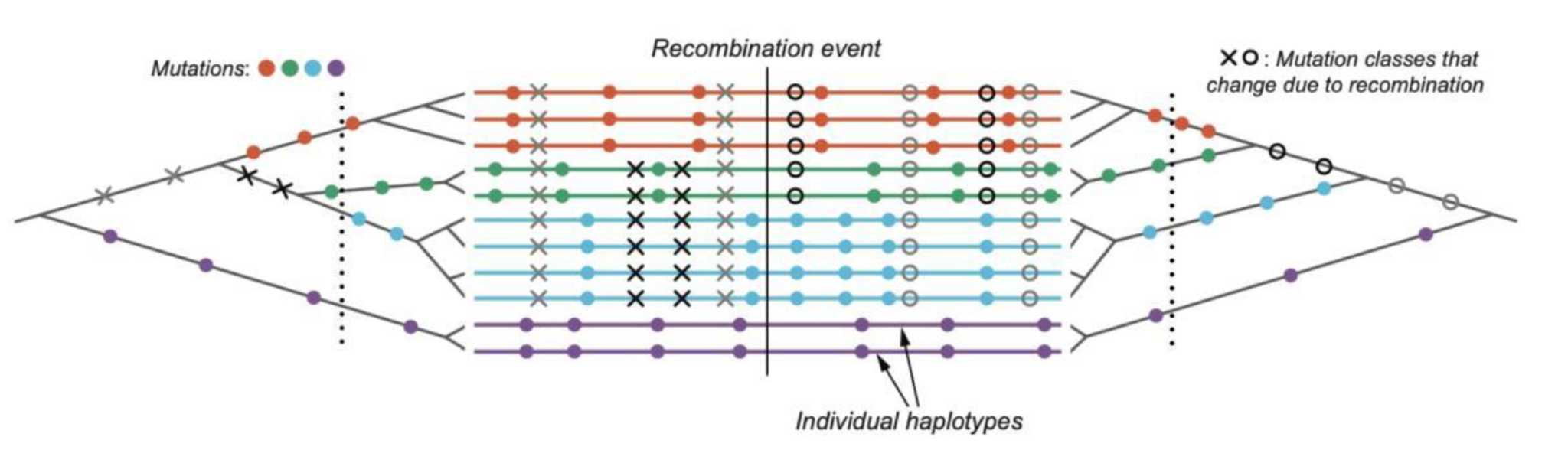
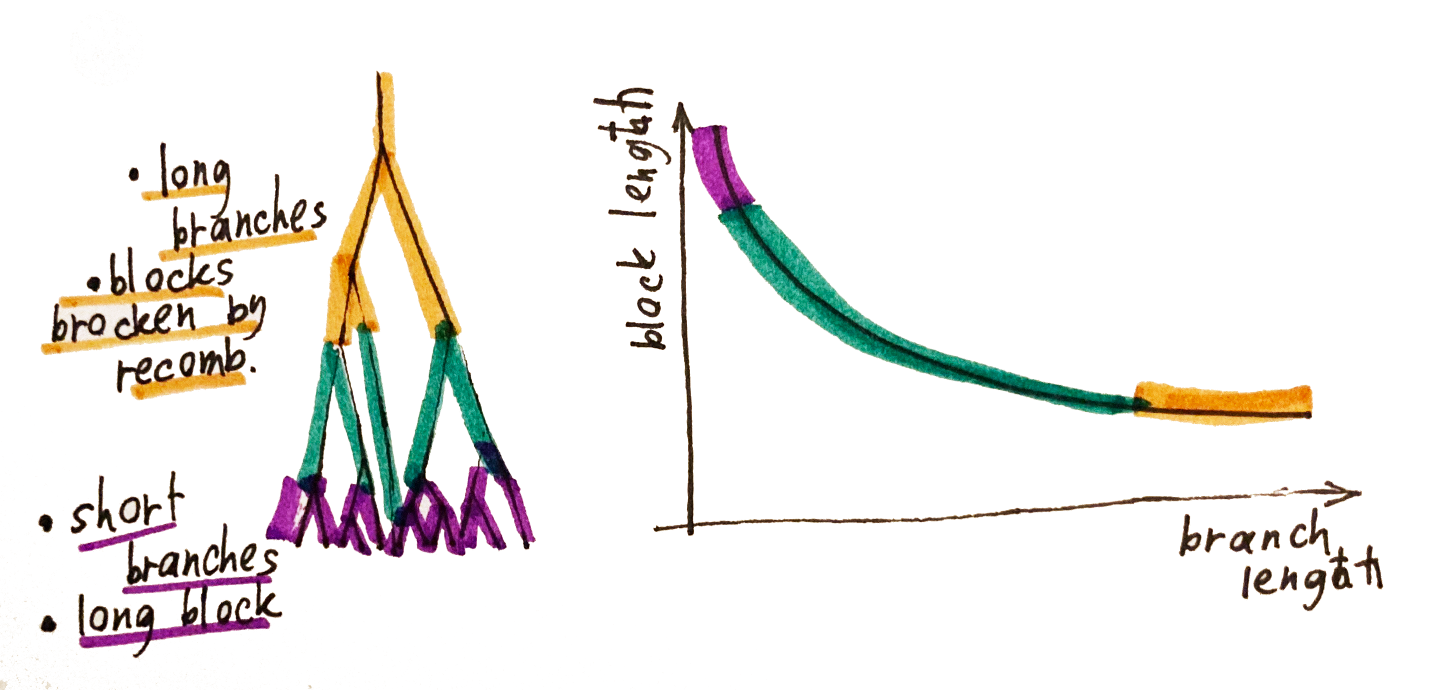


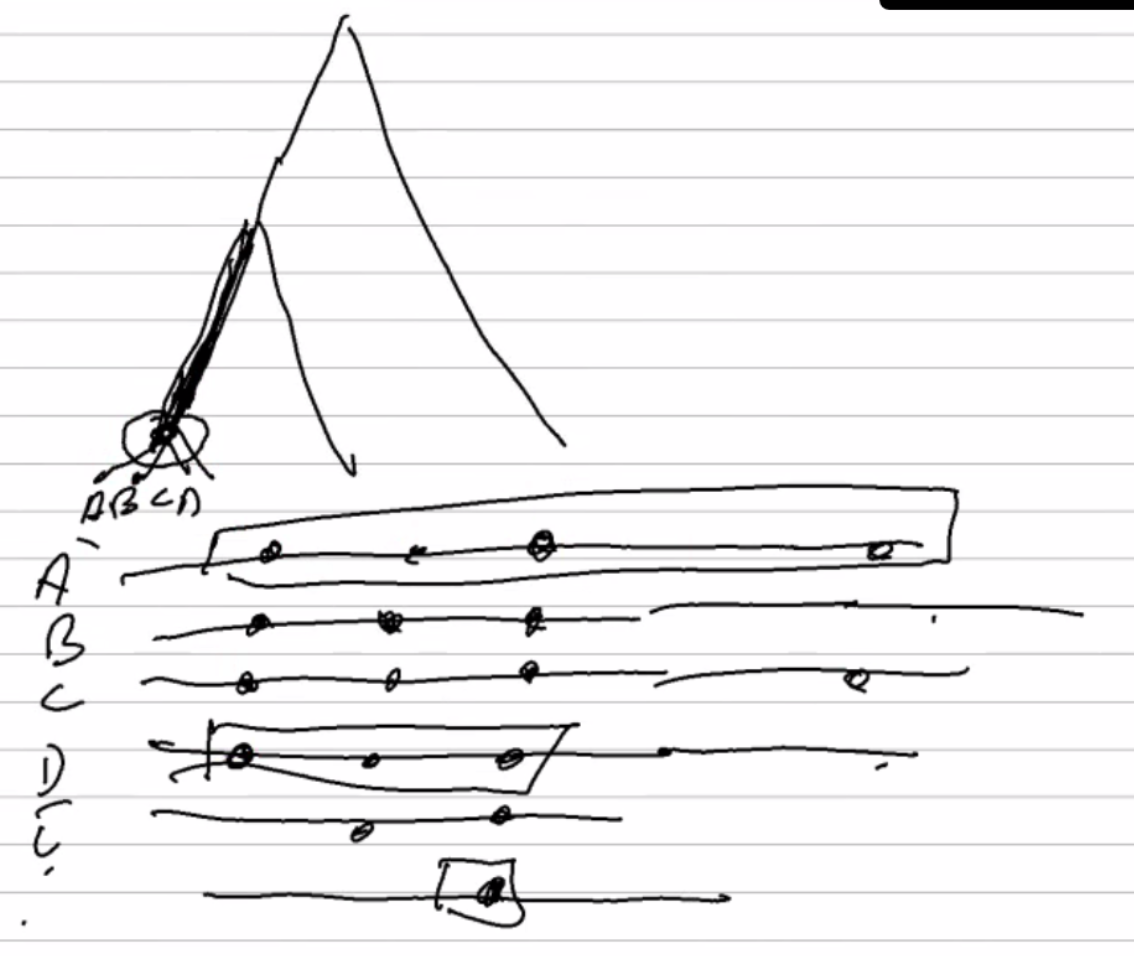
Fig.1

1. In order to be able to identify haplotype block using branch information, we need to understand the relationship between branch length and block length



**Fig.2** Example of one locus, neutral model

1. Secondly, we need to understand how haplotype and branch length are expected to behave under various models. (Simulations here?)
   * 1. false positive - a region with (by chance) an unusually short genealogy
     2. Neutral (above already)
     3. hard sweep
     4. balancing selection
     5. island model

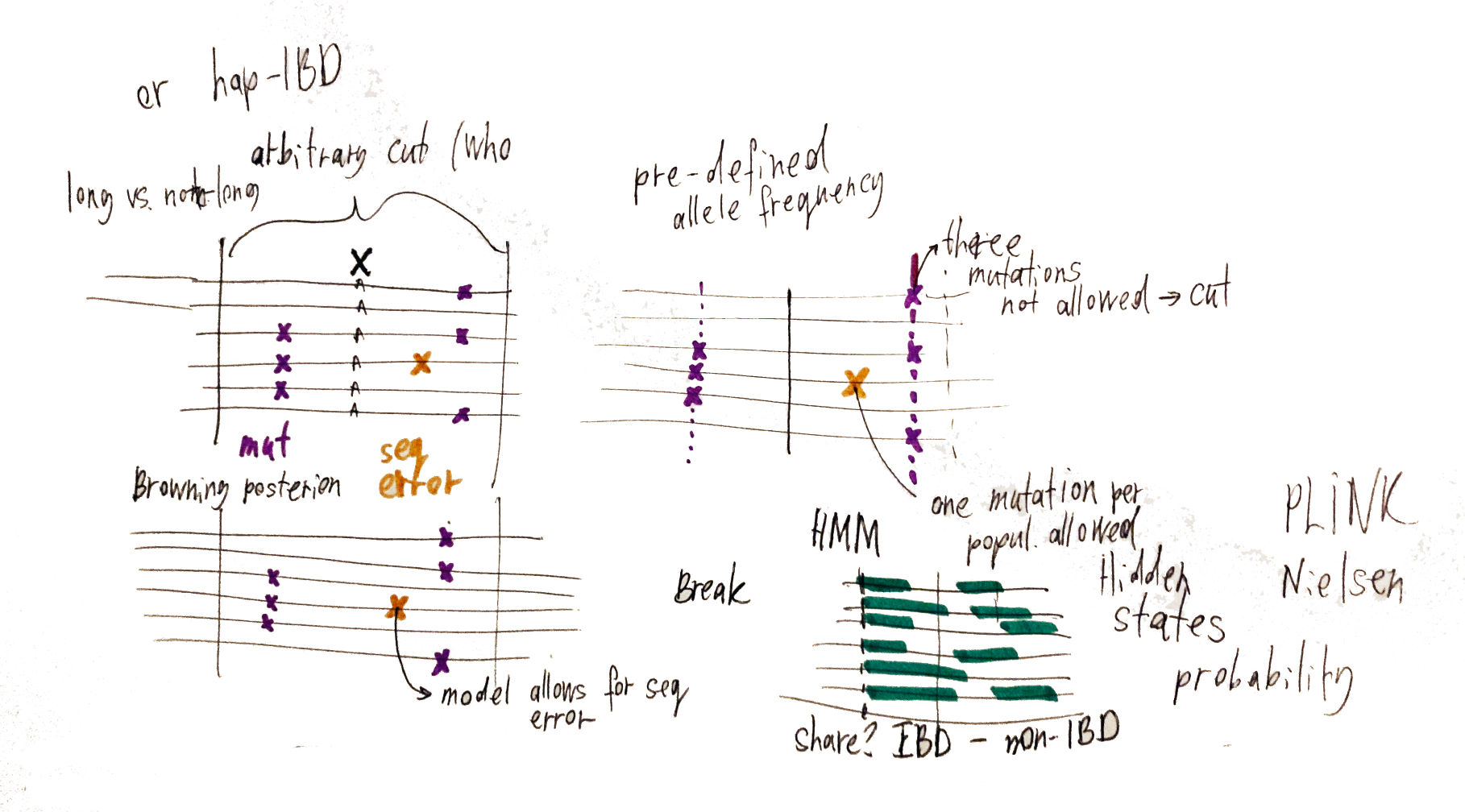


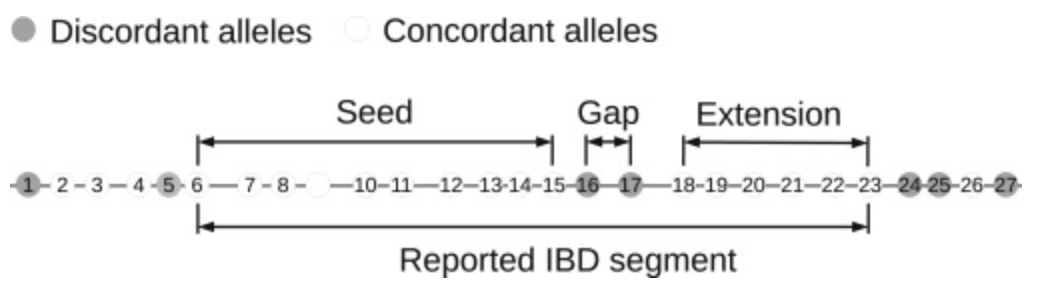
**Fig.3**  Branch length and haplotype block length distribution in the hard sweep

1. Above examples demonstrate how one should approach definition of the haplotype block, while keeping in mind, that it is still a “crude measure”.

#### Practical definition of the haplotype block

1. In this section we consider how haplotype length is defined in practice by different methods and demonstrate how it is incorporated into different methods of inference





**Fig. 4** Haplotype block as defined by different methods and approaches (below figure from Brownings)

1. Some methods simply define haplotype block by setting up a window of arbitrary length (proportional to rec rate) (fineRadSTRUCTURE, w, gIMble)
2. Majority of the statistics incorporate some sort of model-based definition, there are two approximations
   1. Phasing software underlying assumptions about haplotype structure, how could it influence definition of the block length?
      1. HaploBlocker: haplotype block is a sequence of genetic markers that has a predefined minimum frequency.
   2. Sequential Markov coalescent (SMC) is assumed as an approximation, whenever a Hidden Markov Model is used. Limitations of HMM
3. Genealogy-aware algorithms (ARG weaver, tsinfer, Relate, CLUES):
   1. Ancestral Haplotype reconstruction serves as a basis for block length definition (tsinfer) and HMM again
4. ROH?

#### Summary and conclusions

1. ???
2. How can we do better in the inference above