Analyzing quantitative traits with R
QTL mapping
and
Genome Wide Association Study (GWAS)

BTI Discovery & Innovation in the Life Sciences
BOYCE THOMPSON INSTITUTE

May 10th 2017

Topic(s)

- QTL mapping with Rqtl

- GWAS mapping with R: MLM model
  MLMM model (optional)

- GWAS mapping using tassel (optional)

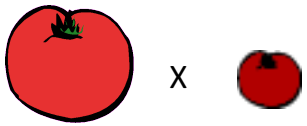
- Course material (scripts, input files and docs):

https://wordpress.com/page/btiplantbioinfocourse.wordpress.com/2471

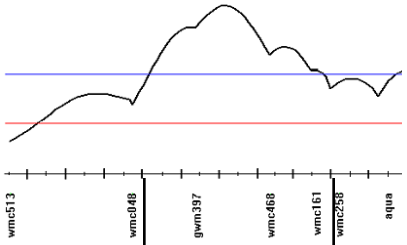
Introduction

QTL mapping

Mapping populations: F2, BC, DH



QTL mapping

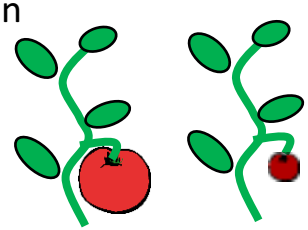


Fine mapping

Physical mapping

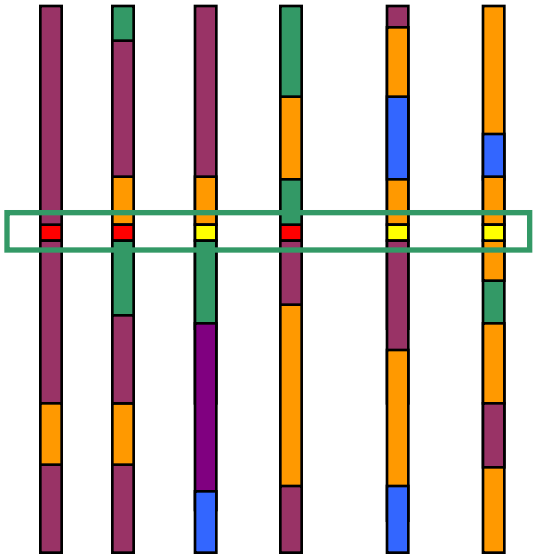
Positional cloning

Functional validation

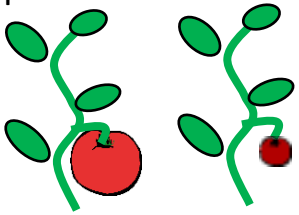


Association mapping (GWAS)

Natural populations

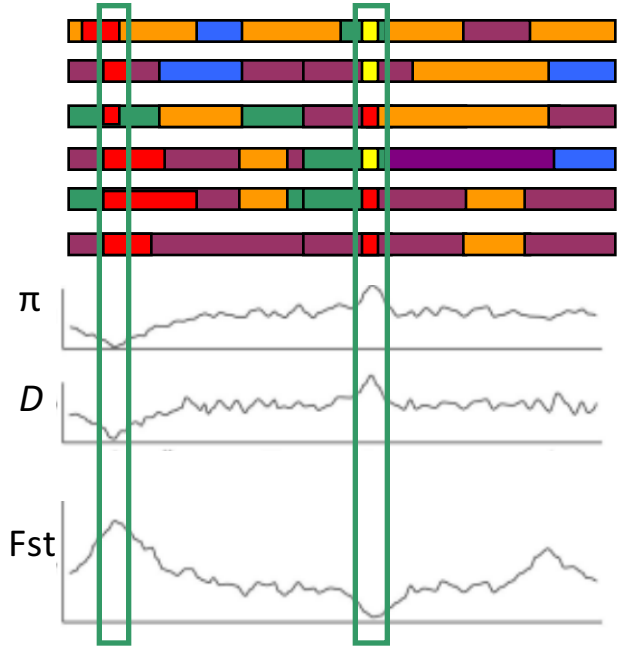


Functional validation



Population genomics

Natural populations



Nucleotidic diversity,

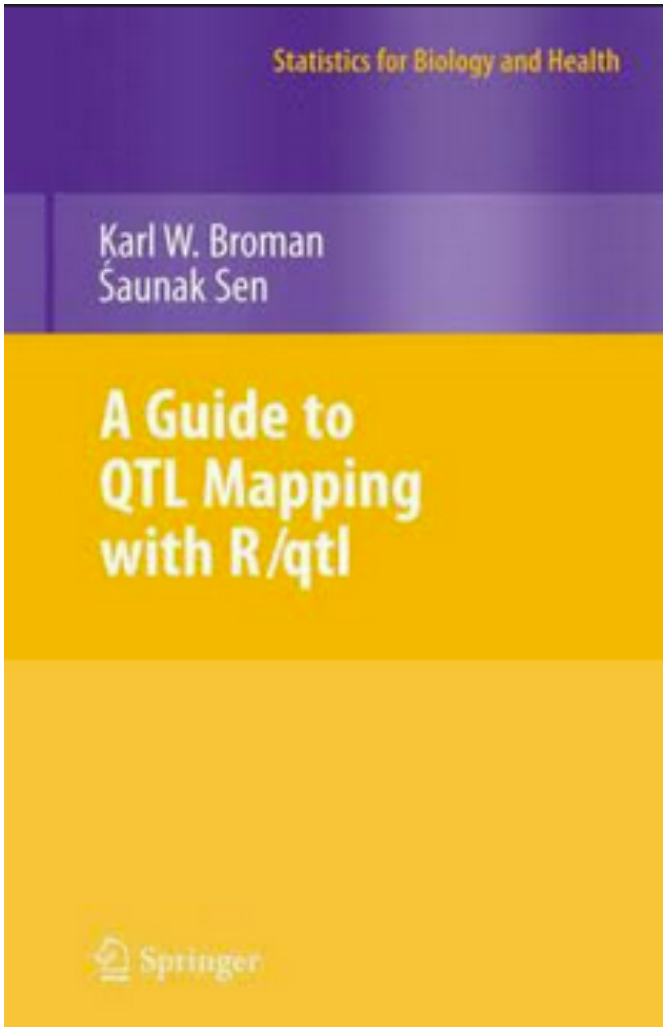
Tajima's *D*,

Fixation index,

dN/dS,

etc…

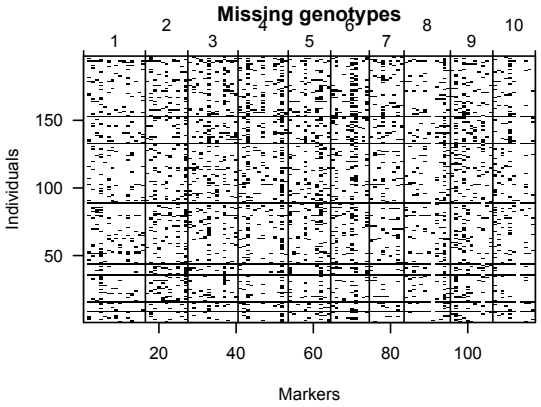
http://www.rqtl.org/

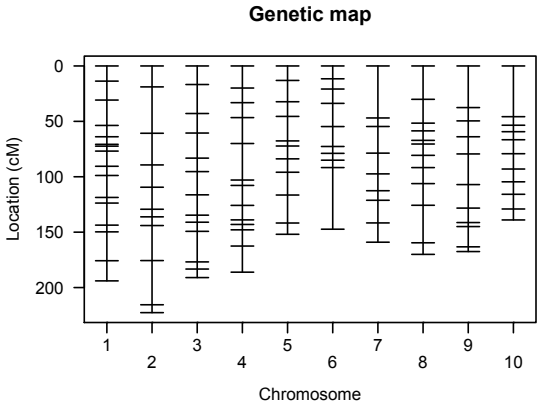


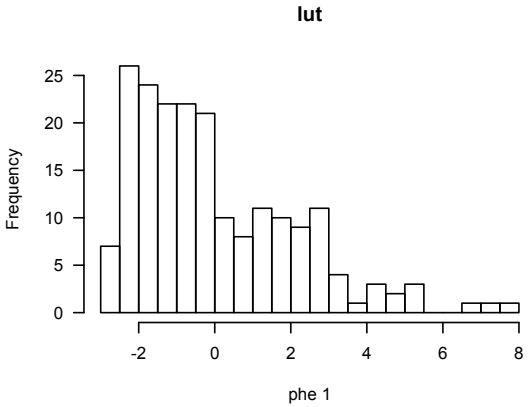
Our test dataset:

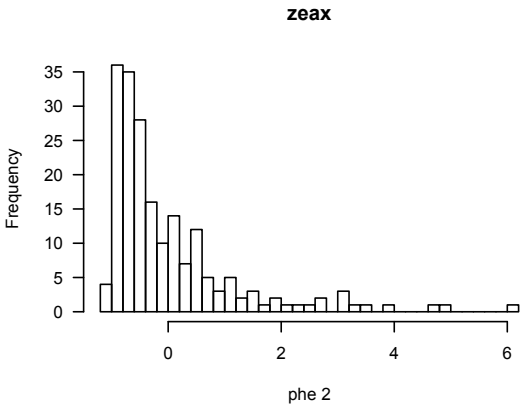
Maize dataset on carotenoid traits

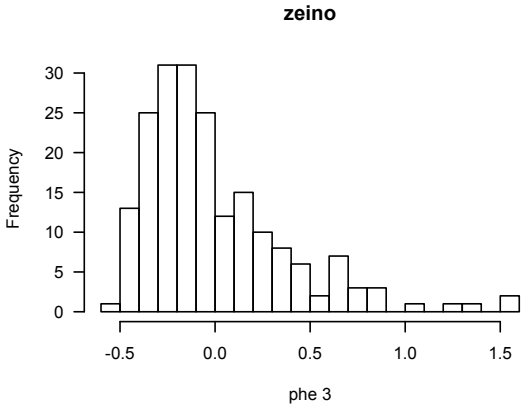
Hands on #1: QTL analysis

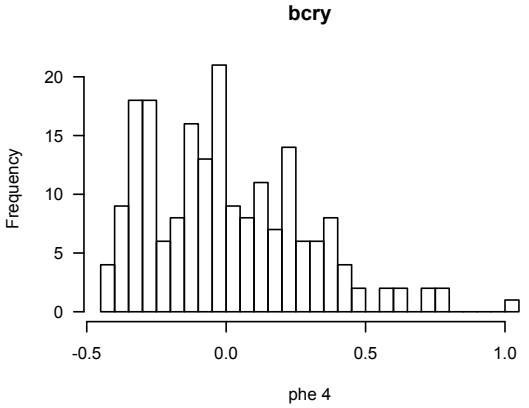


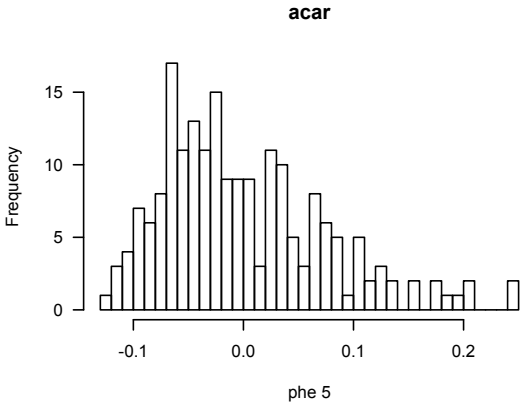


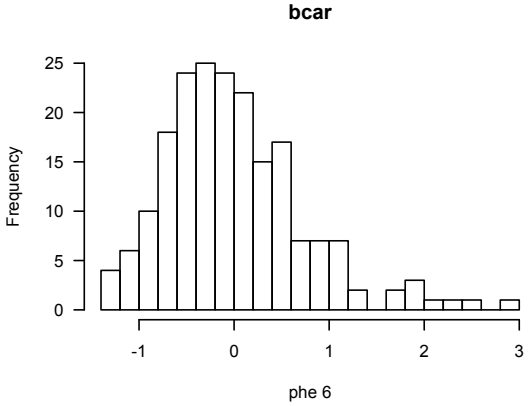




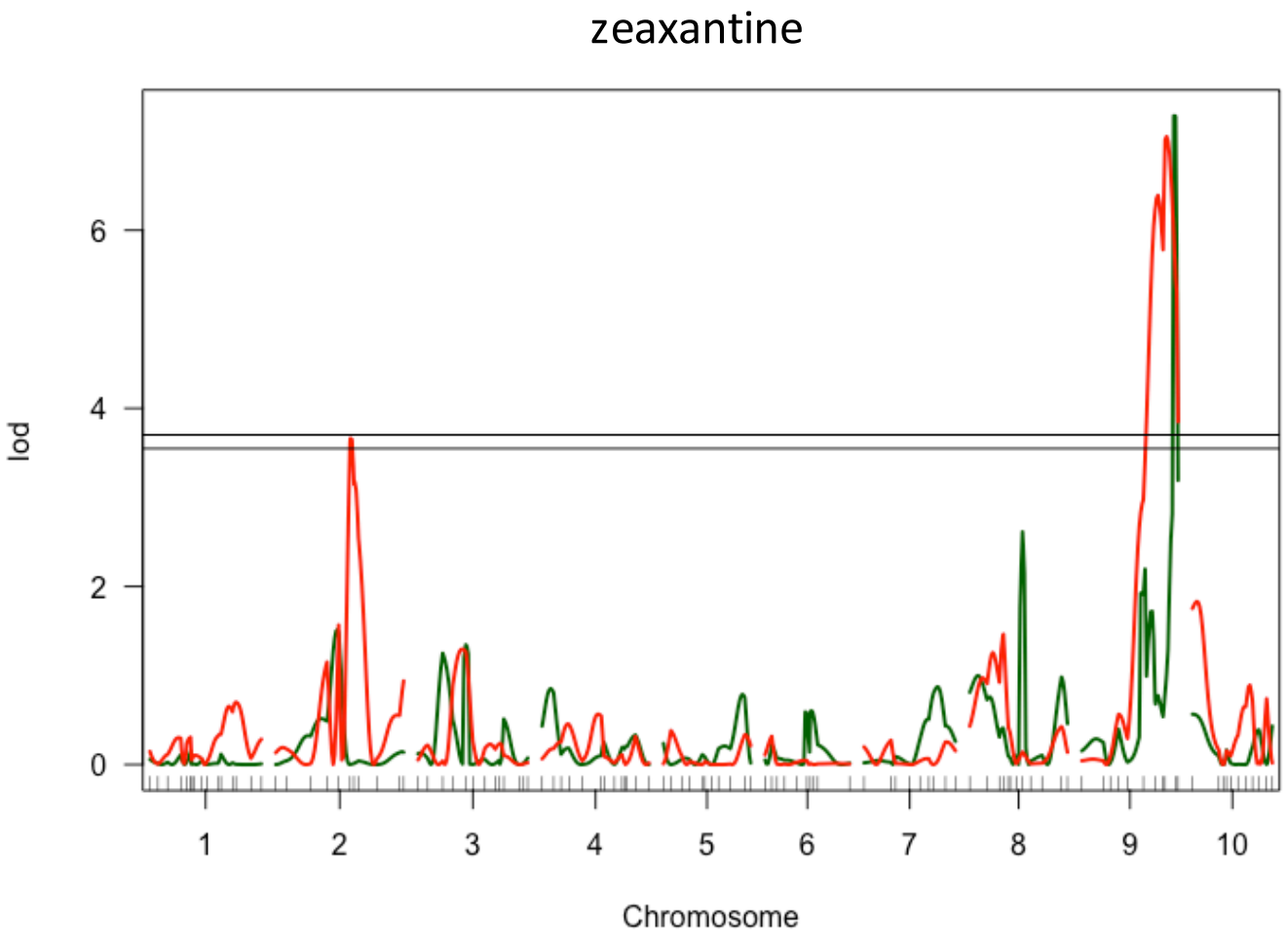








Different models: single locus/standard interval (SIM) or composite interval (CIM)



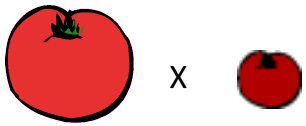
Hands on #1: QTL analysis

Steps and Questions

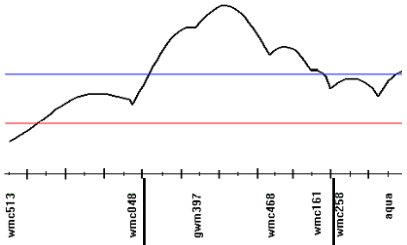
# Step 1: Importing data	-> "ASCarotFamily091014.csv"
# Step 2: Data checking	-> How many individuals? Markers? Phenotypes? -> Type of cross? -> Generate physical map and trait distribution plo -> Generate trait correlation pairs
# Step 3: Single-QTL analysis	-> plot QTL for trait #2, which chromosomes are involved? -> How many QTL with LOD>5?
# Step 4: composite interval mapping	-> plot CIM vs SIM QTL for trait #2, which are the differences? -> How many QTL with LOD>5?
# Step 5: Non-normal phenotypes	-> How are trait distributions?
# Step 6: QTL effects	-> display effect plot, any interactions?

QTL mapping

Mapping populations: F2, BC, DH



QTL mapping

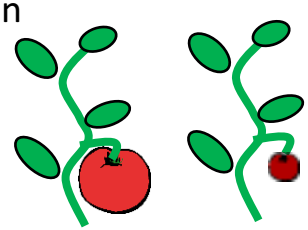


Fine mapping

Physical mapping

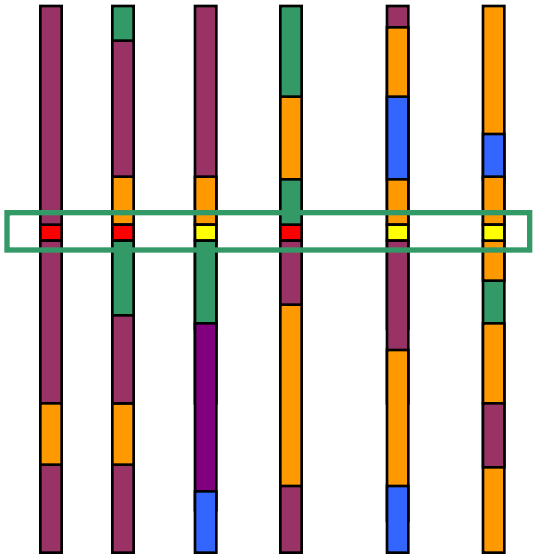
Positional cloning

Functional validation

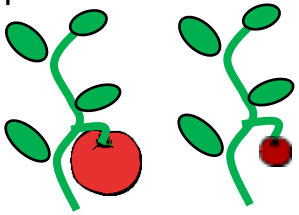


Association mapping (GWAS)

Natural populations

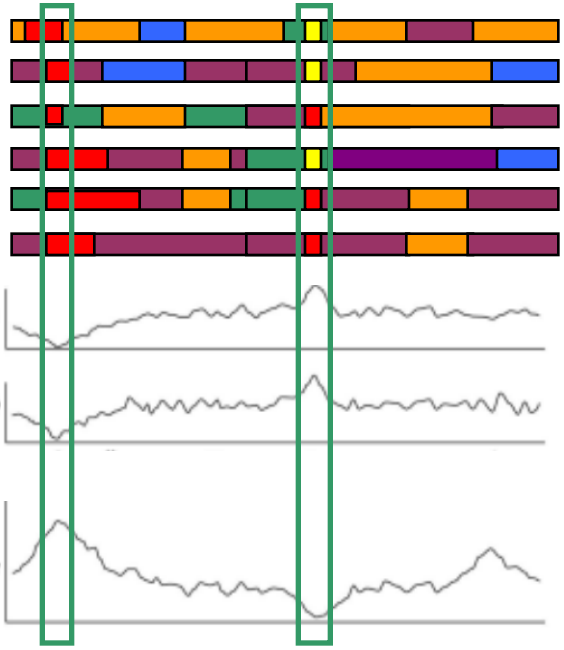


Functional validation



Population genomics

Natural populations



Nucleotidic diversity,

Tajima's *D*,

Fixation index,

dN/dS,

etc…

GWAS, what for?

Same as QTL mapping...

-Study the genetics underlying natural variation

-Understand the genetic architecture of traits of ecological and agricultural importance

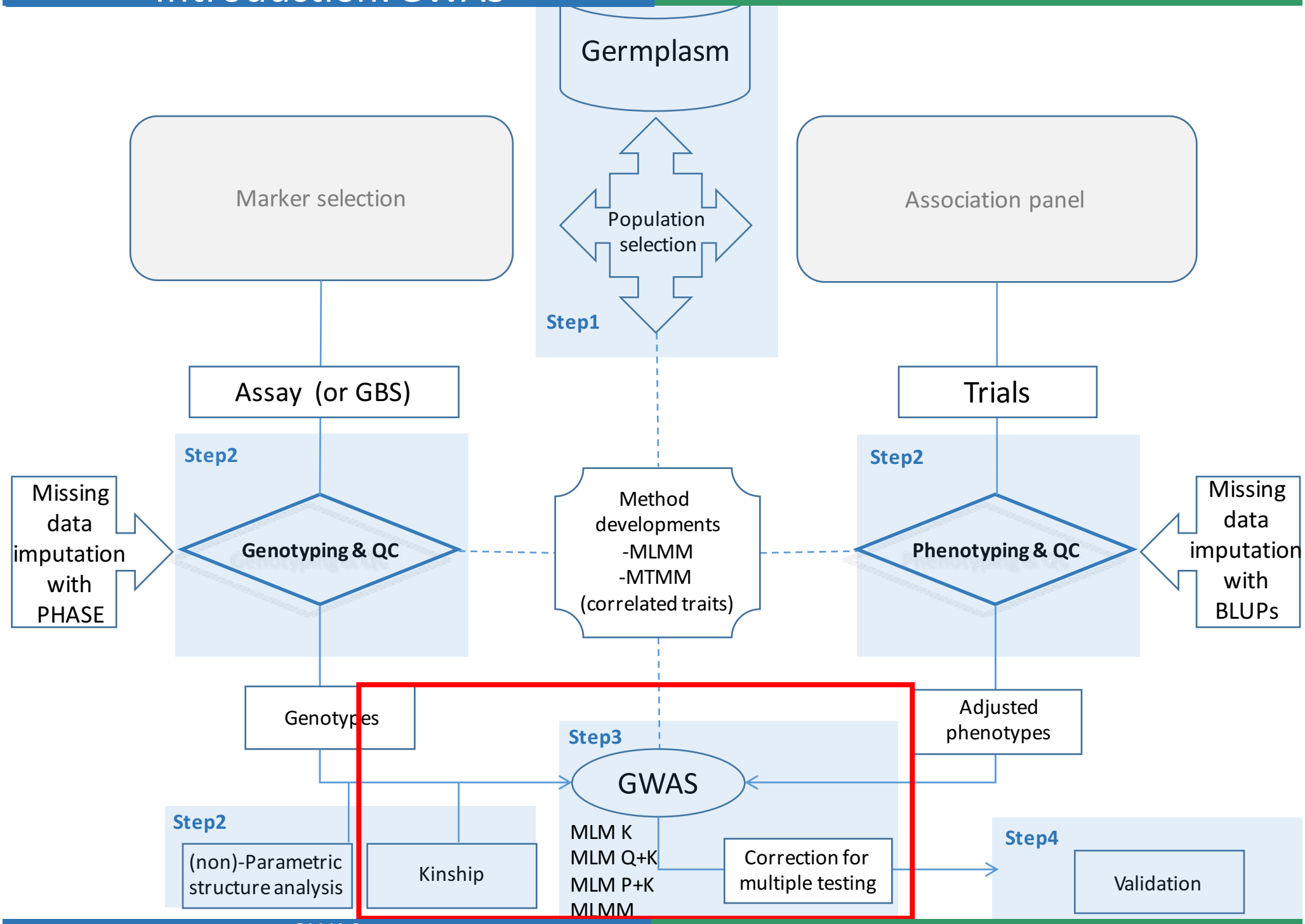
-Identify the genomic regions controlling their genetic variation

GWAS, how?

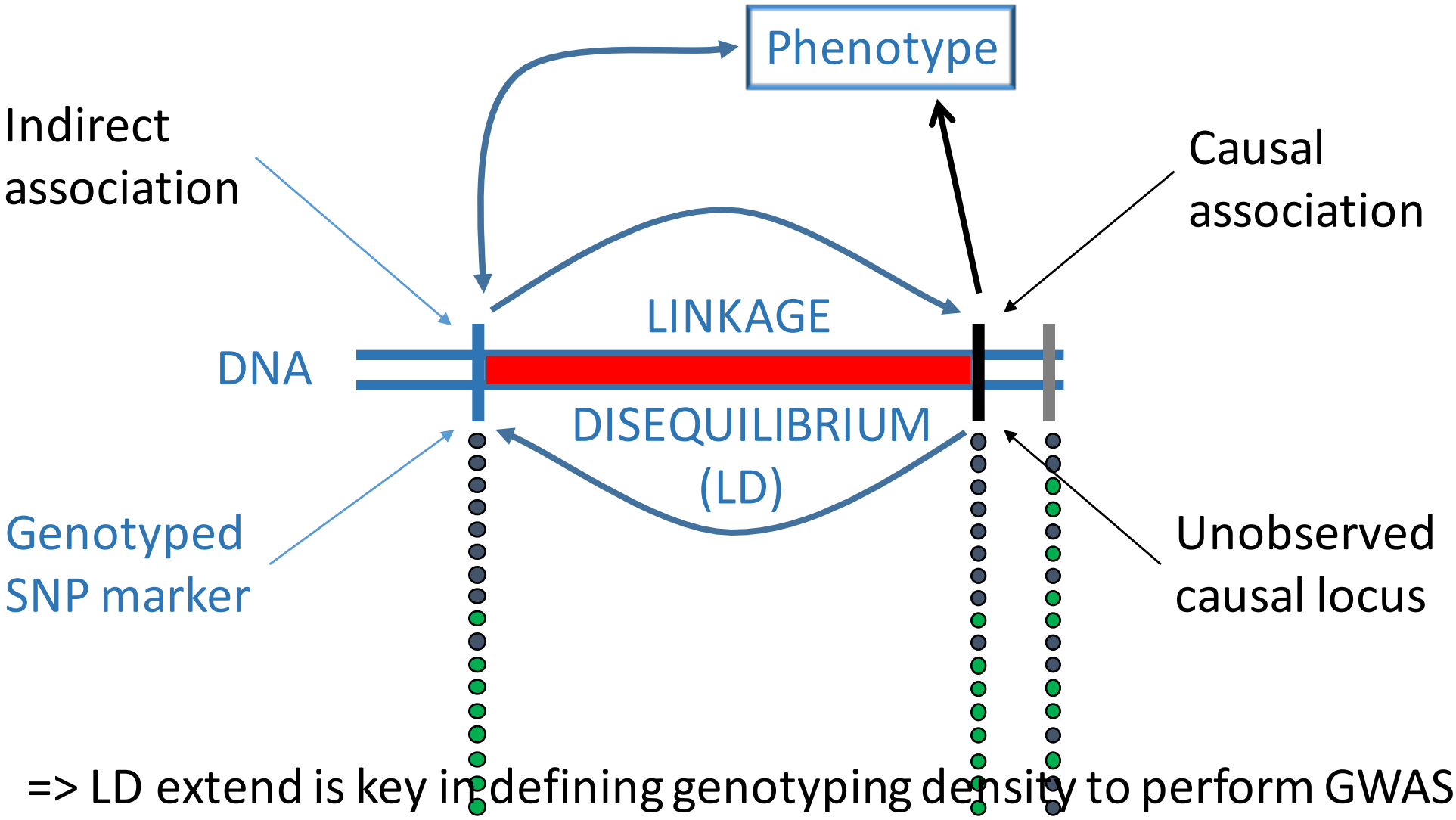
Combine Genotypic and phenotypic data within a population

Population Genotype Phenotype

Accession	SNP1	SNP2	...	SNPi	Trait
1	A	C	...	T	0.546
2	T	C	...	C	-1.745
3	A	G	...	C	-1.298
....	....	....	....	....	....
N	T	C	....		-0.972


Germplasm
Marker selection
Population selection
Association panel
Step1
Assay (or GBS)
Trials
Step2
Step2
Missing data imputation with PHASE
Genotyping & QC
Method developments
-MLMM
-MTMM
(correlated traits)
Phenotyping & QC
Missing data imputation with BLUPs
Genotypes
Step3
GWAS
Adjusted phenotypes
Step2
(non)-Parametric structure analysis
Kinship
MLM K
MLM Q+K
MLM P+K
MLMM
Correction for multiple testing
Step4
Validation

linkage study in a natural population



=> LD extend is key in defining genotyping density to perform GWAS

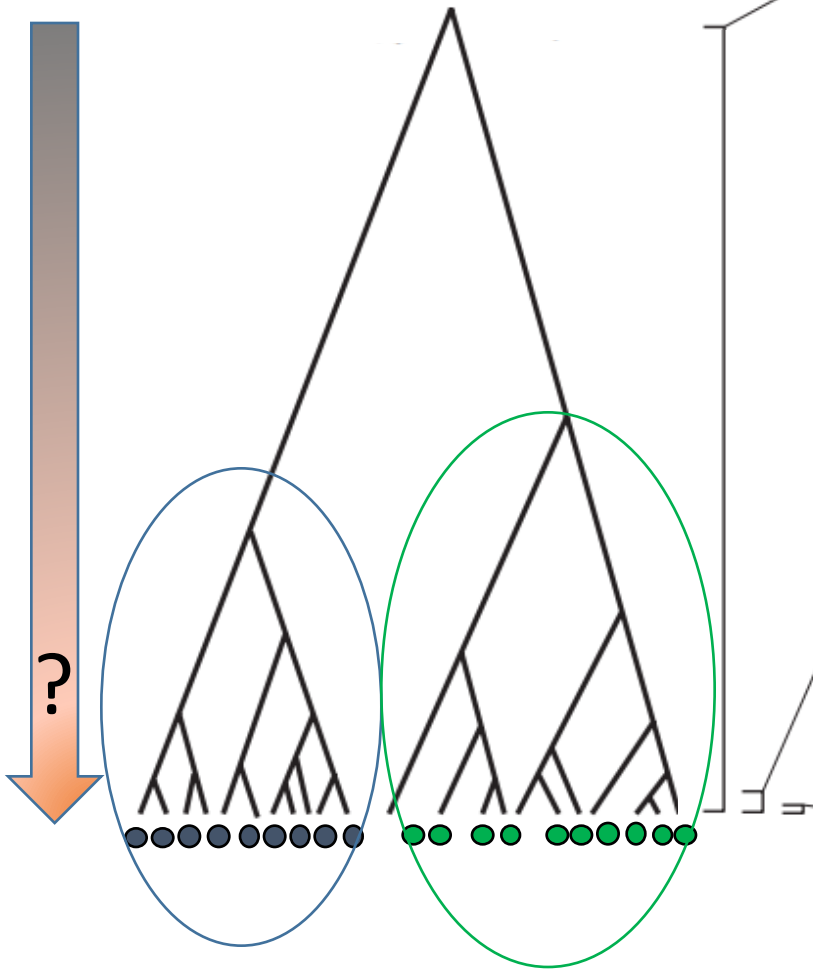
GWAS pros/cons

Many meiotic events….

(2) Population stratification

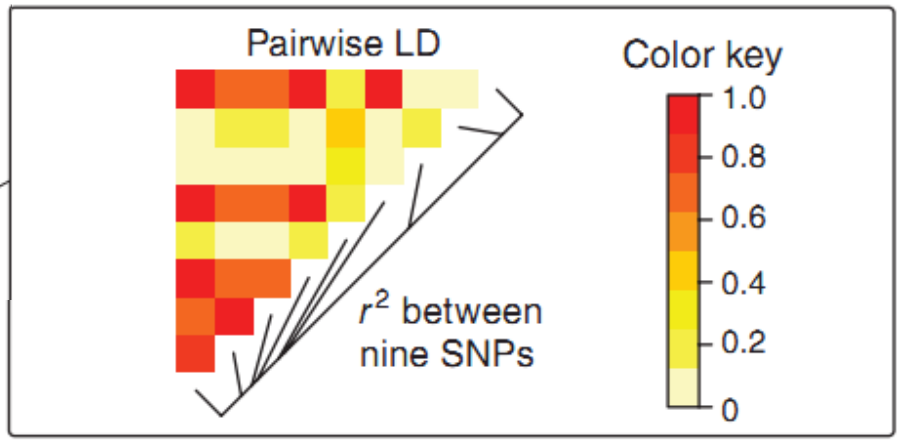
(1) Cryptic relatedness

?

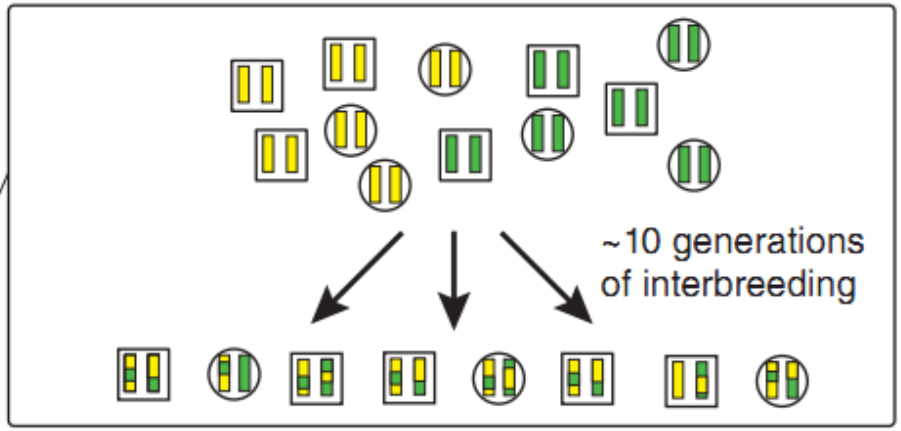


…with a risk false positive

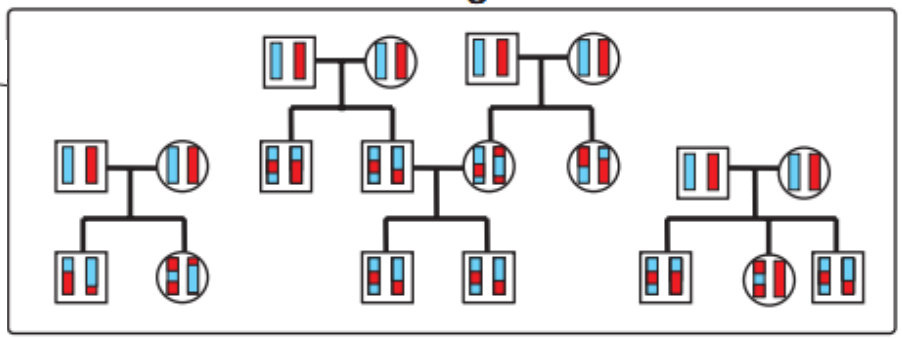
LD



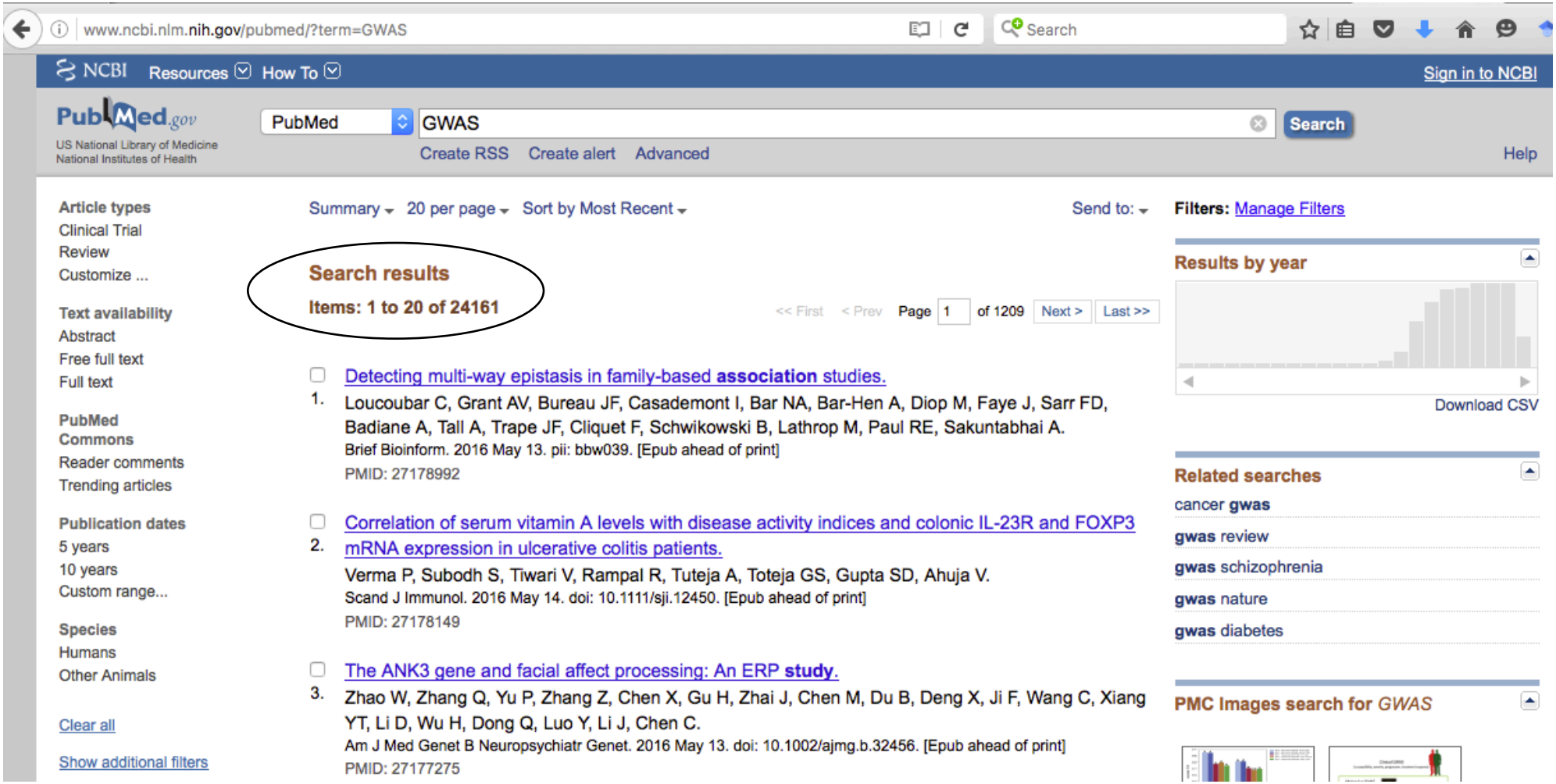
Admixture



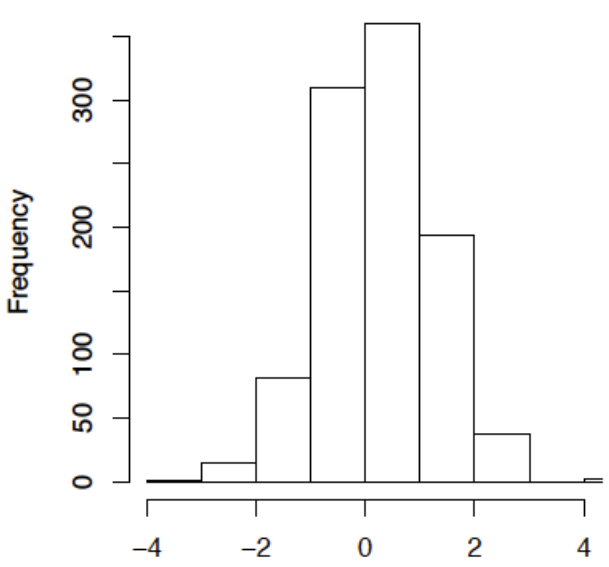
Pedigree

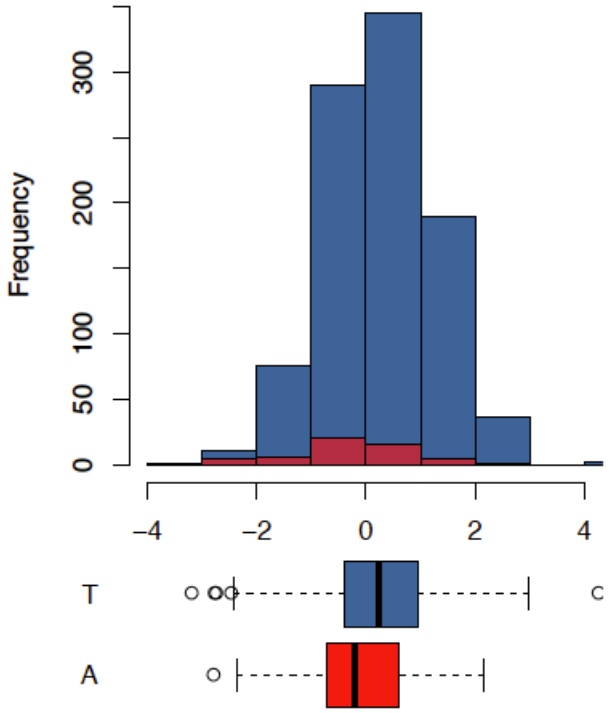


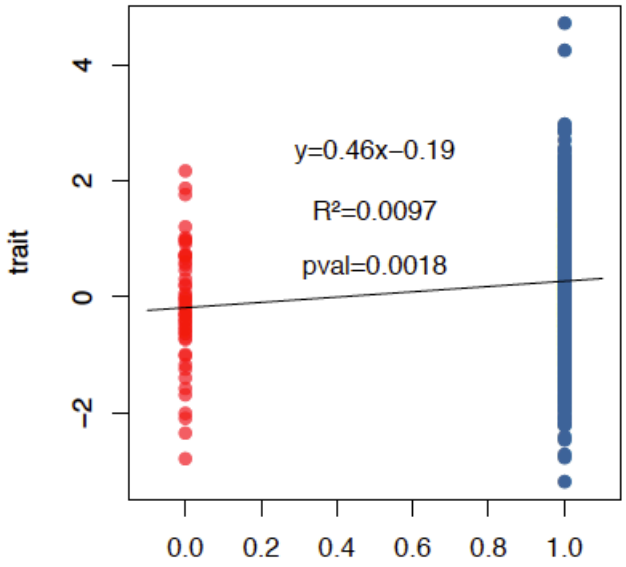
Introduction: GWAS



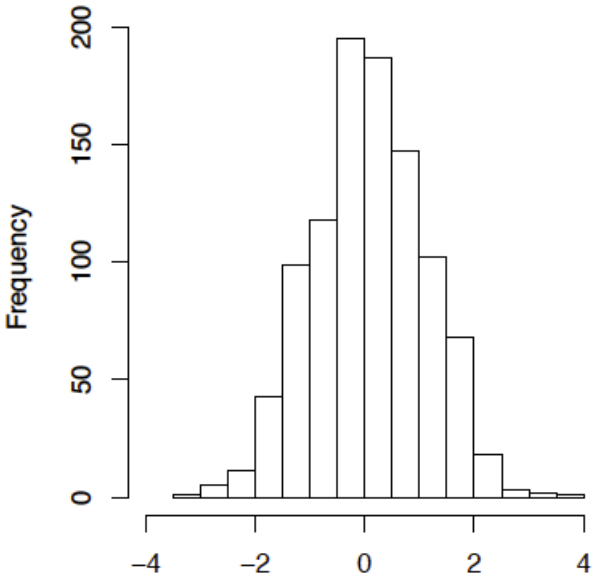
$$N = 1000, SNP = A/T, MAF = 0.05$$

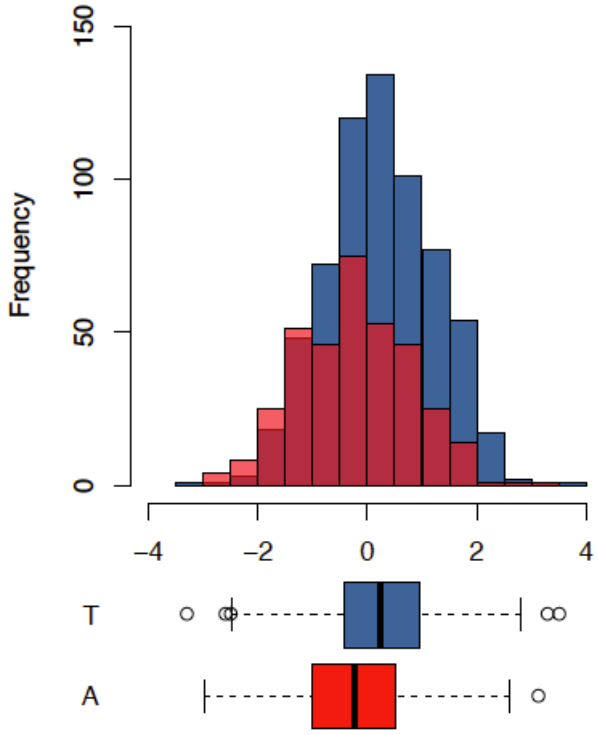


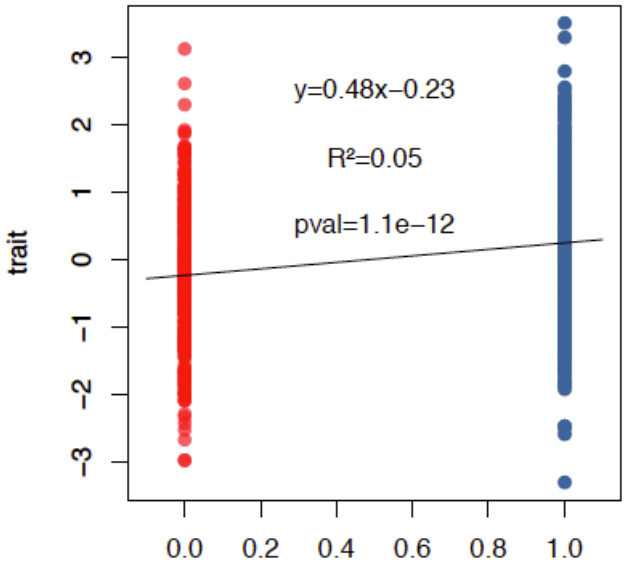




$$N = 1000, SNP = A/T, MAF = 0.35$$



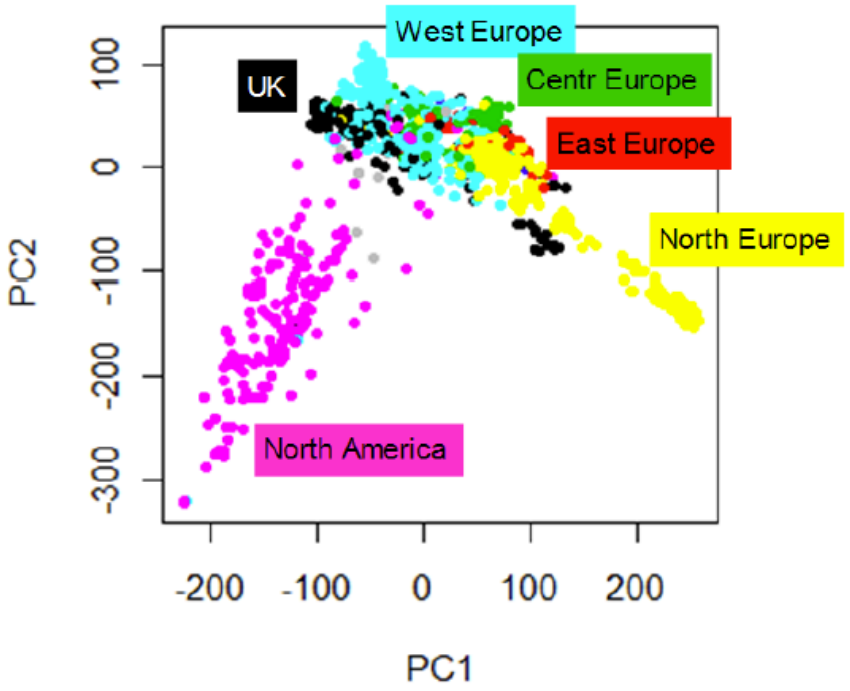




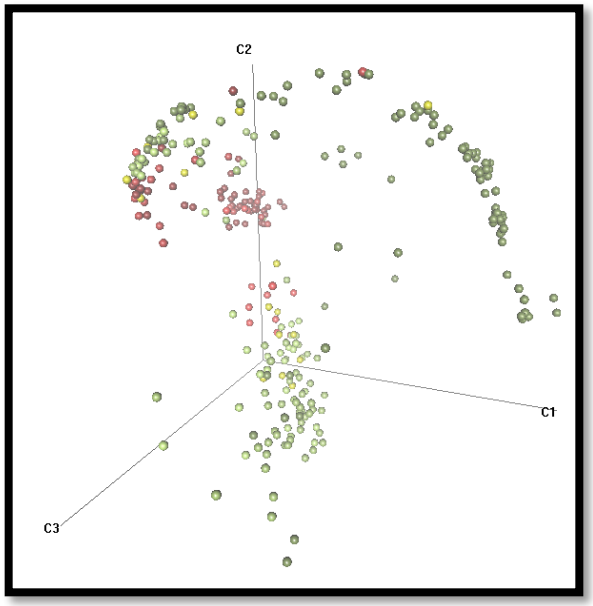
- Isolation by distance: Individuals are not independent

-> Accessions tend to cluster in subpopulations according to geographic origins

Ex: arabidopsis



Ex: tomato



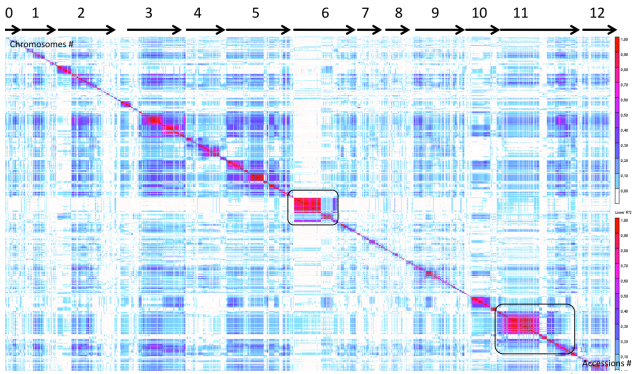
⇒ Allelic frequencies varies between subpopulations and generate long range disequilibrium

Linkage disequilibrium

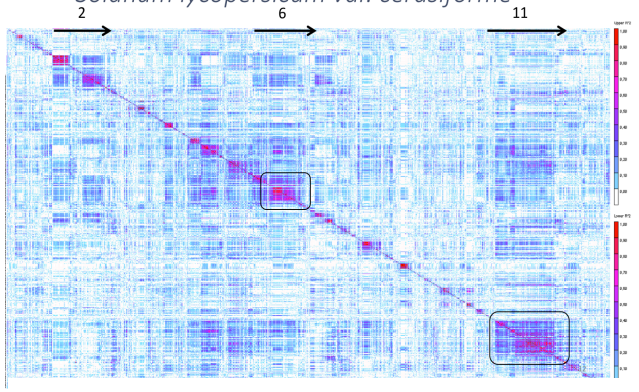
⇒ Cause/Consequence of population structure:

Linkage disequilibrium varies according to genetic groups

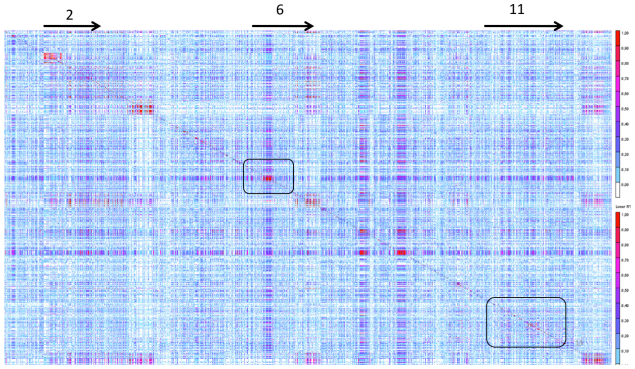
Solanum lycopersicum



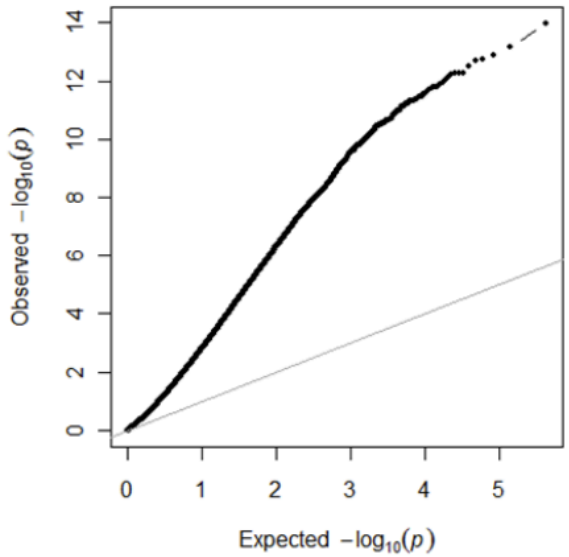
Solanum lycopersicum var. cerasiforme



Solanum pimpinellifolium

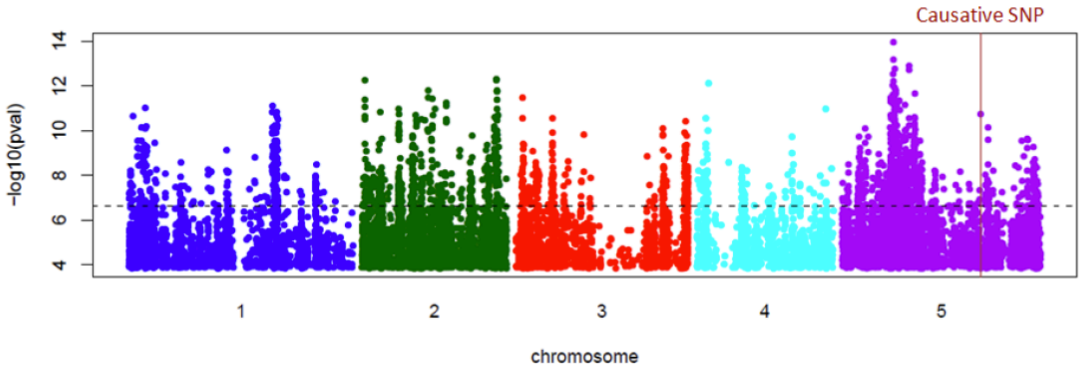


QQ-plot



⇒ Statistical test is inflated

Manhattan -plot



⇒ False positives are generated

Model

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$,
$\mathbf{u} \sim N(0, \sigma_g^2 \mathbf{K})$ & $\epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I})$,
$\mathbf{K}$ is a relationship matrix, e.g. IBD or IBS

- Proposed in association mapping by Yu *et al,* (2006)

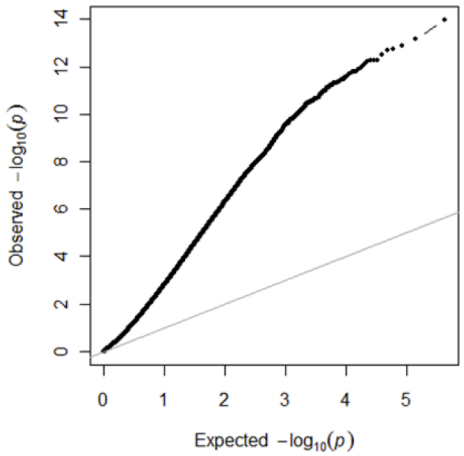
- Implemented in the EMMA R library (Kang *et al,* 2008): slow with large datasets

- Approximate method:
  - -GRAMMAR (Aulchenko *et al*, 2007 http://www.genabel.org/news20120917)
  - -EMMAX (Kang *et al,* 2010 http://csg.sph.umich.edu/kang/emmax/)
  - -P3D (Zhang *et al,* 2010 http://www.maizegenetics.net/tassel)
- Exact method:
  - -FMM (Astle & Balding, http://www.genabel.org/MixABEL/FastMixedModel.html)
  - -Fast-LMM (Lippert et al, 2011 http://research.microsoft.com/en-us/downloads)
  - -GEMMA (Zhou et al, 2012) http://www.xzlab.org/software.html
  - Etc….

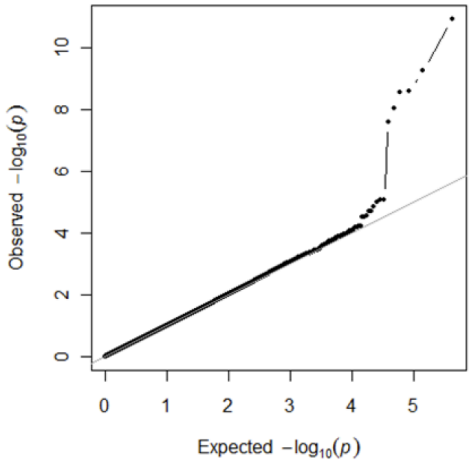
Tools using mixed modeling

QQ-plot

Linear Regression

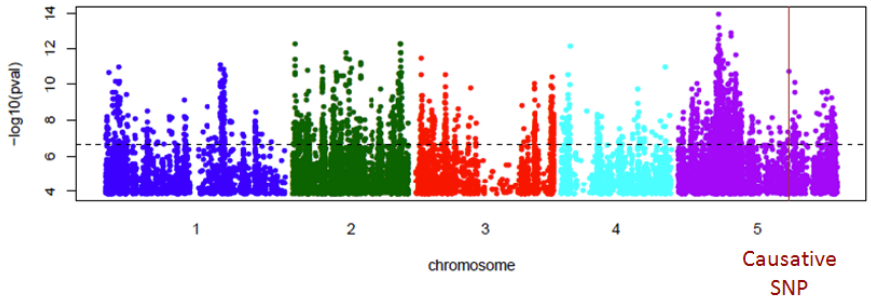


EMMAX

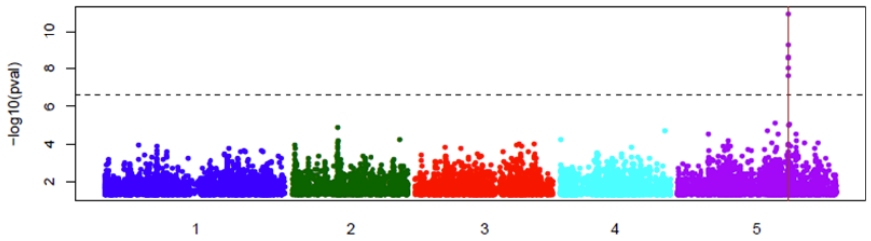


Manhattan -plot

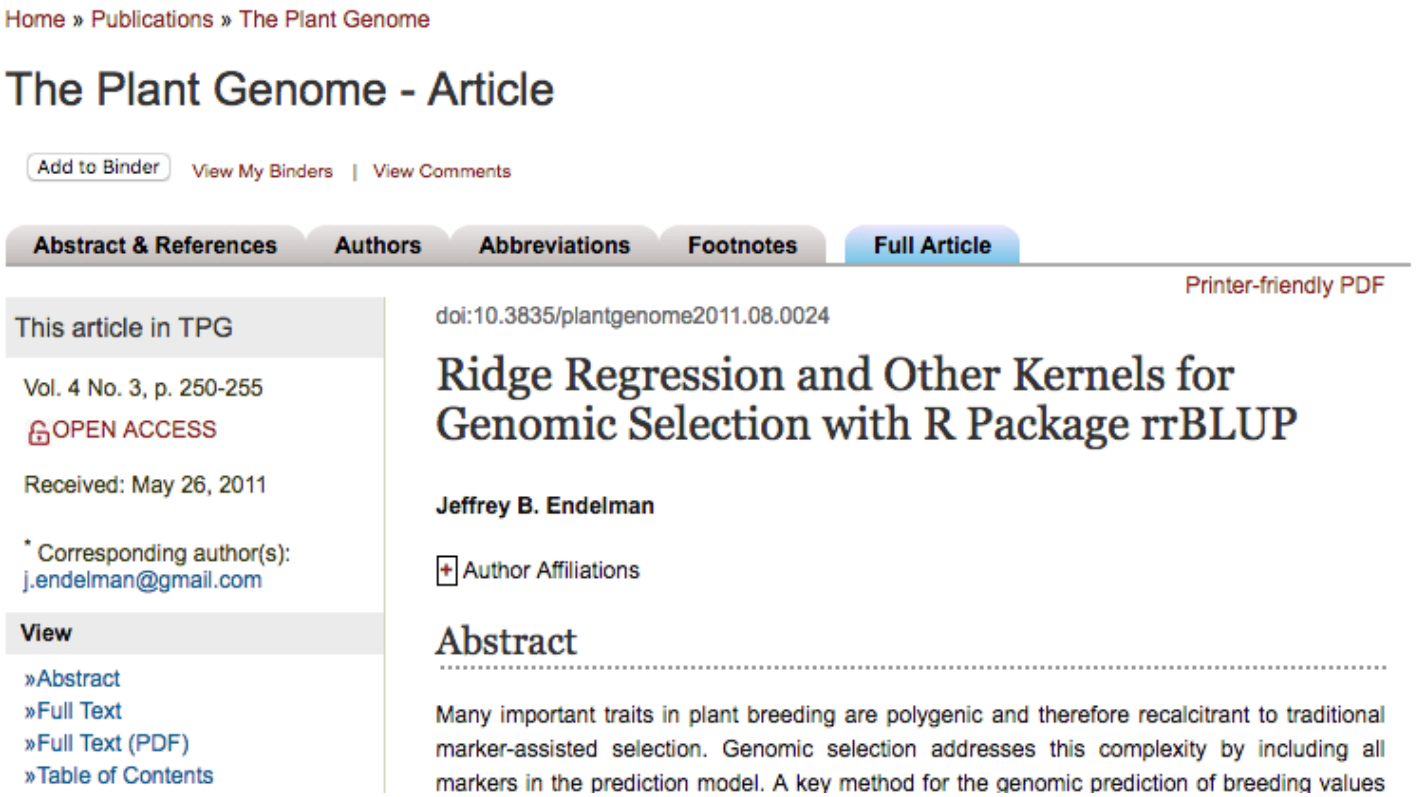
Linear Regression



EMMAX



RR-BLUP package



Home » Publications » The Plant Genome

The Plant Genome - Article

Add to Binder View My Binders | View Comments

Abstract & References | Authors | Abbreviations | Footnotes | Full Article

Printer-friendly PDF

This article in TPG

Vol. 4 No. 3, p. 250-255

OPEN ACCESS

Received: May 26, 2011

* Corresponding author(s):
j.endelman@gmail.com

View

»Abstract
»Full Text
»Full Text (PDF)
»Table of Contents

doi:10.3835/plantgenome2011.08.0024

Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP

Jeffrey B. Endelman

Author Affiliations

Abstract

Many important traits in plant breeding are polygenic and therefore recalcitrant to traditional marker-assisted selection. Genomic selection addresses this complexity by including all markers in the prediction model. A key method for the genomic prediction of breeding values

Our test dataset:

Oat dataset on yield

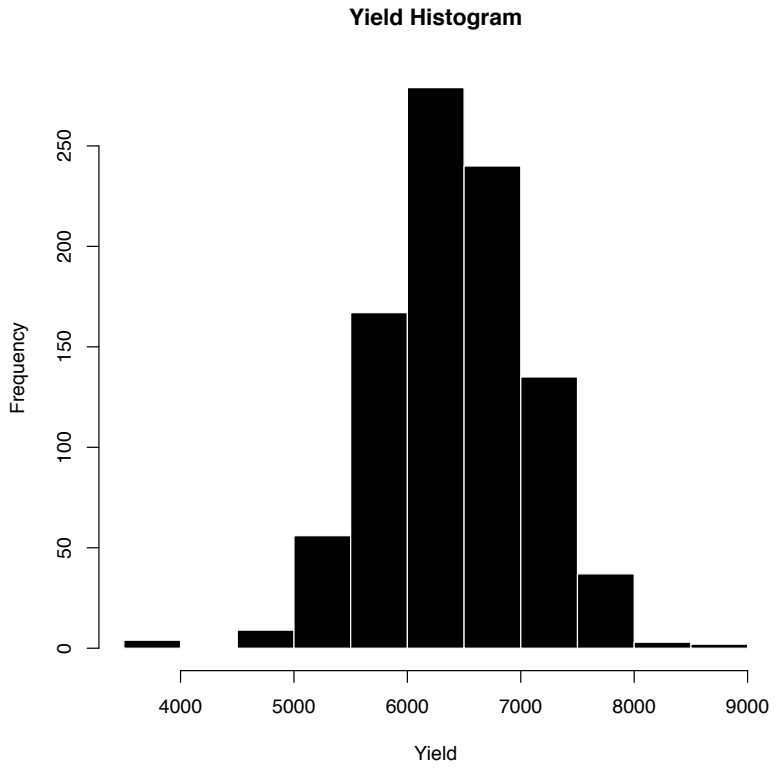
More infos:
https://www.rdocumentation.org/packages/rrBLUP/versions/1.1/topics/RR.BLUP

Hands on # 2: MLM

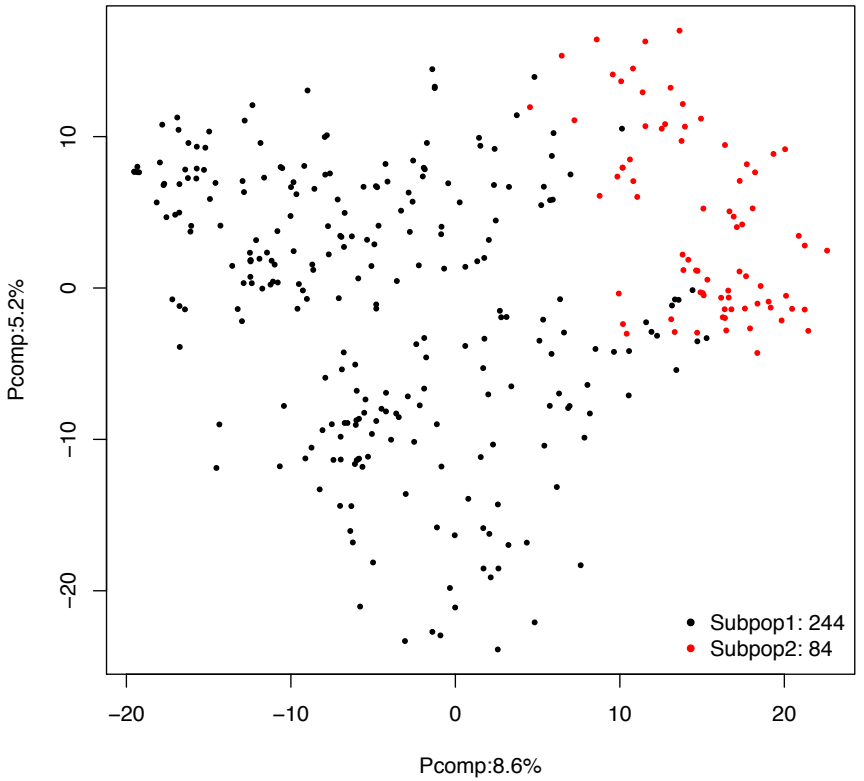
Steps and Questions

#Step 1: Phenotypes- Cleaning	-> generate histogram -> how many observations after filtering
#Step 2: Genotypes- Filtering	
#Step 3: Imputation	-> how many SNP after filtering? -> how many individuals after filtering?
#Step 4: Checking population structure effects	-> Kinship matrix dimensions? -> Variance proportion explained by PC1, PC2 -> Scree plot and PCA plot
#Step 5: Matching phenotype and genotype	-> Dimensions pheno.GWAS object?
#Step 6: Matching genotype and map file	-> Dimensions geno.GWAS object?
#Step 7: Analysis: running the model	-> Generate the 4 models -> What are the differences?
#Step 8: QQ-Manhattan and Correlation Plots	-> Where are the main peaks?
#Step 9: Which are the hits?	-> Which test is more conservative?

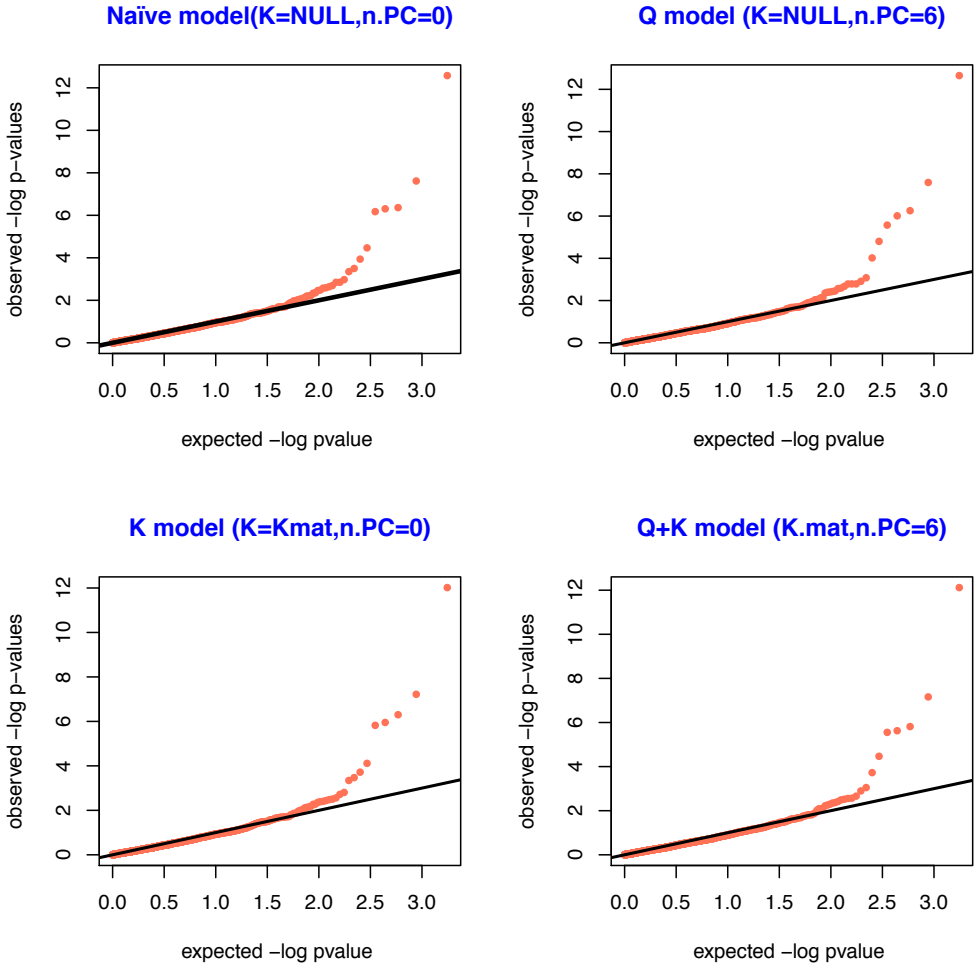
1) Check data distribution and residual homogeneity



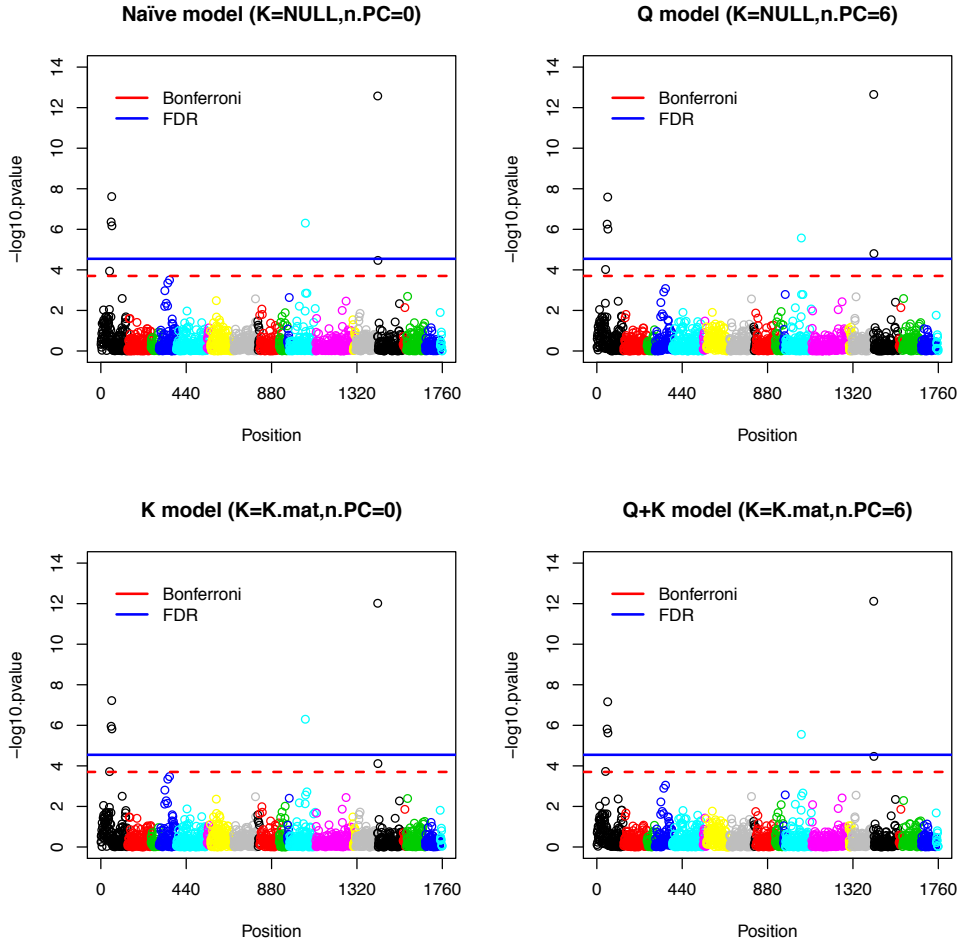
2) Check population structure



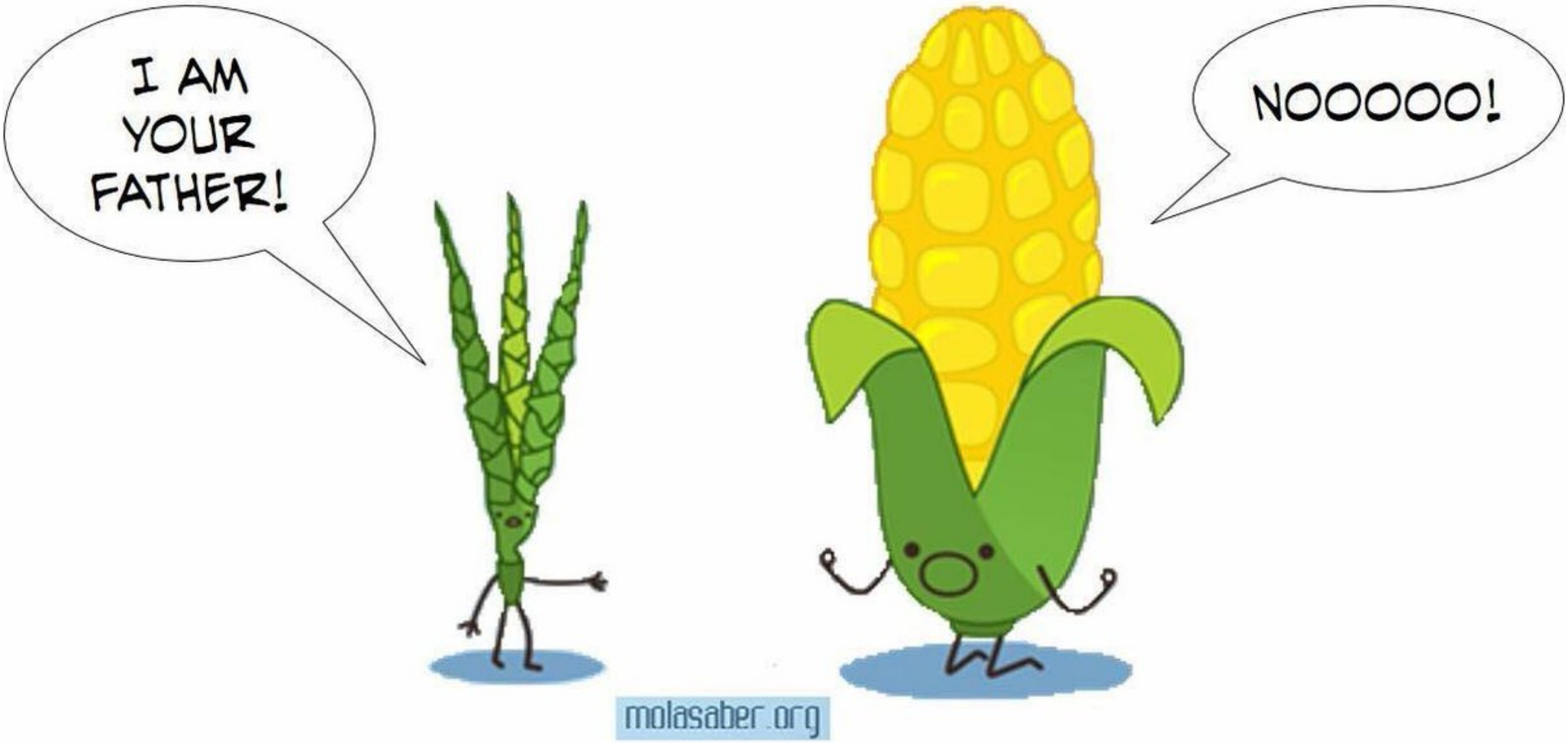
3) Fit models (QQ plot)



4) Identify associations (manhattan plot)



- Two mapping techniques with distinct designs but similar goals
- Both can be applied in various research context a molecular data types (SNP chip, GBS, WGS) , phenotypes (metabolome, RNAseq, agronomic traits)
- Choice of one or the other depends Trait architecture and available resources
- Generic approaches that can be customized accordingly


I AM YOUR FATHER!
NOOOOO!
molasaber.org

NATURE GENETICS | TECHNICAL REPORT

日本語要約

An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations

Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long & Magnus Nordborg

Affiliations | Contributions | Corresponding author

Nature Genetics **44**, 825–830 (2012) | doi:10.1038/ng.2314

Received 16 November 2011 | Accepted 04 May 2012 | Published online 17 June 2012

Model:

Segura et al, Nat. Gen., 2012

Multi-locus mixed-model (MLMM)

R

Dataset:

Sauvage et al, Plant Phys. 2014

-> 160 tomato accessions

-> 6000 SNP

-> 20 metabolic traits

Plant Physiology

HOME | ABOUT | SUBMIT | SUBSCRIPTIONS | ADVERTISE | ARCHIVE | CONTACT US

« Previous | Table of Contents: GENES, DEVELOPMENT, AND... | Next Article »

EXPAND

Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits[1][W]

Christopher Sauvage*, Vincent Segura, Guillaume Bauchet, Rebecca Stevens, Phuc Thi Do, Zoran Nikoloski, Alisdair R. Fernie and Mathilde Causse

+ Author Affiliations

*Address correspondence to christopher.sauvage@avignon.inra.fr.

First Published on June 3, 2014, doi: http://dx.doi.org/10.1104/pp.114.241521

Plant Physiology July 2014 vol. 165 no. 3 1120–1132

Multi locus Model approach

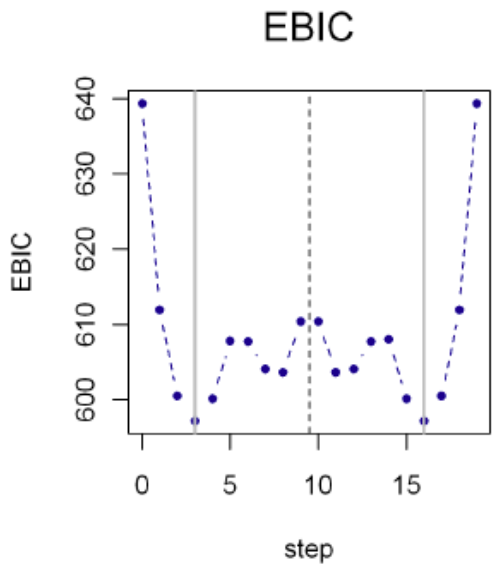
Testing all combinations is not possible...

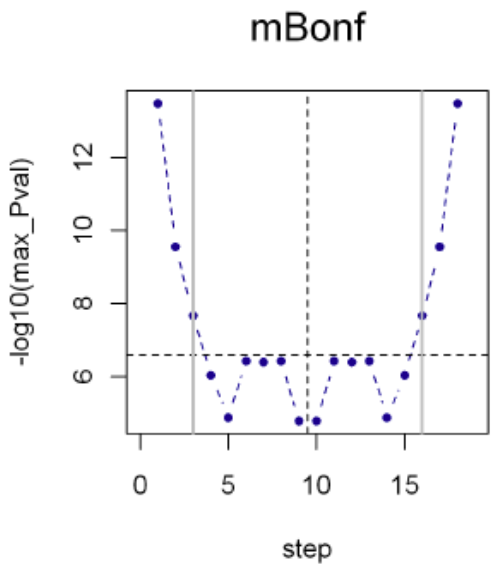
Proposed approach: Decreasing the model space complexity with stepwise regression while controlling population structure confounding with the mixed-model

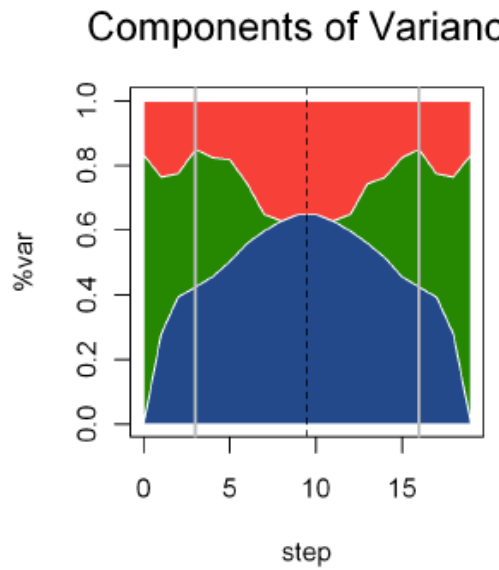
1. Null model (step 0): $\mathbf{y} = \beta_0 + \mathbf{Zu} + \epsilon$
2. Forward inclusion (forward steps 1 to i):
   - Do a GWAS scan with EMMAX, using the REML estimates $\hat{\sigma}^2_g$ & $\hat{\sigma}^2_\epsilon$ from the mixed-model at step $i-1$
   - Select x_i, the most significant SNP, to be included as a cofactor in the mixed-model at step i: $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$
3. Backward elimination (backward steps 1 to i): Remove iteratively the least significant cofactor form the last forward mixed-model until the null model.

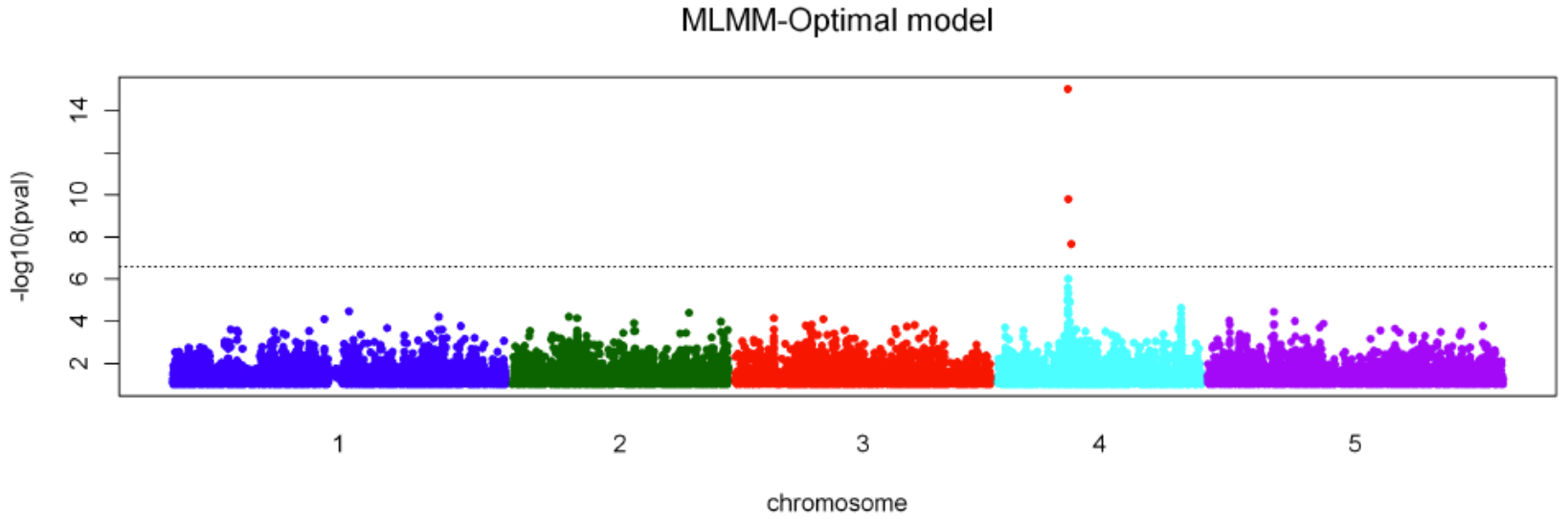
Multi locus Model: multiple testing

- Forward stopping criterion: $\hat{h}^2 = \hat{\sigma}^2_g/(\hat{\sigma}^2_g + \hat{\sigma}^2_\epsilon)$ close to 0
- Best Model Identification:
  - BIC (Schwarz, 1978)
  - Extended BIC: $EBIC = BIC + 2logC^k_m$, where C^k_m is the dimension of the model space (Chen and Chen, 2010)
  - *mBonf*: Select the least parsimonious model whose cofactors are all significant at a threshold of 5% Bonferroni corrected


EBIC
EBIC
600
610
620
630
640
0
5
10
15
step


mBonf
-log10(max_Pval)
6
8
10
12
0
5
10
15
step


Components of Variance
%var
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
step


MLMM-Optimal model
-log10(pval)
2
4
6
8
10
14
1
2
3
4
5
chromosome

Hands on #3: MLMM

#Step 1:

Load datasets (phenotype, genotypes, relationship matrix, snp file)

#Step 2:

Perform GWAS using different models:
Test thresholds and traits:

#Step 3:

Display results using Manhattan plot, QQ-plot and histograms at different steps

#Step 4:

Draw first conclusions on trait architecture

Multi locus Model specificities: take home

- Most GWAS are based on single SNP tests ('large *p*, small *n*' problem)

- ' Wrong model' for phenotypic traits with a polygenic inheritance

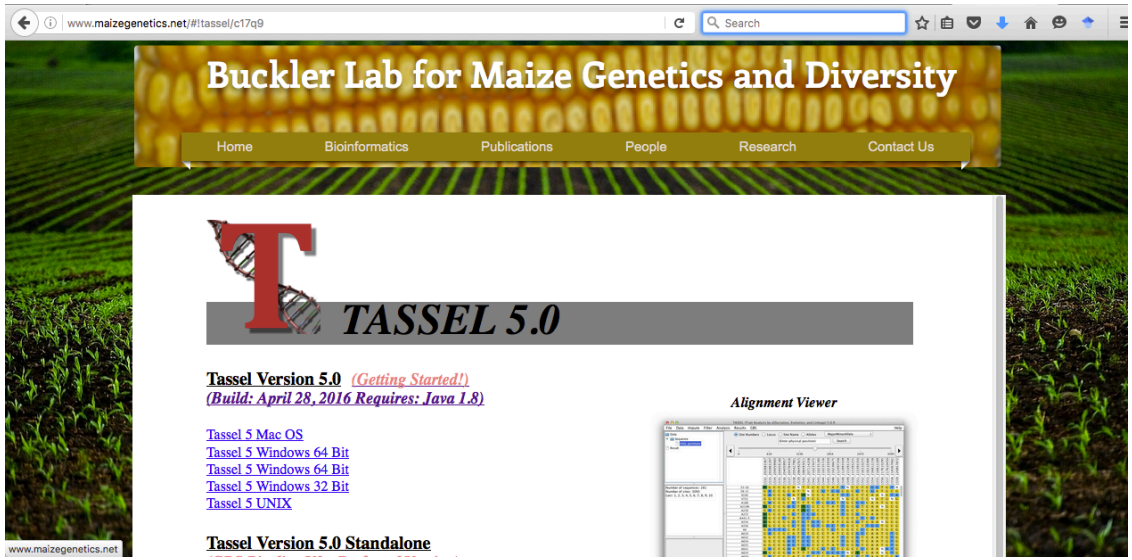
- Similar approach as composite interval mapping in QTL mapping (Jansen & Stam, 1994)

- Increase power

- Clarify/solve synthetic association

Hands on #4: TASSEL (optional)

http://www.maizegenetics.net/tassel



Test dataset :

1-Genotypes*:

GWAS_genotype.map

GWAS_genotype.ped

2- Phenotypes*:

GWAS_phenotypes_tassel

3- Relationship matrix + covariates

GWAS_structureQ1_tassel

GWAS_kinship_tassel