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

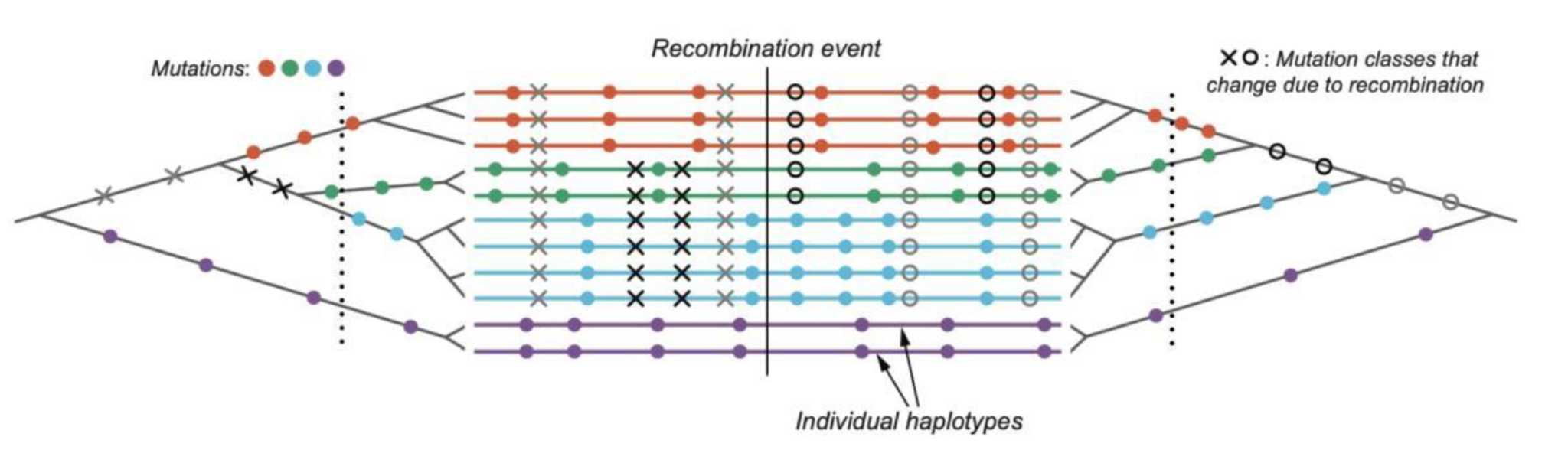
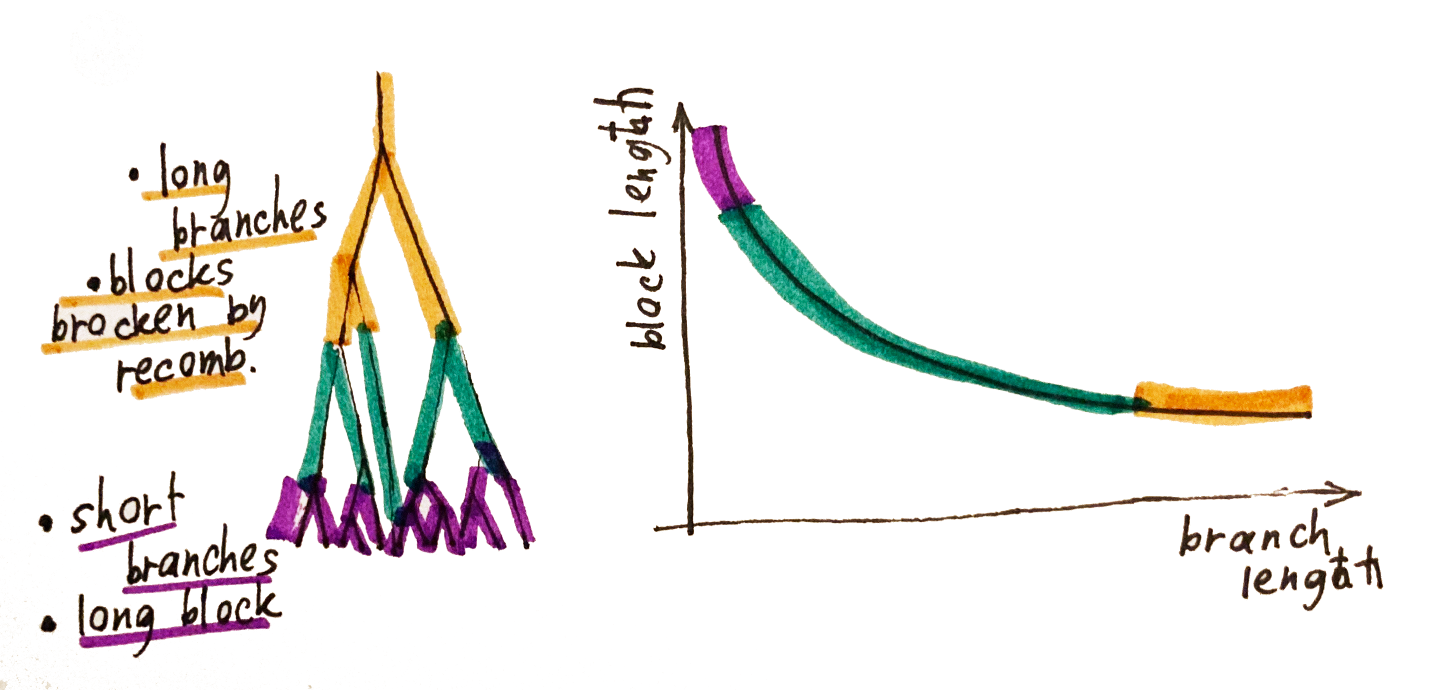


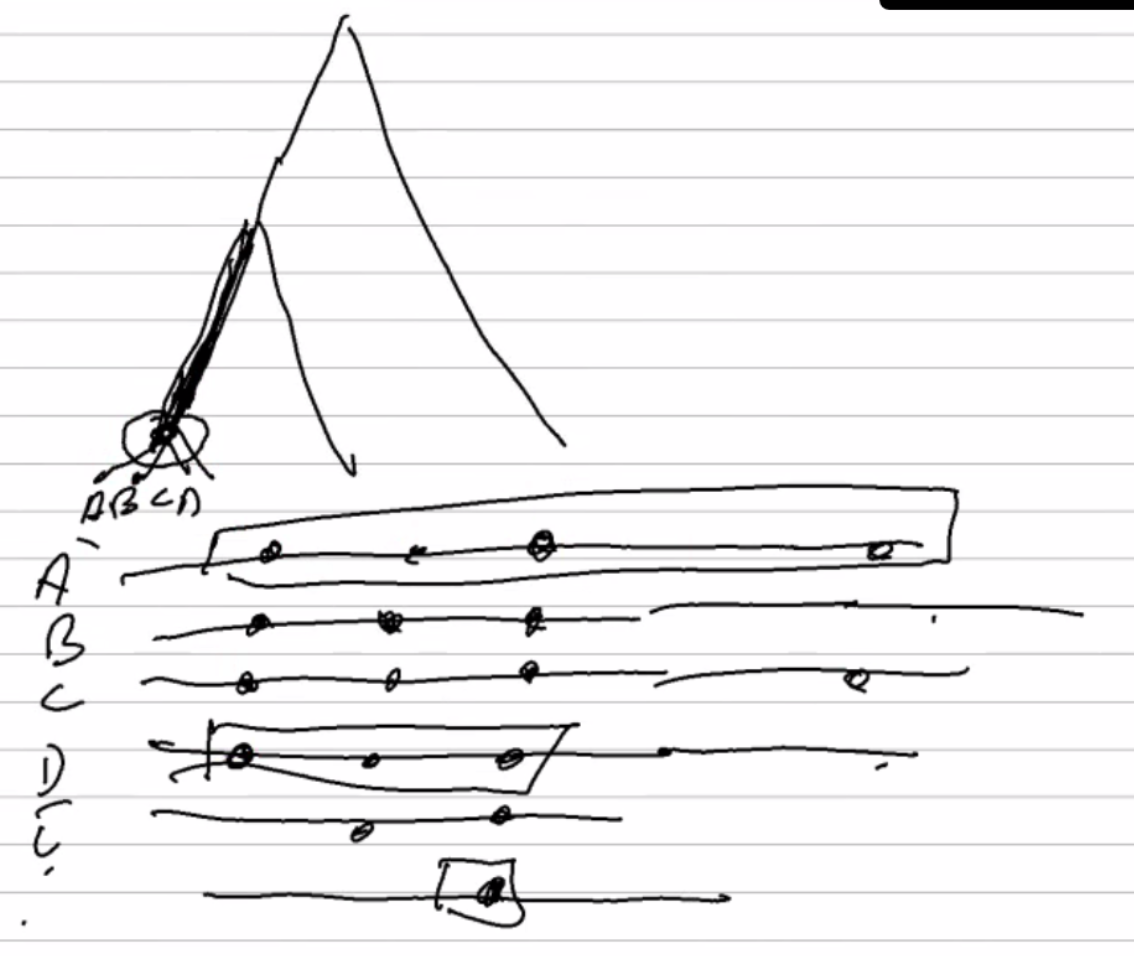
Fig.1

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

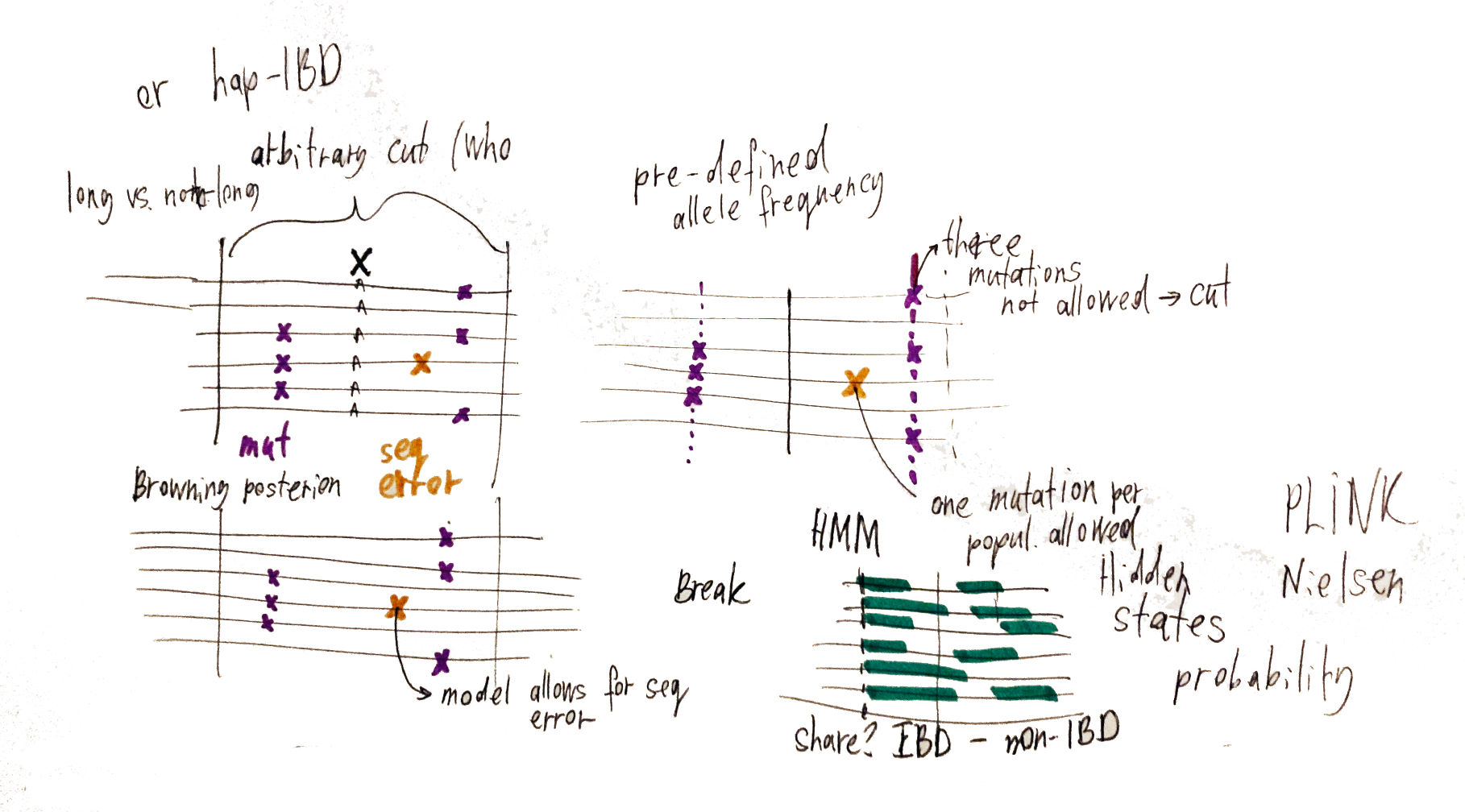


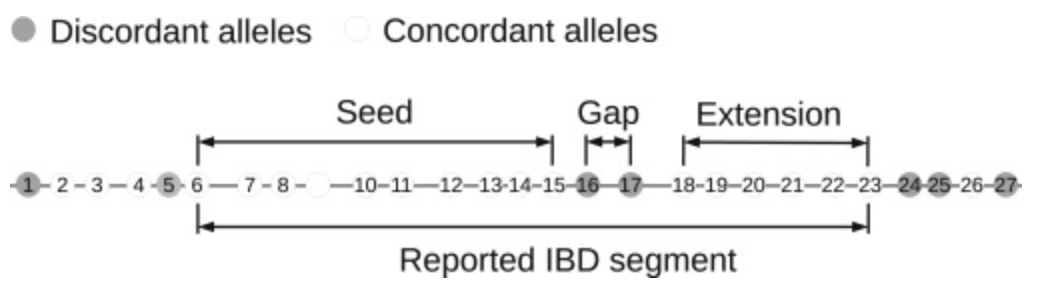
**Fig.3**  Branch length and haplotype block length distribution in the hard sweep

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

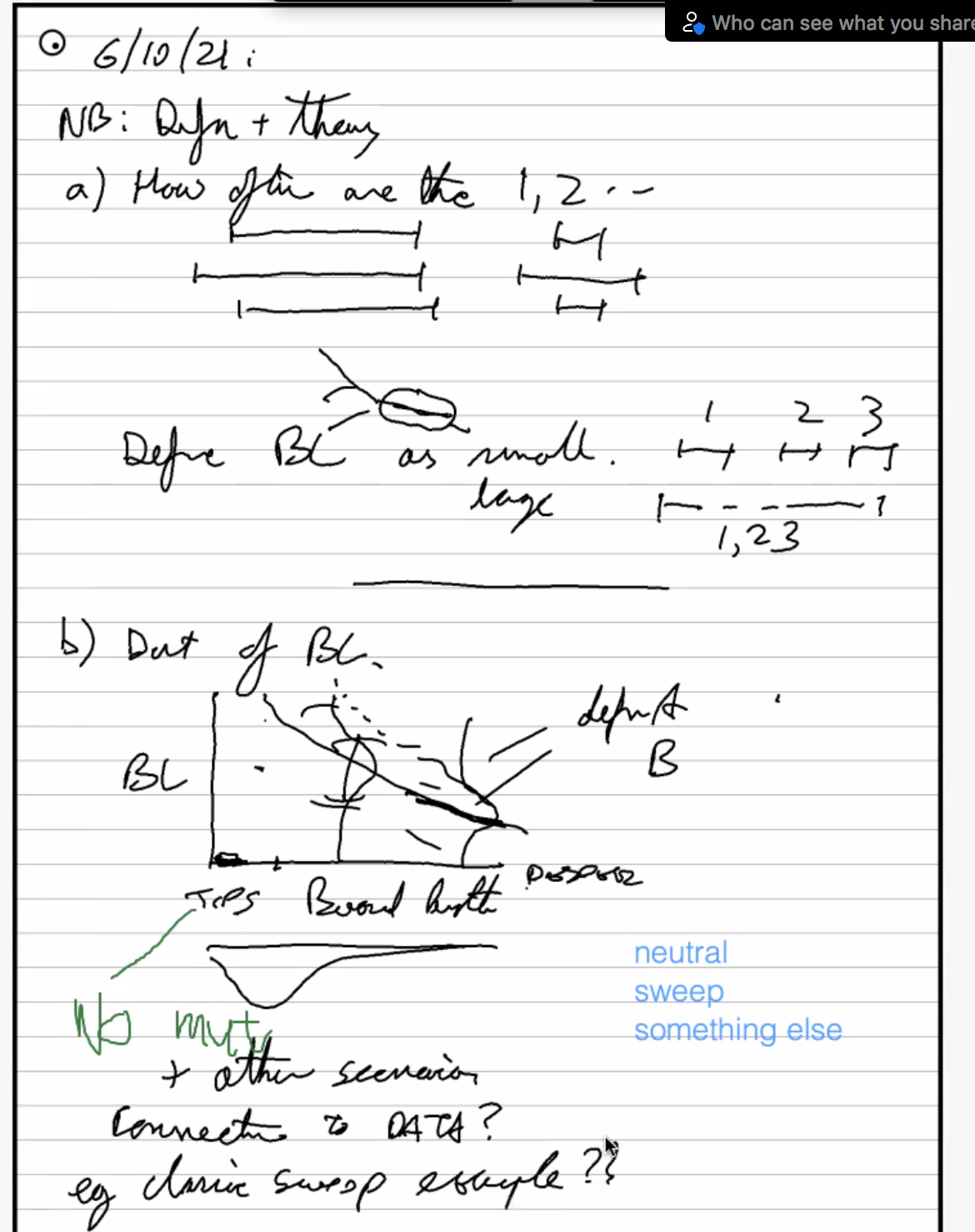
Bits and pieces:

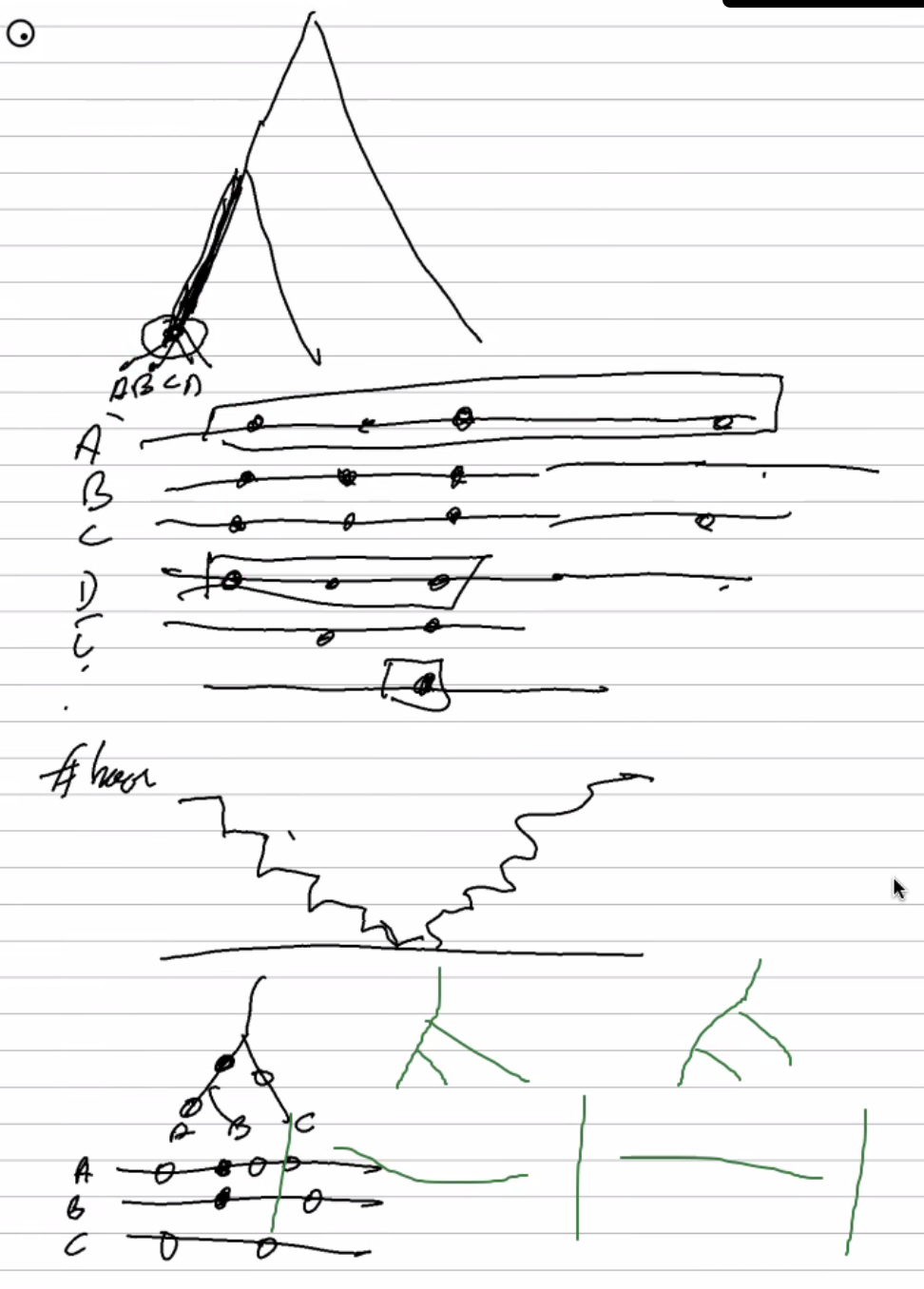
1. Citing them: “An IBD segment for a pair of haplotypes is a segment of DNA inherited from a single common ancestor, with no crossovers occurring within the segment in the lineages of the two haplotypes since the common ancestor. Within a shared IBD segment, sequence identity can be disrupted by mutation and gene conversion. In addition, genotype error can cause two haplotypes to appear to be discordant at a position. At such positions, two “identical by descent” haplotypes are in fact not identical. This non-identity needs to be considered when detecting IBD segments.”
2. *A key point to make is that some branches matter more than others, and some recombinations matter more than others. We can make this point with a simple illustrative example*

That extra information comes from knowing the block length (and how exactly do we define that?) *Well, there is much more information than just block length. 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.*

* 1. How length of the haplotype block is defined? What information does it carry? What definitions are used in the literature?
  2. Importance of haplotype length information for inference
  3. *Emphasise that even a haplotype block is still a crude measure - if we have multiple (or infinitesimal) selected sites we have a much richer (nightmare?) scenario…*

*Selection \_> CLUES focal genealogy -> selection in variable sites*

**



#### Length of the haplotype block: simulation example Example/Simulation/Cartoon

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. Theoretical expectations on length distribution
3. Simulation example: *as illustrations!*

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

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

In this section we consider how haplotype length is defined in practice by different haplotype-based methods and how it’s incorporated into different softwares for ....

Main points to consider:

1. Some methods simply define haplotype block by setting up a window of arbitrary length (proportional to rec rate) (fineRadSTRUCTURE, w, gIMble)
2. Sometimes data doesn’t need to be phased? Does phasing have real advantage for inference? *Maybe a discussion point?*
3. 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
4. Genealogy-aware algorithms (ARG weaver, tsinfer, Relate, CLUES):
   1. How accurate are the algorithms in inferring the “true” blocks?
      1. Human vs. non-model
   2. New and exciting algorithms are incorporating older ones and inherit underlaying assumptions
   3. Ancestral Haplotype reconstruction serves as a basis for block length definition (tsinfer) and HMM again
5. Using haplotype block length in inference of selection (BOX?)
   1. EHH and derivatives (iHS, xpEHH): practical definition of the haplotype block
   2. H12
   3. Is using Relate output is better approach for inferring selection?
6. 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
7. Balancing selection? Hybrid zones?
8. More about length itself?
9. *A key point to make is that some branches matter more than others, and some recombinations matter more than others. We can make this point with a simple illustrative example*

#### 

#### Summary and conclusion

1. How can we do better in the inference above

## 

## 

## Outline/Paragraph Plan

#### Introduction

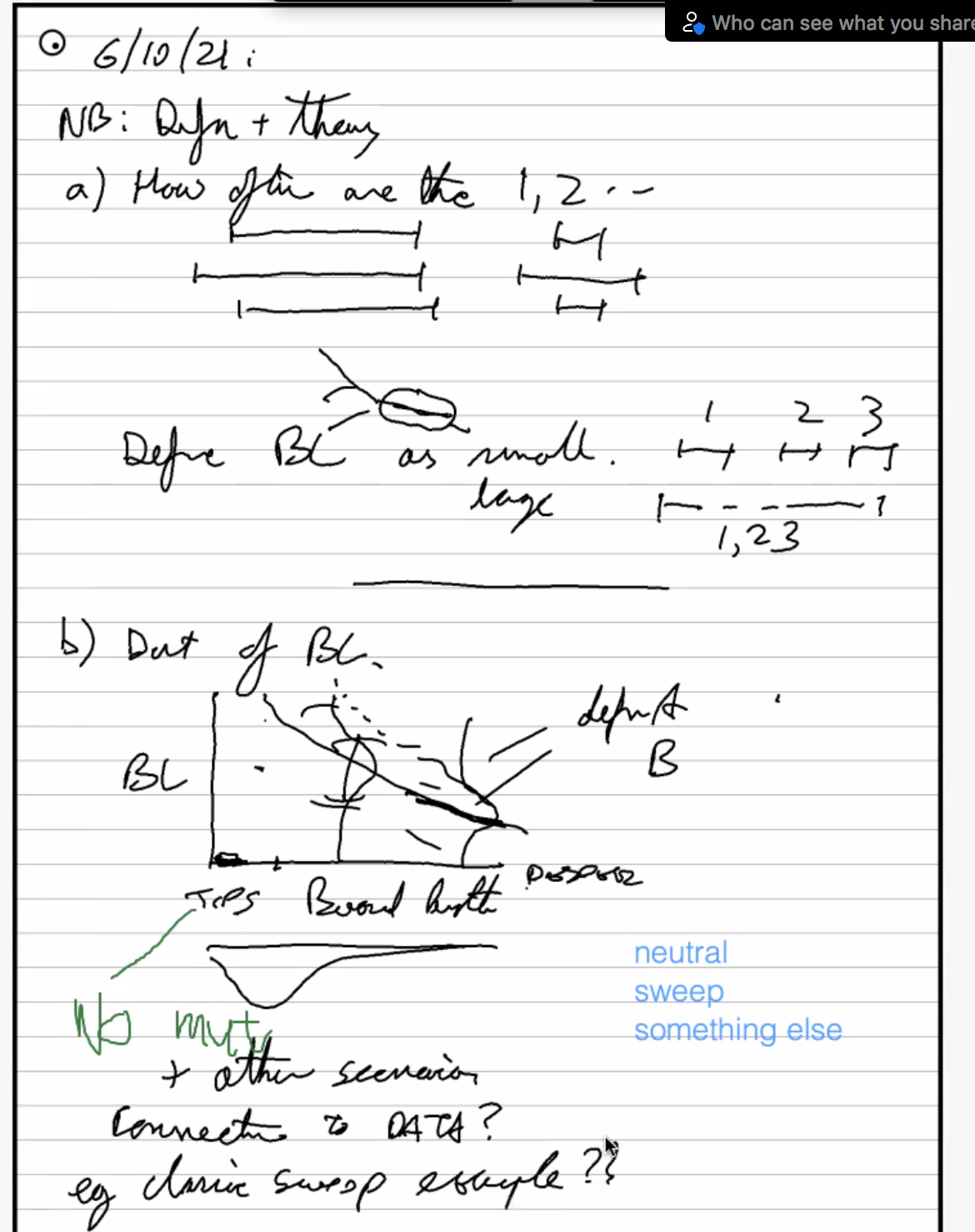
“Haplotype”” and “Haplotype block” are widely used terms, and have increased in importance for several reasons:

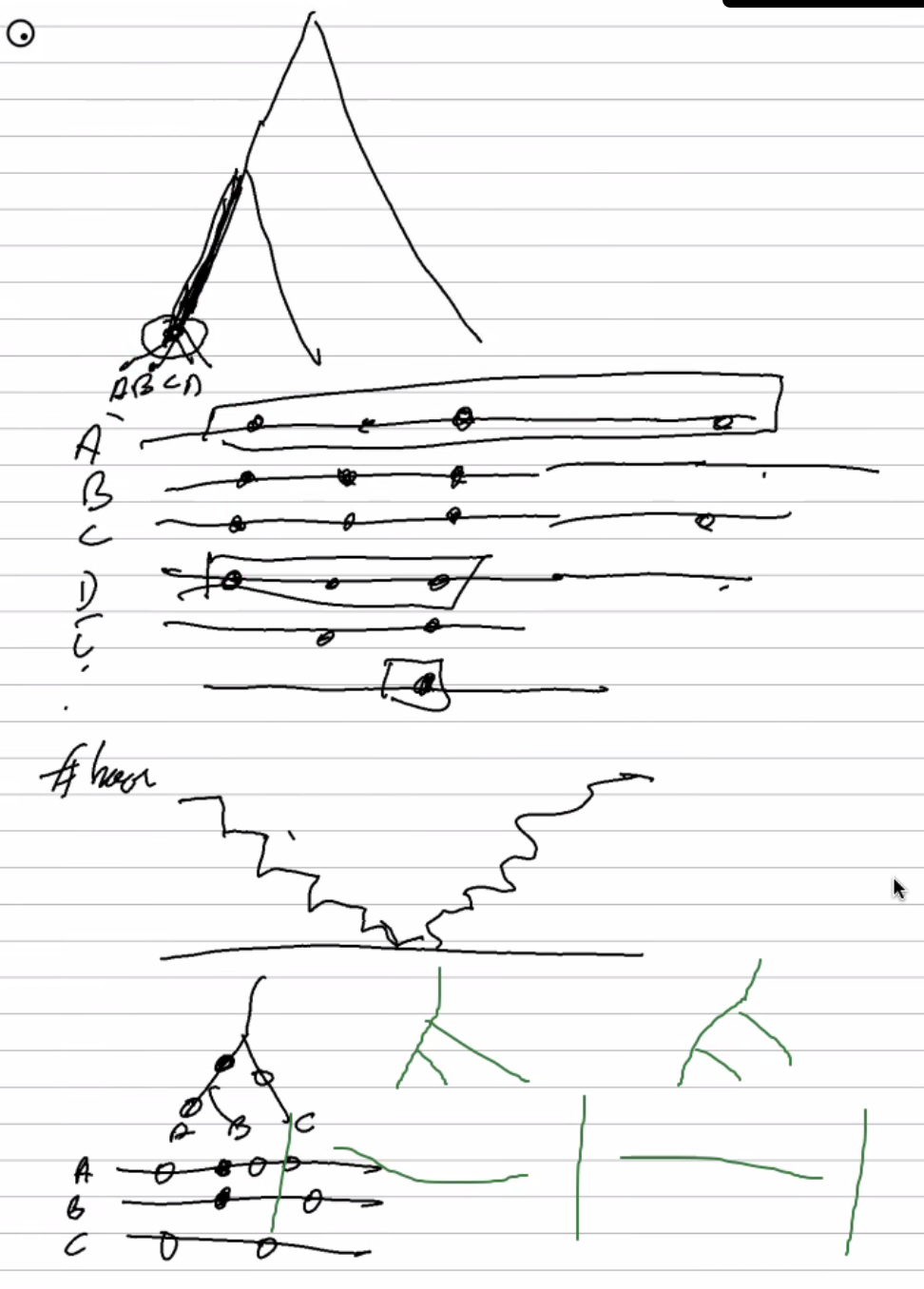
1. Emergence of the new methods for obtaining reliable haplotype information: molecular phasing using linked reads, haplotagging.
2. Limitations of widely used site-based statistics are recognized, need for alternative approaches arises
3. Rapid development of methodology for genealogy inference (tsinfer, 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

1. Original and simple meaning of “haplotype”, other synonyms (What definitions are used in the literature?)
2. Distinguish “haplotype” (simple) from “haplotype block” (subtle)
3. Definition through identity by descent (IBD): correct, but unavailable in practice *not really - one can have the founder genomes in an experiment. More important, it is defined relative to an arbitrary reference population.*
4. Definition through ancestral recombination graph (ARG): blocks descend from the branch, can be detected through carrying mutations with a certain configuration (How variable are they?)
5. That extra information comes from knowing the block length (and how exactly do we define that?) *Well, there is much more information than just block length. 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.*
   1. How length of the haplotype block is defined? What information does it carry? What definitions are used in the literature?
   2. Importance of haplotype length information for inference
   3. *Emphasise that even a haplotype block is still a crude measure - if we have multiple (or infinitesimal) selected sites we have a much richer (nightmare?) scenario…*

*Selection \_> CLUES focal genealogy -> selection in variable sites*

**



#### 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. Theoretical expectations on length distribution
3. Simulation example: *as illustrations!*

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

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

In this section we consider how haplotype length is defined in practice by different haplotype-based methods and how it’s incorporated into different softwares for ....

Main points to consider:

1. Some methods simply define haplotype block by setting up a window of arbitrary length (proportional to rec rate) (fineRadSTRUCTURE, w, gIMble)
2. Sometimes data doesn’t need to be phased? Does phasing have real advantage for inference? *Maybe a discussion point?*
3. 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
4. Genealogy-aware algorithms (ARG weaver, tsinfer, Relate, CLUES):
   1. How accurate are the algorithms in inferring the “true” blocks?
      1. Human vs. non-model
   2. New and exciting algorithms are incorporating older ones and inherit underlaying assumptions
   3. Ancestral Haplotype reconstruction serves as a basis for block length definition (tsinfer) and HMM again
5. Using haplotype block length in inference of selection (BOX?)
   1. EHH and derivatives (iHS, xpEHH): practical definition of the haplotype block
   2. H12
   3. Is using Relate output is better approach for inferring selection?
6. 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
7. Balancing selection? Hybrid zones?
8. More about length itself?
9. *A key point to make is that some branches matter more than others, and some recombinations matter more than others. We can make this point with a simple illustrative example*

#### 

#### Summary and conclusion

1. How can we do better in the inference above

IBD scans look for genomic regions that have a significantly higher than average number of IBD segments. If the genome were completely neutral, and there are no biases in detecting IBD segments or estimating their centiMorgan (cM) lengths, the expected number of IBD segments exceeding some cM length threshold would be constant across the genome.

(Text draft)

## Introduction

0. (Maybe we need to start introduction from discussing long, linked reads technology and then move on to better phasing to haplotype blocks)

1. “Haplotype”” and “Haplotype block” are widely used terms, and have increased in importance for several reasons:
2. Emergence of the new methods for obtaining reliable haplotype information: molecular phasing using linked reads, haplotagging.

It will soon be possible to use haplotype-based methods which were previously only available for model organisms. We will discuss the methods which are likely to gain popularity in the near future, discuss their advantages and disadvantages and propose further directions for method development.

1. Limitations of widely used site-based statistics are recognized, need for alternative approaches arises
2. Rapid development of methodology for genealogy inference (tsinfer, Relate)
3. Phasing and imputation depend on assumptions about haplotype structure
4. Haplotypes are increasingly used in inference, of selective sweeps, introgression, and population structure.

## Definition & Theory

1. Original and simple meaning of “haplotype”, other synonyms
2. 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.

1. 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?)

1. How length of the haplotype block is defined? What information does it carry? What definitions are used in the literature?
2. 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
* false positive - a region with (by chance) an unusually short genealogy
* island model
* balancing selection

## Practical definition of the haplotype block

1. Approaches to infer haplotype blocks: phasing, haplotype. We focus here on non model organisms
   1. Phasing software underlying assumptions

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:

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): length of derived haplotype is longer then length of ancestral one, based on LD. Limitation: point estimate for one focal gene. iHS is implemented in …
3. Mention xpEHH
4. w?
   1. Is using Relate output is better approach ?
5. Using haplotype block in population structure inference
   * 1. Definition of the haplotype block is inference of population structure
     2. Relation to ADMIXTURE methods
     3. fineSTRUCTURE
     4. Special case: RAD data

## Summary and conclusion

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

To address this haplotype-structure limitation, researchers have introduced haplotype-based summary statistics such as *H*1, *H*12, and

*H*1 measures the frequency of the most abundant haplotype, which assumes high values in a hard sweep owing to local reductions in heterozygosity. Similarly, *H*2 measures the frequency of the second most common haplotype.

By using discrete time points and enumerating tree topologies, ARGweaver approximates the continuous state space of the SMC by a finite set, which permits the use of standard dynamic programming algorithms for hidden Markov models (HMMs) in ARG inference. The main innovation of ARGweaver is to use an MCMC algorithm to sample only a portion of the ARG at a time, in such a manner that the sequence of sampled ARGs is guaranteed to eventually converge to the desired posterior distribution. ARGweaver enables the recovery of the distribution of the local genealogies (topology and branch lengths), recombination breakpoints, recombination rates, time to most recent common ancestry (TMRCA) and other derived statistics, as well as the recovery of allele ages. It can additionally accommodate unphased data, missing genotypes, and ancient DNA samples. The main disadvantage of ARGweaver is its computational cost, particularly as the number of individuals increases. The method can currently only be applied to a few dozen individuals at a time.

tsinfer

tsinfer [[109](https://www.sciencedirect.com/science/article/pii/S0168952519302690?via=ihubm#bb0545)] is an ultra-fast, heuristic ARG inference method that scales to hundreds of thousands of complete genomes. The method requires an input of biallelic sample haplotypes, phased data, and ancestral/derived states. It proceeds by reconstructing ancestral genome fragments for each site, and then inferring the relationships among these fragments according to an ancestral copying process. After explaining all haplotypes in this manner, the method outputs a tree sequence. Notably, tsinfer does not explicitly infer an ARG but rather a sequence of local gene trees, described by their topologies only. tsinfer was shown to be substantially faster and more scalable than ARGweaver and RENT+, with comparable topological accuracy to ARGweaver [[109](https://www.sciencedirect.com/science/article/pii/S0168952519302690?via=ihubm#bb0545)]. Importantly, however, these comparisons were based on tree topologies only and ignore the absence of branch lengths in tsinfer’s reconstructed ARG. Additionally, like RENT+, tsinfer infers a single-point estimate of an ARG, rather than allowing for gene tree uncertainty.

It is worth noting that RELATE and tsinfer have one major similarity in that they both use approaches based on a Li-and-Stephens algorithm [[111](https://www.sciencedirect.com/science/article/pii/S0168952519302690?via=ihubm#bb0555)] to estimate ancestral relatedness.

Recently, Stern *et al*. introduced a clever method, called CLUES, that extends the approach of Coop and Griffiths to allow for recombination [[127](https://www.sciencedirect.com/science/article/pii/S0168952519302690?via=ihubm#bb0635)]. CLUES infers both the selection coefficient and historical allele frequency trajectory for a specific allele and nucleotide site of interest. This likelihood-based method uses ARGweaver to efficiently sample coalescent trees from the posterior distribution of ARGs at the locus of interest. It then uses a HMM to marginalize out the allele frequency trajectory conditioned on these trees. Because the ARGs sampled by ARGweaver reflect the assumption of selective neutrality, CLUES treats them as ‘proposals’ only and uses an importance sampling scheme to estimate the selection coefficient. Stern *et al*. estimated the selection coefficients at various pigmentation-associated variants and found evidence of adaptation, consistent with previous work.

## Summary and conclusion

### 

## References

<https://www.frontiersin.org/research-topics/8750/haplotype-analysis-applied-to-livestock-genomics>

2. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.