

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

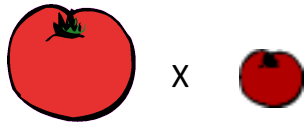
Topic(s)

- 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>

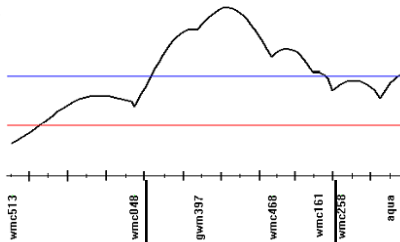
Introduction

QTL mapping

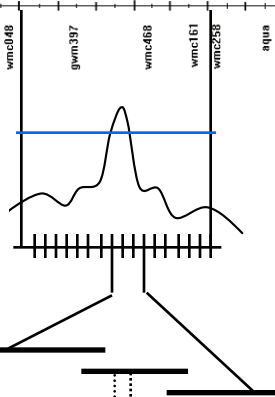
Mapping populations: F2, BC, DH



QTL mapping



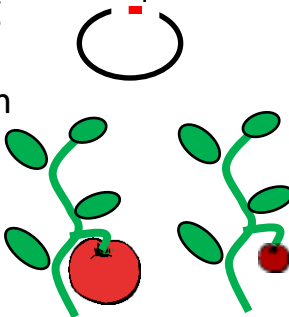
Fine mapping



Physical mapping

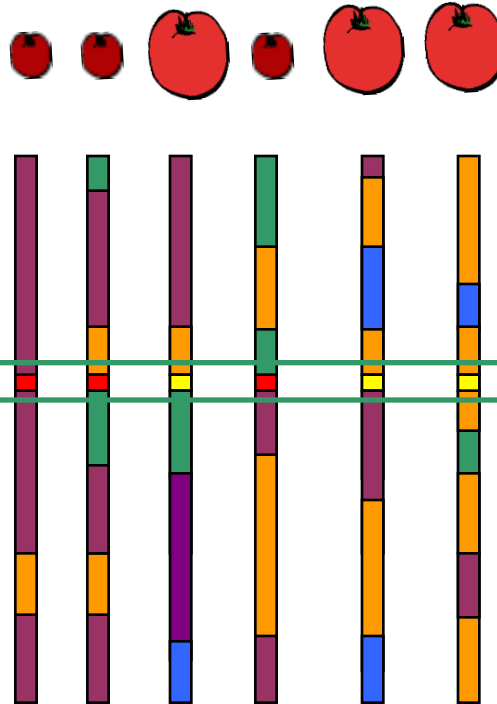
Positional cloning

Functional validation

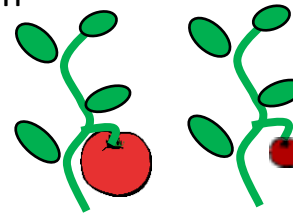


Association mapping (GWAS)

Natural populations

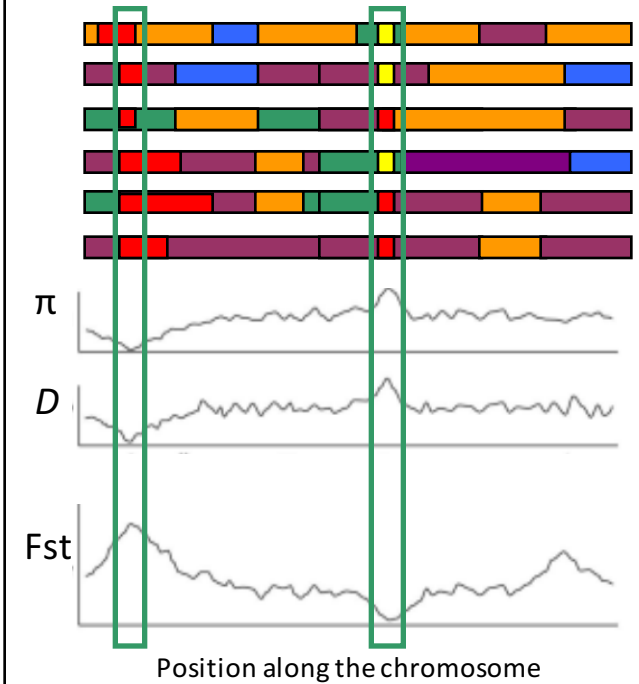


Functional validation



Population genomics

Natural populations



Nucleotide diversity,

Tajima's D ,

Fixation index,

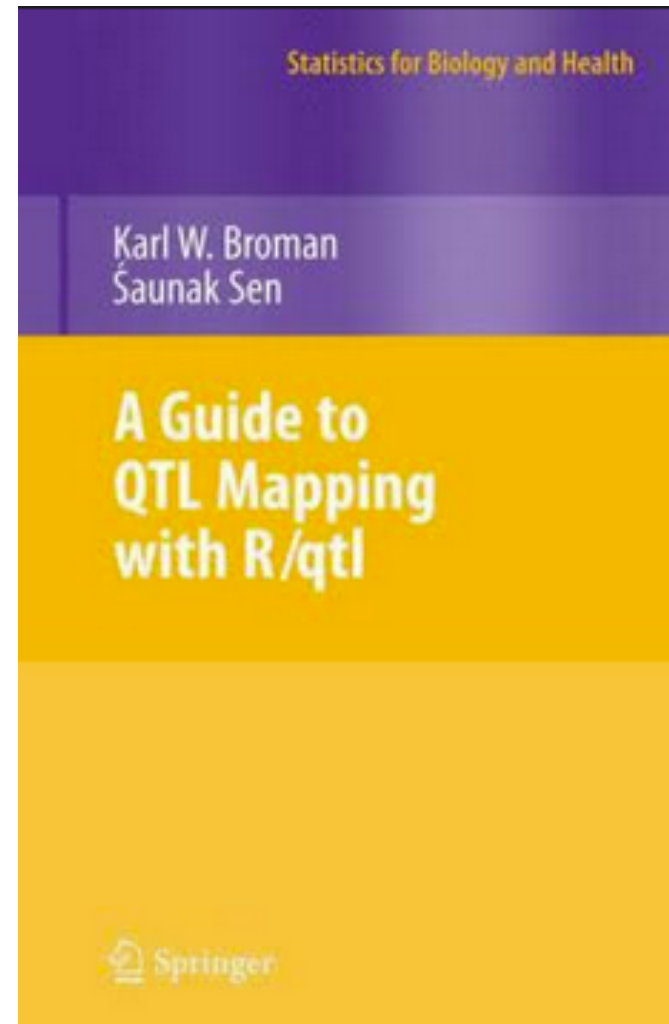
dN/dS ,

etc...

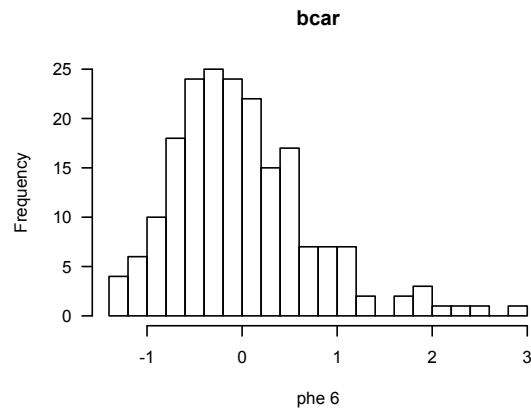
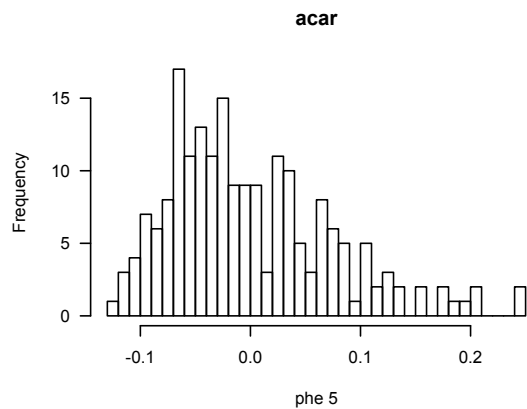
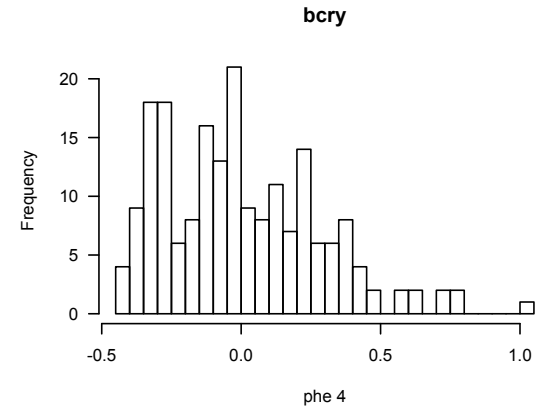
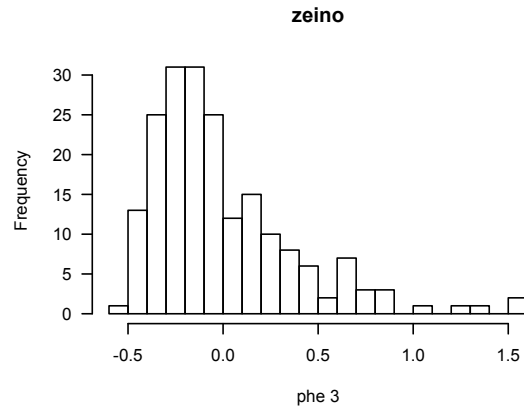
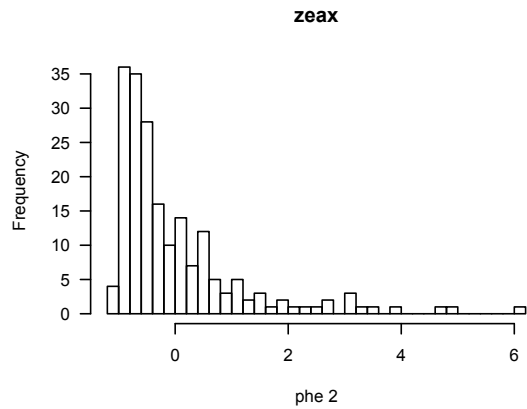
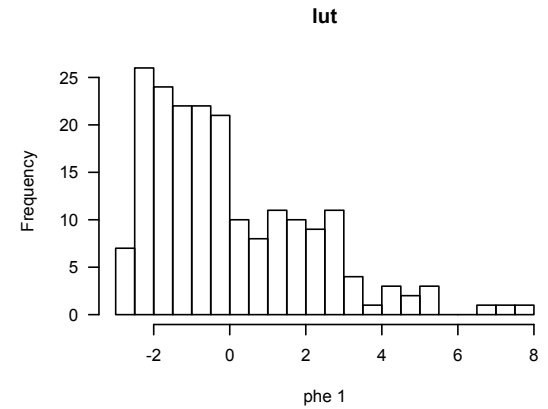
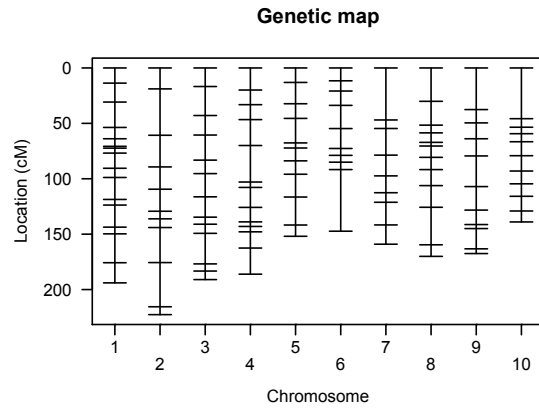
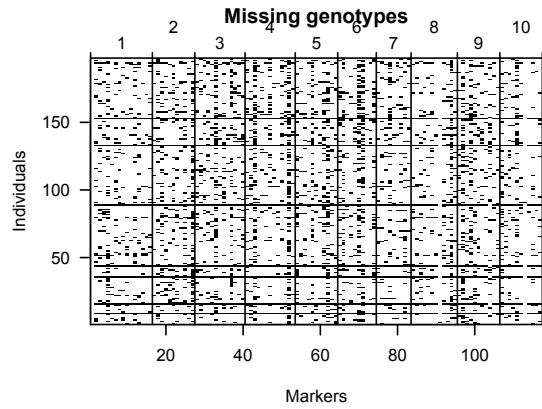
<http://www.rqtl.org/>

Our test dataset:

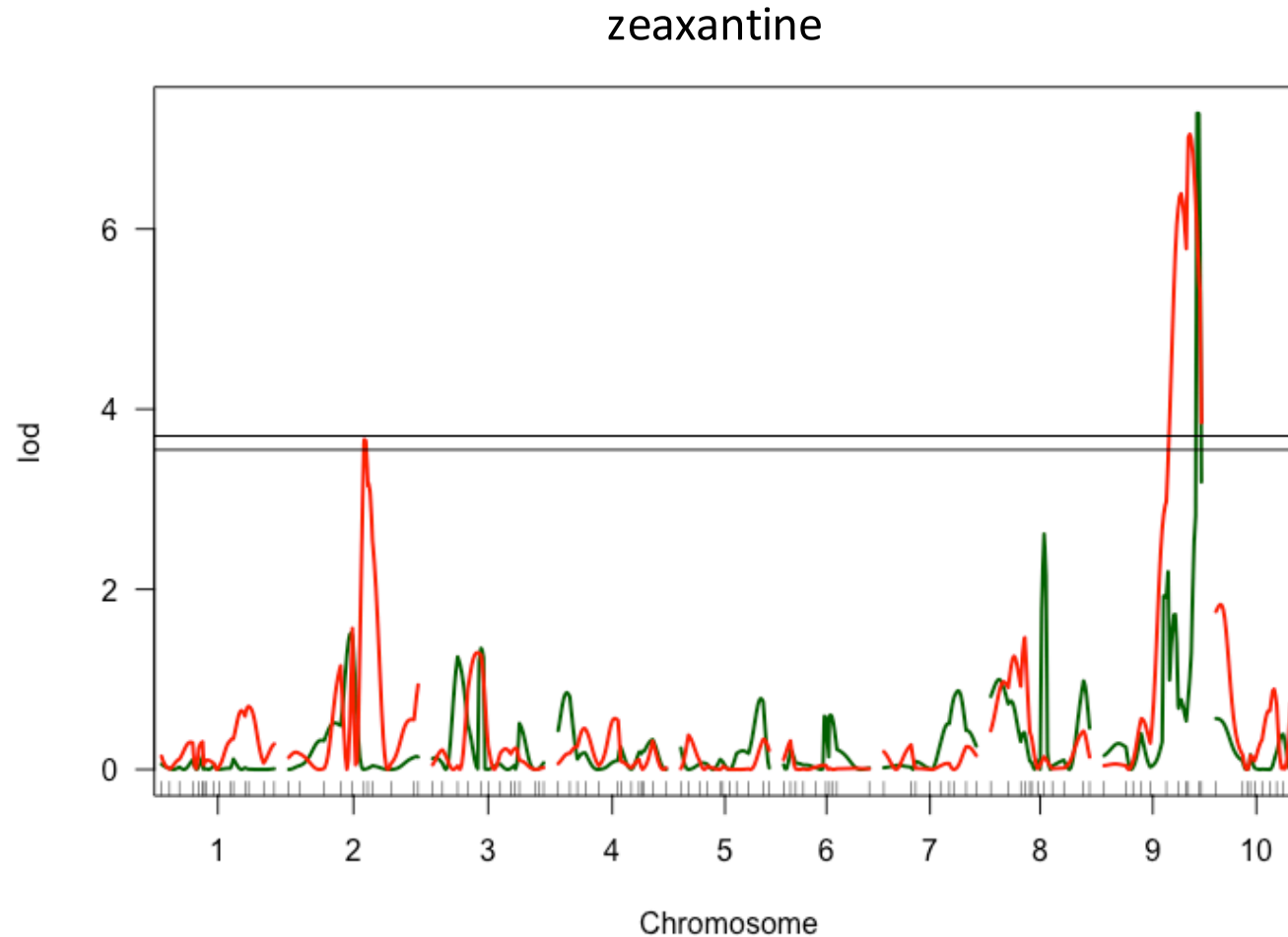
Maize dataset on carotenoid traits



Hands on #1: QTL analysis



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



Step 1: Importing data

-> "ASCarotFamily091014.csv"

Step 2: Data checking

-> How many individuals? Markers? Phenotypes?

-> Type of cross?

-> Generate physical map and trait distribution plot

-> 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

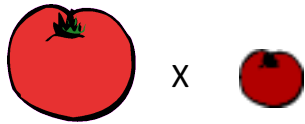
-> How are trait distributions?

Step 6: QTL effects

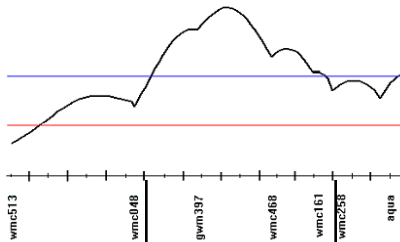
-> display effect plot, any interactions?

QTL mapping

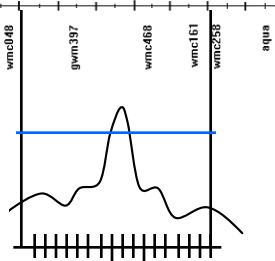
Mapping populations: F2, BC, DH



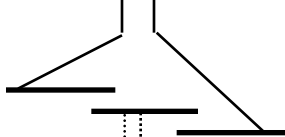
QTL mapping



Fine mapping



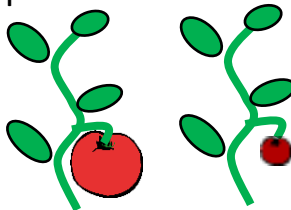
Physical mapping



Positional cloning

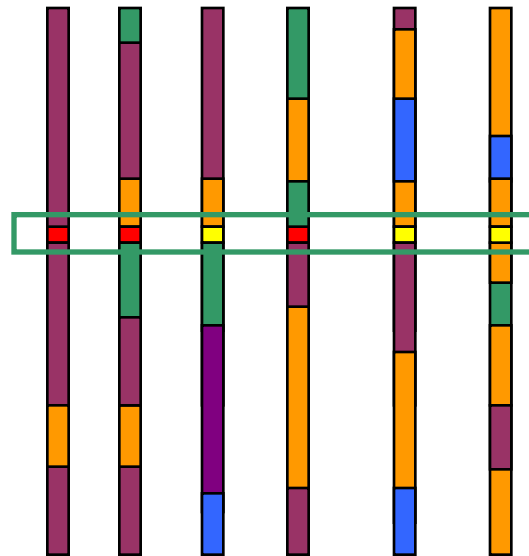


Functional validation

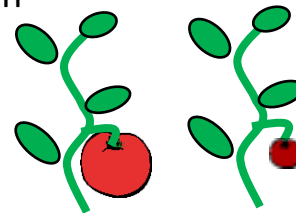


Association mapping (GWAS)

Natural populations

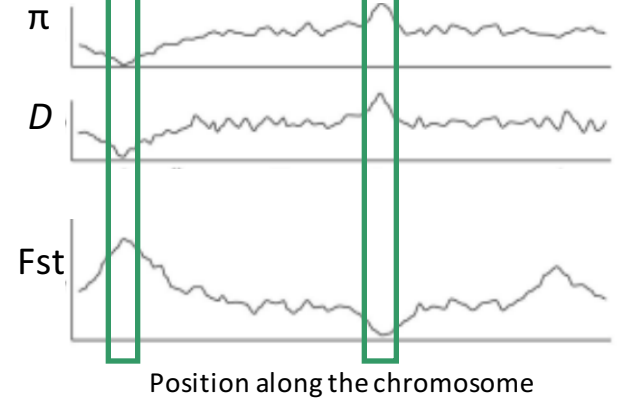
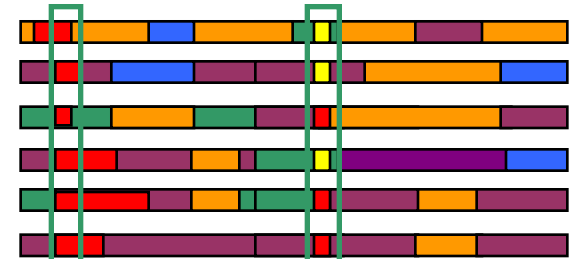


Functional validation



Population genomics

Natural populations



Nucleotide 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?

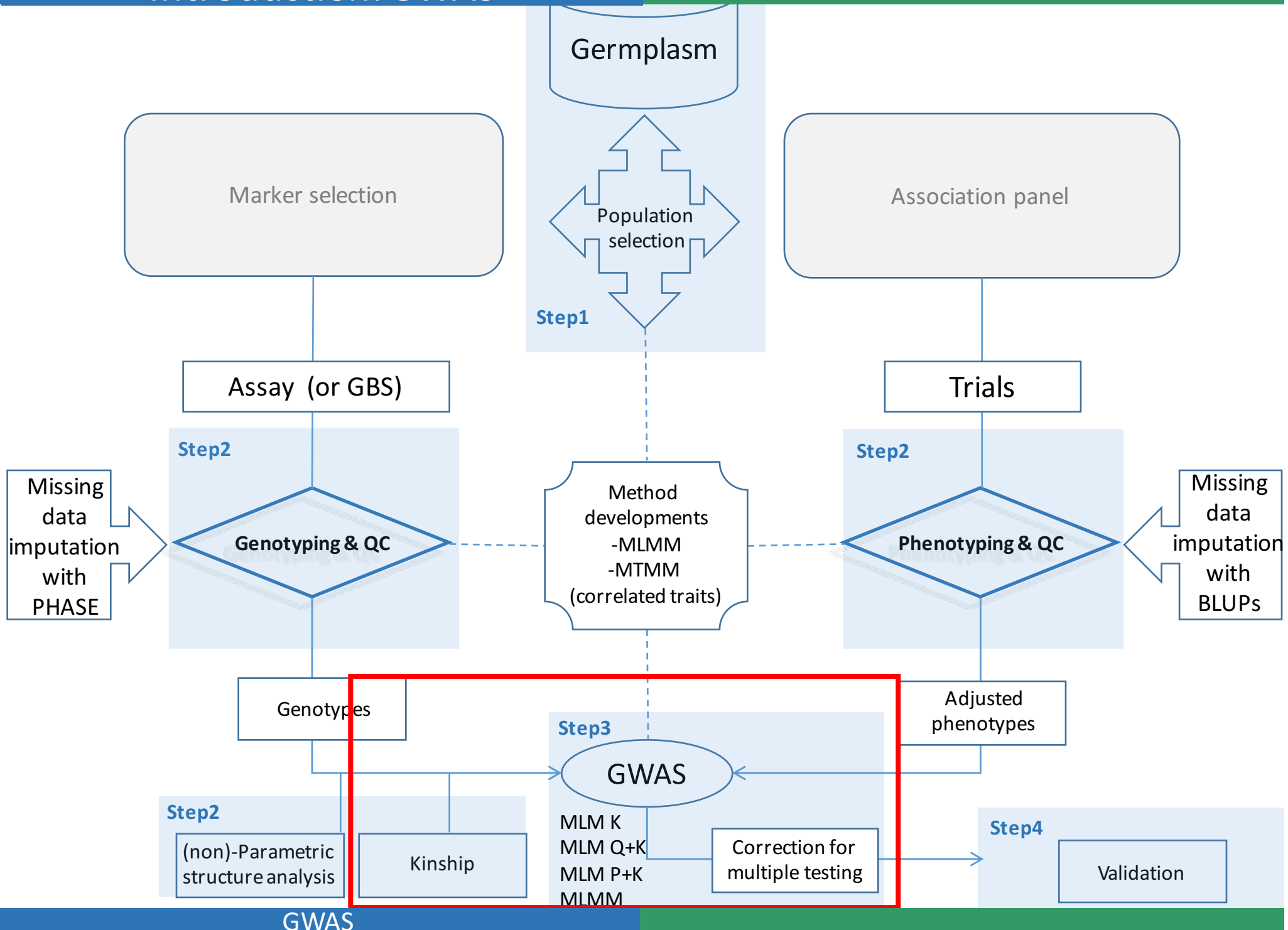
Combine Genotypic and phenotypic data within a population

Population

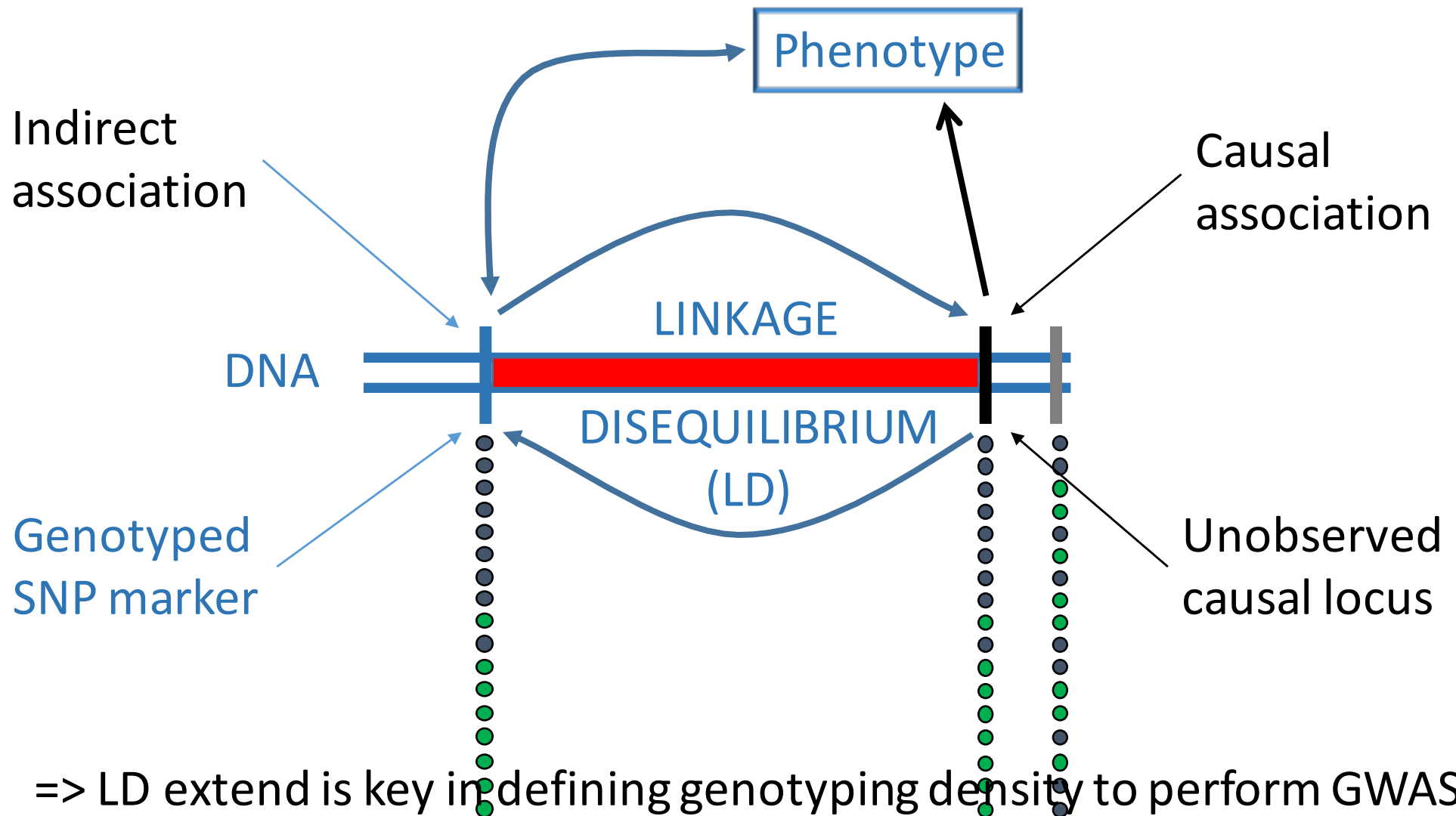
Genotype

Phenotype

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

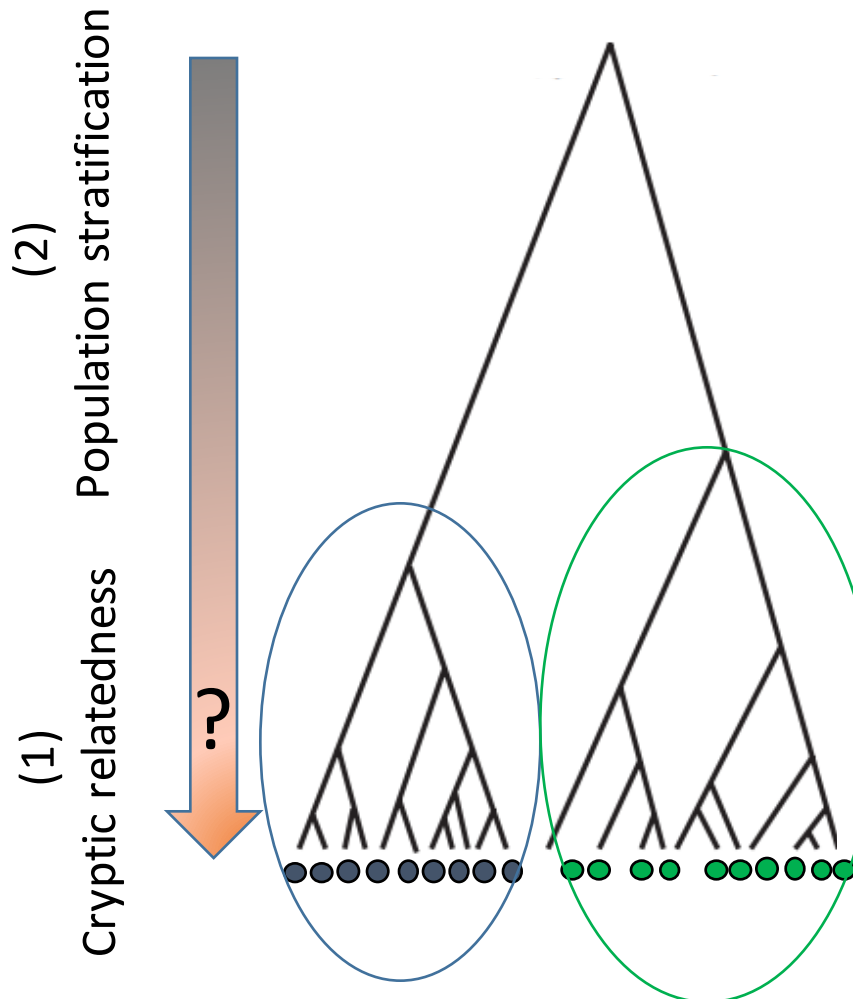


linkage study in a natural population

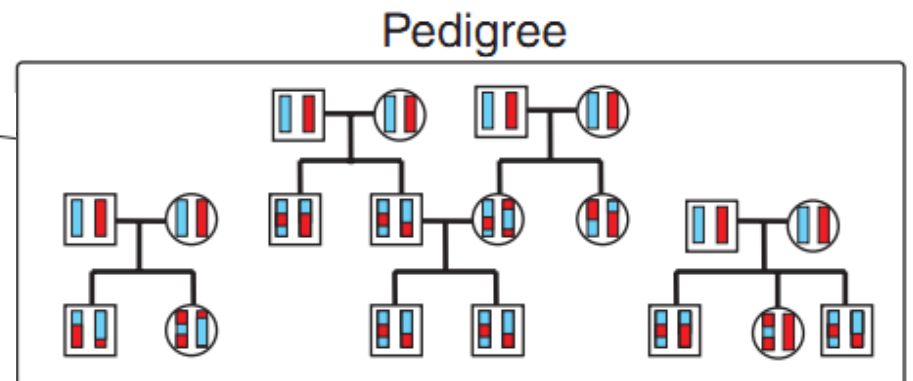
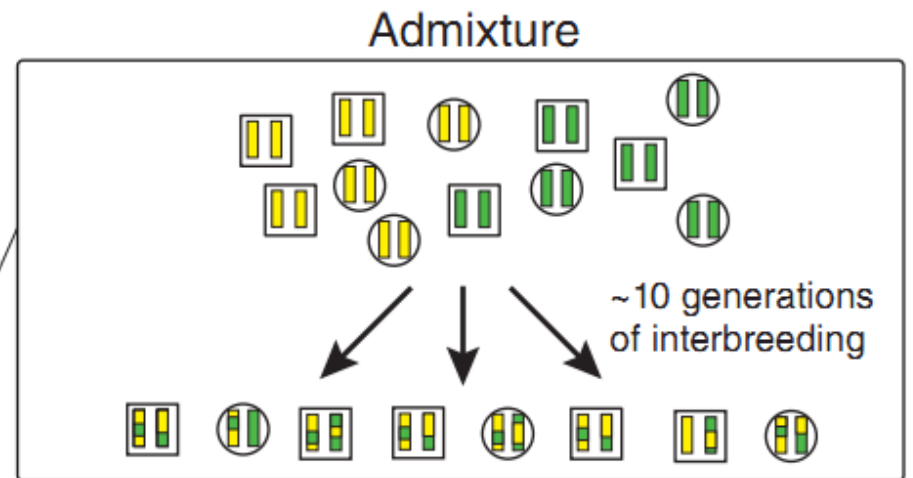
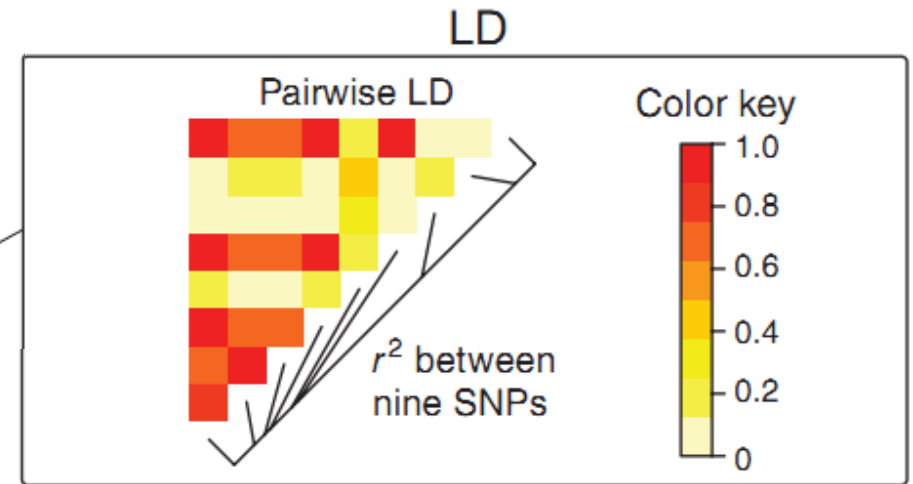


GWAS pros/cons

Many meiotic events....



...with a risk false positive



Introduction: GWAS

www.ncbi.nlm.nih.gov/pubmed/?term=GWAS

NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed GWAS Search

Create RSS Create alert Advanced Help

Article types: Clinical Trial, Review, Customize ...

Text availability: Abstract, Free full text, Full text

PubMed Commons: Reader comments, Trending articles

Publication dates: 5 years, 10 years, Custom range...

Species: Humans, Other Animals

[Clear all](#)

[Show additional filters](#)

Summary 20 per page Sort by Most Recent

Send to: Filters: [Manage Filters](#)

Search results

Items: 1 to 20 of 24161

<< First < Prev Page 1 of 1209 Next > Last >>

- ☐ [Detecting multi-way epistasis in family-based association studies.](#)
Loucoubar C, Grant AV, Bureau JF, Casademont I, Bar NA, Bar-Hen A, Diop M, Faye J, Sarr FD, Badiane A, Tall A, Trape JF, Cliquet F, Schwikowski B, Lathrop M, Paul RE, Sakuntabhai A. Brief Bioinform. 2016 May 13. pii: bbw039. [Epub ahead of print]
PMID: 27178992
- ☐ [Correlation of serum vitamin A levels with disease activity indices and colonic IL-23R and FOXP3 mRNA expression in ulcerative colitis patients.](#)
Verma P, Subodh S, Tiwari V, Rampal R, Tuteja A, Tuteja GS, Gupta SD, Ahuja V. Scand J Immunol. 2016 May 14. doi: 10.1111/sji.12450. [Epub ahead of print]
PMID: 27178149
- ☐ [The ANK3 gene and facial affect processing: An ERP study.](#)
Zhao W, Zhang Q, Yu P, Zhang Z, Chen X, Gu H, Zhai J, Chen M, Du B, Deng X, Ji F, Wang C, Xiang YT, Li D, Wu H, Dong Q, Luo Y, Li J, Chen C. Am J Med Genet B Neuropsychiatr Genet. 2016 May 13. doi: 10.1002/ajmg.b.32456. [Epub ahead of print]
PMID: 27177275

Results by year

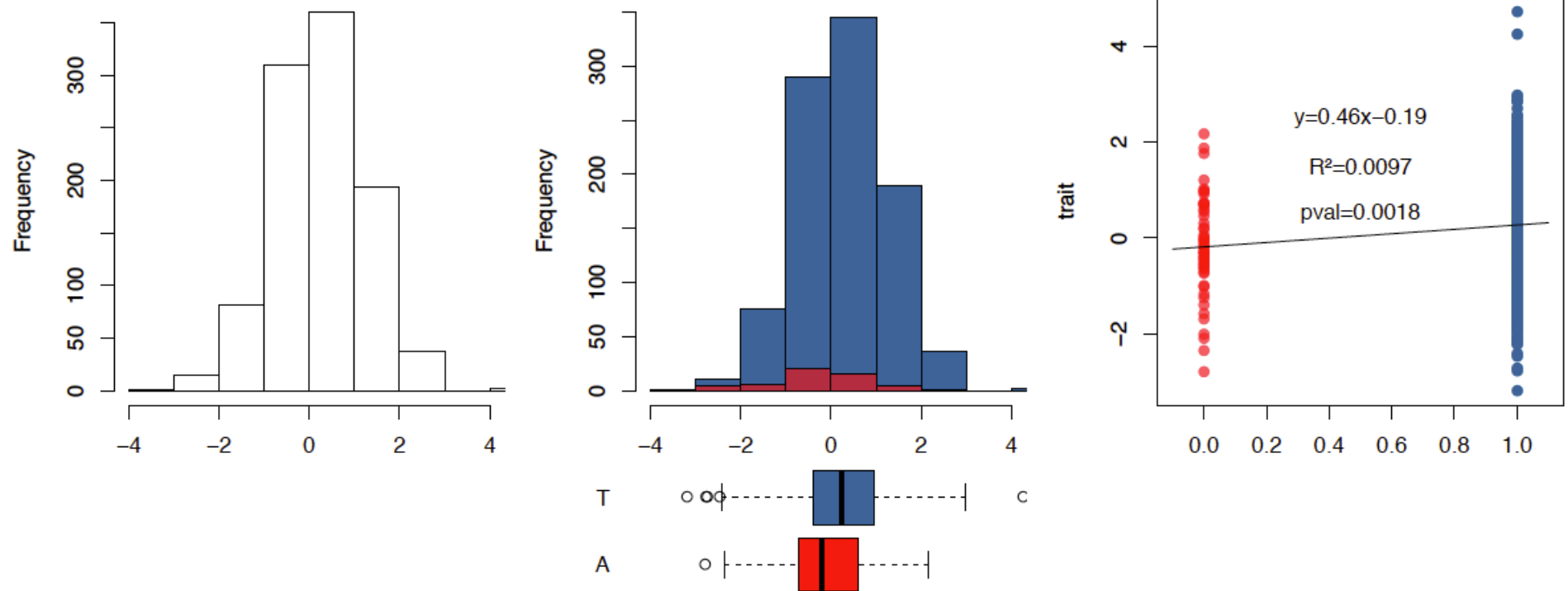
Download CSV

Related searches

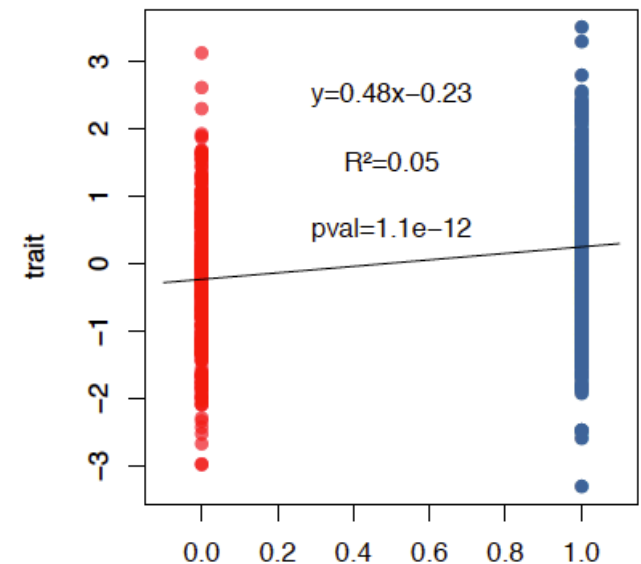
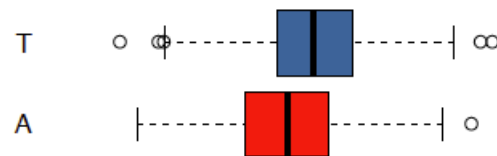
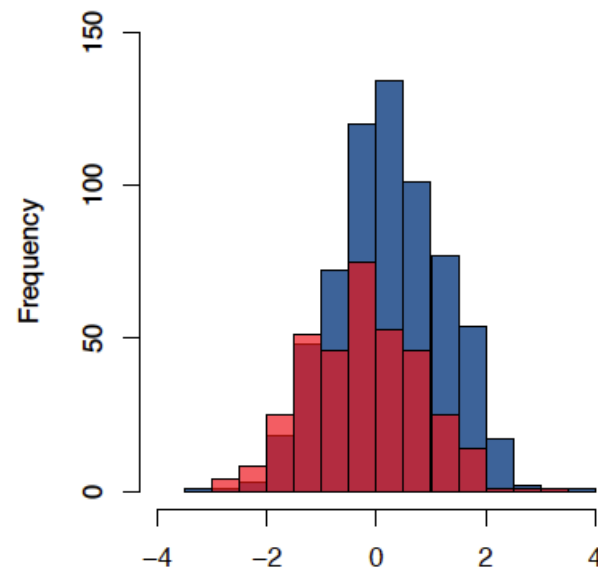
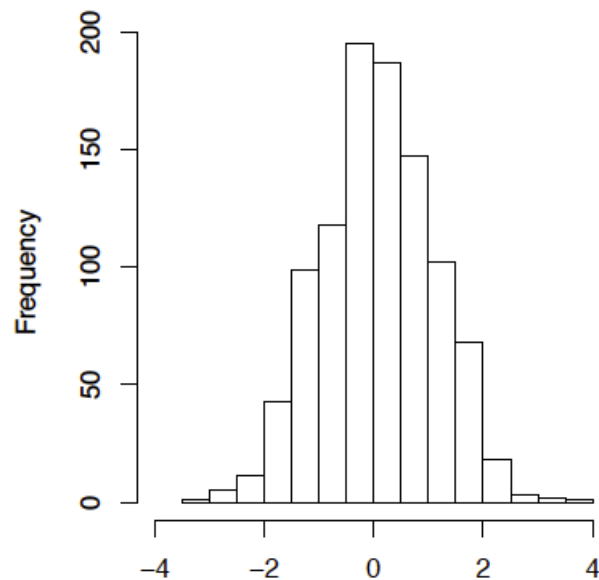
- cancer gwas
- gwas review
- gwas schizophrenia
- gwas nature
- gwas diabetes

PMC Images search for GWAS

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

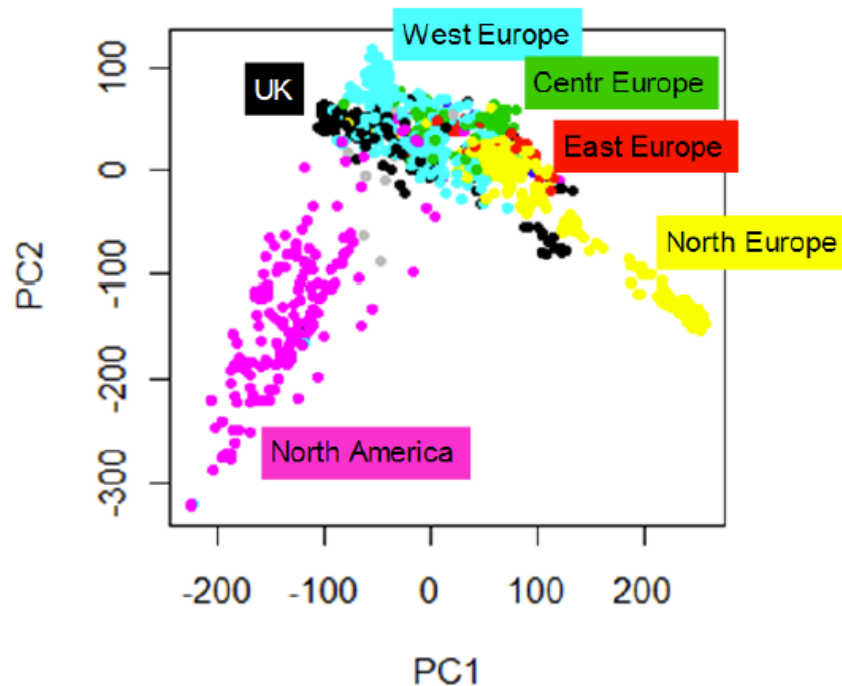


$$N = 1000, SNP = A/T, MAF = 0.35$$

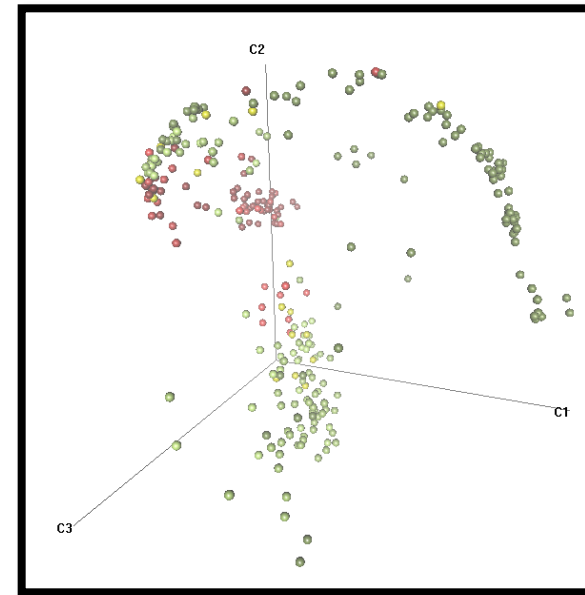


- Isolation by distance: Individuals are not independent
- > Accessions tend to cluster in subpopulations according to geographic origins

Ex: arabidopsis

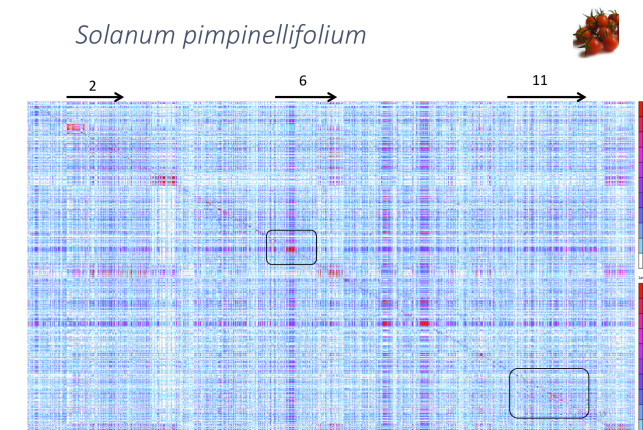
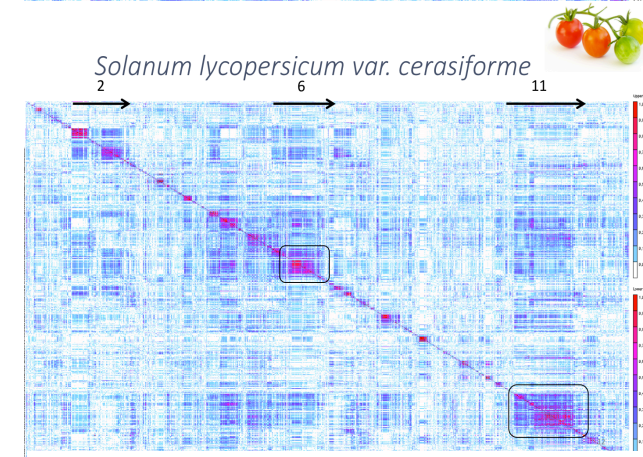
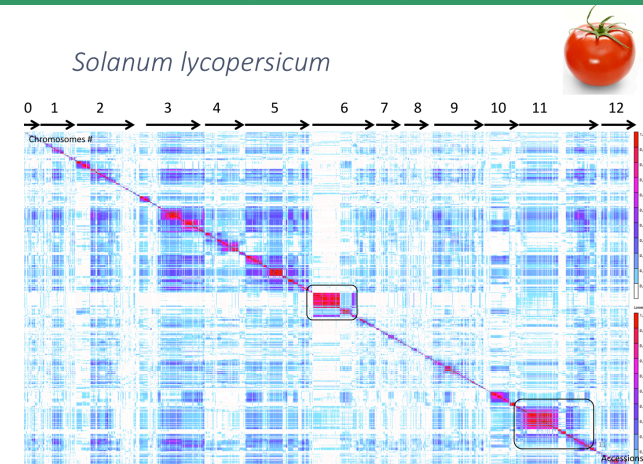


Ex: tomato

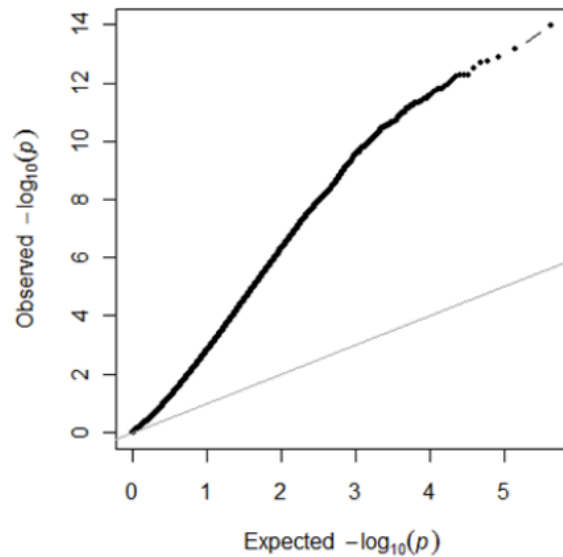


⇒ Allelic frequencies varies between subpopulations and generate long range disequilibrium

⇒ Cause/Consequence of population structure:
Linkage disequilibrium varies according to genetic groups

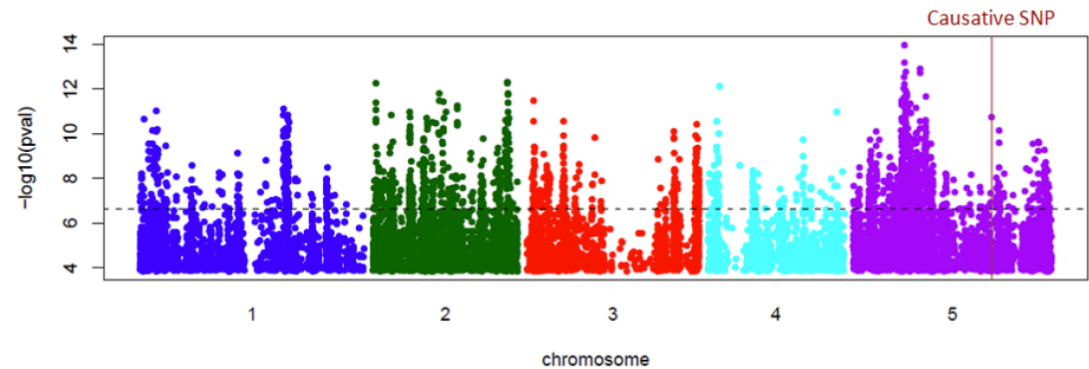


QQ-plot



⇒ Statistical test is inflated

Manhattan -plot



⇒ False positives are generated

Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

$$\mathbf{u} \sim N(0, \sigma_g^2 \mathbf{K}) \text{ \& } \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}),$$

$\mathbf{K}$ 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

The screenshot displays three overlapping web browser windows. The top window shows the Buckler Lab for Maize Genetics and Diversity website. The middle window shows the Bioconductor website with the GWASTools package highlighted. The bottom window shows the PLINK 1.90 beta download page.

Buckler Lab for Maize Genetics and Diversity

Home Bioinformatics Publications People Research Contact Us

Tassel Version (Build: April 2)

Tassel 5 Mac OS
Tassel 5 Window
Tassel 5 Window
Tassel 5 Window
Tassel 5 UNIX

Tassel Version

www.maizegenetics.net

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

GWASTools

platforms all downloads 1.00
build ok commits 1.00

Tools for Genome Wide Association Studies

Bioconductor version: Release (4.0.0)

Classes for storing very large G data sets.

Author: Stephanie M. Gogarten, Caitlin McHugh, Ian Painter, Xitong Wang

Maintainer: Stephanie M. Gogarten, u.washington.edu

Citation (from within R, enter `citation("Bioconductor::GWASTools")`)

PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab at Mount Sinai School of Medicine, and others. (What's new?) (Credits.) (Methods paper.)

Binary downloads

Operating system ¹	Build		
	Stable (beta 3.36, 31 Mar)	Development (16 Apr)	Old ² (v1.07)
Linux 64-bit	download	download	download
Linux 32-bit	download	download	download
OS X (64-bit)	download	download	download
Windows 64-bit	download	download	download
Windows 32-bit	download	download	download

¹: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
²: These are just mirrors of the binaries posted at <http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml>.

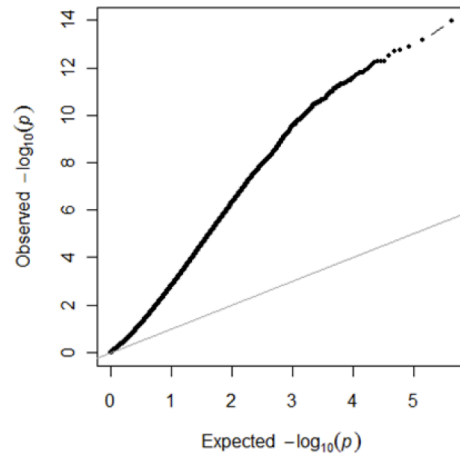
Source code, compilation instructions, and the like are on the [developer page](#).

The following documented PLINK 1.07 flags are not supported by 1.90 beta 3:

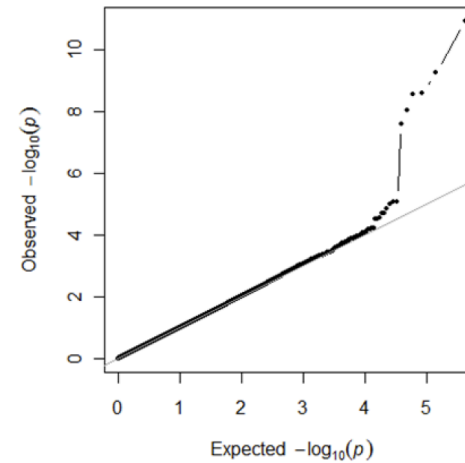
- `--qual-geno-scores3`
- `--segment4`
- `--dfam`
- `--p2, --genedrop`
- `--hap, --hap-window, --hap-snp5`
- `--proxy-assoc, --proxy-impute5`
- `--cnv-list, --cfile, --gfile`
- `--R`
- `--id-dict, --id-match6`
- `--compress, --decompress7`

QQ-plot

Linear Regression

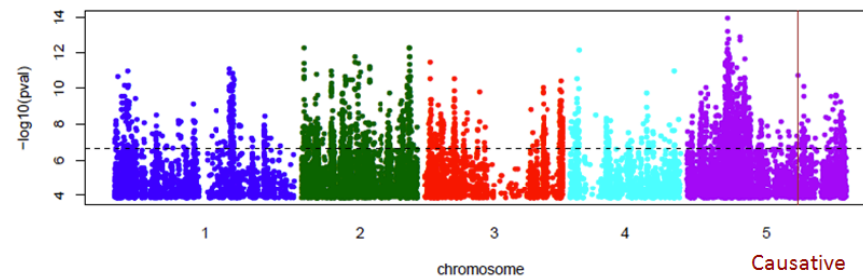


EMMAX

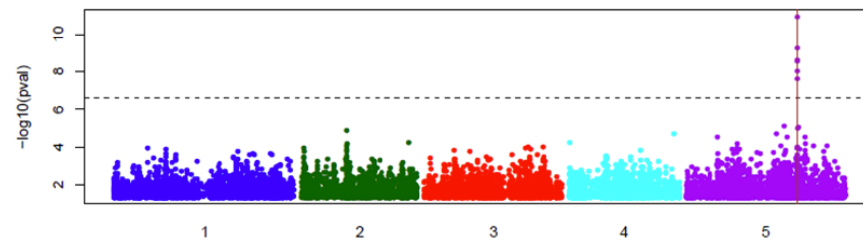


Manhattan -plot

Linear Regression



EMMAX



RR-BLUP package

Our test dataset:

Oat dataset on yield

More infos:

<https://www.rdocumentation.org/packages/rrBLUP/versions/1.1/topics/RR.BLUP>

[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

#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?

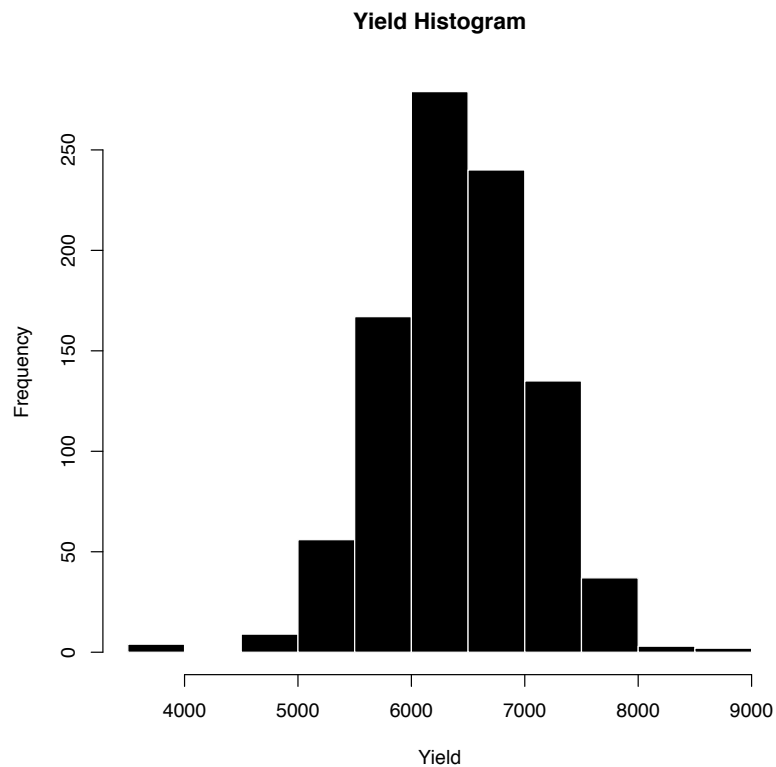
#Step 8: QQ-Manhattan and Correlation Plots

-> Where are the main peaks?

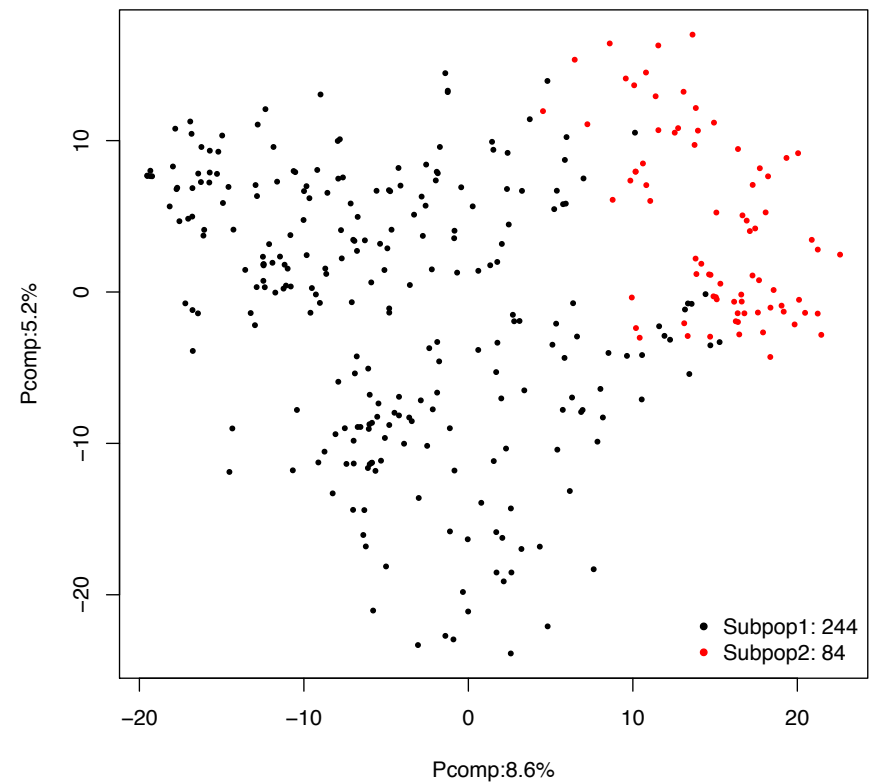
#Step 9: Which are the hits?

-> Which test is more conservative?

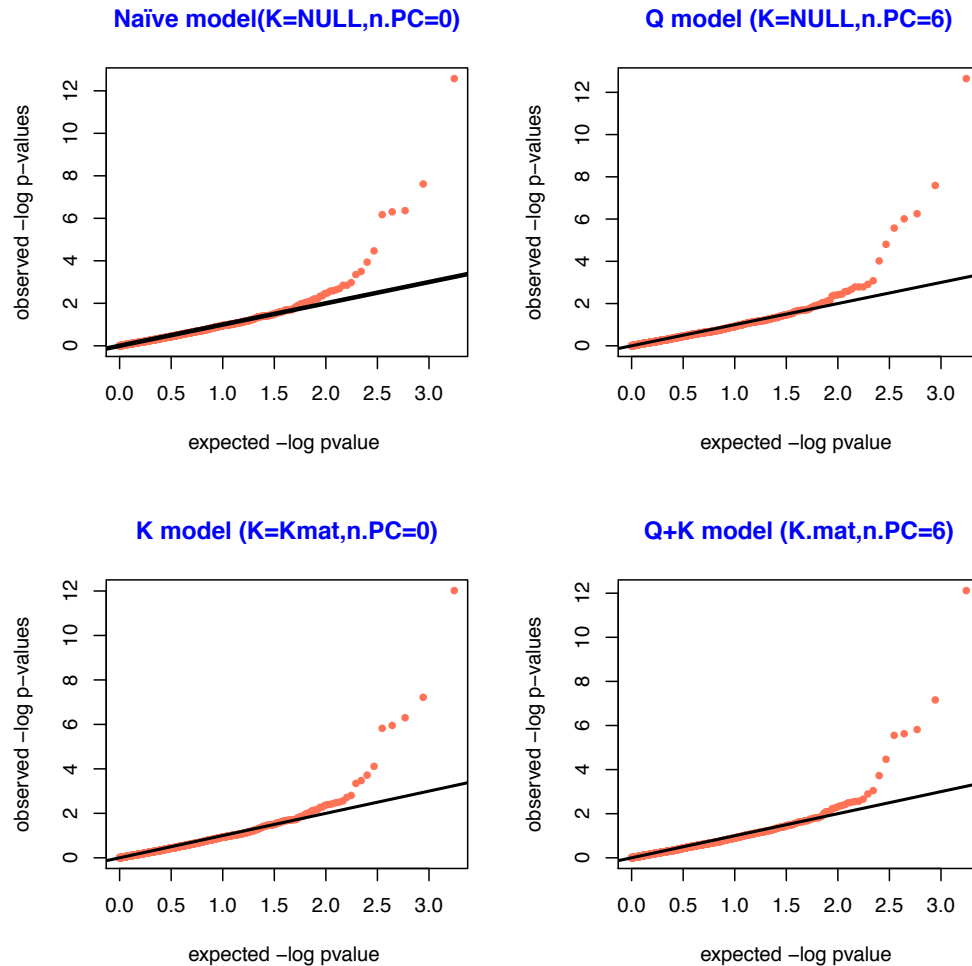
1) Check data distribution and residual homogeneity



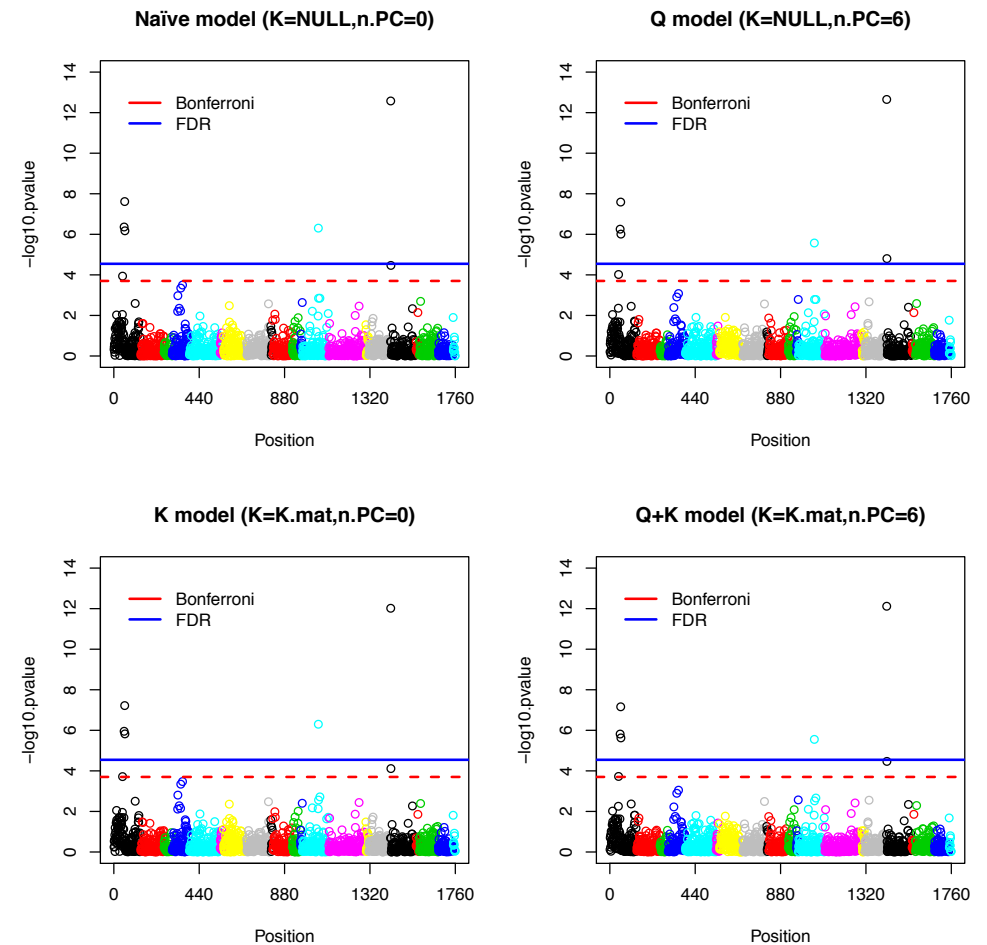
2) Check population structure



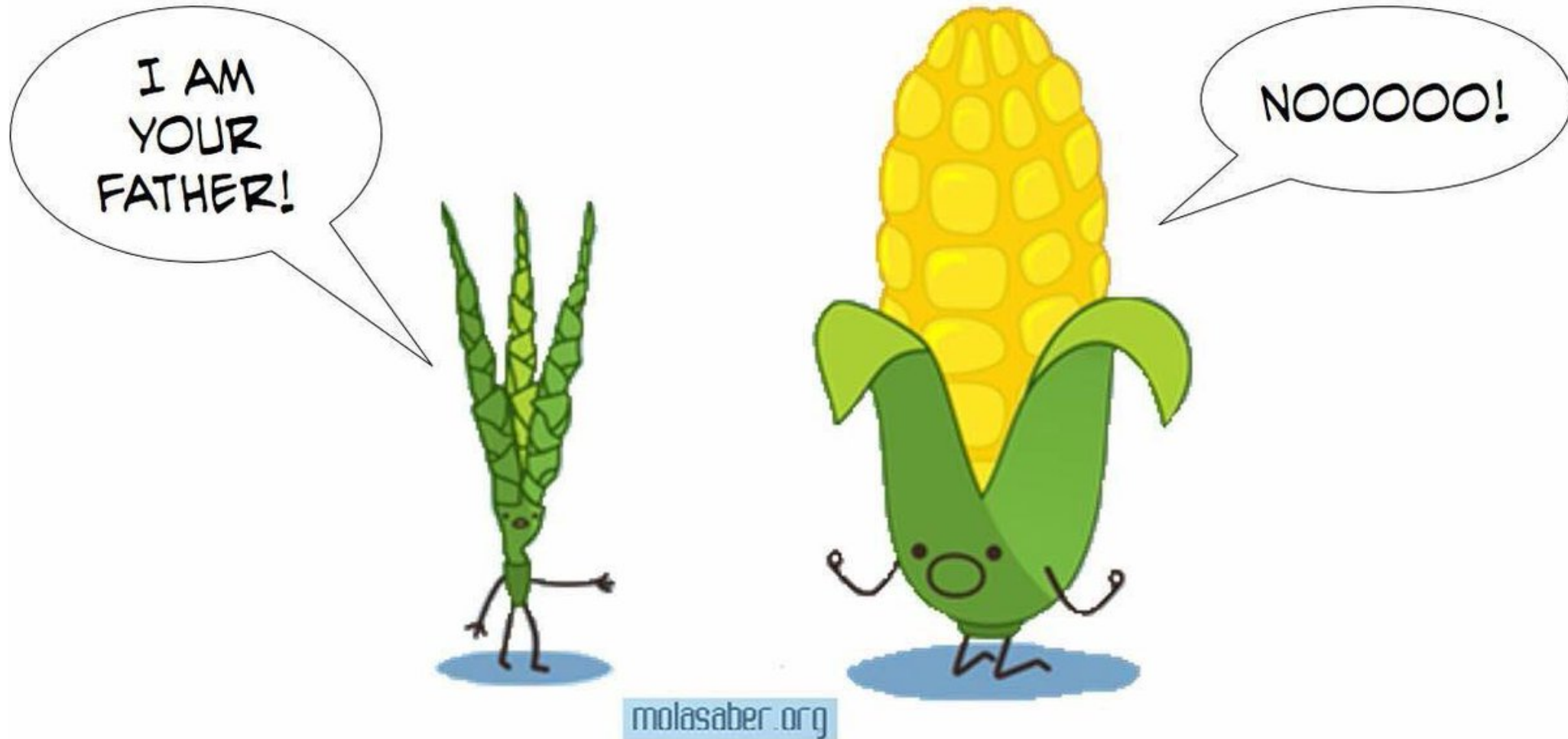
3) Fit models (QQ plot)



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



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

Dataset:

Sauvage et al, Plant Phys. 2014

- > 160 tomato accessions
- > 6000 SNP
- > 20 metabolic traits

Model:

Segura et al, Nat. Gen., 2012

Multi-locus mixed-model (MLMM)

R

Plant Physiology

[HOME](#) | [ABOUT](#) | [SUBMIT](#) | [SUBSCRIPTIONS](#) | [ADVERTISE](#) | [ARCHIVE](#) | [CONTACT US](#)



« Previous

Table of Contents: GENES, DEVELOPMENT, AND...

Next Article »

EXPAND ➔

© 2014 American Society of Plant Biologists. All Rights Reserved.

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

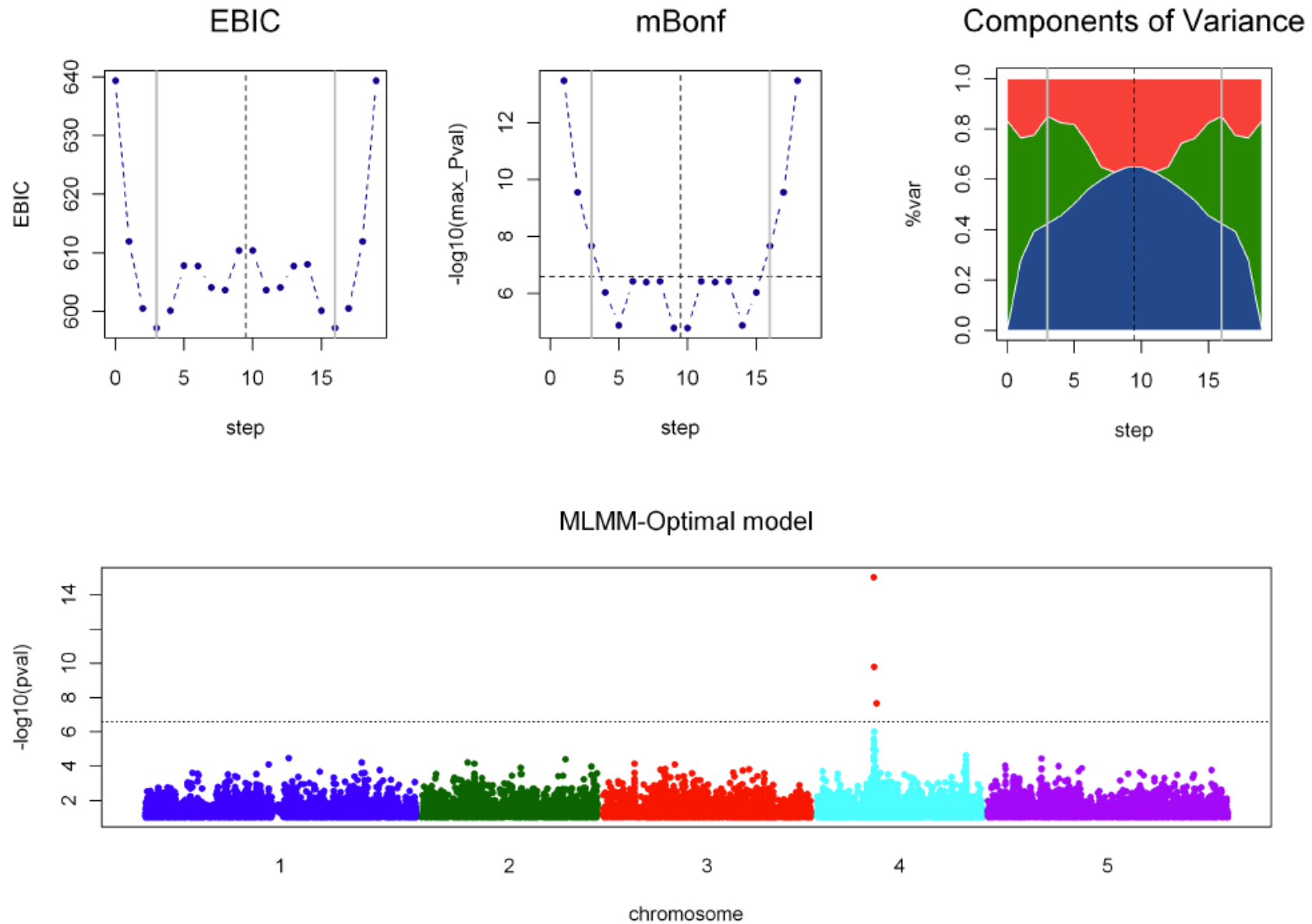
Testing all combinations is not possible...

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

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

Multi locus Model: multiple testing

- Forward stopping criterion: $\hat{h}^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_\epsilon^2)$ close to 0
- Best Model Identification:
 - BIC (Schwarz, 1978)
 - Extended BIC: $EBIC = BIC + 2\log C_m^k$, where C_m^k 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



#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

<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