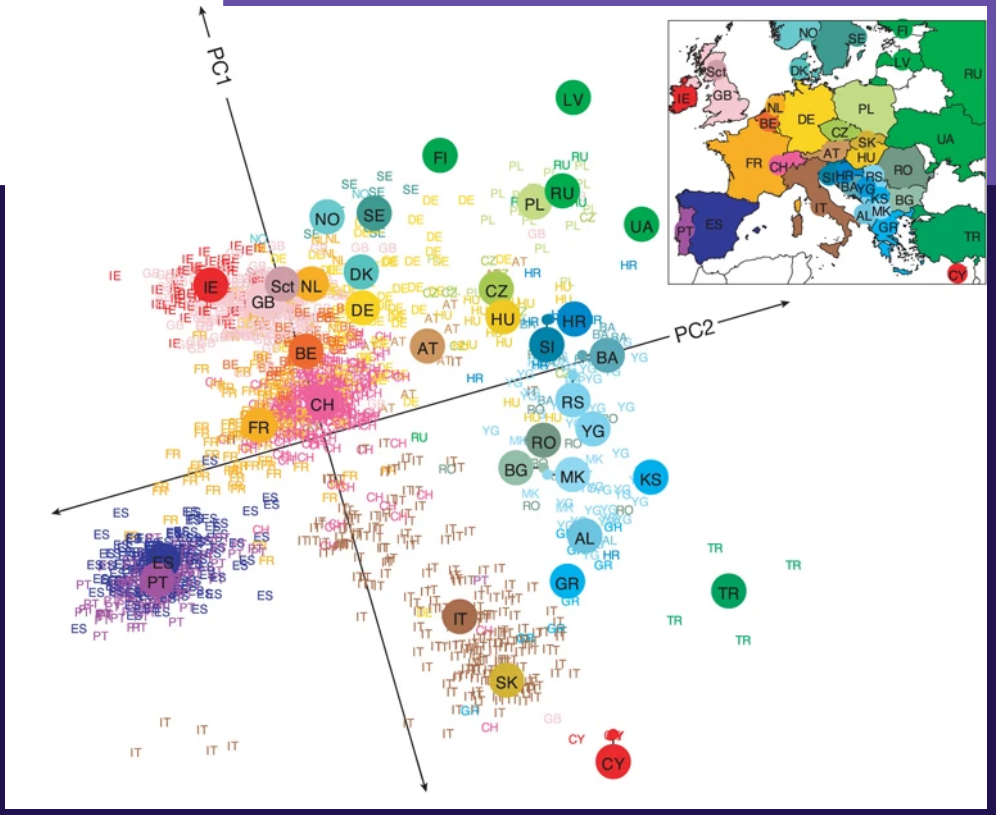
Physalia

Courses

Genome scans

29/05/2025



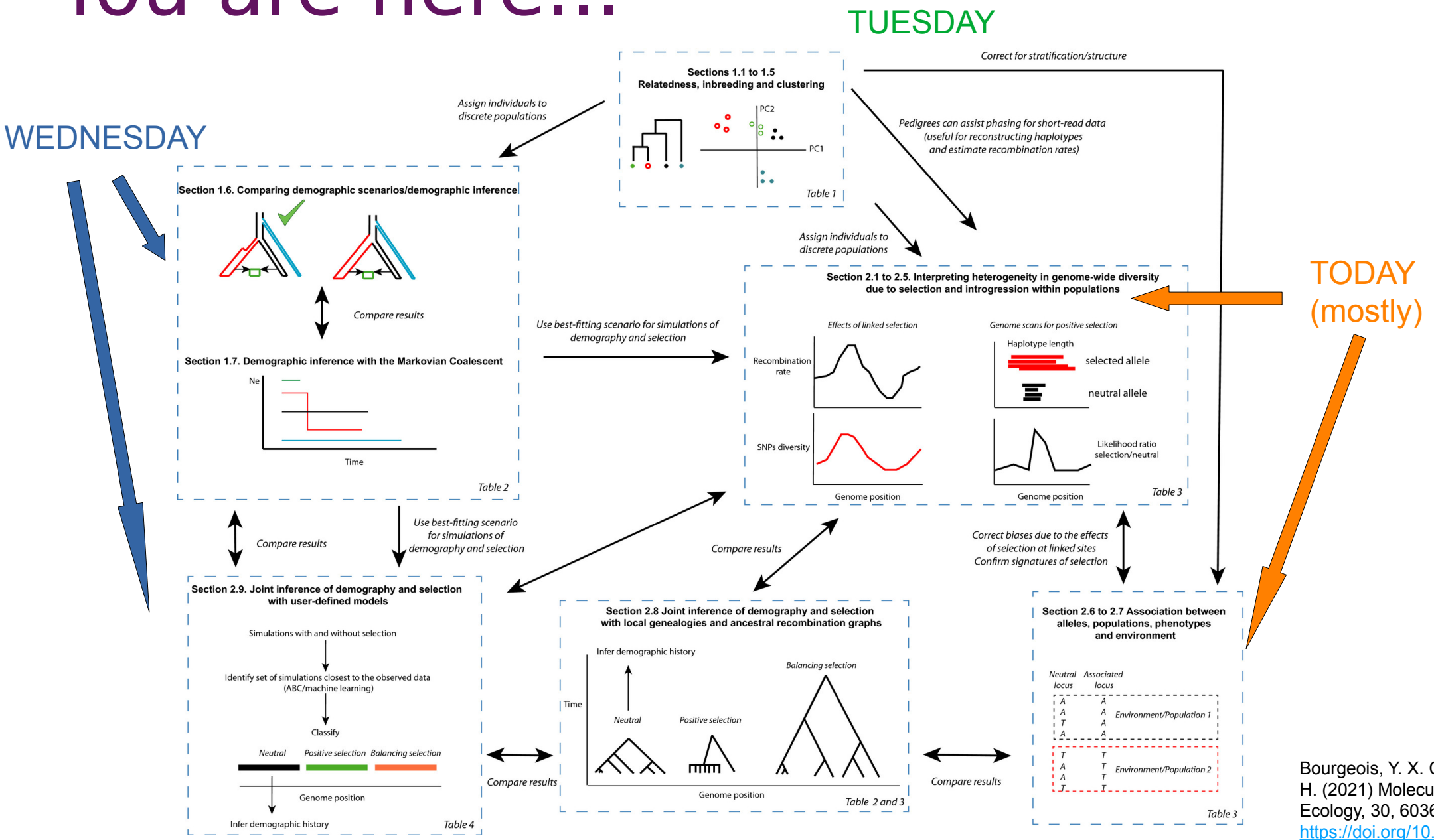
Physalia course

Yann Bourgeois, Thibault Leroy

Goals for today's lecture

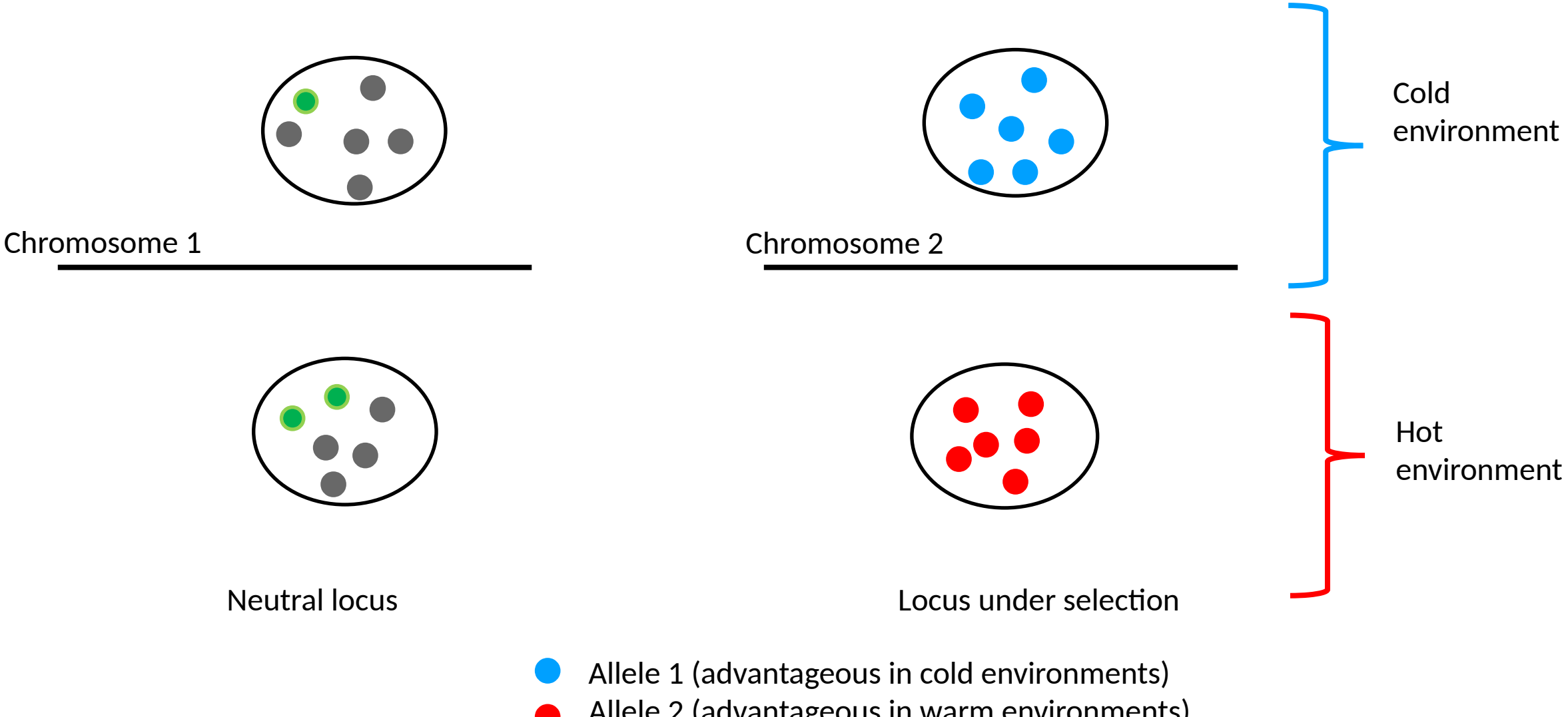
- What is a genome-wide association study (GWAS)?
- How can we detect selection in a single population?
- How do we detect divergent selection?
- A few methods to detect balancing selection.
- The perils of interference and linkage.
- As usual: a few limitations to keep in mind.

You are here...



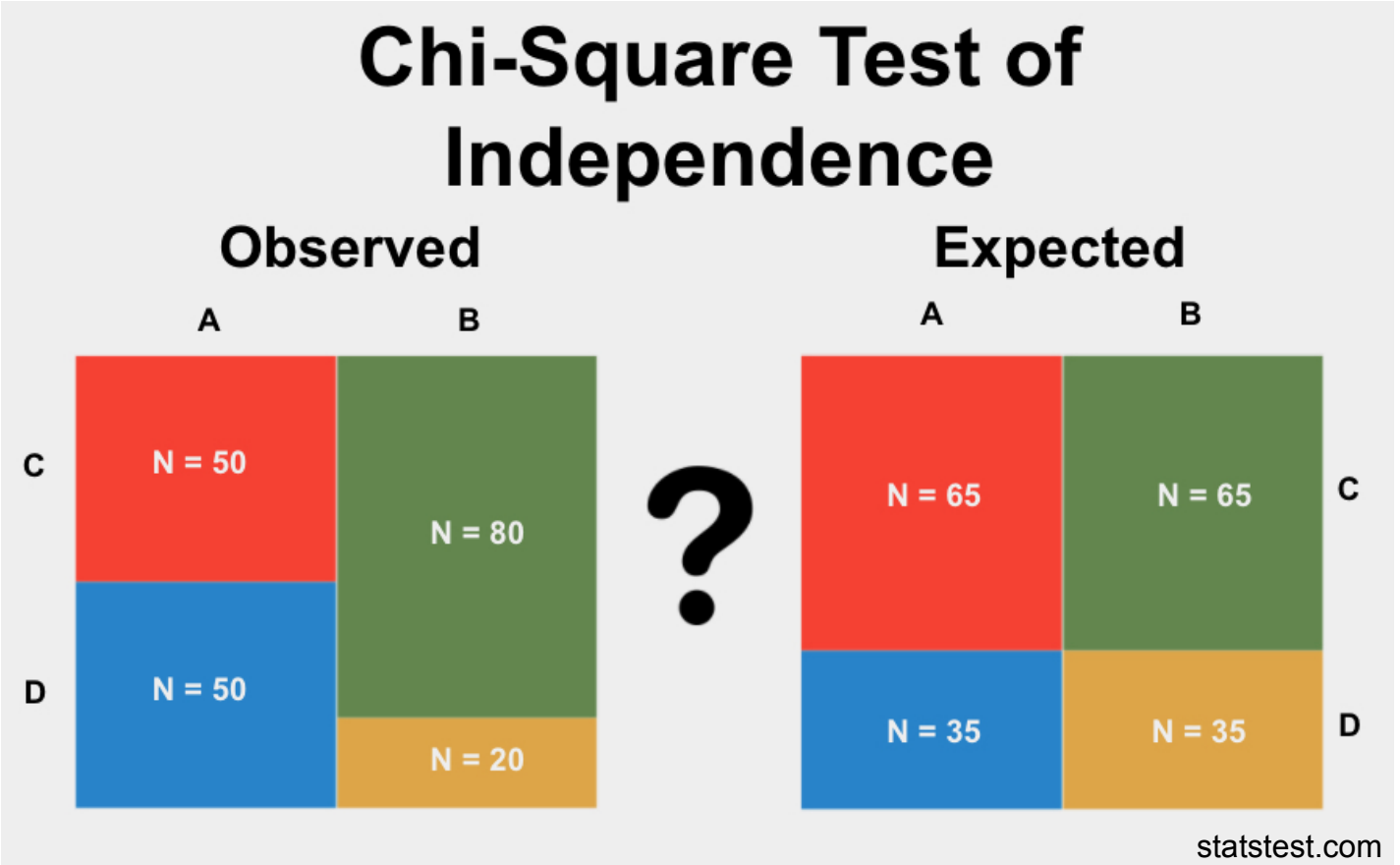
Bourgeois, Y. X. C., & Warren, B. H. (2021) Molecular Ecology, 30, 6036–6071. https://doi.org/10.1111/mec.15989

Genome scans for association/adaptation



Genome Wide Scans of Association (GWAS)

- The simplest: χ^2 test (--assoc in PLINK1.9, cf workshop)
- Works only for discrete categories
- Categories are not ordered
- Not possible to directly account for structure or covariables.



Genome Wide Scans of Association (GWAS)

- Basic principle: a linear model of association.

Continous phenotype

intercept

genotype

Effect size (slope)

$$Y = \alpha + X \cdot \beta + \epsilon$$

residuals

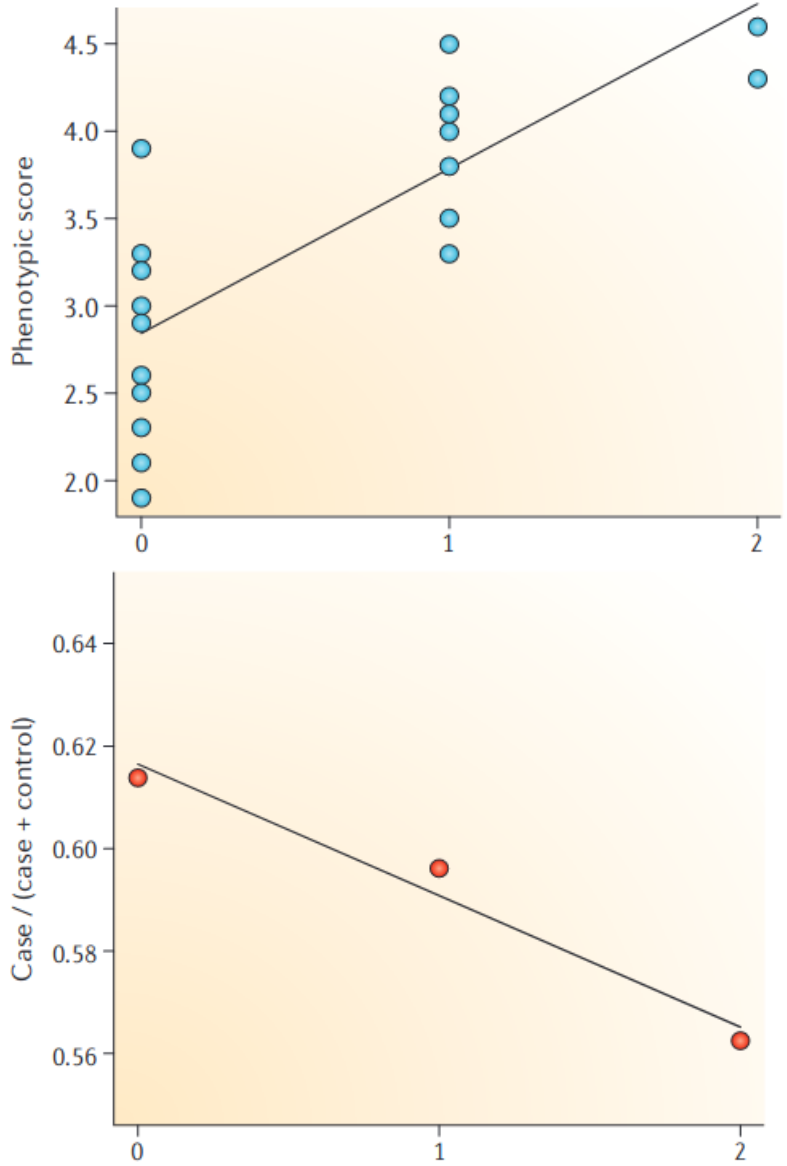
Continuous phenotype:
Linear regression

$$\ln\left(P/(1-P)\right) = \alpha + X \cdot \beta + \epsilon$$

Probability(case)

ln(odds ratio)

Binary phenotype (case-control):
Logistic regression



Balding Nat. Rev. Genet. 2010

Genome Wide Scans of Association (GWAS)

- We can add confounding variables and complexify the model if desired.
- Many statistical methods to increase power
- Unless you are working with very large panels, do not overinterpret differences between methods.
- More about this later…

$$Y = \alpha + X \cdot \beta + u + \epsilon$$

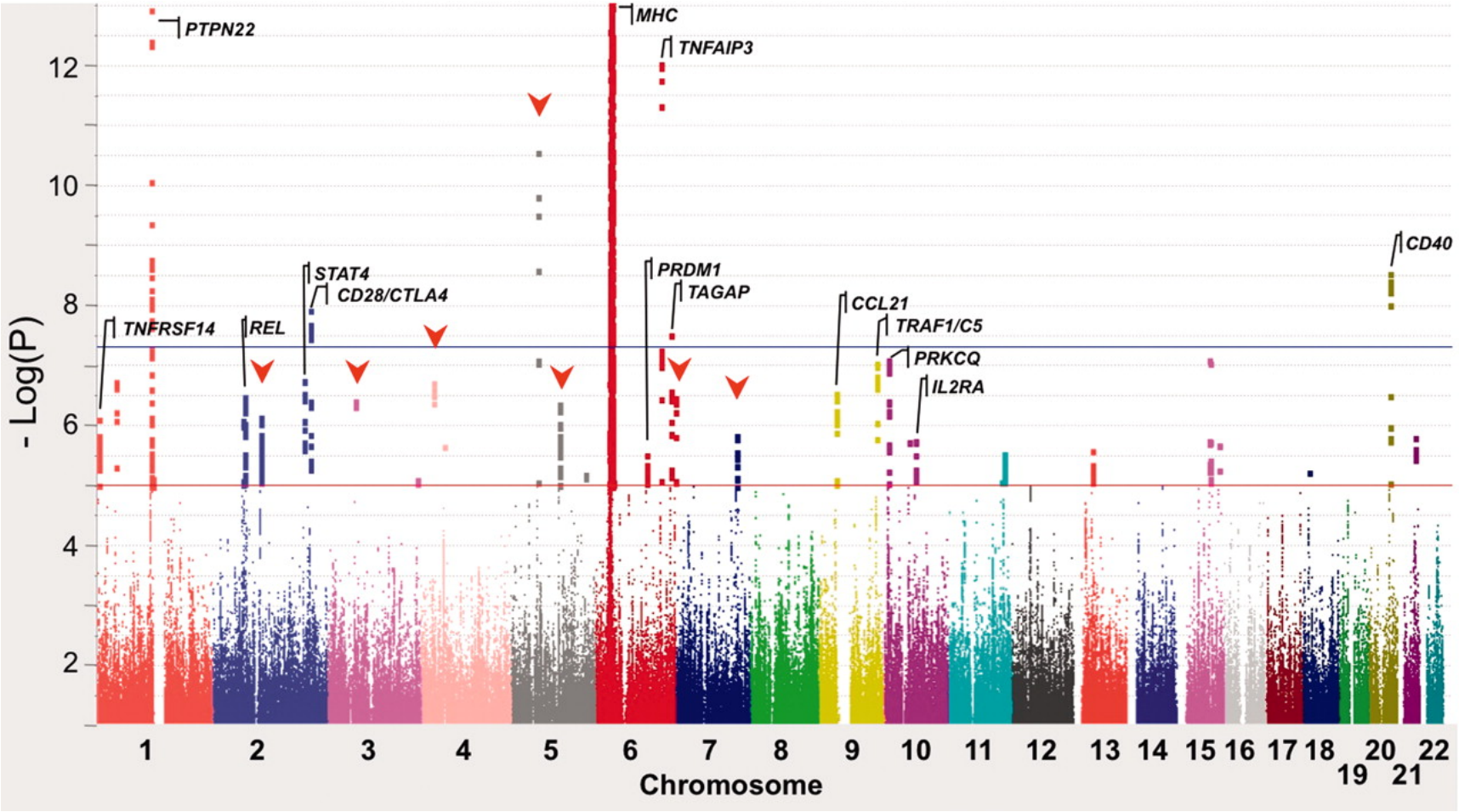
Correction for population structure/relatedness
Can also be latent factors (LFMM2)

About **LFMM2**: Kevin Caye, Basile Jumentier, Johanna Lepeule, Olivier François, LFMM 2: Molecular Biology and Evolution, Volume 36, Issue 4, April 2019, Pages 852–860, https://doi.org/10.1093/molbev/msz008

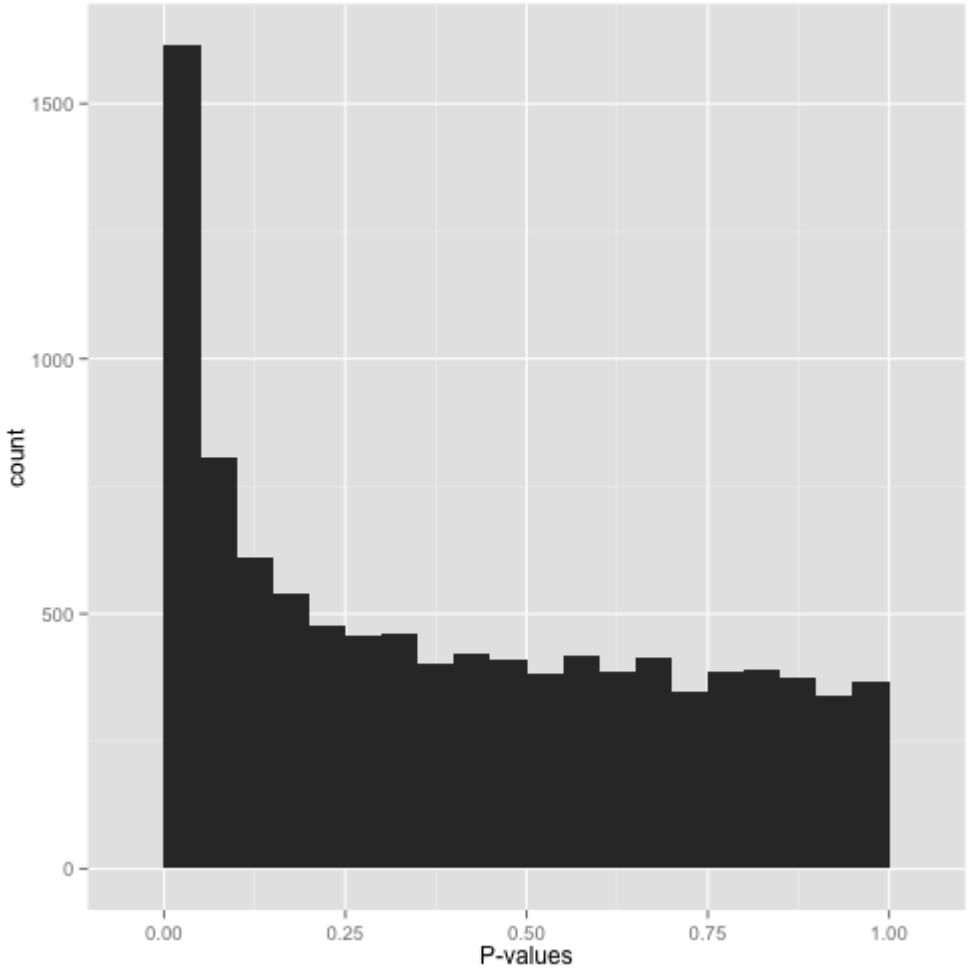
Genome Wide Scans of Association (GWAS)

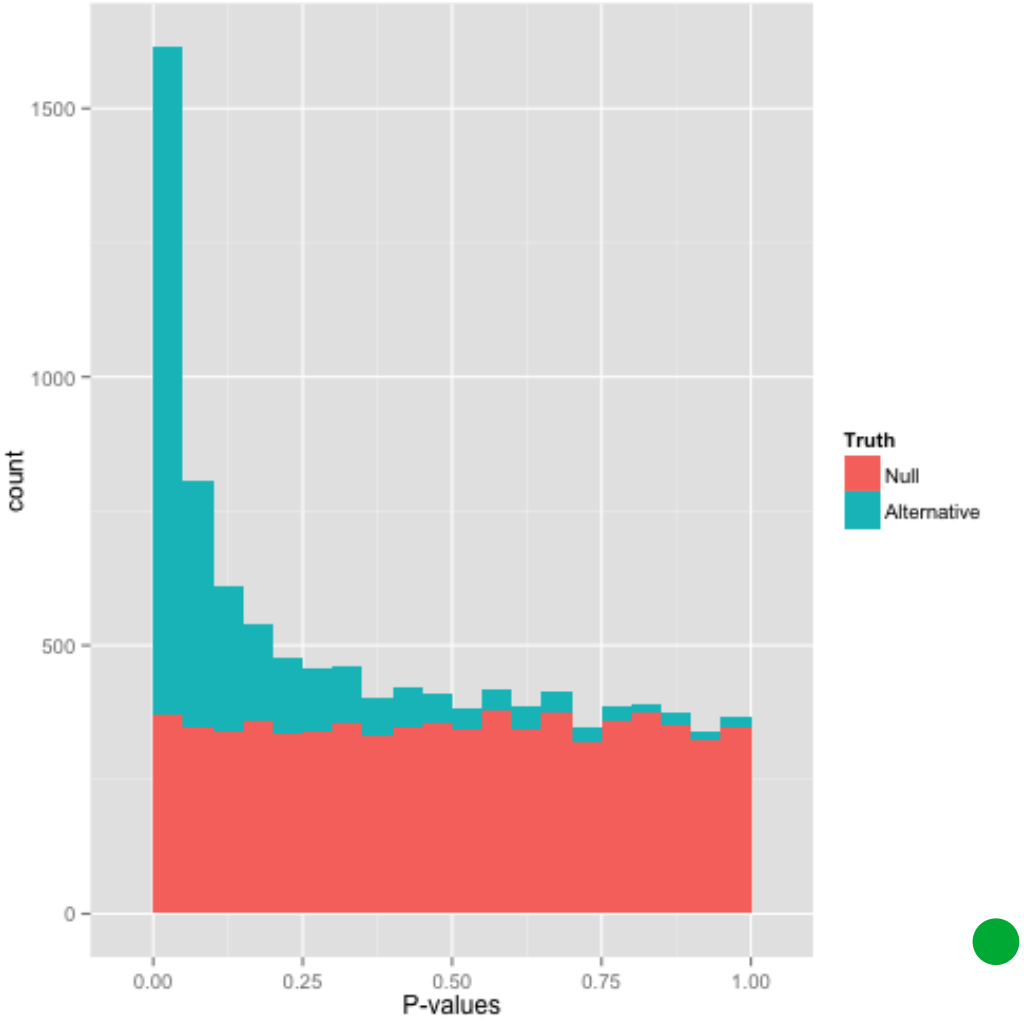
- Here a scan for association with rheumatoid arthritis
- Cohorts with 100,000 or even million individuals

- Useful to a point
- For polygenic traits: need high sample size



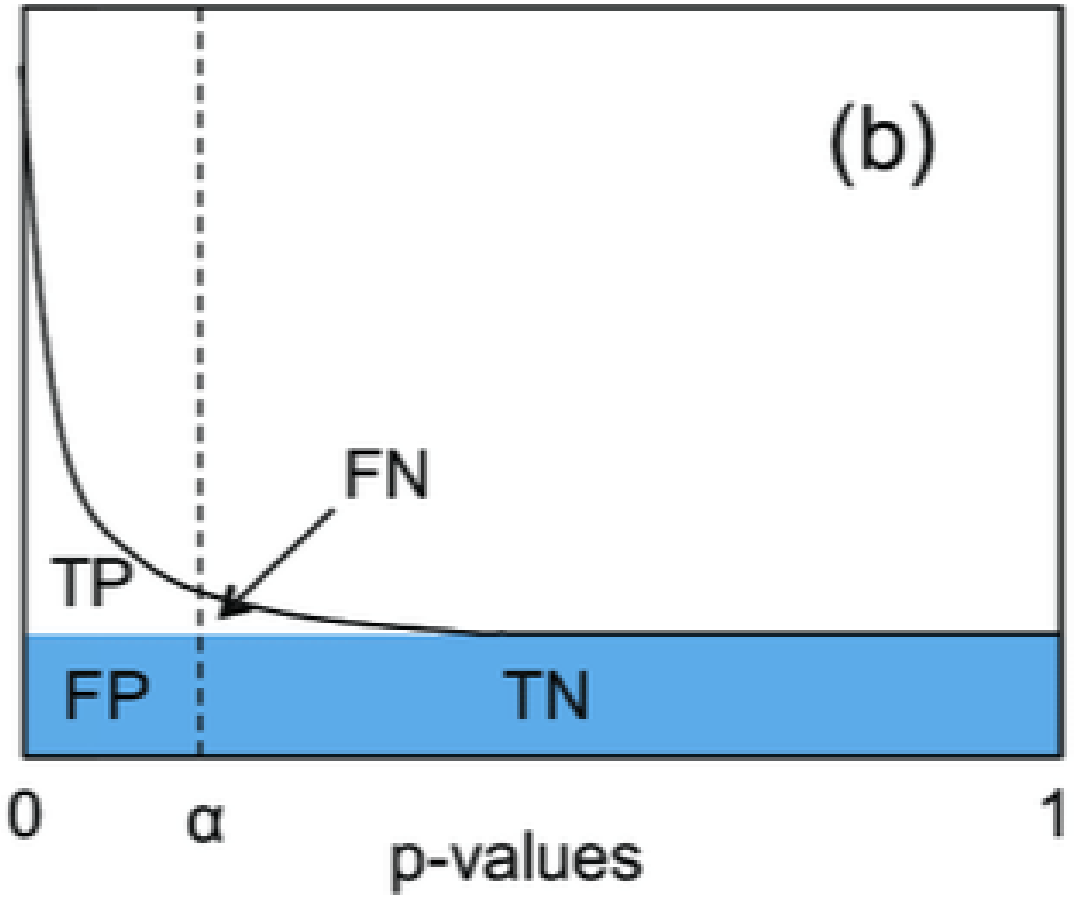
Always check your p-value distribution (not just for GWAS)

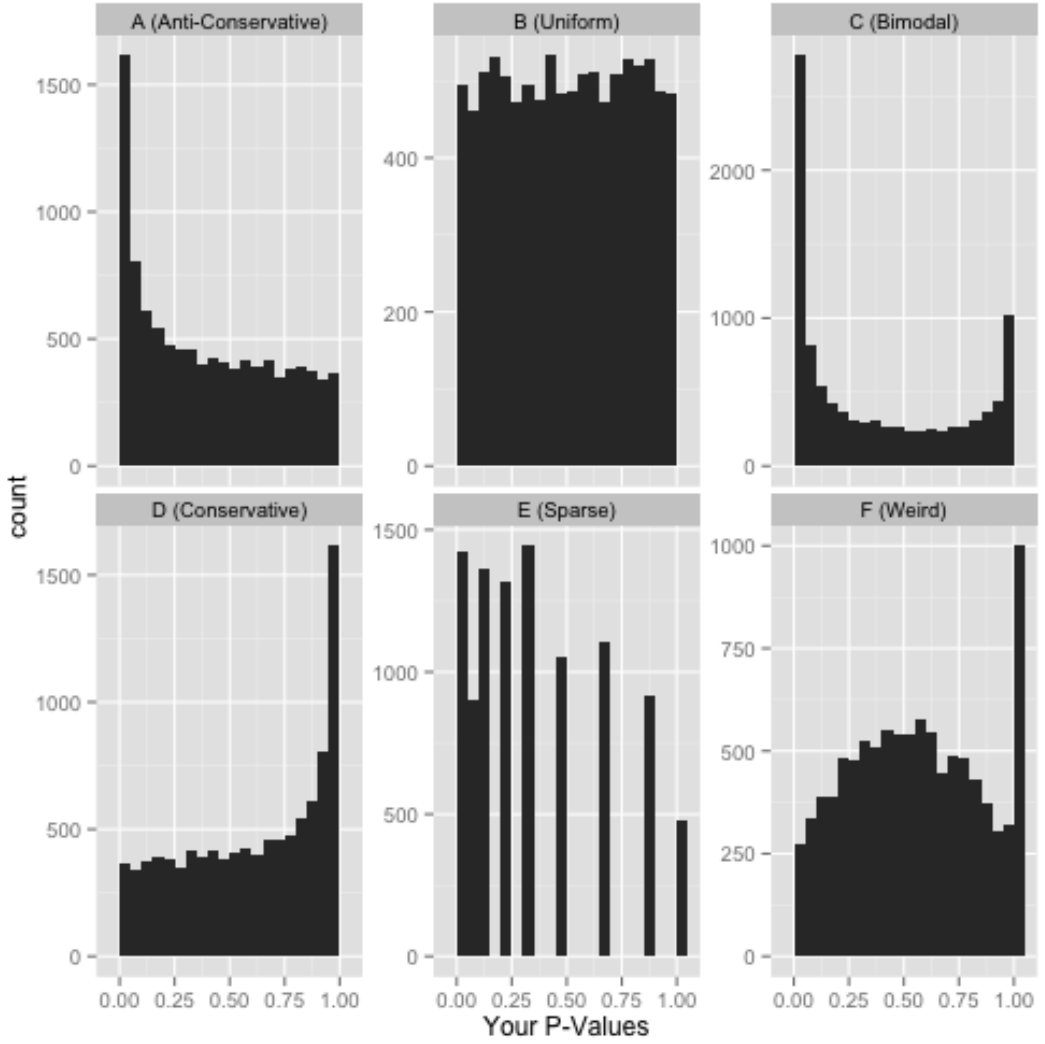




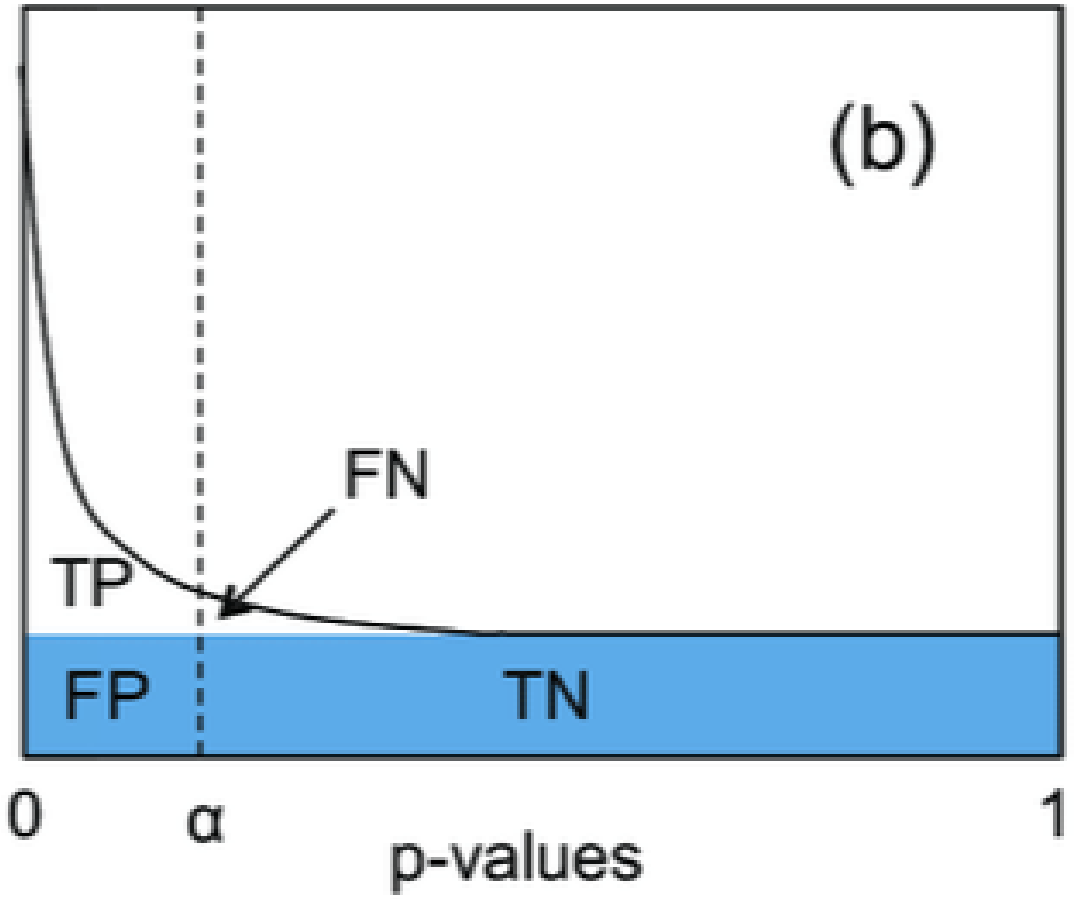
http://varianceexplained.org/statistics/interpreting-pvalue-

Always check your p-value distribution (not just for GWAS)





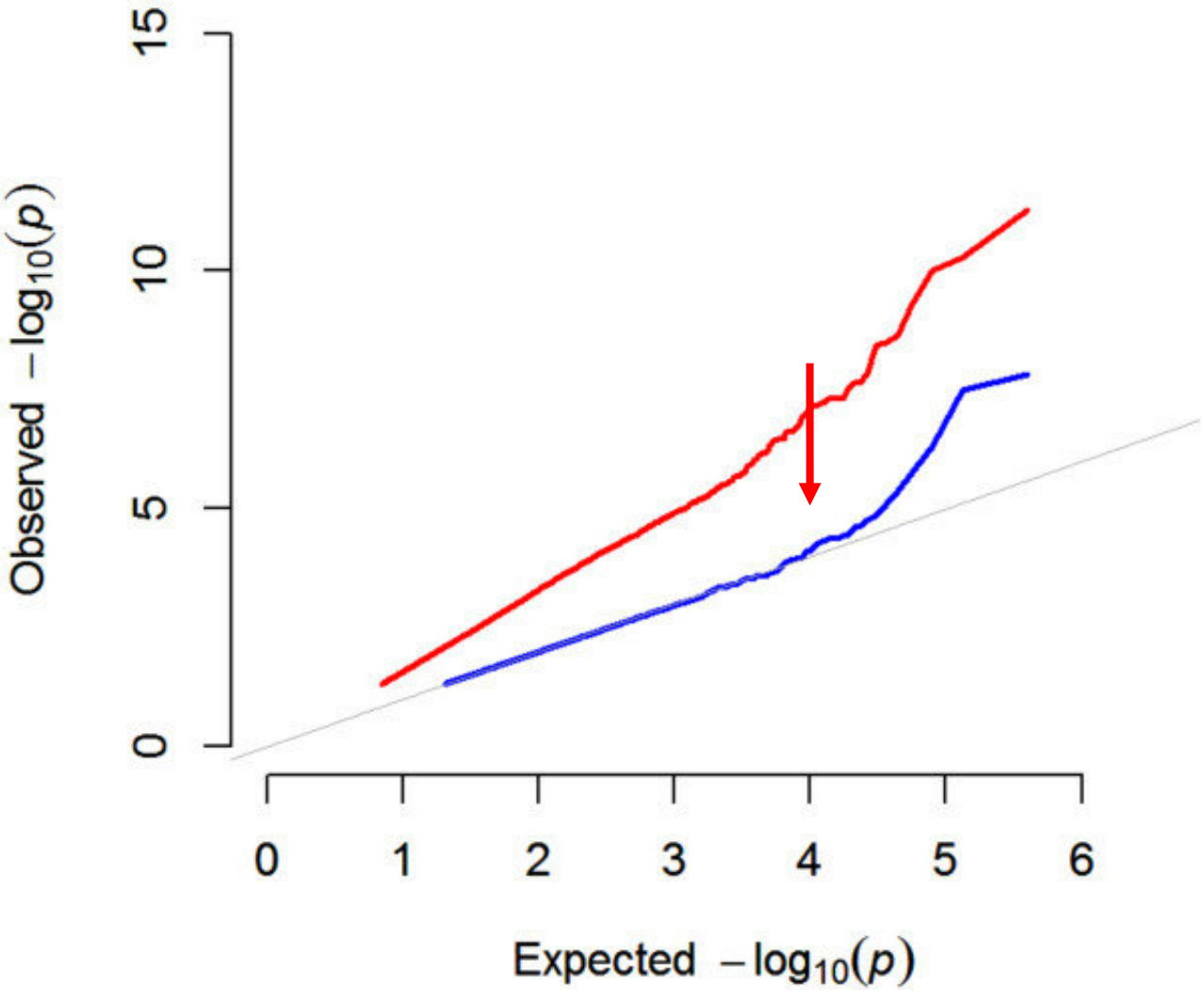
Always check your p-value distribution (not just for GWAS)



- You can apply a ‘brutal’ correction for multiple testing
- Bonferroni correction: multiply your *p*-value by the number of tests
- More subtle methods: *q*-value, the proportion of false positives for that particular p-value
- Local False Discovery Rate (FDR): the probability that H0 is actually true given this particular p-value.
- In R package qvalue
- Requires the distribution of p-values to be uniform.

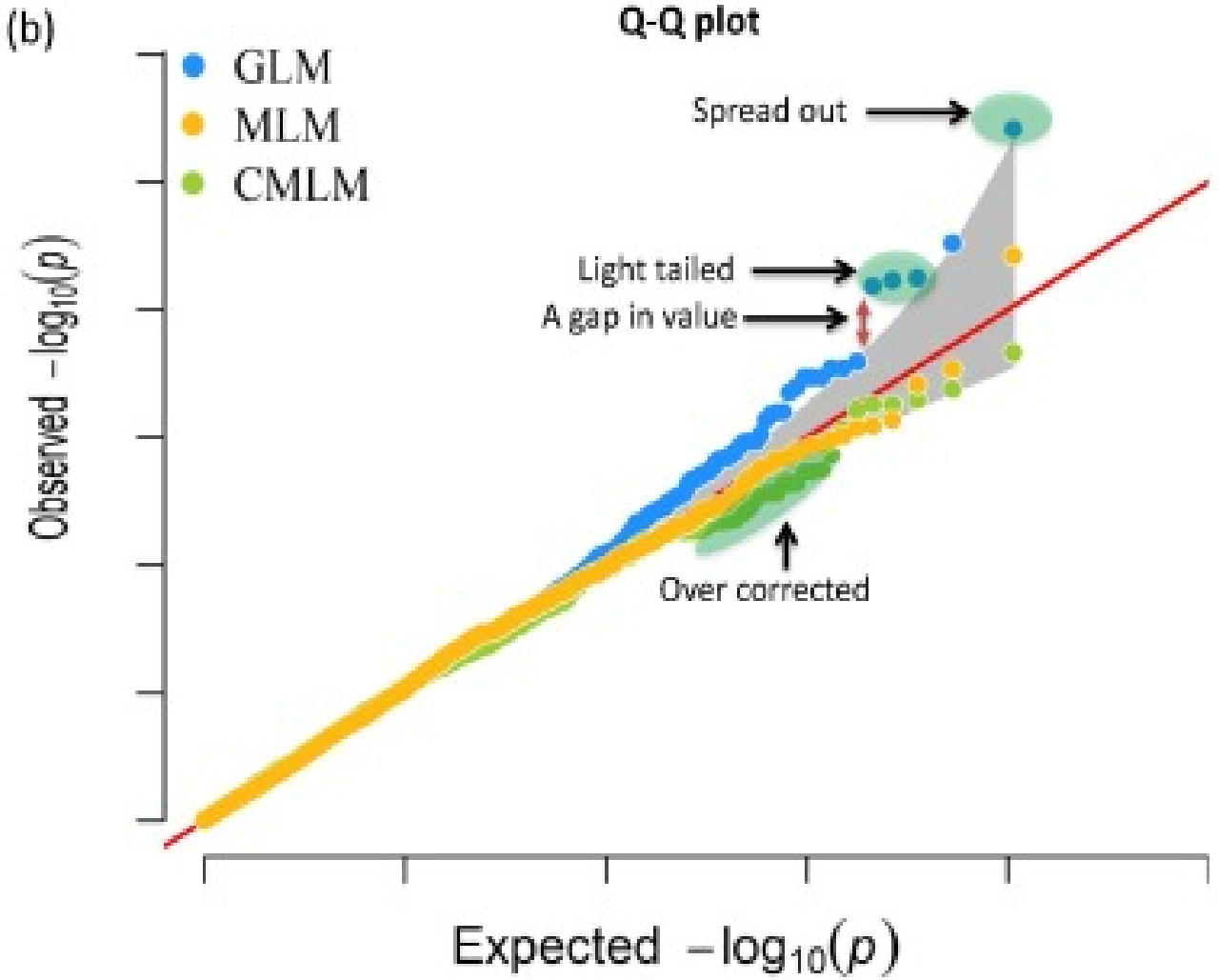
https://doi.org/doi:10.18129/

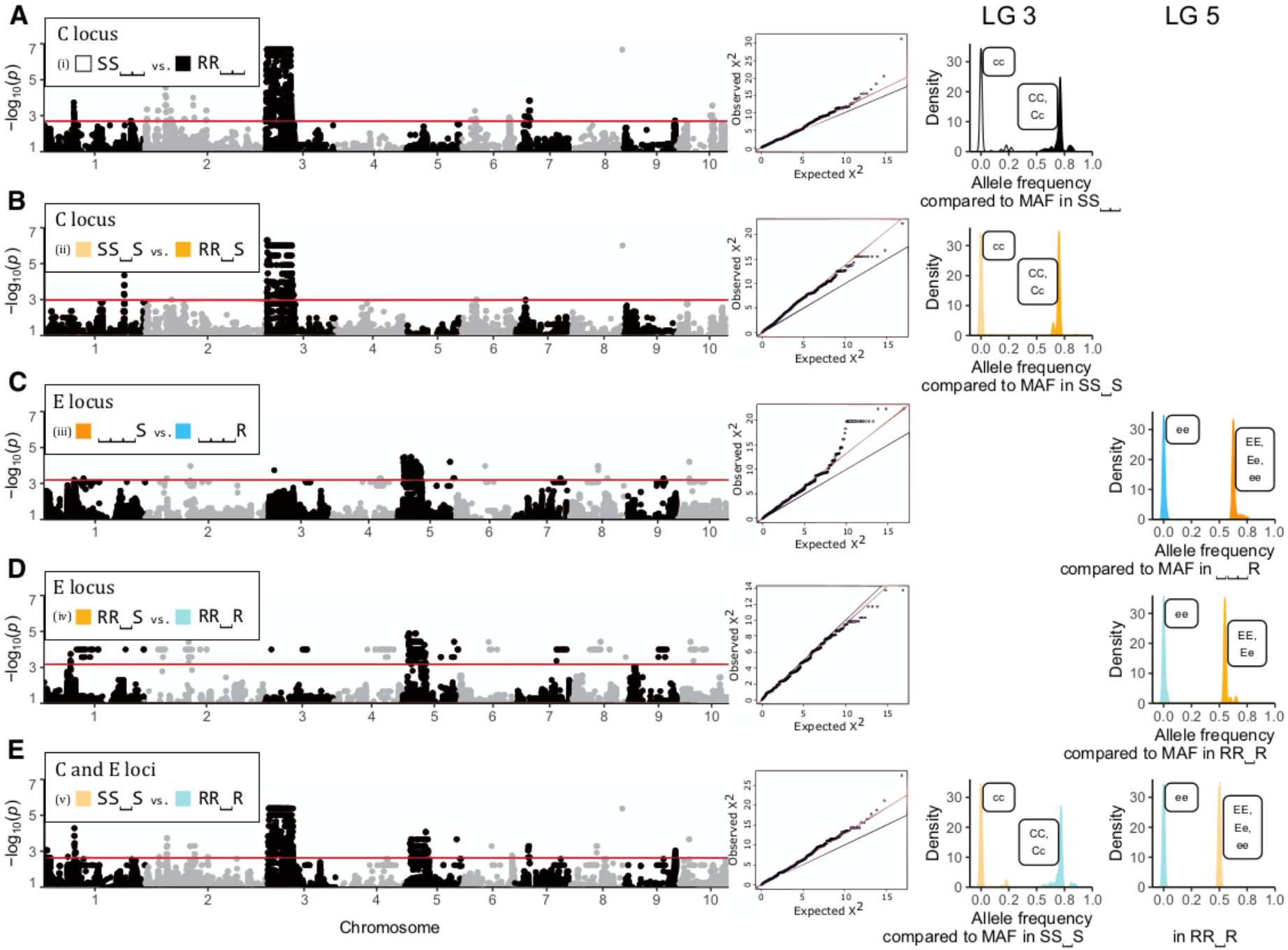
Correcting for p-value inflation



- The expectation is one drawn assuming a χ^2 distribution
- The ratio between observed and expected χ^2 values gives you a correction factor, called λ.
- Genomic control

Always check your p-value distribution (not just for GWAS)




A
C locus
(i) SS_ vs. RR_
B
C locus
(ii) SS_S vs. RR_S
C
E locus
(iii) _S vs. _R
D
E locus
(iv) RR_S vs. RR_R
E
C and E loci
(v) SS_S vs. RR_R
−log10(p)
Chromosome
Observed X2
Expected X2
LG 3
LG 5
Density
cc
CC, Cc
ee
EE, Ee, ee
EE, Ee
Allele frequency compared to MAF in SS_
Allele frequency compared to MAF in SS_S
Allele frequency compared to MAF in _R
Allele frequency compared to MAF in RR_R
in RR_R

Back to quantitative genetics

- Polygenic inheritance
- Refers to quantitative characteristics controlled by cumulative effects of many genes
- Each locus still follows Mendel's rules
- May be influenced by environmental factors
- Kind of works, and simplifies things…

2 loci with 2 alleles control spine number:

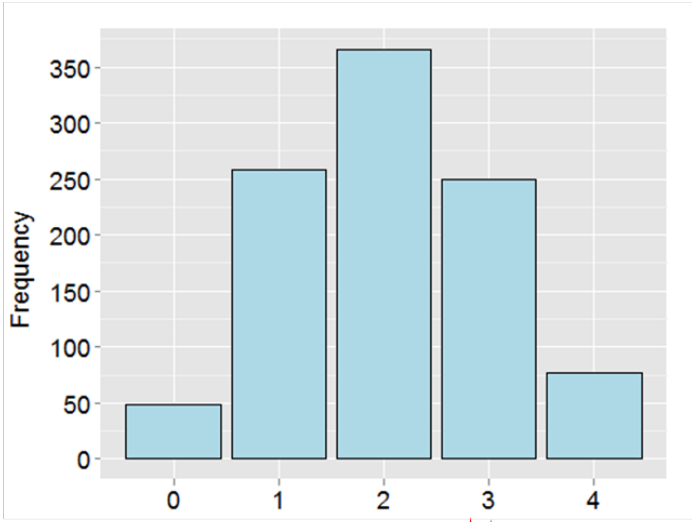
Locus 1: 0 +1

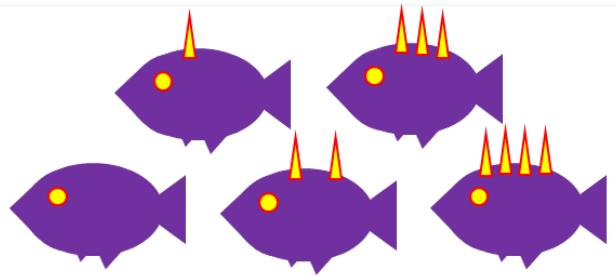
Locus 2: 0 +1

Nine genotype combinations:

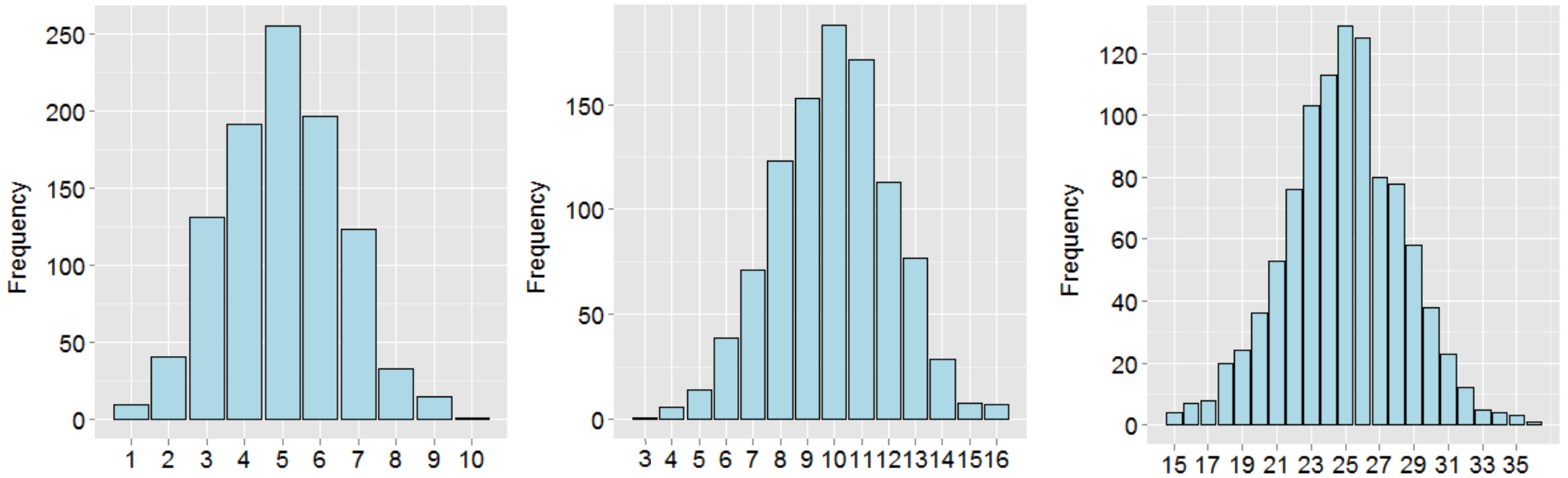
0 0 / 0 0	0 +1 / 0 0	+1 +1 / 0 0
No spines	**1 spine**	**2 spines**
0 0 / 0 +1	0 +1 / 0 +1	+1 +1 / 0 +1
1 spine	**2 spines**	**3 spines**
0 0 / +1 +1	0 +1 / +1 +1	+1 +1 / +1 +1
2 spines	**3 spine**	**4 spines**

If p = 0.5 at both loci, we expect phenotypic distribution to be:





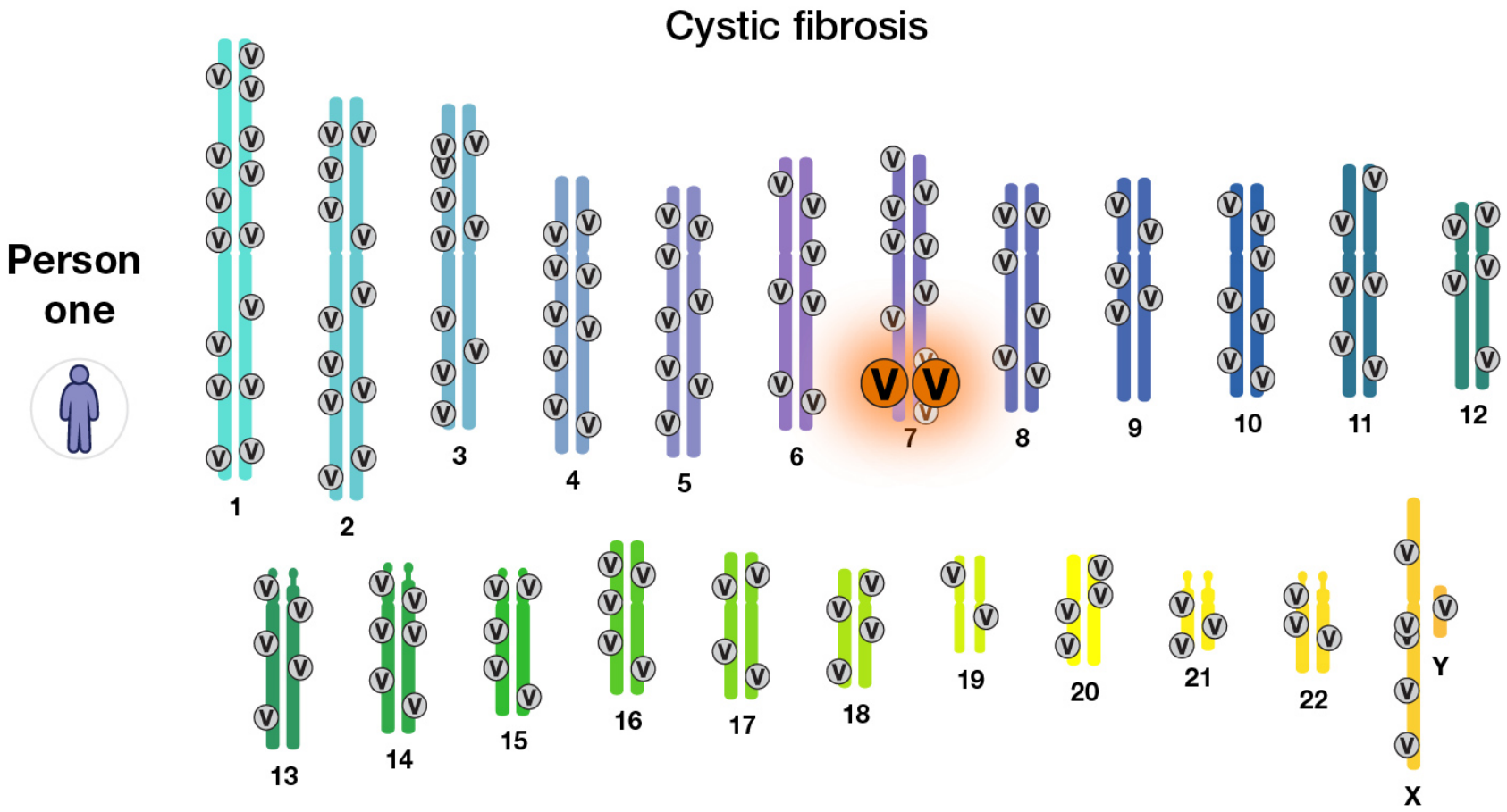
Back to quantitative genetics



- As more loci control the trait, it becomes difficult to describe the trait in terms of individual loci
- Central Limit theorem: the combination of random draws converges to a normal distribution.

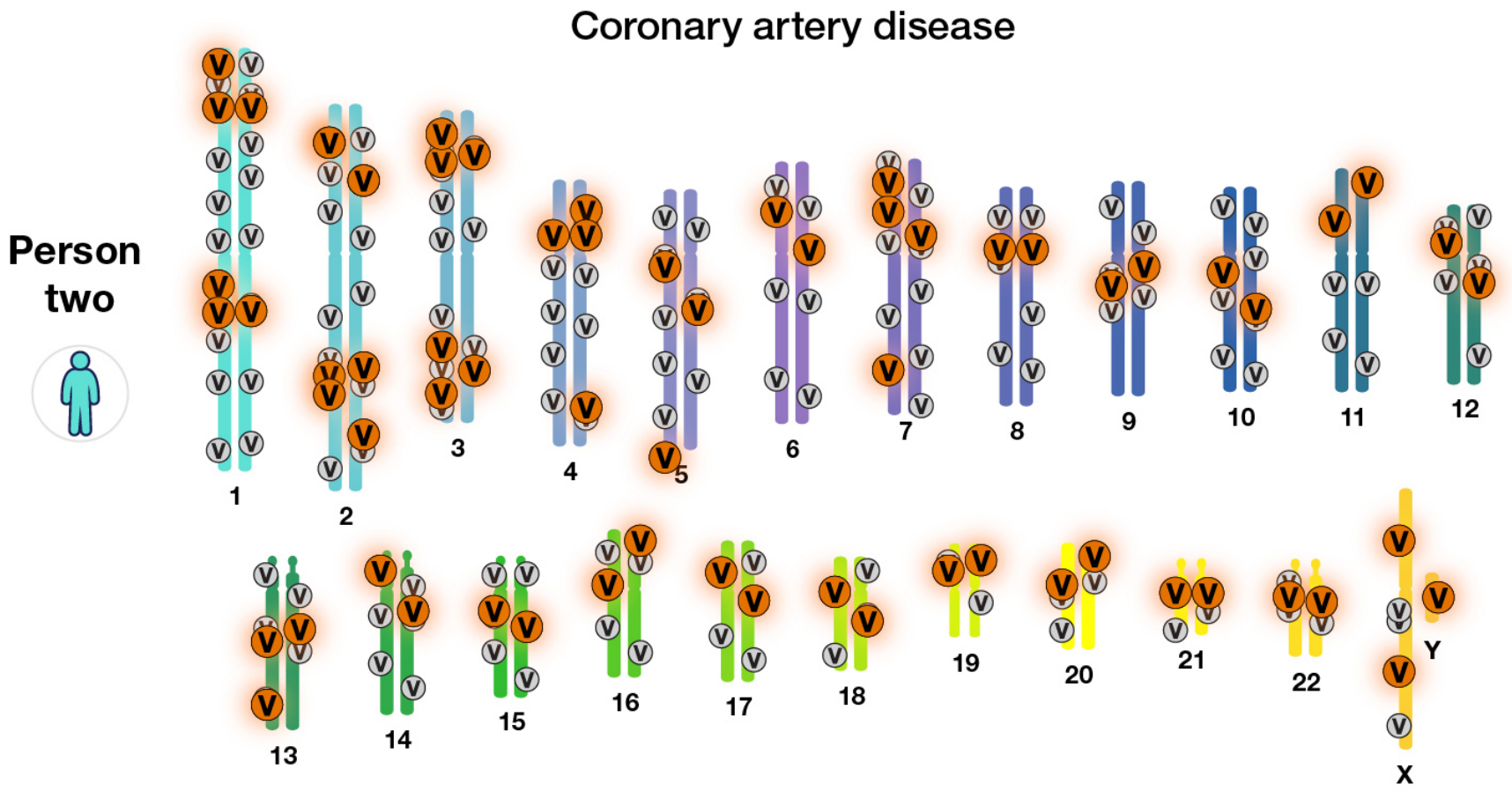
Polygenic risk scores

- We assume additive effects
- The risk is therefore additive.
- Neglects epistasy.
- Using genotypes at associated loci to predict an outcome
- Here with a single locus



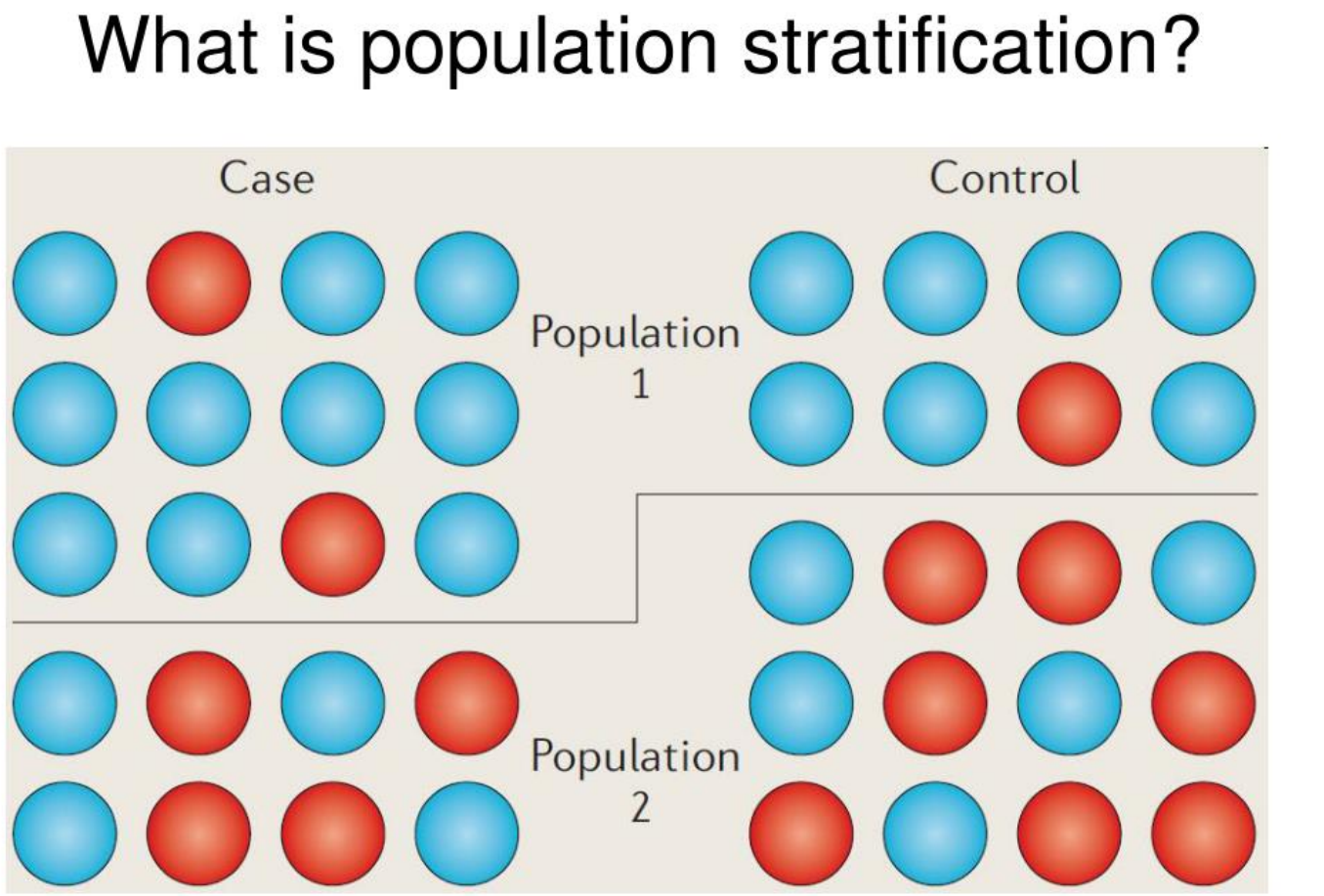
Polygenic risk scores

- Using genotypes at associated loci to predict an outcome
- Here with many loci



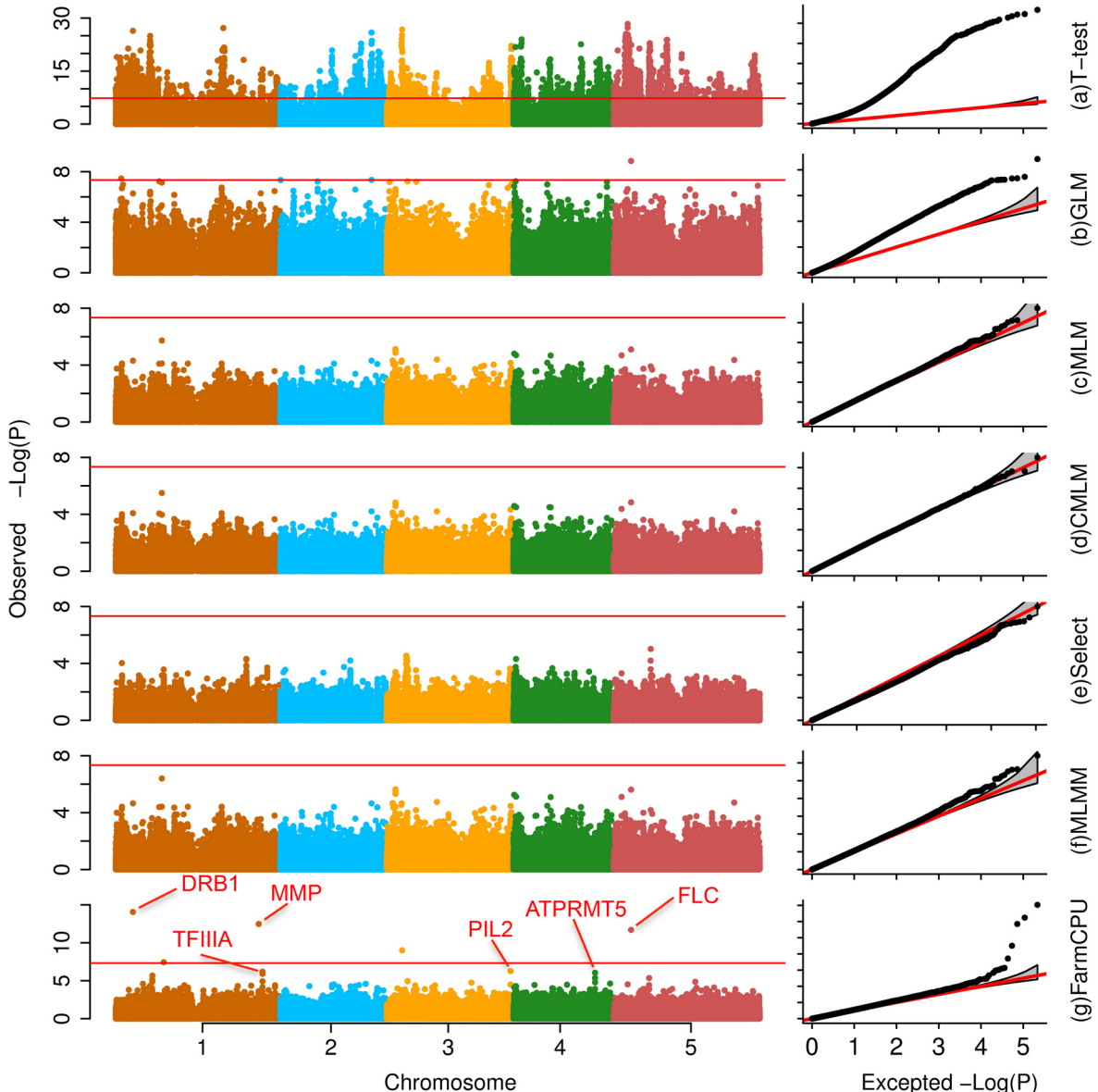
Genome Wide Scans of Association (GWAS): limitations and pitfalls

- Small effect sizes
- Biases due to population structure (stratification)
- Ethnic bias
- Story-telling
- Look at the effect size, not just the p-value

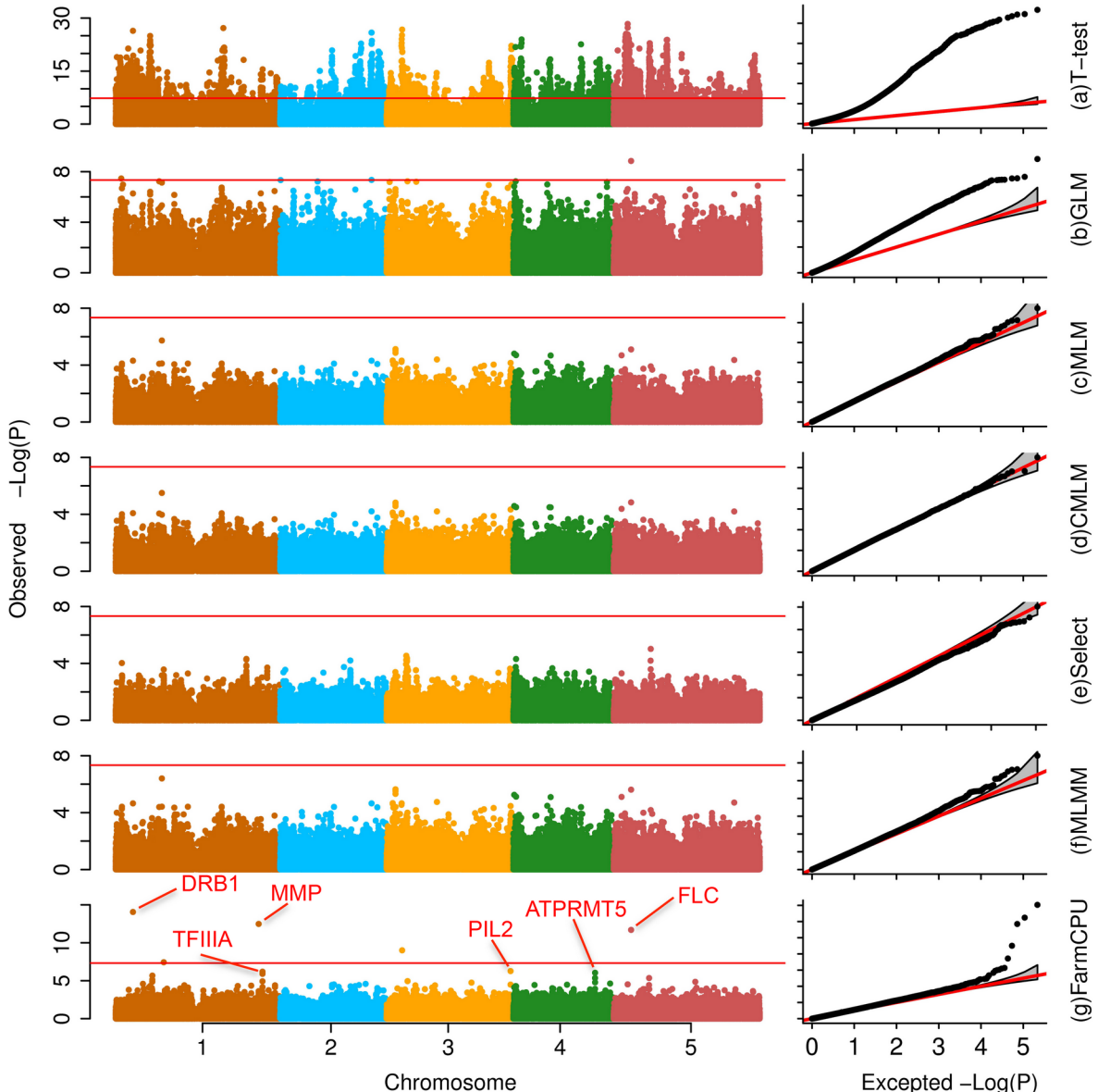


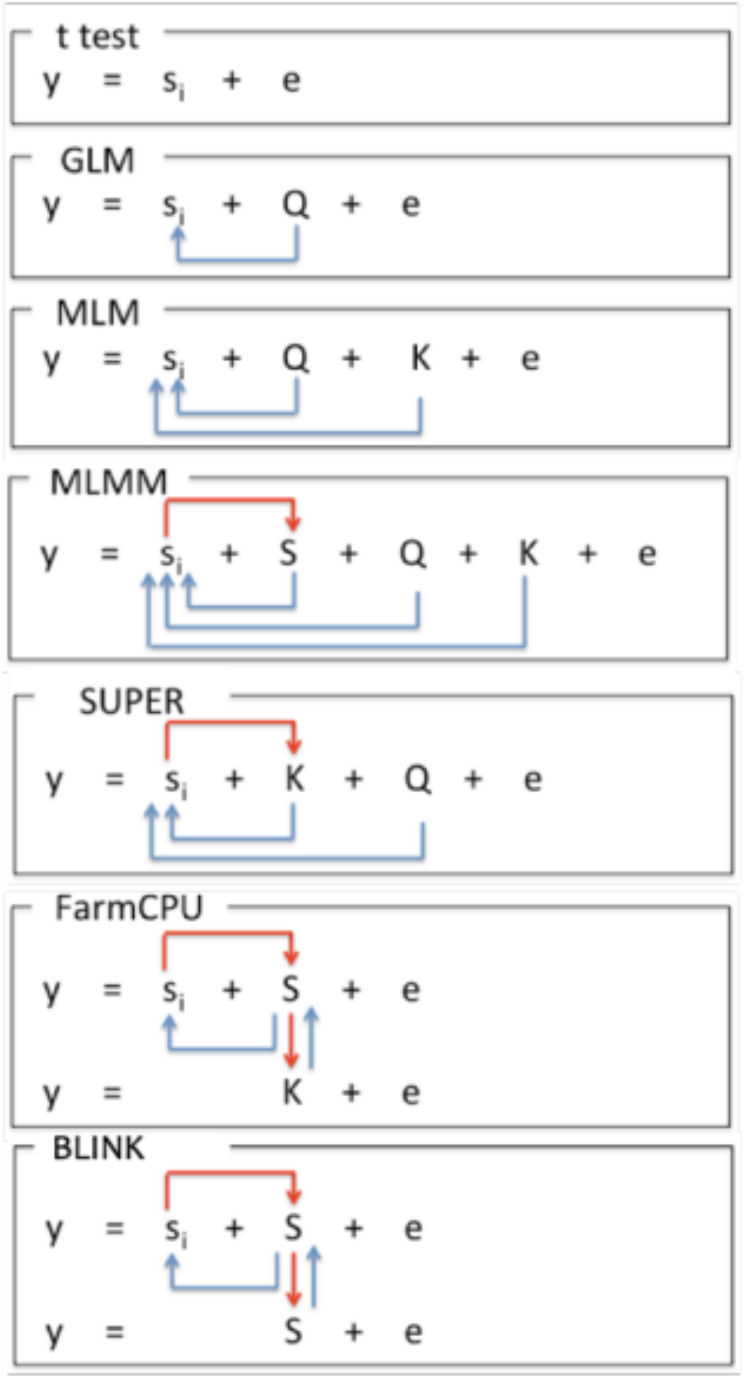
Balding, Nature Reviews Genetics 2010

A few GWAS models



A few GWAS models





S_i: Testing marker Q: Population structure K: Kinship

S: Pseudo QTNs Q: Population structure Arrow: adjustment

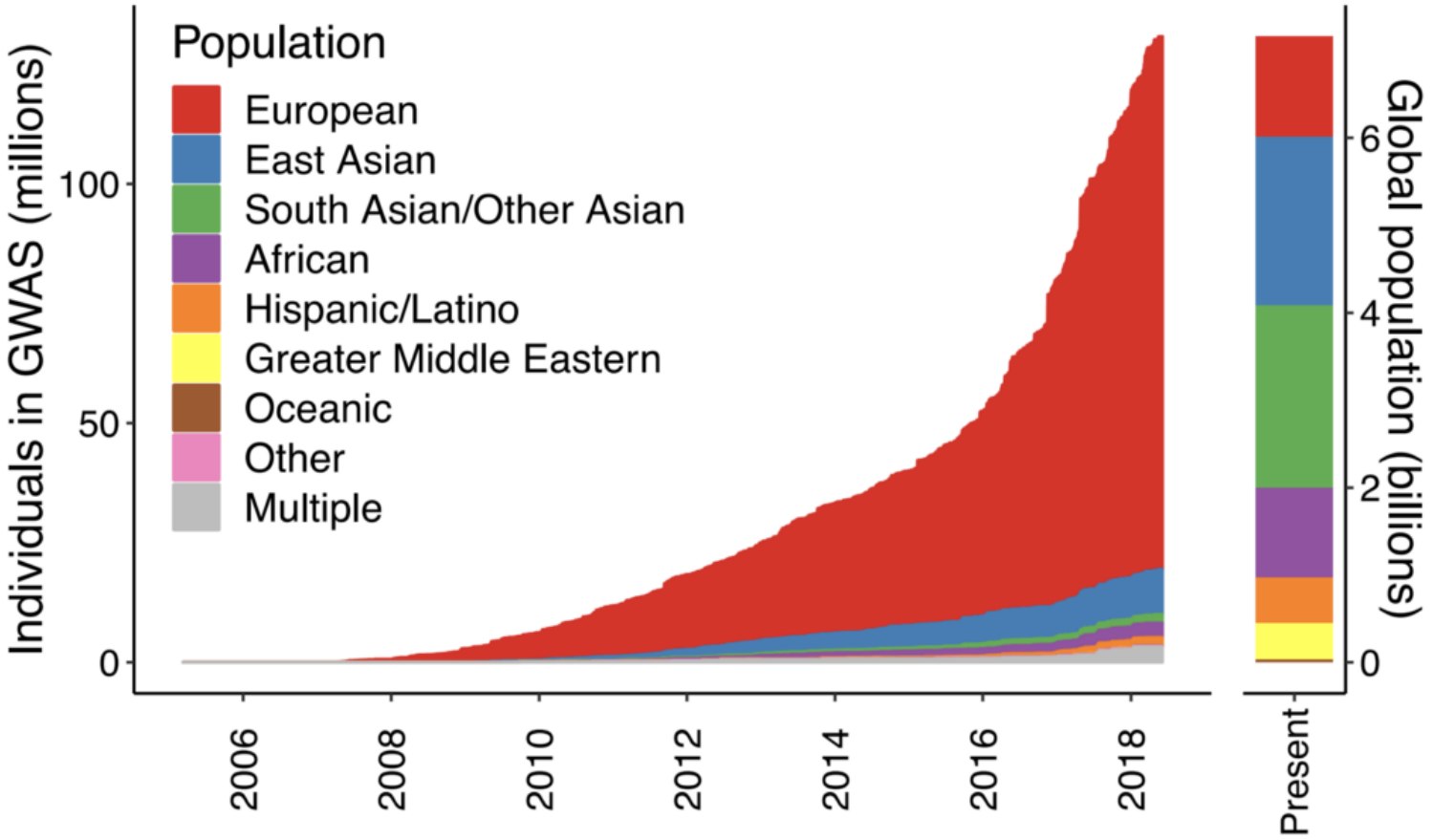
Abstract

False positives in a Genome-Wide Association Study (GWAS) can be effectively controlled by a fixed effect and random effect Mixed Linear Model (MLM) that incorporates population structure and kinship among individuals to adjust association tests on markers; however, the adjustment also compromises true positives. The modified MLM method, Multiple Loci Linear Mixed Model (MLMM), incorporates multiple markers simultaneously as covariates in a stepwise MLM to partially remove the confounding between testing markers and kinship. To completely eliminate the confounding, we divided MLMM into two parts: Fixed Effect Model (FEM) and a Random Effect Model (REM) and use them iteratively. FEM contains testing markers, one at a time, and multiple associated markers as covariates to control false positives. To avoid model over-fitting problem in FEM, the associated markers are estimated in REM by using them to define kinship. The P values of testing markers and the associated markers are unified at each iteration. We named the new method as Fixed and random model Circulating Probability Unification (FarmCPU). Both real and simulated data analyses demonstrated that FarmCPU improves statistical power compared to current methods. Additional benefits include an efficient computing time that is linear to both number of individuals and number of markers. Now, a dataset with half million individuals and half million markers can be analyzed within three days.

The second strategy changes the definition of kinship among individuals. Only the associated genetic markers are used as pseudo Quantitative Trait Nucleotides (QTNs) to derive kinship instead of all, or a random sample of genetic markers. Pseudo QTNs are expected to closely track some of the causative QTNs, and are selectively used to derive kinship for a specific testing marker. Whenever a pseudo QTN is correlated with the testing marker, it is excluded from those used to derive kinship. In the FaST-LMM-Select method, a pseudo QTN is considered correlated if it is within a 2Mb interval on either side of the testing marker[23]. Instead of using a 2Mb interval, the Settlement of MLM Under Progressively Exclusive Relationship (SUPER) method applies a threshold on Linkage Disequilibrium (LD) between the pseudo QTNs and the testing marker. Selectively including and/or excluding pseudo QTNs to derive kinship for a specific testing marker improves statistical power compared to deriving a overall kinship from all, or a random sample of genetic markers[24].

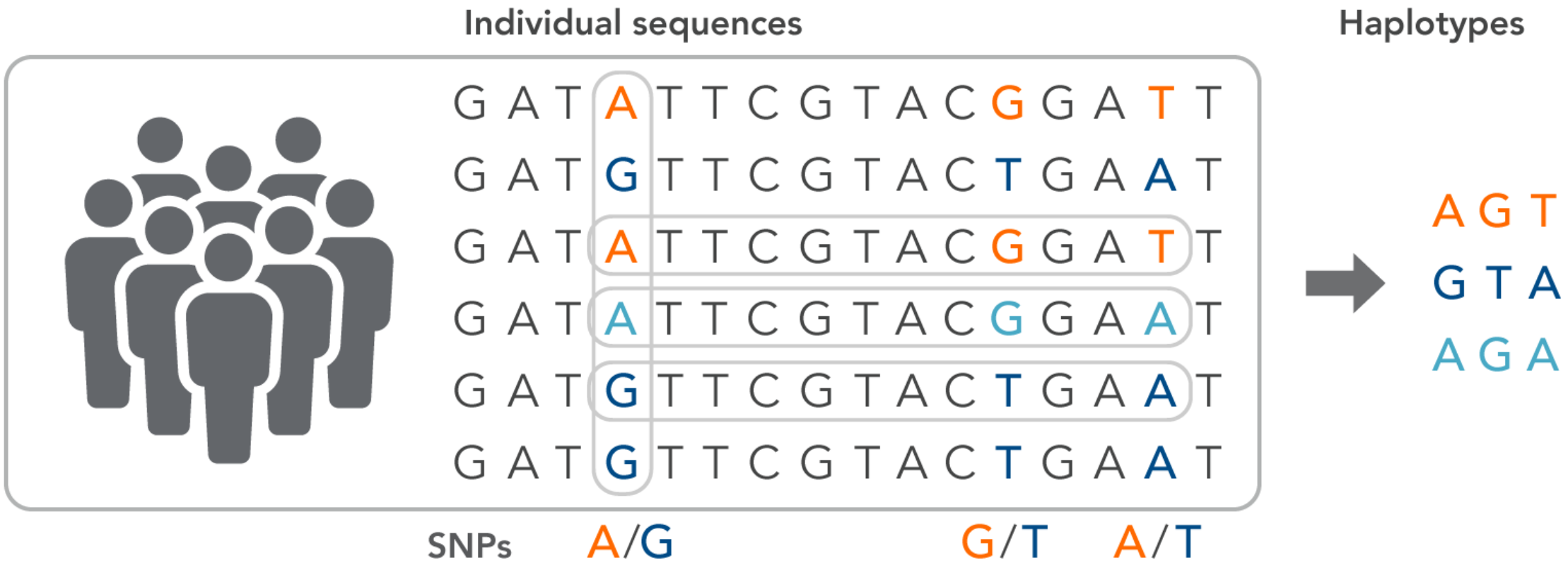
Genome Wide Scans of Association (GWAS): limitations and pitfalls

- Small effect sizes
- Biases due to population structure (stratification)
- Ethnic bias
- Story-telling
- Look at the effect size (β), not just the p-value.



Martin et al. Nat Genet 2019.

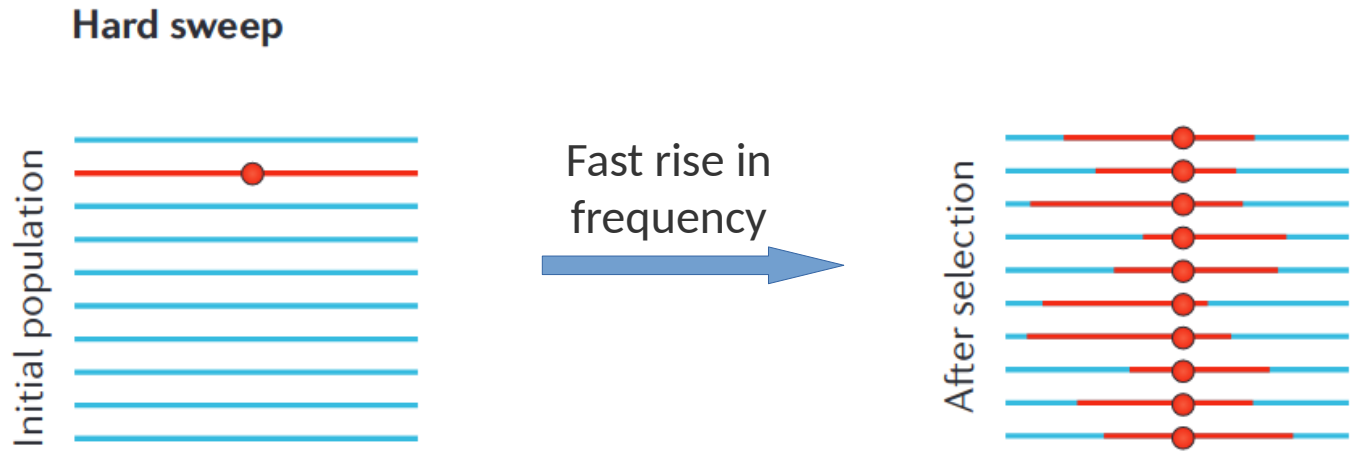
Scans for selection. Quick reminder



https://www.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know

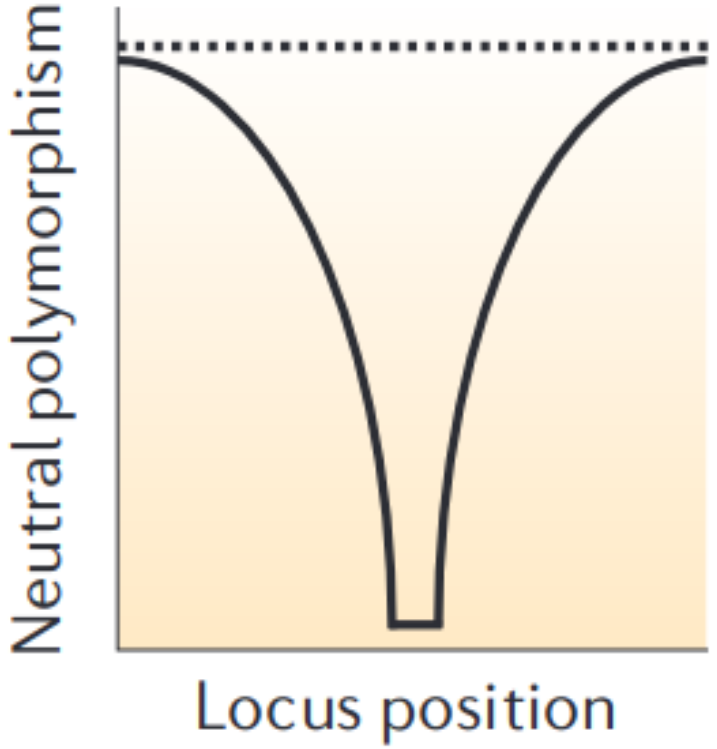
A (hard) selective sweep

- The association between alleles across individuals usually breaks down because of recombination
- Recent selection leads to a fast increase in frequency of the selected allele and those at the vicinity (selective sweep)
- One consequence: drop in ‘how different’ alleles are near the selected locus



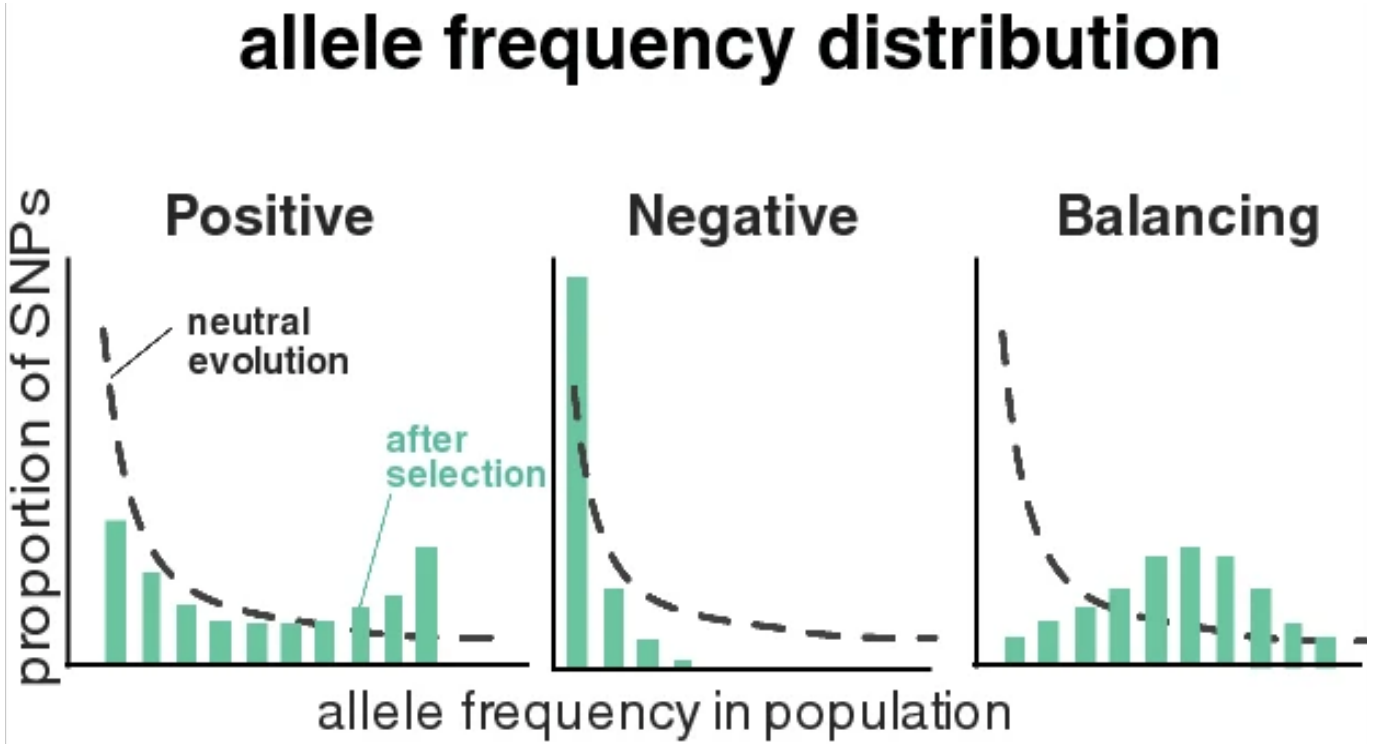
Cutter and Payseur Nat. Rev. Genet. 2013

Effect on diversity statistics



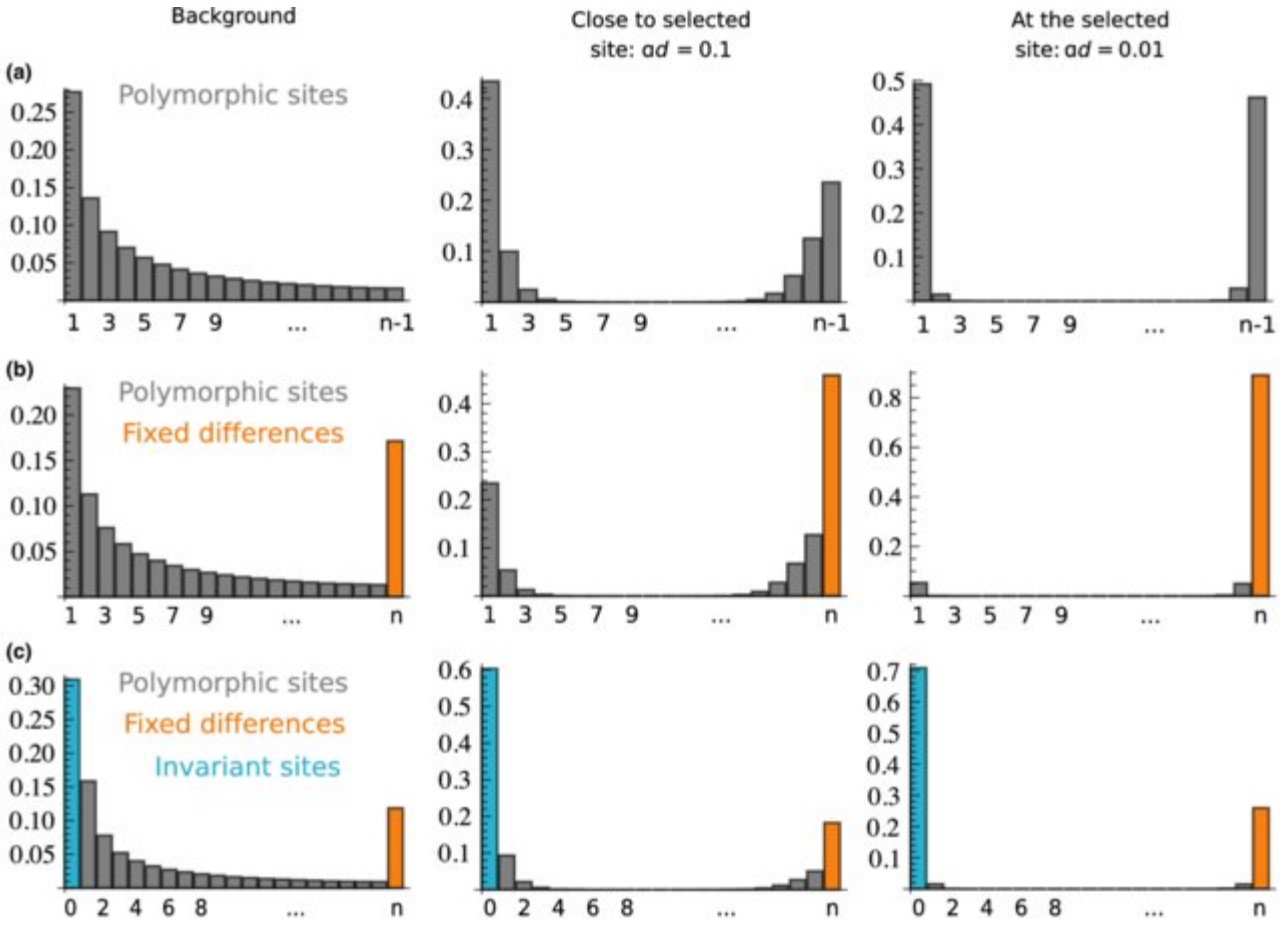
Effect on the Allele Frequency Spectrum

- A selective sweep leads to a change in the frequency of derived alleles
- This signature can be detected by contrasting local and global (or simulated) 'neutral' spectra
- Can be described by diversity statistics, such as Tajima's *D*



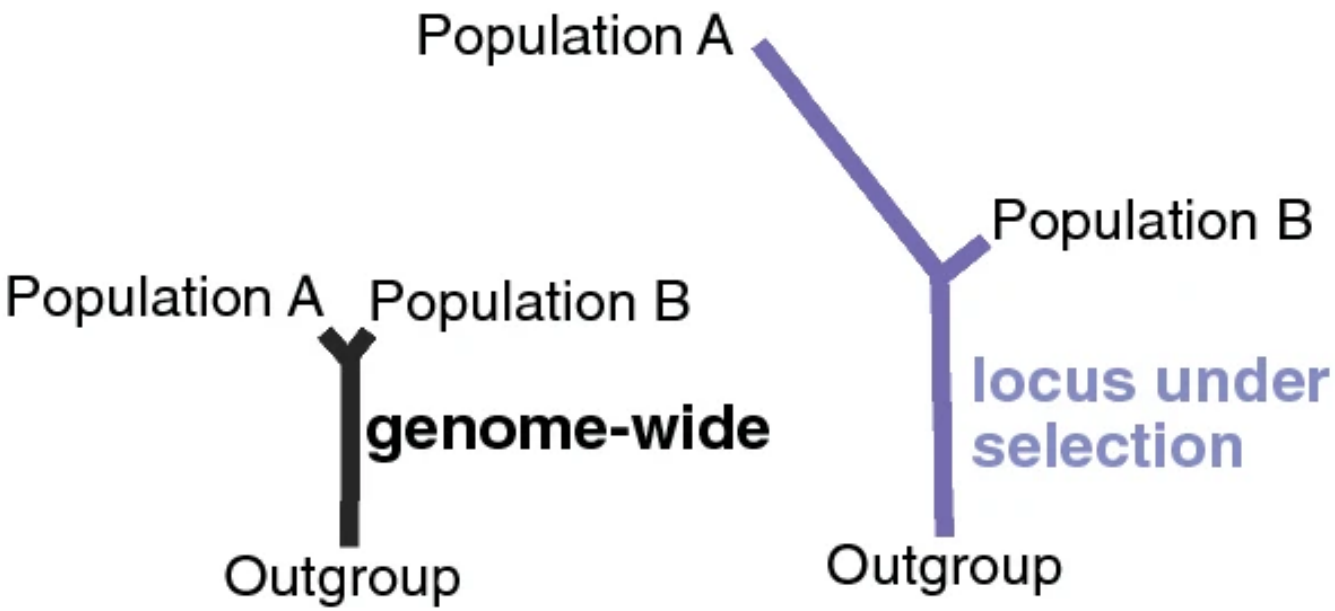
Effect on the Allele Frequency Spectrum

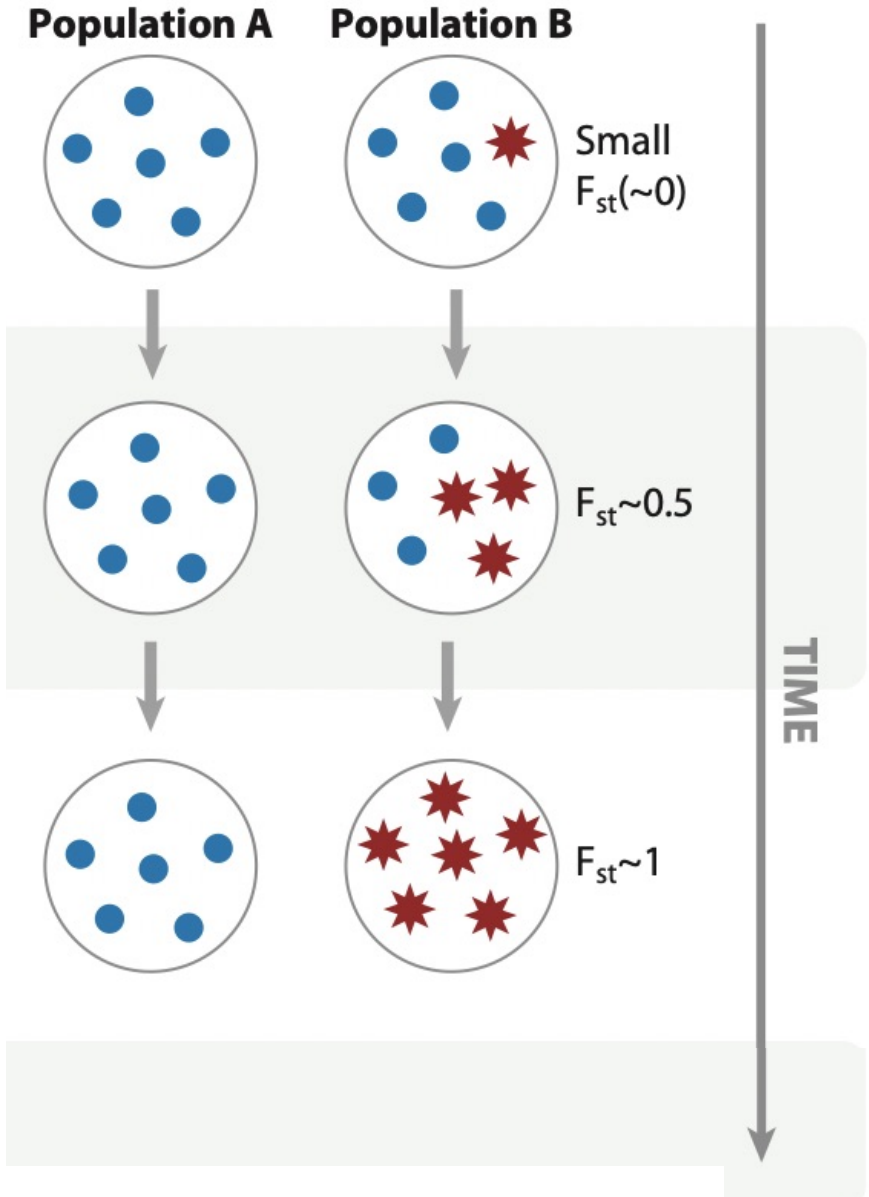
- A selective sweep leads to a change in the frequency of derived alleles
- This signature can be detected by contrasting local and global (or simulated) ‘neutral’ spectra
- Can be described by diversity statistics, such as Tajima’s *D*
- The entire spectrum can be used more efficiently: Composite Likelihood Ratios (CLR) estimated by SweepFinder/SweeD.



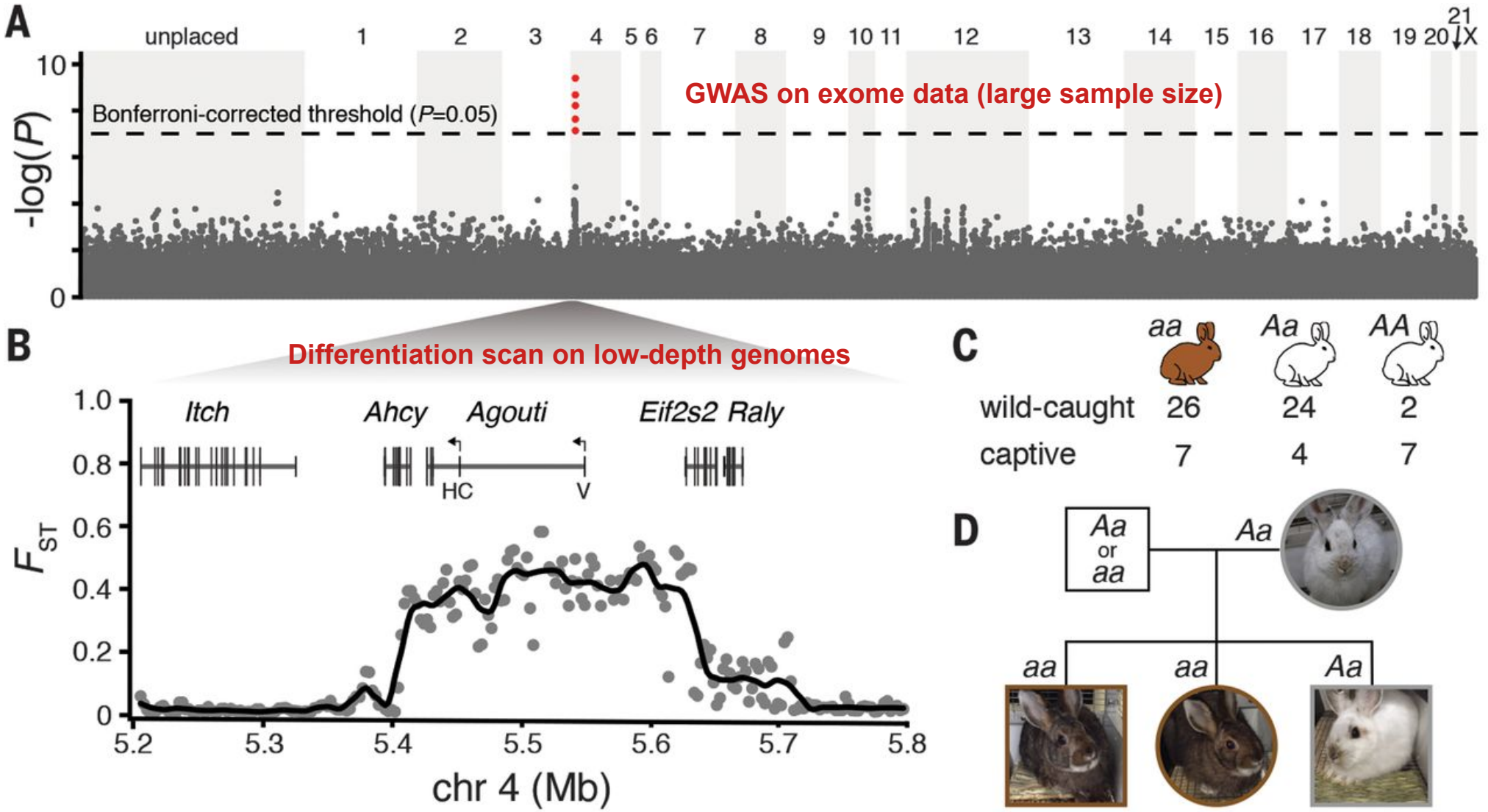
Effect on differentiation statistics

- Increase in F_{ST}
- Increase in Population Branch Tests
- More sophisticated methods exist (e.g. BAYPASS). See Friday's lecture on GEA.



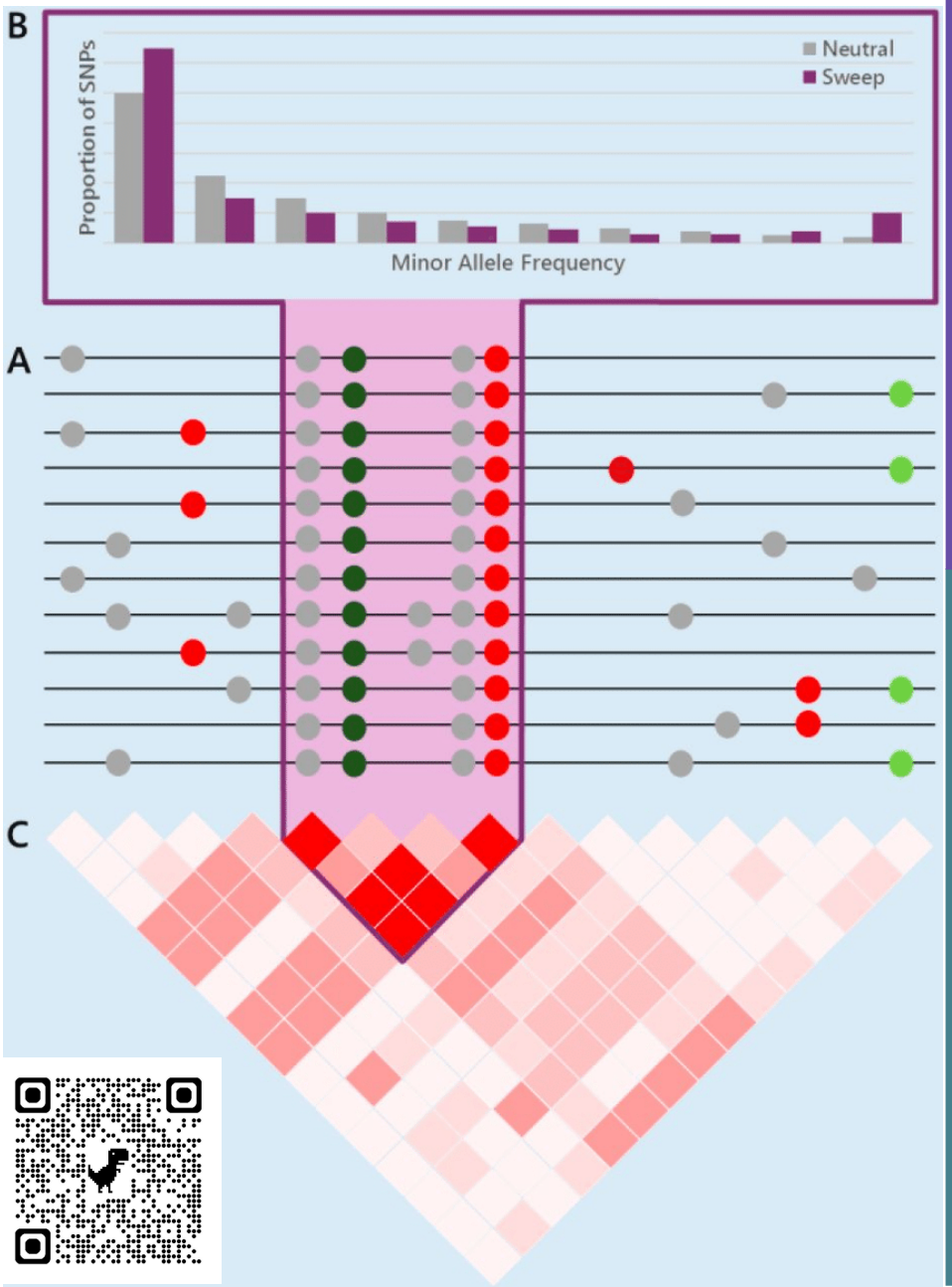


Scans of Differentiation



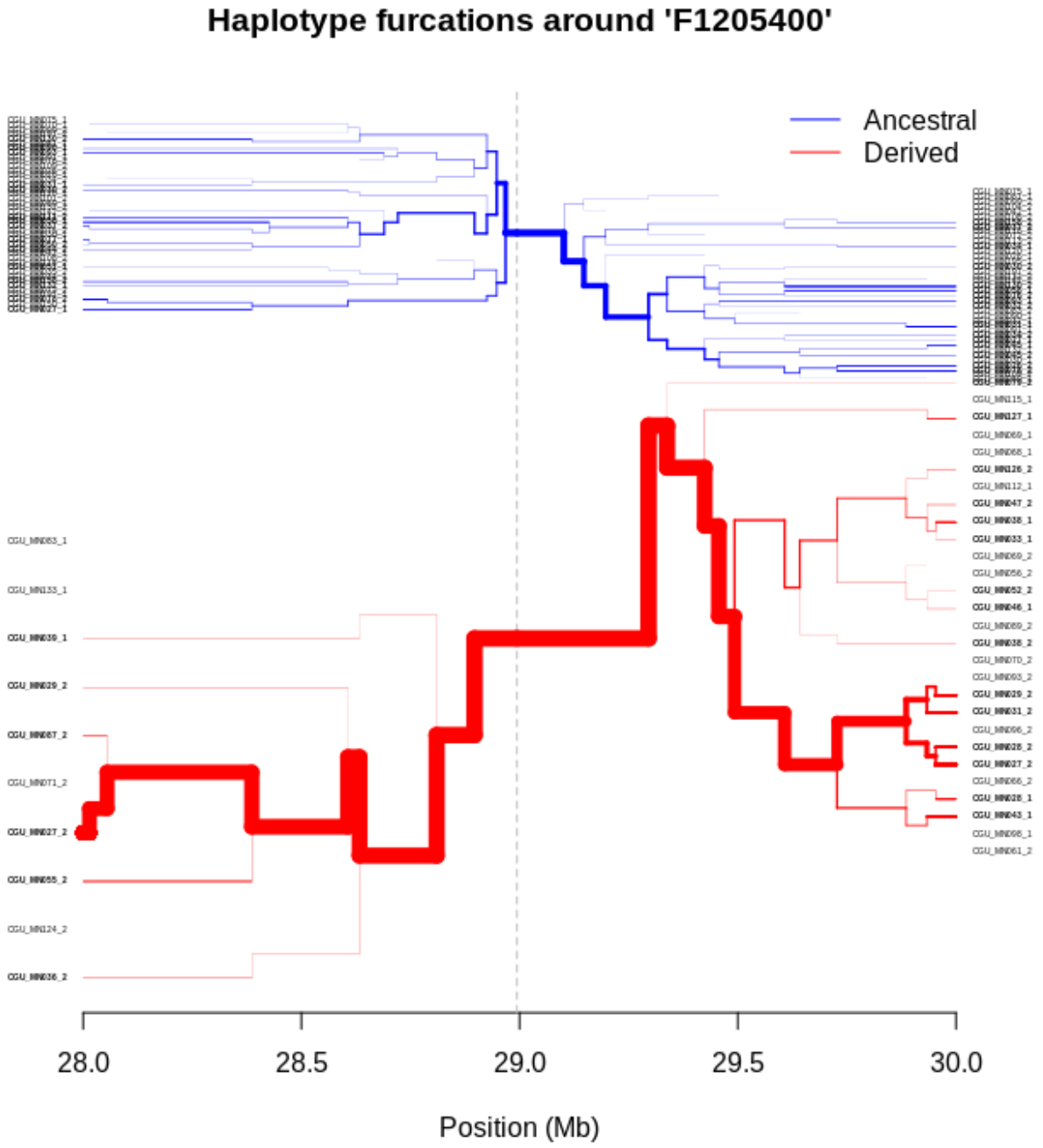
Effect on Linkage

- The association between alleles across individuals breaks down because of recombination
- Recent selection leads to a fast increase in frequency of the selected allele and those at the vicinity (selective sweep)
- Selected haplotypes are frequent but have had less time to recombine with non-selected ones
- Can be detected by methods based on haplotype extension (Sabeti et al. 2006, see methods such as rehh, in the workshop). OmegaPlus is also based on LD.
- Requires phasing to be efficient.



Effect on Linkage

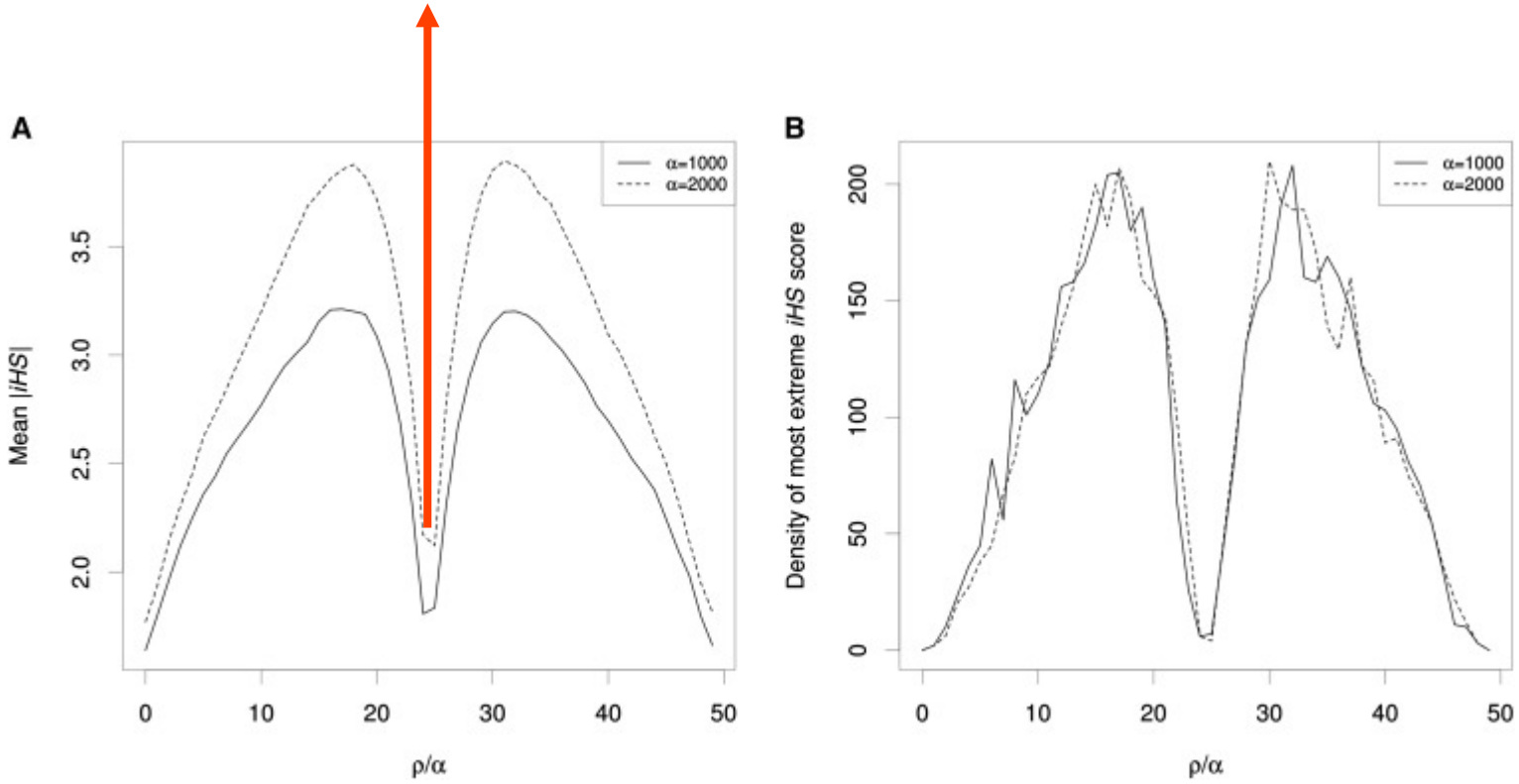
- The association between alleles across individuals breaks down because of recombination
- Recent selection leads to a fast increase in frequency of the selected allele and those at the vicinity (selective sweep)
- Can be detected by methods based on haplotype extension (e.g. iHS score, in rehh, see workshop)
- Possible to compare extension across populations instead of within (XP-EHH, Rsb).



Effect on Linkage

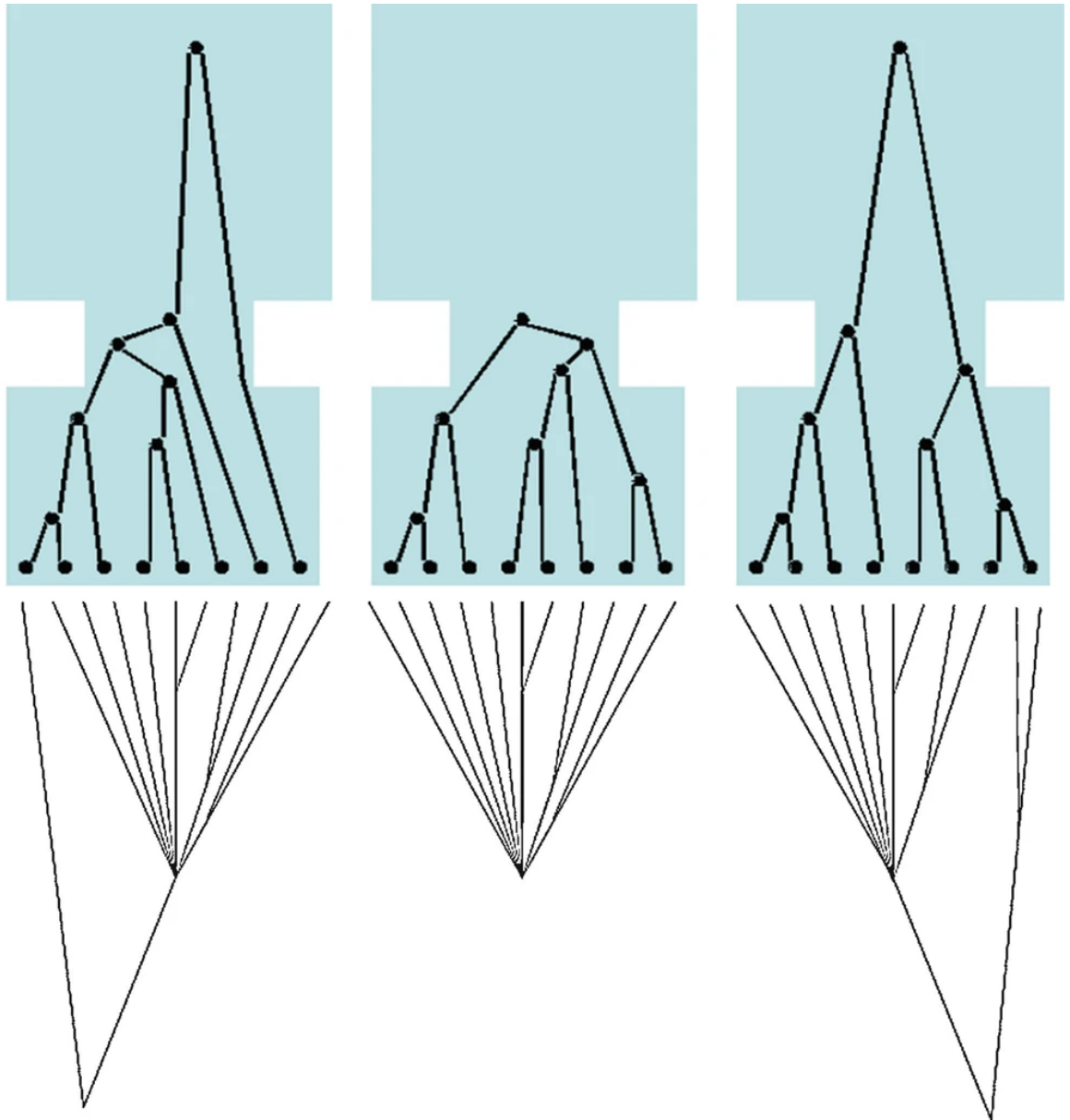
- The association between alleles across individuals breaks down because of recombination
- Recent selection leads to a fast increase in frequency of the selected allele and those at the vicinity (selective sweep)
- Can be detected by methods based on haplotype extension (e.g. rehh, see workshop)

If the sweep is fixed, there is only the derived allele at the centre, and the iHS score can not be computed

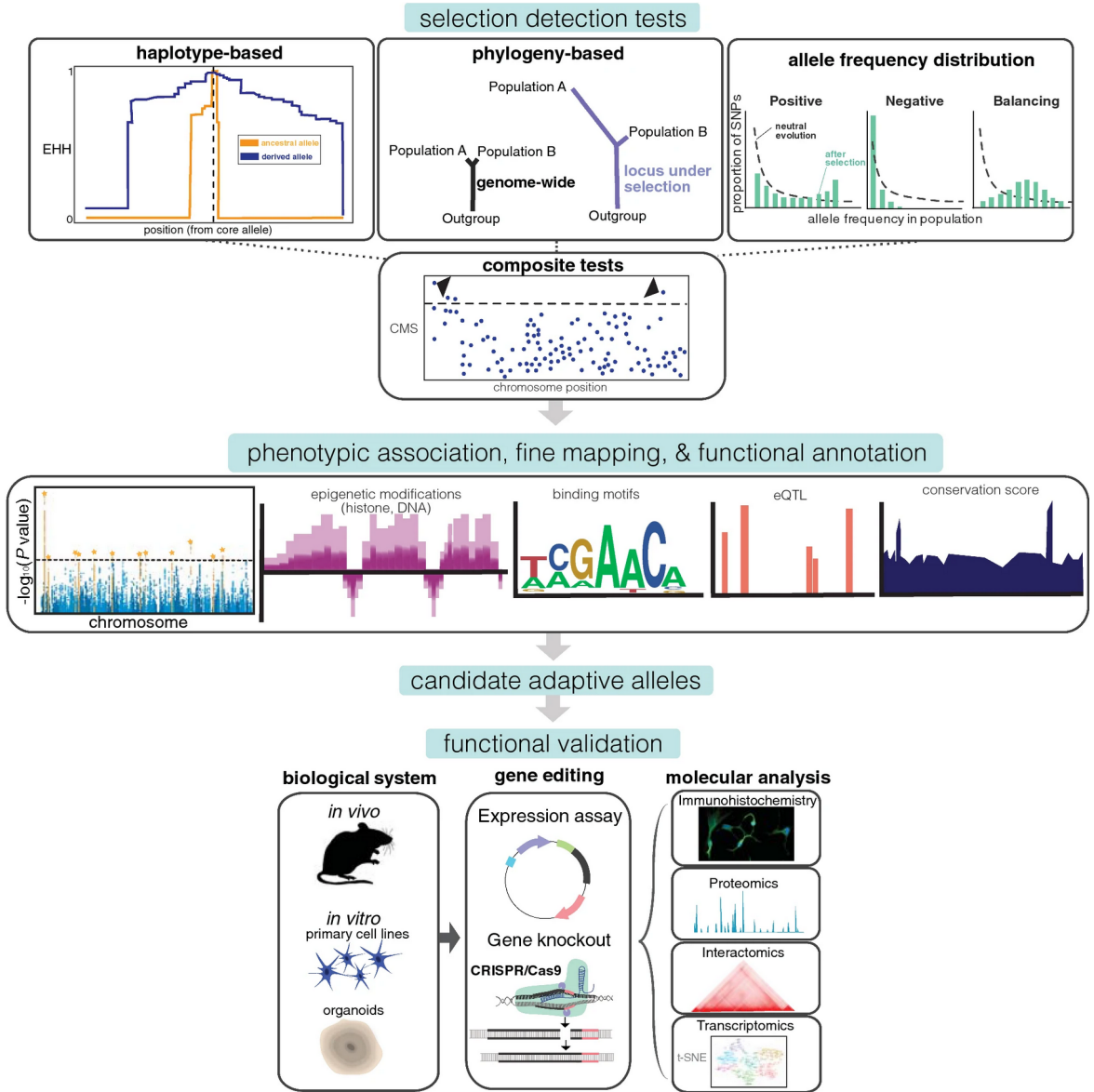


Effect on local genealogies, and a problem...

- But selection acts locally and demography over the whole genome.
- Right?? … …
- Issue of variance...



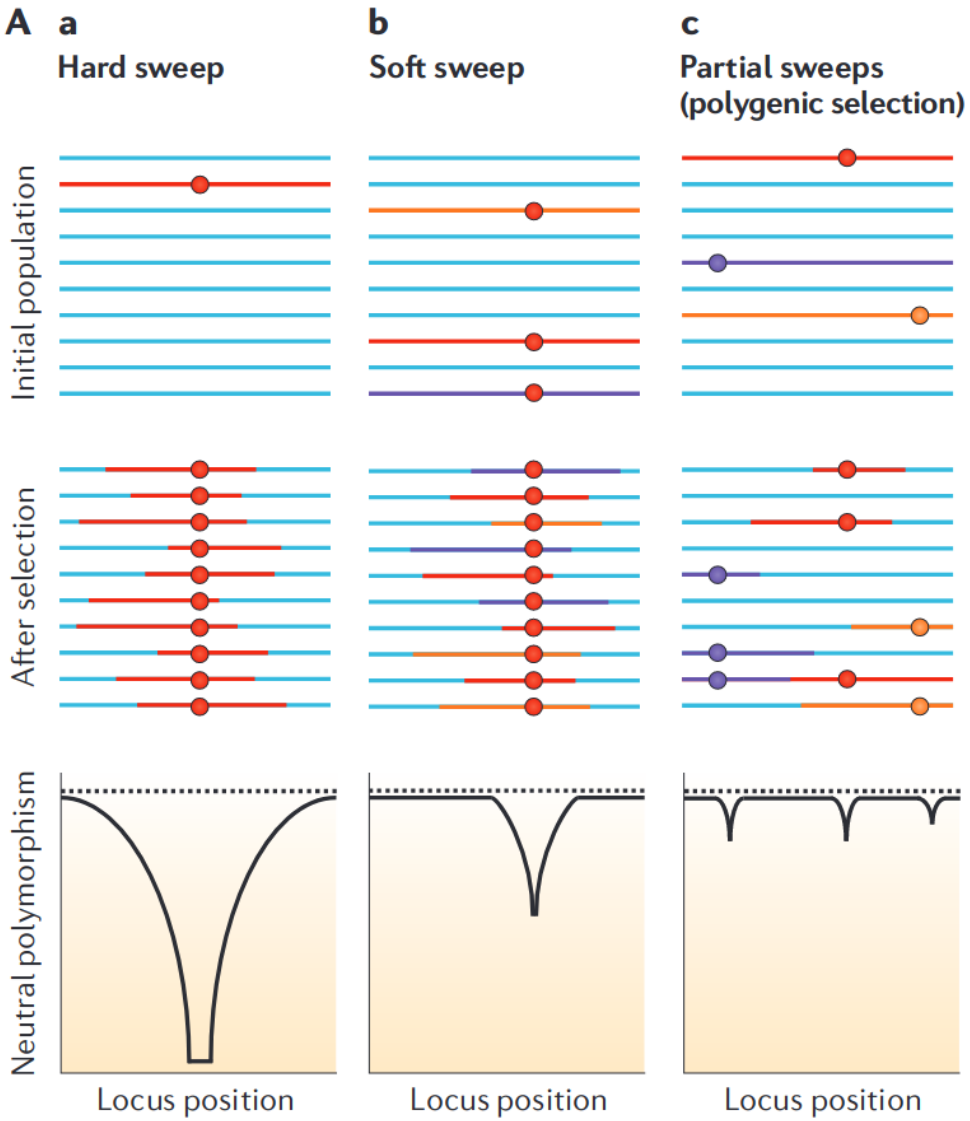
The ideal situation



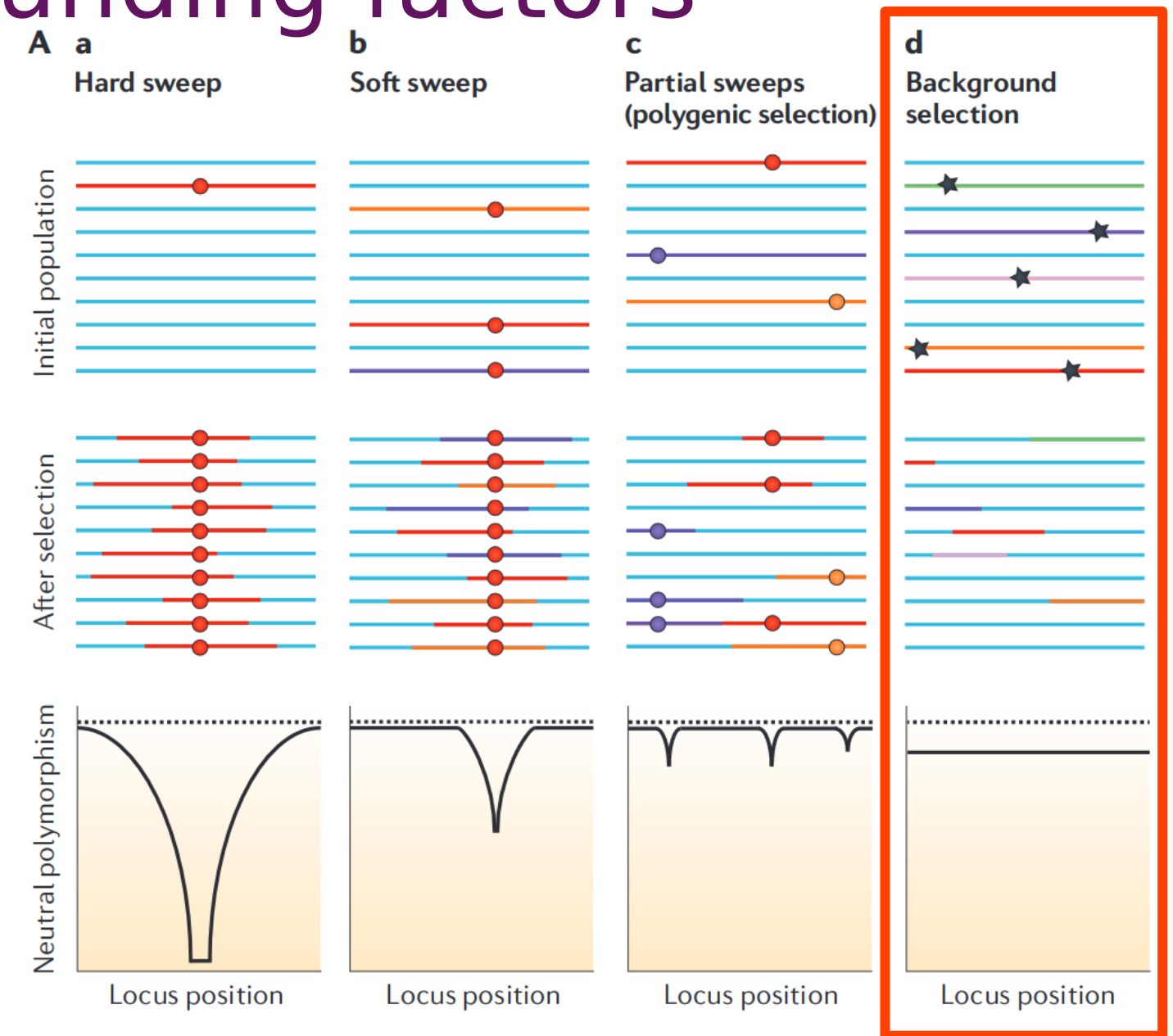
This part can be challenging beyond a few model species
Recent developments (e.g. Nanopore) make it easier

This part is not yet fully doable for many species
One can however use extensive databanks from model species (e.g. bank of mutants in Arabidopsis).
Risk of storytelling.
Risk of lack of functional conservation

More complicated. Soft and partial sweeps

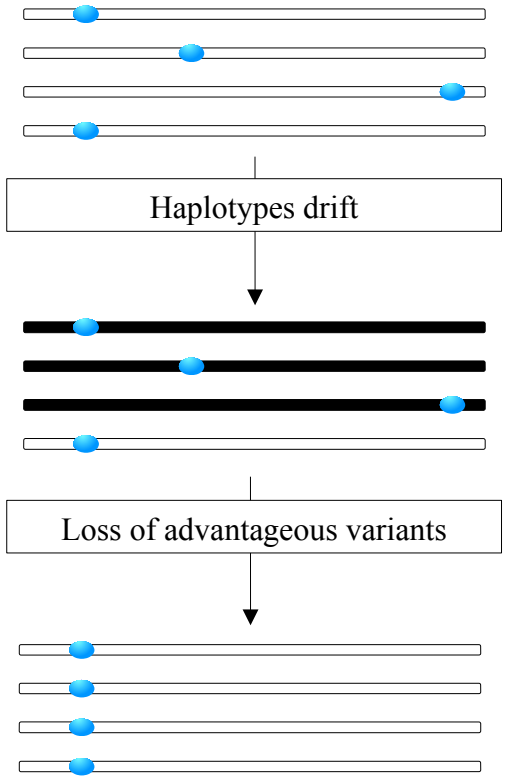


Confounding factors

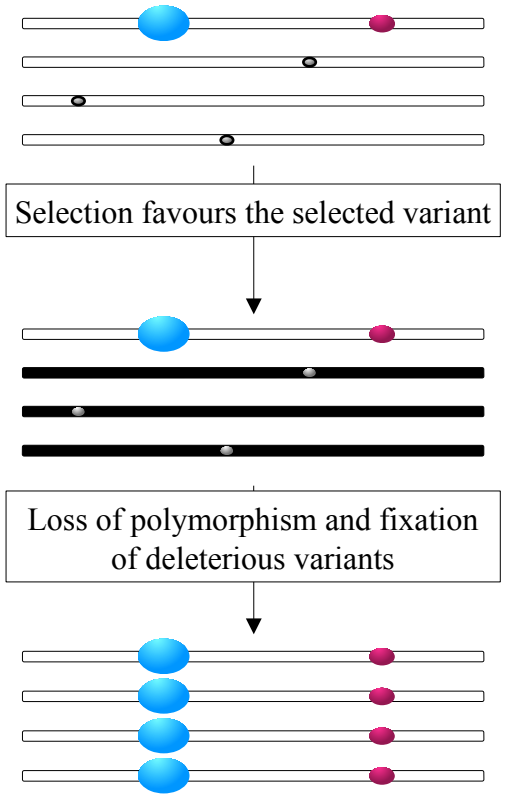


‘Linked’ selection

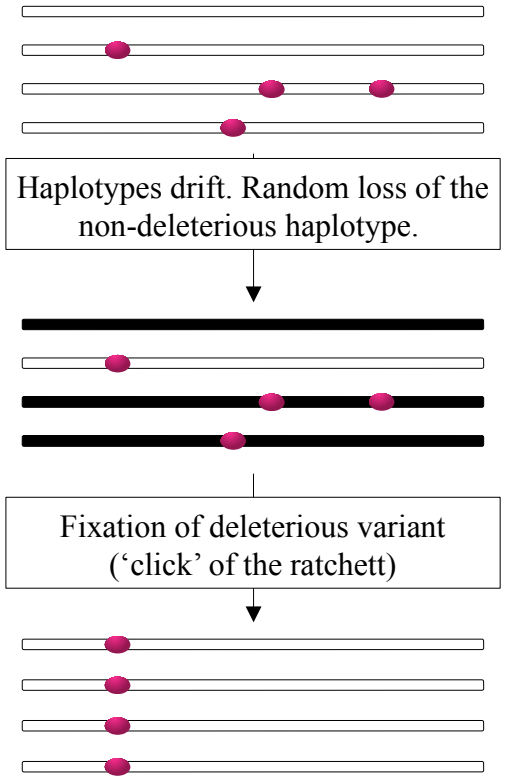
Hill-Robertson interference
(Hill & Robertson 1966)



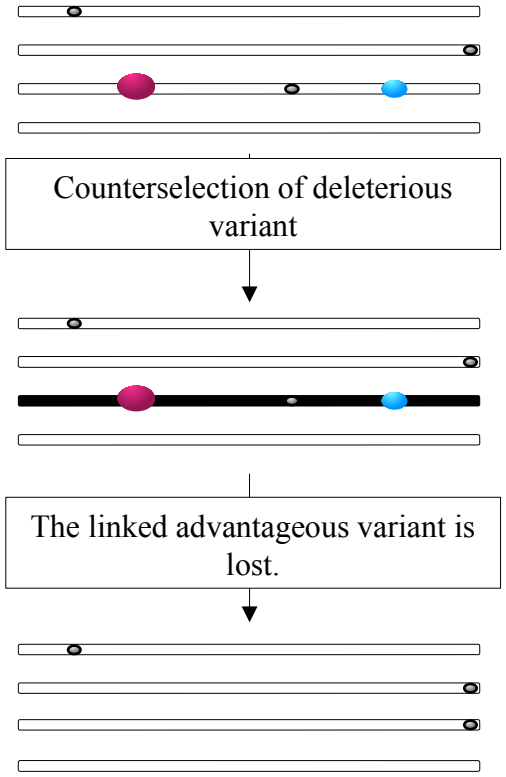
Selective sweep
(Maynard-Smith & Haig 1974)



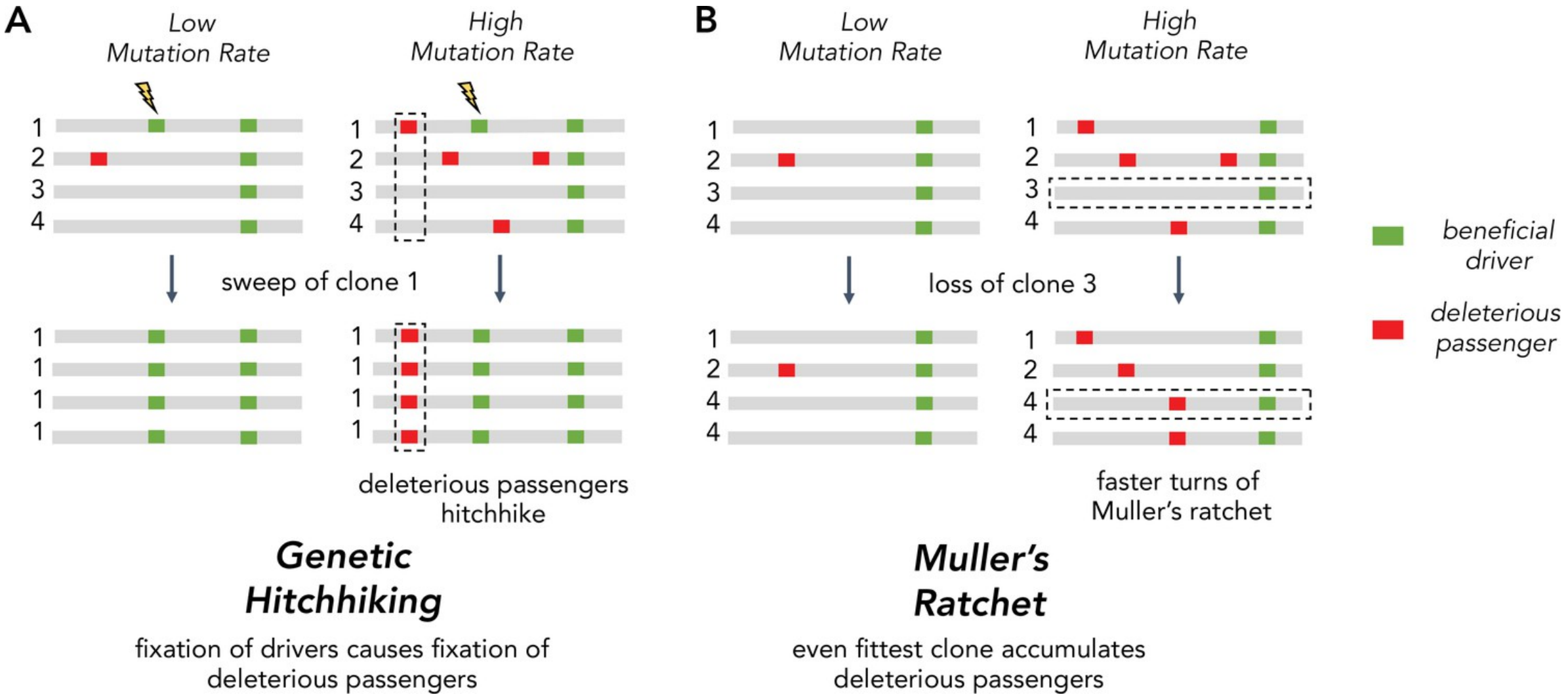
Müller’s ratchett
(Muller 1932)



Background selection
(Charlesworth et al. 1993)



‘Linked’ selection

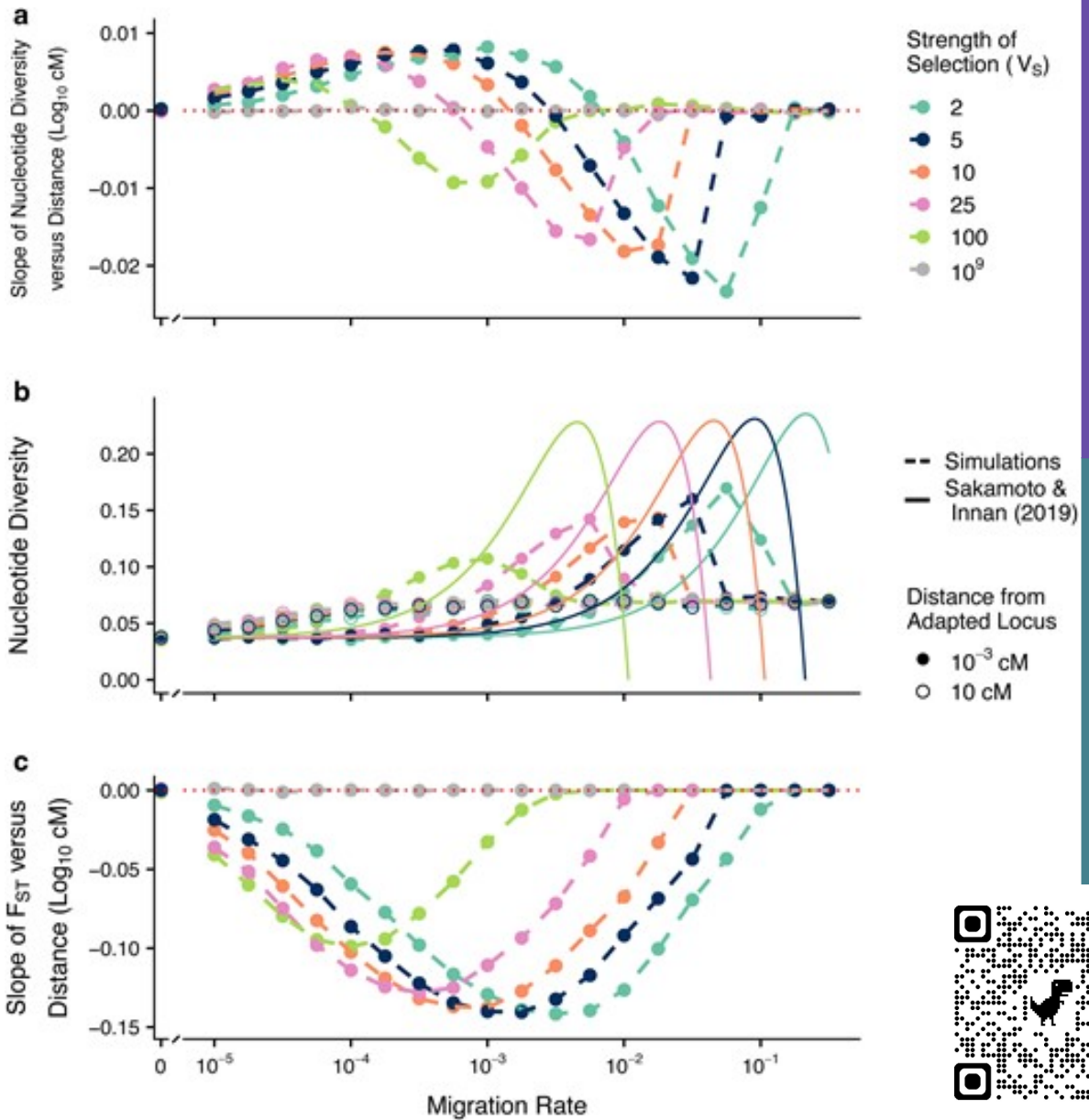


Note that we consider populations that sexually reproduce, but if recombination is locally low relative to mutation, we converge to this clonal case

https://doi.org/10.7554/eLife.67790

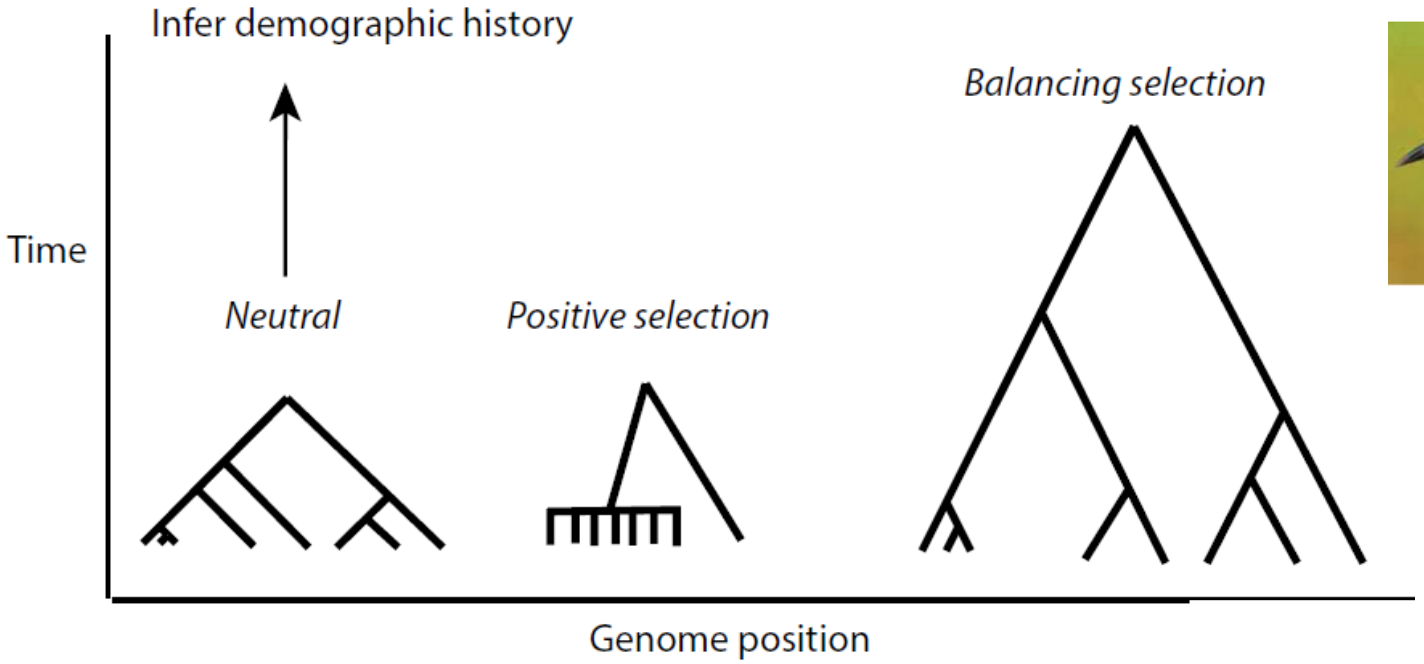
Confounding factors

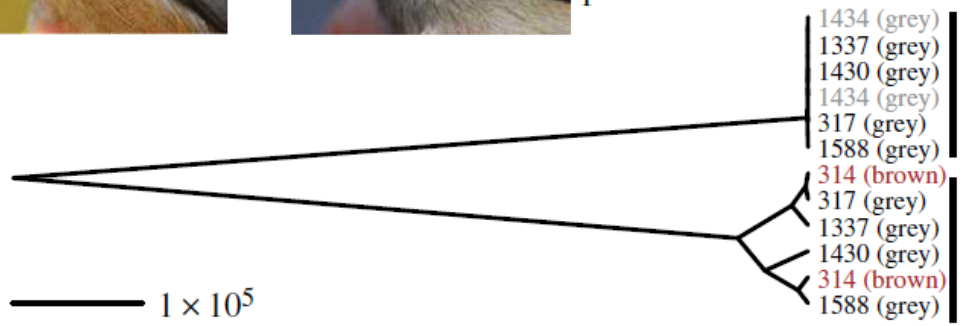
- Local adaptation might show more complicated patterns
- This is due to the interplay between selection, recombination, and migration
- Recent simulation studies suggest that both peaks and throughs in diversity are possible
- Think of the scale at which the sweep may have happened.
- Whenever possible, run simulations to get a sense of what you may expect/detect.



Reconstruct past genealogies and population structure

- Coalescence rate and topology will change if there is selection at a gene
- Genealogies at distant loci are less similar than ones at closely linked loci (because of recombination breaking associations between alleles within a chromosome)

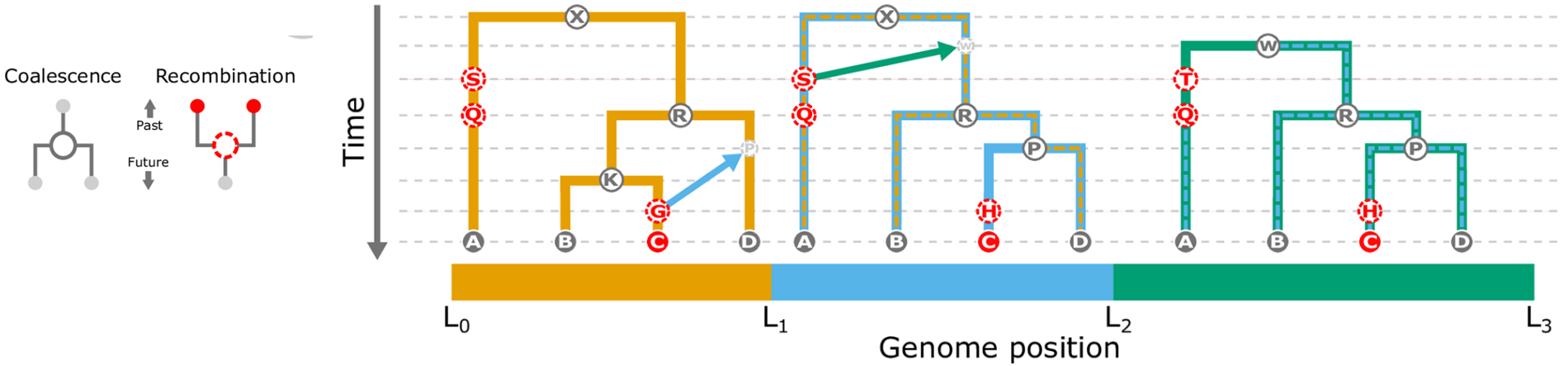




Bourgeois et al. 2017 RSOS

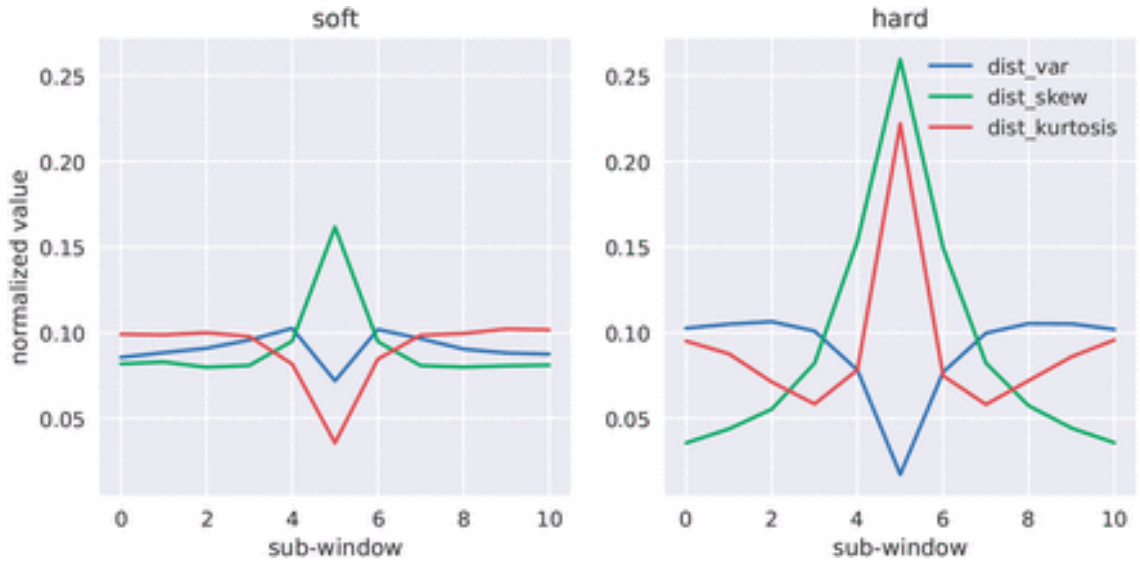
Reconstruct past genealogies and population structure

- One can reconstruct genealogies using polymorphisms along the genome
- The rate at which new branches merge from present to past is inversely correlated with population size (small populations -> high rates of **coalescence**)
- Local deviations in genealogies and recombination can be fully described (ARGWeaver, Relate)
- Can be used to reconstruct allele frequency over time and infer *s* (CLUES2)



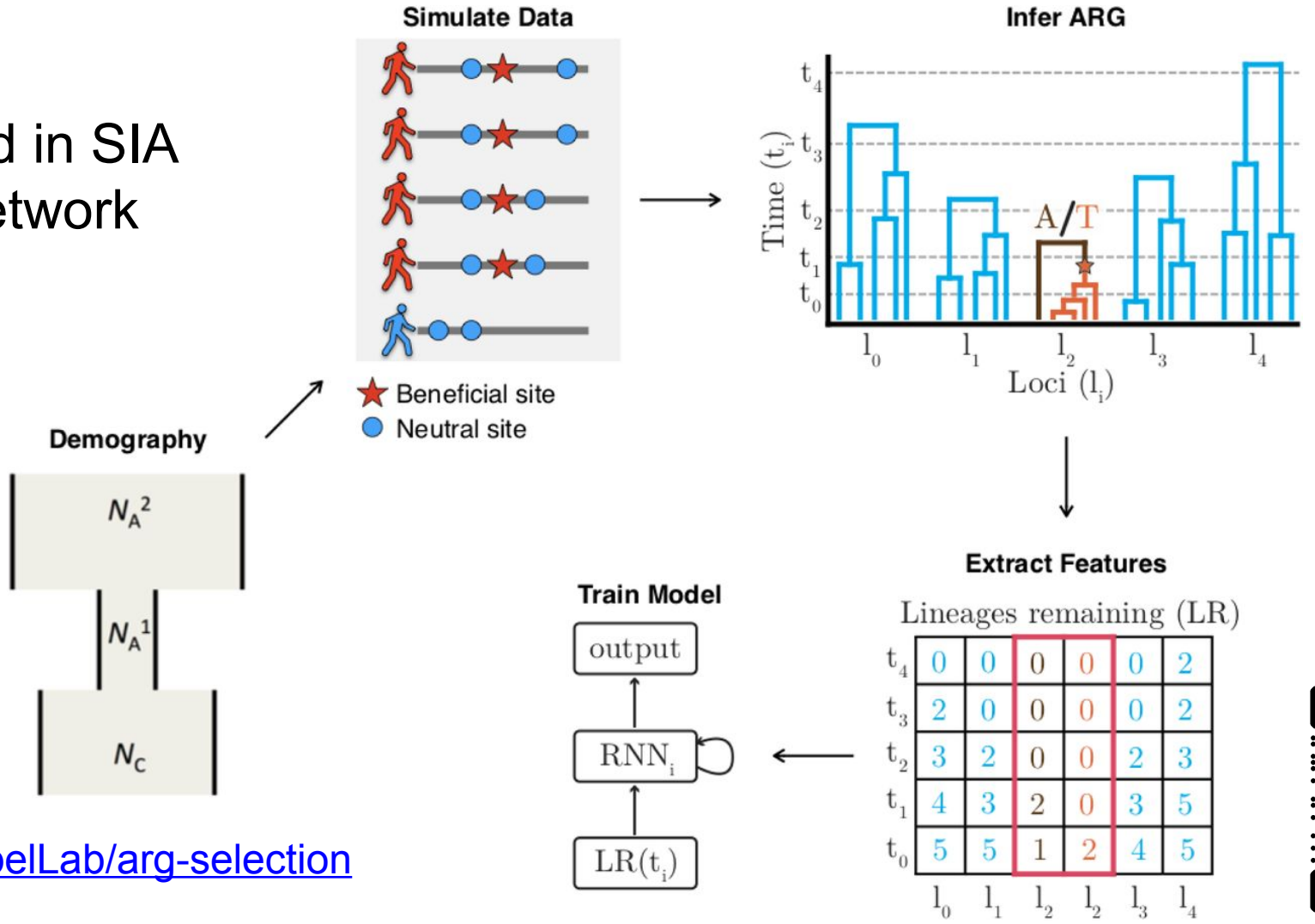
Integration of signals: machine learning (summary statistics)

- Machine and deep learning share a principle with ABC (and they overlap)
- We simulate neutral and selected datasets, obtain summary statistics, and compare them to observations through a trained algorithm.
- Possible advantage compared to ABC: fewer issues with the choice of summary statistics.
- diploS/HIC and S/HIC



Integration of signals: machine learning (Ancestral Recombination Graphs)

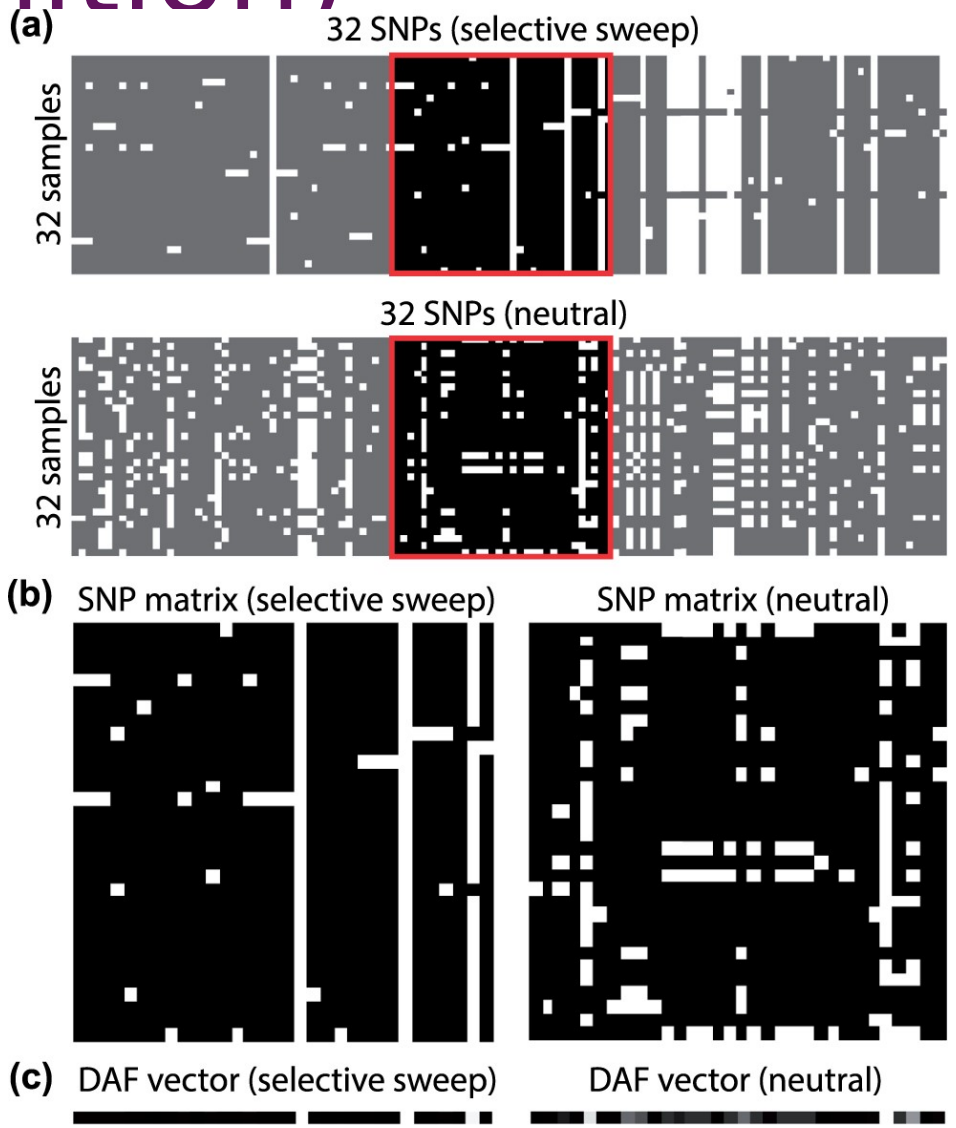
Method implemented in SIA
Recurrent Neural Network



https://github.com/CshlSiepelLab/arg-selection

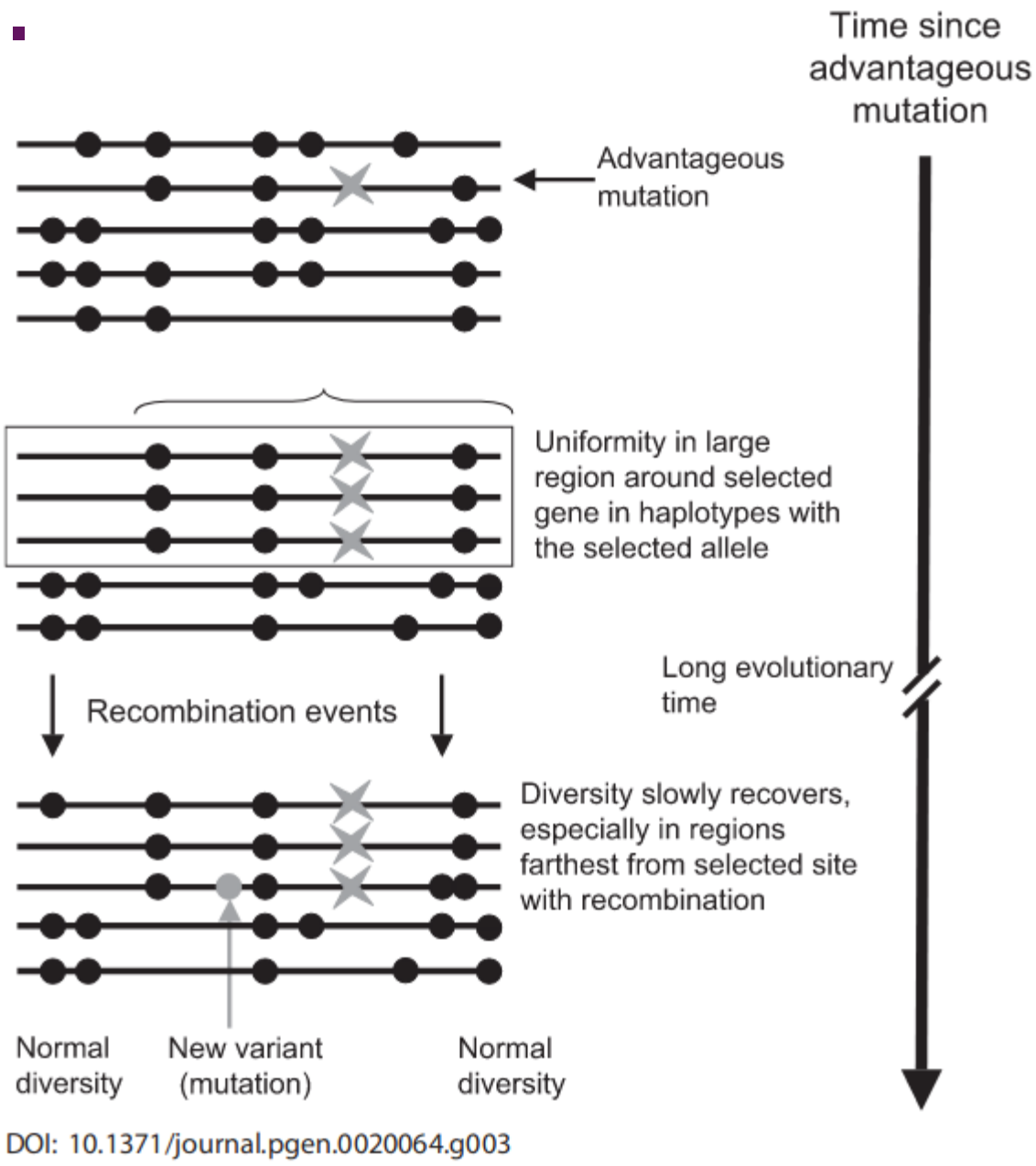
Integration of signals: machine learning (image recognition)

- Direct image recognition is also seducing in principle to get rid of summary statistics
- Take the alignment and train the algorithm to recognize images that fit with a sweep
- Fast- NN (takes the vector of frequencies as an image).
- SweepNet (takes the alignment as an image).



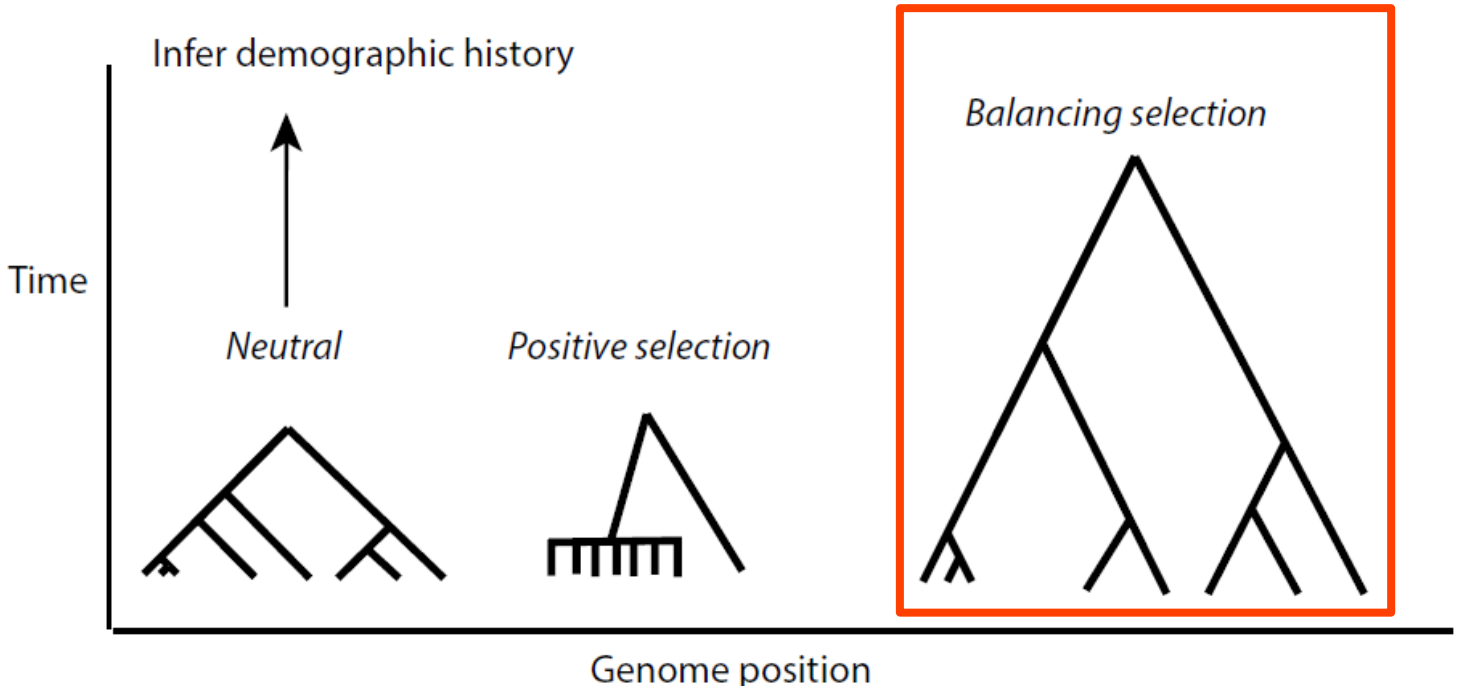
Balancing selection.

- This is an umbrella term, covering processes that maintain genetic diversity over time at a locus. Especially interesting to people working on diversifying processes, such as host-parasite interaction, or self-incompatibility.
- Think of a partial sweep that never gets fixated.
- The main challenge here is the loss of linkage over evolutionary times, which may make the signal very narrow.



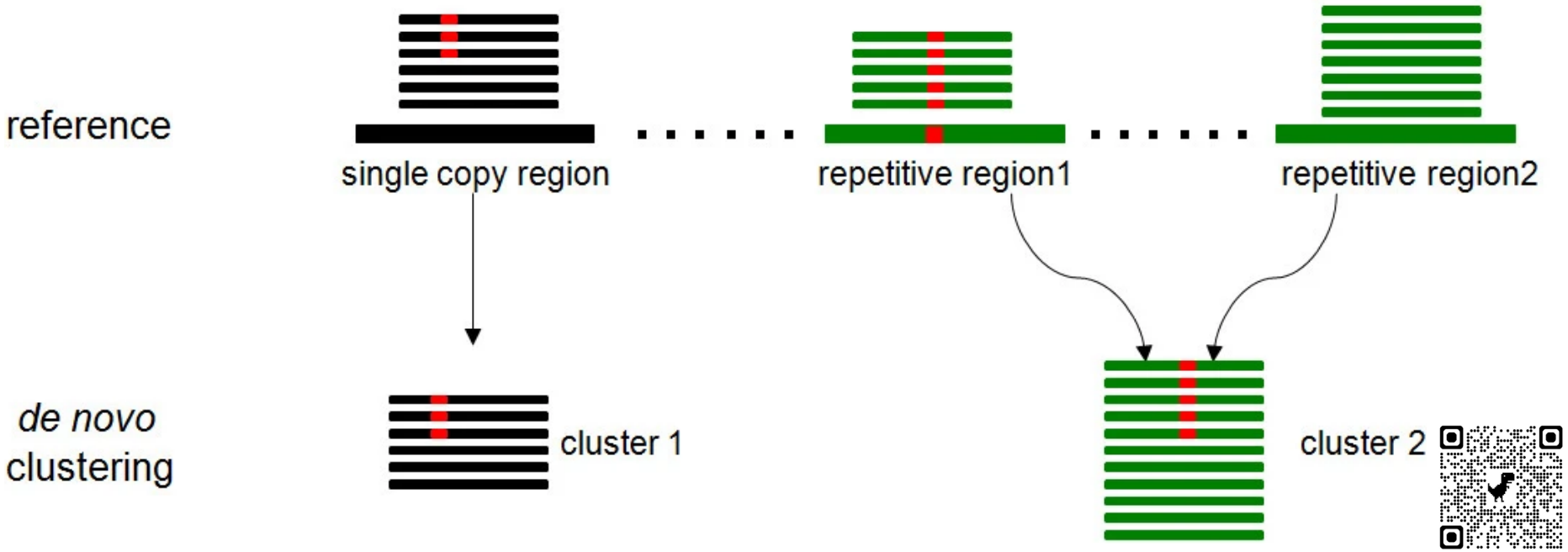
Long-term balancing selection.

- The genealogies near a locus under long-term balancing selection should be older. ARGs can be useful here.
- However, this is also dangerous: misalignment in your (short read) data along undetected paralog loci may lead to false signals
- Very old balancing selection can also involve very divergent haplotypes, making alignment impossible.



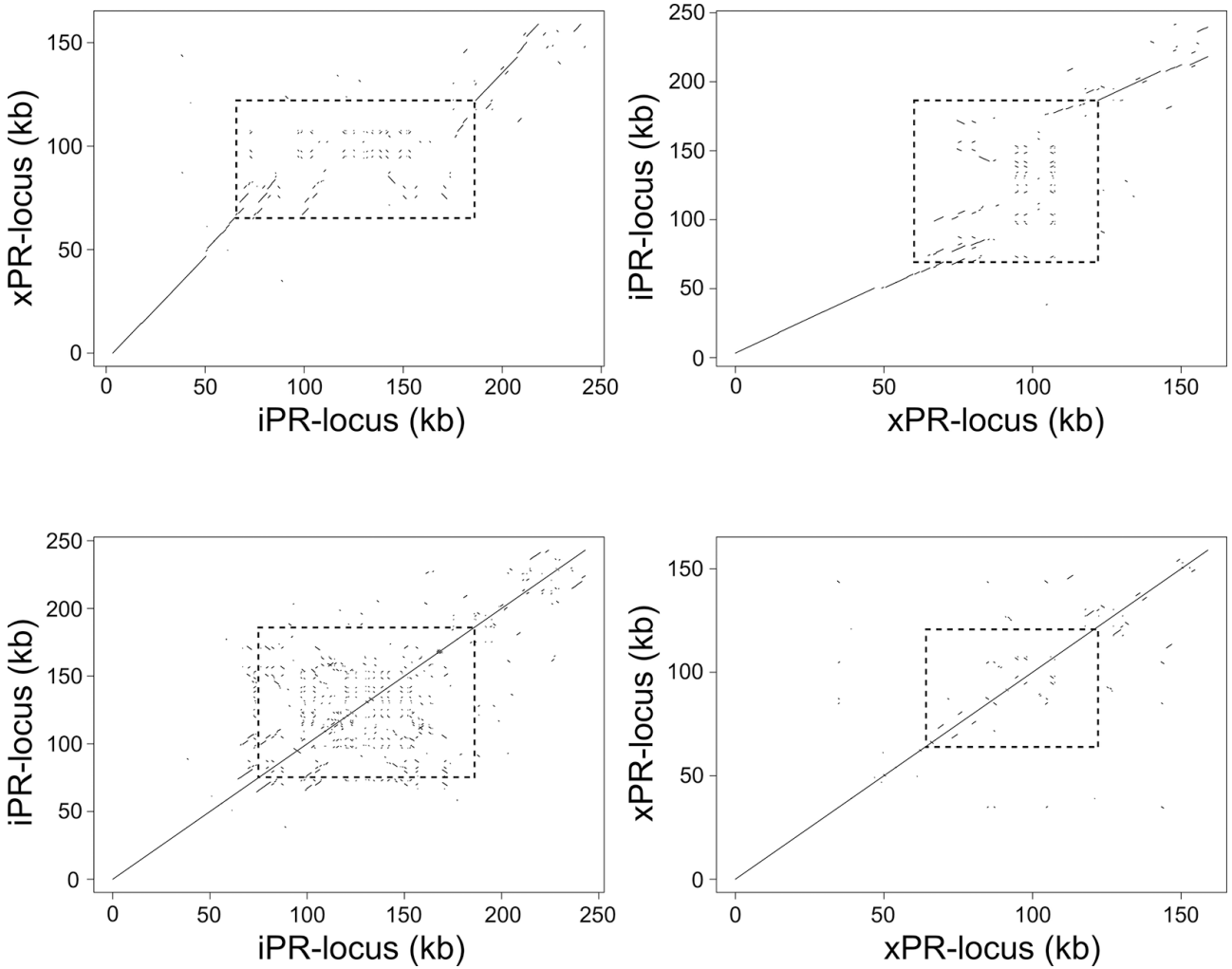
Long-term balancing selection.

- The genealogies near a locus under long-term balancing selection should be older. ARGs can be useful here.
- However, this is also dangerous: misalignment in your (short read) data along undetected paralog loci may lead to false signals
- Very old balancing selection can also involve very divergent haplotypes, making alignment impossible.



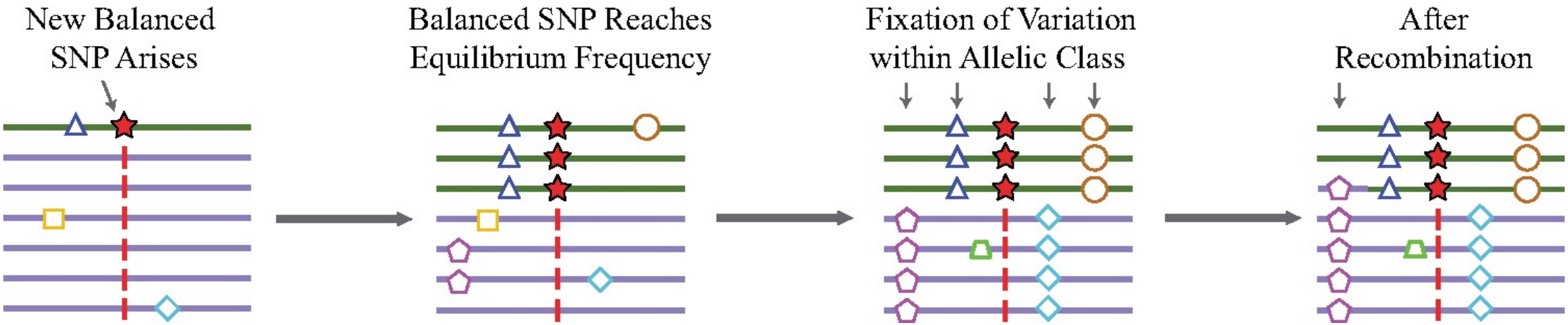
Long-term balancing selection.

- The genealogies near a locus under long-term balancing selection should be older. ARGs can be useful here.
- However, this is also dangerous: misalignment in your (short read) data along undetected paralog loci may lead to false signals
- Very old balancing selection or gene conversion can also lead to very divergent haplotypes, making alignment impossible.



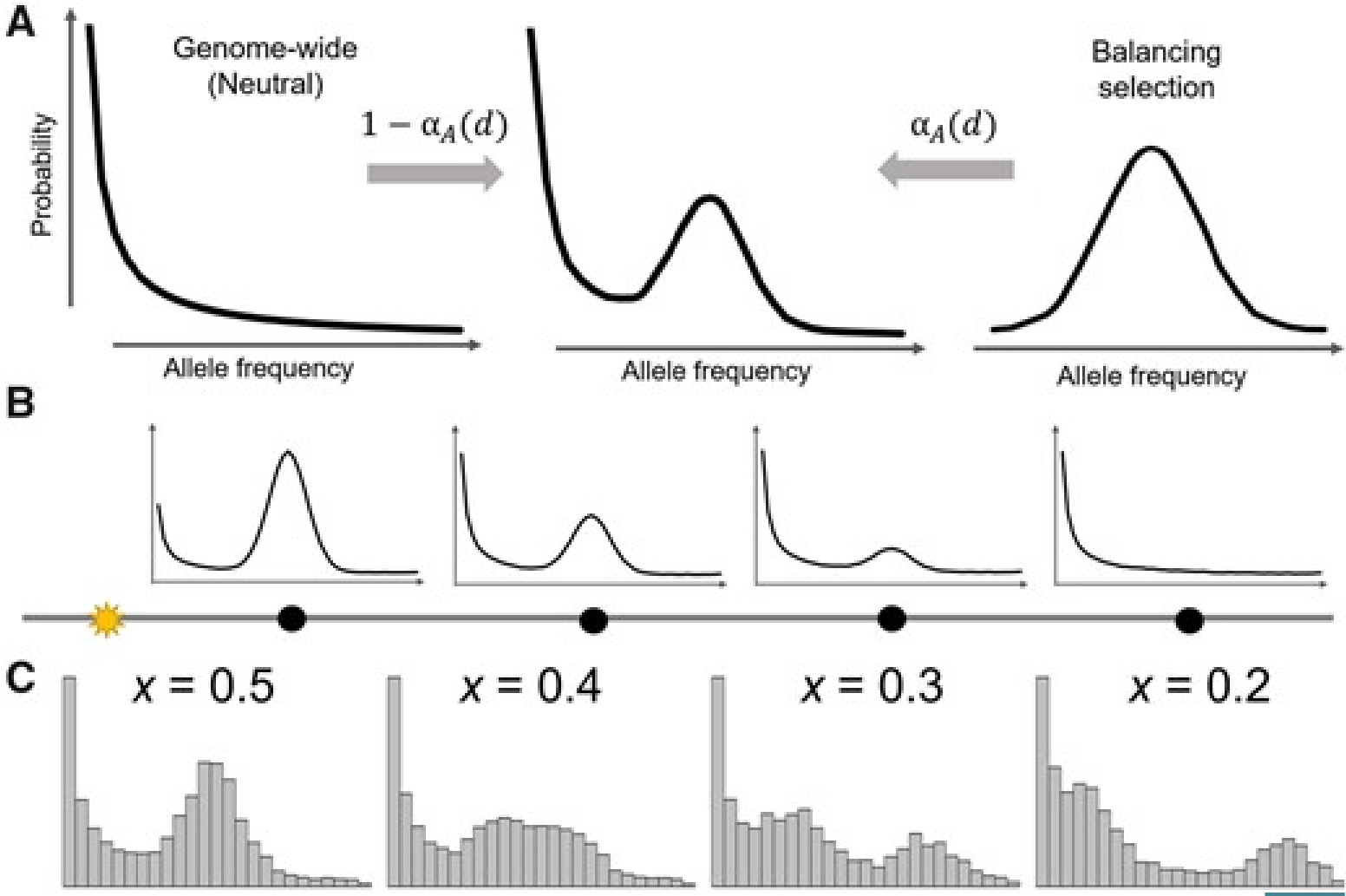
Detecting Balancing selection: LD.

- The β statistic method can work without specifying ancestral and derived alleles (useful in non-model species)
- It is fast.
- Same issues of mappability of short reads to take into account.



Detecting Balancing selection: Spectrum.

- BalLeRMix+. Method can work without specifying ancestral and derived alleles (useful in non-model species)
- It is fast. Robust to the choice of window's size.
- Benefits from an outgroup to include divergence with polymorphism.



Detecting Balancing selection: Spectrum.

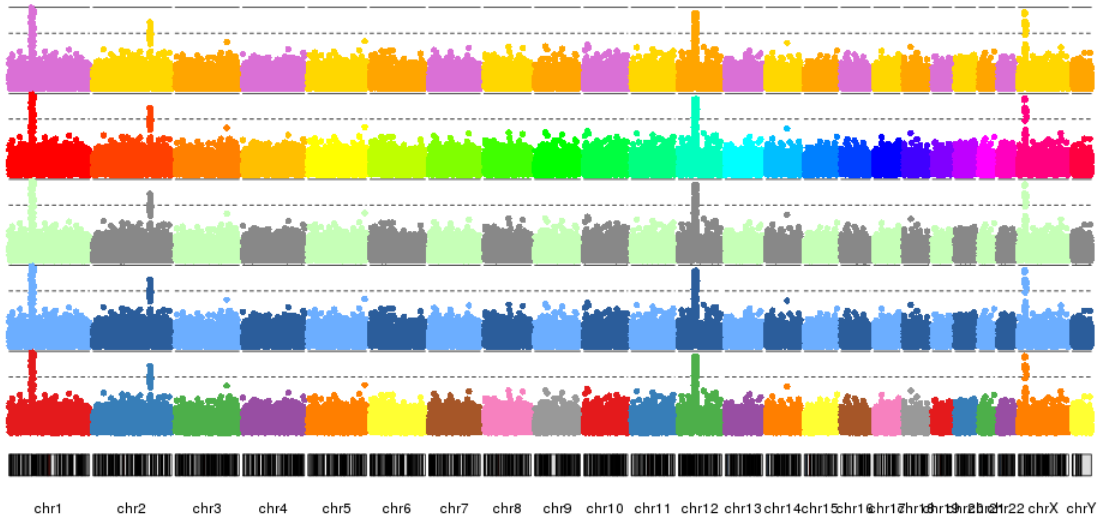
- Those are two relatively user-friendly examples
- For a full list of recent methods (2023), see:
- Bárbara D Bitarello, Débora Y C Brandt, Diogo Meyer, Aida M Andrés, Inferring Balancing Selection From Genome-Scale Data, Genome Biology and Evolution, Volume 15, Issue 3, March 2023, evad032, https://doi.org/10.1093/gbe/evad032

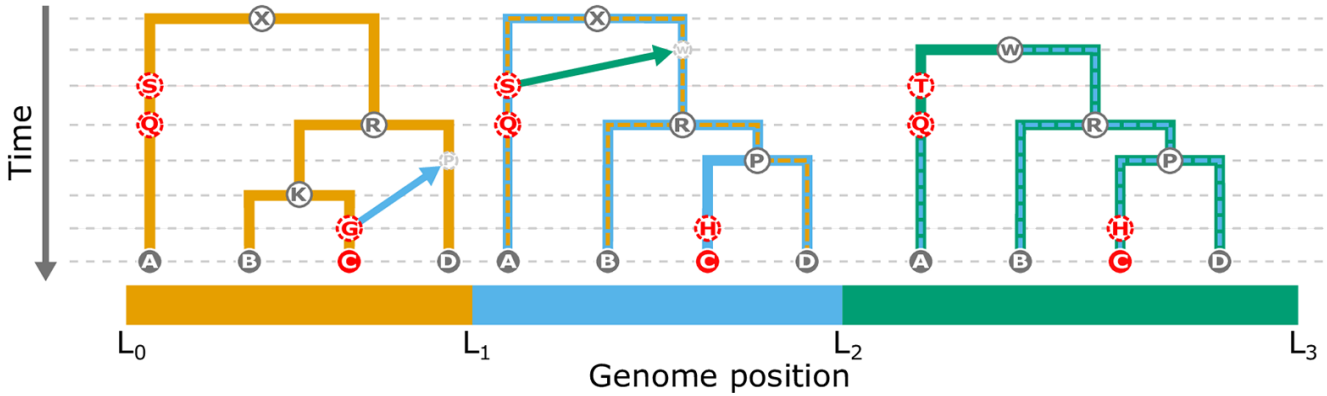
Table 4

Available Software that Implements Tests for BLS

Software	Methods or Tests Implemented	URL
Betascan (python)	$\beta^{(1)*}; \beta^{(1)}; \beta^{(1)*}_{std}; \beta^{(2)}; \beta^{(2)}_{std}$	https://github.com/ksiewert/BetaScan
BALLET (C)	$T_1; T_2$	http://degiorgiogroup.fau.edu/ballet.html
NCD-Statistics (R)	$NCD1; NCD2$	https://github.com/bbitarello/NCD-Statistics
balselr (R)	$NCD1; NCD2$	https://github.com/bbitarello/balselr
MuteBaSS (python)	$NCD2_{trans}$; $NCD2_{mid,trans}$; $NCD2_{opt,trans}$; HKA_{trans}	https://github.com/bioXiaoheng/MuteBaSS
MULLET (C)	$T_{1,trans}; T_{2,trans}$	http://degiorgiogroup.fau.edu/mullet.html
BalLeRMix (python)	B_0 ; $B_{0,MAF}$; B_1; B_2; $B_{2,MAF}$	https://github.com/bioXiaoheng/BalLeRMix
BalLeRMix + (python)	B_0; $B_{0,MAF}$; B_1; B_2; $B_{2,MAF}$	https://github.com/bioXiaoheng/BallerMixPlus
BaSe (python)	Artificial and convoluted neural networks.	https://github.com/ulasisik/balancing-selection
ANGSD	Tajima's *D*; Fu & Li's *F*; Fu & Li's *D*; Fay's *H*; Zheng's *H*	https://github.com/ANGSD/angsd
tskit (python)	Tajima's *D*	https://github.com/tskit-dev/tskit
Tsel (R)	T_{sel}	https://blogs.cornell.edu/clarklabblog/clark-lab/software
C code	HKA	https://github.com/alanrogers/hka
C code	HKA	https://github.com/andrewkern/hka
C code	HKA	https://bio.cst.temple.edu/!~tuf29449/hka_manual

Highly polygenic selection.





Implemented in PALM

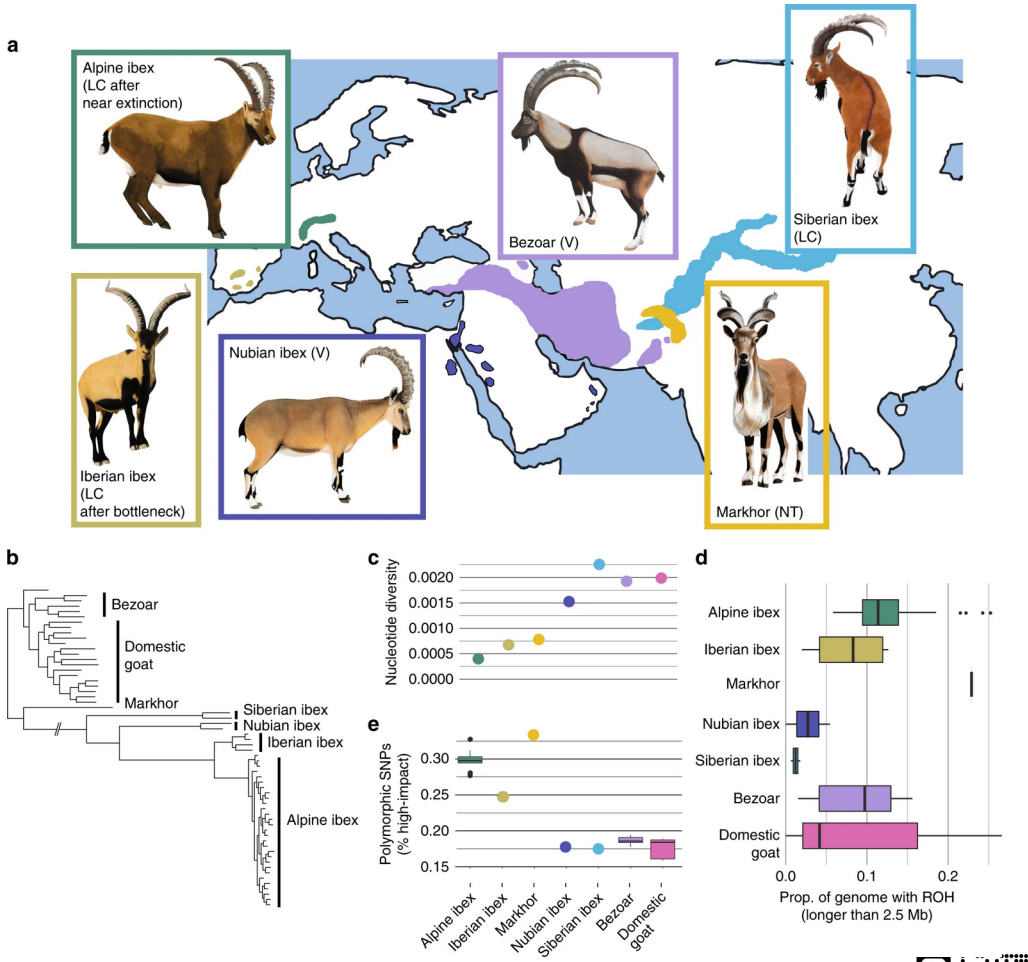
Obtain effect sizes for SNPs associated with a trait of interest

Reconstruct allele trajectories through time and infer selective coefficients

Obtain genome-wide genealogies

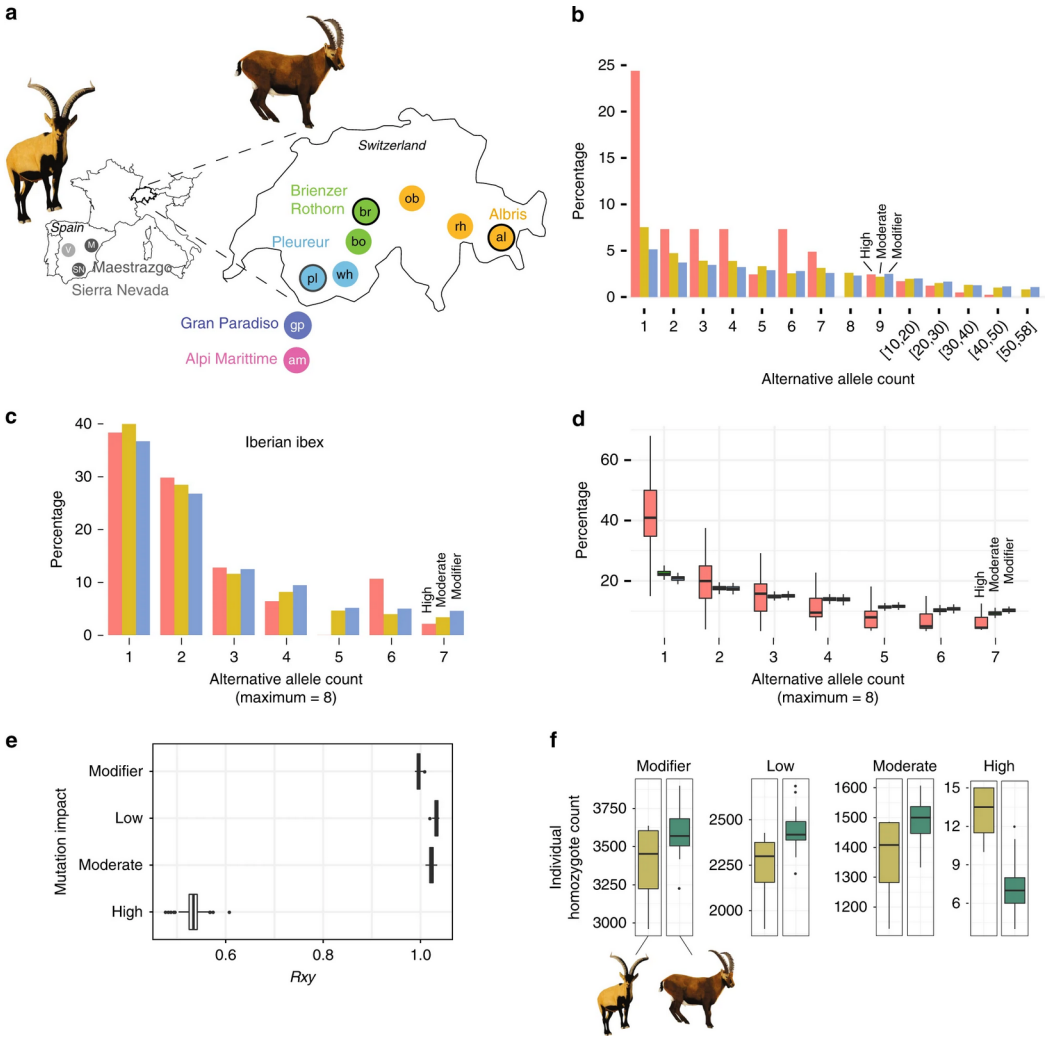
A final note: the concept of mutation load

- Not everything in selection is positive…
- The genomic burden of deleterious variants is being more and more studied in conservation.
- Demographic processes are important.
- May be extremely important in adaptability.



A final note: the concept of mutation load

- Not everything in selection is positive…
- The genomic burden of deleterious variants is being more and more studied in conservation.
- Demographic processes are important.
- May be extremely important in adaptability.



A few questions now...

Which methods or statistics are the most suited to detect ancient balancing selection in a genome scan?

A - Extended Haplotype Homozygosity (EHH)

B - Population Branch Statistic (PBS)

C - β statistic based on LD

D - BalLeRMix based on the allele frequency spectrum

E - Local ancestry painting (e.g., PCAdmix)

A few questions now...

Which methods or statistics are the most suited to detect ancient balancing selection in a genome scan?

A - Extended Haplotype Homozygosity (EHH)

B - Population Branch Statistic (PBS)

C - β statistic based on LD

D - BalLeRMix based on the allele frequency spectrum

E - Local ancestry painting (e.g., PCAdmix)

A few questions now…

What pattern is typically expected near a locus under a recent hard selective sweep?

A - Low genetic diversity near the selected locus

B - Increase in heterozygosity near the selected locus

C - Extended haplotype homozygosity around the selected allele

D - An increase in Tajima’s D values

E - Decrease in linkage disequilibrium near the selected locus

A few questions now...

What pattern is typically expected near a locus under a recent hard selective sweep?

A - Low genetic diversity near the selected locus

B - Increase in heterozygosity near the selected locus

C - Extended haplotype homozygosity around the selected allele

D - An increase in Tajima's D values

E - Decrease in linkage disequilibrium near the selected locus

A few questions now...

Which of the following statements about GWAS is correct?

A - Chi-squared tests are ideal for continuous traits

B - Logistic regression is used for binary phenotypes

C - GWAS can be affected by population structure

D - Bonferroni correction decreases the false negative rate

E - GWAS assumes strong selection at each tested locus

A few questions now...

Which of the following statements about GWAS is correct?

A - Chi-squared tests are ideal for continuous traits

B - Logistic regression is used for binary phenotypes

C - GWAS can be affected by population structure

D - Bonferroni correction decreases the false negative rate

E - GWAS assumes strong selection at each tested locus