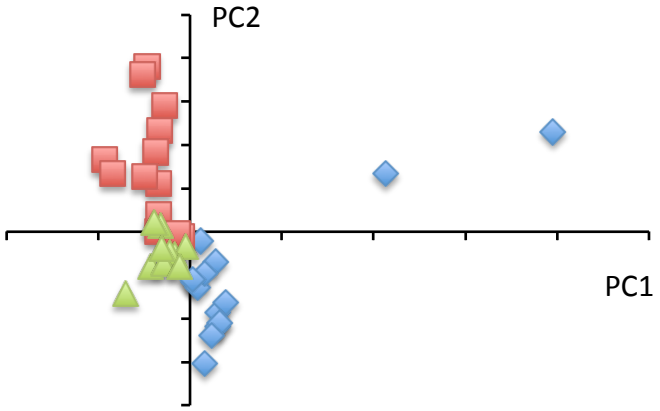
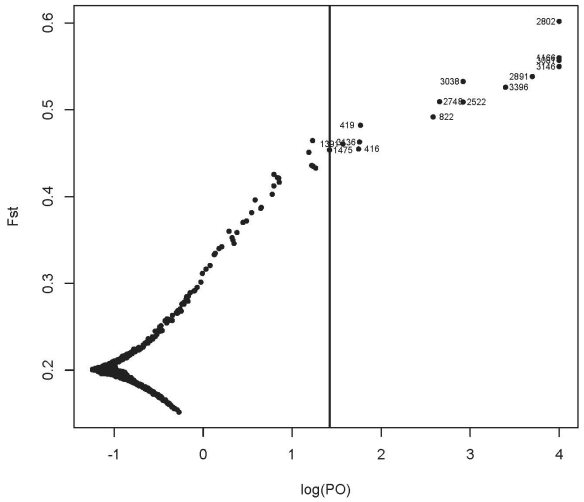
Analyses of genotype datasets

- There are many applications for a SNP/genotype dataset!

- In exercises tomorrow morning, we will focus on 2 different approaches:
  1. PCA – used to find large-scale differences between individuals/populations across the genome or transcriptome.
  2. F_{ST} outlier scan – used to find individual SNPs that are more divergent between populations than expected in a neutral model and thus potentially acted upon by differential natural selection.





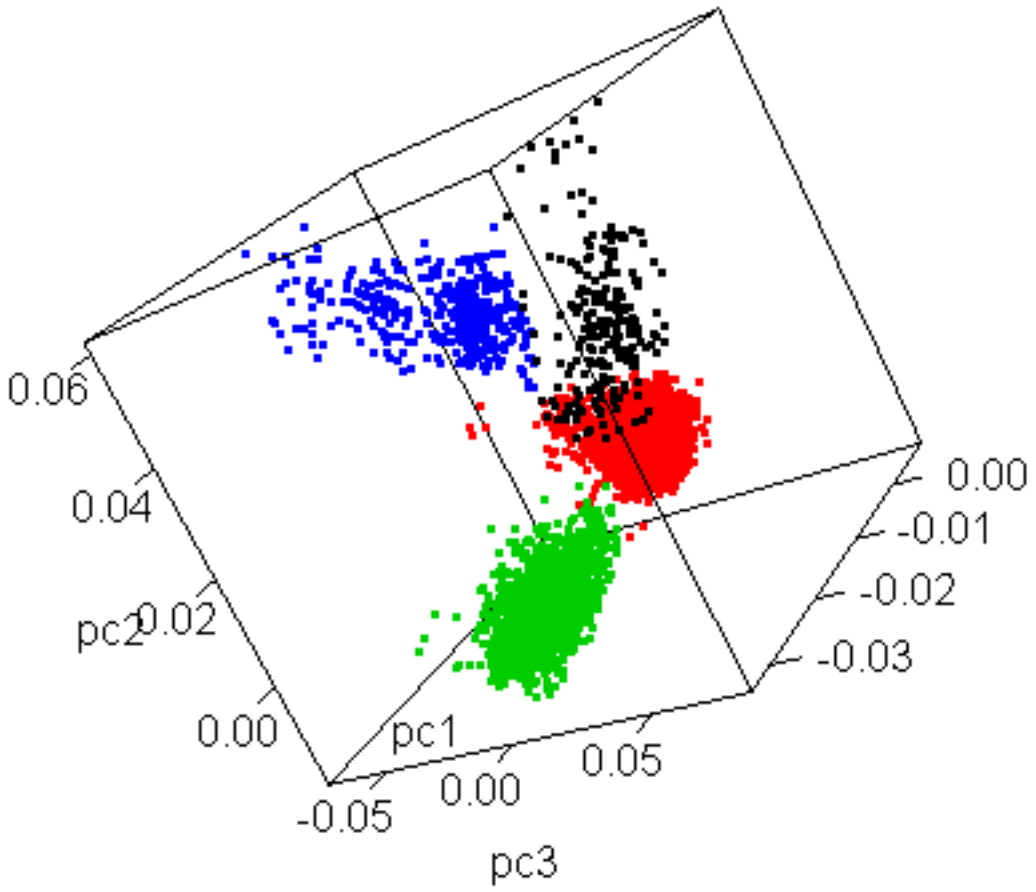
Principal components analysis

Reduces all the variation within each individual to one data point in (N-1) dimensional space (N is number of individuals)

Each dimension is uncorrelated

PC1 explains most of the variation, then PC2 etc..

Individuals more similar to each other group together.

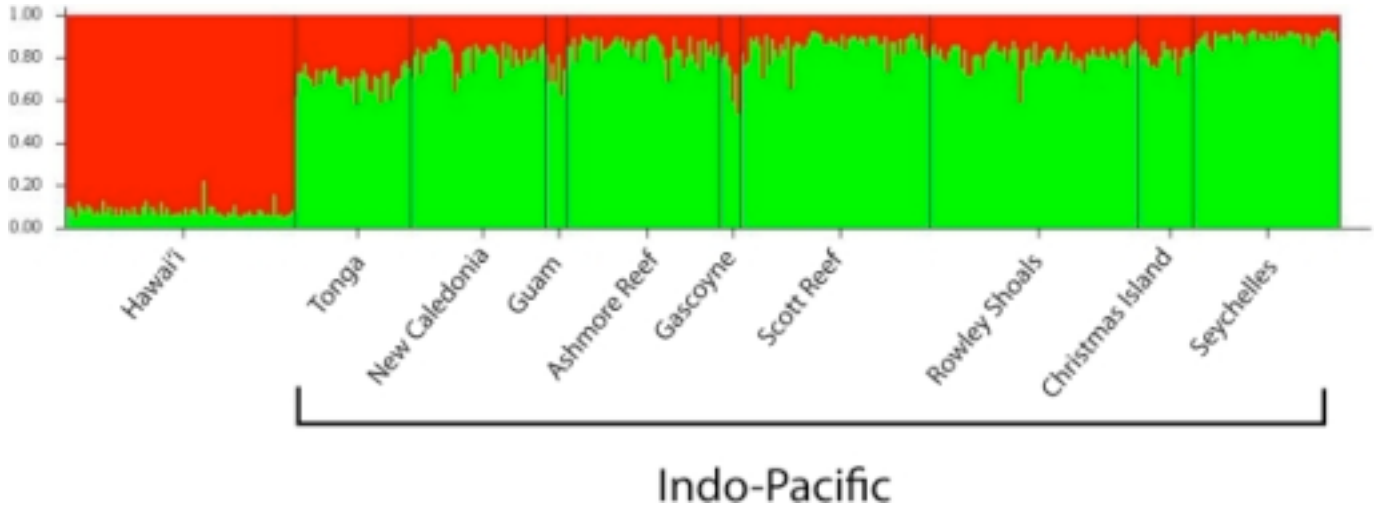


STRUCTURE

Bayesian clustering of individuals, maximizing HWE

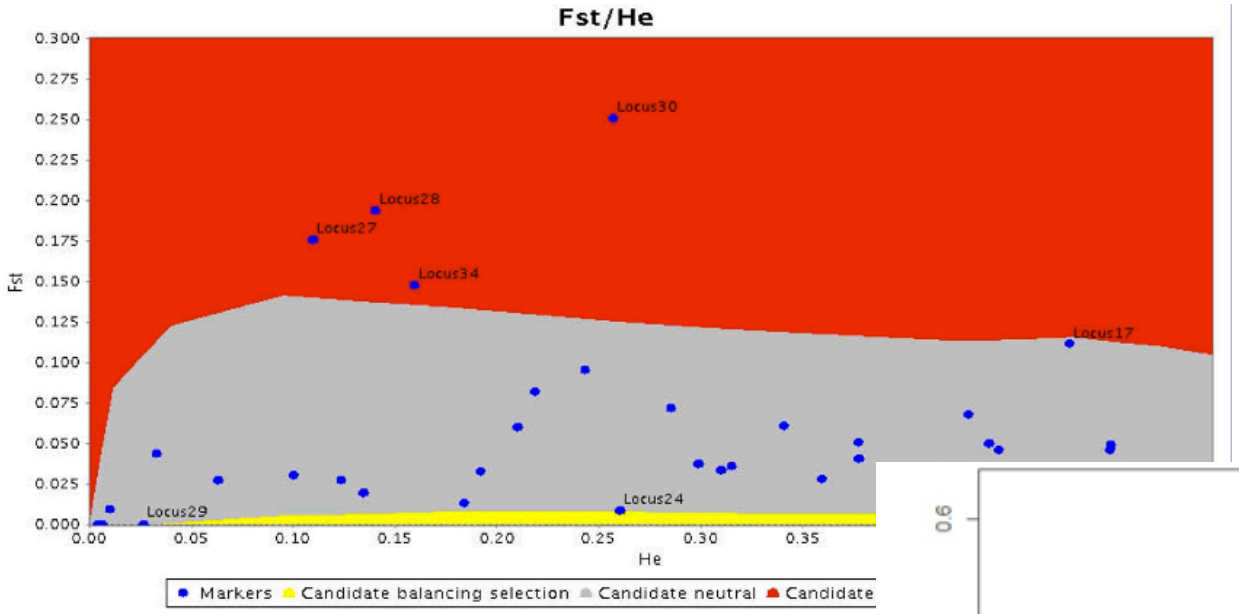
Can model introgression

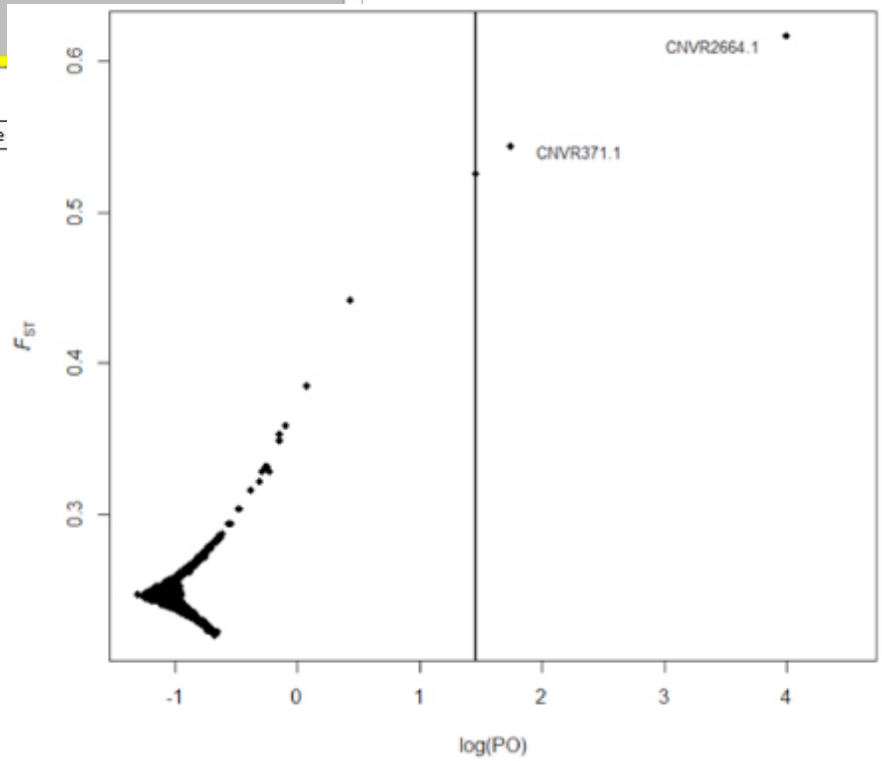
Needs *a priori* information about number of populations



Gaither et al. 2011

Outlier analyses



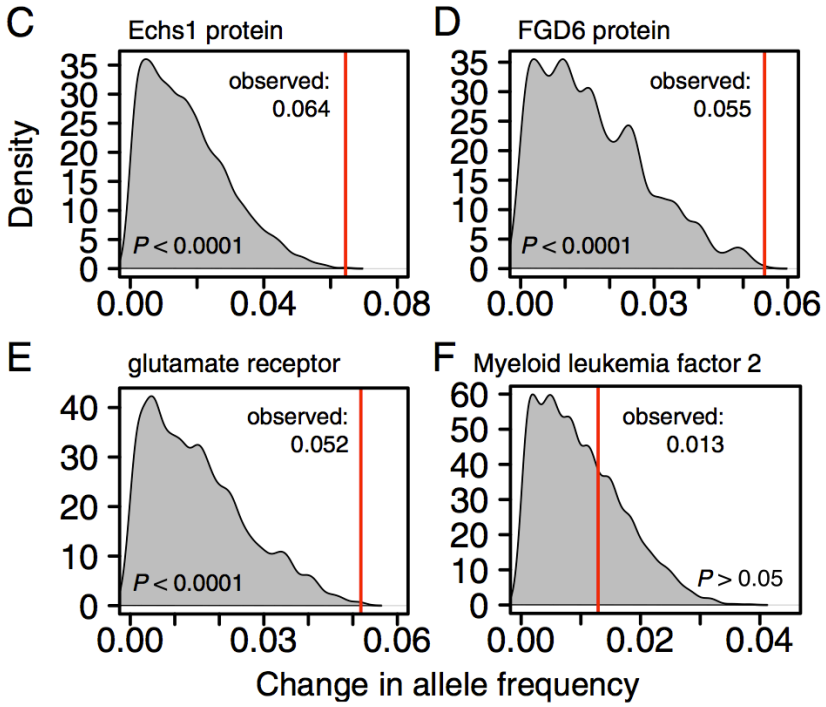


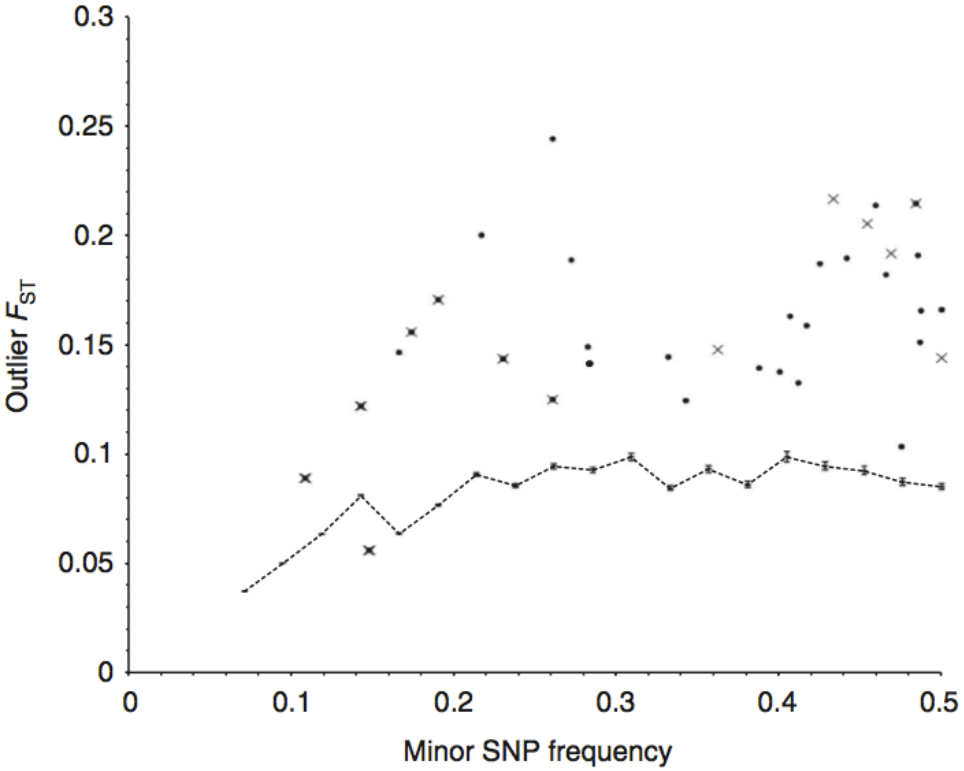
- Divergence vs heterozygosity plots – take any SNP above conf.int as outlier. Will ALWAYS find outliers.

- Bayesian methods to search for loci affected by positive selection (PO= posterior odds).

Permuted Outlier analyses

- By randomizing datasets, it is possible to test for statistical significance of F_{ST} or allele frequency changes





Pespeni et al. 2013 PNAS

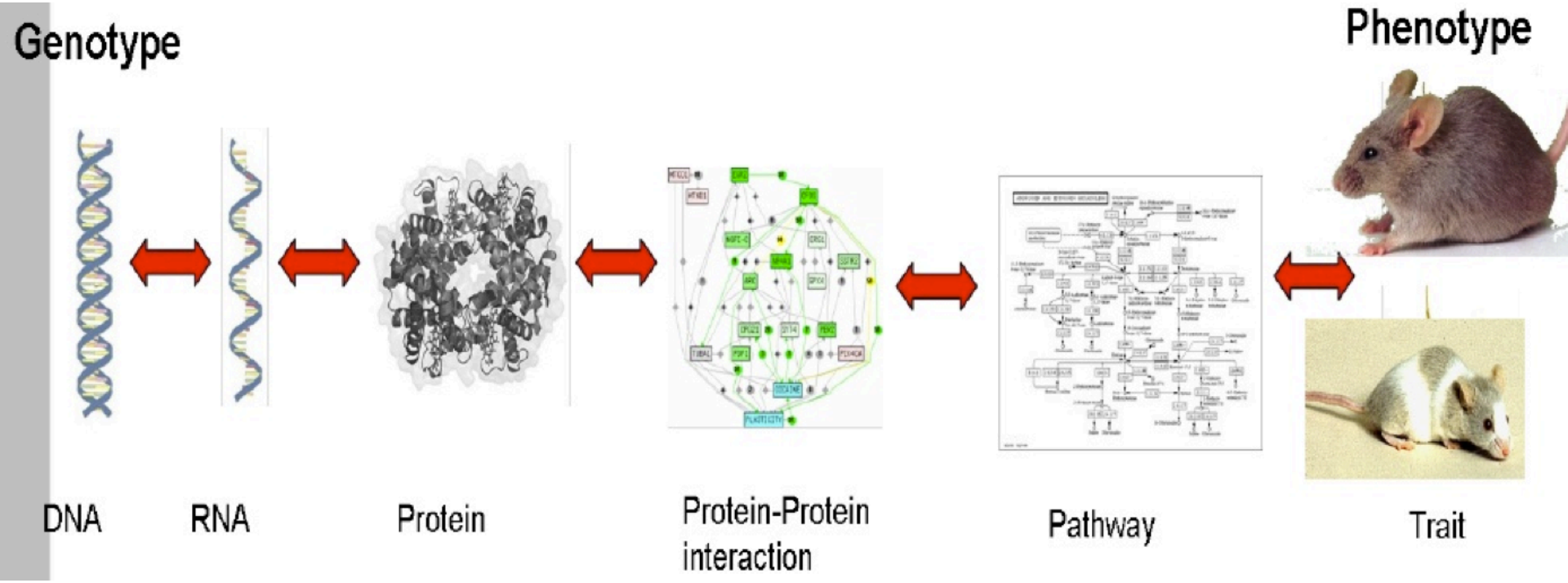
De Wit et al. 2014 Nature Comms

Genotype-Phenotype interactions

- In many cases, we want to know which loci are involved in traits.

- Two approaches for studying this:
  - Quantitative Trait Locus mapping
    - Lab crossing experiments of inbred lines
  - Genome-wide association scans
    - Natural populations

Gene expression is a phenotype

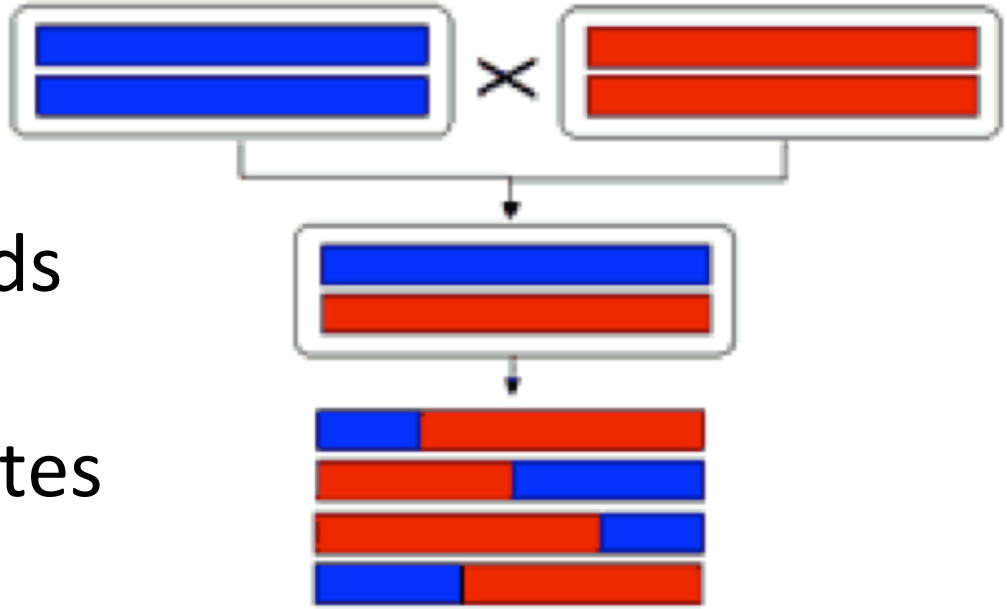


QTL mapping

Parents

F1 hybrids

F1 gametes



Can be used to infer CAUSATION

However, need inbred parent lines, and F2 offspring

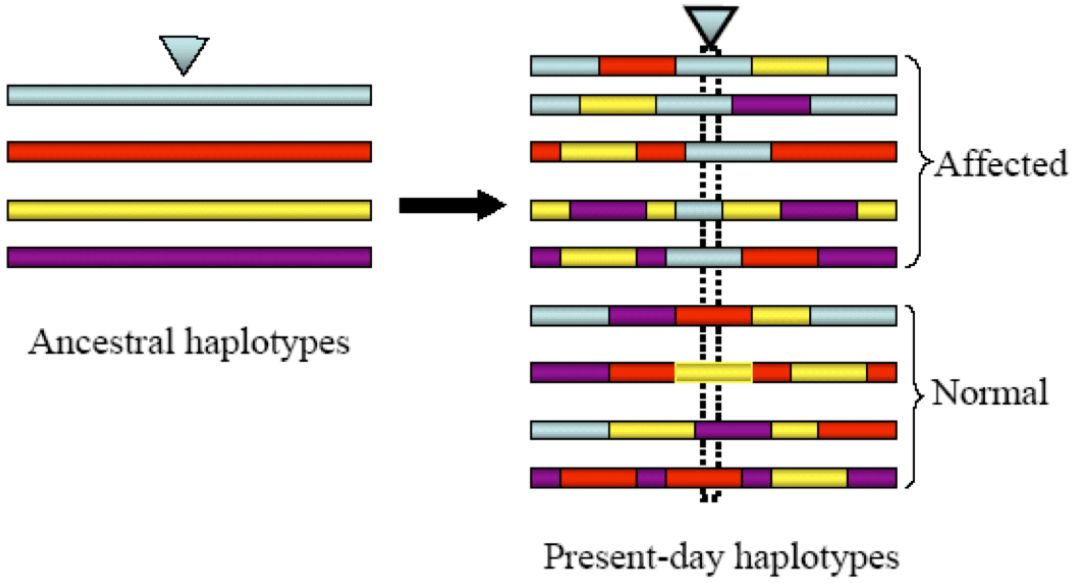
Also, linkage blocks are in general large (100s of kb), so hard to map to a single gene.

What is a good trait?

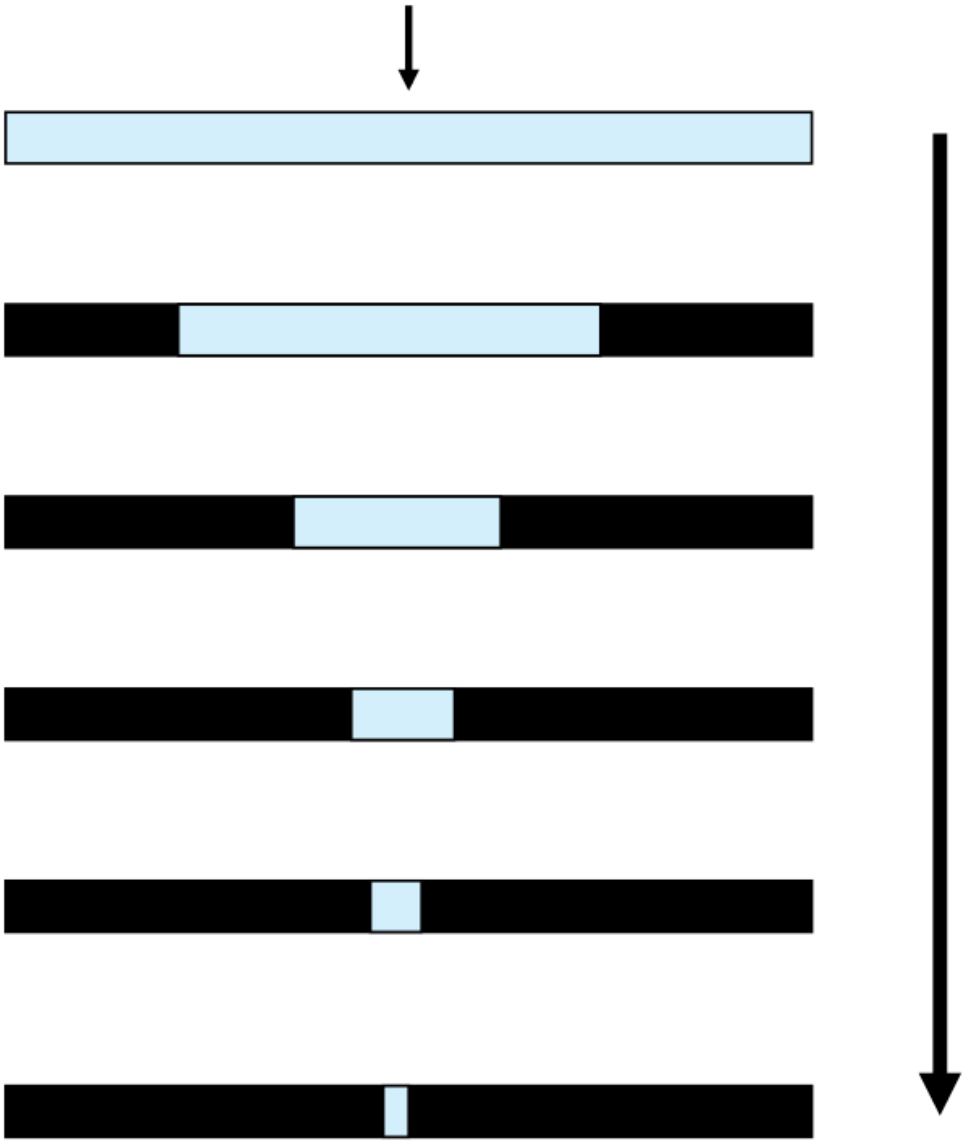
- Well-defined, quantifiable
- Affected by few loci of large effect
- Covariation with other traits?
  - Multivariate QTL analysis (Cheng et al. 2013 Genetics).

Association Mapping

- Uses historic recombination
- Correlation is not causation
- Analysis of many alleles
- High marker density
- Necessitates subsequent candidate gene testing



Every new mutation arises in Linkage Disequilibrium

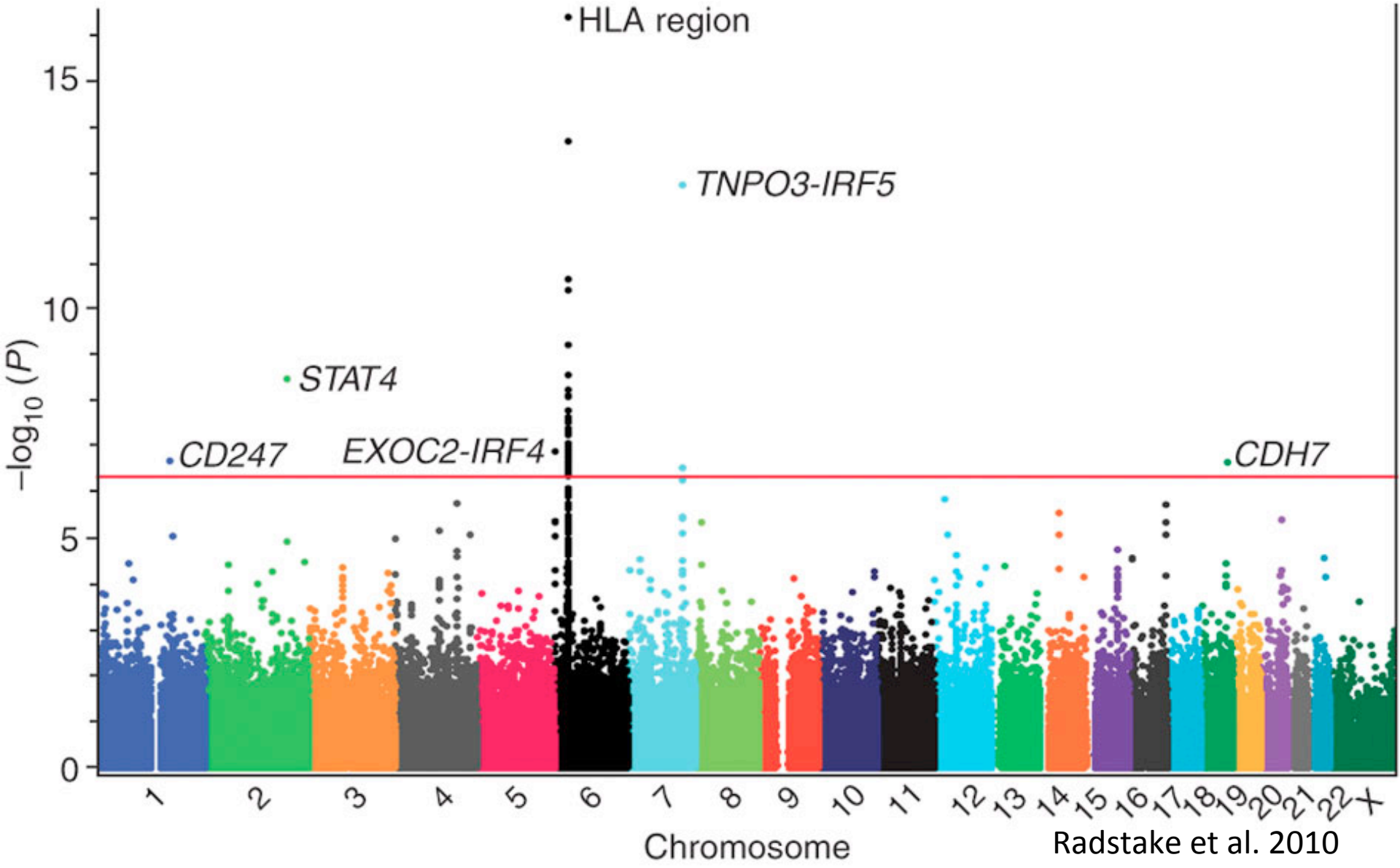


The longer a mutation has been in a population , the smaller the linkage block surrounding it.

- Requires more closely spaced markers to find regions of interest.

+ Fine mapping of outlier loci is possible (down to certain parts of exons in some cases)

Genome-Wide Association Scans



Radstake et al. 2010