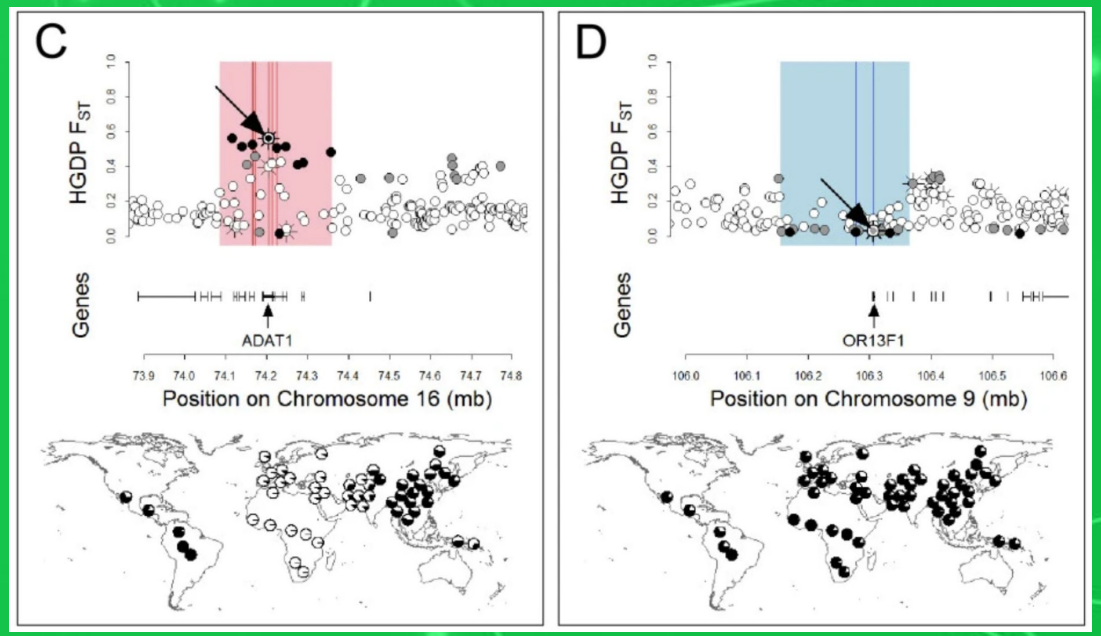
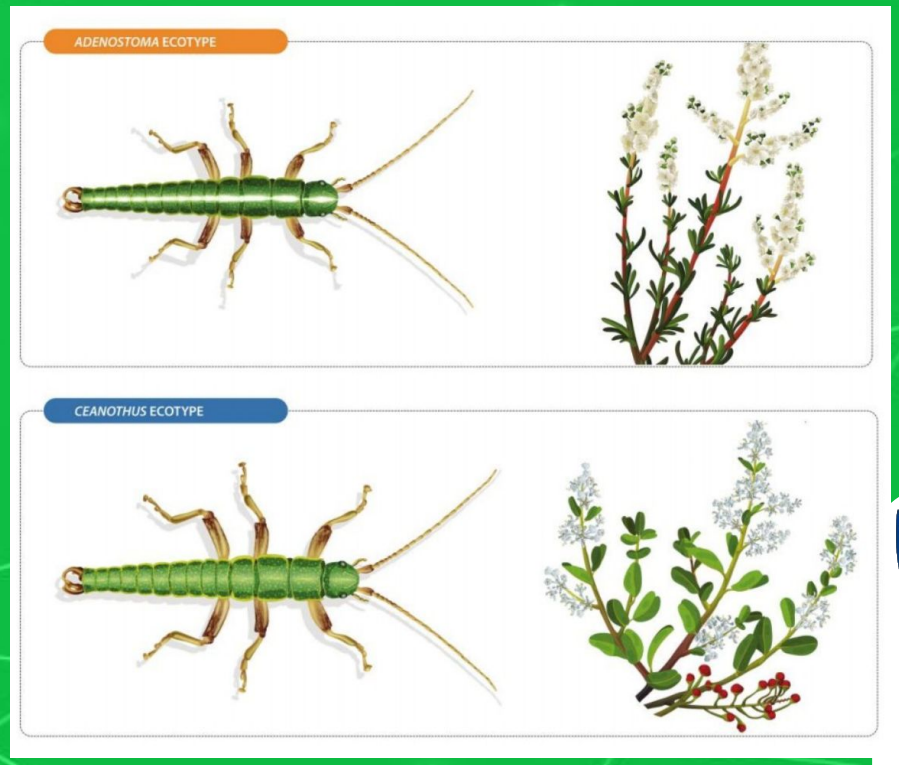
Centre for Genomic Research

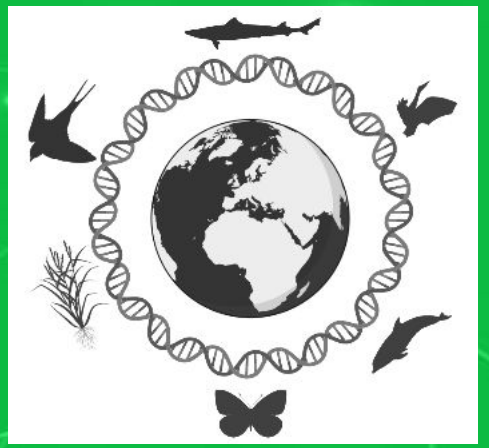
UKRI Natural Environment Research Council

NEOF NERC ENVIRONMENTAL OMICS FACILITY

Population Genomics - Day 2







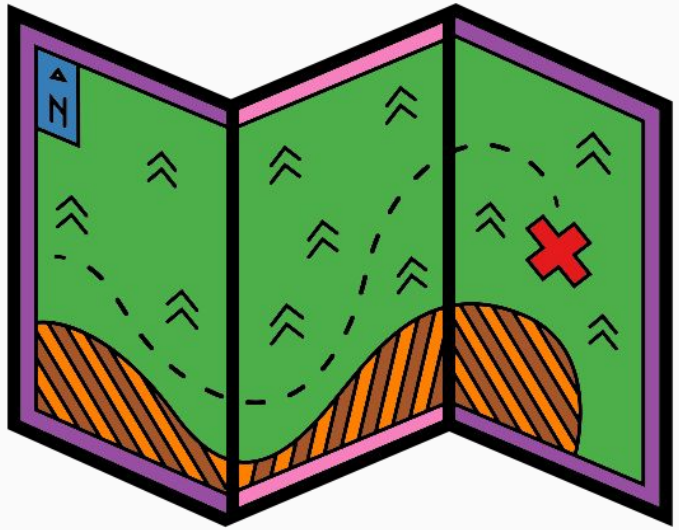
University of Liverpool

The University Of Sheffield.

Outline for Today

Detecting Selection

GWAS



Detecting Selection using genetic markers

- Do not require prior knowledge of candidates
- Generally using some sort of population genetics statistic to measure divergence: allele frequency, relative differentiation (FST), absolute divergence (Dxy), nucleotide diversity (π)
- Many methods: some simple, some more sophisticated involving models and simulation
- Single or multiple population comparisons

Detecting Selection using genetic markers

FST outliers tests

Loci with extreme allele frequencies relative to the background

Bayescan – Foll & Gaggiotti (2008) http://cmpg.unibe.ch/software/BayeScan/download.html

Detects outliers displaying greater differentiation than expected by fitting two models (with and without selection) for each locus

Population structure tests

Loci excessively related to population structure (candidates for local adaptation)

PCAdapt – Luu et al. (2016) https://cran.r-project.org/web/packages/pcadapt/index.html

Use PCA of genotypes to identify outliers

Genetic-Environment Association tests

Loci with allele frequencies correlated with environmental variables

BayEnv & BayEnv2 – Coop et al. (2010) https://bitbucket.org/tguenther/bayenv2_public

Null model is based on estimated patterns of covariance in allele frequencies between populations – estimated from all data and then used to test particular loci.

BayPass – Gautier (2015) http://www1.montpellier.inra.fr/CBGP/software/baypass/

BayPass extends BayEnv model and allows alternate covariate models to test for association with population-specific variables

Latent Factor Mixed Models (LFMM & LFMM2)

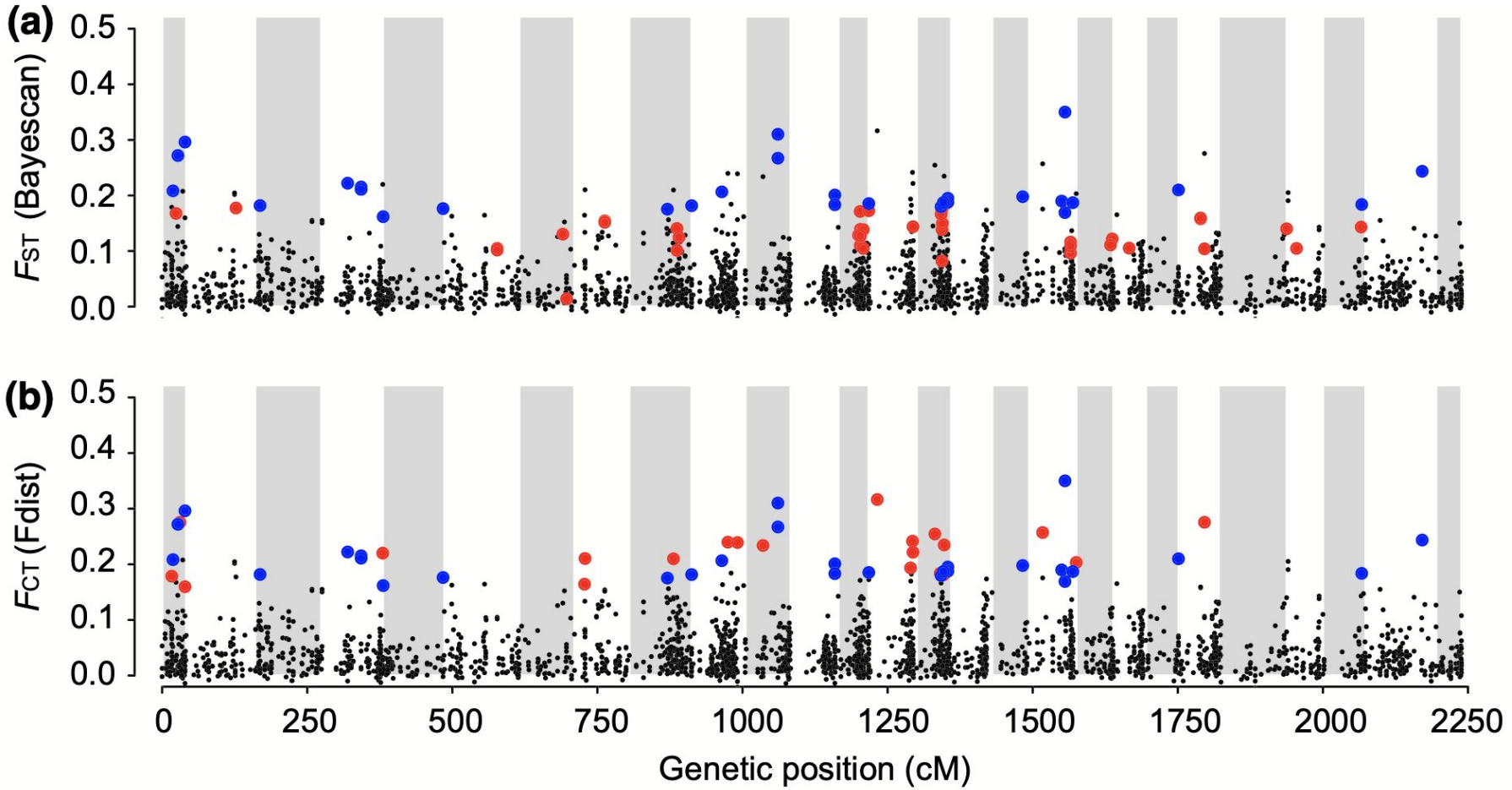
Frichot et al. (2013) http://membres-timc.imag.fr/Olivier.Francois/lfmm/index.htm

Caye et al. (2019) https://academic.oup.com/mbe/article/36/4/852/5290100

Association between loci and environmental gradients

Detecting Selection using genetic markers

Locus by locus genome scans - outlier tests



Moore *et al.* 2014 Molecular Ecology

North American Atlantic salmon - Genome scan among regions

Genome Scans Methods

Some considerations:

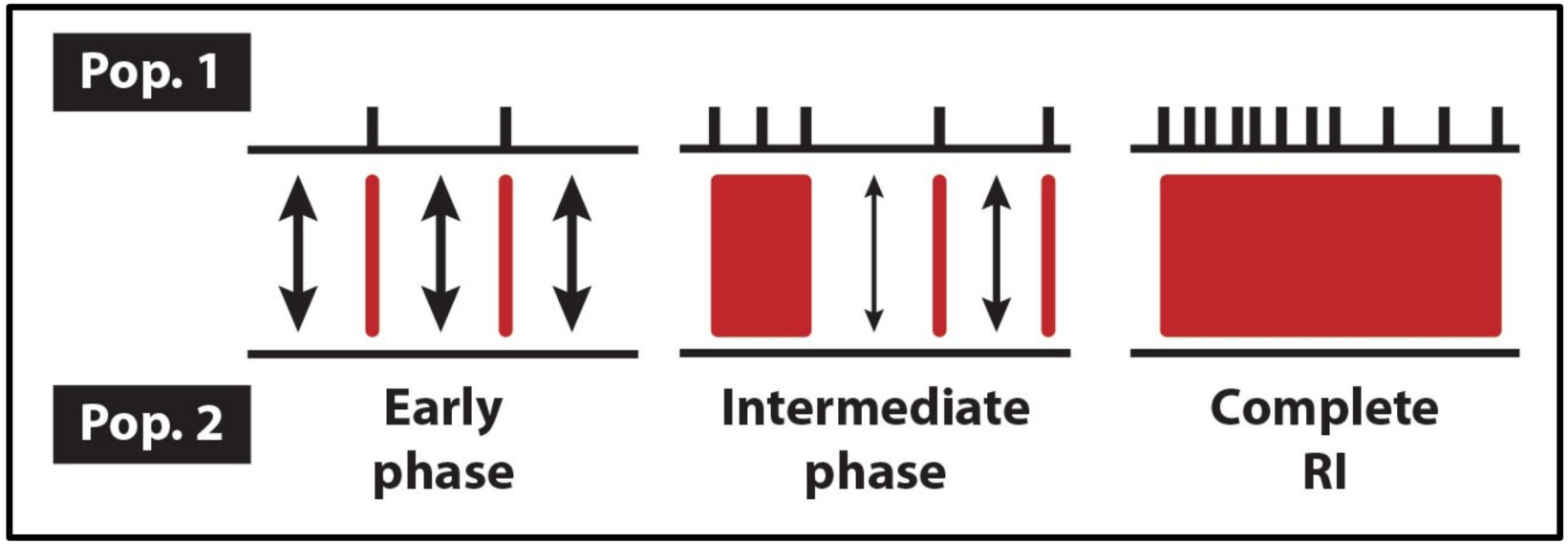
- Replicate population pairs for more robust analysis
- Outlier tests may only be sensitive to large- effect loci & find traits with simple genetic architecture
- Useful investigative tool, but may not give the whole picture
- Validation of candidates: gene expression, GWAS, common-garden experiments, knock-outs

Genomic islands

Genic model of speciation with gene flow

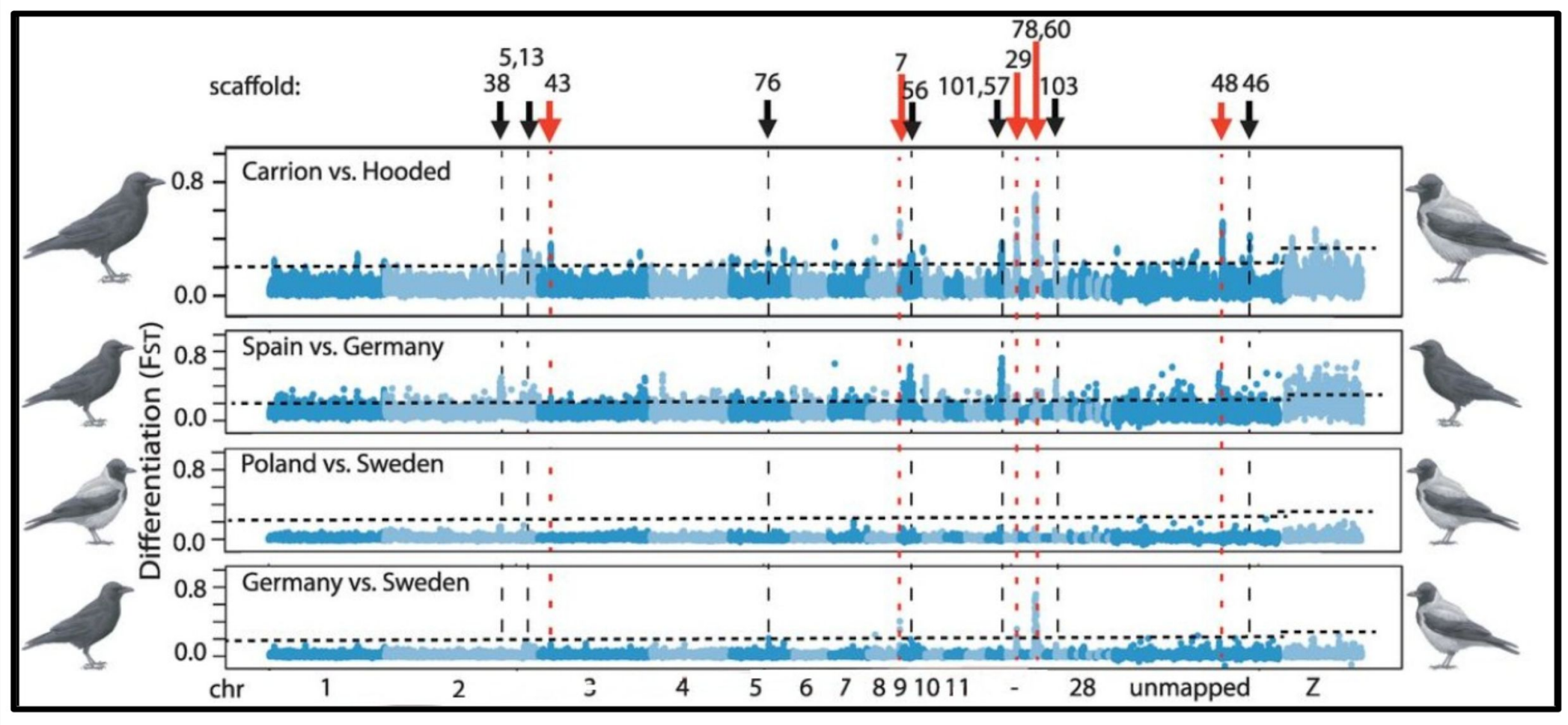
1) Early genic phase: Few localized regions of accentuated differentiation (‘genomic islands’)

2) Intermediate genomic phase: differentiation become genome-wide



Wu 2001 J Evol Biol; Mallet 1995 TREE

Genomic islands



Wu 2001 J Evol Biol; Mallet 1995 TREE

Methods

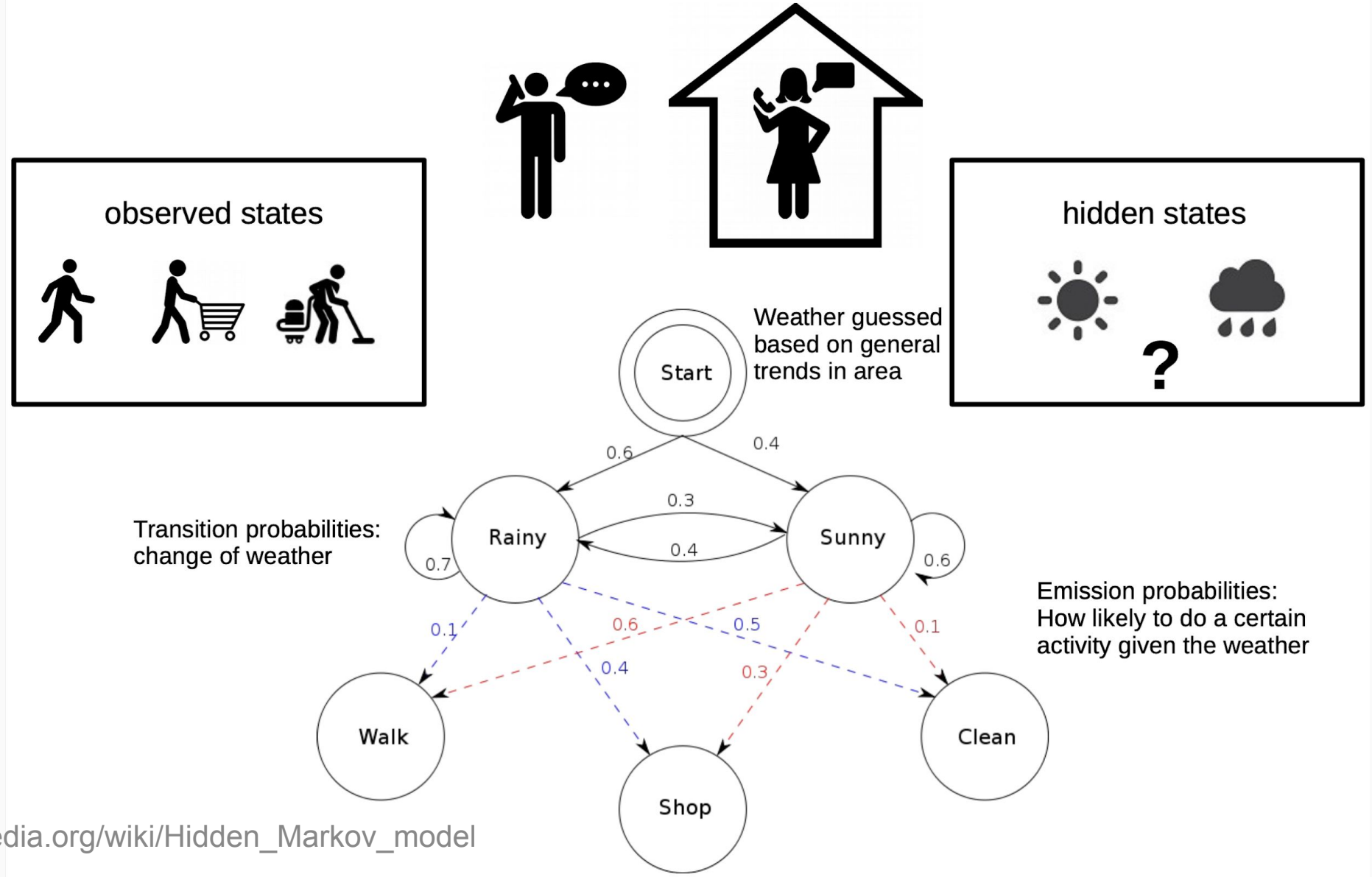
Sliding window approaches issues:

- Choice of window size not trivial
- Window size can have strong impact on number and size of regions
- Random fluctuations of the test statistic in a delimited window might lead to the detection of a cluster when there is none

Hidden Markov Models (HMM)

- Probabilistic models for linear sequence 'labeling'
- Statistical model in which the system is modeled as a Markov process with hidden states (Markov process: probability of subsequent state depends only on previous state)
- Explicitly model dependencies among neighbouring markers

Hidden Markov Models

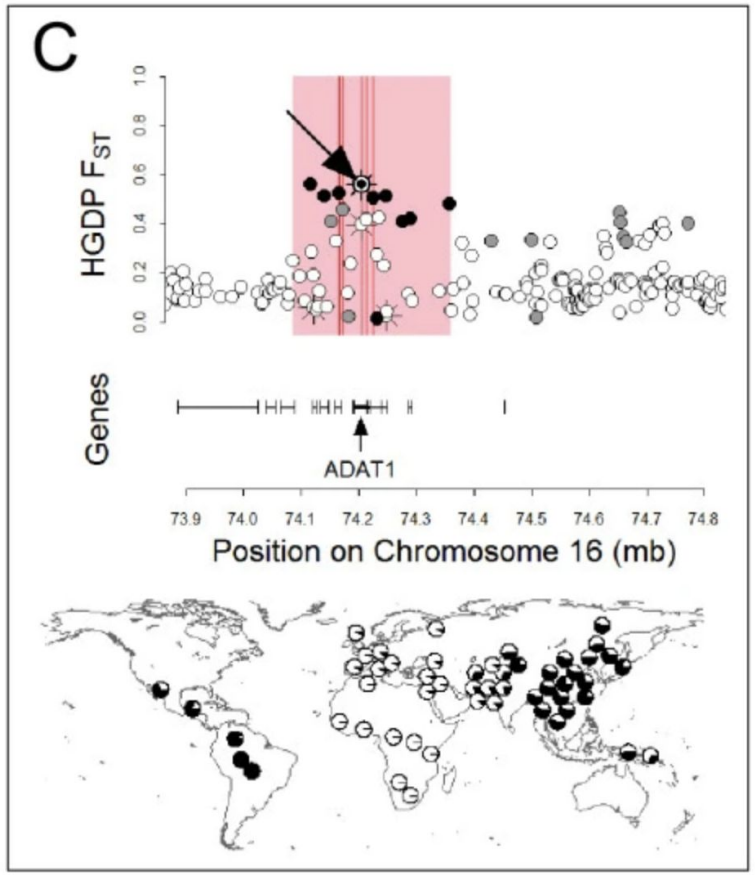


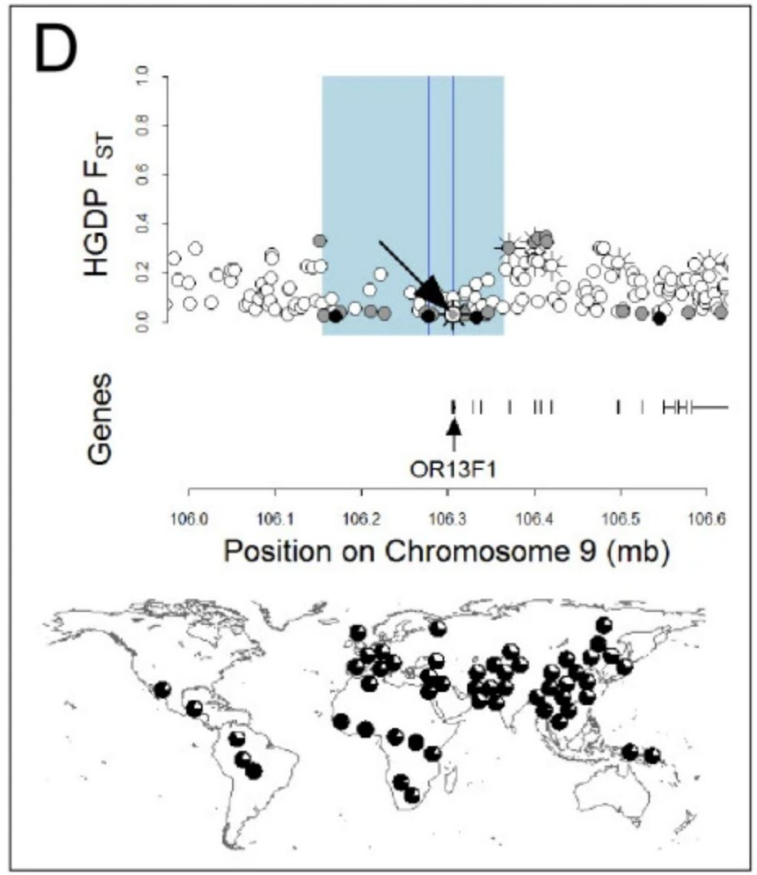
https://en.wikipedia.org/wiki/Hidden_Markov_model

Example: genomic islands of differentiation in human populations

Used HMM to identify regions of high or low differentiation

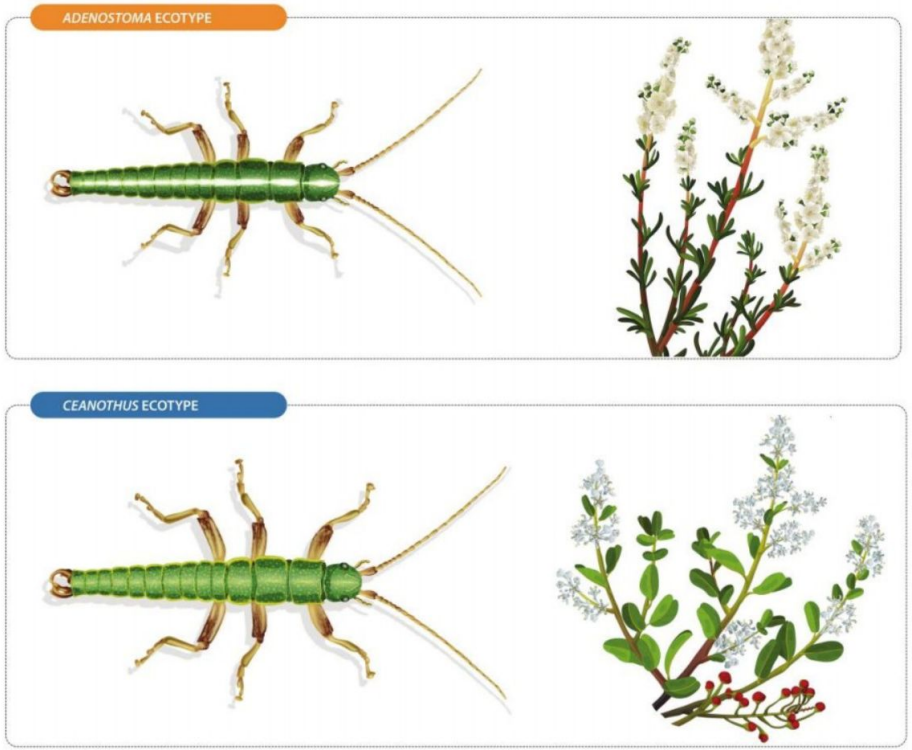
Examined spatial distribution of significant SNPs





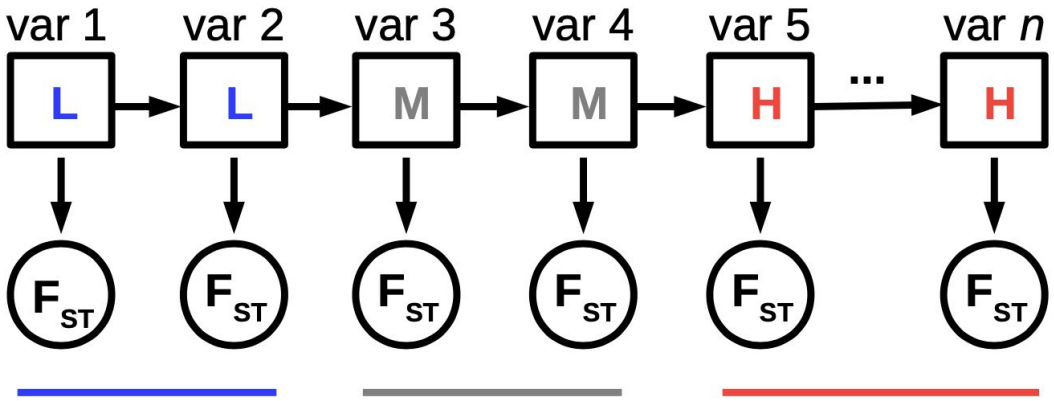
Hofer, T *et al.* https://doi.org/10.1186/1471-2164-13-107

Example: genetic differentiation in Timema stick insect ecotypes (workbook chapter 9)



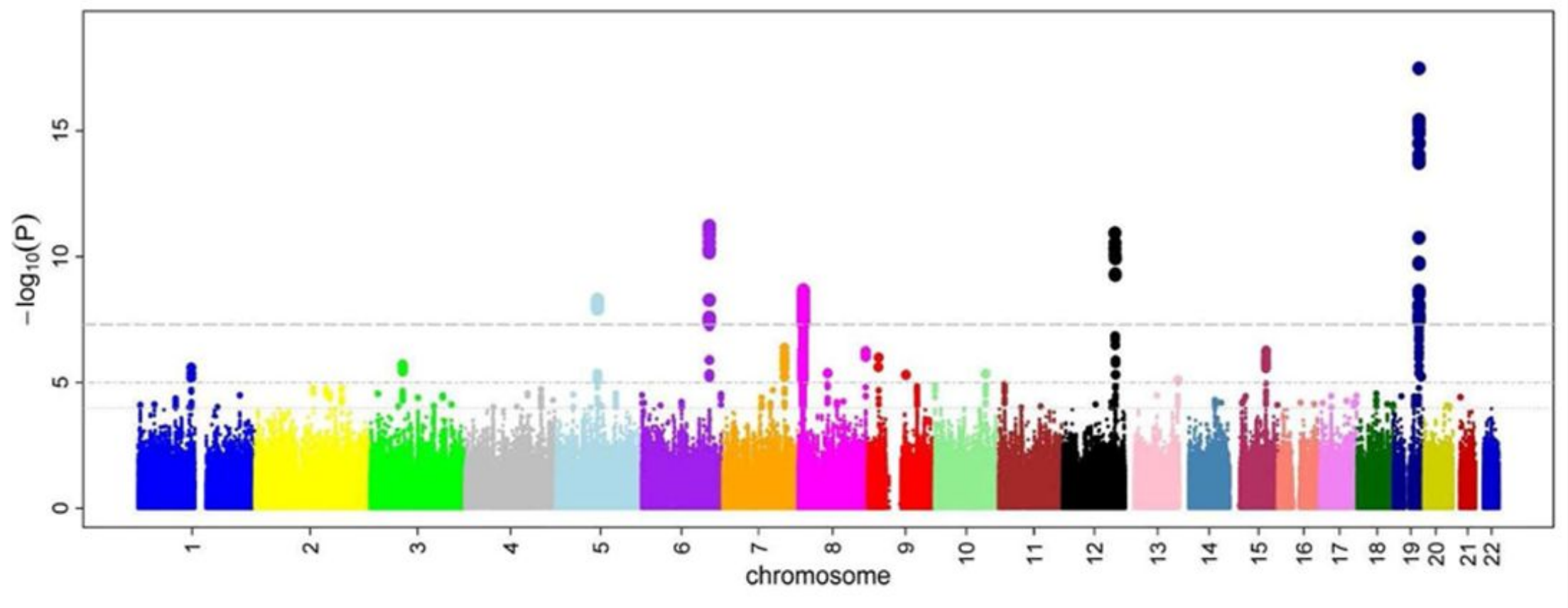
Observed: F_{ST} across genome (SNPs)

Hidden: 3 differentiation states – low (L), medium (M), and high (H)



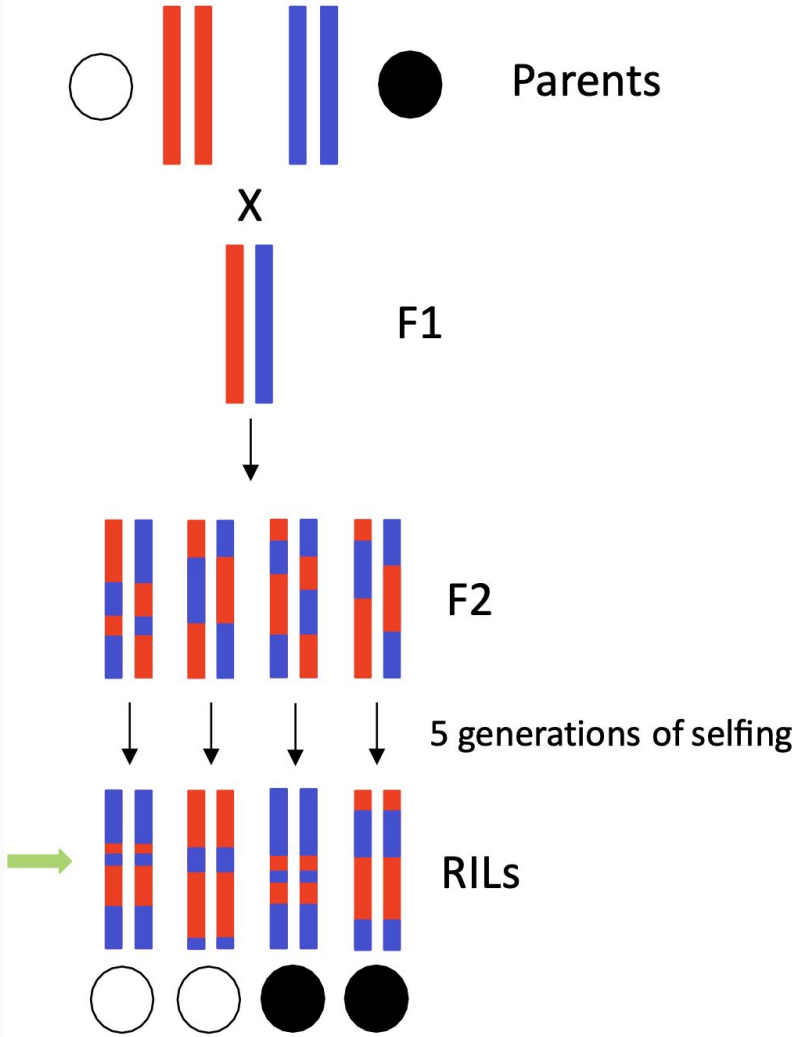
delimitation contiguous genetic regions
number, size, distribution

Genome Wide Association Studies



https://en.wikipedia.org/

QTL mapping: example of design



Advantage:

'low number' of marker needed

Problems:

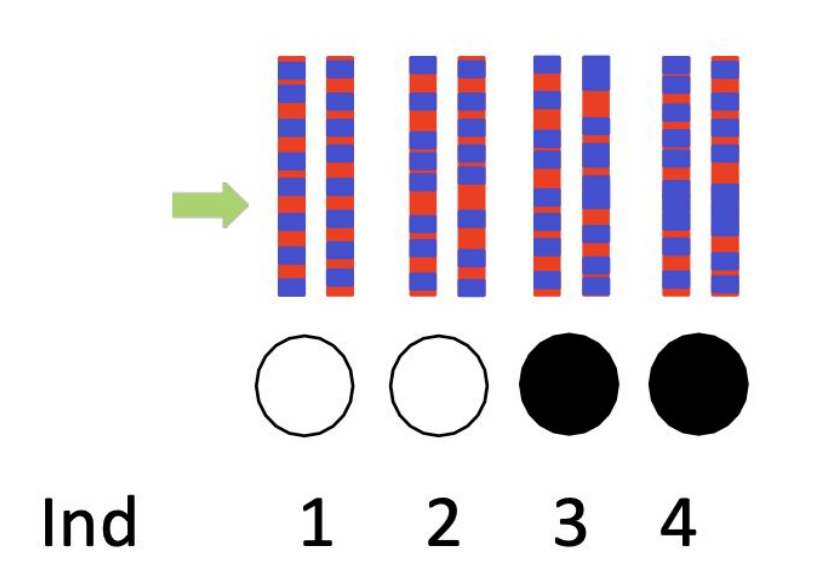
Take time to make crosses (ex: 2–3years in *A. thaliana* with short generation time)

Discovery of allele not 'evolutionary meaningful' = rare variants

Lack of recombination = QTL are usually huge genomic regions ~ 200 genes

→ GWAS (Genome–wide association study)

GWAS: making use of recombination occurring in natural populations



Advantages:

Smaller genomic regions identified
No need for crosses

Problems:

Need for a high number of markers, well distributed over the genome and in linkage disequilibrium with causal variants

Sensitive to rare allele (MAF cut–off)

Population structure = demographic history

GWAS software

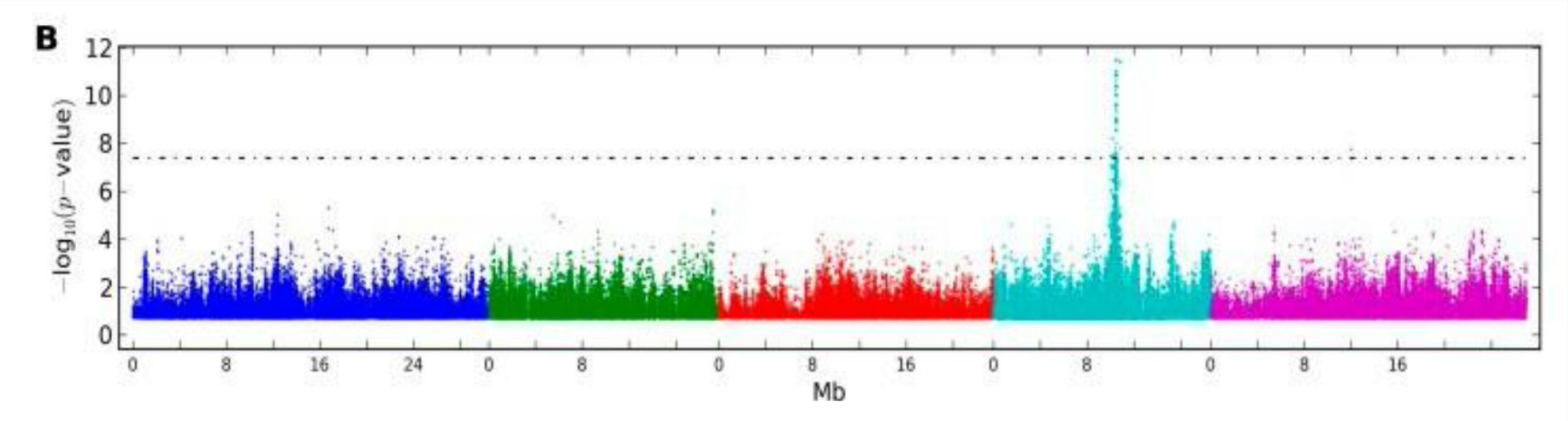
EMMAX, GenABEL, GAPIT, GWAspi, gemma, piMASS ... and many others!

2 main types:
Single marker association: e.g. GenABEL
Multi-marker association: e.g. gemma - session in this workshop

Single marker methods

e.g. MMAX, GenABEL

Association computed for every marker independently along the genome

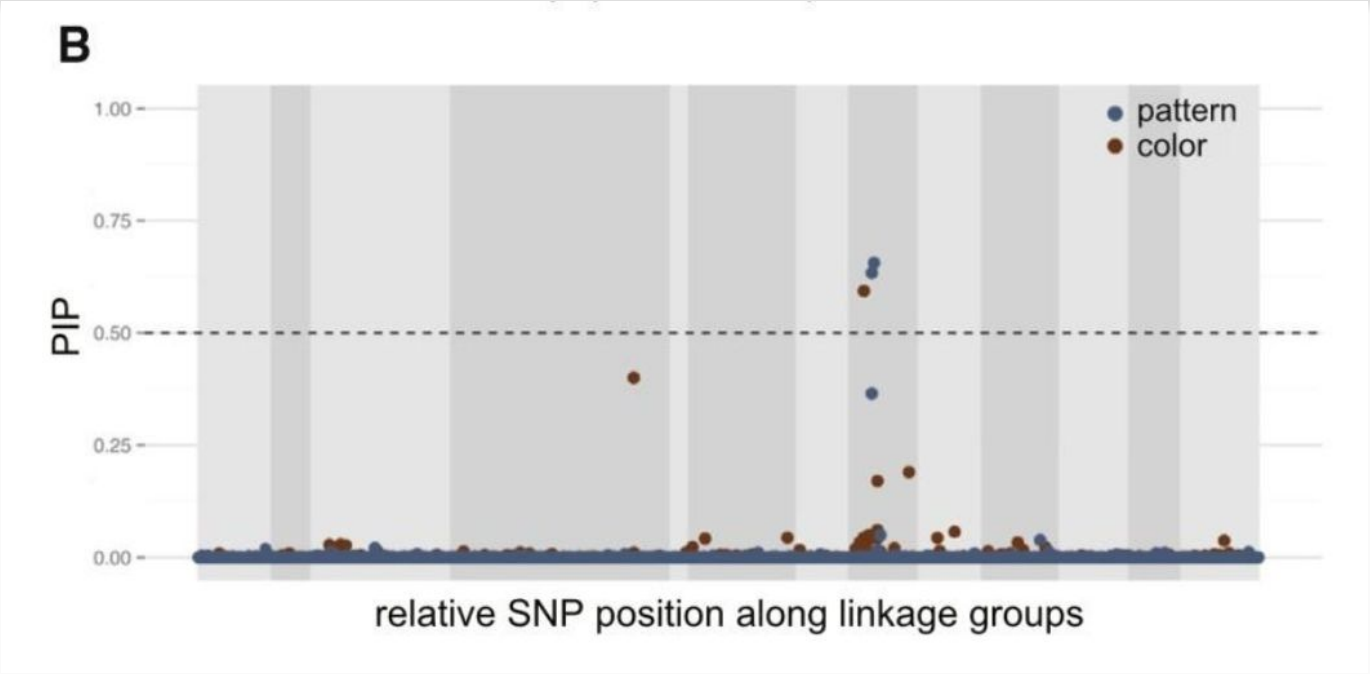


Advantages: Very fast (30 seconds to 5 minutes)

Caveats: Results with LD for downstream analysis (enrichment in GO term...)

Multi-marker methods

e.g. gemma, piMASS
Association computed with a combination of markers



Advantages: Results without LD for downstream analysis (enrichment in genetic features, GO etc...)
Caveats: Bayesian framework and MCMC = takes time(1.5 days per run)

Marker with the highest association score != causal variant

The marker with the highest association score can sometimes be the causal variant but:

- It is only the most associated marker in your dataset and the causal variant might not be in it
- Most GWAS only use bi-allelic SNPs, but structural variants may be crucial (indels, inversions, etc)
- Epistatic interactions

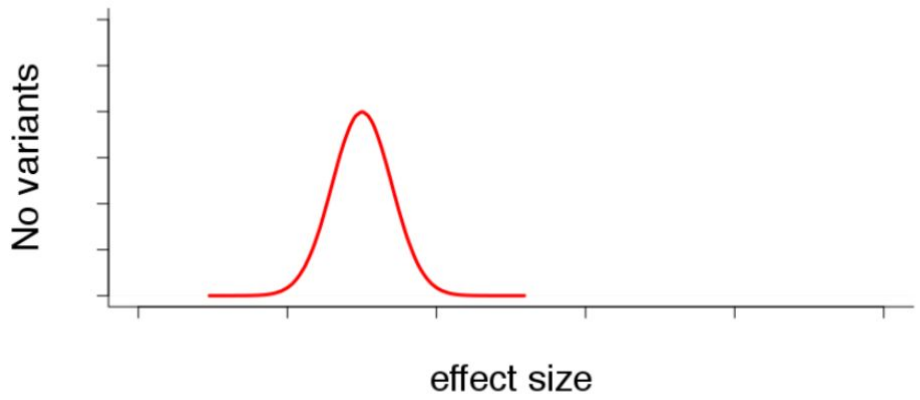
GWAS MULTI-SNP MODELS

Linear Mixed Model (LMM)

Assume polygenic basis:

all variants affect the phenotype

Effect sizes normally distributed

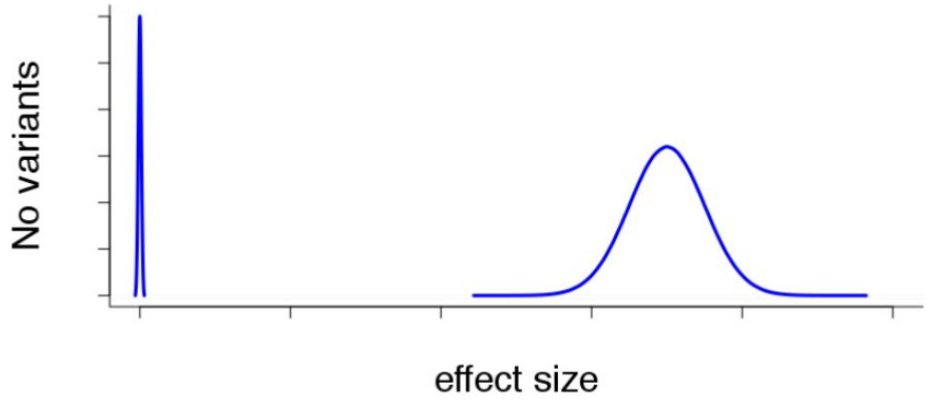


Bayesian Variable Selection Regression model (BVSR)

Assume mono/oligogenic basis:

a small proportion of variants affect the phenotype

Effect sizes as mixture of point mass at 0 and normal distribution



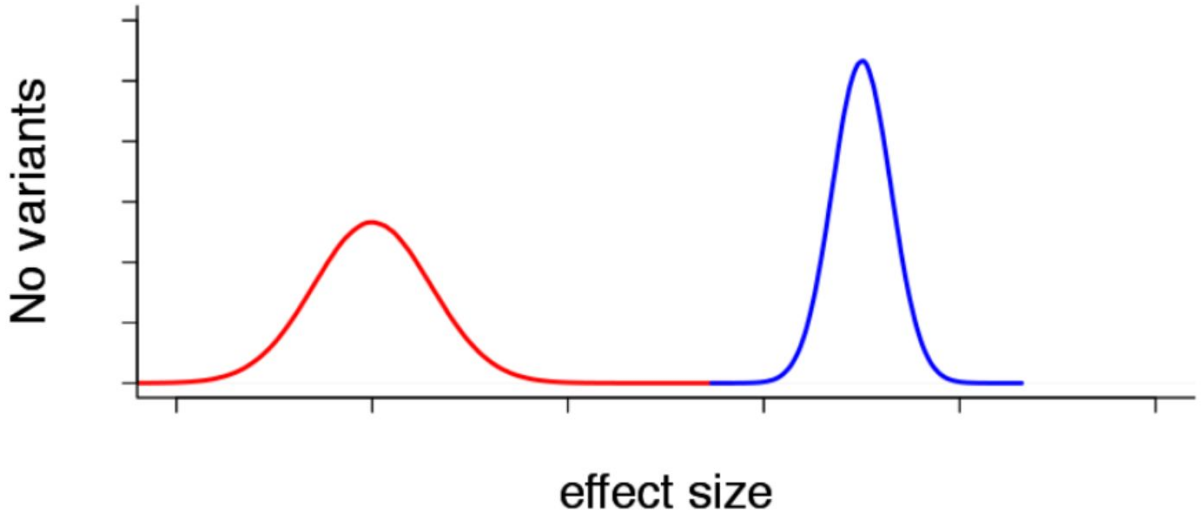
GWAS MULTI-SNP MODELS

Hybrid general model:
Bayesian Sparse Linear Mixed Model
(BSLMM)

Mixture of polygenic (LMM) and mono/oligogenic basis (BVSR)

Two distribution of effect sizes:
1) small effect size of all variants (α)
2) additional large effect size of some variants (β)

effect size of a given variant = $\alpha_i + \beta_i$



GEMMA
Genome-wide Efficient Mixed Model Association

Three models:

- Univariate Linear Mixed Model (LMM)
- Multivariate Linear Mixed Model (mvLMM)
- **Bayesian-Sparse Linear Mixed Model (BSLMM)**

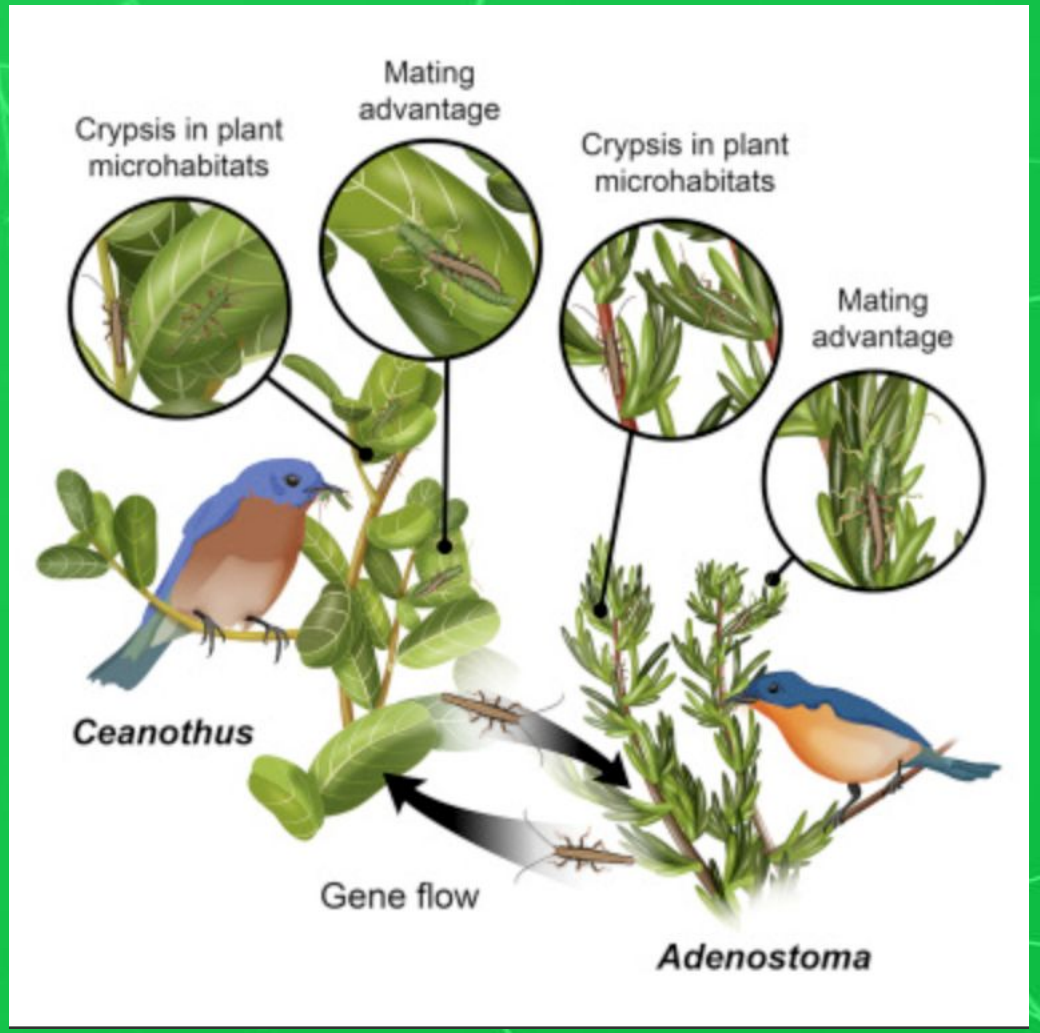
Manual – read it!
www.xzlab.org/software/GEMMAmanual.pdf

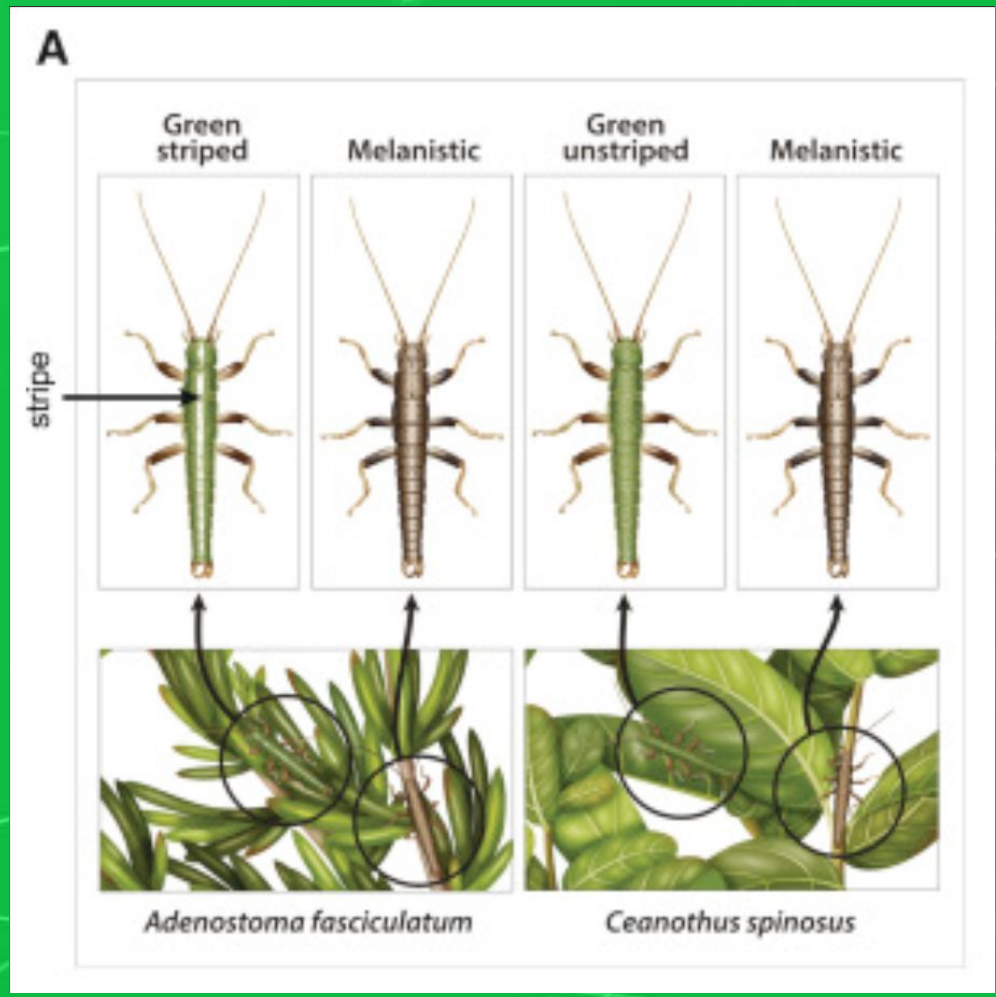
Publications

- Xiang Zhou and Matthew Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 44: 821–824. http://goo.gl/pFb7Qy
- Xiang Zhou and Matthew Stephens (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods. 11(4): 407–409. http://goo.gl/9pWM1Y
- **Xiang Zhou, Peter Carbonetto and Matthew Stephens (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics. 9(2): e1003264. http://goo.gl/YStR2a**

Also, BSLMM accounts for sample relatedness and population stratification

GEMMA example – genetic basis of phenotypic trait (workbook chapter 10)





Thank you!

Questions?