Simulating sweeps and adaptation of quantitative traits in evolution experiments using SLiM, part I

Changyi Xiao & Neda Barghi

Overview

Day1

- Detecting selection target
  - Drift
  - statistical methods (CMH-test Chisq-test)
- Reconstruct haplotype blocks
  - Linkage (correlated frequency change)
  - characterize selection (e.g. selection coefficient)

Day 2

- SLiM simulation
  - Model selection
- Model Fitting
  - connection between simulation and experiments
  - summary statistics (e.g. Jaccard index)

Overview

Day1

- Detecting selection target
  - Drift
  - statistical methods (CMH-test Chisq-test)
- Reconstruct haplotype blocks
  - Linkage (correlated frequency change)
  - characterize selection (e.g. selection coefficient)

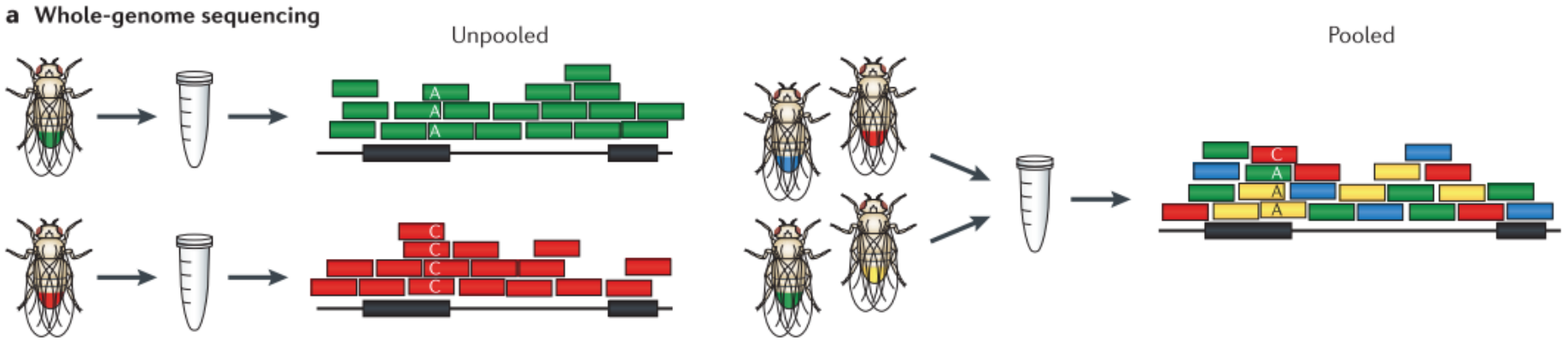
Day 2

- SLiM simulation
  - Model selection
- Model Fitting
  - connection between simulation and experiments
  - summary statistics (e.g. Jaccard index)

Day 1_1
Detecting selection target from empirical data

Poolseq data

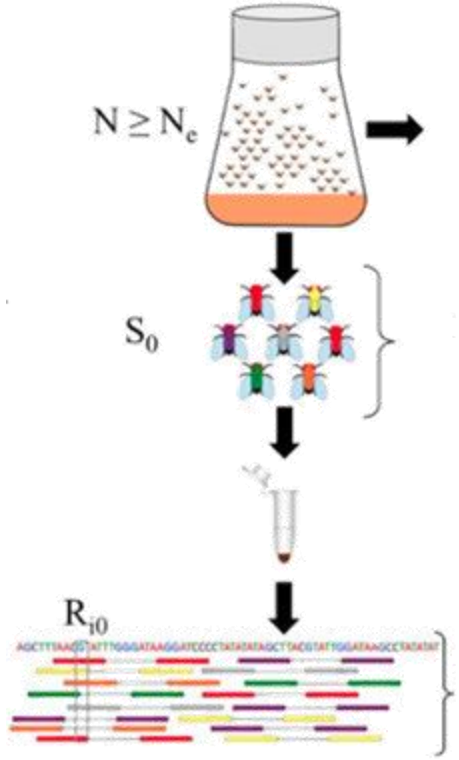
- Obtain information for markers at a population level
  - No linkage/haplotype blocks info!
- Sequence (NGS) multiple individuals together to get population data
  - Cheaper than individual sequencing



Schlötterer et al. 2014

Poolseq data

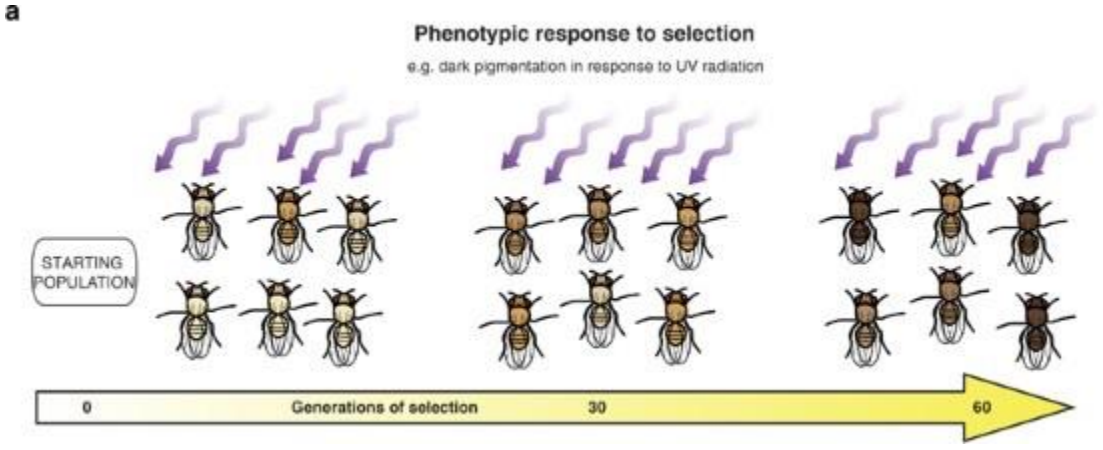
- Requires depth > 50x for reliable AF
  - disentangle low-frequency variants from sequencing errors
- Sampling needs to be appropriate according to census size
- Larger sample size is helpful in minimizing the sampling error



(Jonas et al, 2016)

Experimental evolution

- Experimental evolution is the study of evolutionary (phenotypic and genomic) changes occurring in experimental populations as a consequence of conditions (environmental, demographic, genetic, social) imposed by the experimenter.
- The opportunity to study evolutionary processes experimentally in real time.
- Mutation, genetic drift, and gene flow can act together with selection.



Teotónio et al. 2017

Kawecki et al. 2012

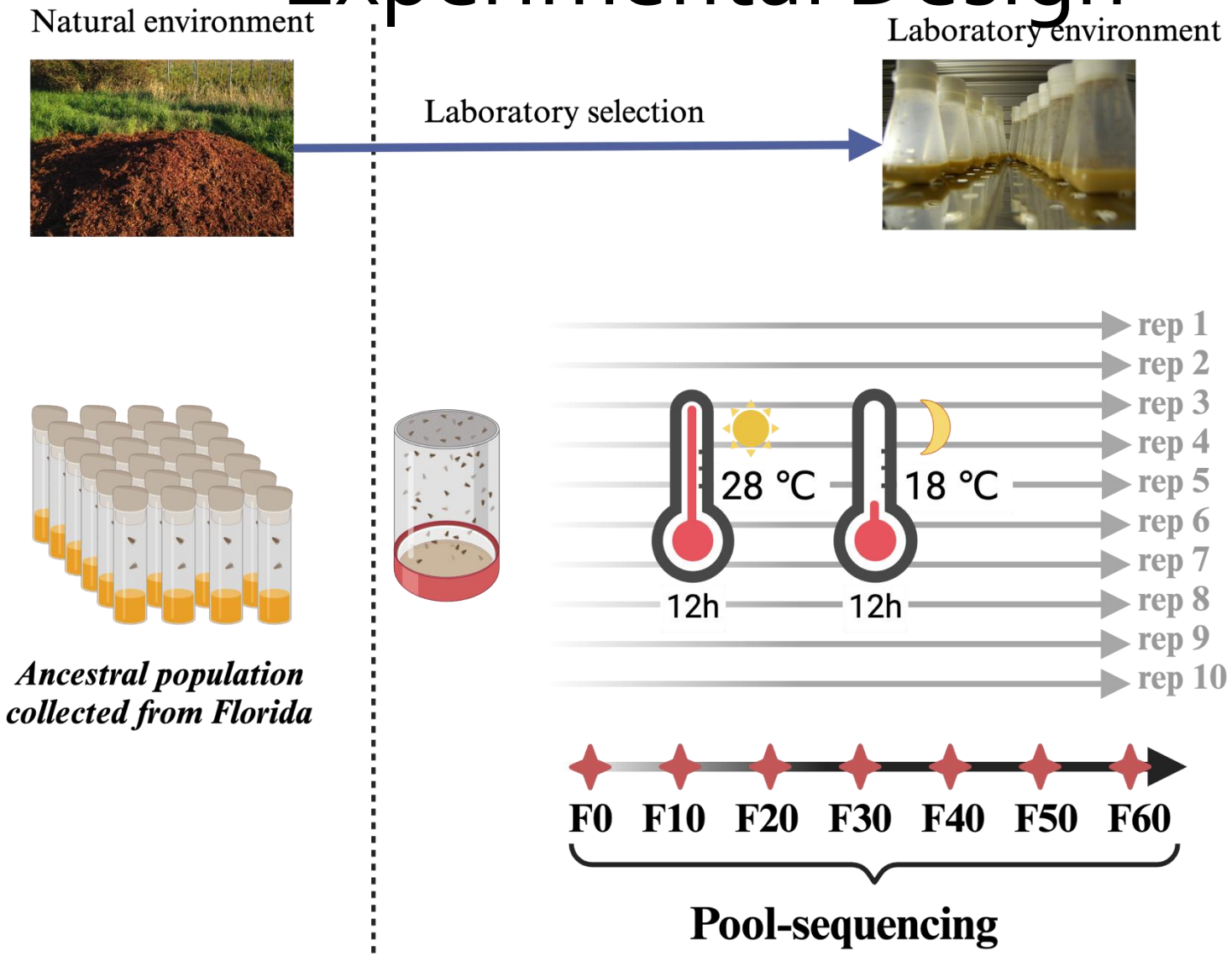
DNA data in EE

- Studying the genomic responses under a controlled environment (Experimental evolution)
- Discern (linked) selection from drift using theoretical models by exploiting the whole trajectory
- Focus on the dynamics of the selected alleles to learn about the adaptive scenario, to study clonal interference…
- Use of time series data

E&R framework

- Establishment of a reliable variants catalogue (standing genetic variation)
- Candidates underlying the latent selected trait can be identified as having consistent allele frequency changes across replicates in a controlled environment (parallel response)
  - But not always
- Contrast evolved and ancestral populations

Experimental Design



Protocol

- PCA
- Drift
- Chisq-test
- CMH-test
- Allele frequency correlation
- Selection coefficient

Quality control

- Make sure the SNPs you are tracking are reliable.

- Good alignment:
  - mapping quality (read)
  - base quality (position)
  - proper paired (read)

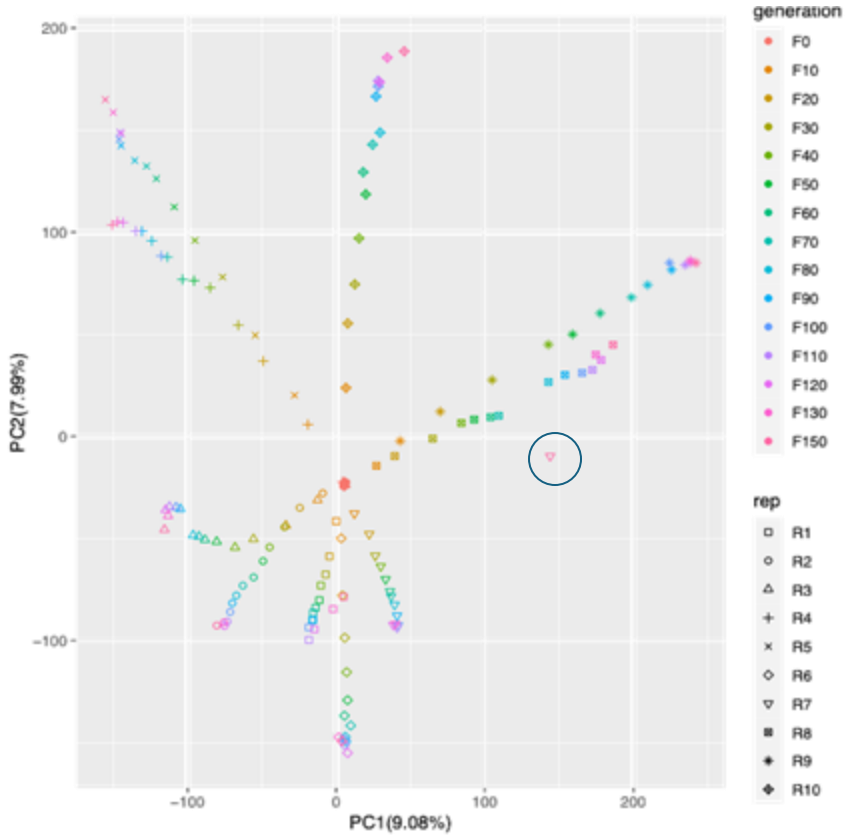
- Reliable AF estimation:
  - Coverage
  - Low frequency variants

$$AF = \frac{AD}{COV}$$

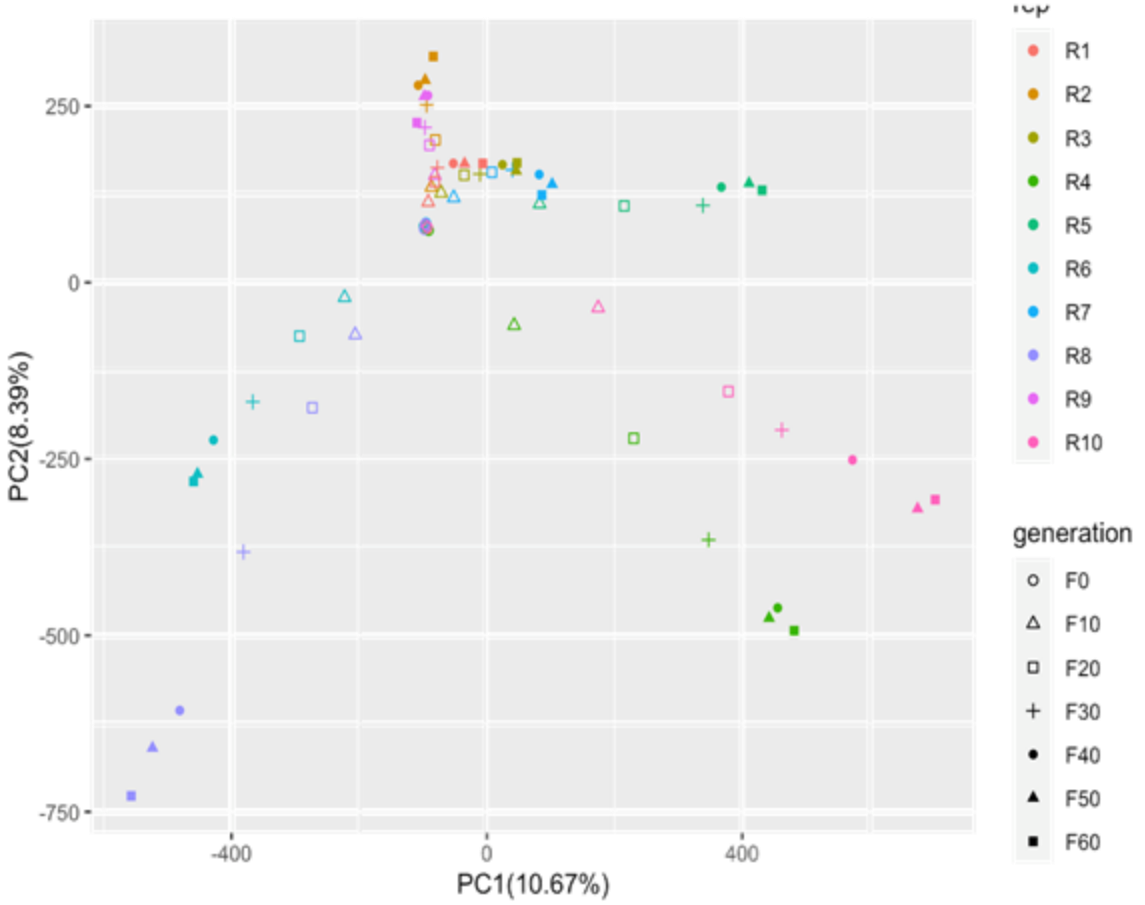
Quality control: PCA

- Check the clustering of your samples with a PCA on transformed frequencies
- Does the samples' name match?
- Is there any contamination in the sample?
- Is there other unexpected pattern?

- We transform the frequencies to normalize AFC: SNPs with AF closer to 0 or 1 are expected to change less



Hands on: QC



What is your expectation in the PCA results?

Allele frequency can change other than selection-- Genetic Drift

- Changes in allele frequencies due to stochastic (independent of external and heritable factors) sampling variation (in offspring number) in a finite population
- Several models have been proposed. The most commonly used models are the Wright-Fisher model and the Moran model.

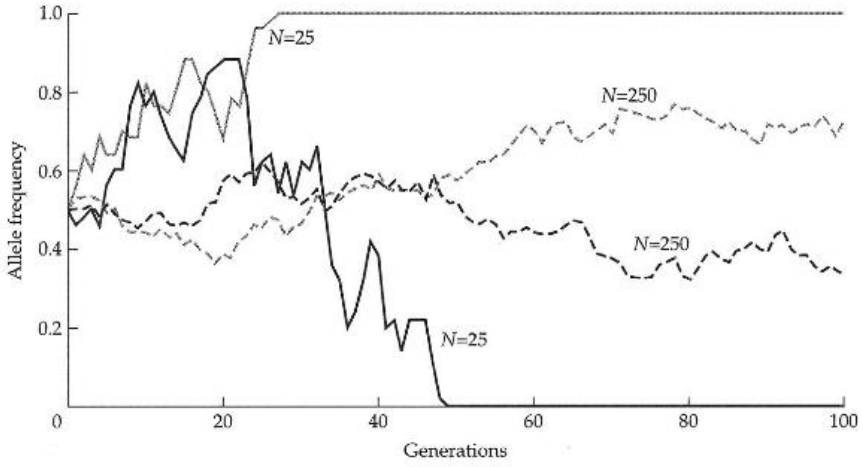
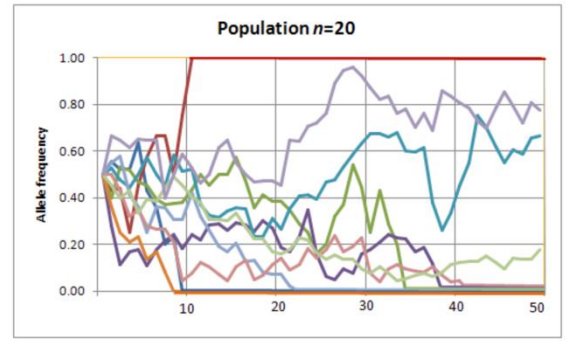


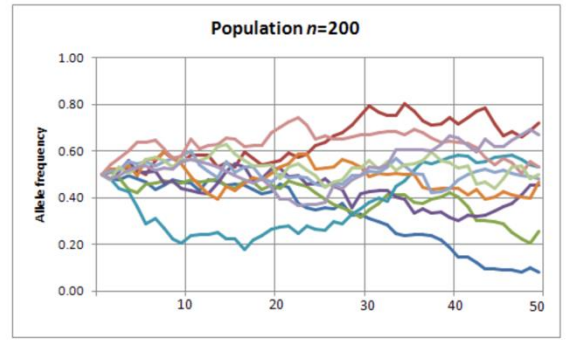
Figure 2.4 Changes in frequencies of alleles subject to random genetic drift in populations of different sizes (N). In each generation, $2N$ genes were sampled with replacement from the previous generation. For each population size, two replicates are presented. It is assumed that the effective population size N_e is equal to the actual size N.

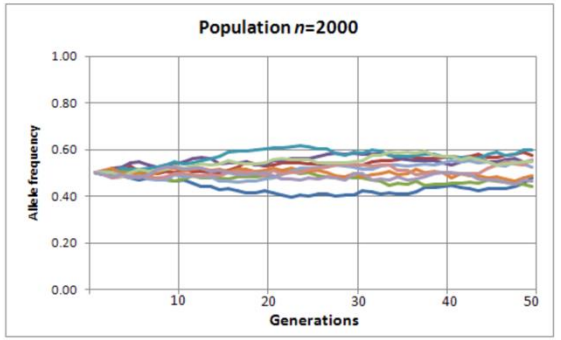
Bodmer and Cavalli-Sforza 1976

Estimating Ne—accounting for drift

- limited population size in E&R studies
- allele frequency (genetic components):
  - directional change by selection
  - random change by drift
- the larger Ne, the smaller AFC by drift
- estimate Ne in E&R studies to account for drift







```
##  [1] 434 298 285 254 240 271 264 165 359 189
```

Hands on: Ne estimation

What is the Ne for each replicate?

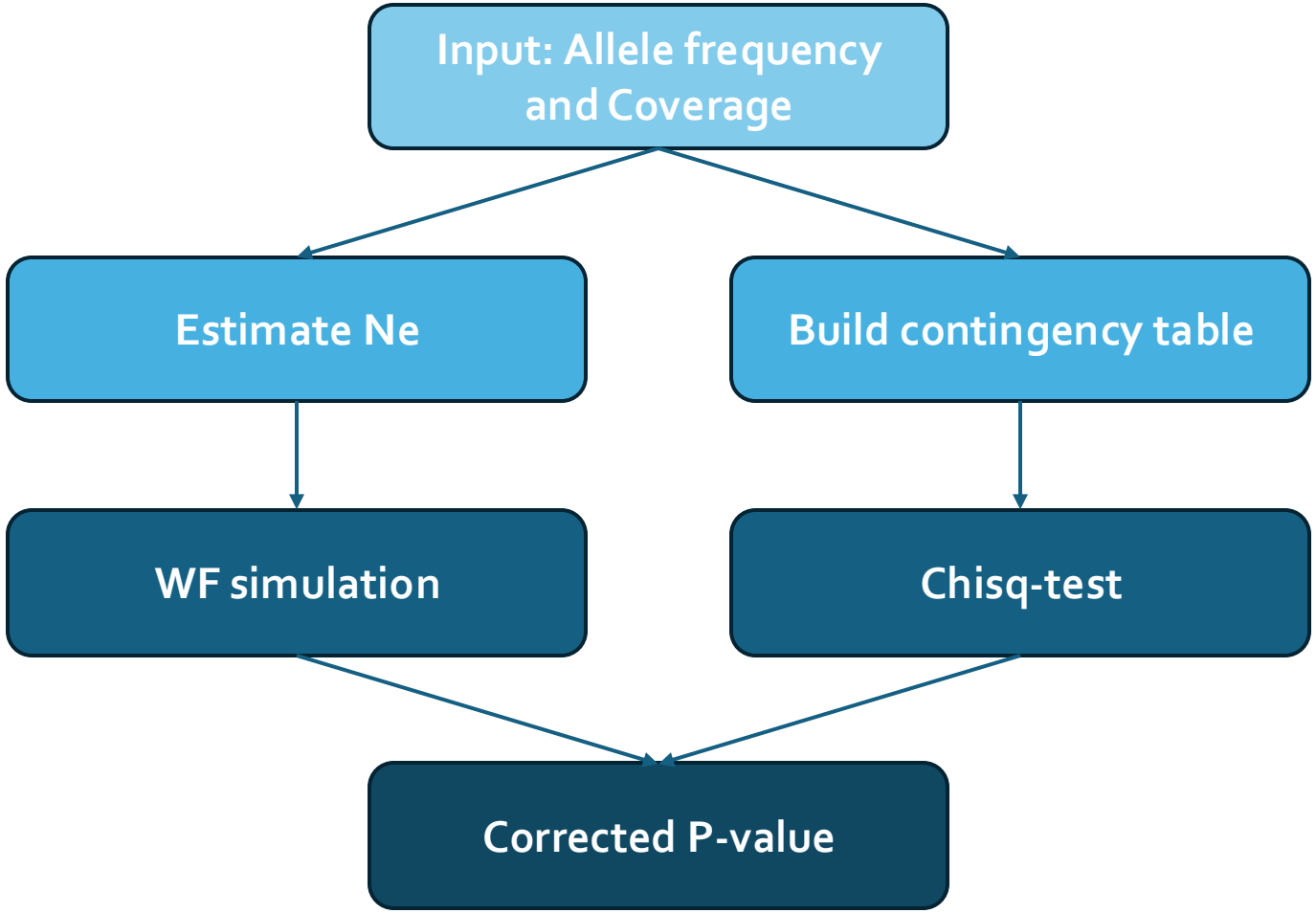
Keep in mind: autosome and X chromosome

Detecting significant allele frequency change

- beneficial allele will be kept due to selection and increase its AF
- deleterious allele will be purged out and decrease AF
- neutral sites will not respond to selection:
  - how will these sites respond in a single replicate? Chisq-test
  - how about across replicates? CMH-test

- In E&R studies:
  - AFC between evolved and ancestral populations larger than by drift alone

Chisq-test in detecting selection



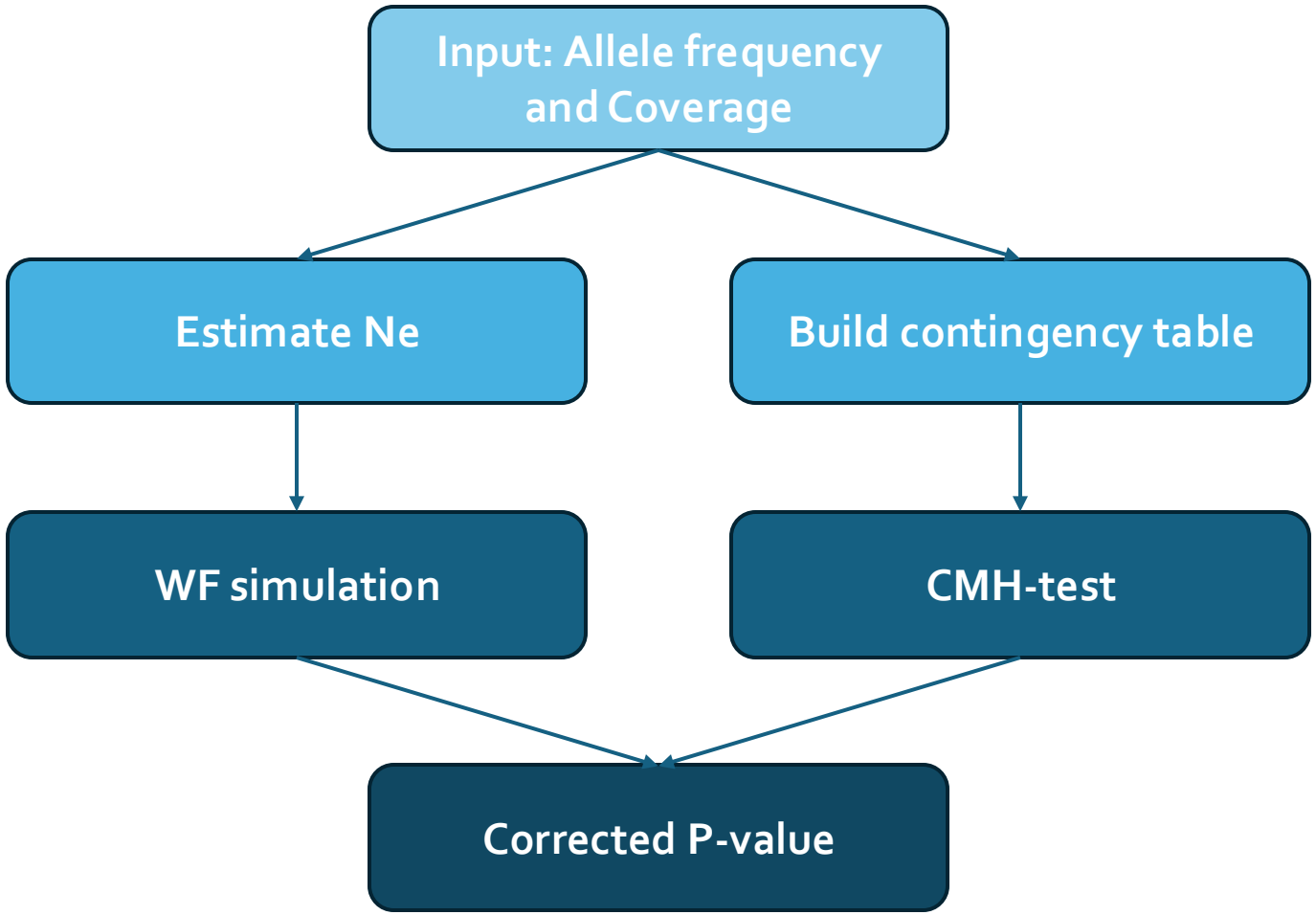
Per SNP

	Reference allele	Alternative Allele
Ancestral	2	48
Evolved	13	47

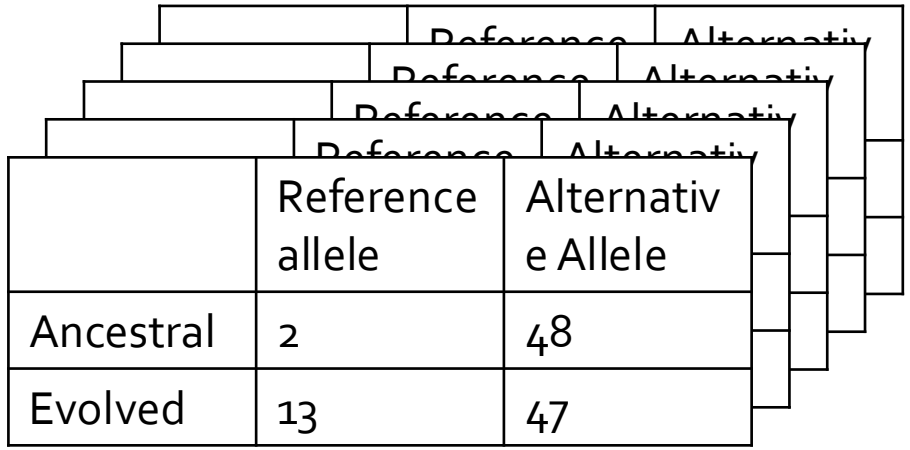
#allele count

Replicates-specific
SNPs with significant allele frequency change

CMH-test in detecting selection



K contingency tables

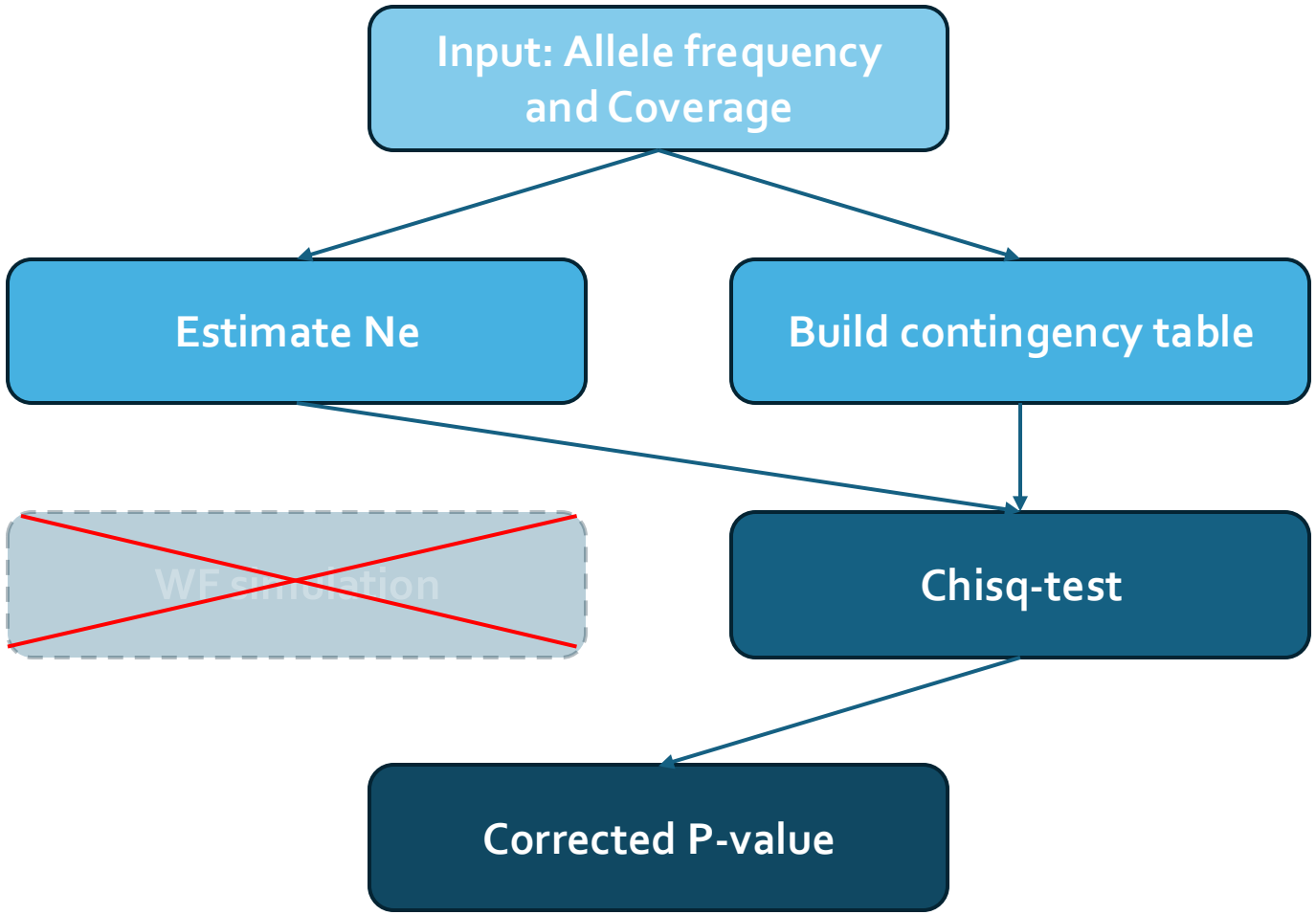


	Reference allele	Alternative Allele
Ancestral	2	48
Evolved	13	47

#allele count

Replicates-specific
SNPs with significant allele frequency change/change parallelly across replicates

Adapted Chisq-test in detecting selection



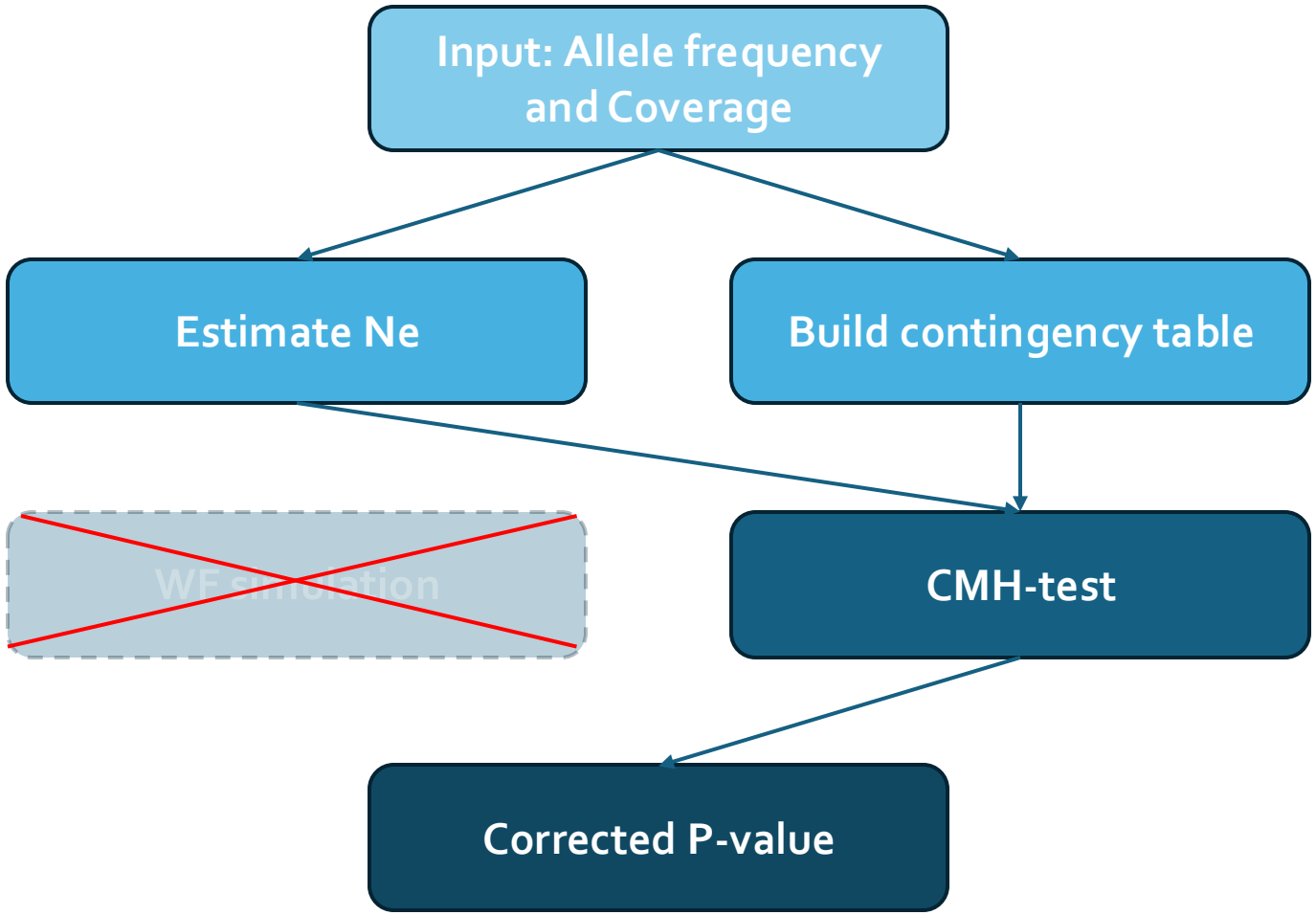
Per SNP

	Reference allele	Alternative Allele
Ancestral	2	48
Evolved	13	47

#allele count

Replicates-specific
SNPs with significant allele frequency change

Adapted CMH-test in detecting selection



K contingency tables

	Reference allele	Alternative Allele
Ancestral	2	48
Evolved	13	47

#allele count

Replicates-specific
SNPs with significant allele frequency change/change parallelly across replicates

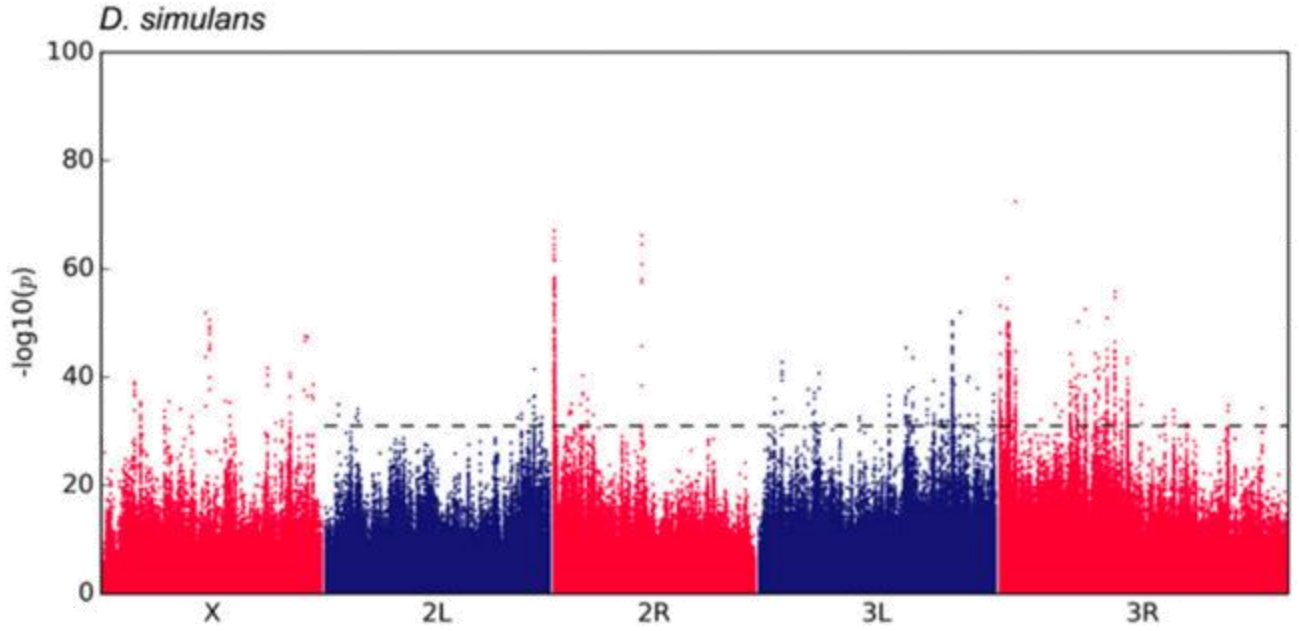
The ACER package

- Classic CMH-test and classic Chisq-test:
  - Don't consider drift, sampling, and pool-seq
  - Need to generate expected AF based on the Ne and starting AF using neutral simulations (drift)
  - Can only test two time points

- ACER:
  - Modified version of CMH-test and chisq-test
  - Takes drift, sampling, and pool-seq into account – more efficient (faster)
  - Allows including intermediate time points (if available)

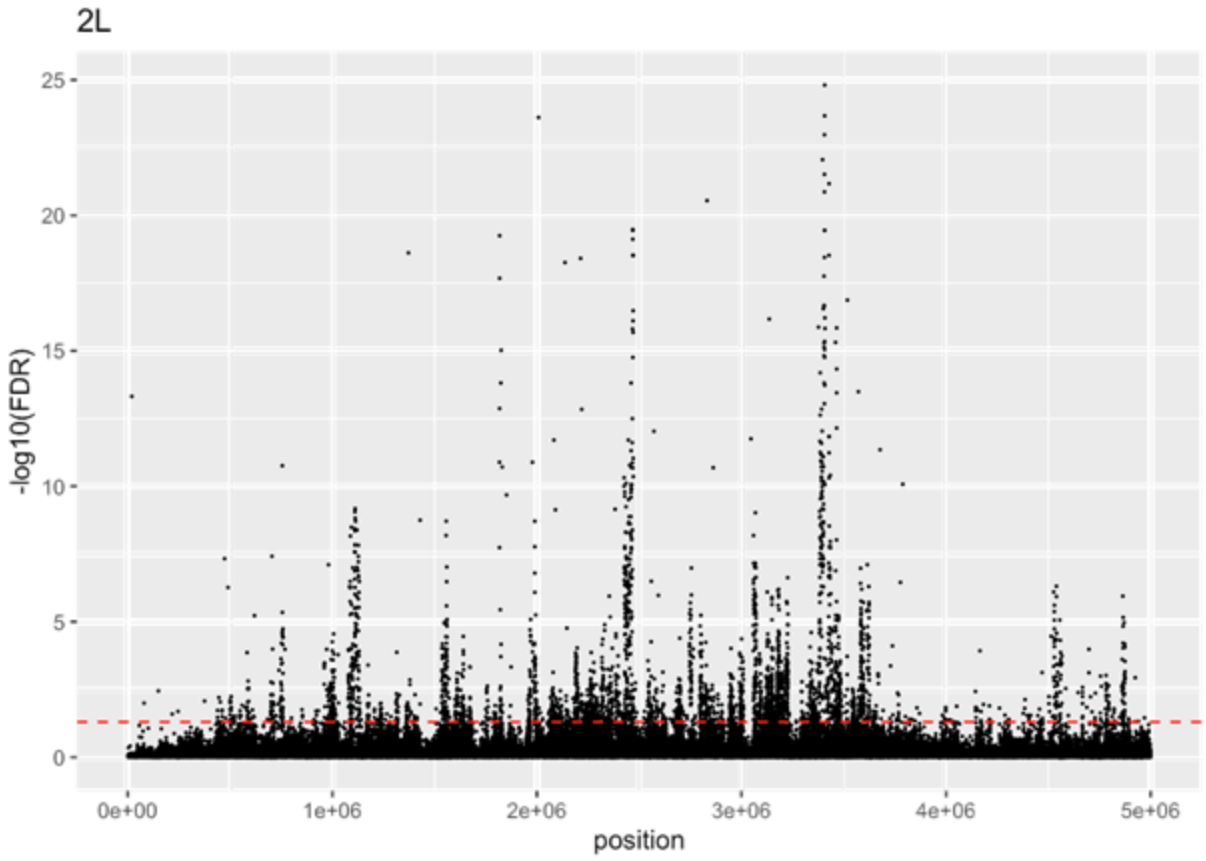
How to detect biological outliers?

- Manhattan plot



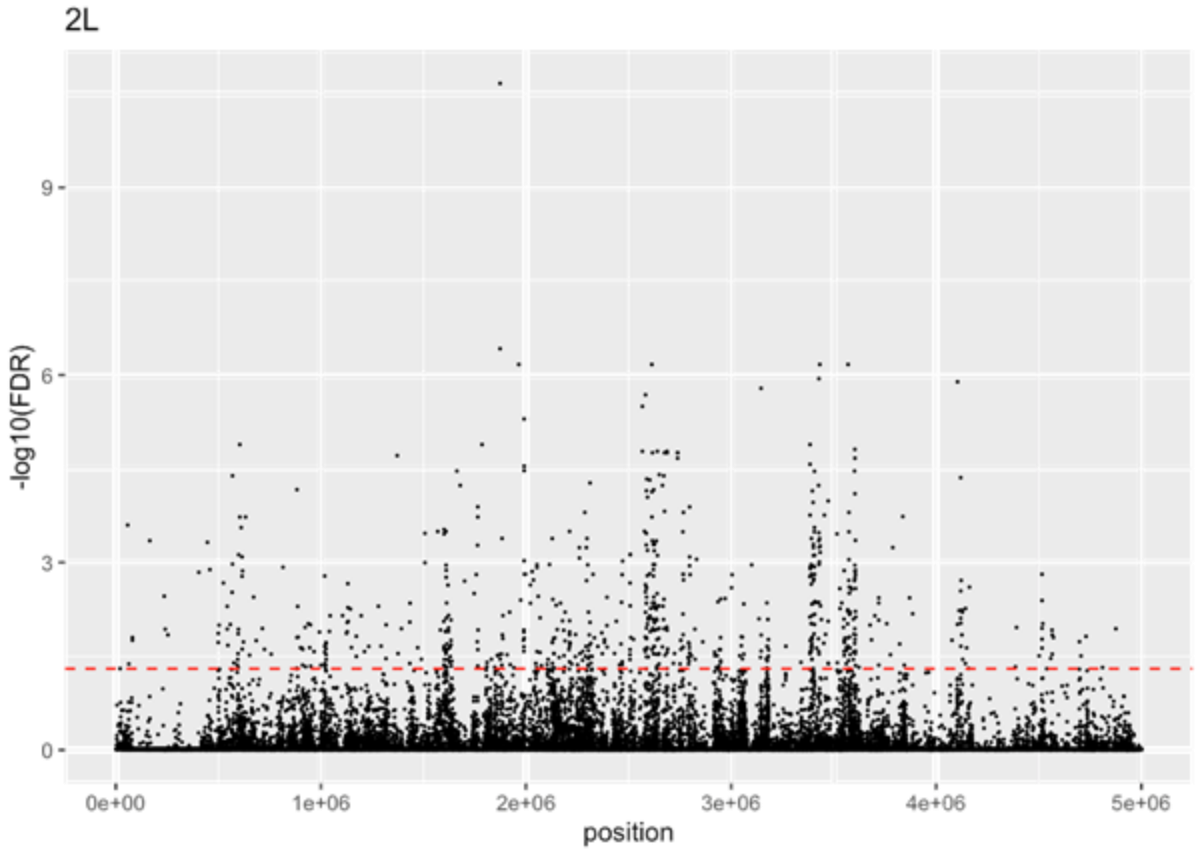
Neda Barghi, Raymond Tobler, Viola Nolte, Christian Schlötterer, Drosophila simulans: A Species with Improved Resolution in Evolve and Resequence Studies, G3 Genes|Genomes|Genetics, Volume 7, Issue 7, 1 July 2017, Pages 2337–2343, https://doi.org/10.1534/g3.117.043349

Hands on: adapted CMH



Without intermediate timepoints : 4.1

Hands on: adapted X2



Without intermediate timepoints – 5.1

Including intermediate time points

- Usually results in a more conservative testing (less significance)
- Available methods: ACER-$\chi 2$, ACER-CMH, GLM, LM, BBGP, WFABC
- Code available at sections 4.2 (cmh) and 5.2 (X2)

Recap

- estimate Ne to account for drift
- adapted-CMH to detect consistent AFC across replicates (parallel)
- adapted X2 to detect significant AFC in single replicate
- more conservative when considering intermediate time points
- FDR is important!

Thanks!

Any questions?

What happens with linkage?

Day1_2
Haplotype reconstruction

Changyi Xiao & Neda Barghi

Overview

Day1

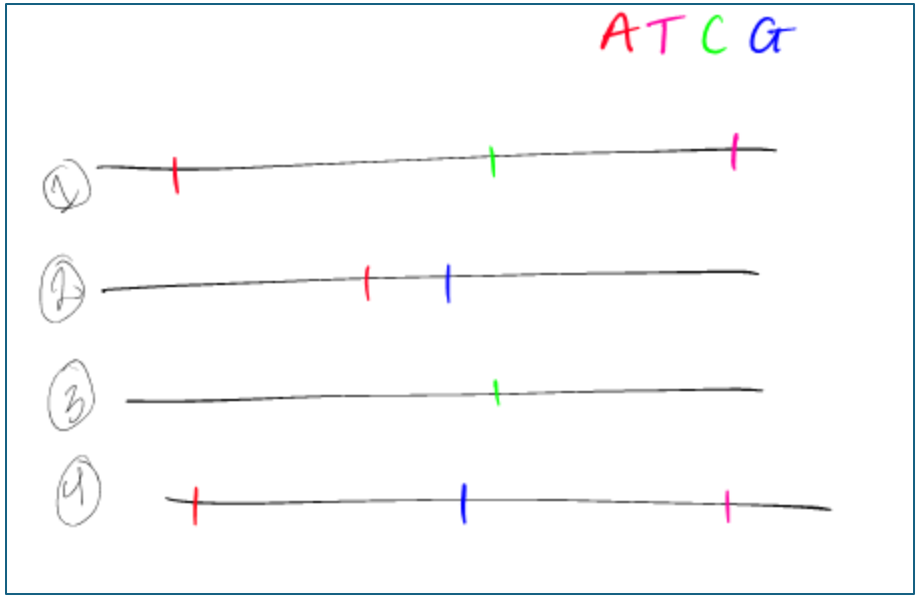
- Detecting selection target
  - Drift
  - statistical methods (CMH-test Chisq-test)
- Reconstruct haplotype blocks
  - Linkage (correlated frequency change)
  - characterize selection (e.g. selection coefficient)

Day 2

- SLiM simulation
  - Model selection
- Model Fitting
  - connection between simulation and experiments
  - summary statistics (e.g. Jaccard index)

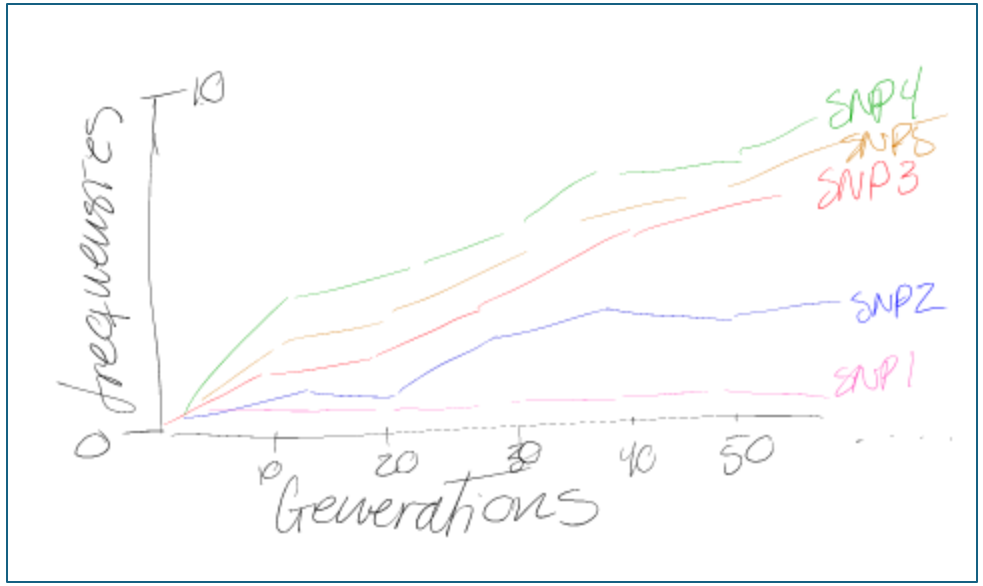
Haplotypes

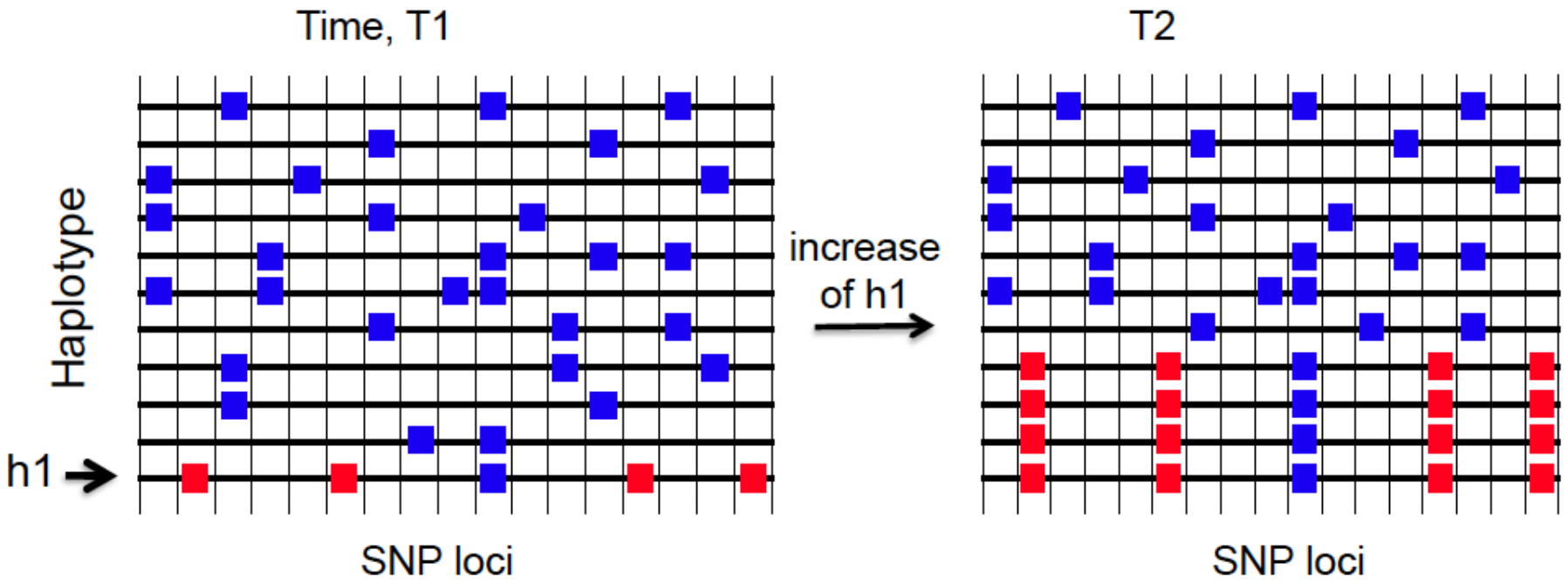
- Composition of genetic variants (SNPs) which are inherited together on a stretch of DNA
  - Due to linkage disequilibrium (LD)
  - Can be large stretches which depend on the LD landscape



Why would we care about haplotypes?

- Haplotype response depends on the joint effects of the underlying SNPs (non-independent)
- Narrow down the selection target
  - Many responding dependent SNPs
- Quantify the number of selection targets
  - Characterization of the selection response




Time, T1
T2
Haplotype
h1
increase
of h1
SNP loci
SNP loci

In the absence of haplotype data …

- The resolution of LD information in Pool-seq is only up to the read length
  - Much shorter than actual haplotypes and targets of selection

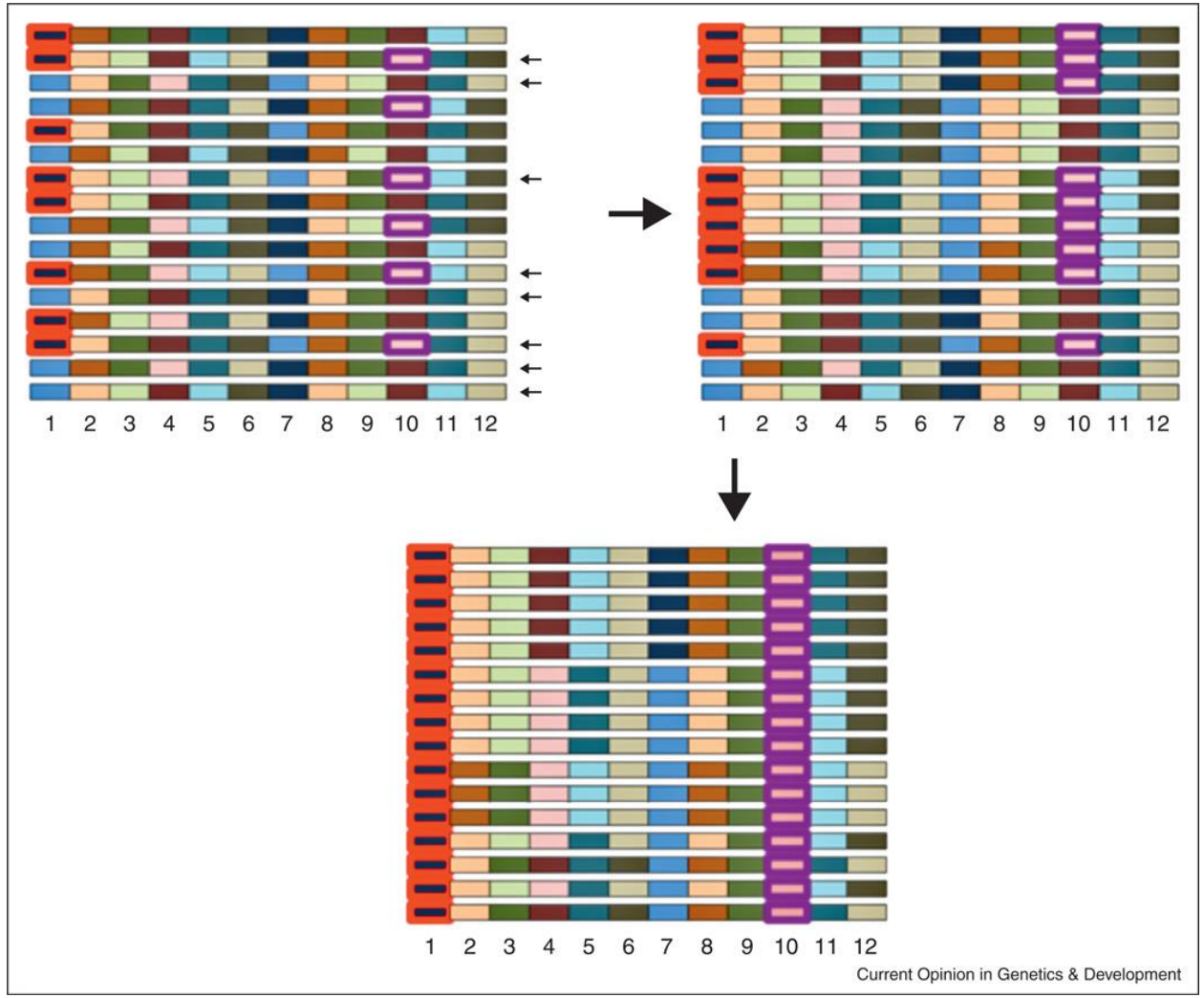
- How could we maneuver this problem?

… and presence of time series pool-sed data

- Time series of pooled individuals allows us to track the frequency of SNPs across generation
  - SNPs located on the same haplotypes tends to have correlated frequency change

- → This information can be used to reconstruct haplotypes

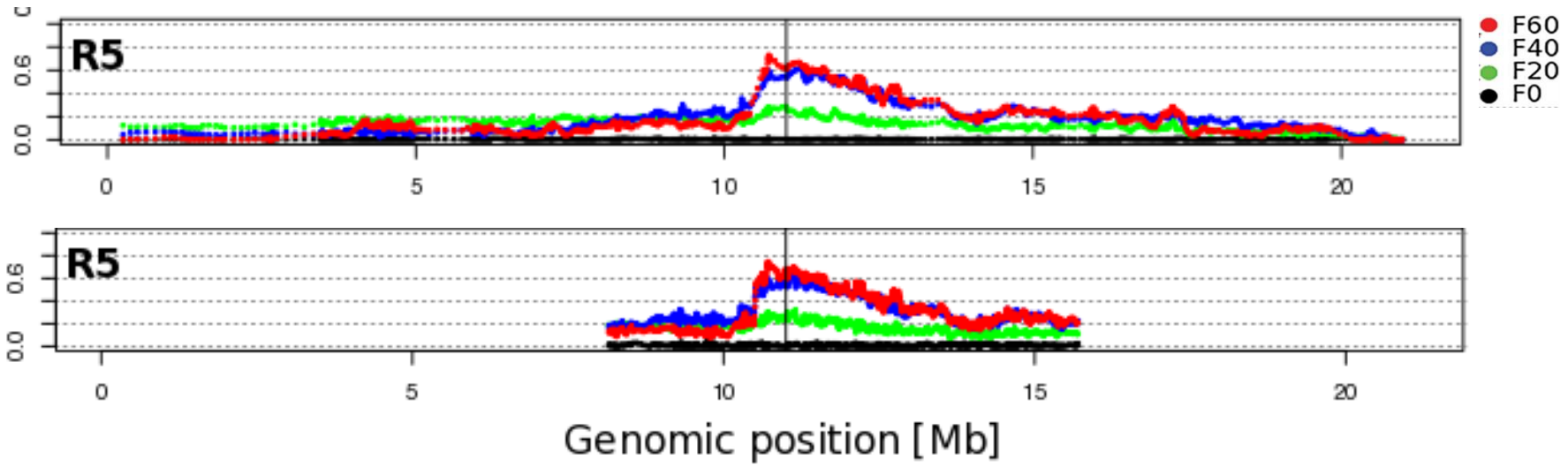
The effect of long-range LD on linked selection



Turner and Nuzhdin 2013

Reconstruction of selected haplotype blocks from Pool-Seq

- Frequency trajectories of linked (selected and hitchhiking) SNPs are correlated across time and replicates.

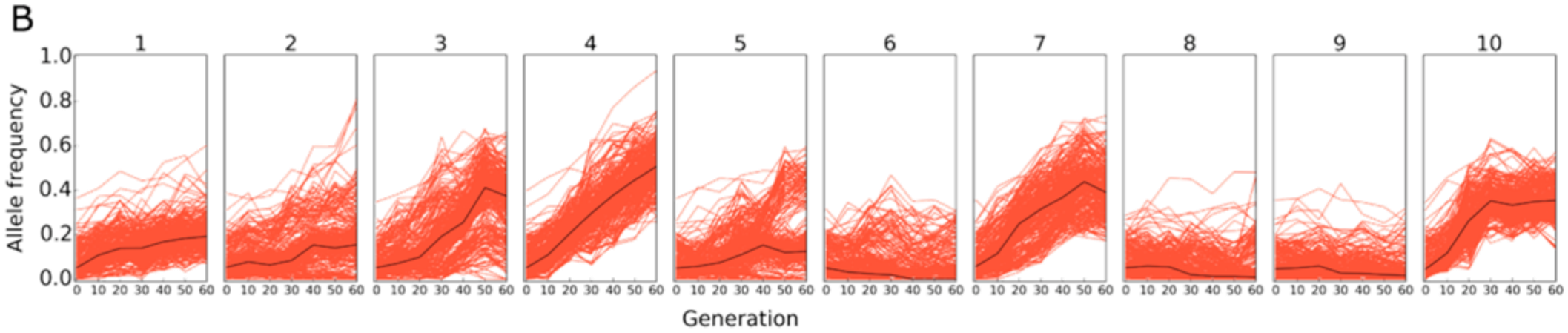


Franssen et al. 2016

Plotting allele frequency trajectories

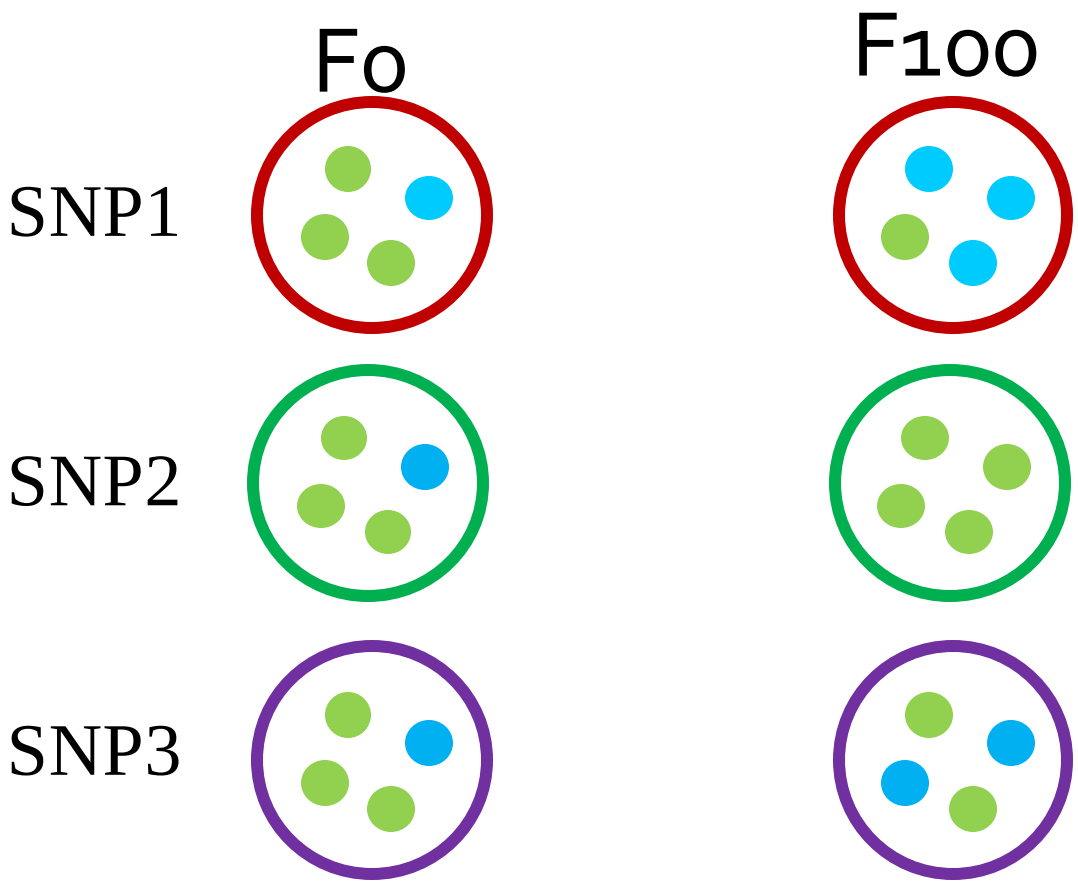
Making sure that alleles from the same haplotype backgrounds during clustering

- Importance of polarization （ref/alt, rising, minor）
  - hitchhiking

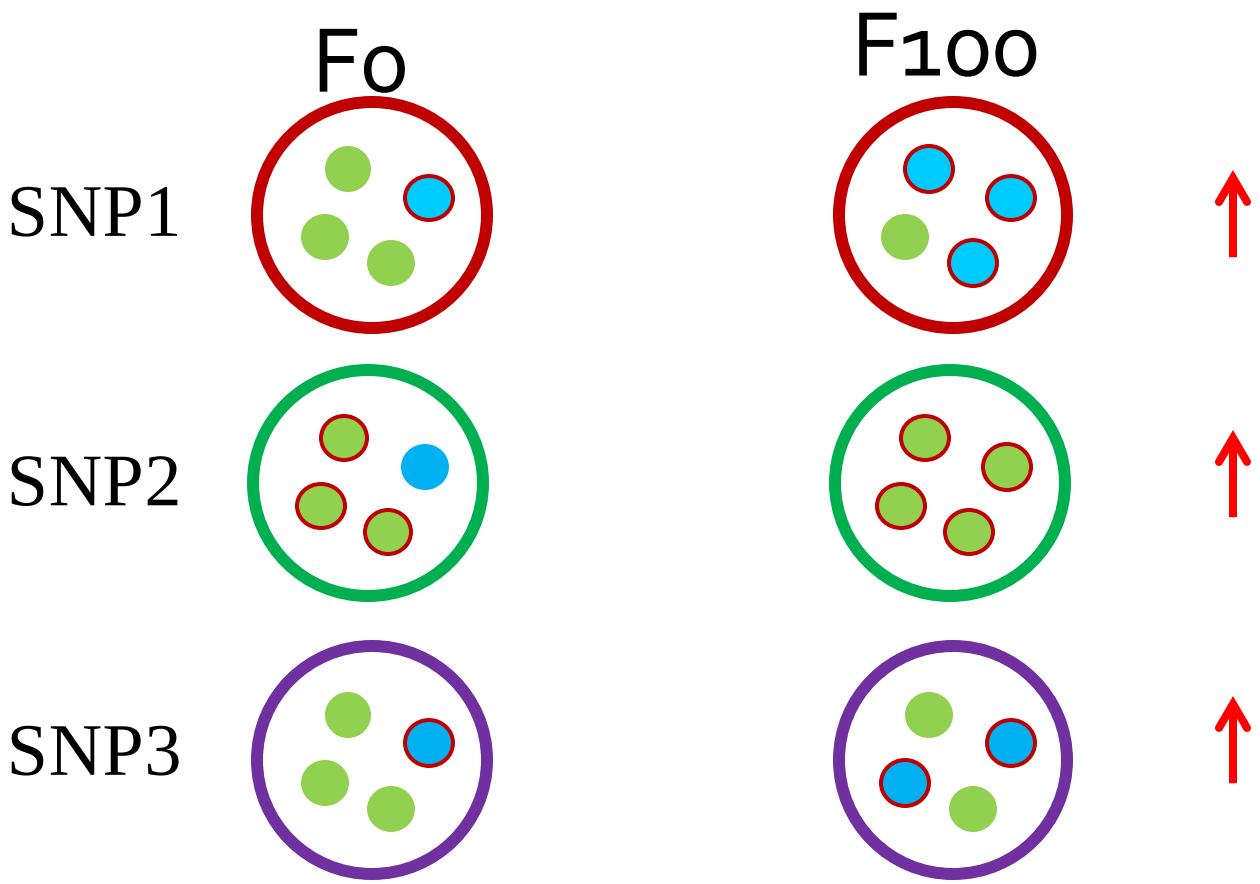


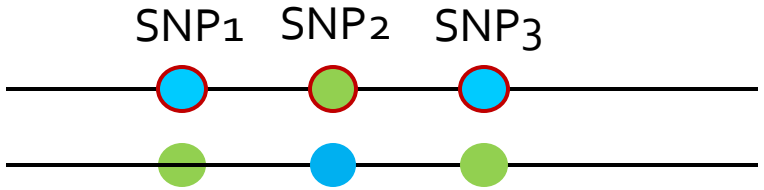
Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. Nat Rev Genet 21, 769–781 (2020). https://doi.org/10.1038/s41576-020-0250-z

Polarization: determine the allele of interest



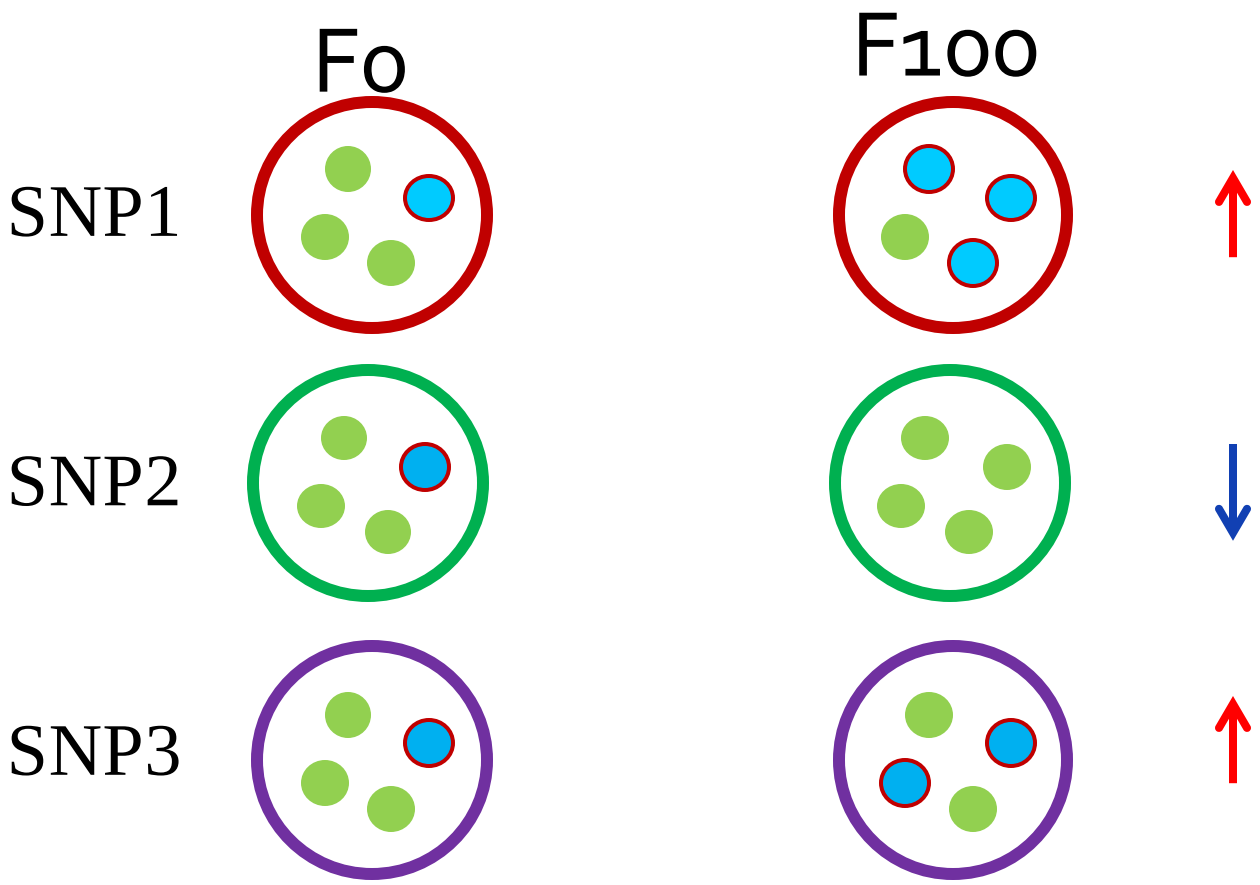
Polarization 1: Rising Allele

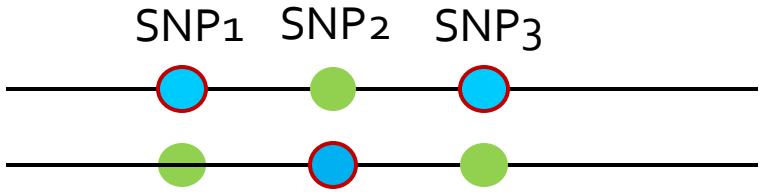




SNPs that located on the same haplotypes tends to respond consistently

Polarization 2: Minor allele





Introduction to haplotype reconstruction

- HaploReconstruct (Franssen et.al., 2017)
  - Reconstructs selected haplotypes from replicated time series using pool-seq data
  - identification of selected haplotypes through correlated frequencies of alleles
  - No individual sequencing of founder chromosomes are needed

http://cran.nexr.com/web/packages/haploReconstruct/index.html

Experimental design

- Simulated data of *Drosophila simulans*
  - 10 replicates
  - over 60 generations (data every 10th generation)
  - Chromosome 2 & 3

Exercice 0

- How many targets of selection have there been simulated per chromosome?
- Is there a relationship between starting allele frequency and selection coefficient?

Exercice 1

- Run the clustering and explore the output
- What is the difference between stringent and relaxed clustering?
- What could min.cl.cor mean?
  - Franssen et.al. (2017), Mol Bio Evol, table 1
- How many blocks are reconstructed with min.cl.cor = 0.6/0.2
- How is the number of marker SNPs affected by min.cl.cor?
- How is the haplotype block length affected min.cl.cor?

Exercice 2

- How many of your targets are assigned to a haplotype block with min.cl.cor = 0.6/0.2
- Compare to the number of reconstructed haplotype blocks. (min.cl.cor = 0.6/0.2)
  - How can you explain these results?
- For the min.cl.cor = 0.6 results: Investigate the frequency and s of detected vs. missed targets.
  - Do you observe any interesting patterns? What does this mean?

Exercice 3

- Change the window size parameter from 3 Mb to 0.5 Mb and 10 Mb (min.cl.cor = 0.2)
- Call perform_clustering() with the changed window size
  - l.win.size
- How does window size affect your haplotype reconstruction?
- Repeat with min.cl.cor = 0.6, what do you observe here?

Discussion

- How do you pick a “good” window size?
- How do we pick a “good” minimum correlation?

Problems encountered

WINDOW SIZE

- too big: less resolution, more computationally expensive
- too small: restrict haplotype length, but perhaps not enough SNPs

CORRELATION

- too stringent correlation: Too many blocks, blocks are artificially split, regions of selection are over-estimated.
- too relaxed correlation: Too few blocks, blocks are artificially merged, regions of selection are under-estimated

Importance of the parameters

Parameters that affect the clustering

- Filtering of snp (SNP density)
- Minimum correlation
- Windows size

Parameters for the clustering

- Allele frequencies
- Number of replicates
- Generation information
- Selected snps
- window size
- minimum correlation

Window size

- Fit window size to data-set:
  - number of SNPs
  - cmh-score (proxy for SNPs effect size)
- Approach: (within chromosome)
  - same fraction of total candidate SNP effect size per window (to make haplotype reconstruction comparable)
  - Enough SNPs to reconstruct a block

median normalized CMH score sum (MNCS) of 1%

$$MNCS = \text{median}\left(\left(\sum -\log(p)_{\mathbf{window}}\right)/\left(\sum -\log(p)_{total}\right)\right)$$

Minimum correlation

- How are we sure our haplotypes are properly correlated?
  - Stringent correlation & too relaxed correlation will produce incorrect blocks
- Where to stop?
  - Correlation of blocks we wouldn't expect to correlate (unliked blocks)
- Haplotype blocks from different chromosomes behave independently (No linkage)
  - need of different chromosomes to compare!

Analyze the correlation between blocks (focal vs background)!

Use correlation between background blocks to establish a minimum.

Haplovalidate approach overview

Define window size	Small stringer blocks	Extend the blocks
According to our data • number of SNP • p-values	Given the window size, reconstruct haplotypes with high allele frequency trajectory	If correlation with blocks close by is bigger than with blocks in other chromosomes, extend

Haplovalidate: auto-optimize the parameters for the clustering

Otte KA, Schlötterer C. Detecting selected haplotype blocks in evolve and resequence experiments. Mol Ecol Resour. 2021 Jan;21(1):93-109. doi: 10.1111/1755-0998.13244. Epub 2020 Sep 6. PMID: 32810339; PMCID: PMC7754423.

Discussion

- Some caveats
  - For selected haplotypes
  - Parameter sensitive: mainly on SNP density (currently working on it)
  - Assumption of parallel trajectories = same haplotype
  - Multiple datasets

How to plot haplotype trajectories

- Haplotype information: chromosome, starting position, end position, contributing snps
- Get allele trajectories from all SNPs from the haplotype block
- Estimate median AF for each time point
- Plot median AF
- (Optional) Plot a line for each SNP trajectory on the background

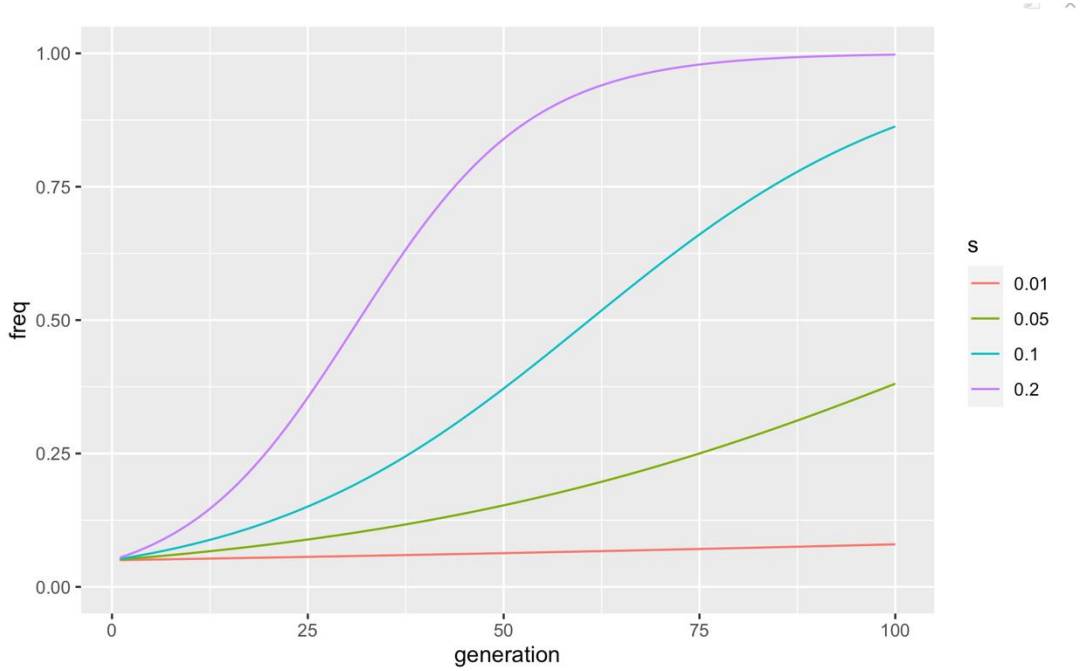
HANDS ON: plotting haplotype trajectories

How would you plot the trajectories of your haplotype blocks?

Characterizing the haplotype blocks

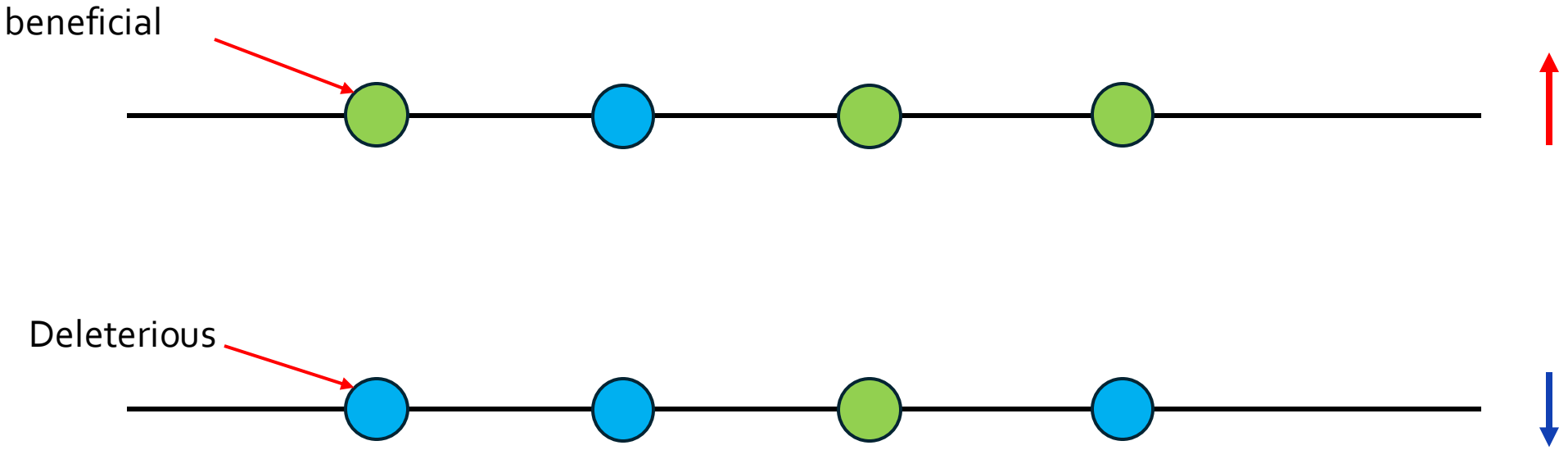
- How strong the selective response of each haplotype blocks?
  - Estimate selection coefficient based on the Allele frequency change.

The allele frequency change
determined by
Initial frequency
Selection coefficient



Allele frequency change of haplotypes

- Determined by the joint effects of contributing loci that located in the haplotype
- Selection acting on the individual level



Estimate the selection coefficient of the haplotypes

- Haplotype information: chromosome, starting position, end position, contributing snps Don’t forget Polarizing!!!
- Get allele trajectories from all SNPs from the haplotype block
- Estimate median AF for each time point
- Estimate the selection coefficient from median AF change

Thanks for the attention!!

Thanks for the attention!