GWAS

Arthur Korte, Fri 4th March

About me (Disclaimer)

PhD in Molecular Biology (TU Munich)

Postdoc in Population Genetics
(Gregor Mendel Institute of Molecular Plant Biology, Vienna)

Junior professor
Center for Computational Theoretical Biology, University Würzburg

CTB Center for Computational and Theoretical Biology

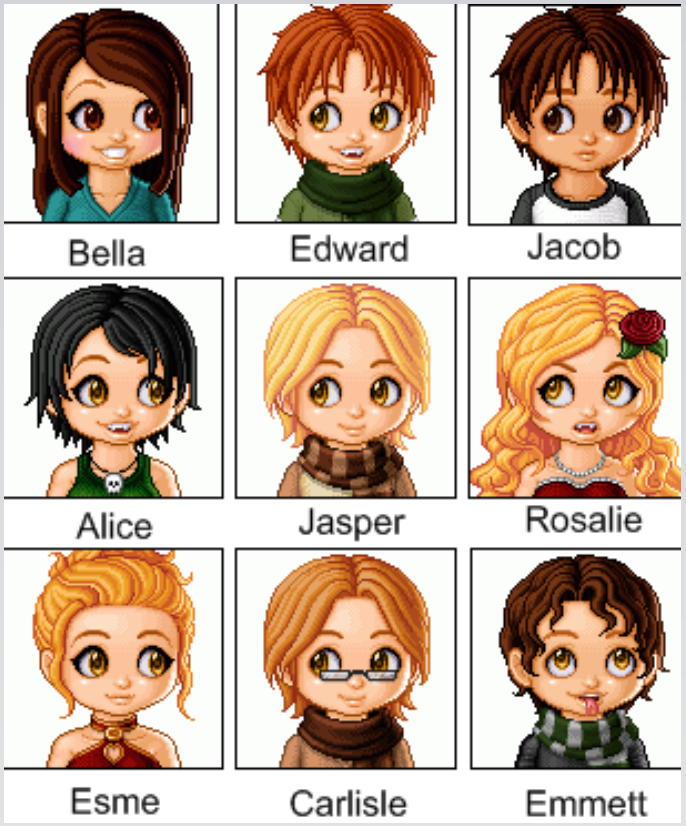
Content

1. Lecture I : Introduction to GWAS and population structure
2. Practical I: GWAS with custom R scripts

3. Lecture II: Different GWAS models / software
4. GWAS with webtools
5. Practical II

Introduction to GWAS and population structure

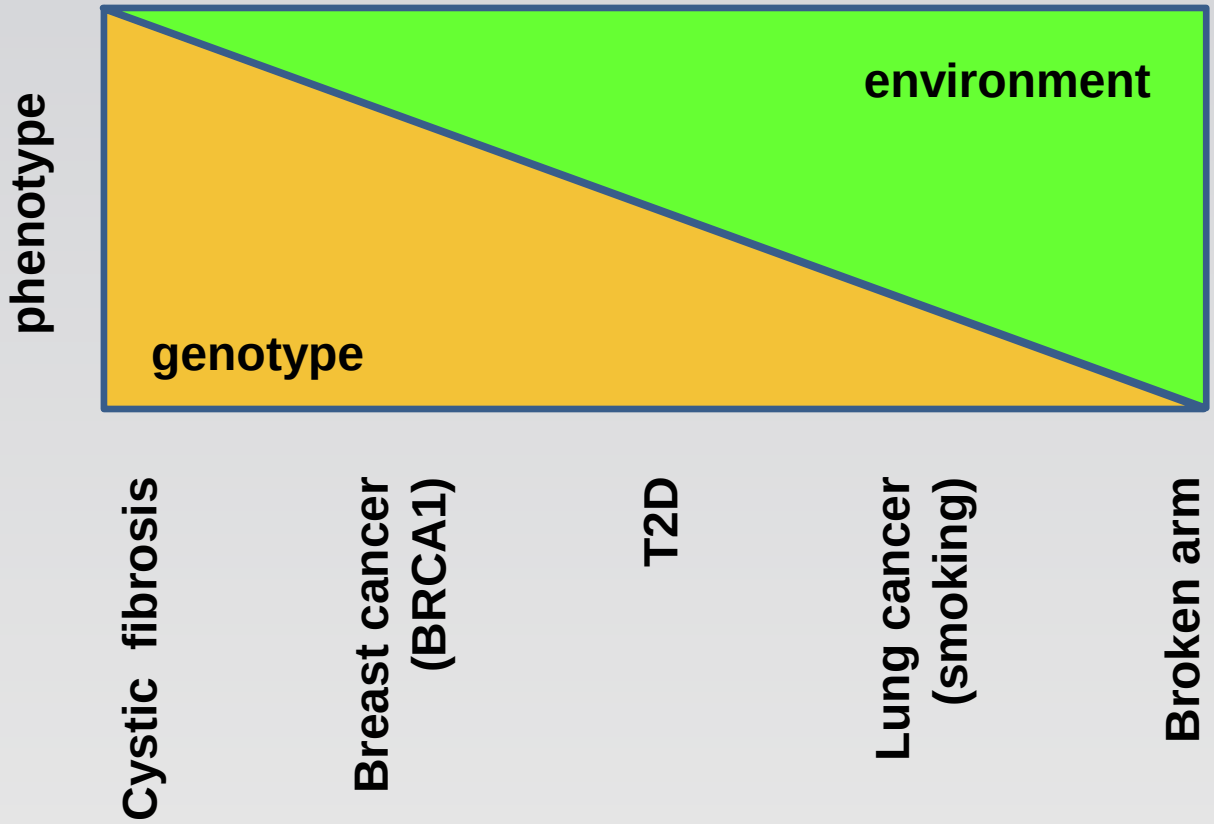
GWAS

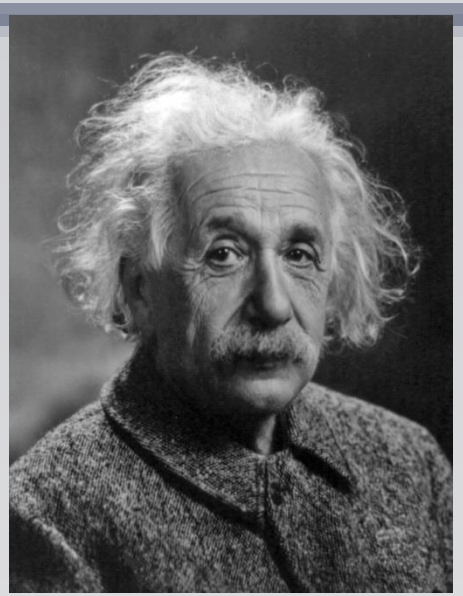


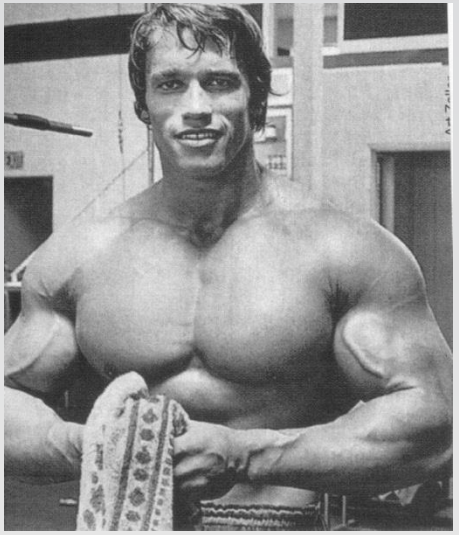
Associations between phenotype and genotype

Y = G + E + GxE

Center for Computational and Theoretical Biology



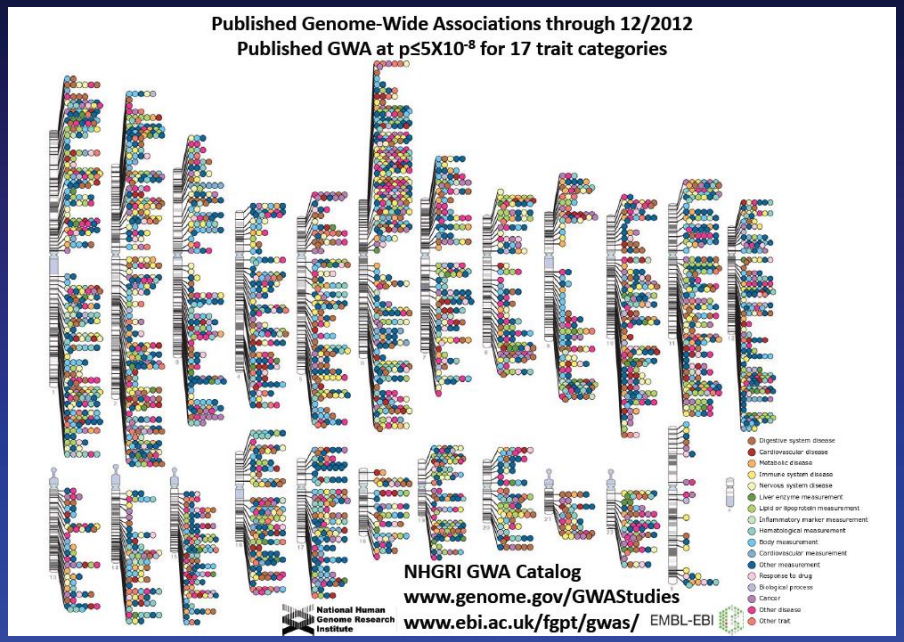




A Catalog of Published Genome-Wide Association Studies

Center for Computational and Theoretical Biology

Published GWA Reports, 2005 – 6/2012



1350

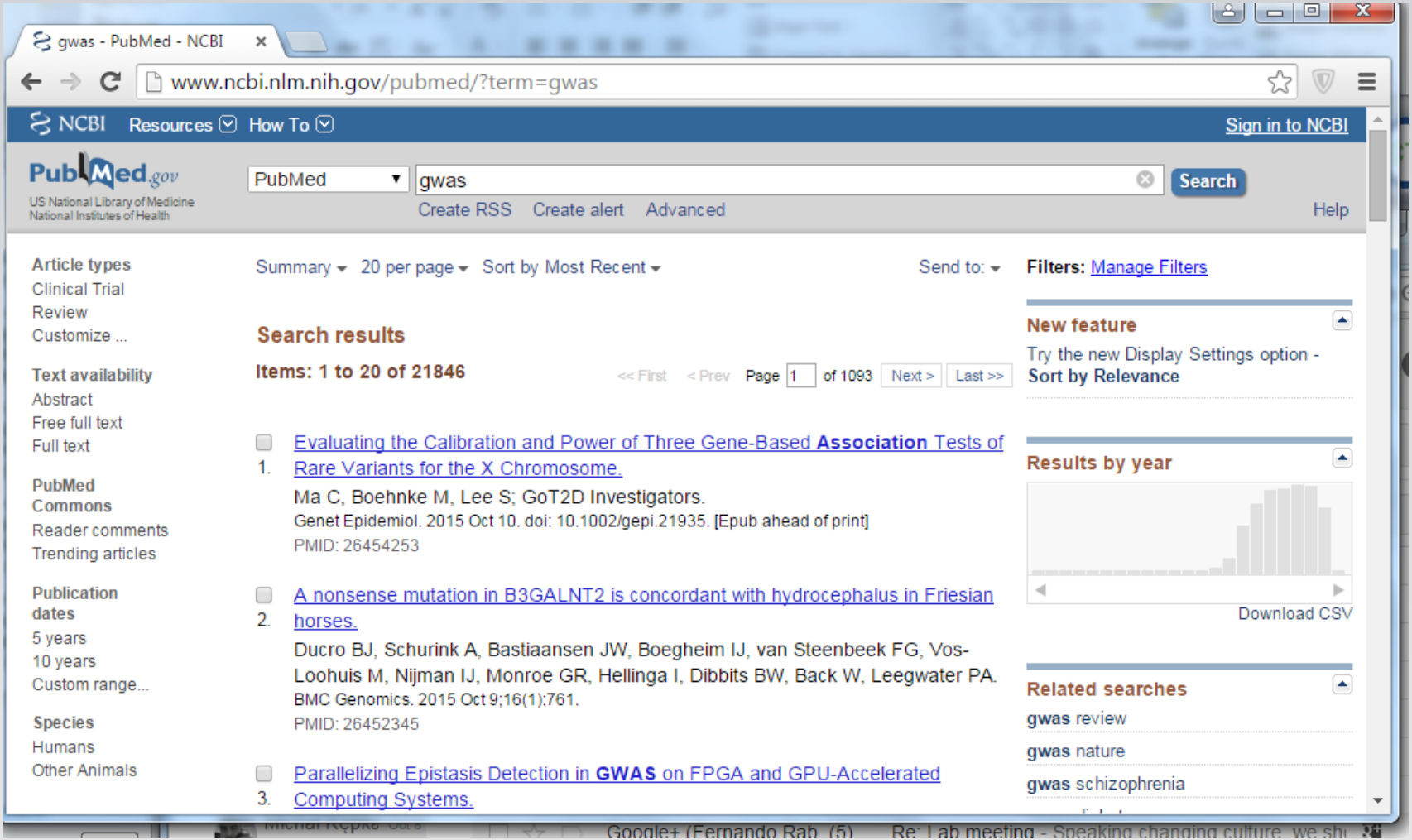
Total Number of Publications

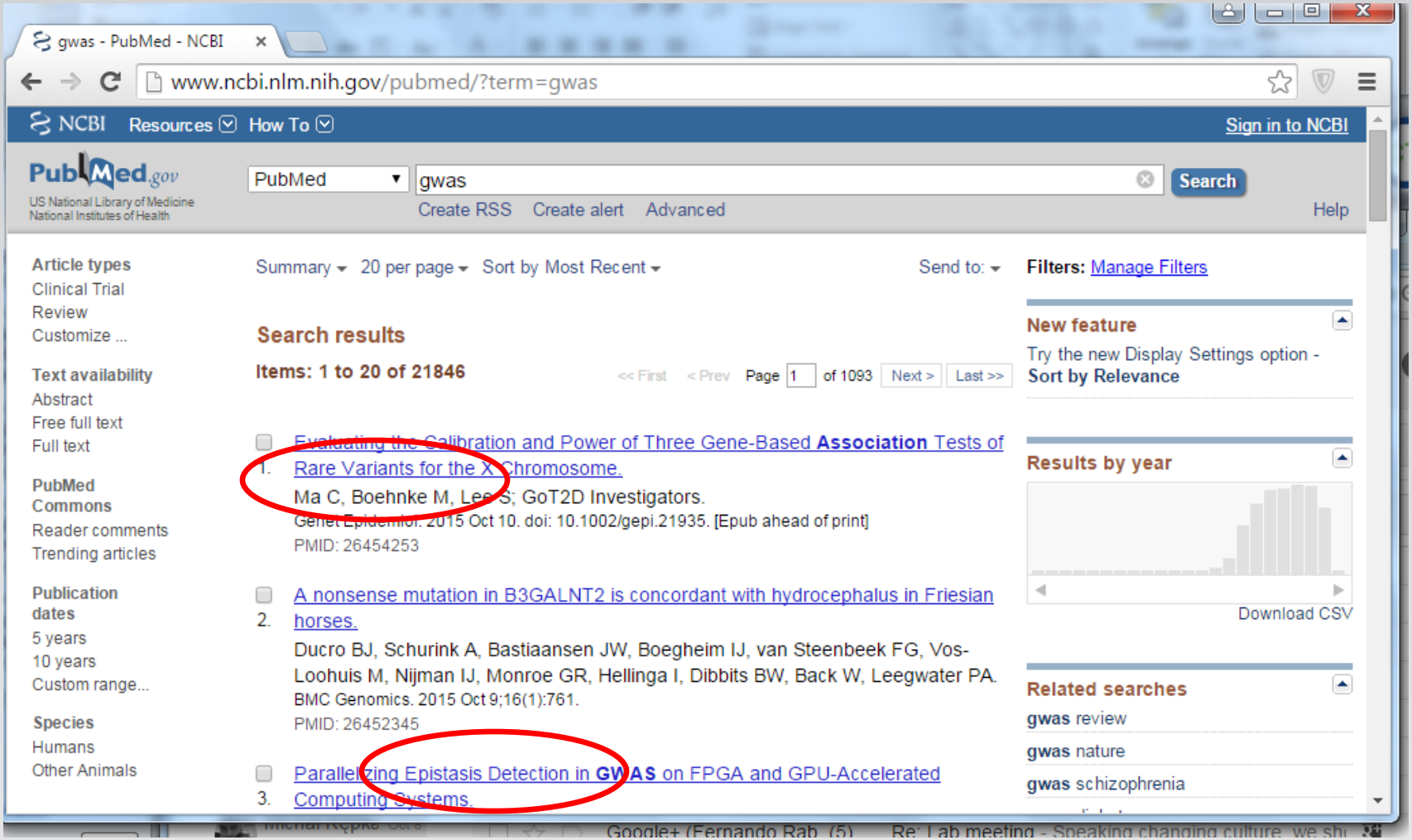
1400
1200
1000
800
600
400
200
0

2005 2006 2007 2008 2009 2010 2011 2012

Calendar Quarter

http://www.genome.gov/gwastudies/


gwas - PubMed - NCBI
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
Summary 20 per page Sort by Most Recent
Send to:
Filters: Manage Filters
Search results
Items: 1 to 20 of 21846
<< First < Prev Page 1 of 1093 Next > Last >>
1. Evaluating the Calibration and Power of Three Gene-Based Association Tests of Rare Variants for the X Chromosome.
Ma C, Boehnke M, Lee S; GoT2D Investigators.
Genet Epidemiol. 2015 Oct 10. doi: 10.1002/gepi.21935. [Epub ahead of print]
PMID: 26454253
2. A nonsense mutation in B3GALNT2 is concordant with hydrocephalus in Friesian horses.
Ducro BJ, Schurink A, Bastiaansen JW, Boegheim IJ, van Steenbeek FG, Vos-Loohuis M, Nijman IJ, Monroe GR, Hellinga I, Dibbits BW, Back W, Leegwater PA.
BMC Genomics. 2015 Oct 9;16(1):761.
PMID: 26452345
3. Parallelizing Epistasis Detection in GWAS on FPGA and GPU-Accelerated Computing Systems.
New feature
Try the new Display Settings option - Sort by Relevance
Results by year
Download CSV
Related searches
gwas review
gwas nature
gwas schizophrenia

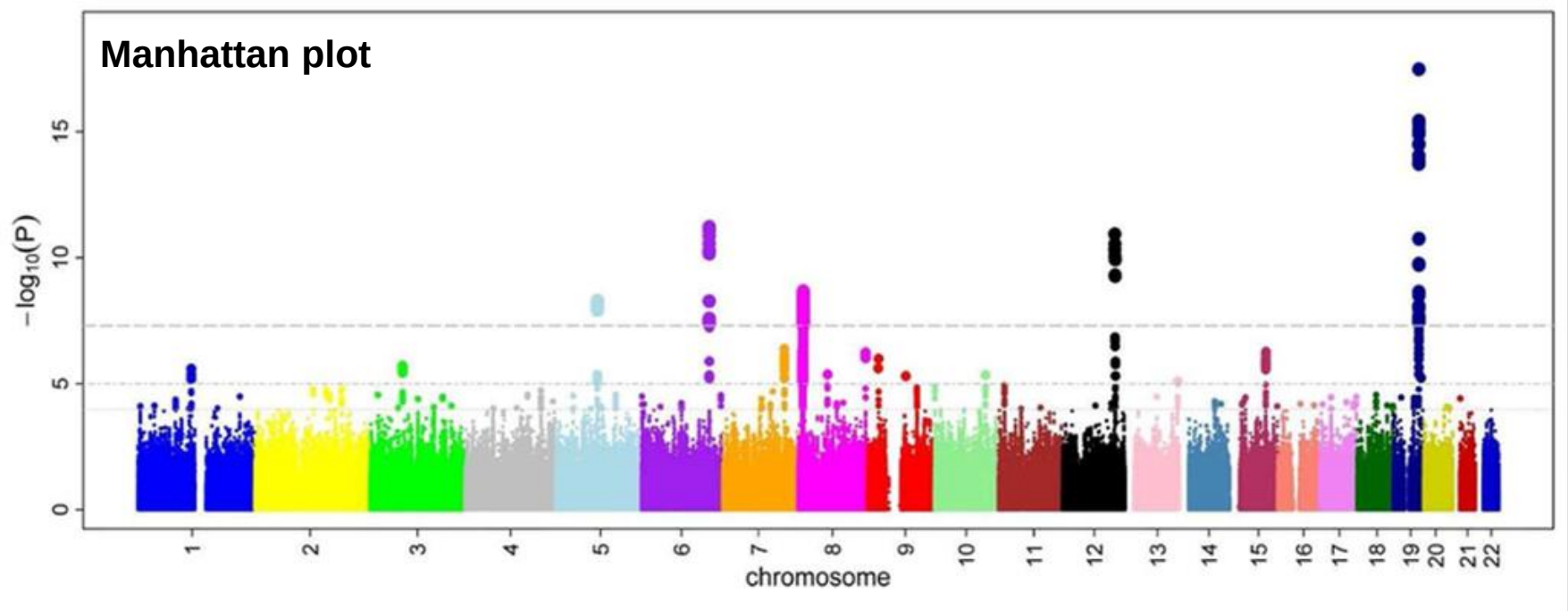

gwas - PubMed - NCBI
www.ncbi.nlm.nih.gov/pubmed/?term=gwas
NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
gwas
Search
Create RSS Create alert Advanced
Help
Article types
Clinical Trial
Review
Customize ...
Text availability
Abstract
Free full text
Full text
PubMed Commons
Reader comments
Trending articles
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Summary 20 per page Sort by Most Recent
Send to:
Filters: Manage Filters
Search results
Items: 1 to 20 of 21846
<< First < Prev Page 1 of 1093 Next > Last >>
1. Evaluating the Calibration and Power of Three Gene-Based Association Tests of Rare Variants for the X Chromosome.
Ma C, Boehnke M, Lee S; GoT2D Investigators.
Genet Epidemiol. 2015 Oct 10. doi: 10.1002/gepi.21935. [Epub ahead of print]
PMID: 26454253
2. A nonsense mutation in B3GALNT2 is concordant with hydrocephalus in Friesian horses.
Ducro BJ, Schurink A, Bastiaansen JW, Boegheim IJ, van Steenbeek FG, Vos-Loohuis M, Nijman IJ, Monroe GR, Hellinga I, Dibbits BW, Back W, Leegwater PA.
BMC Genomics. 2015 Oct 9;16(1):761.
PMID: 26452345
3. Parallelizing Epistasis Detection in GWAS on FPGA and GPU-Accelerated Computing Systems.
New feature
Try the new Display Settings option - Sort by Relevance
Results by year
Download CSV
Related searches
gwas review
gwas nature
gwas schizophrenia

Why GWAS ?

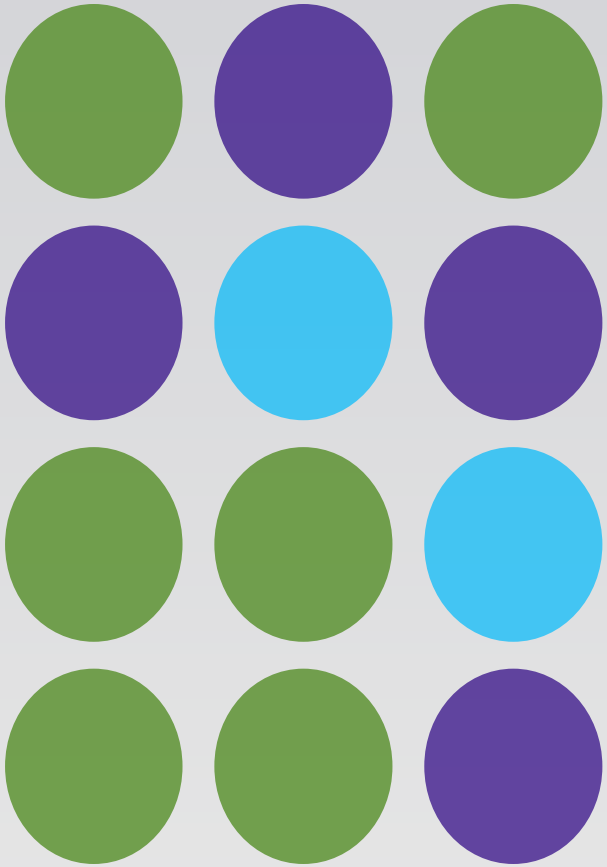
How do the genetic differences in a population translate to distinct phenotypic differences ?

Just do it !

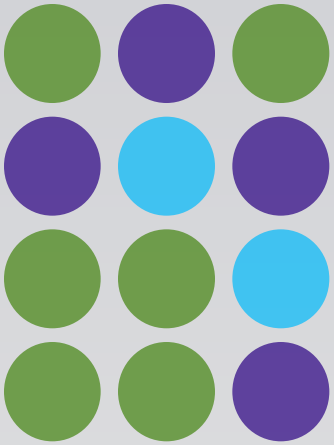
The outcome


Manhattan plot
$-\log_{10}(P)$
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
chromosome

Introduction: GWAS



Introduction: GWAS



•••AGCCTG----TGCACTAAGACT•••

•••AGCCTG----TGCACTAAGACT•••

•••AGCCTG----TGCACTAAGAGT•••

•••AGCCTG----TGCACTAAGACT•••

•••AGCCTGAGTGTGCACTAAGAGT•••

•••AGCCTGAGTGTGCACTAAGAGT•••

•••AGCCTGAGTGTGTACTAAGACT•••

•••AGCCTGAGTGTGTACTAAGAGT•••

•••AGCCTGAGTGTGTACTAAGACT•••

•••AGCCTGAGTGTGTACTAAGAGT•••

•••AGCCTGAGTGTGTACTAAGACT•••

•••AGCCTGAGTGTGTACTAAGACT•••

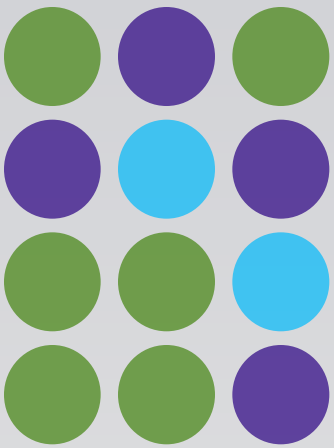
SNPs

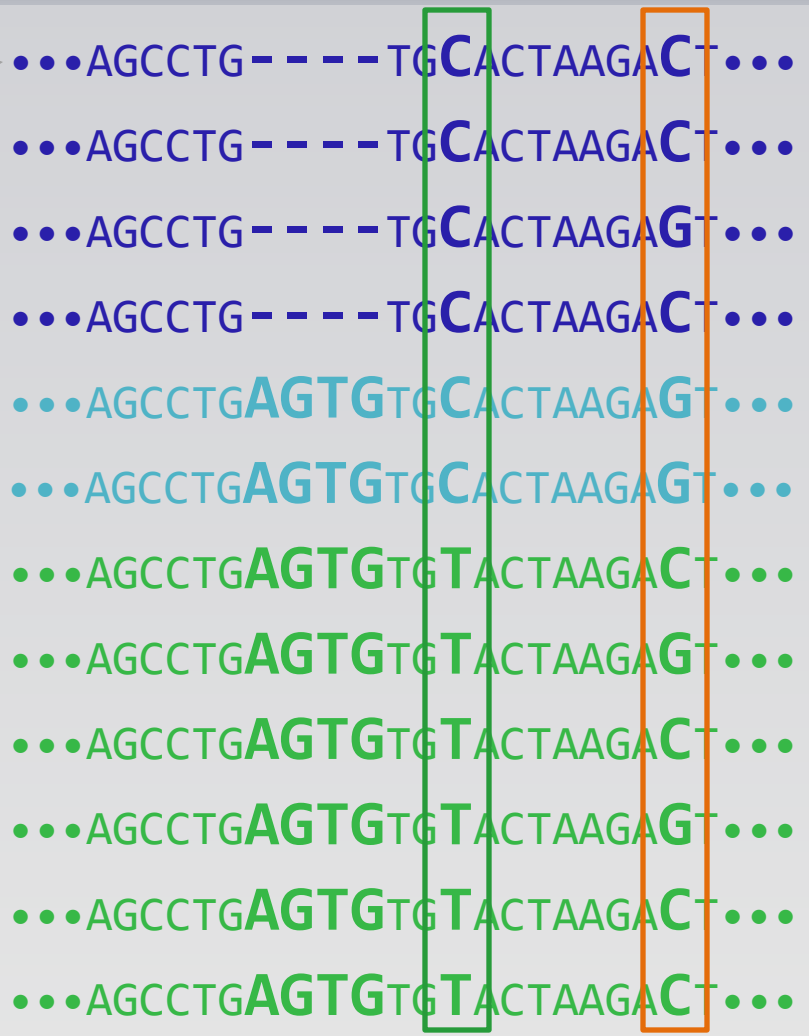
Indels

CNVs

Epigenetic markers

Introduction: GWAS





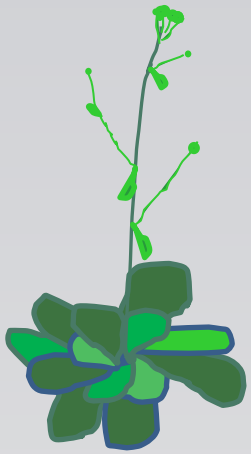
SNPs

Indels

CNVs

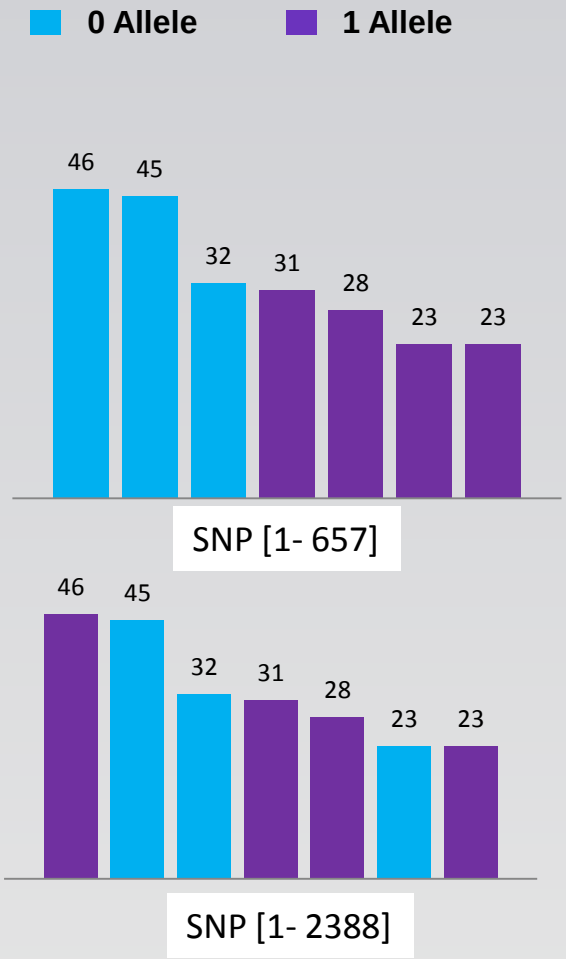
Epigenetic markers

Introduction: GWAS (linear model)



ecotype	phenotype
6909	23
6911	32
7344	28
7436	23
8233	45
8234	46
8673	31

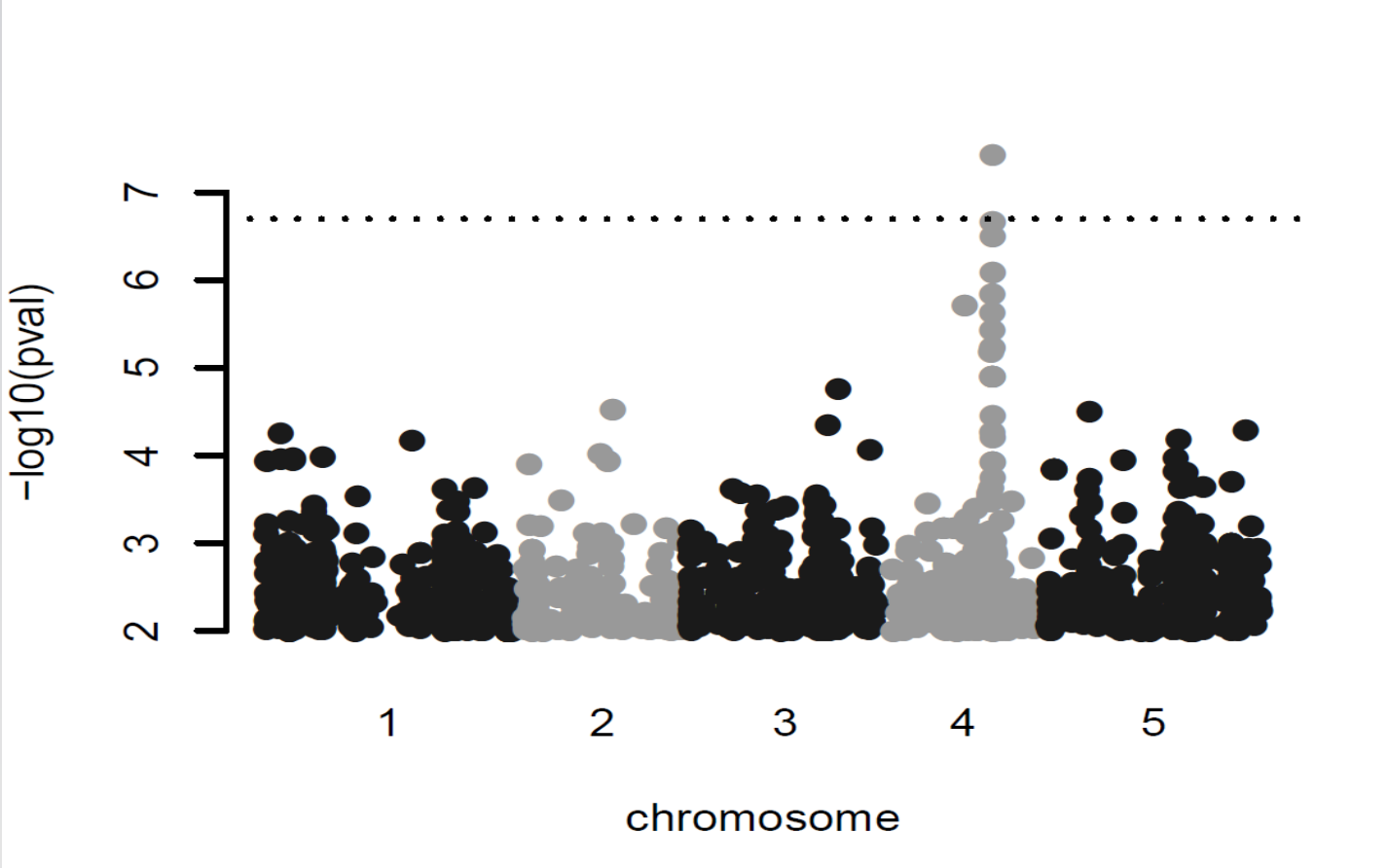
SNP [1- 657]	SNP [1- 2388]
0	0
1	1
0	0
0	1
1	1
1	0
0	0



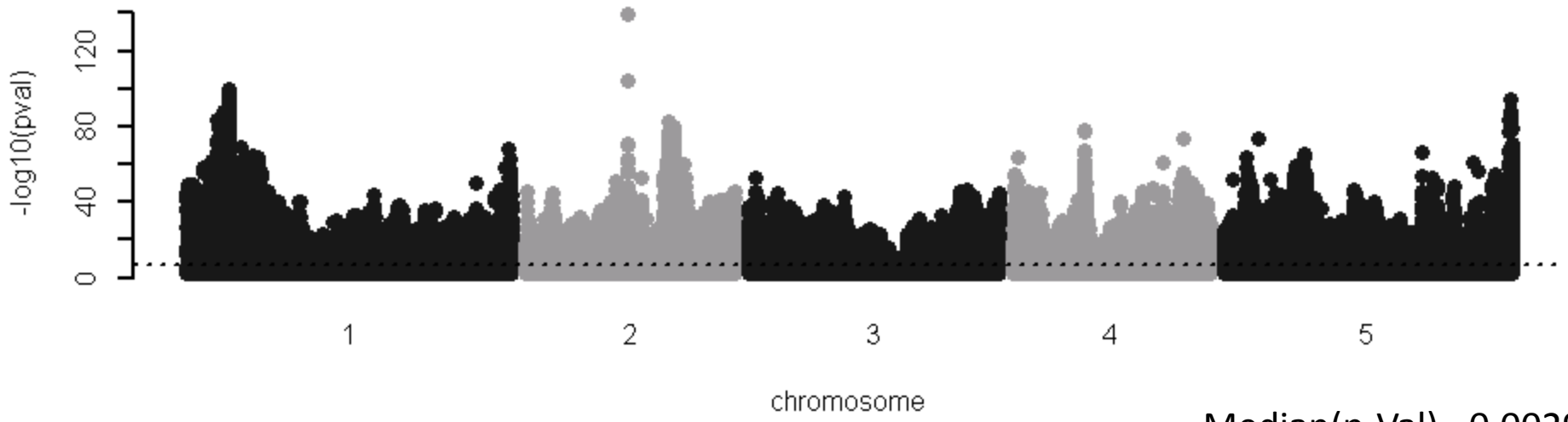
$$H_0 : Y = \beta_0 + \varepsilon$$
$$H_1 : Y = \beta_0 + \beta_1 G + \varepsilon$$

Test for each SNP, if including the SNP in the model can explain the phenotype

Introduction: GWAS (linear model)



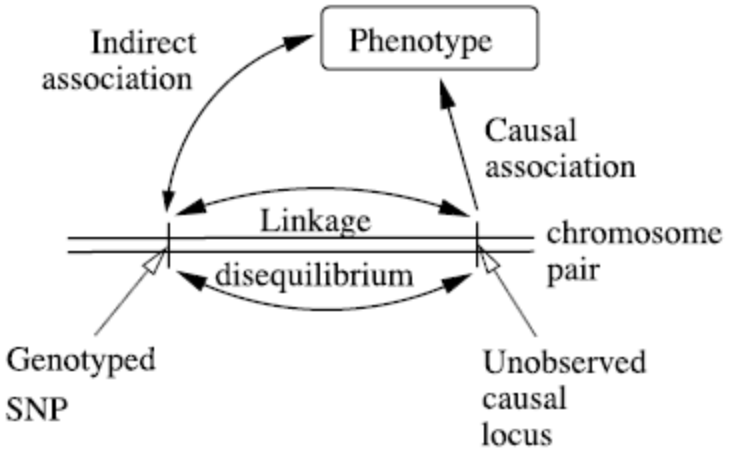
GWAS (linear model) for 925 accessions on flowering time

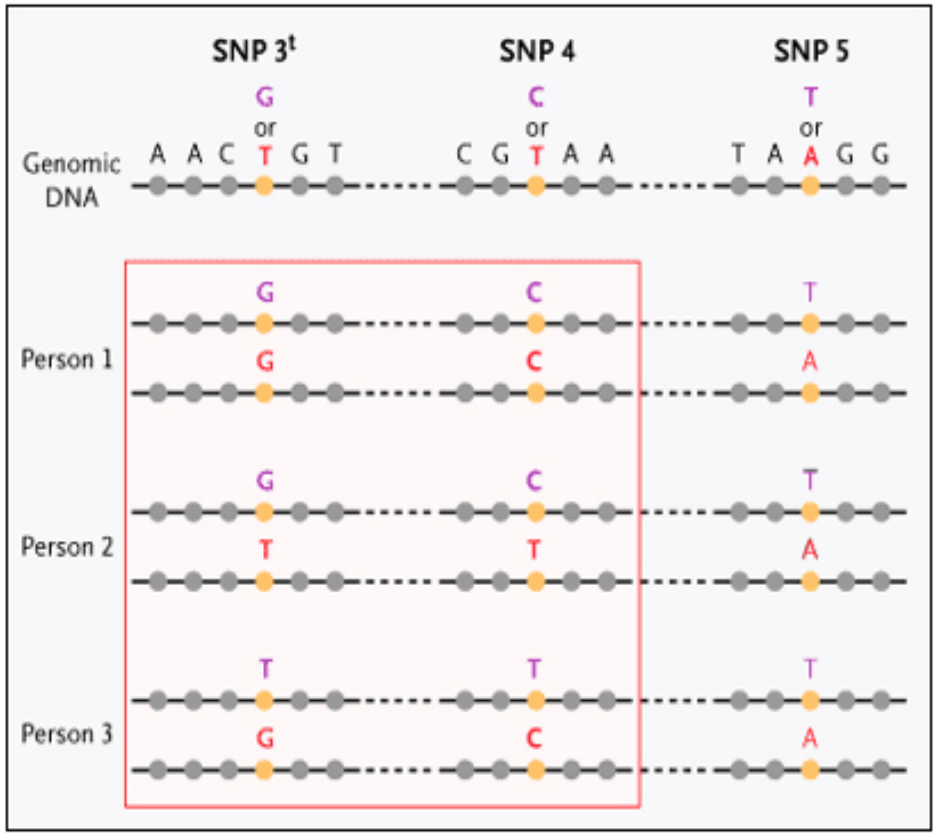


Median(p-Val) =0.00298

~30 % of the genome is causative for flowering time ?!?

Introduction: LD

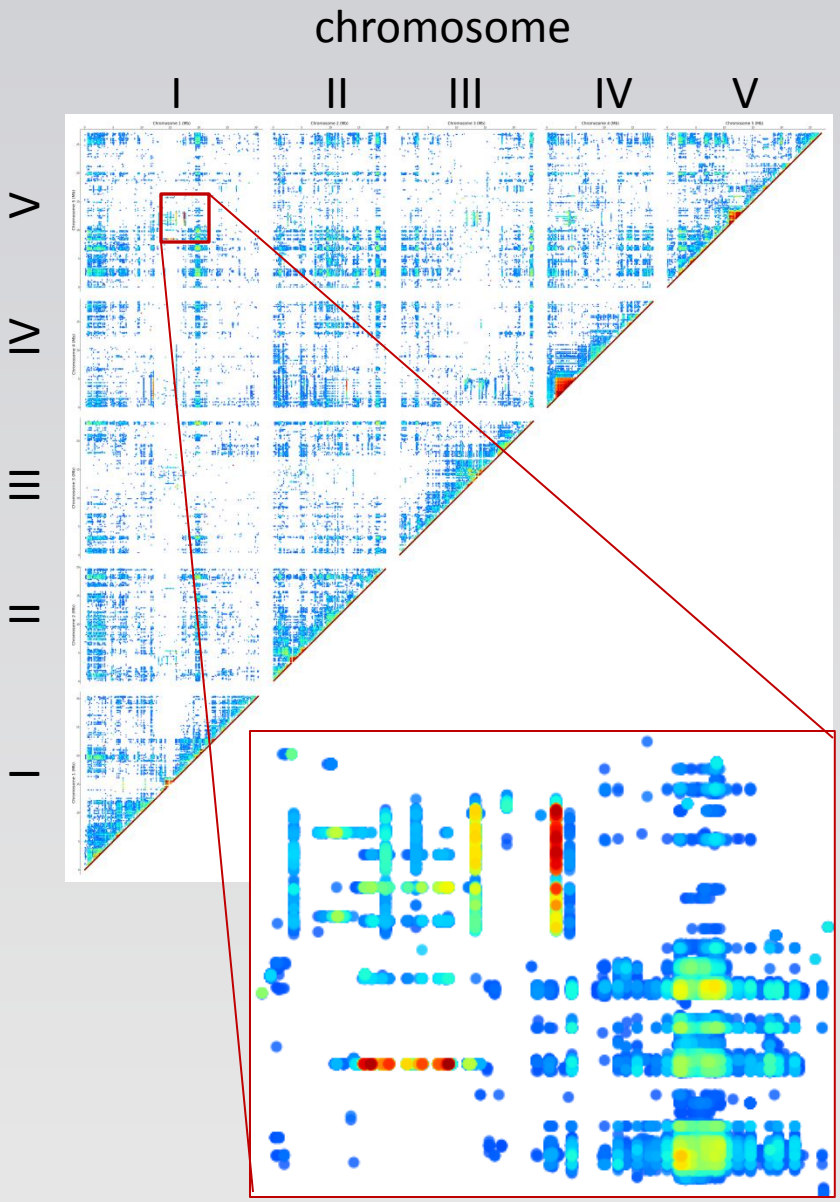




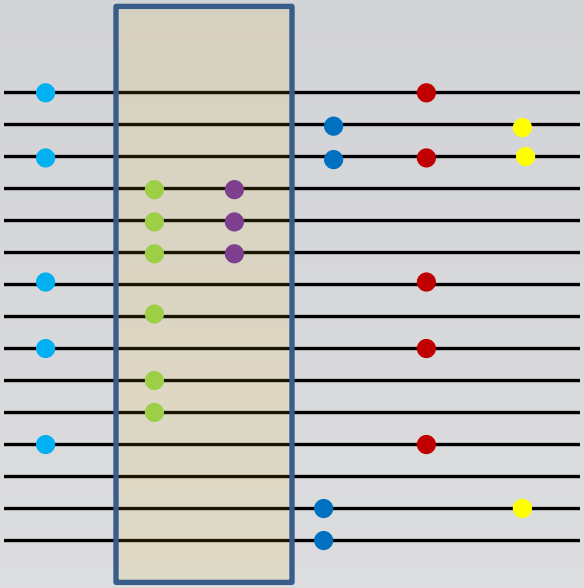
Due to Linkage disequilibrium one SNP can serve as a proxy for others, but also 'rent' power from them

Christensen and Murray *N Engl J Med* (2007)

Pair wise correlation of SNPs across t



Long et all Nat Genet 2013



p_1(●) = 0.4

q_1(●) = 0.2

χ_{11}(●●) = 0.2

$D = \chi_{11} - p_1 q_1$

$D = 0.12$

$D' = D/D_{max}$

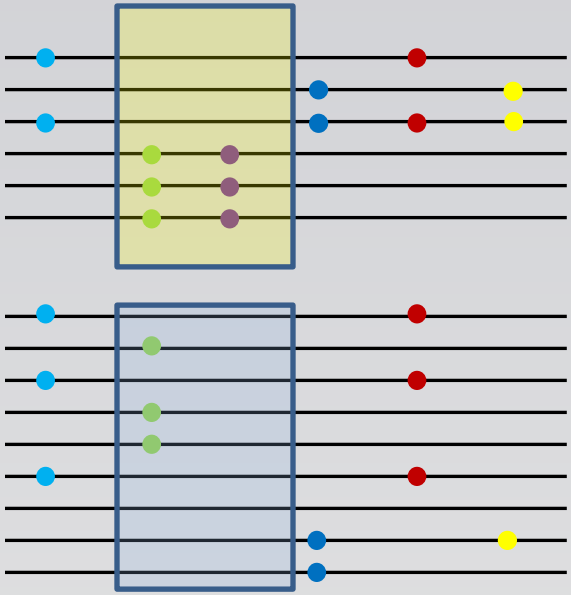
$$D_{max} = \begin{cases} \min(p_1 q_{1,}\ p_2 q_2) \text{ when } D < 0 \\ \min(p_1 q_{2,}\ p_2 q_1) \text{ when } D > 0 \end{cases}$$

$D' = 1$

$r^2 = D^2 / p_1 q_1 p_2 q_2$

$r^2 = 0.375$

Linkage Disequilbrium



Subpopulation 1

Subpopulation 2

Creates LD	Breaks down LD
Selection	Recombination
Admixture	gene conversion
Drift	

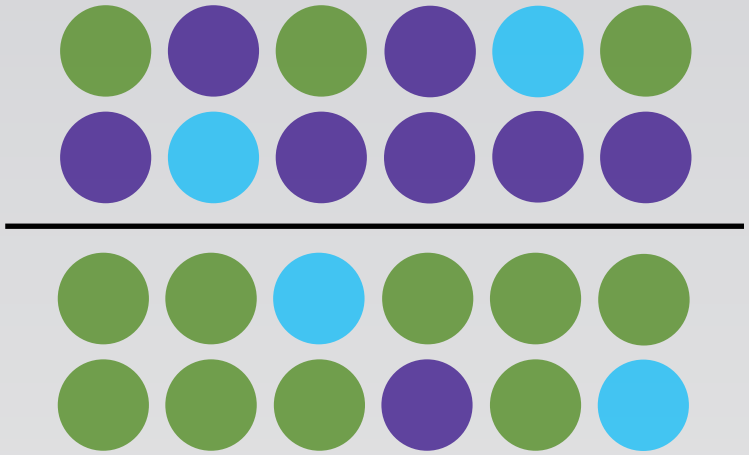
LD between two SNPs can differ in different Subpopulations

Population structure

Confounding due to population structure may arise if it correlates with the trait in question

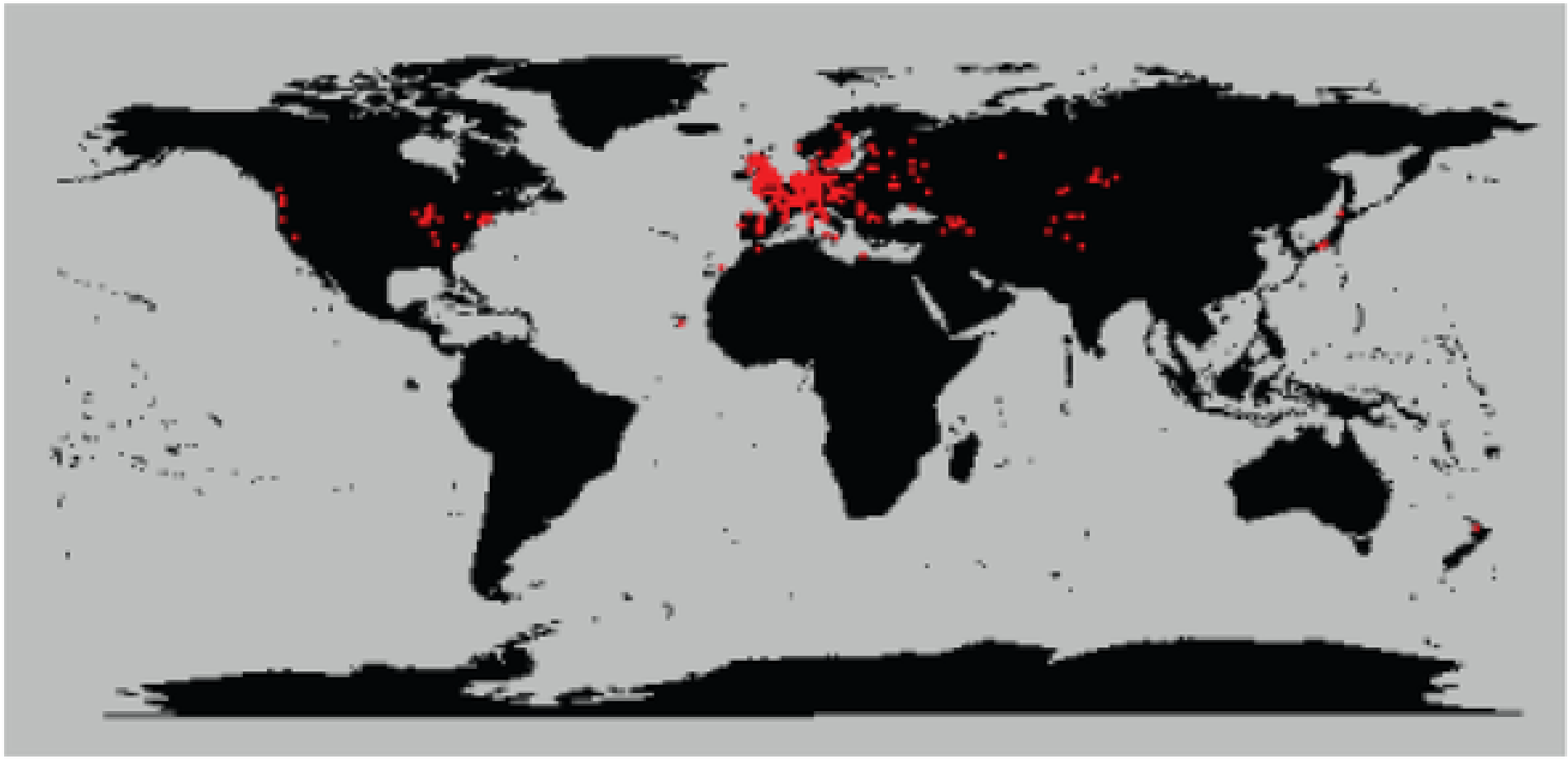
Sub-population 1

Sub-population 2



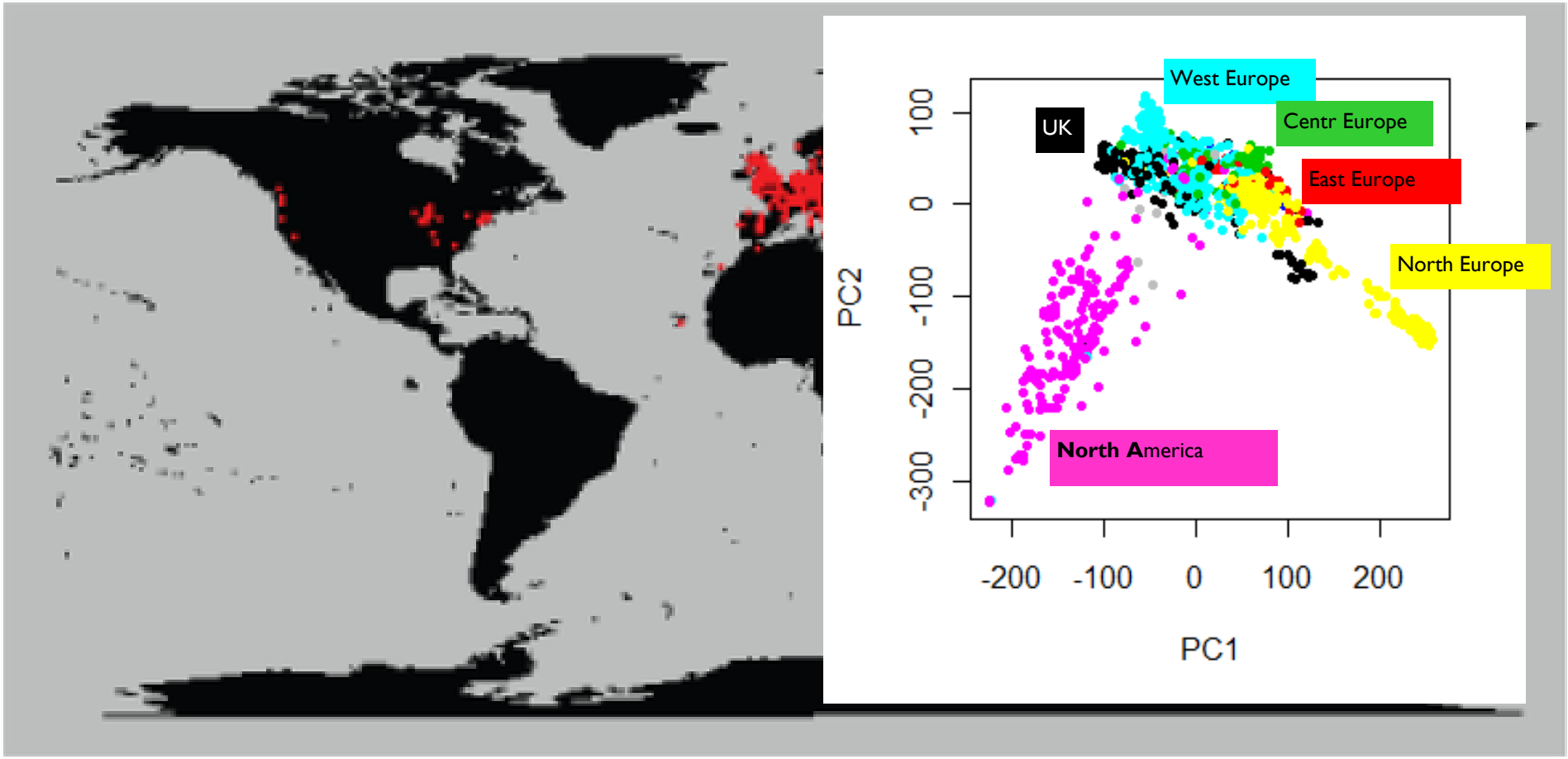
The penetrance of Cystic firosis is ~ 1:2.000 for Europeans, for Asians the penetrance is ~ 1:90.000.

Population structure in *Arabidopsis thaliana*



Platt *et al*. *PloS Genet.*(2010)

Population structure in *Arabidopsis thaliana*

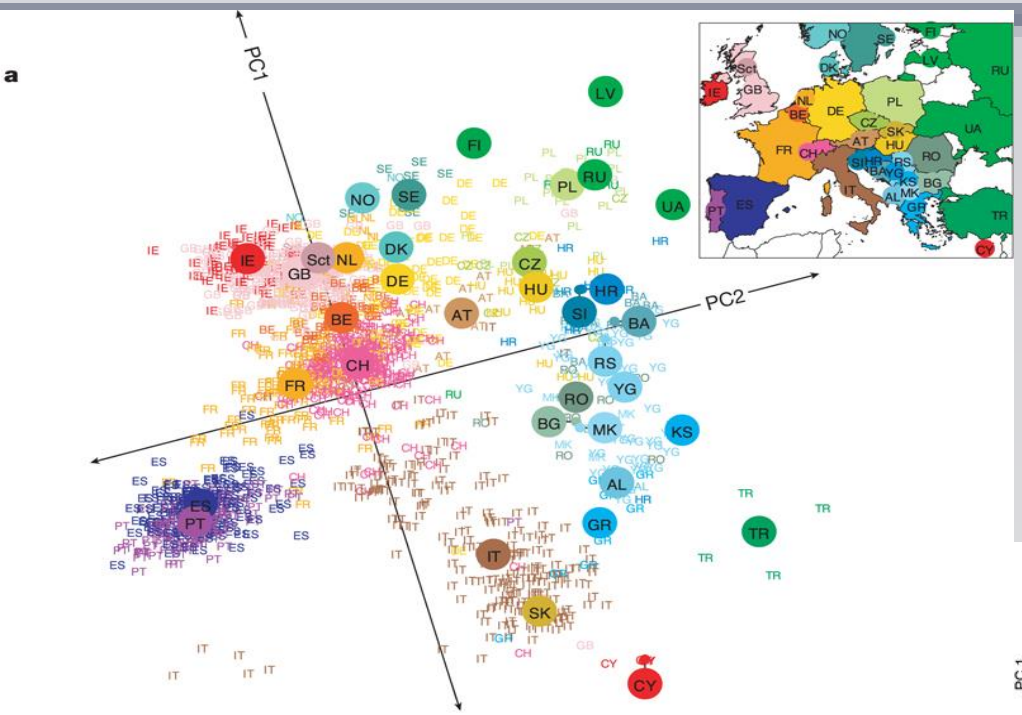


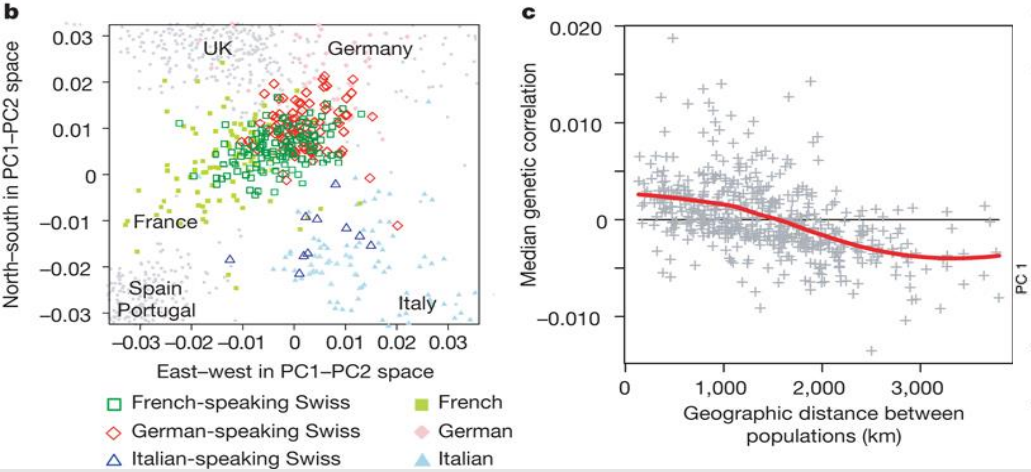
Closely related individuals share not only the causative variants, but have a lot of non-causative SNPs in LD

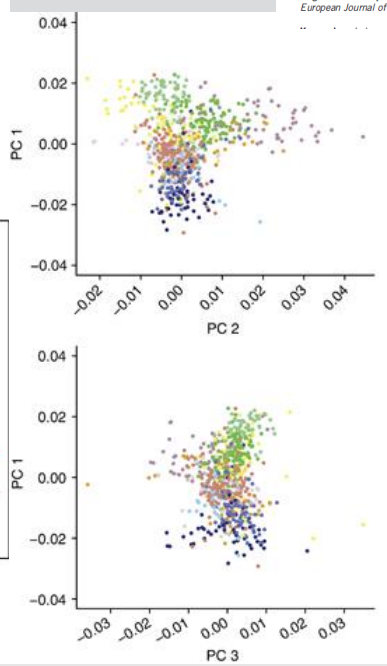
Platt *et al*. *PloS Genet.*(2010)

Structure in human data

Center for Computational and Theoretical Biology







EJHG Open

European Journal of Human Genetics (2013), 1–7
© 2013 Macmillan Publishers Limited. All rights reserved 1018-4813/13
www.nature.com/ejhg

ARTICLE

The Genome of the Netherlands: design, and project goals

Dorret I Boomsma[*,1,22], Cisca Wijmenga[2,22], Eline P Slagboom[3,22], Morris A Swertz[2,22], Lennart C Karssen[4], Abdel Abdellaoui[1], Kai Ye[5], Victor Guryev[5,6], Martijn Vermaat[7,8,9], Freerk van Dijk[7], Laurent C Francioli[10], Jouke Jan Hottenga[1], Jeroen FJ Laros[7,8,9], Qibin Li[11], Yingrui Li[11], Hongzhi Cao[11], Ruoyan Chen[11], Yuanping Du[11], Ning Li[12], Sujie Cao[12], Jessica van Setten[10], Androniki Menelaou[10], Sara L Pulit[10], Jayne Y Hehir-Kwa[13], Marian Beekman[14], Clara C Elbers[10], Henrhiy Byelas[2], Anton JM de Craen[14], Patrick Deelen[2], Martijn Dijkstra[2], Johan T den Dunnen[8,9], Peter de Knijff[8,9], Jeanine Houwing-Duistermaat[17], Vyacheslav Koval[18], Karol Estrada[18], Albert Hofman[4], Alexandros Kanterakis[2], David van Enckevort[2], Hailiang Mai[7], Mathijs Kattenberg[1], Elisabeth M van Leeuwen[4], Pieter BT Neerincx[2], Ben Oostra[19], Fernanodo Rivadeneira[18], Eka HD Suchiman[3], Andre G Uitterlinden[18], Gonneke Willemsen[1], Bruce H Wolffenbuttel[20], Jun Wang[1,13,14,22], Paul IW de Bakker[10,22], Gert-Jan van Ommen[21,22] and Cornelia M van Duijn[*,4,22]

Within the Netherlands a national network of biobanks has been established (Biobanking and Biomolecular Research Infrastructure-Netherlands (BBMRI-NL)) as a national node of the European BBMRI. One of the aims of BBMRI-NL is to enrich biobanks with different types of molecular and phenotype data. Here, we describe the Genome of the Netherlands (GoNL), one of the projects within BBMRI-NL. GoNL is a whole-genome-sequencing project in a representative sample consisting of 250 trio-families from all provinces in the Netherlands, which aims to characterize DNA sequence variation in the Dutch population. The parent–offspring trios include adult individuals ranging in age from 19 to 87 years (mean = 53 years; SD = 16 years) from birth cohorts 1910–1994. Sequencing was done on blood-derived DNA from uncultured cells and accomplished coverage was 14–15x. The family-based design represents a unique resource to assess the frequency of regional variants, accurately reconstruct haplotypes by family-based phasing, characterize short indels and complex structural variants, and establish the rate of *de novo* mutational events. GoNL will also serve as a reference panel for imputation in the available genome-wide association studies in Dutch and other cohorts to refine association signals and uncover population-specific variants. GoNL will create a catalog of human genetic variation in this sample that is uniquely characterized with respect to micro-geographic location and a wide range of phenotypes. The resource will be made available to the research and medical community to guide the interpretation of sequencing projects. The present paper summarizes the global characteristics of the project.
European Journal of Human Genetics advance online publication, 29 May 2013; doi:10.1038/ejhg.2013.118

Novembre, Nature 2008

HOW TO CONTROLL FOR STRUCTURE ?

- Genomic controll (Devlin & Roeder 1999, Biometrics)
- Structured association (Pritchard et al. 2000, Am.J.Hum.Genet.)
- Principal-components approach (Price et al. 2006, Nature Genet.)
- Mixed-model approach (Yu et al. 2006, Nature Genet.; Kang et al. 2008, Genetics)
- FaST-LMM -Select (Listgarten et al. 2013, Nature Genet.)
- ....

The mixed model is the only one that works

- Mixed model handles different levels of relatedness generally:

$$Y = X\beta + u + \epsilon, \; u \sim N(0, \sigma_g K), \; \epsilon \sim N(0, \sigma_e I)$$

LM: $Y = X\beta + \varepsilon$

Introduced by R.A. Fischer (1918) and further developed by Henderson in the 1950s

The mixed model

$$LMM: y = \mu + X\beta + u + \varepsilon$$

- y is a n x 1 vector of the observed phenotype
- μ is a fixed effect representing the phenotypic mean
- X is an n x q matrix of fixed effects
- β is a q x 1 vector of the coefficients of the fixed effects
- *u* is the random effect of the mixed model with $var(u) = \sigma_g K$
- K is the n x n kinship matrix inferred from genotypes
- ε is a n x n matrix of residual effects with $var(\varepsilon) = \sigma_e I$

The overall phenotypic variance-covariance matrix can be represented as

$$V = \sigma_g K + \sigma_e I$$

Pseudo heritability $\hat{h}^2 = \hat{\sigma}_g^2/(\hat{\sigma}_g^2+\hat{\sigma}_e^2)$

$$\text{LMM: } y = \mu + X\beta + u + \varepsilon$$

Yu et al. (2006) : solving the model with best linear unbiased predictor (BLUP) via the Henderson equation (Henderson 1984)

Kang et al. (2008) : direct estimation of the variance components (σ_g and σ_e) by maximizing the restricted likelihood (EMMA)

$$l_F(y;\beta,\sigma,\delta) = \frac{1}{2}\left[-n\,\log(2\pi\sigma^2) - \log|H| - \frac{1}{\sigma^2}(y - X\beta)'H^{-1}(y - X\beta)\right] \quad (2)$$

$$l_R(y;\sigma,\delta) = l_F(y;\hat{\beta},\sigma^2,\delta) + \frac{1}{2}\left[q\log(2\pi\sigma^2) + \log|X'X| - \log|X'H^{-1}X|\right] \quad (3)$$

Kang et al. (2010) : fast heuristic approximation that assumes σ_g/σ_e is constant (EMMAX)

Zhang et al. (2010): P3D

Lippert et al.(2011): FaST-LM

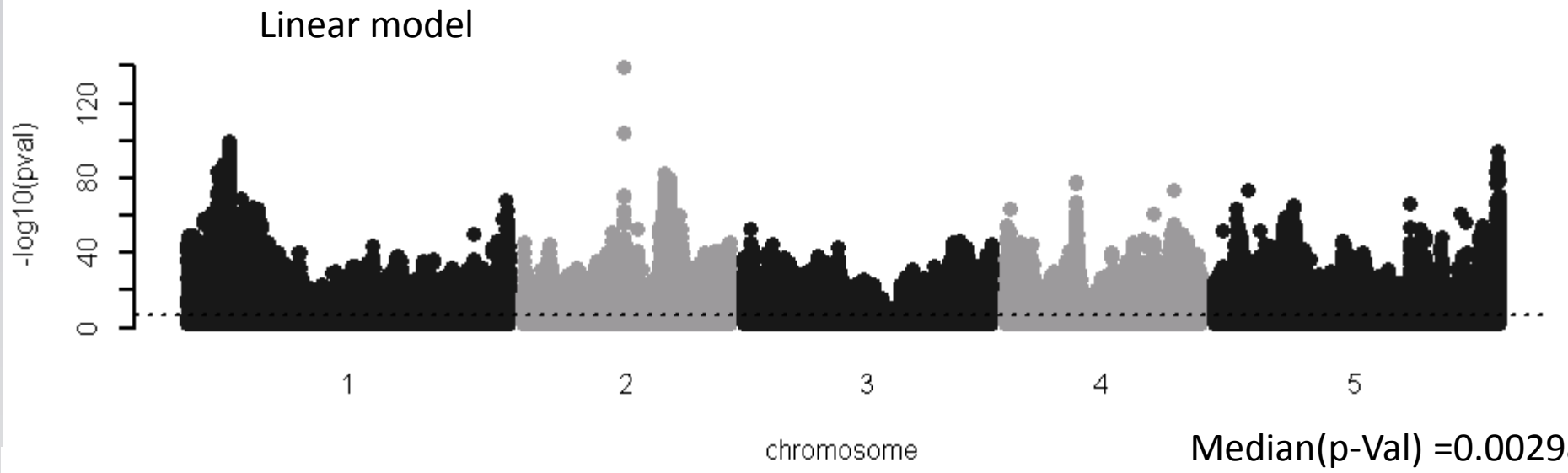
Model Assumptions

There are four principal assumptions which justify the use of linear regression :
(i) linearity of the relationship between dependent and independent variables
(ii) independence of the errors
(iii) constant variance of the errors
(iv) normality of the error distribution

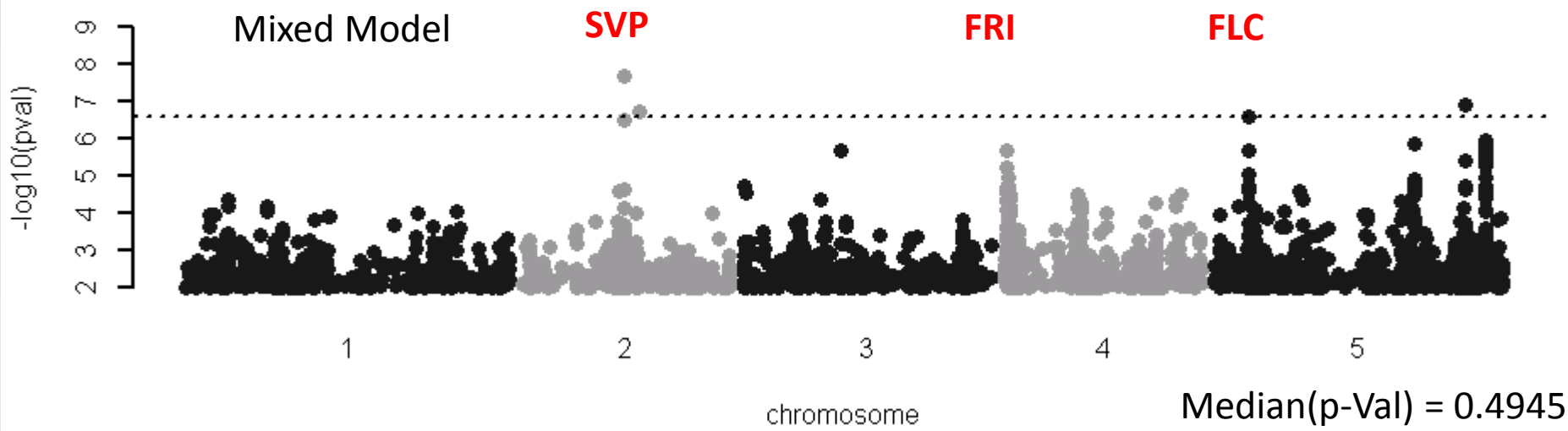
If any of these assumptions is violated model predictions can be wrong, which makes the interpretation of the results inefficient or even misleading.

One alternative could be non-parametric models (e.g. Kruskal-Wallis or Anderson-Darling test) or generalized linear models

It works !

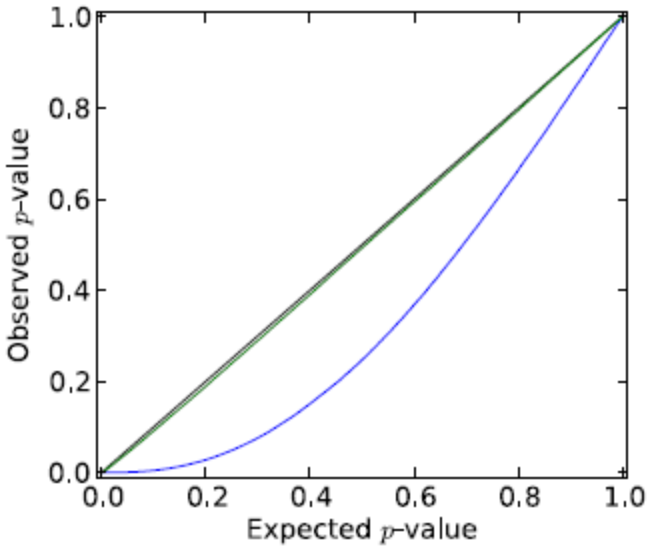


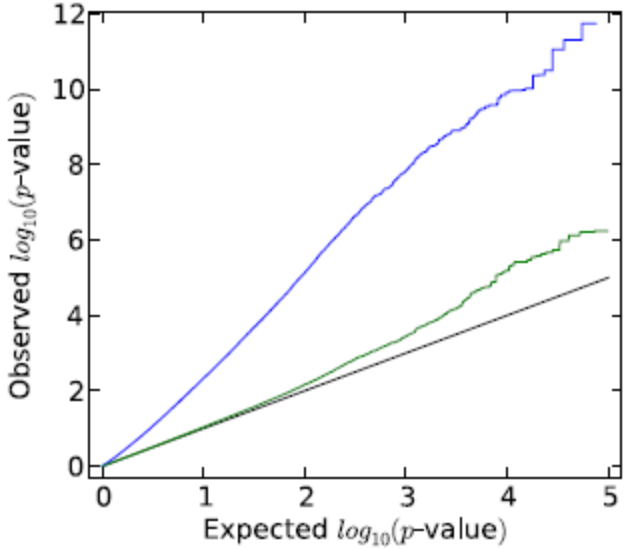
Median(p-Val) =0.00298



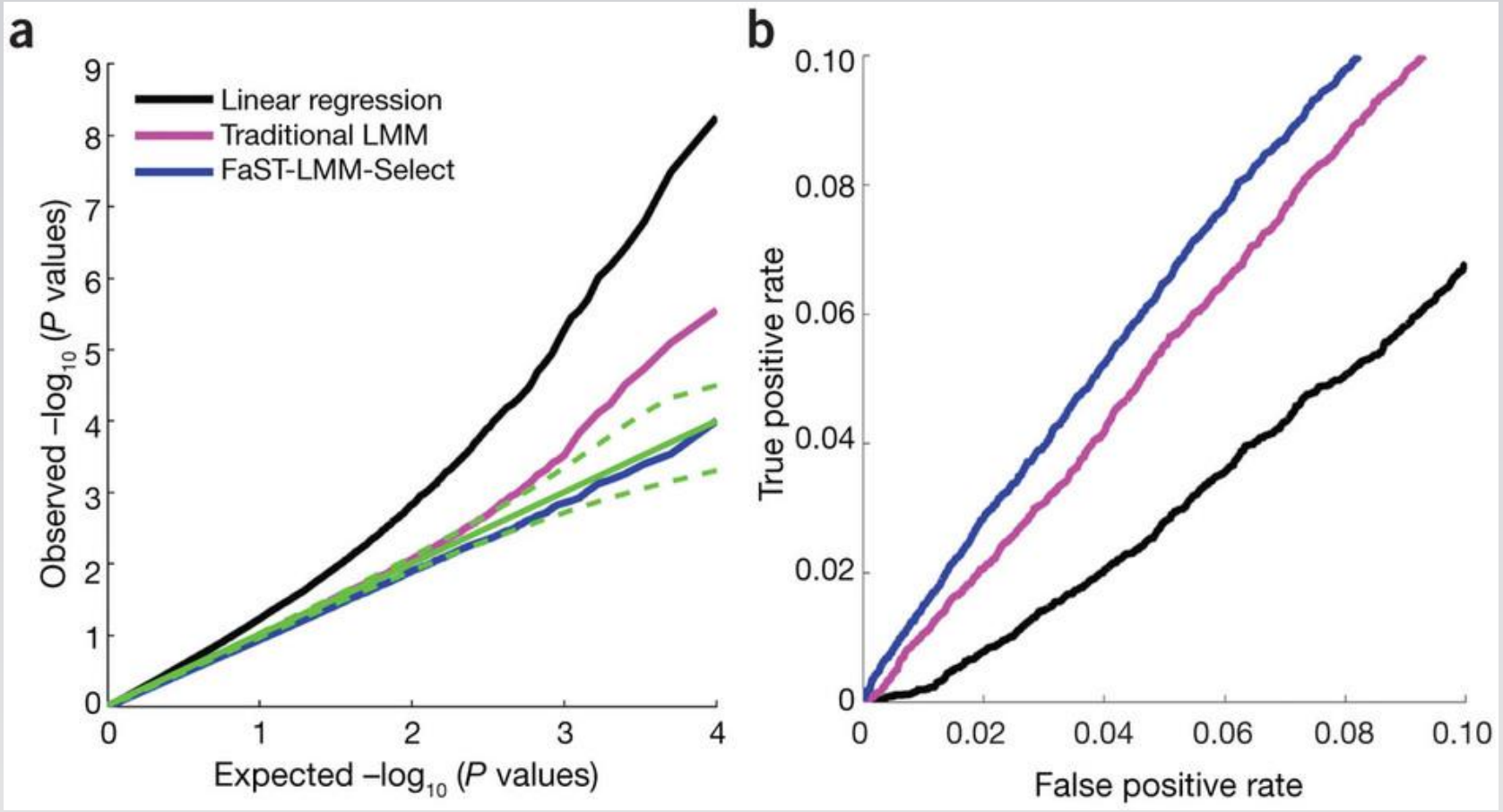
Median(p-Val) = 0.49452

qq-plots





- We still have false positives — and have probably introduced false negatives...

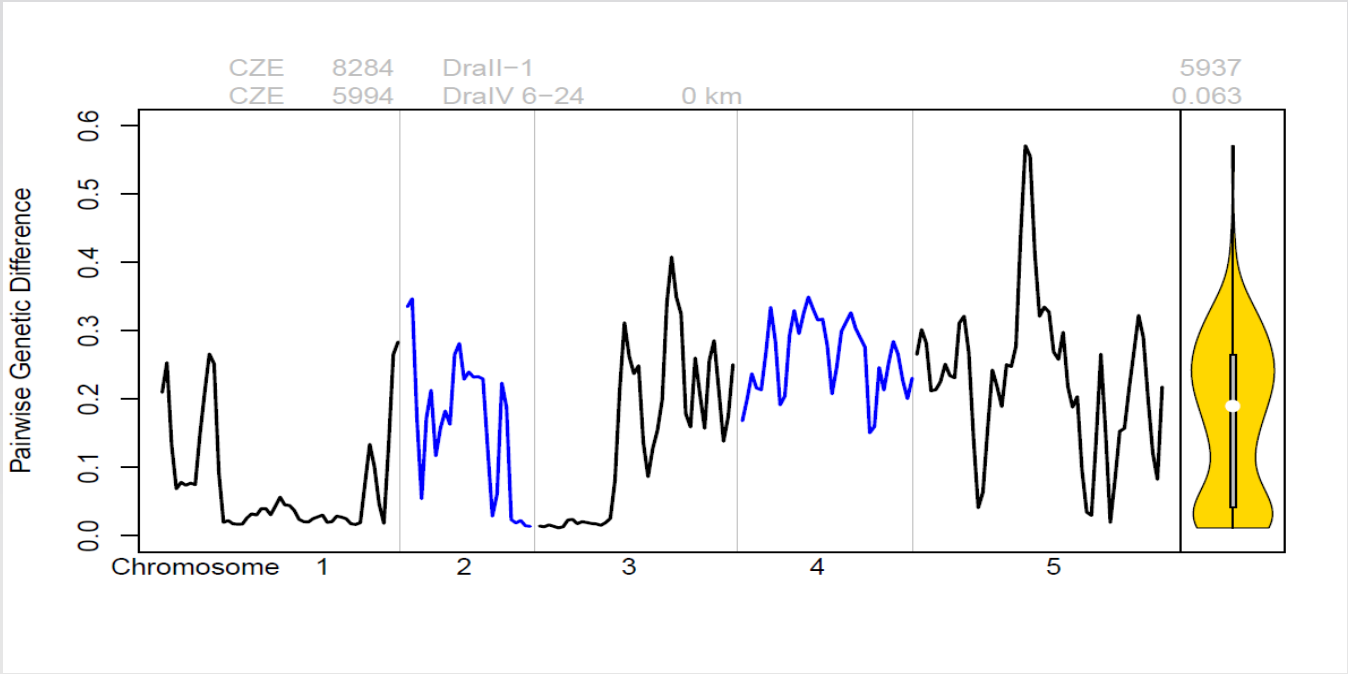

a
Linear regression
Traditional LMM
FaST-LMM-Select
Observed −log10 (P values)
Expected −log10 (P values)
0
1
2
3
4
5
6
7
8
9
b
True positive rate
False positive rate
0
0.02
0.04
0.06
0.08
0.10

Kinship matrix

Kinship matrix could be calculated as IBD – matrix, but this assumses the contribution of each loci to the trait is independend of its allele frequency.

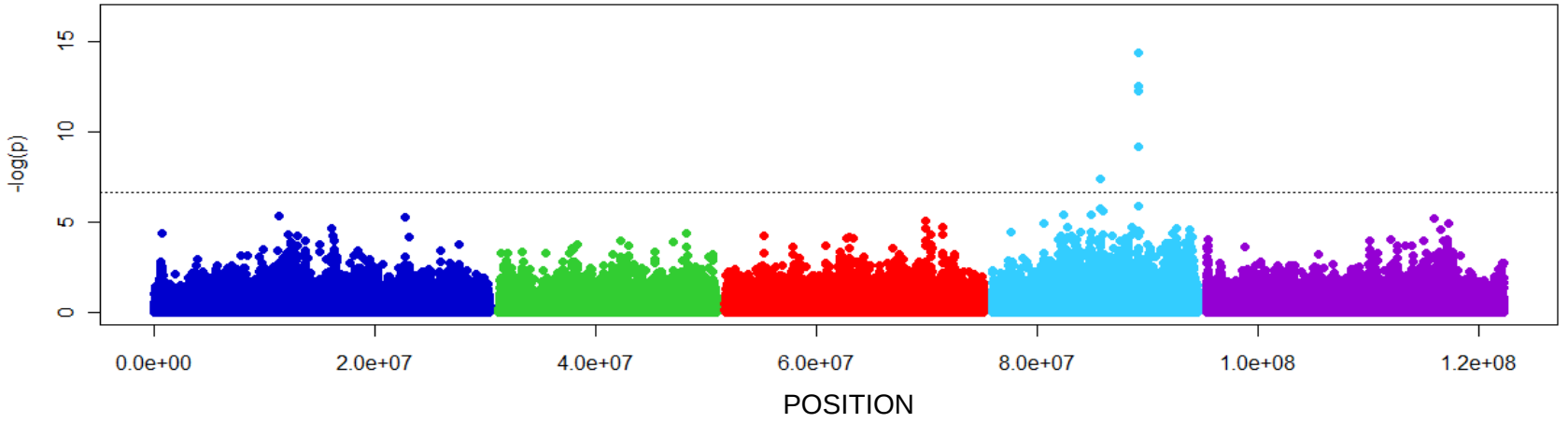
A nother possibility is the calculation of an IBS matrix, estimated using all SNPs (infinitesimal model)

Additionally methods are proposed to select only subsets of SNPs to calculate the kinship matrix (see e.g. Listgar ten et al. 2013)

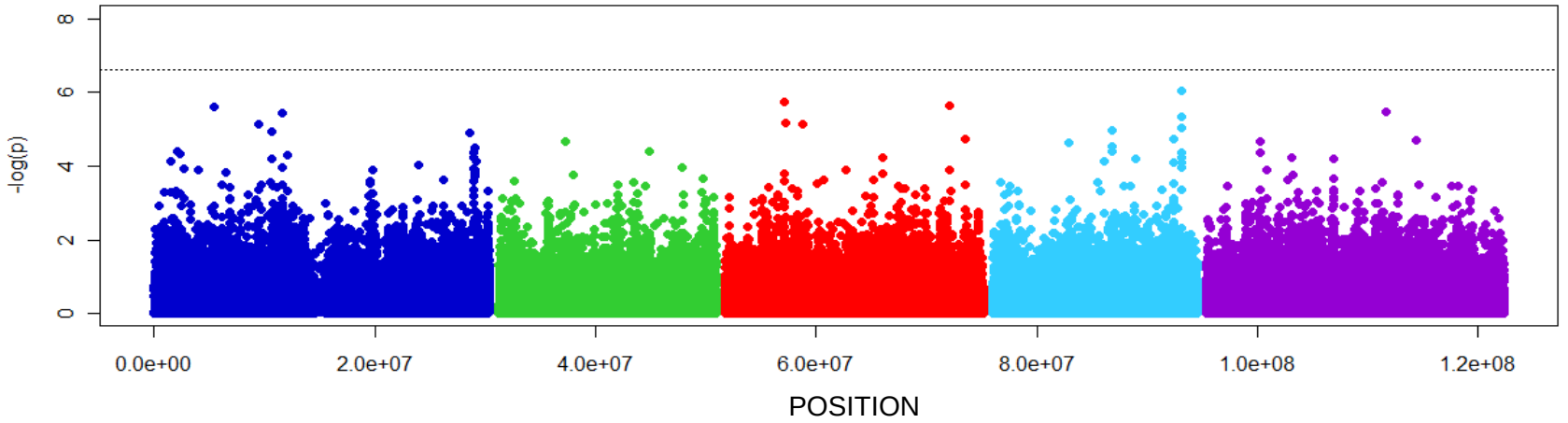


Examples of real traits

A: GWA analysis of hypersensitive response to bacterial elicitor



B: GWA analysis of germination on MS medium



Considerations

•Trait architecture (simple vs. complex traits, how many causative Alleles, allecis heterogenity, epistasais etc….)

• Phenotypic distribution of the trait (normal ?, outliers ?)

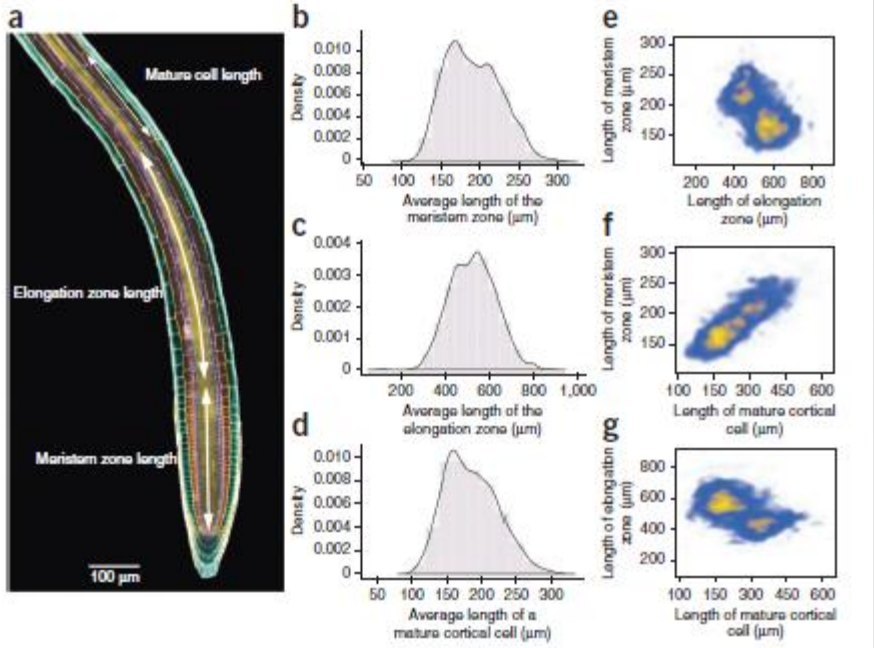
• Correlation between the phenotype and non-genetic parameters

GWAS on cellular traits

nature genetics

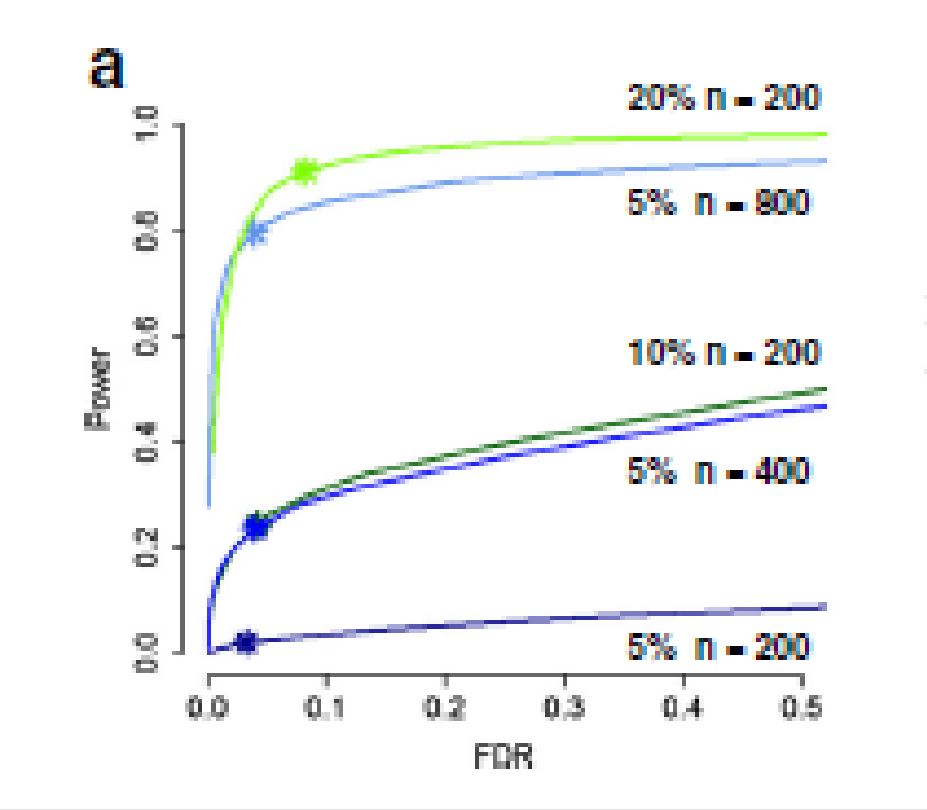
Genome-wide association study using cellular traits identifies a new regulator of root development in *Arabidopsis*

Mónica Meijón[1,2], Santosh B Satbhai[1], Takashi Tsuchimatsu[1] & Wolfgang Busch[1]



Meijon *et al.* Nature Genetics (2014)

Power / FDR

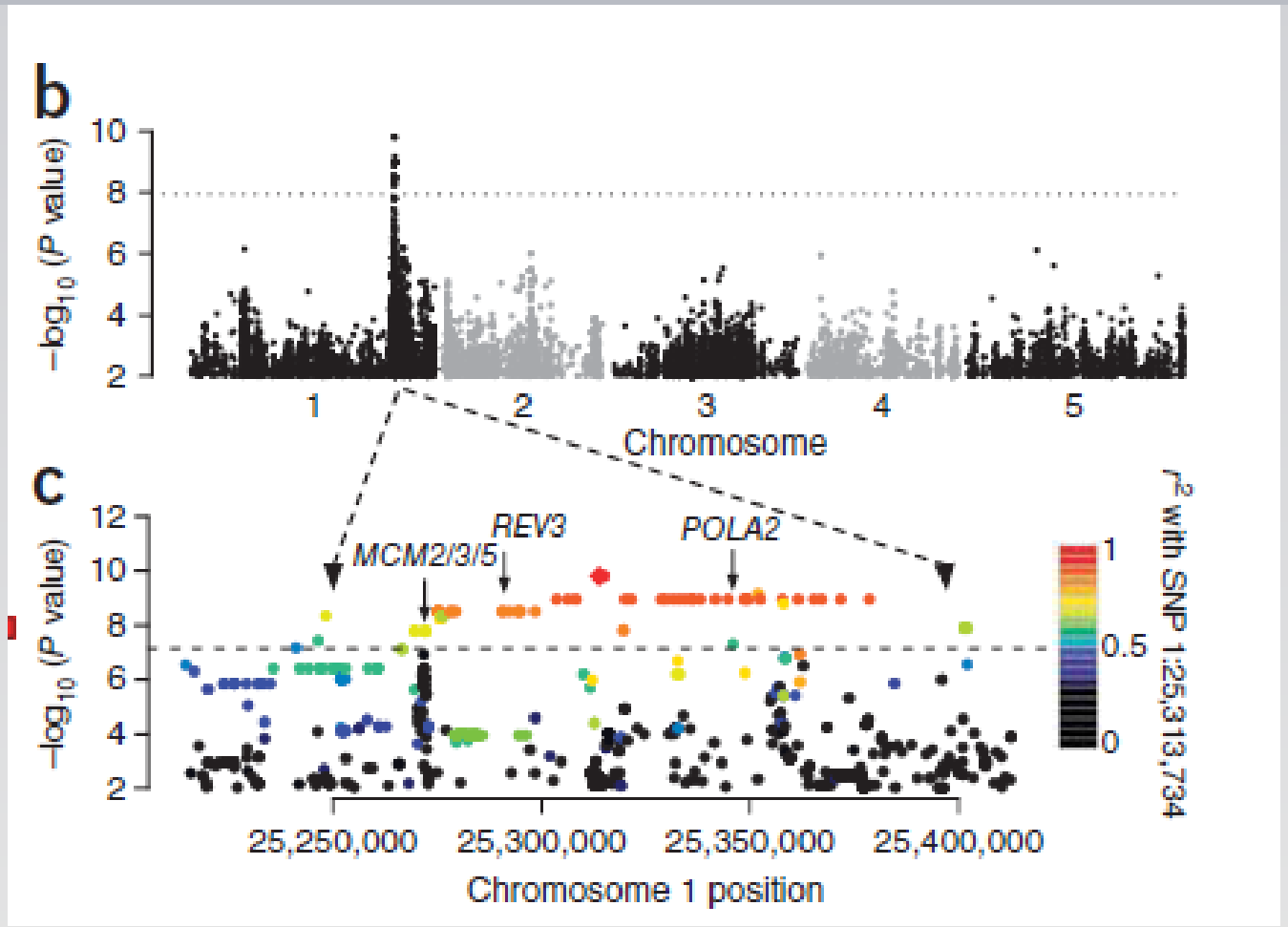


Korte and Farlow (Plant Methods) 2013

Another problem

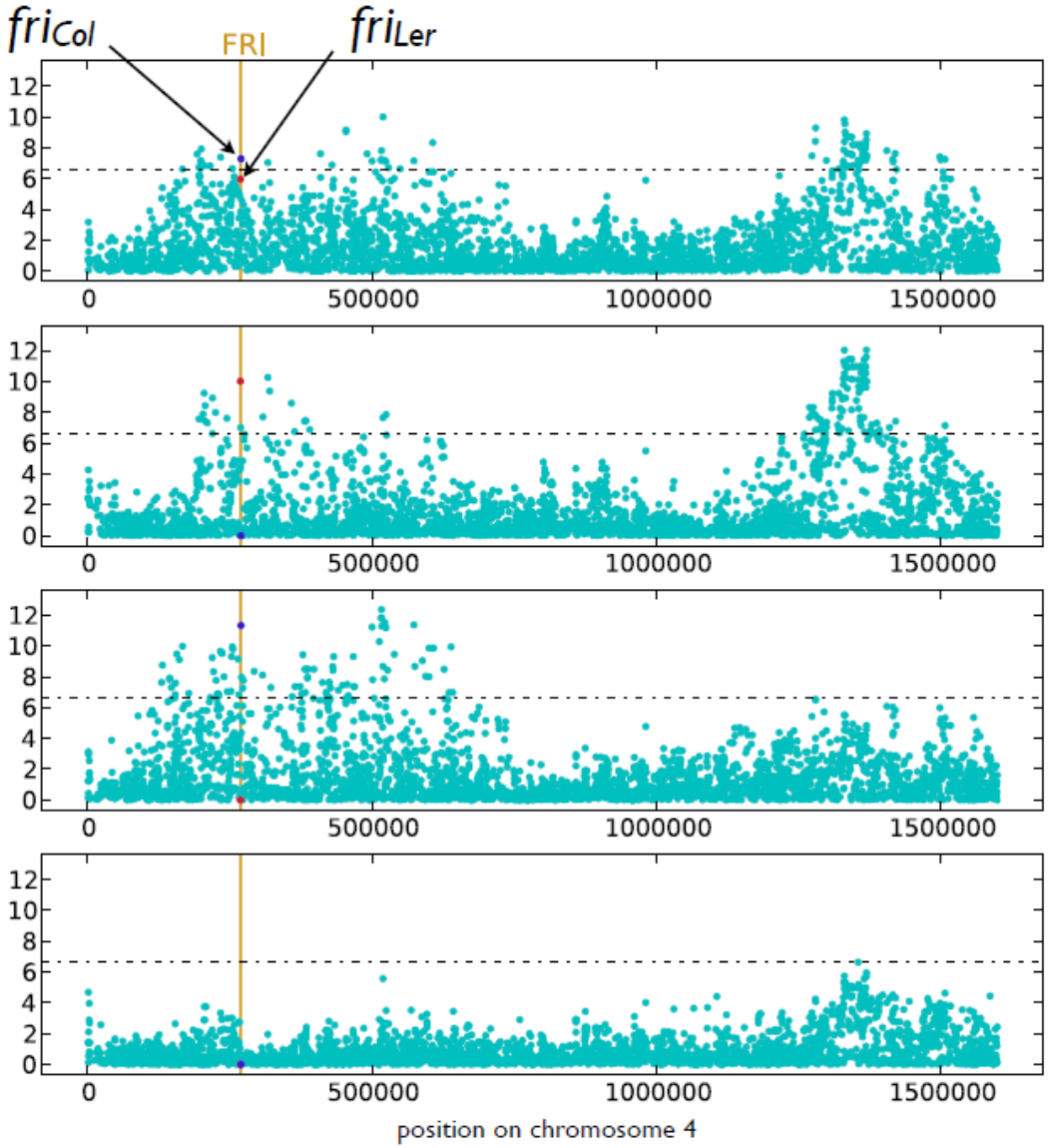
- We have seen that it is usually difficult to decide which peaks are significant
- In addition, peaks are often complex, making it difficult to pinpoint causal sites

GWAS peaks can be complex



Long *et al.* (Nature Gentics) 2013

Example of a complex GWAS peak



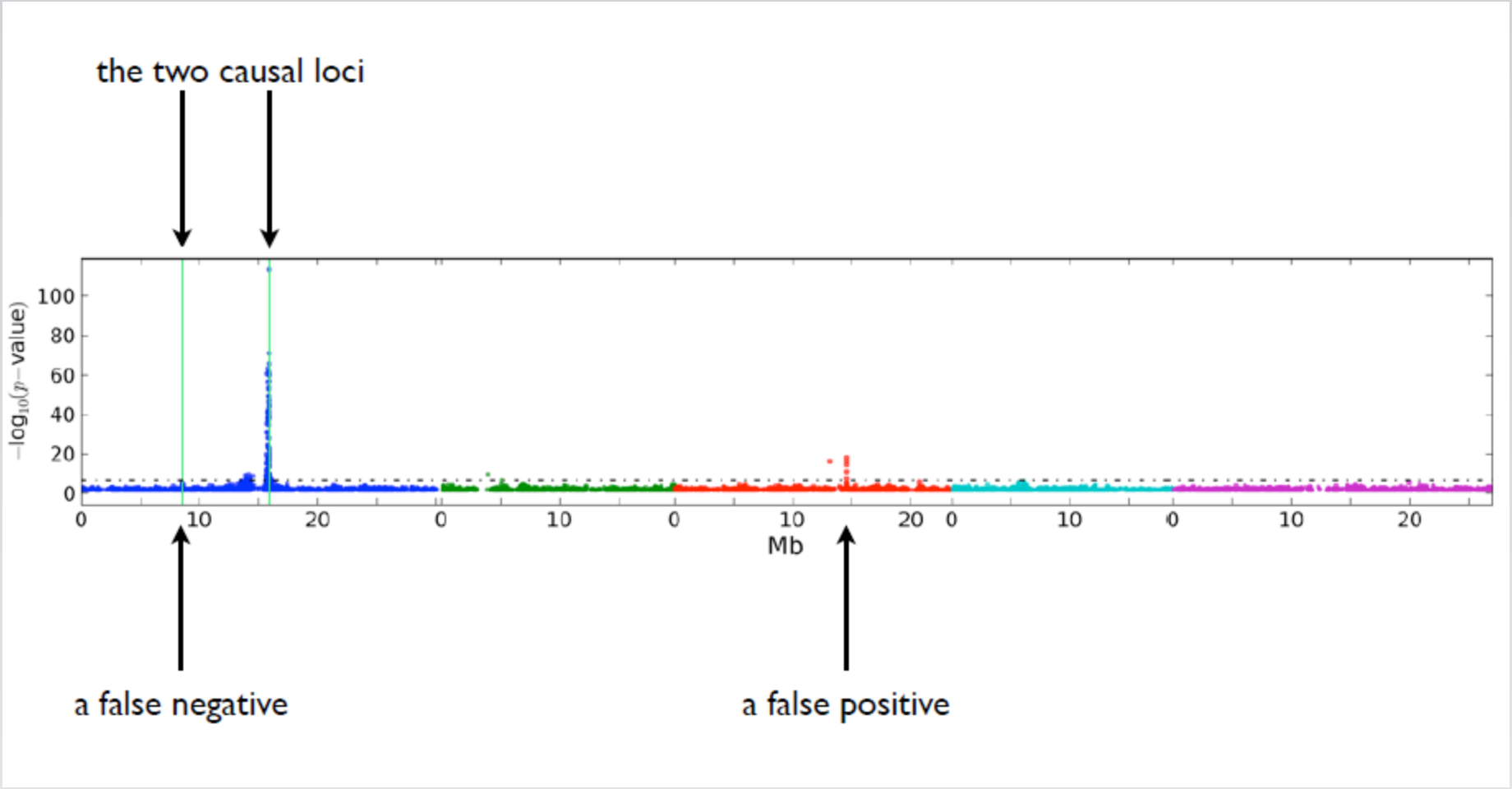
Atwell et al. 2010

Conditions under which GWAS will be positively misleading

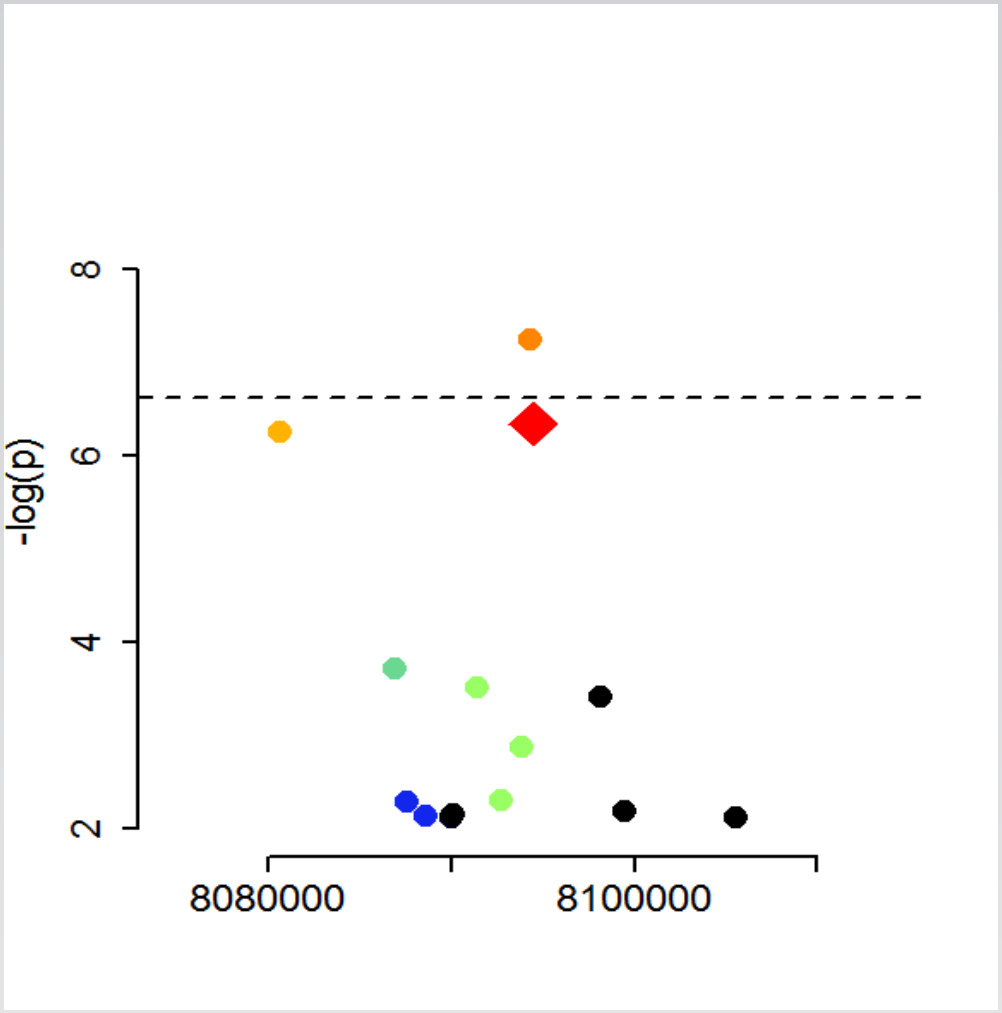
- Correlation between causal factors and unlinked non-causal markers
- More than one causal factor
- Epistasis

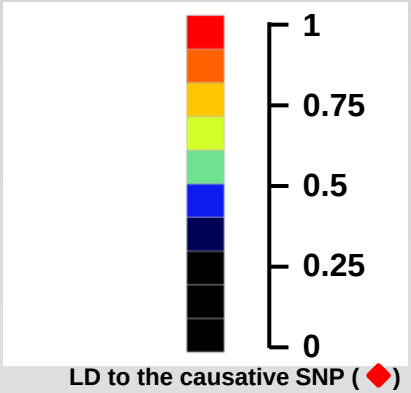
(Platt et al., *Genetics* 2010)

Sometimes it is bad !



And it can become worse !





- **A non-causative SNP can show stronger association with a given phenotype than the actual causative SNP.**

Korte and Farlow *Plant Methods* (2013)

- Population structure *per se* is neither necessary or sufficient for GWAS to go wrong
- The fundamental problem is *not* population structure, it is model mis-specification — modeling something multifactorial using a single locus
- Population structure greatly aggravates the problem, however

COMMENT

The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] *and Magnus Nordborg*[3,4]

The authors argue that population structure per se is not a problem in genome-wide association studies — the true sources are the environment and the genetic background, and the latter is greatly underappreciated. They conclude that mixed models effectively address this issue.

> "population structure is not the fundamental source of the problem, and removing it is not the solution"

Thanks to dramatically decreasing genotypin
sequencing costs, genome-wide association s
(GWASs) are becoming the default method for stu
the genetics of natural variation. The increasing
ber and diversity of GWASs will require appro
statistical analysis methods. The most basic prob
assessing the significance of an association in the l
confounding effects that may cause spurious associ

The aspect of this problem that has receiv
most attention is the danger of false positives in
tured populations. If the study population is a m
of populations that differ with respect to allele fre
cies as well as the trait of interest, spurious correl

Korte and Farlow *Plant Methods* 2013, **9**:29
http://www.plantmethods.com/content/9/1/29

PLANT METHODS

REVIEW **Open Access**

The advantages and limitations of trait analysis with GWAS: a review

Arthur Korte*† and Ashley Farlow†

Abstract

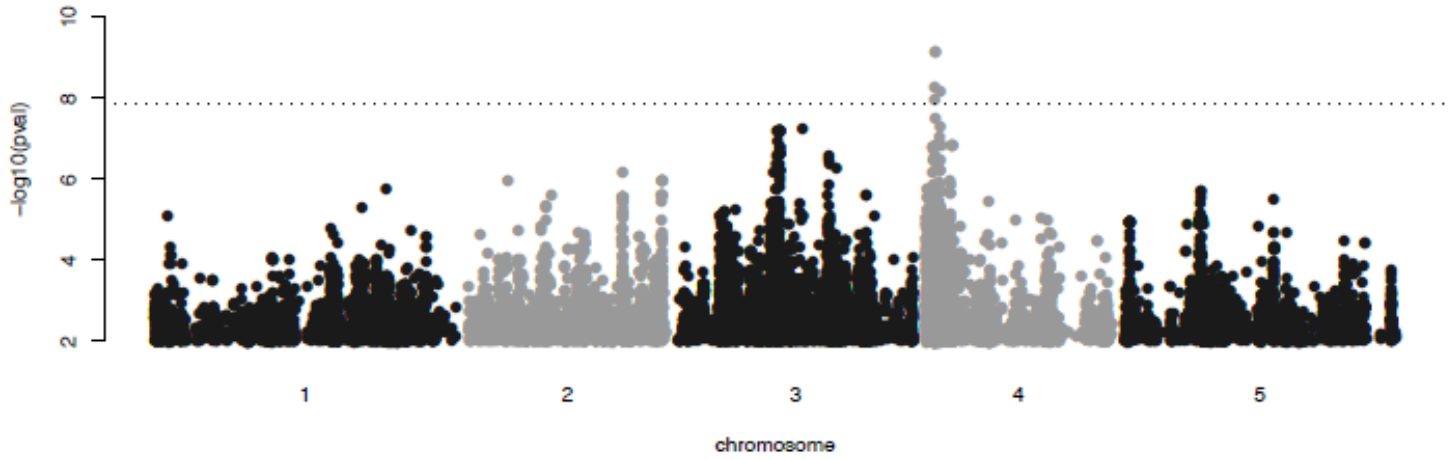
Over the last 10 years, high-density SNP arrays and DNA re-sequencing have illuminated the majority of the genotypic space for a number of organisms, including humans, maize, rice and *Arabidopsis*. For any researcher willing to define and score a phenotype across many individuals, Genome Wide Association Studies (GWAS) present a powerful tool to reconnect this trait back to its underlying genetics. In this review we discuss the biological and statistical considerations that underpin a successful analysis or otherwise. The relevance of biological factors including effect size, sample size, genetic heterogeneity, genomic confounding, linkage disequilibrium and spurious association, and statistical tools to account for these are presented. GWAS can offer a valuable first insight into trait architecture or candidate loci for subsequent validation.

Keywords: GWAS, *Arabidopsis*, Mixed model, Effect size, Genetic heterogeneity

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Slug damage GWAS identifies peak consistent with AOP2/3



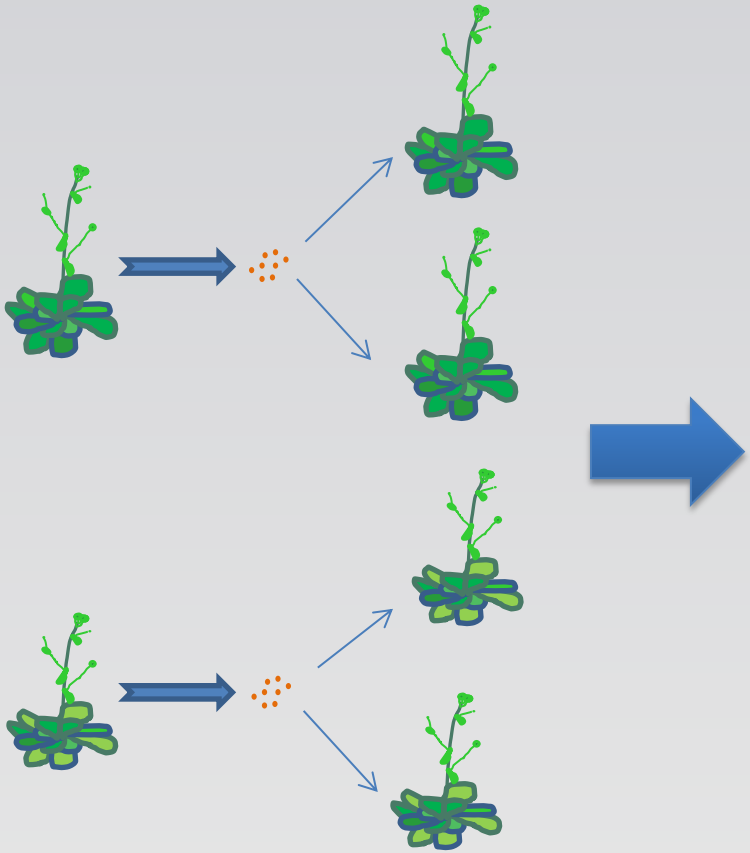
GWAS in *Arabidopsis thaliana*

CTB Center for Computational and Theoretical Biology

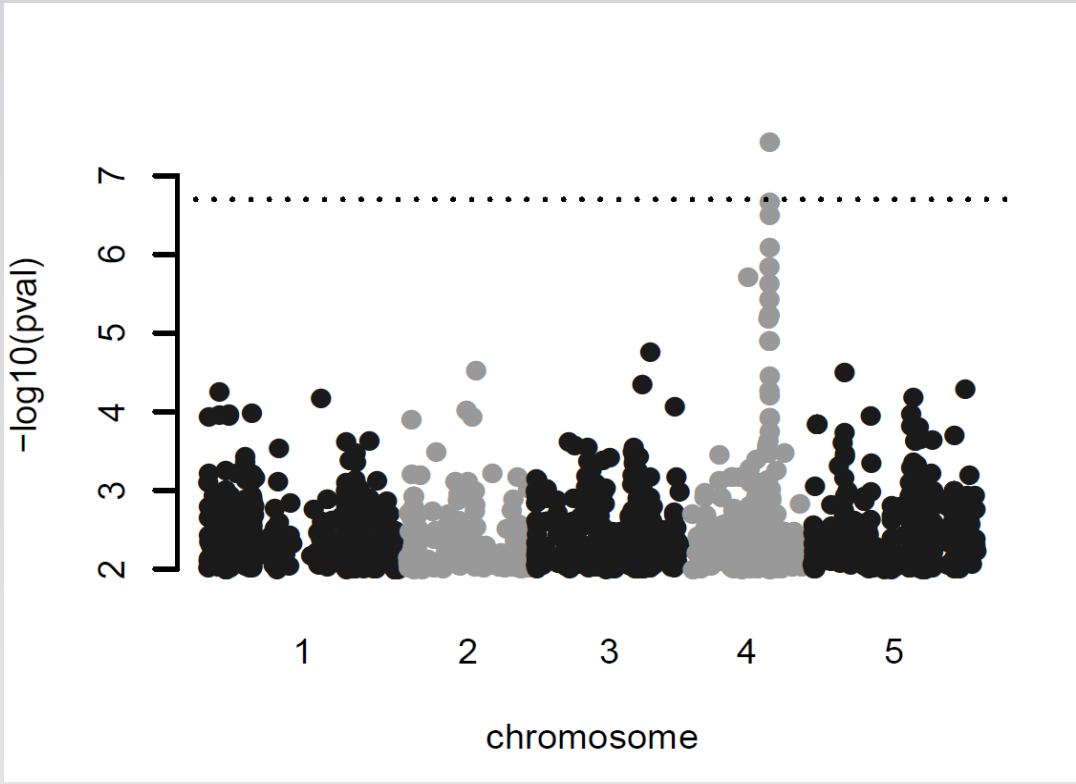
GWAS in *Arabidopsis thaliana*

Center for Computational and Theoretical Biology

Phenotype



GWAS



- **Robust (replicated) phenotypes in controlled environments**
- **Follow-up studies are easy**

www.1001genomes.org

Center for Computational and Theoretical Biology

1001 Genomes
A Catalog of *Arabidopsis thaliana* Genetic Variation

Home | Data Providers | Accessions | Tools | Software | Data Center | About

Welcome to the 1001 Genomes Project

AUTHOR CONTRIBUTIONS

Project coordination

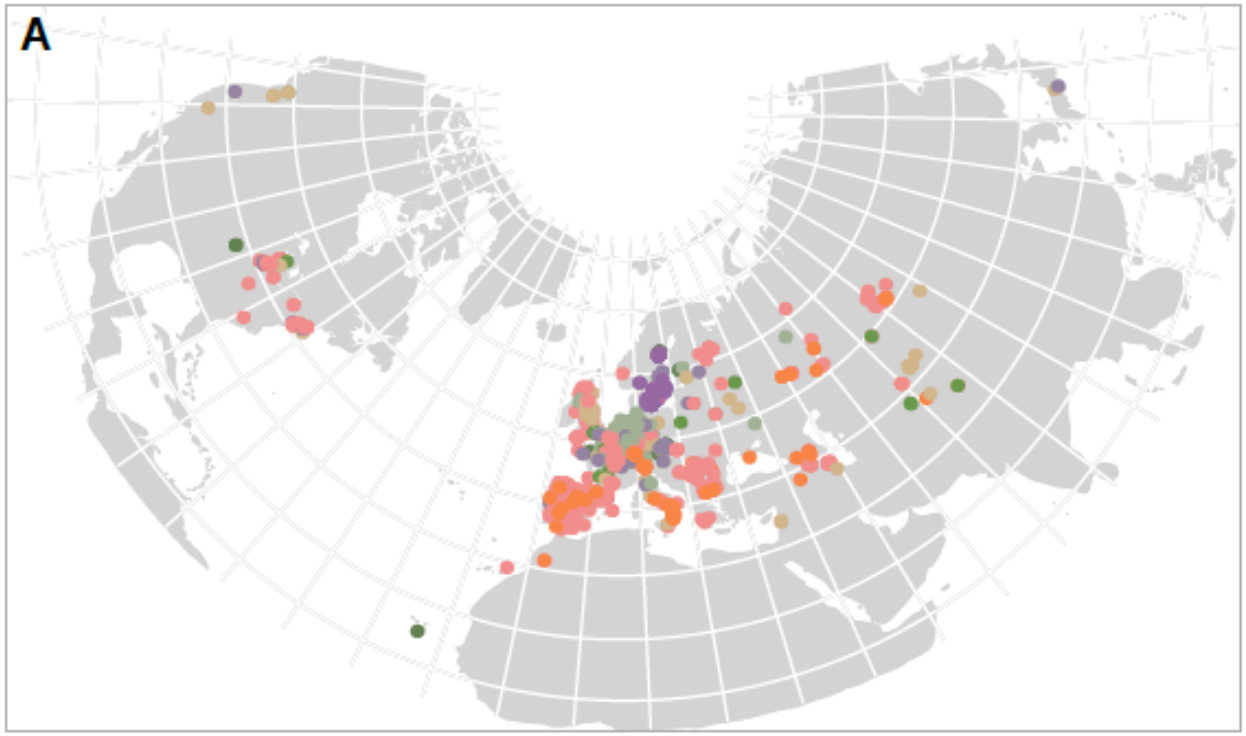
Joy Bergelson	UCHICAGO	jbergels@uchicago.edu
Joe R. Ecker	SALK	ecker@salk.edu
Magnus Nordborg	GMI	magnus.nordborg@gmi.oeaw.ac.at
Mitchell Sudkamp	MONSANTO	mitchell.sudkamp@monsanto.com
Detlef Weigel	MPI	weigel@weigelworld.org

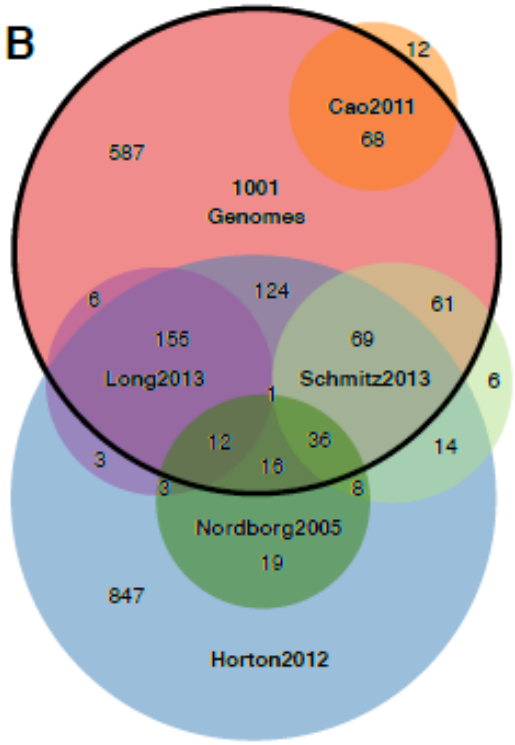
Go »

80 strains (D. Weigel lab, MPI)
195 strains (J. Ecker lab, Salk)
180 strains (M. Nordborg Lab, GMI)

The 1001 Genomes Vision

Origins of the 1,001 Genomes accessions

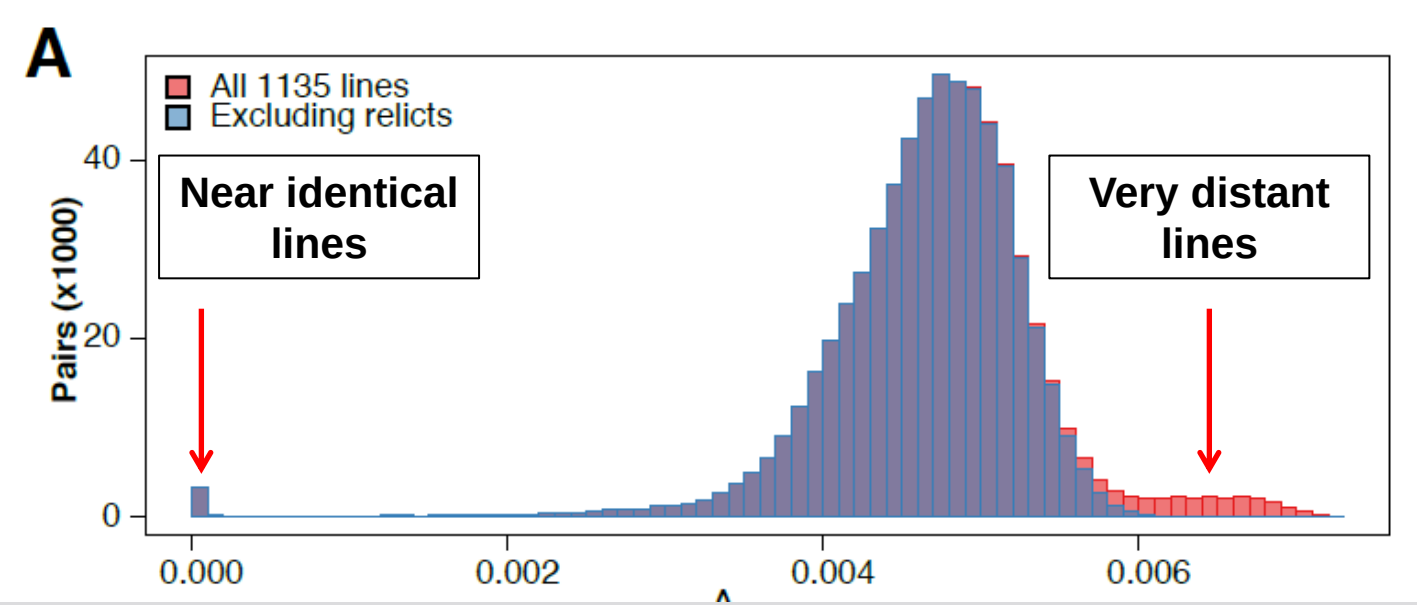




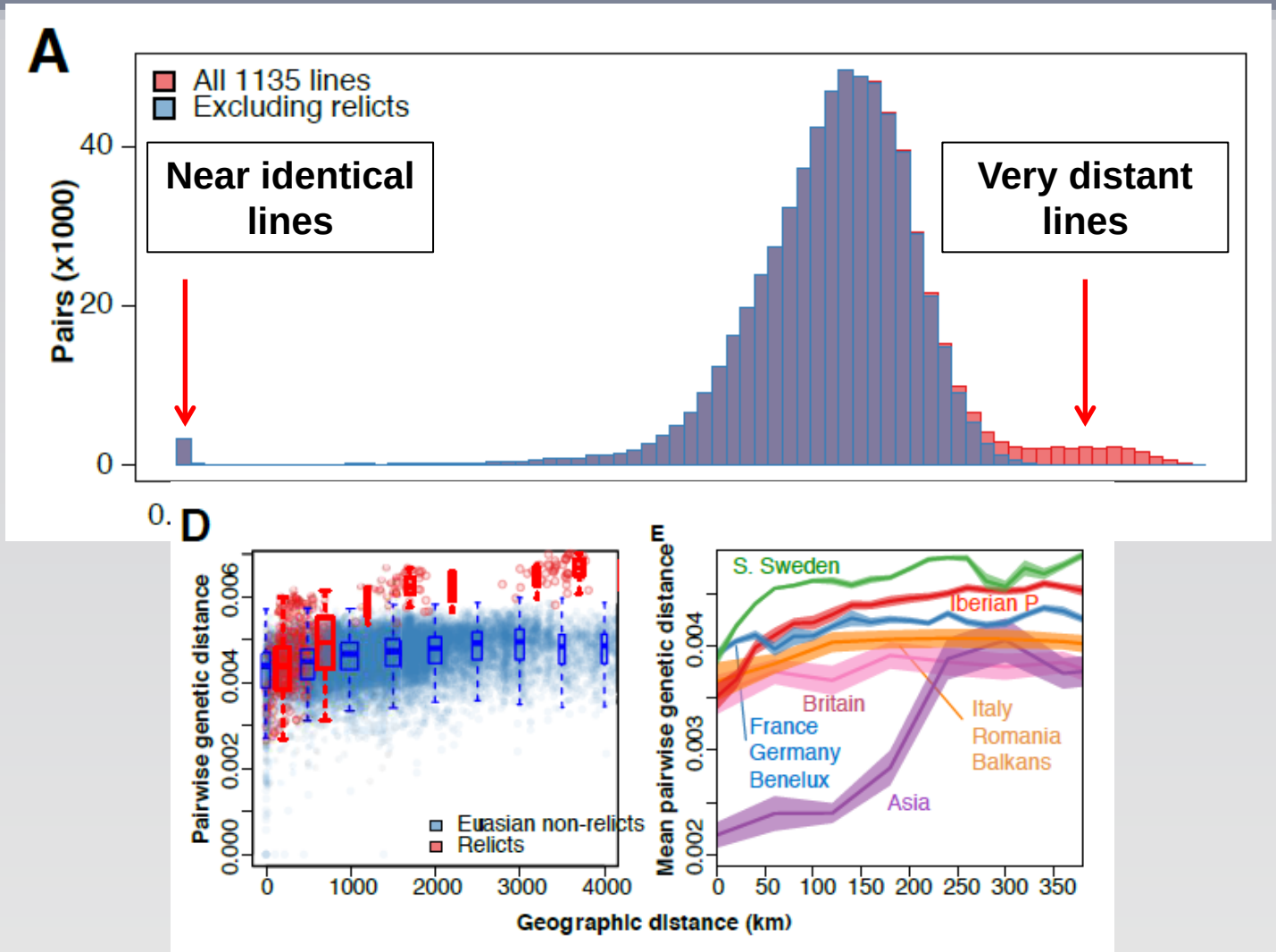
Full genomes for a world wide collection of 1135 different natural inbred lines

The 1001 Genomes consortium, ***submittted***

Genetic and geographic distance



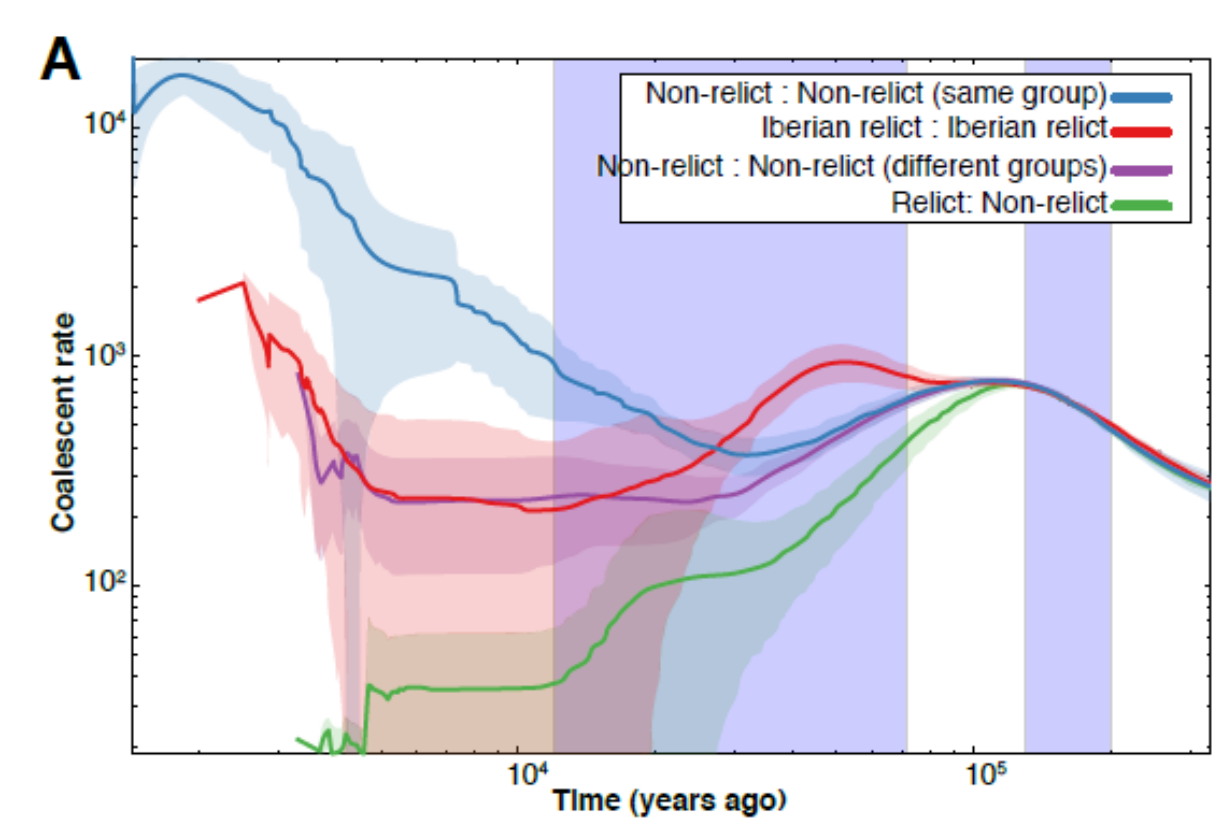
Genetic and geographic distance



Isolation by distance seems to explain the observed diversity, but differs greatly between different subsets

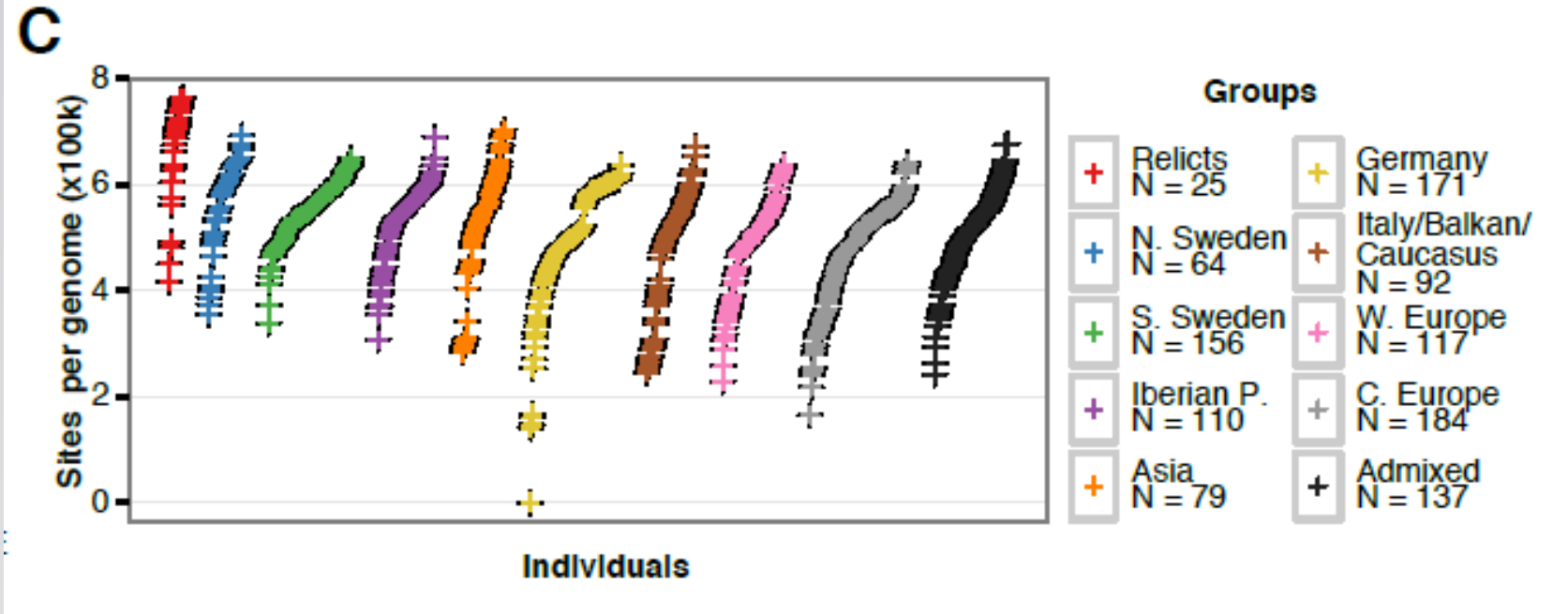
Effect of the last glacial maximum

Inferred coalescence rates using MSMC (Schiffels and Durbin 2014)



Clear seperation of groups during the last glacial maximum

Genetic diversity in different groups



The total (sequenced) population contains 25 relicts and 8 different groups of *modern* accessions

Short summary

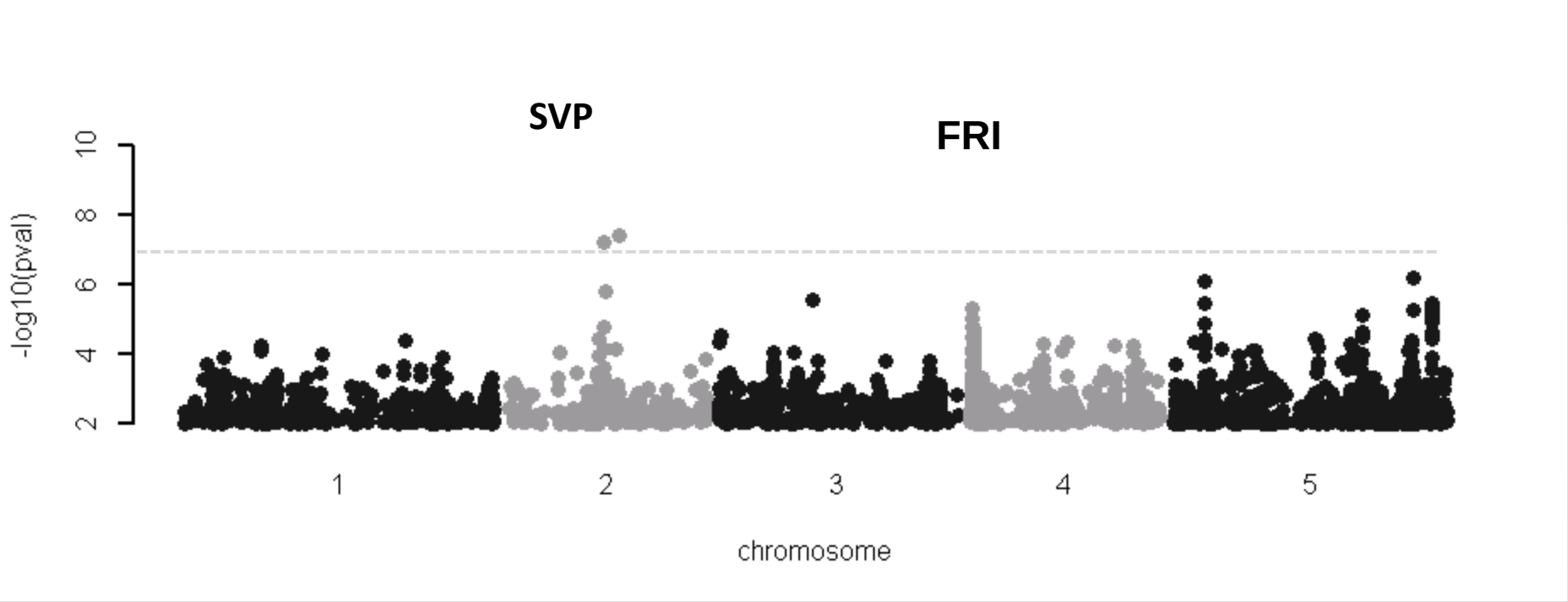
1135 sequenced natural inbred lines reveal the global pattern of polymorphism in *Arabidopsis thaliana*

The 1001 Genomes Consortium*

*Lists of participants and their affiliations appear at the end of the paper.

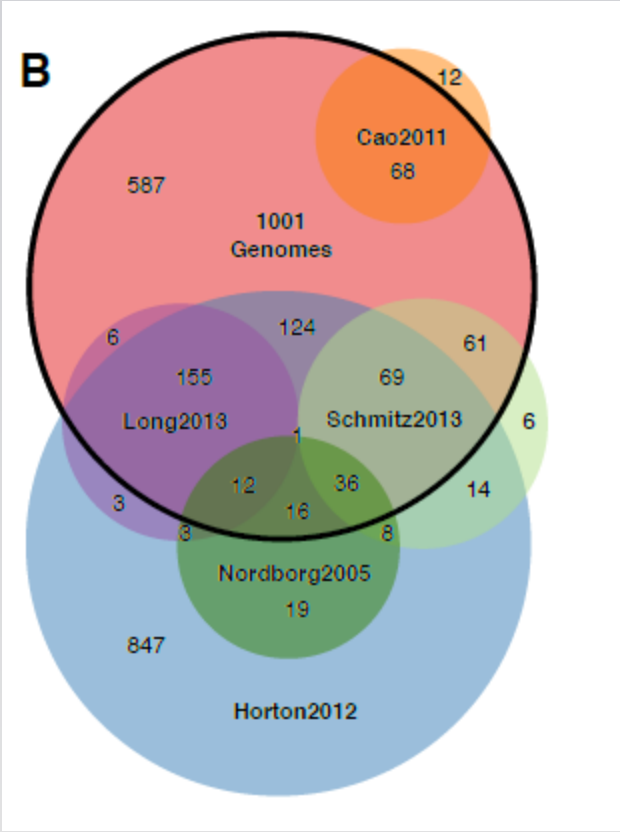
Data : 1,135 high quality genomes with more than 10 M SNP and 500k structural variants

GWAS: 250k data



250 k SNPs
(1SNP/500 bp)

Inputed dataset for GWAS



Full genome data for 1,135 accessions (10M SNPs)

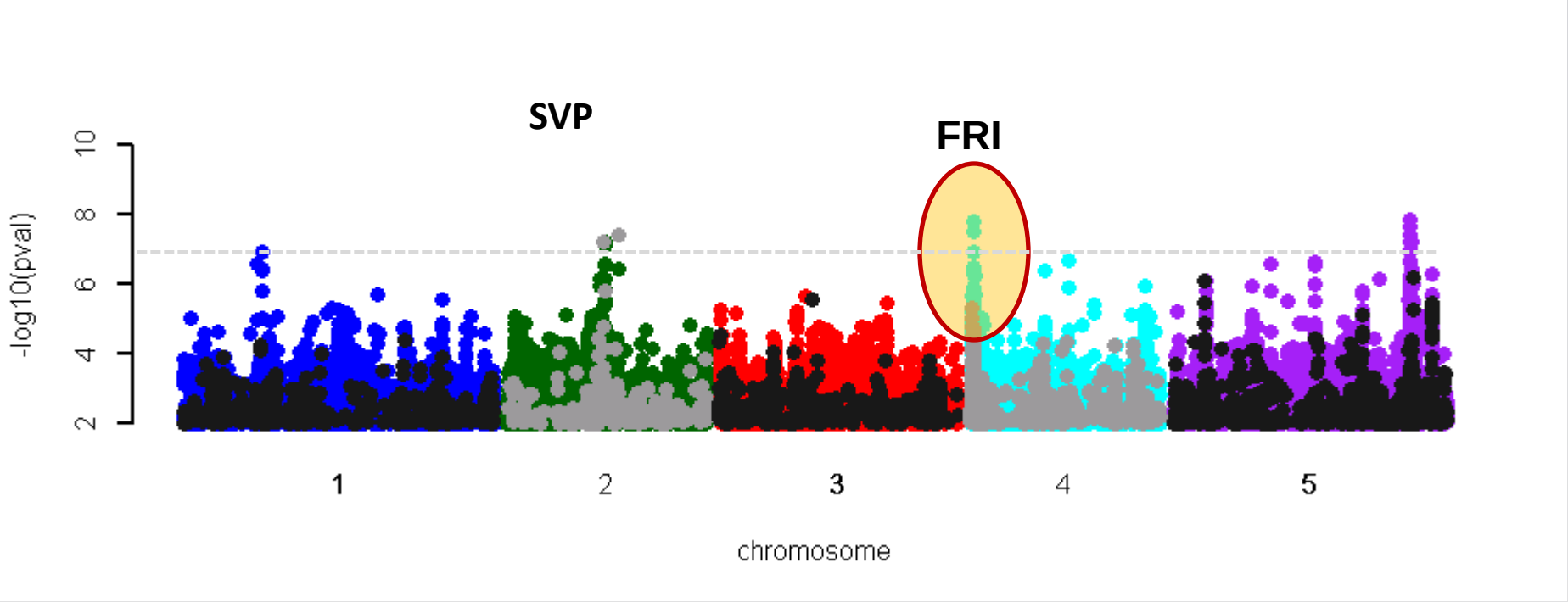
+

250k SNPs for 1,300 accessions (Horton *et al.* 2012)

Generation of an imputed data set using BEAGLE

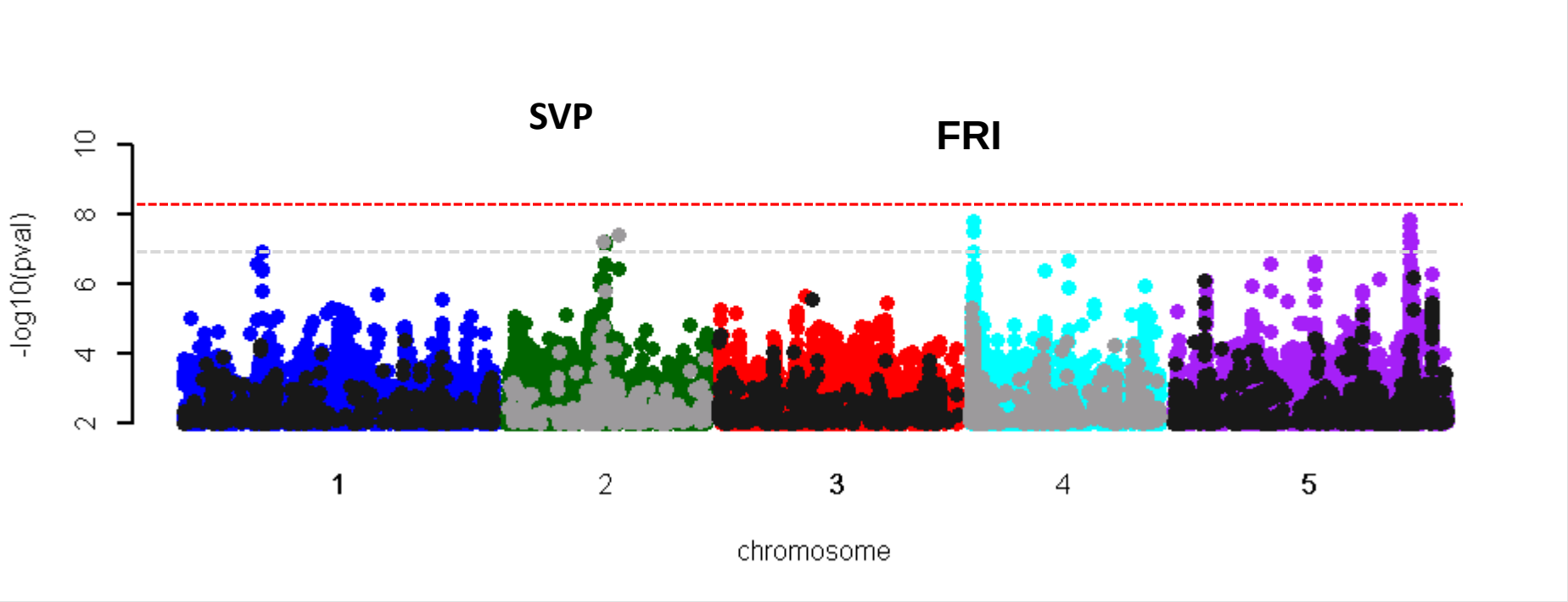
for 2,029 accessions on 10 M SNPs

GWAS: Full sequence data



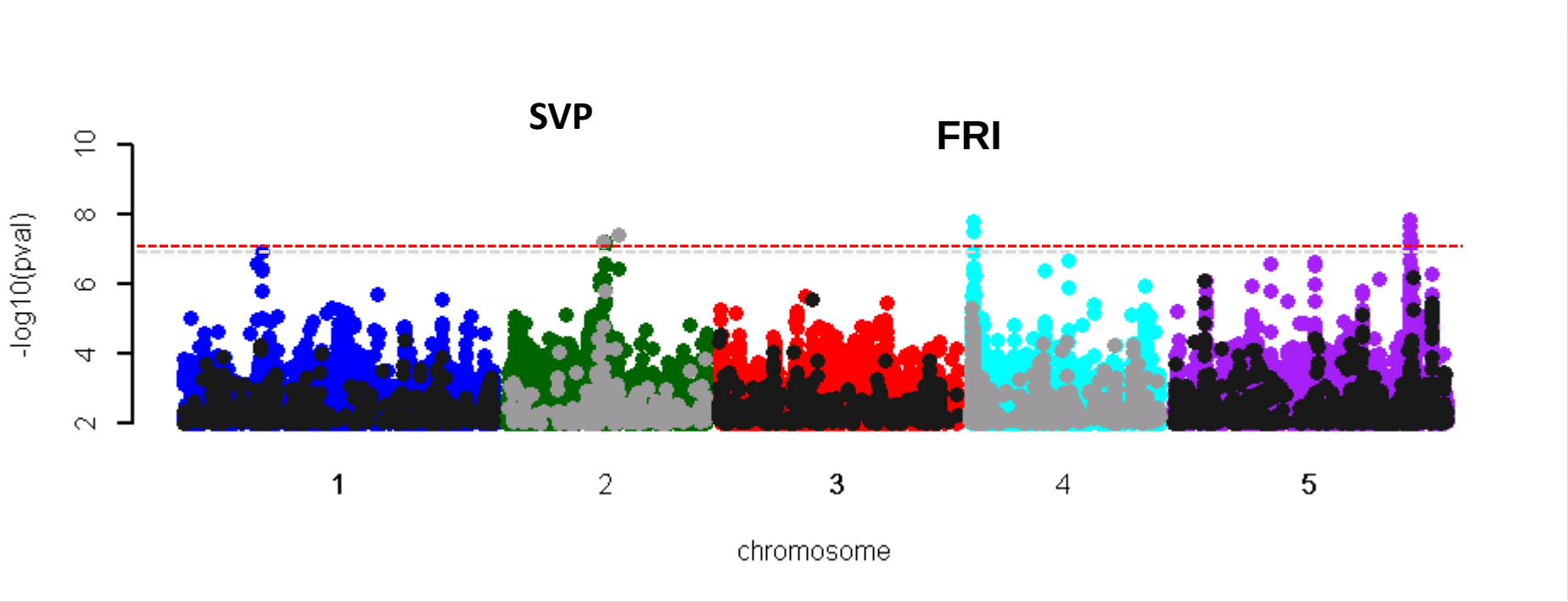
10 M SNPs

GWAS: Full sequence data



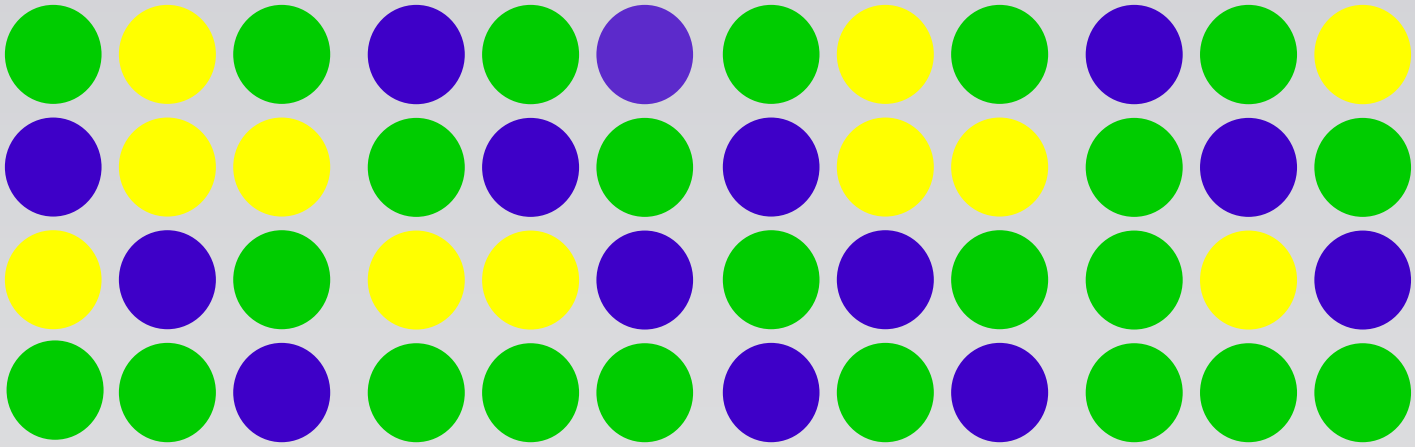
Threshold for multiple testing : Bonferoni

GWAS: Full sequence data



Permutation based threshold

Practical part Ia



Your project :
Identify genetic factors that are important for making dots blue

Dataset : 48 individuals with phenotypes, and chip-data for 1.000 homozygous SNPs

The data


```
> rm(list=ls())
> load('Y_dots.rda')
> load('X_dots.rda')
> load('K_dots.rda')
> source('emma.r')
> source('gwas.r')
> ls()
 [1] "amm_gwas"                  "emma.delta.ML.dLL.w.Z"     "emma.delta.ML.dLL.wo.Z"
 [4] "emma.delta.ML.LL.w.Z"      "emma.delta.ML.LL.wo.Z"     "emma.delta.REML.dLL.w.Z"
 [7] "emma.delta.REML.dLL.wo.Z" "emma.delta.REML.LL.w.Z"    "emma.delta.REML.LL.wo.Z"
[10] "emma.eigen.L"              "emma.eigen.L.w.Z"          "emma.eigen.L.wo.Z"
[13] "emma.eigen.R"              "emma.eigen.R.w.Z"          "emma.eigen.R.wo.Z"
[16] "emma.kinship"              "emma.ML.LRT"               "emma.MLE"
[19] "emma.MLE.noX"              "emma.REML.t"               "emma.REMLE"
[22] "emma.test"                 "K_dots"                    "X_dots"
[25] "Y_dots"
>
```

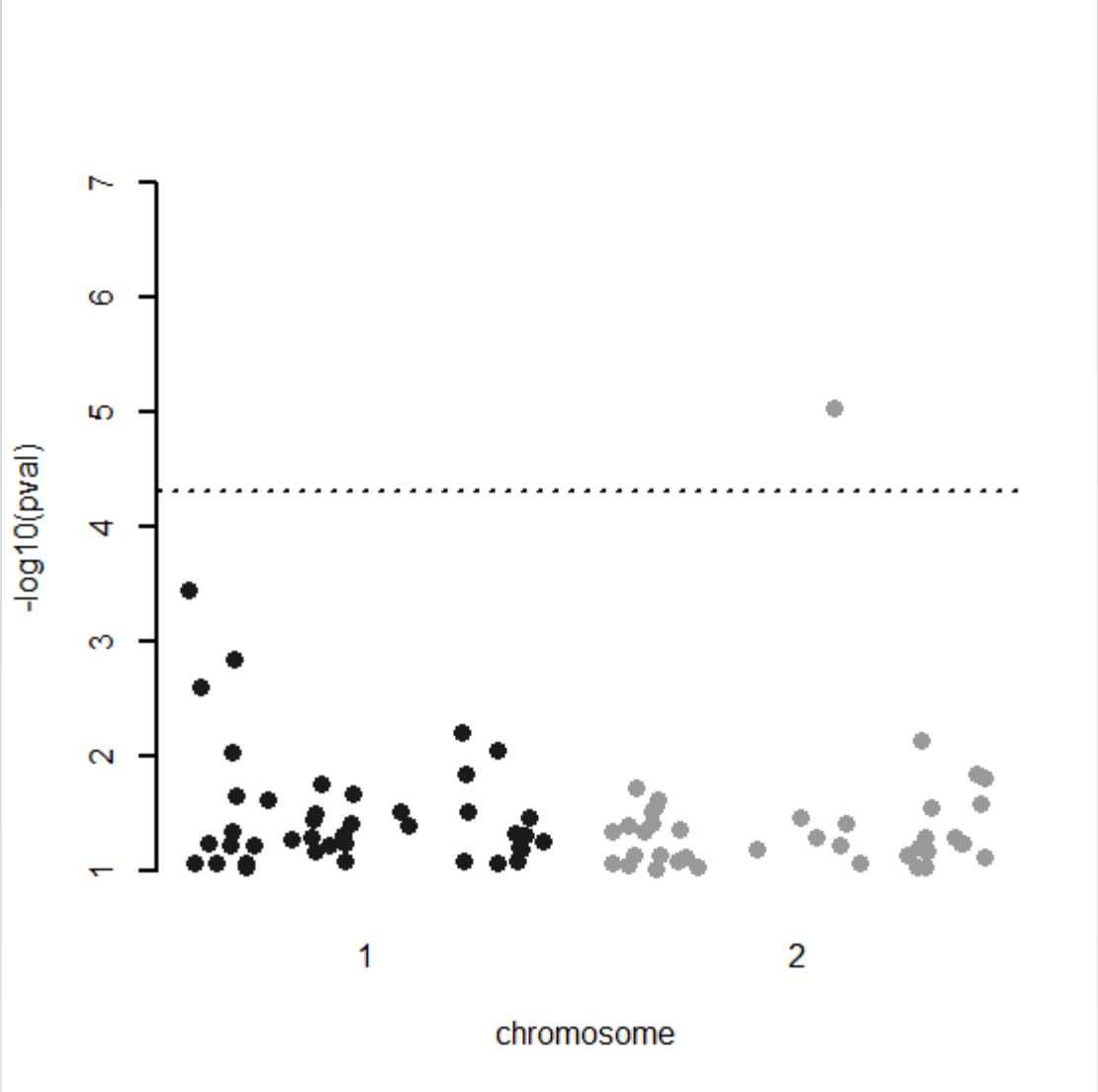
The gwas.r script


```
amm_gwas<-function(Y,X,K,p=0.001,n=2,run=T,calculate.effect.size=FALSE,
include.lm=FALSE,use.SNP_INFO=FALSE,update.top_snps=FALSE,gen.data='binary') {
```

```
> GWAS_dots<-amm_gwas(Y_dots,X_dots,K_dots,n=3)
GWAS performed on 48 ecotypes,  0 values excluded
SNP_INFO file created
pseudo-heritability estimate is  0.4381761
>
```

```
> source('plots_gwas.r')
> plot_gwas(GWAS_dots,lower.limit=0.1)
>
```

Manhattan plot



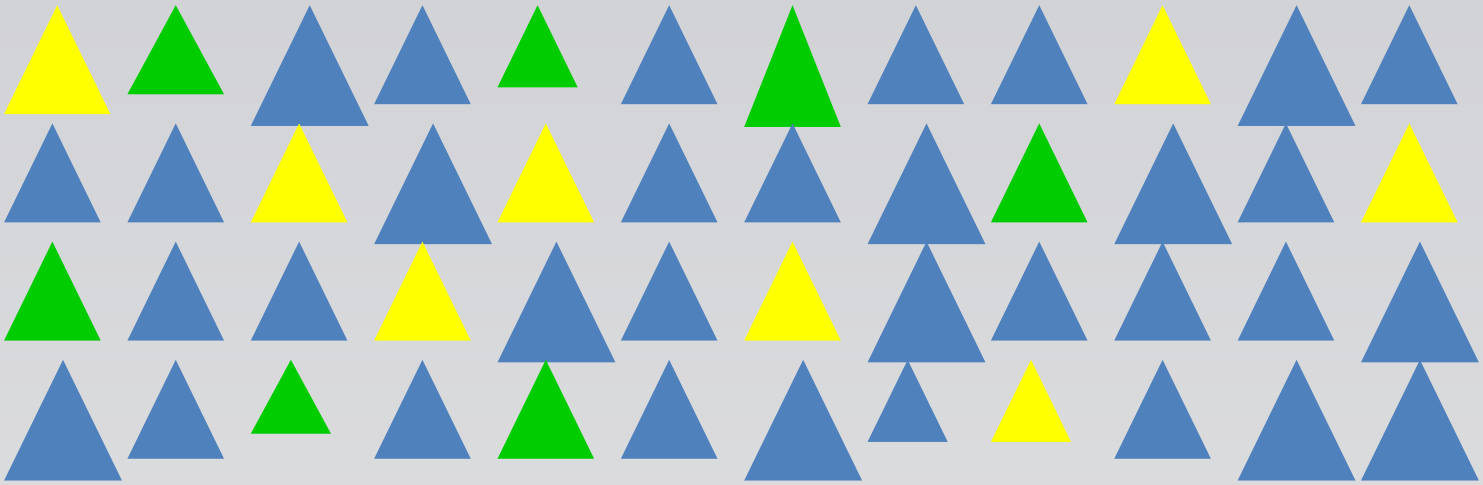
The gwas.r script

Center for Computational and Theoretical Biology

ideas.txt | grc_2015.txt | tals_Tübingen.txt | find_clust.r | mtmm_test.r | gwas.r | AT4G04740_t.txt | plots_

```
1  ##############################################################################
2  ###   AMM   -- R script for GWAS corecting for population structure (simila
3  ###
4  #######
5  #
6  ##
7  ##
8  ##
9  #
10
11
12 ##REQUIRED DATA & FORMAT
13
14 #requires functions from the original emma function (Kang et al. 2008, Gene
15 #source('emma.r')
16 #PHENOTYPE - Y: a n by 1 matrix, where n=number of individuals and the rown
17
18 #GENOTYPE - X: a n by m matrix, where n=number of individuals, m=number of
19 #KINSHIP - K: a n by n matrix, with rownames(K)=colnames(K)=individual name
20 #each of these data being sorted in the same way, according to the individu
21 #
22 #
23 #SNP INFORMATION - SNP_INFO: a data frame having at least 3 columns:
24 # - 1 named 'SNP', with SNP names (same as colnames(X)),
25 # - 1 named 'Chr', with the chromosome number to which belong each SNP
26 # - 1 named 'Pos', with the position of the SNP onto the chromosome it belo
```

Practical part Ib

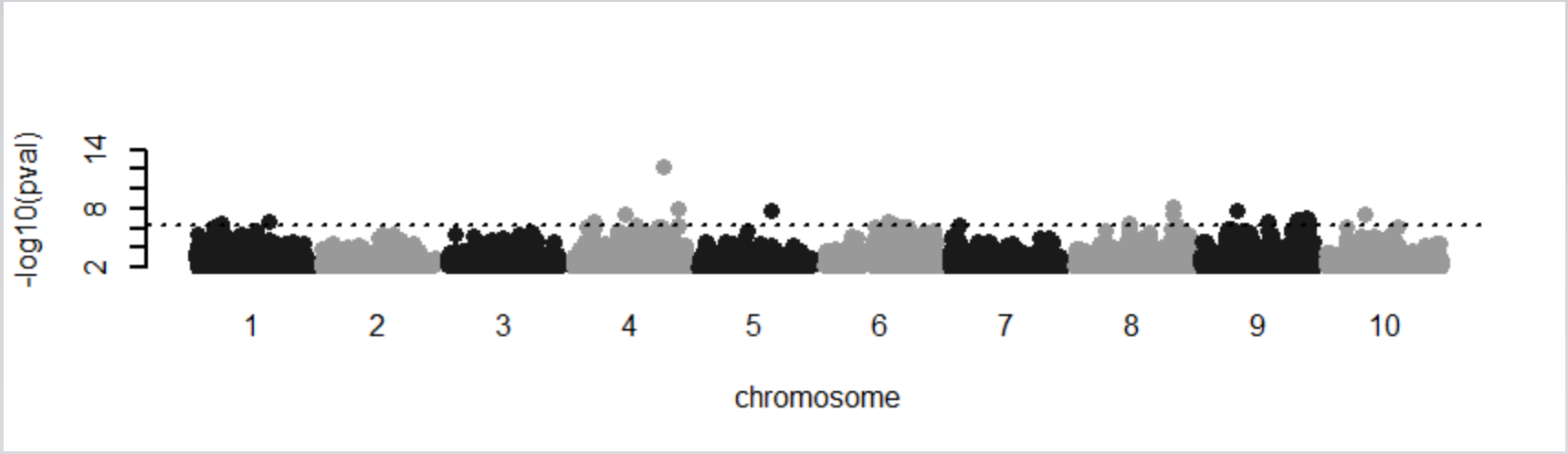


Your project :
Identify genetic factors that are important for the size of triangles.

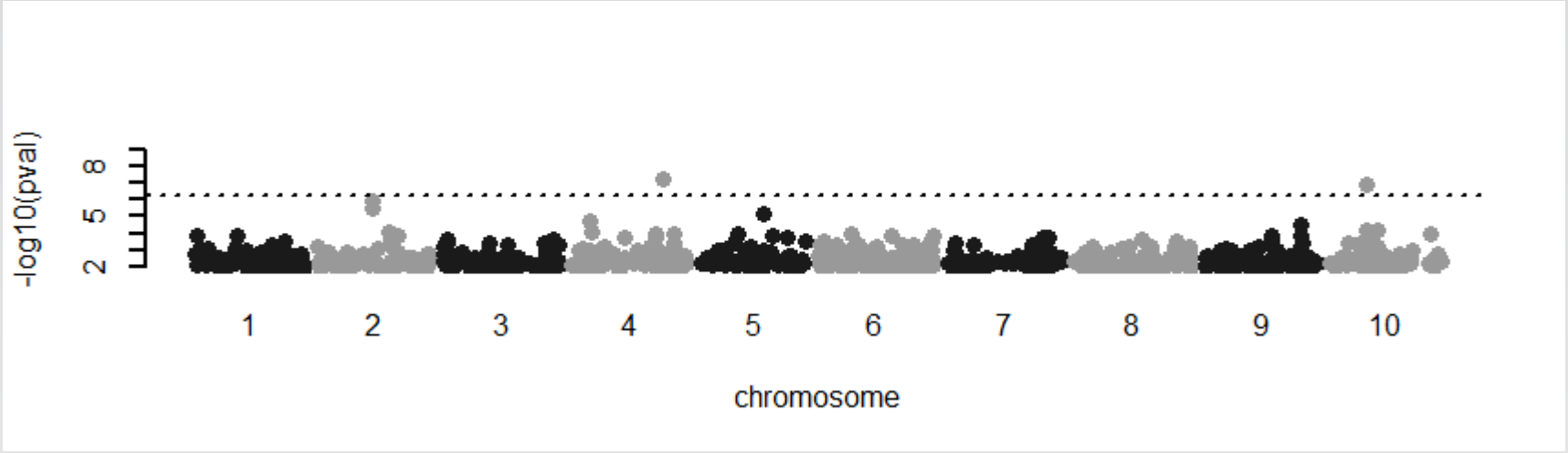
Dataset : 480 individuals with 2 phenotypes, and full sequencing data

Y_tri.rda, X_tri.rda, X_imp.rda,K_tri.rda,SNP_tri.rda

Manhattan plots triangles

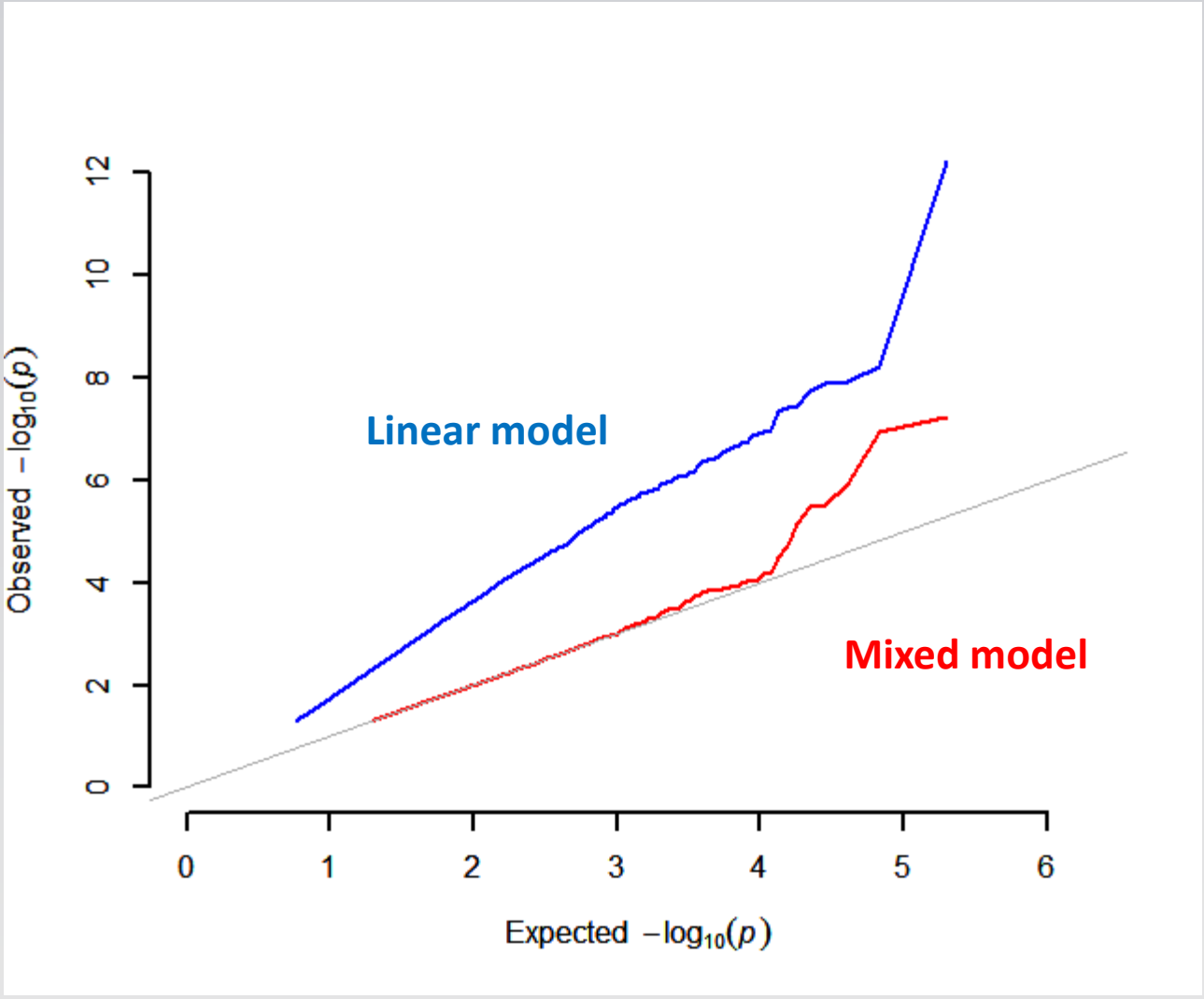


Linear model



Mixed model

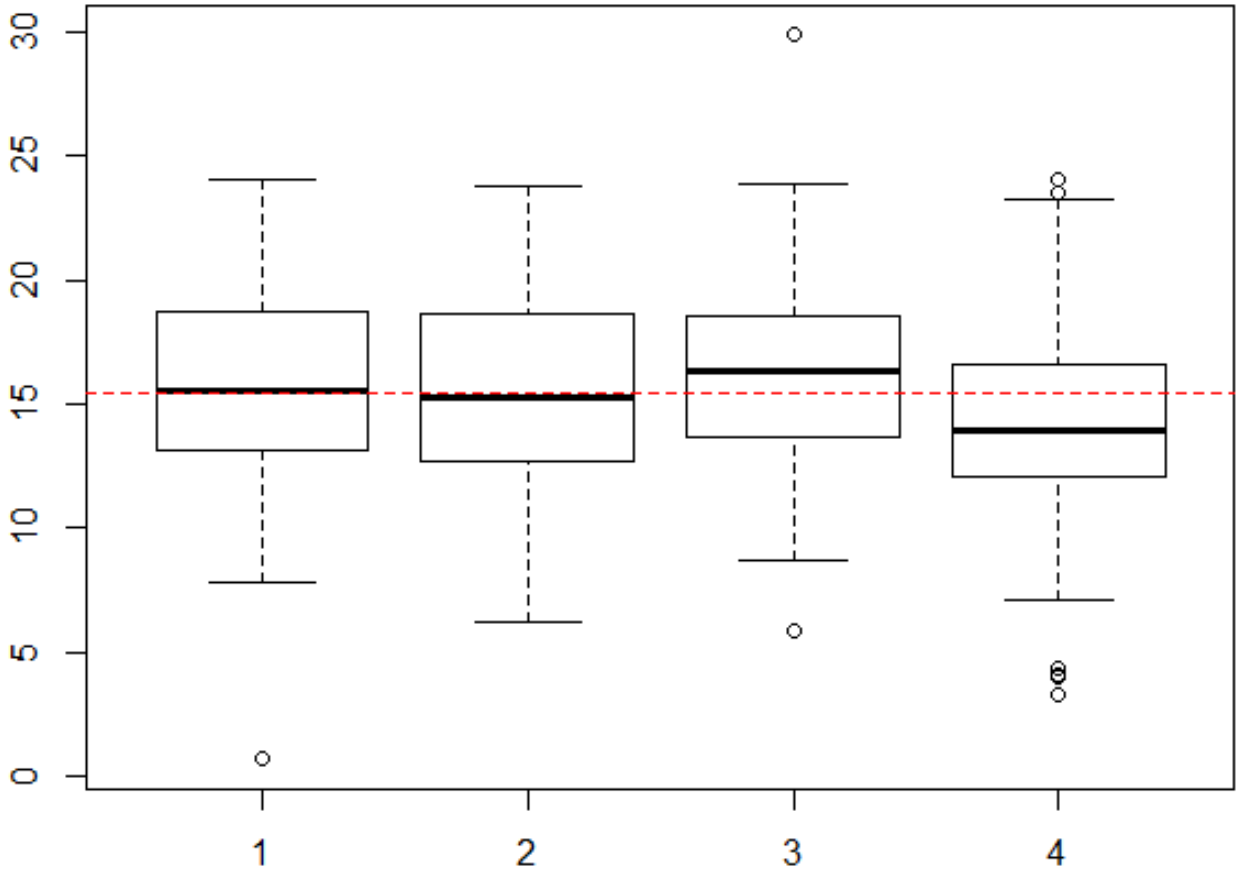
qq_plots triangles



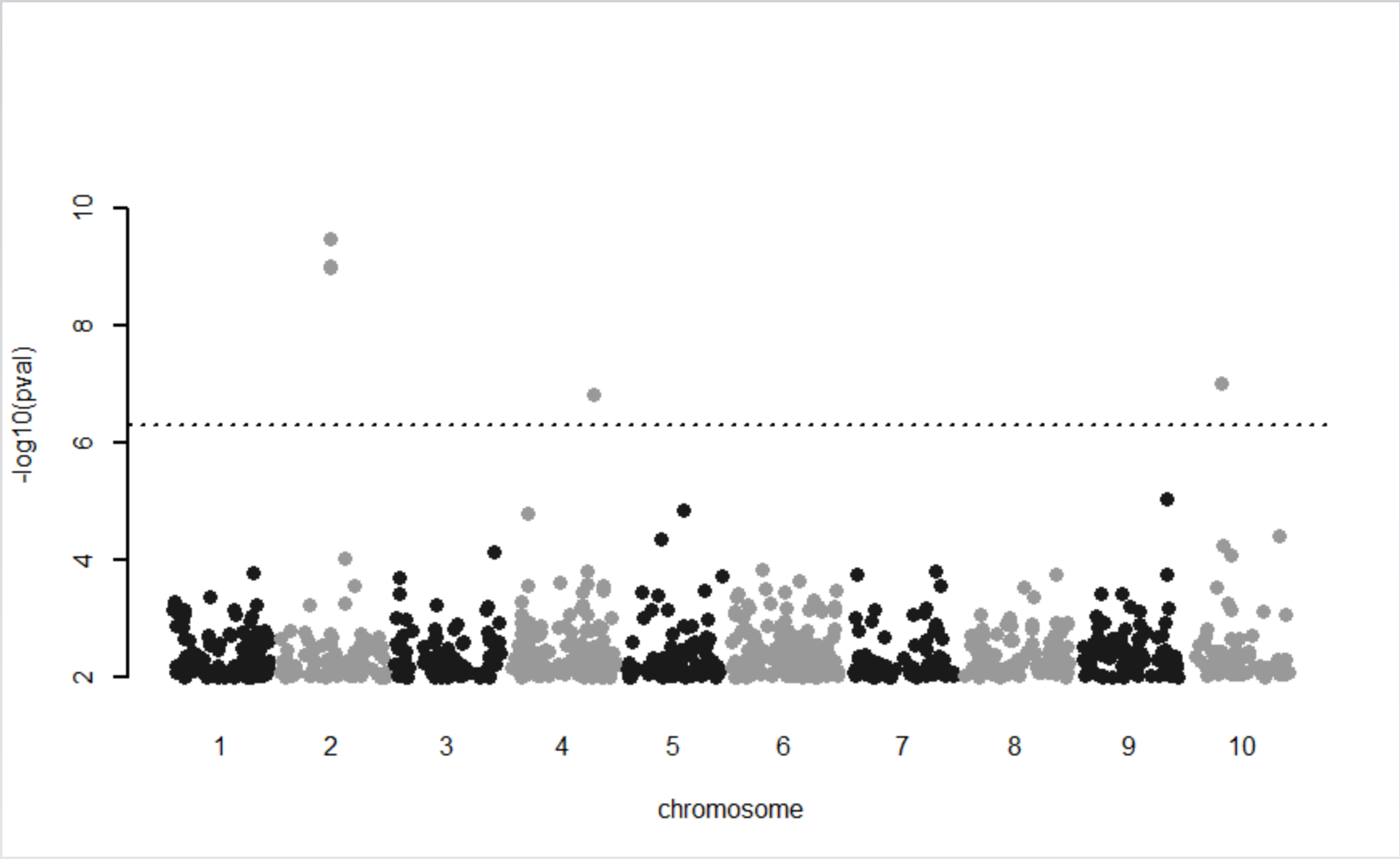
Batch effects


```
> summary(lm(size~batch,data=Y_tri))

Call:
lm(formula = size ~ batch, data = Y_tri)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2259  -2.6554  -0.1001   2.9074  14.5458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.1951     0.4466  36.264   <2e-16 ***
batch        -0.2909     0.1631  -1.784   0.0751 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.994 on 478 degrees of freedom
Multiple R-squared:  0.006613,  Adjusted R-squared:  0.004535
F-statistic: 3.182 on 1 and 478 DF,  p-value: 0.07508
```

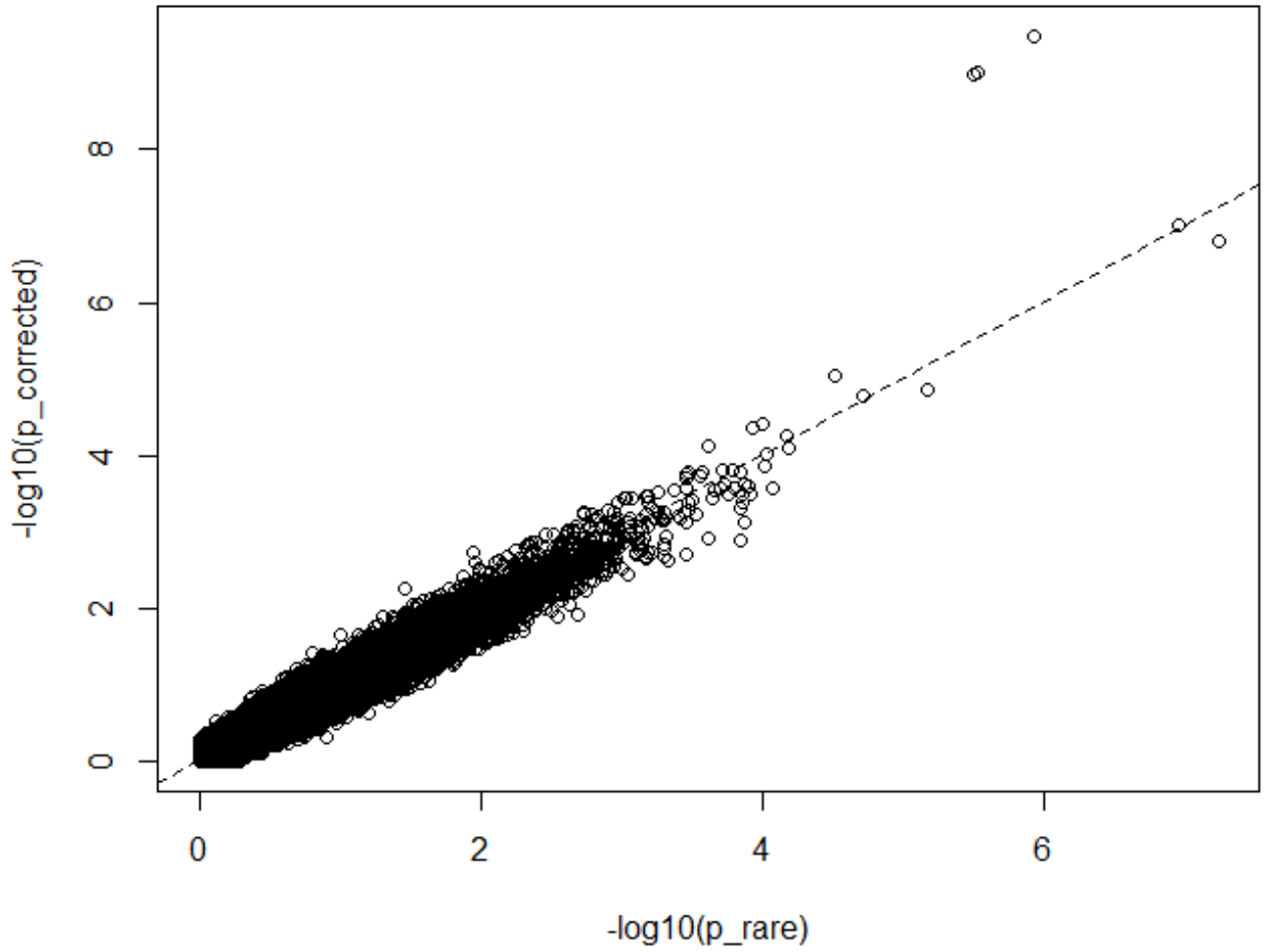
Batch effects



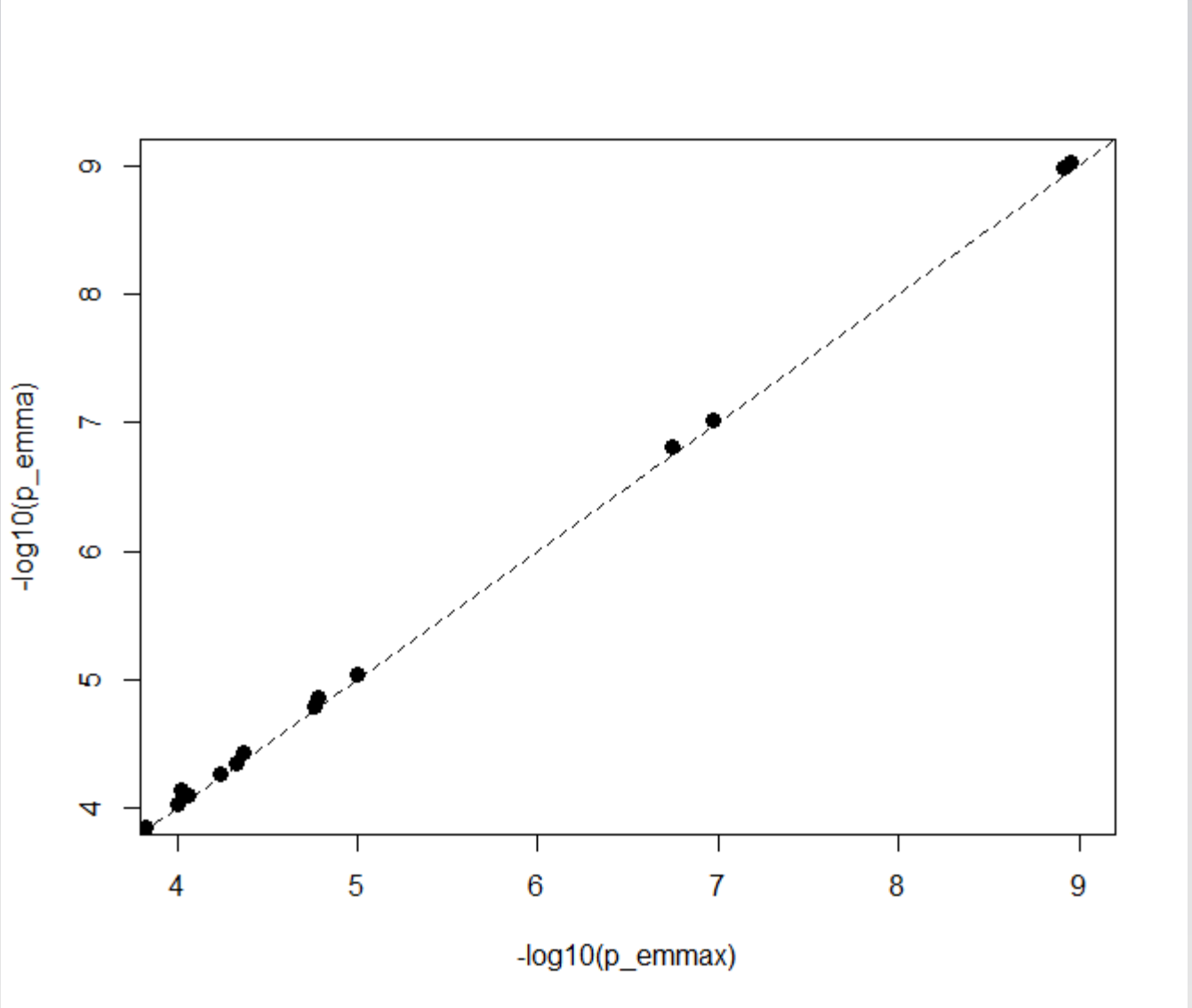
Batch effects



Batch effects



Emma vs. EmmaX



lunch

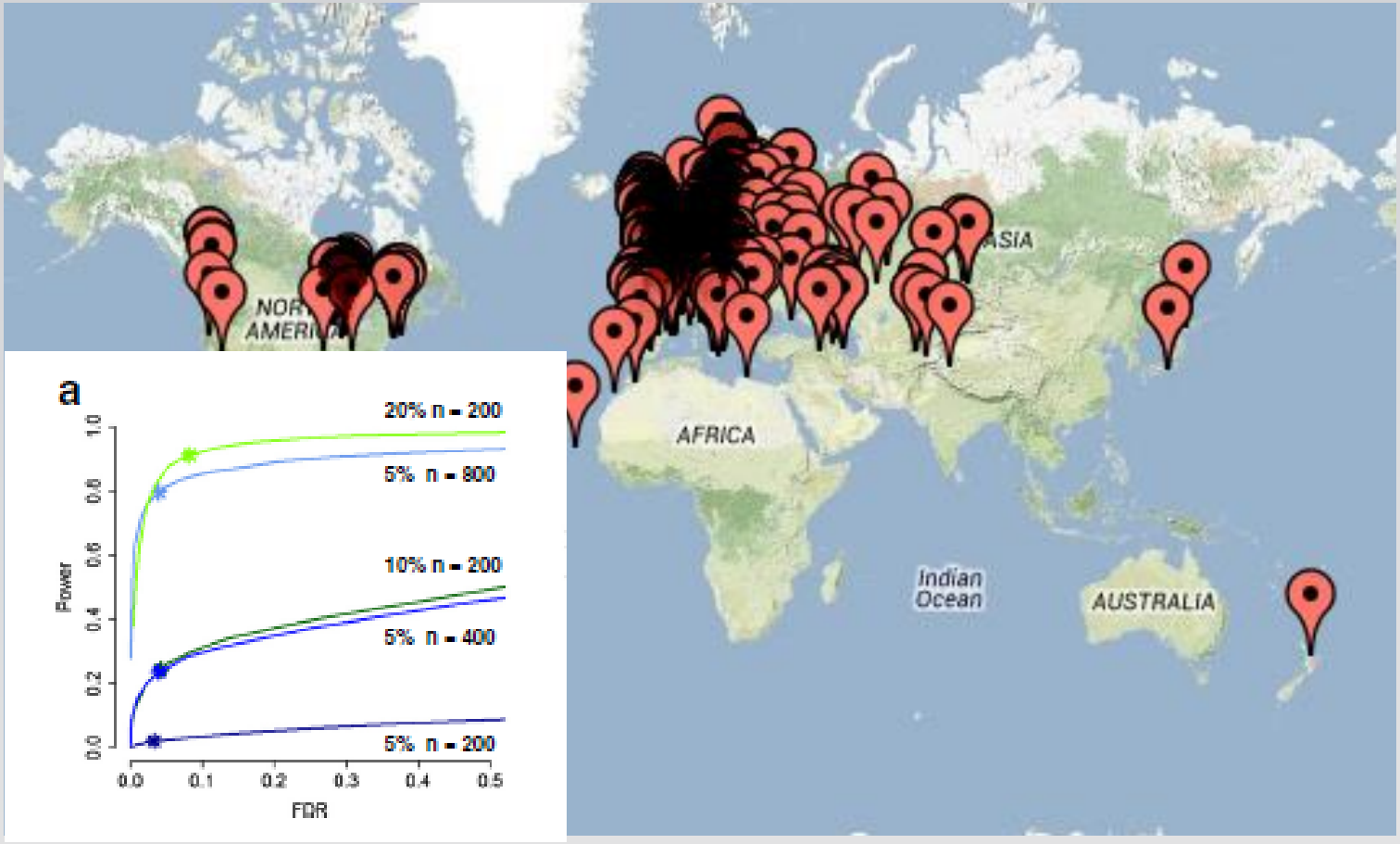
GWAS models / software / method development

Center for Computational and Theoretical Biology

Which lines should I phenotype?

- **> 1,000 different lines completely sequenced**
- **more than 10M SNPs (imputed for over 2,000 lines) and 500K structural variants**

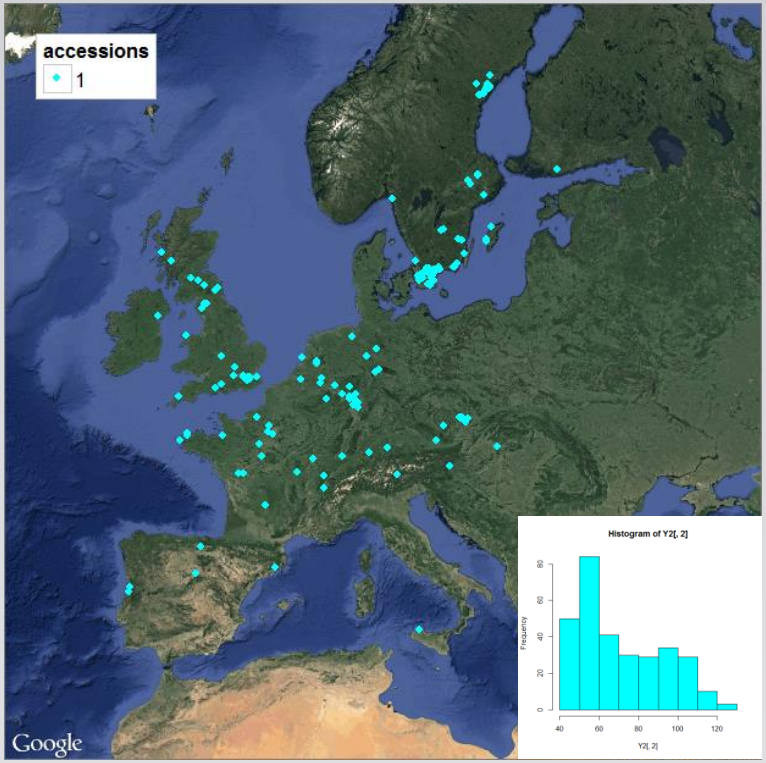
Which lines should I phenotype?

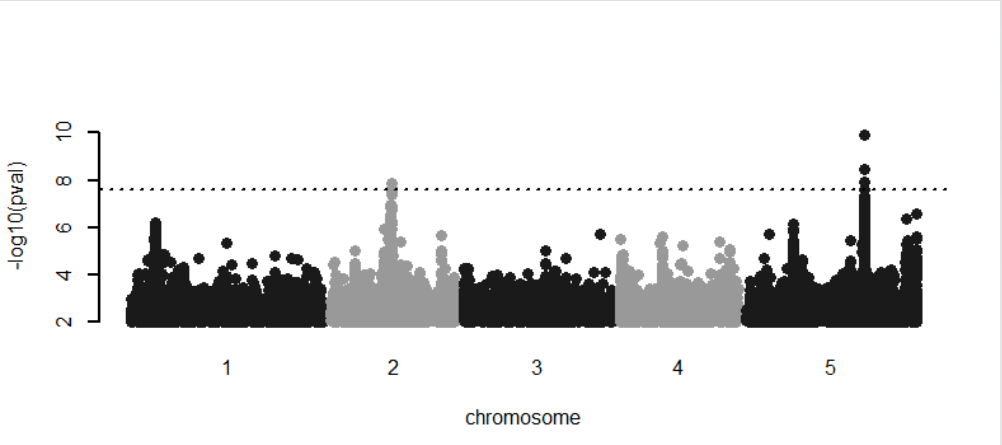


- **> 1,000 different lines completely sequenced**
- **more than 10M SNPs (imputed for over 2,000 lines) and 500K structural variants**

GWAS on flowering time (200 accessions)

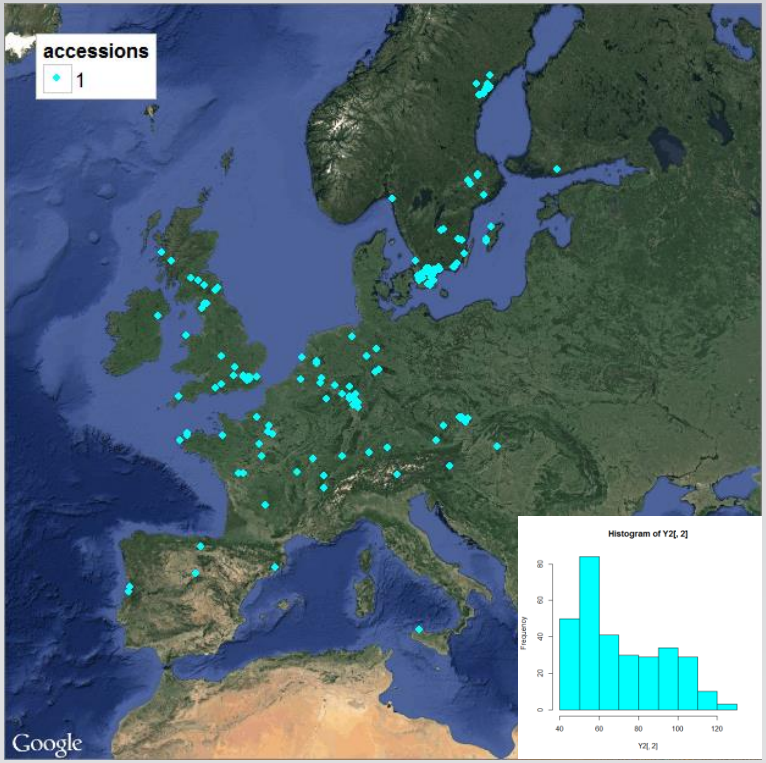
Center for Computational and Theoretical Biology

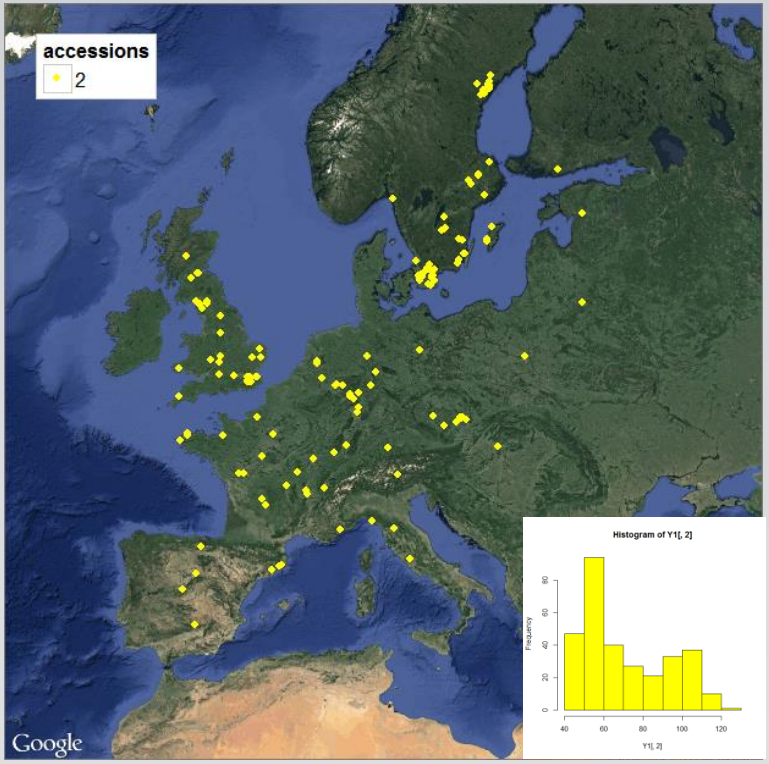


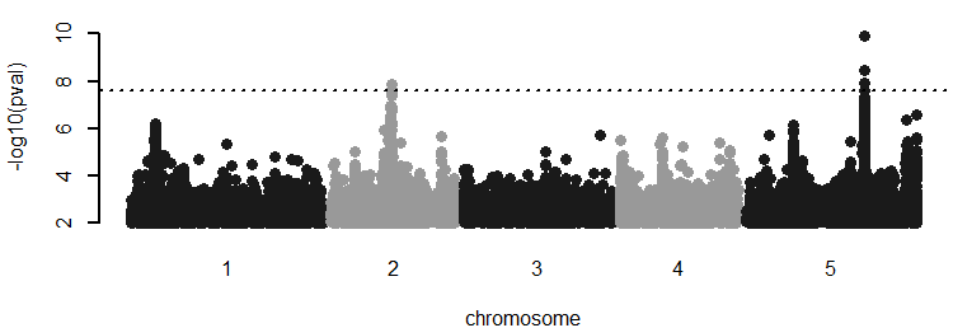


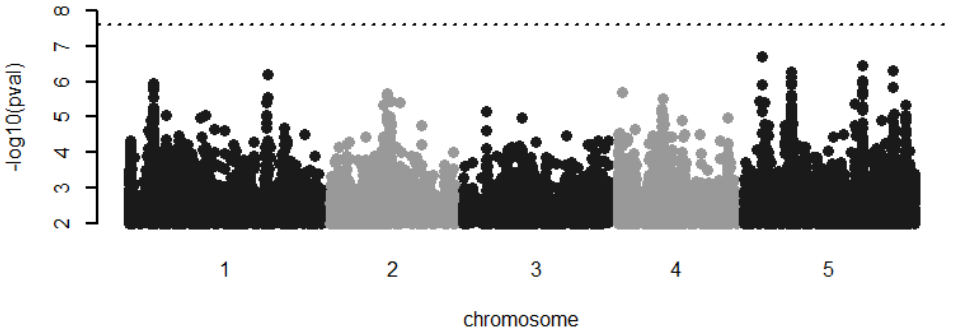
GWAS on flowering time (200 accessions)

Center for Computational and Theoretical Biology





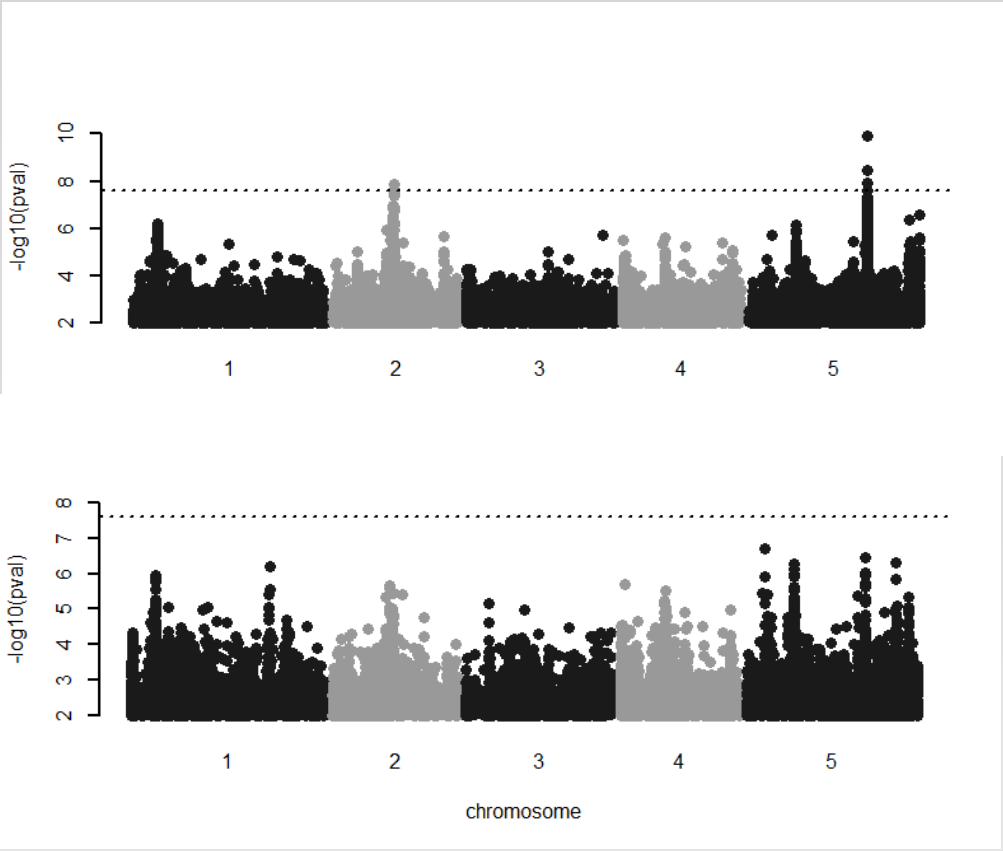




GWAS on flowering time (200 accessions)

Center for Computational and Theoretical Biology

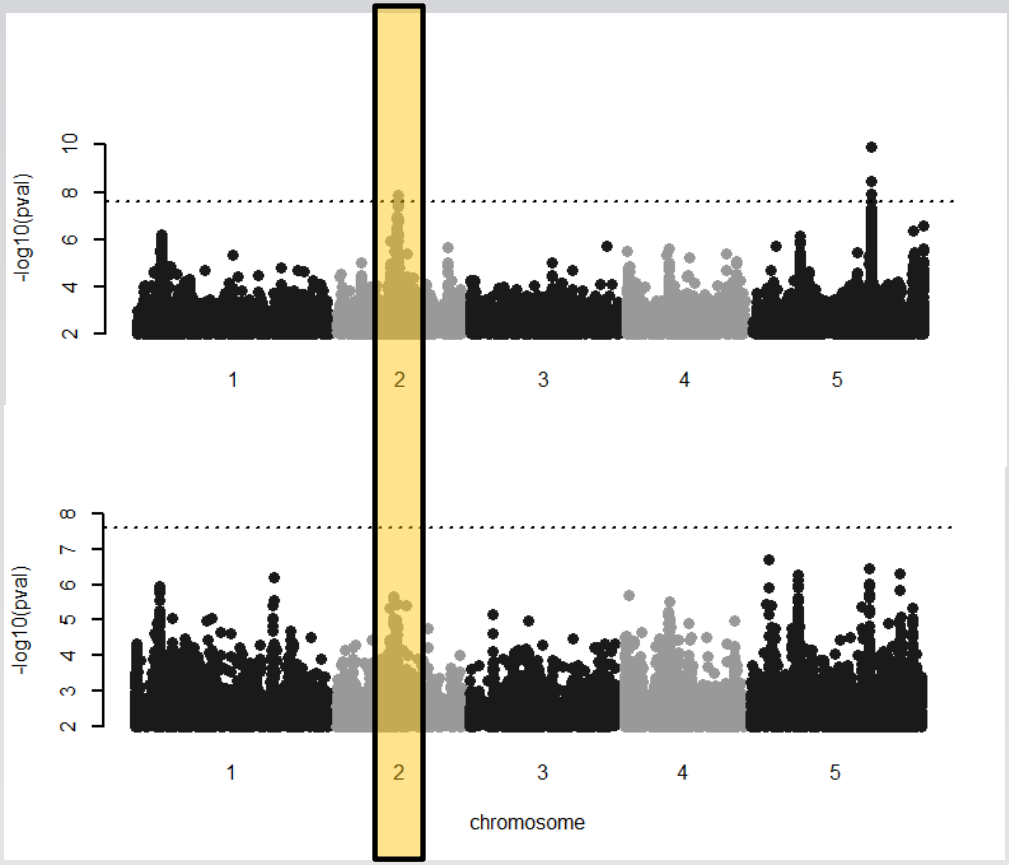
1000 differnet random subsets of 200 accessions

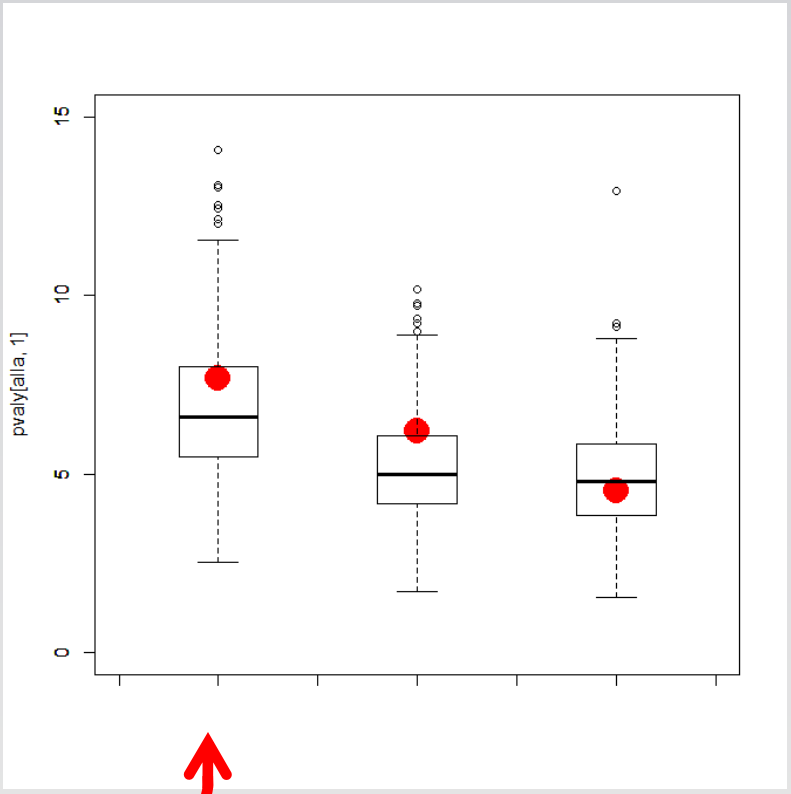


GWAS on flowering time (200 accessions)

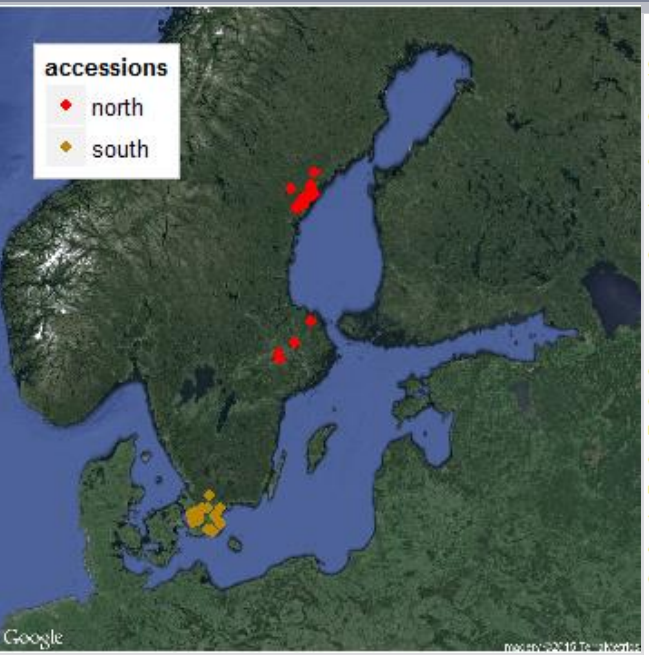
Center for Computational and Theoretical Biology

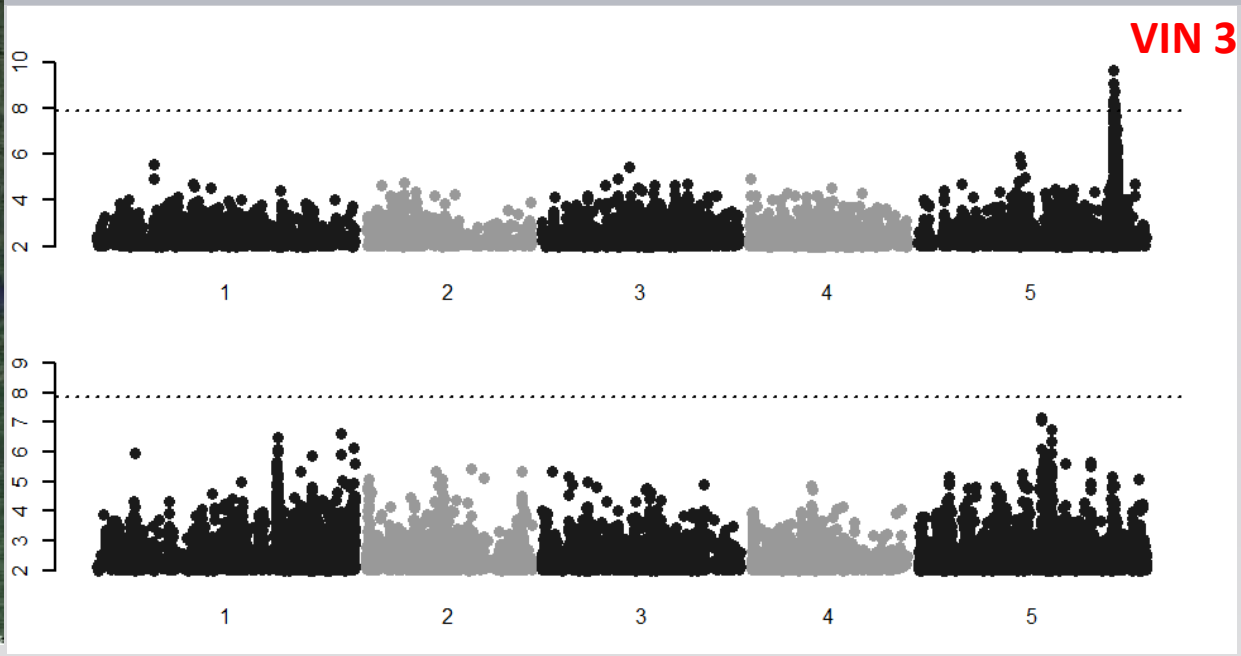
1000 differnet random subsets of 200 accessions





GWAS in local subsets

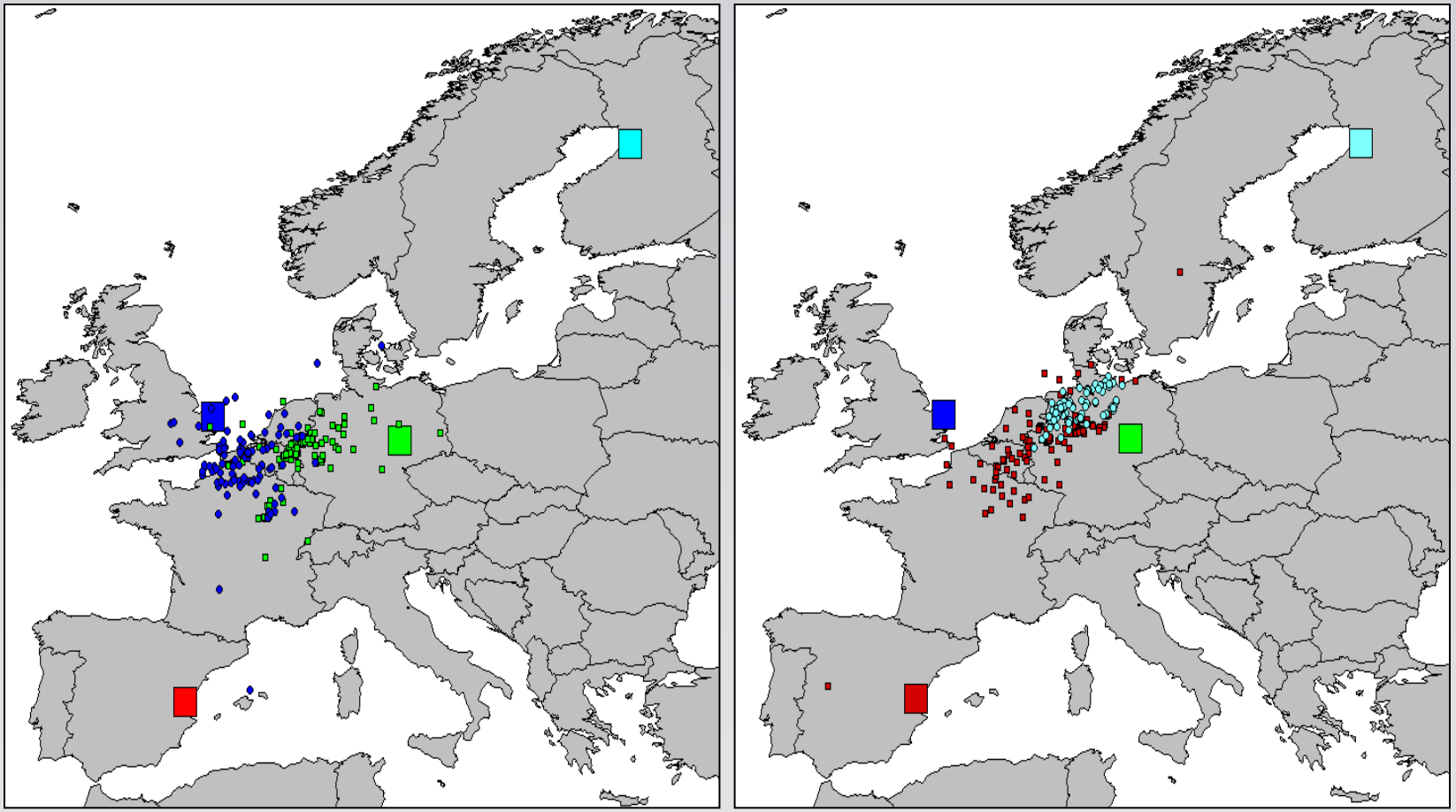




Different results in different subsets

Local Adaptation

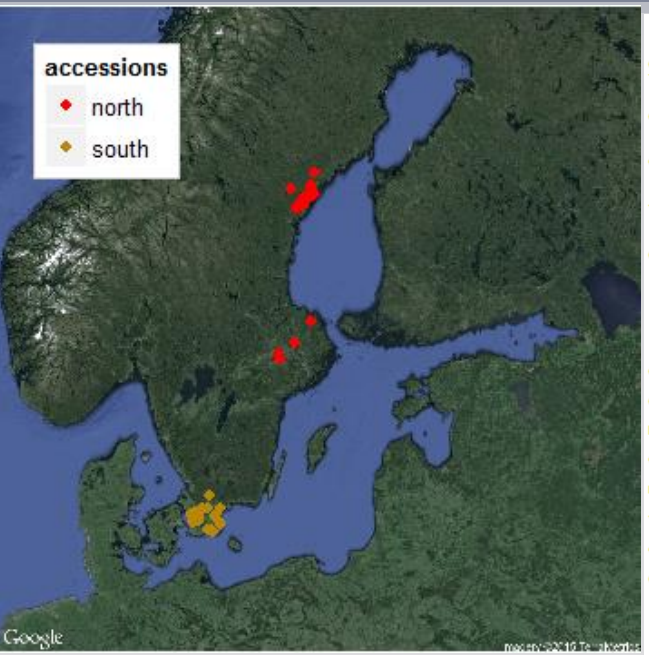
CTB Center for Computational and Theoretical Biology

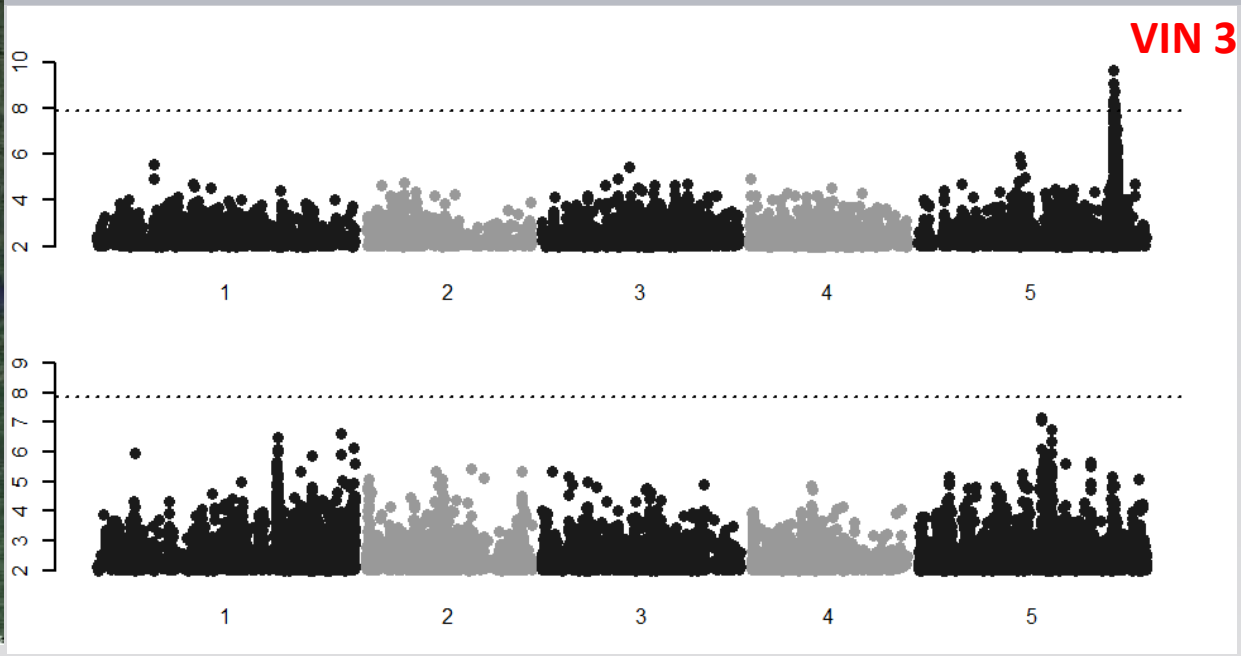


Genes providing fitness under natural conditions are mostly local

Fournier-Level *et al.* 2011, Science

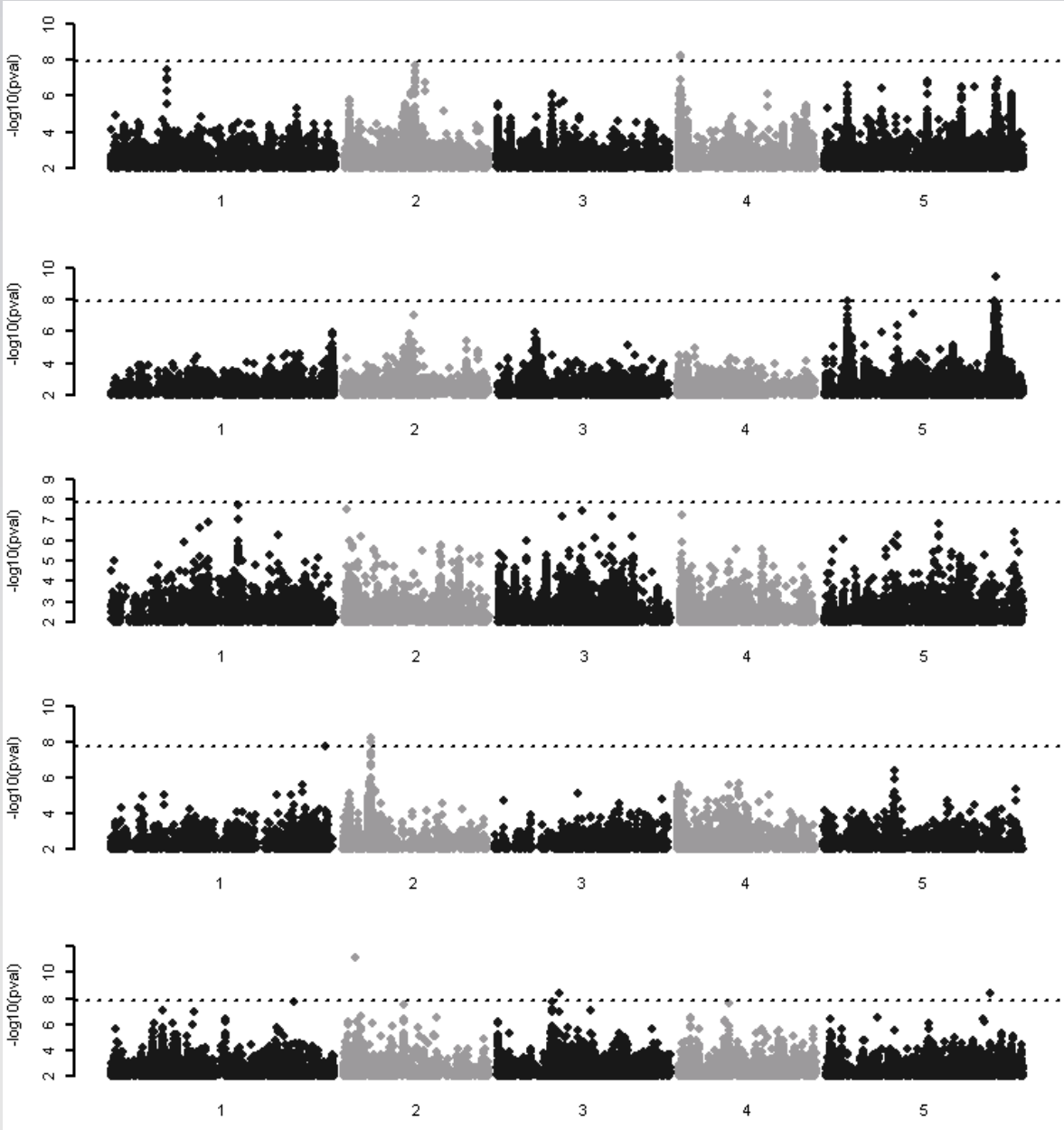
GWAS in local subsets





GWAS results in different subsets

Center for Computational and Theoretical Biology



All

SWE

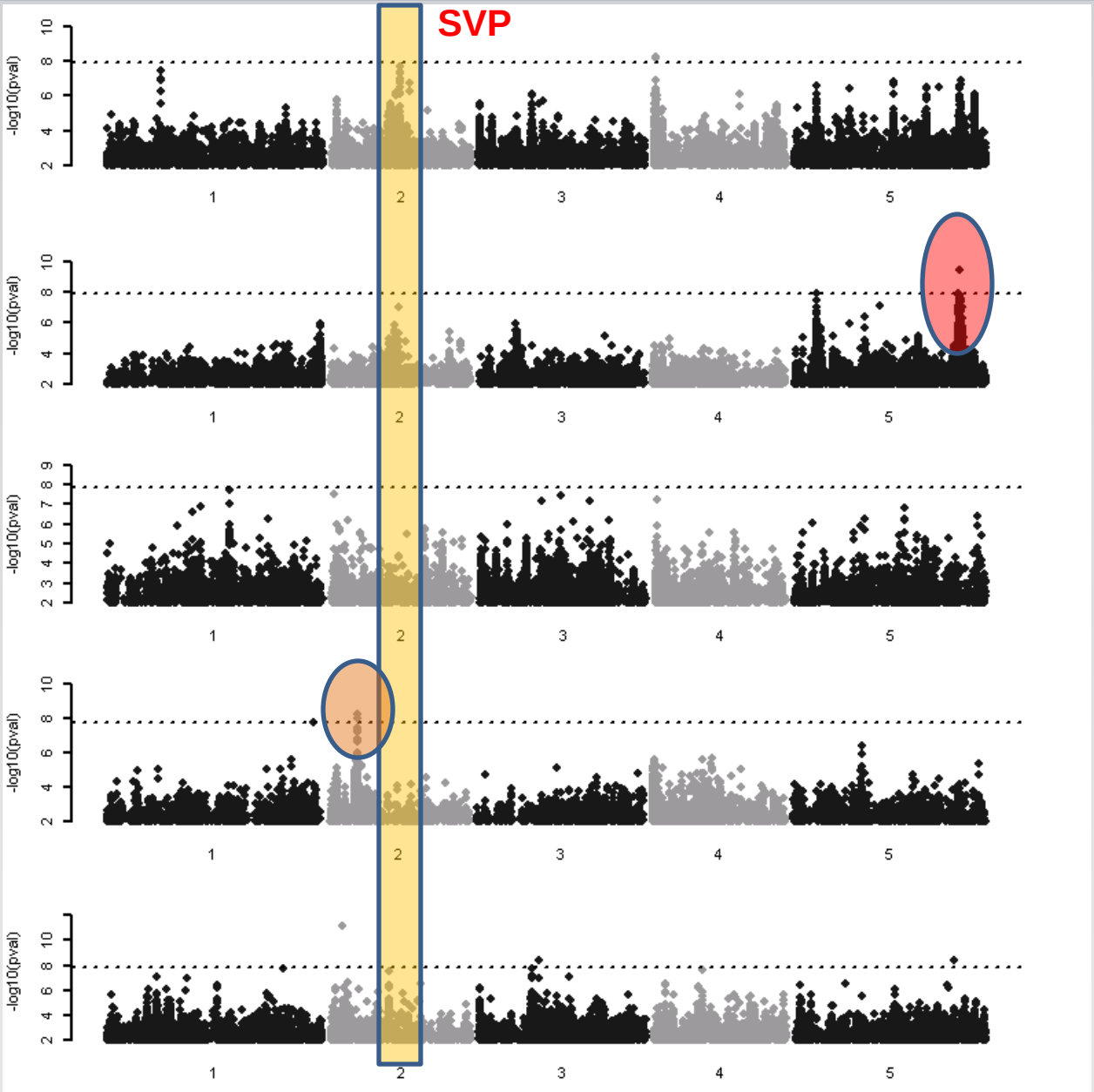
FRA

CZE

UK

GWAS results in different subsets

CTB Center for Computational and Theoretical Biology



All

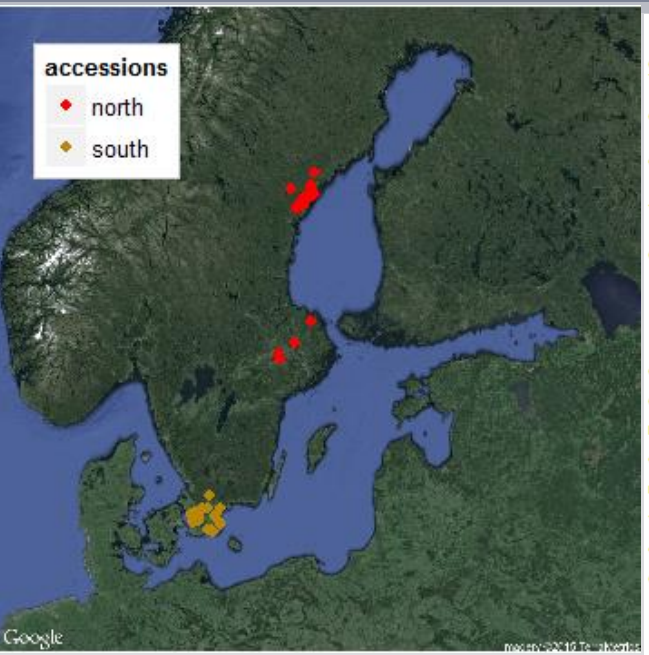
SWE

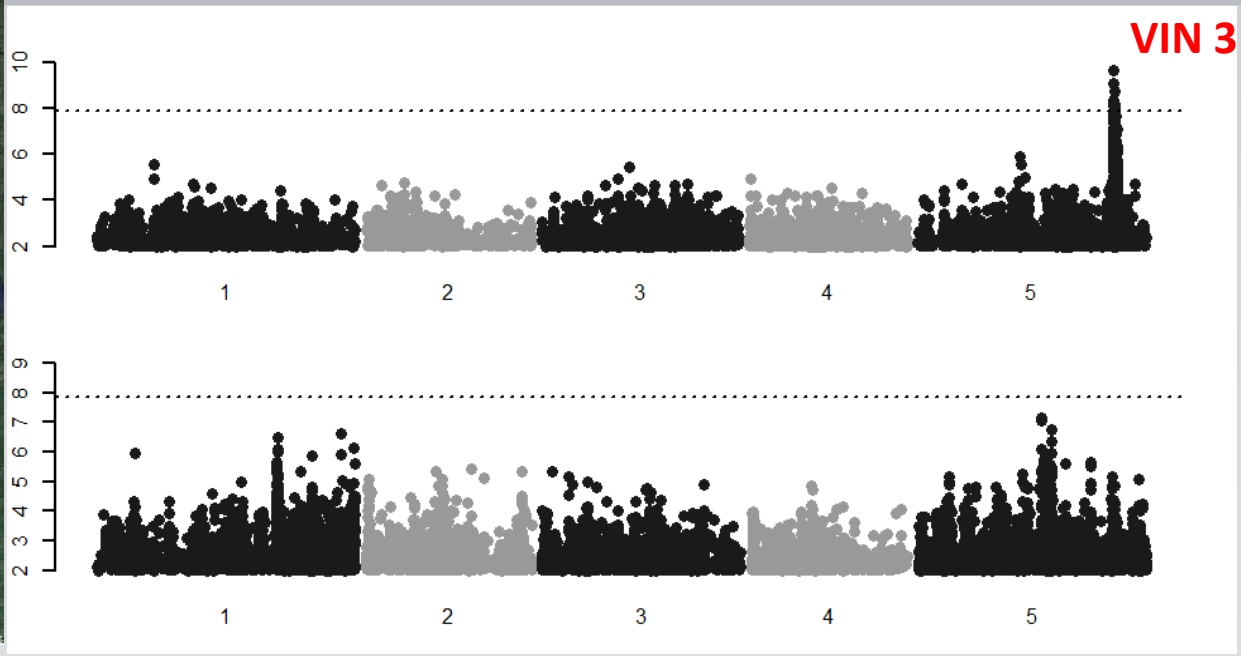
FRA

CZE

UK

GWAS in local subsets

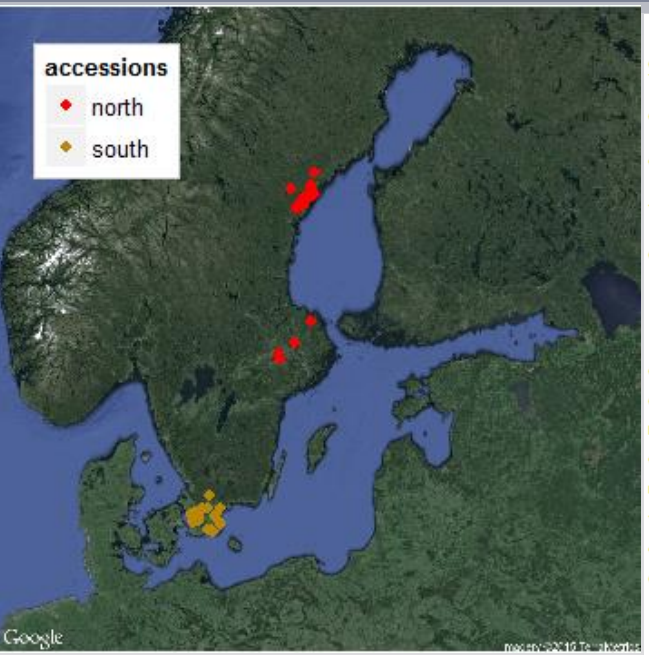


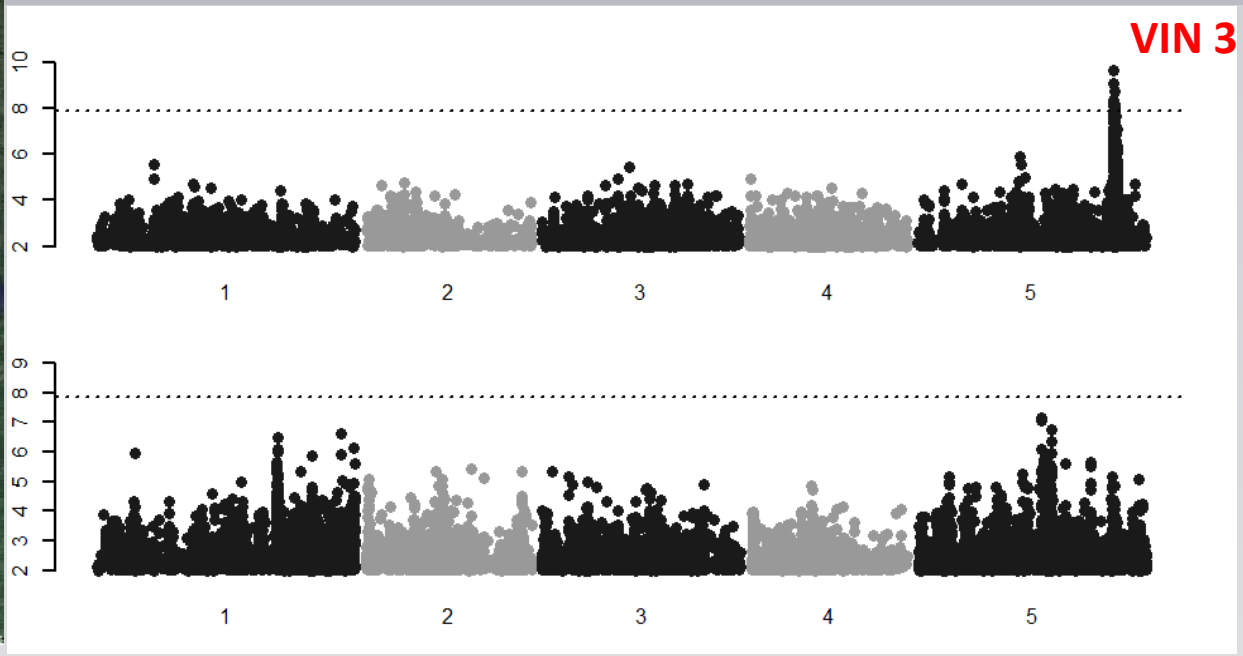


Why ?

- **Differences in Allele frequency of the causative marker**

GWAS in local subsets

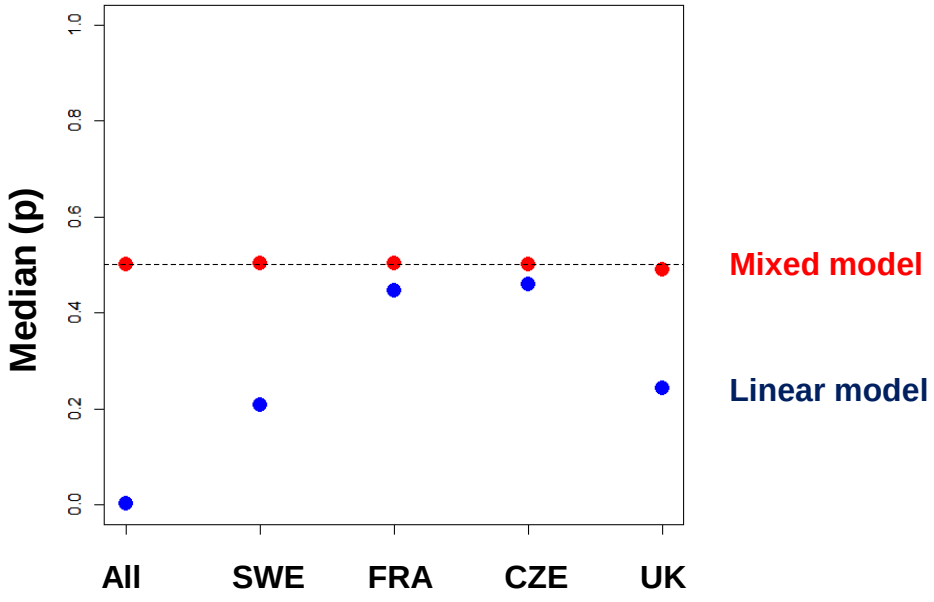




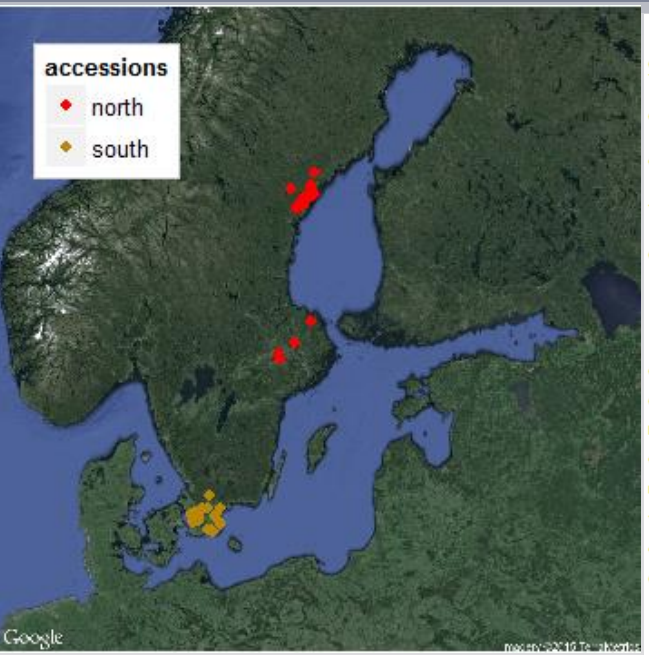
Why ?

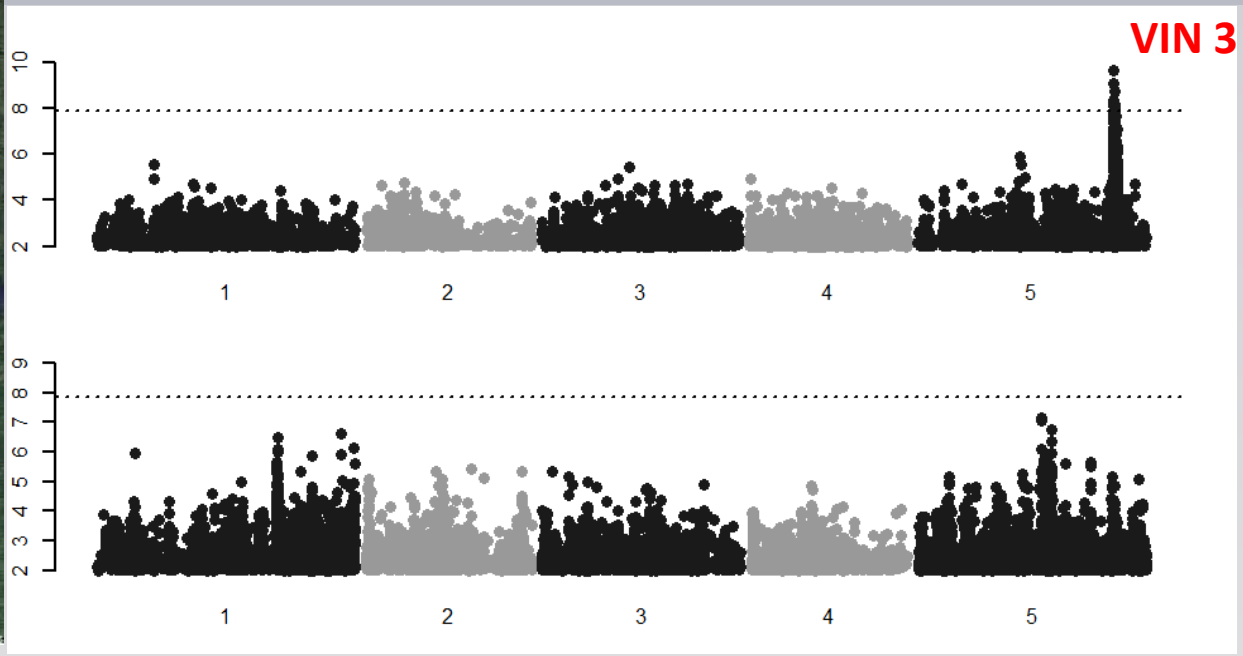
- **Differences in Allele frequency of the causative marker**
- **Artefact of the mixed model**

Population structure controll in subsets



GWAS in local subsets



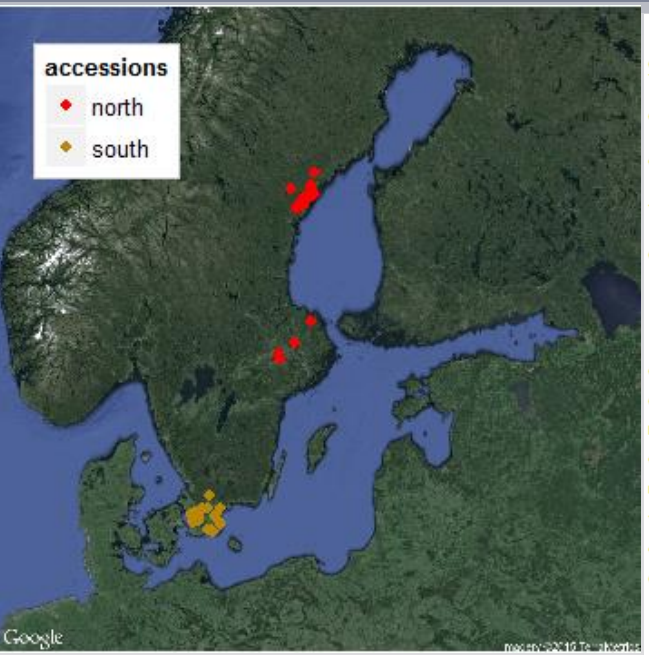


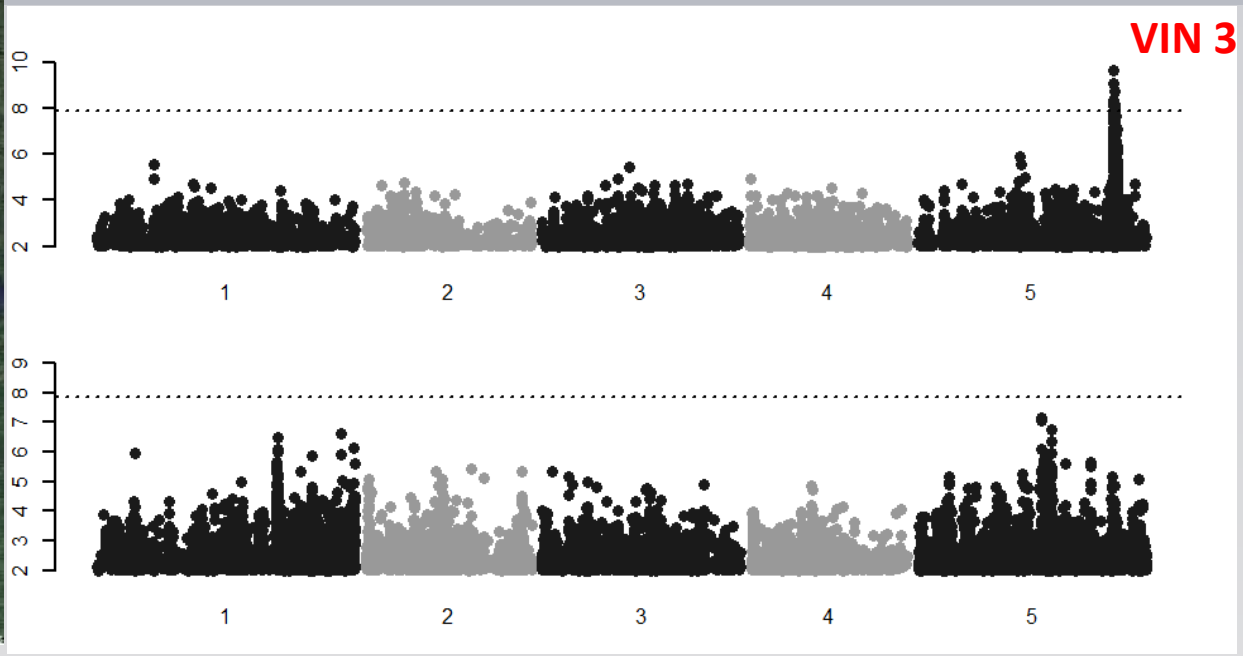
Why ?

- **Differences in Allele frequency of the causative marker**
- **Artefact of the mixed model**
- **The associated SNP is tagging different haplotypes**

GWAS in local subsets

CTB Center for Computational and Theoretical Biology





Why ?

- **Differences in Allele frequency of the causative marker**
- **Artefact of the mixed model**
- **The associated SNP is tagging different haplotypes**
- **Epistasis**

What about Epistasis ?

Center for Computational and Theoretical Biology

We cannot test genome-wide SNP by SNP Epistasis !

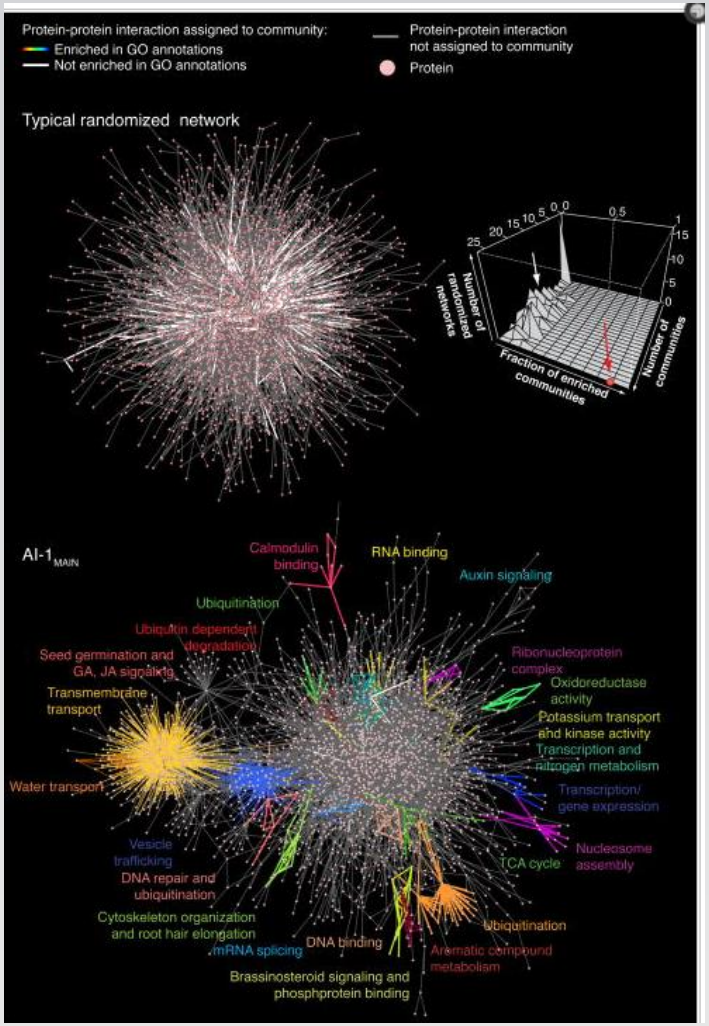
Evidence for Network Evolution in an *Arabidopsis* Interactome Map

Arabidopsis Interactome Mapping Consortium*†

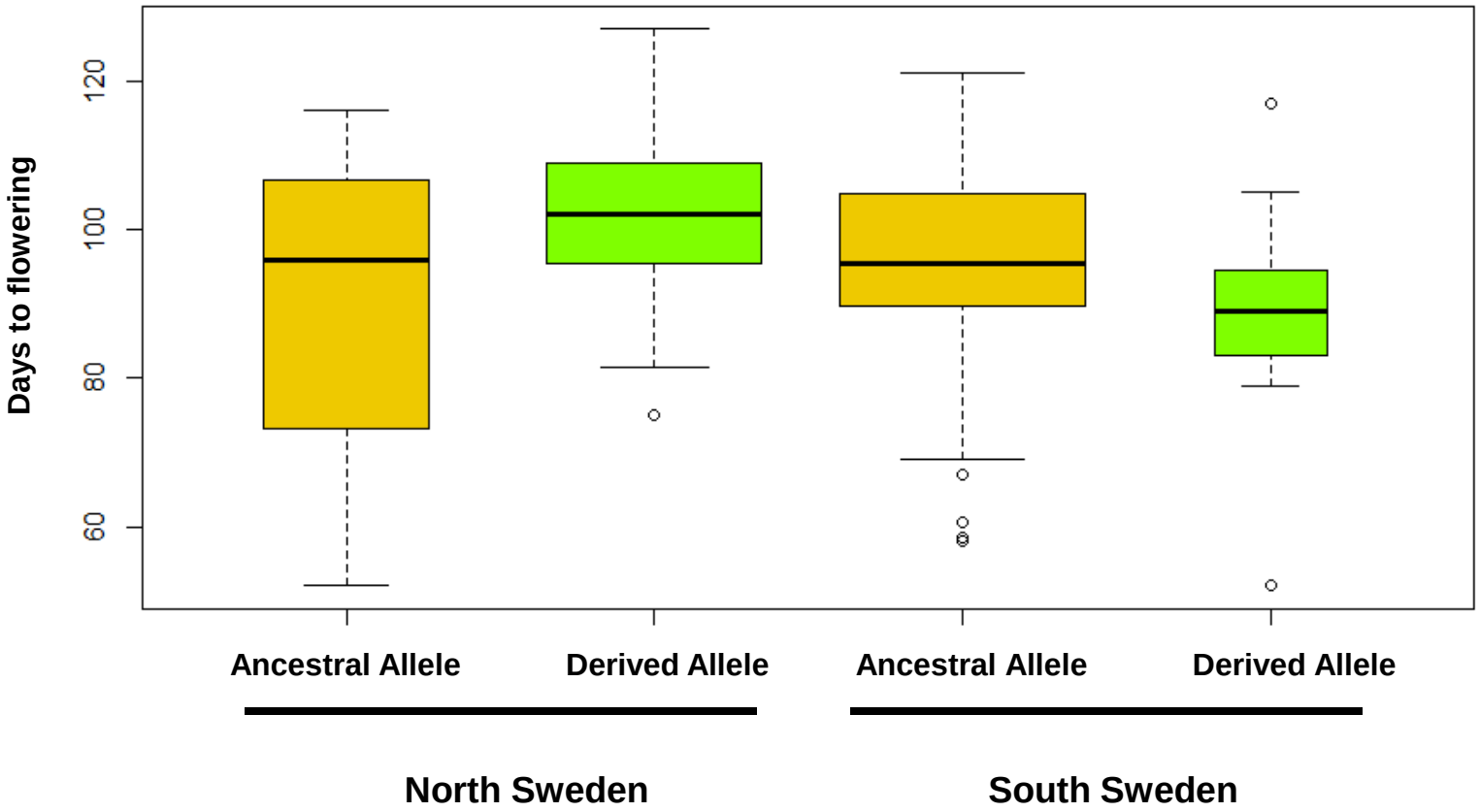
Plants have unique features that evolved in response to their environments and ecosystems. A full account of the complex cellular networks that underlie plant-specific functions is still missing. We describe a proteome-wide binary protein-protein interaction map for the interactome network of the plant *Arabidopsis thaliana* containing about 6200 highly reliable interactions between about 2700 proteins. A global organization of plant biological processes emerges from community analyses of the resulting network, together with large numbers of novel hypothetical functional links between proteins and pathways. We observe a dynamic rewiring of interactions following gene duplication events, providing evidence for a model of evolution acting upon interactome networks. This and future plant interactome maps should facilitate systems approaches to better understand plant biology and improve crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a variety of plant systems. Yet, more than 60% of the protein-coding genes of the model plant *Arabidopsis thaliana* (hereafter *Arabidopsis*) remain functionally uncharacterized. Knowledge about the biological organization of macromolecules in complex and dynamic "interactome" networks is lacking for *Arabidopsis* (fig. S1 and tables S1 and S2), depriving us of an understanding of how genotype-to-phenotype relationships are mediated at the systems level (*1*).

*All authors with their affiliations and contributions are listed at the end of the paper.
†To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu; ecker@salk.edu; pascal_braun@dfci.harvard.edu; david_hill@dfci.harvard.edu

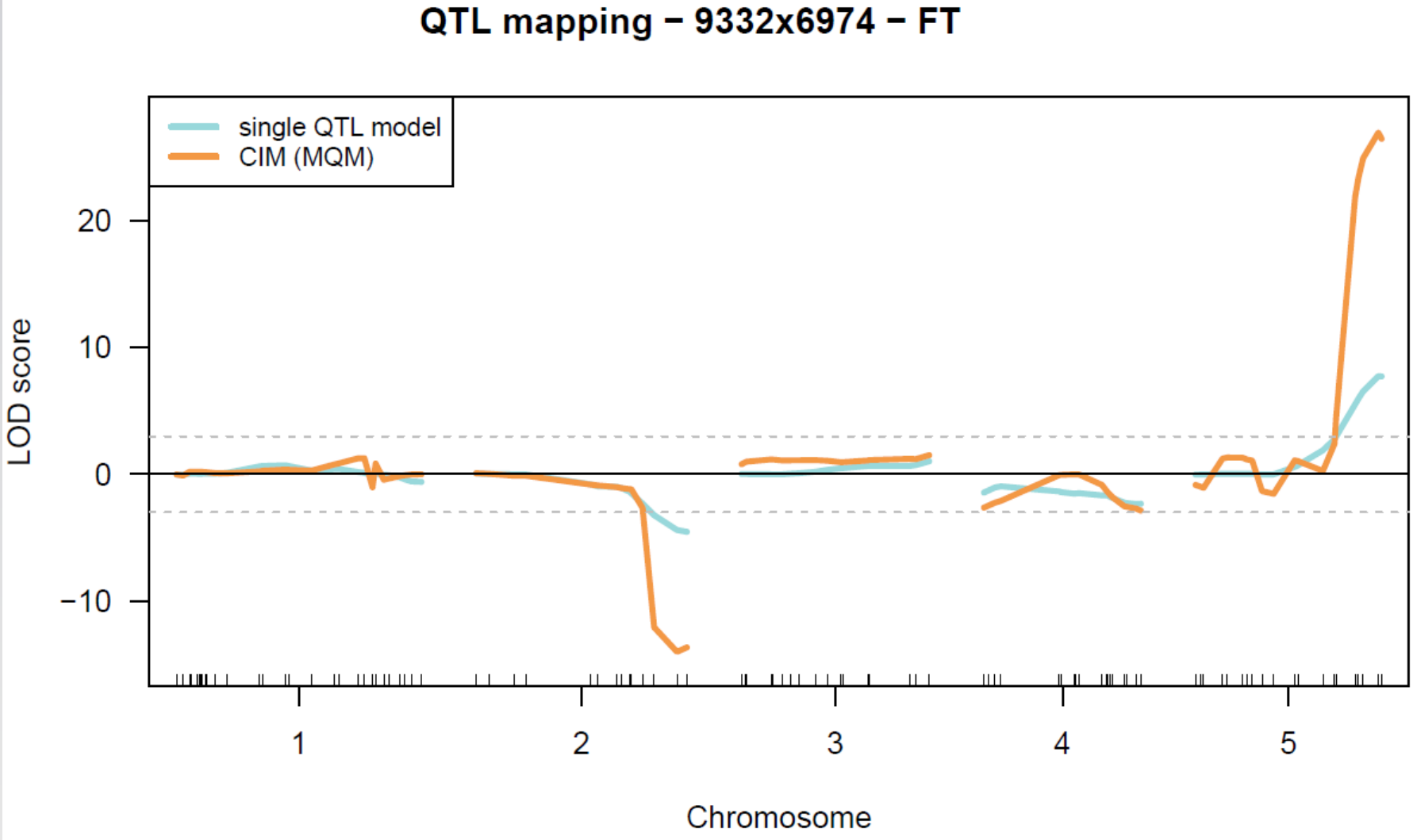


Effect of the *VIN3* Allele on Flowering time

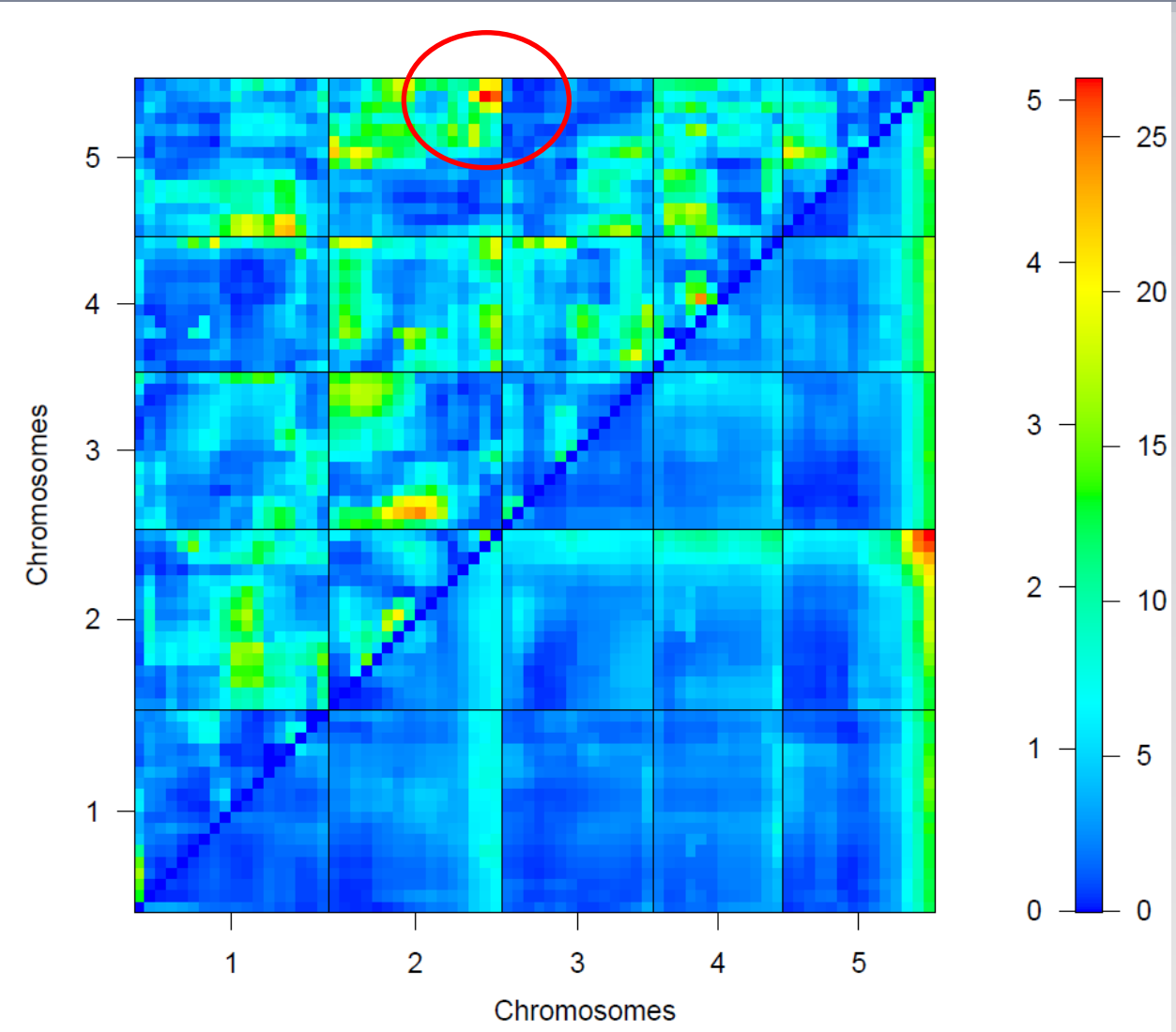


The effect of the respective allele depends on the genetic background

QTL mapping in crosses



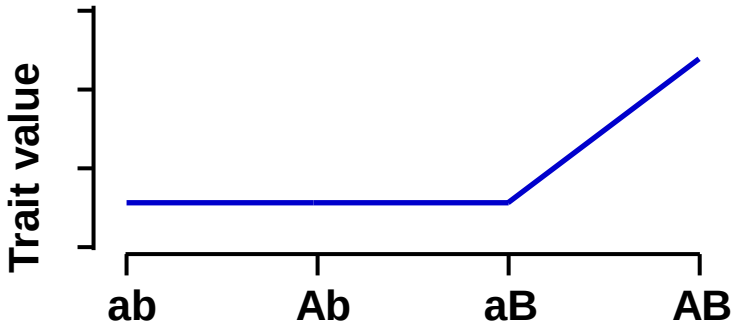
QTL mapping in crosses (2d plot)



Epistatic interaction in the QTL analysis

GWAS models to test for epistasis

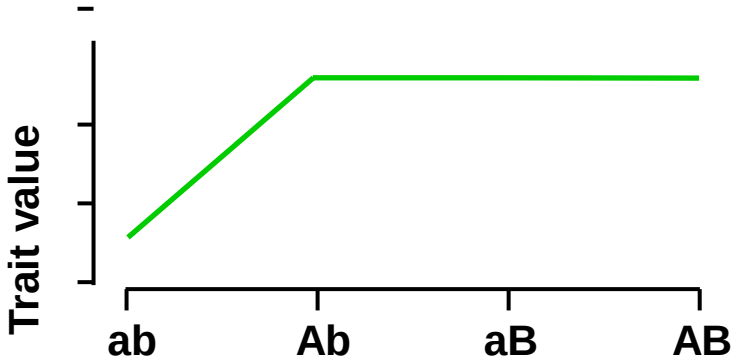
SNP by SNP epistasis

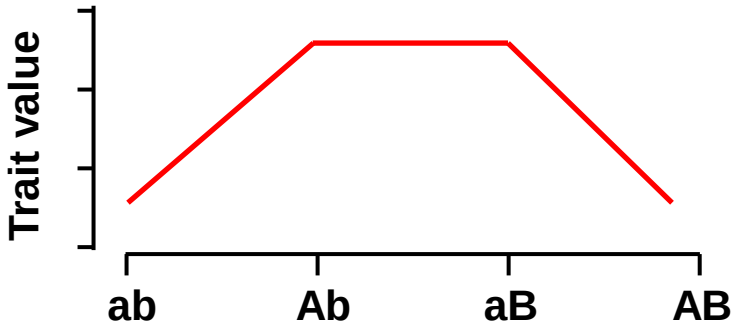


Genome-wide epistasis with the lead SNP in *VIN3*

3 different models with two homozygous SNPs

AND, OR and XOR

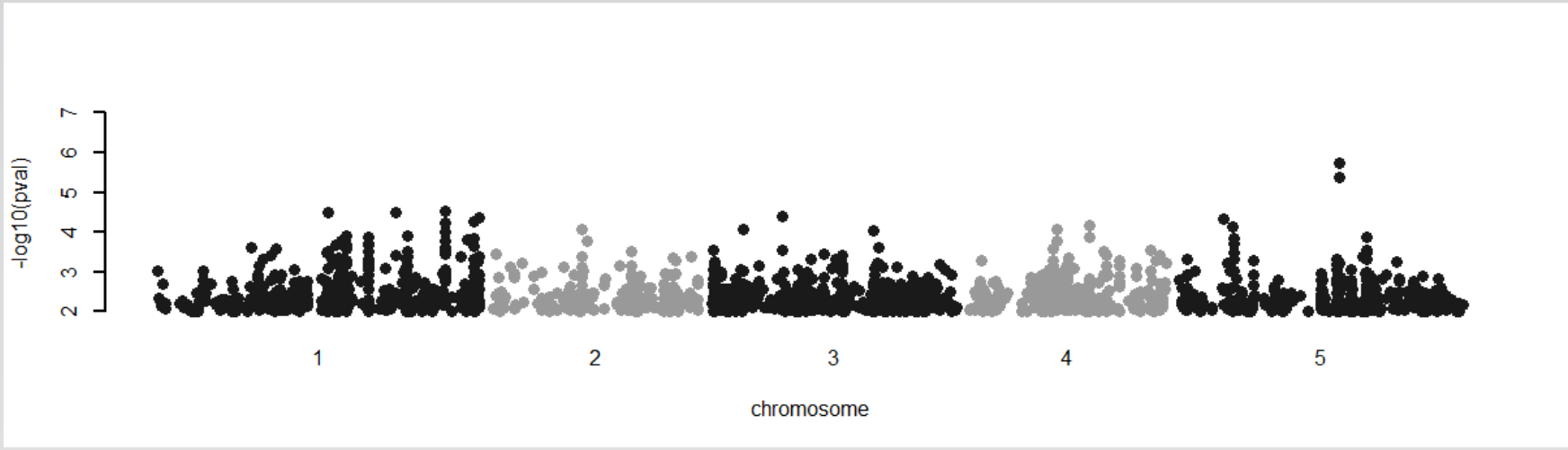




GWAS models to test for epistasis

Center for Computational and Theoretical Biology

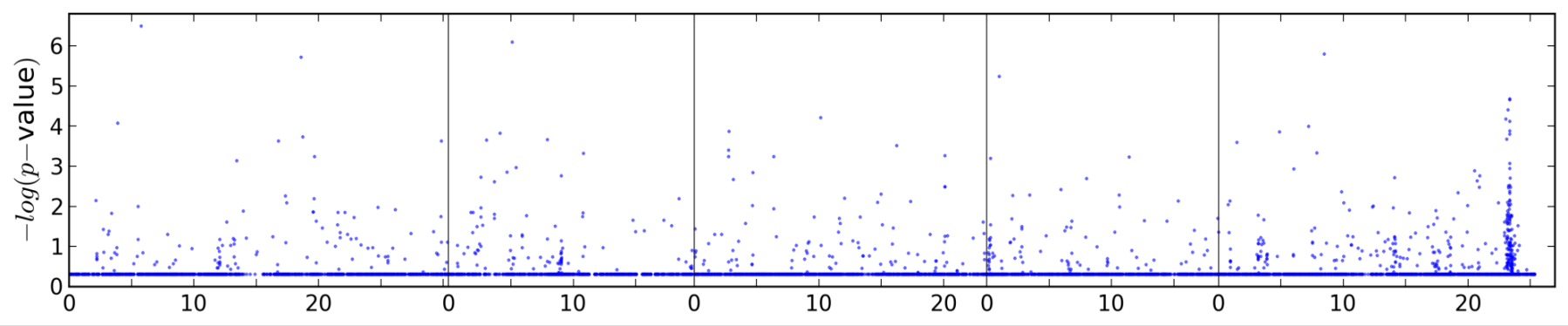
SNP by SNP epistasis



GWAS models to test for epistasis

Center for Computational and Theoretical Biology

SNP-by-genomic background epistasis



Different SNPs in *VIN3* seems to interact with the genomic background

Summary

There are different reason why associations don‘t replicate across different subsets of accessions :

- **False Positives**
- **No local varation of the causative Allele**
- **Effect depends on the genetic background (Epistasis)**
- **...**

American Journal of Medical Genetics 111:88–95 (2002)

Invited Paper

Genotype–Phenotype Correlation in Cystic Fibrosis: The Role of Modifier Genes

Francesco Salvatore,[1]* Olga Scudiero,[1] and Giuseppe Castaldo[1,2]

[1]*CEINGE Scarl and Dipartimento di Biochimica e Biotecnologie Mediche Università di Napoli "Federico II," Naples, Italy*

[2]*Facoltà di Scienze Matematiche, Fisiche e Naturali, Università del Molise, Isernia, Italy*

More than 1,000 mutations have been identified in the cystic fibrosis (CF) transmembrane regulator (CFTR) disease gene. The impact of these mutations on the protein and the wide spectrum of CF phenotypes prompted a series of Genotype–Phenotype correlation studies. The *CFTR* genotype is invariably correlated with pancreatic status—in about 85% of cases with pancreatic insufficiency and in about 15% of cases with pancreatic sufficiency. The correlations between the *CFTR* genotype and pulmonary, liver, and gastrointestinal expression are debatable. The heterogeneous phenotype in … mutations in α-1 antitrypsin (*A1AT*) and mannose binding lectin genes were found to be independent risk factors for liver disease in CF patients. The body of evidence available suggests that the variegated CF phenotype results from complex interactions between numerous gene products.

KEY WORDS: cystic fibrosis; modifier genes; Genotype–Phenotype correlation

TECHNICAL REPORTS

nature genetics

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte[1,4], Bjarni J Vilhjálmsson[1,2,4], Vincent S
Magnus Nordborg[1,2]

Genome-wide association studies (GWAS) are a standard approach for studying the genetics of natural variation. A ma concern in GWAS is the need to account for the complicate dependence structure of the data, both between loci as well between individuals. Mixed models have emerged as a gener and flexible approach for correcting for population structur

TECHNICAL REPORT

nature genetics

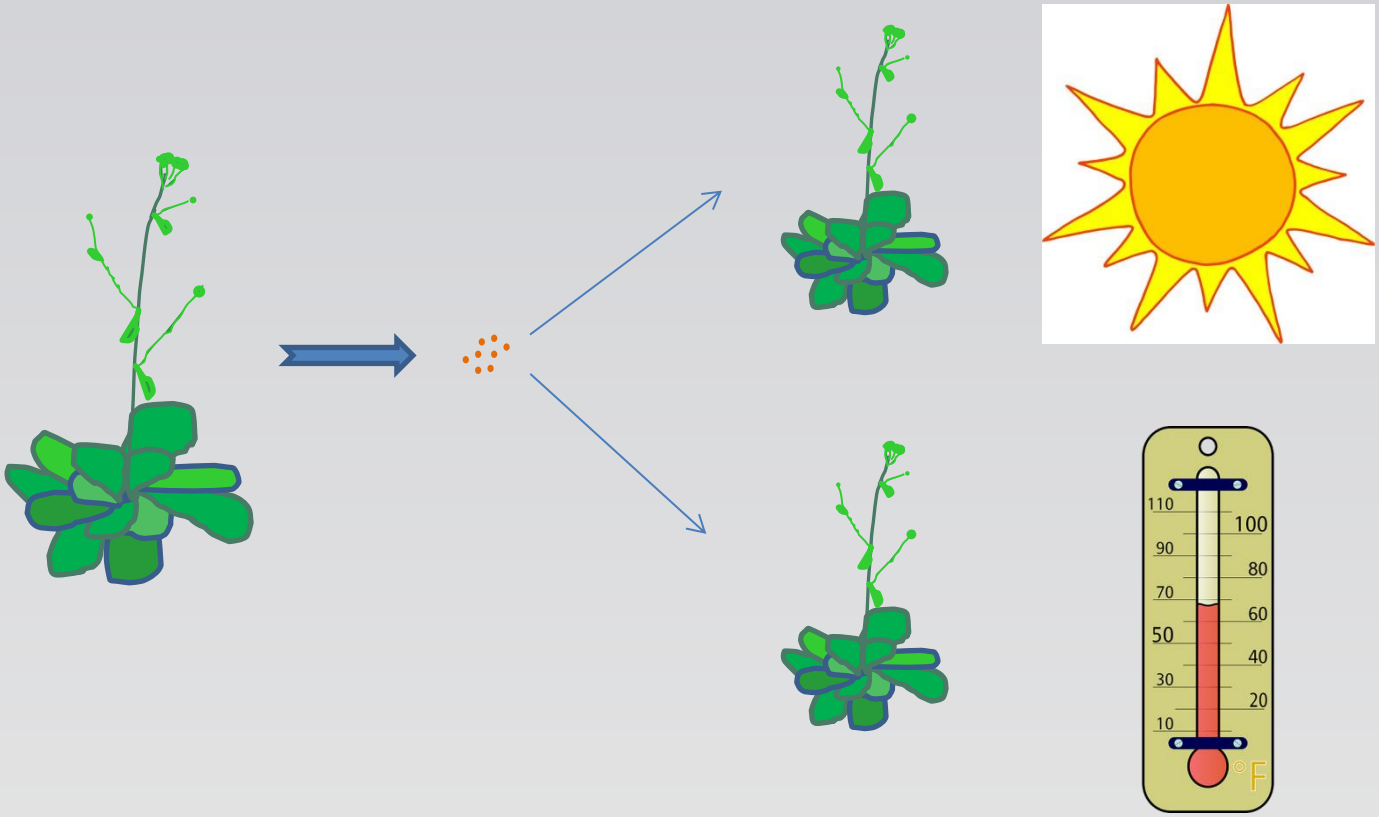
An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations

Vincent Segura[1,2,4], Bjarni J Vilhjálmsson[1,3,4], Alexander Platt[1,3], Arthur Korte[1], Ümit Seren[1], Quan Long[1] & Magnus Nordborg[1,3]

Population structure causes genome-wide linkage disequilibrium between unlinked loci, leading to statistical confounding in genome-wide association studies. Mixed models have been shown to handle the confounding effects of

the cluster memberships and principal-component loadings of individuals, respectively. Whereas these approaches are expected perform well when the population structure is simple, they may p form poorly when the structure is more complex: for example, wh

Combining traits for a joint GW



Perform a combined analysis on the data

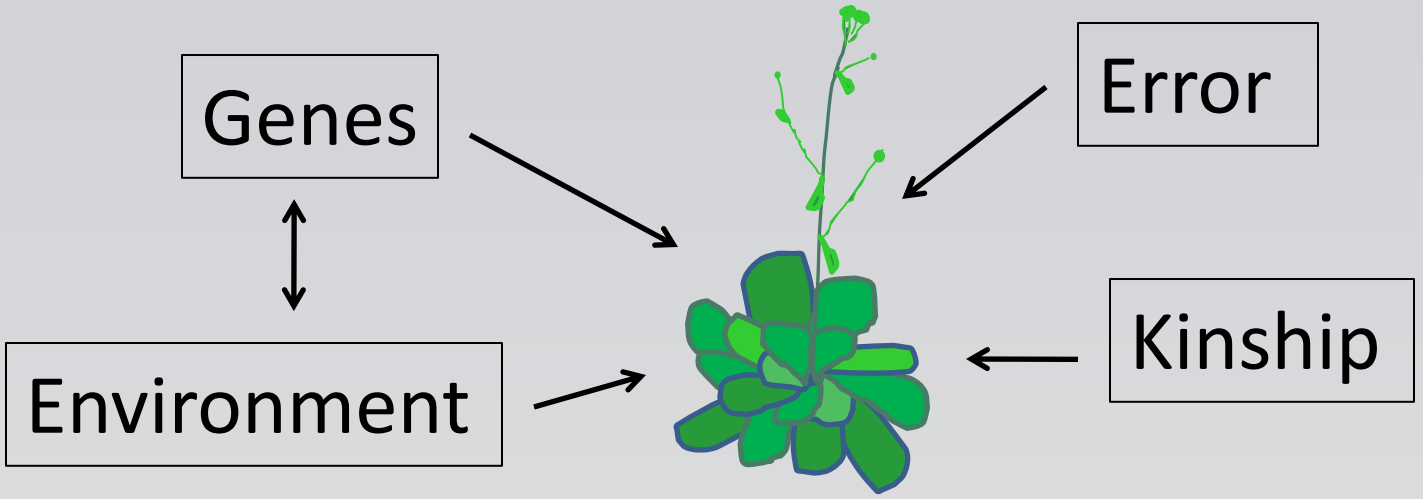
Many phenotypes are correlated

Center for Computational and Theoretical Biology

	X5_FT10	X6_FT16	X7_FT22	X43_FLC	X44_FRI	X80_LN10	X81_LN16	X82_LN22	X158_Siliq	X159_Siliq	X168_Chlo	X169_Chlo	X281_Stor	X282_Stora
X5_FT10	1	0.821332	0.780659	0.317835	0.280113	0.860099	0.697789	0.710058	-0.11299	-0.25084	0.066188	0.227002	0.313736	0.32493
X6_FT16	0.821332	1	0.882998	0.407995	0.354112	0.785828	0.829433	0.756154	-0.15999	-0.3158	0.094969	0.116744	0.352554	0.336236
X7_FT22	0.780659	0.882998	1	0.432569	0.332482	0.752075	0.767188	0.897305	-0.16028	-0.36557	0.113572	0.108052	0.358229	0.407532
X43_FLC	0.317835	0.407995	0.432569	1	0.399599	0.361573	0.431313	0.431833	0.128897	0.177295	0.060705	0.041051	-0.07754	-0.08906
X44_FRI	0.280113	0.354112	0.332482	0.399599	1	0.248296	0.33493	0.249646	-0.16378	-0.079	0.13947	0.031799	0.190994	0.294031
X80_LN10	0.860099	0.785828	0.752075	0.361573	0.248296	1	0.667059	0.657481	-0.17255	-0.31484	0.061653	0.207983	0.329016	0.3094
X81_LN16	0.697789	0.829433	0.767188	0.431313	0.33493	0.667059	1	0.6967	-0.10834	-0.17035	0.102676	0.141131	0.34098	0.296626
X82_LN22	0.710058	0.756154	0.897305	0.431833	0.249646	0.657481	0.6967	1	-0.19009	-0.32481	0.15016	0.09012	0.286067	0.363039
X158_Silique.16	-0.11299	-0.15999	-0.16028	0.128897	-0.16378	-0.17255	-0.10834	-0.19009	1	0.623108	0.036657	0.099576	-0.09551	-0.13473
X159_Silique.22	-0.25084	-0.3158	-0.36557	0.177295	-0.079	-0.31484	-0.17035	-0.32481	0.623108	1	0.070875	-0.11316	-0.3201	-0.30419
X168_Chlorosis.16	0.066188	0.094969	0.113572	0.060705	0.13947	0.061653	0.102676	0.15016	0.036657	0.070875	1	0.357333	-0.07354	0.010447
X169_Chlorosis.22	0.227002	0.116744	0.108052	0.041051	0.031799	0.207983	0.141131	0.09012	0.099576	-0.11316	0.357333	1	0.006146	0.077456
X281_Storage.7.days	0.313736	0.352554	0.358229	-0.07754	0.190994	0.329016	0.34098	0.286067	-0.09551	-0.3201	-0.07354	0.006146	1	0.844788
X282_Storage.28.days	0.32493	0.336236	0.407532	-0.08906	0.294031	0.3094	0.296626	0.363039	-0.13473	-0.30419	0.010447	0.077456	0.844788	1

Correlation of two traits can be due to shared genetics or a shared environment

MTMM



$$Y = \begin{pmatrix} Y1 \\ y2 \end{pmatrix} = \beta_0 + \beta_1 T + \beta_2 G + \beta_3 G*T + u + \varepsilon$$

$H_0 : Y = T$

$H_1 : Y = T + G$

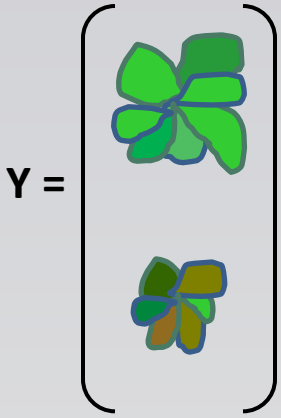
$H_2 : Y = T + G + GT$

Full p-value : GLS between H_2 and H_0

Trait specific p-value : GLS between H_2 and H_1

Trai common p-value : GLS between H_1 and H_0

MTMM (Multi-trait mixed model)

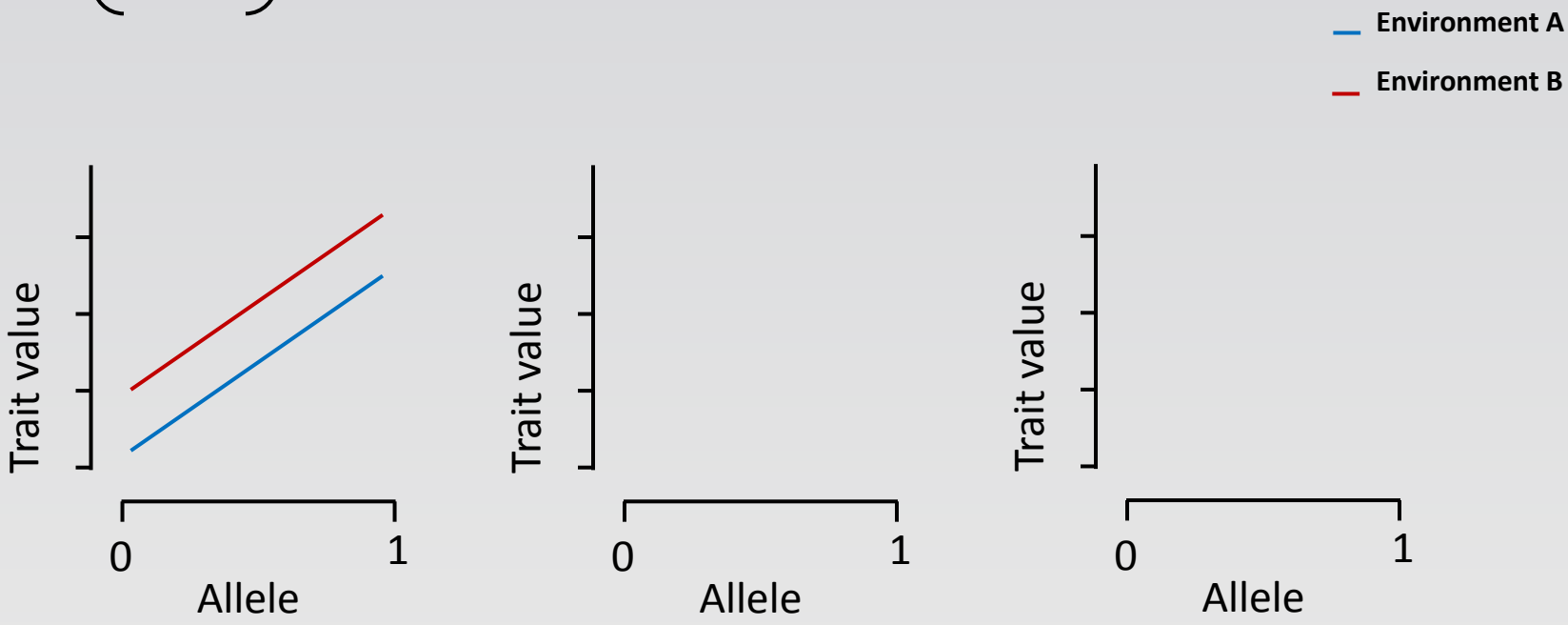


3 different tests :

Trait common : *same effect on the trait in both environments*

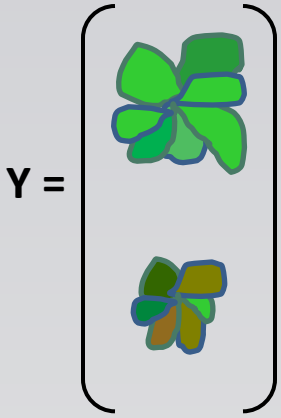
Trait specific : *effect only in one environemnt /*
opposite effect in the two environments

Full test : combination of common and specific effect



(same trait in two environments | longitudinal data)

MTMM (Multi-trait mixed model)

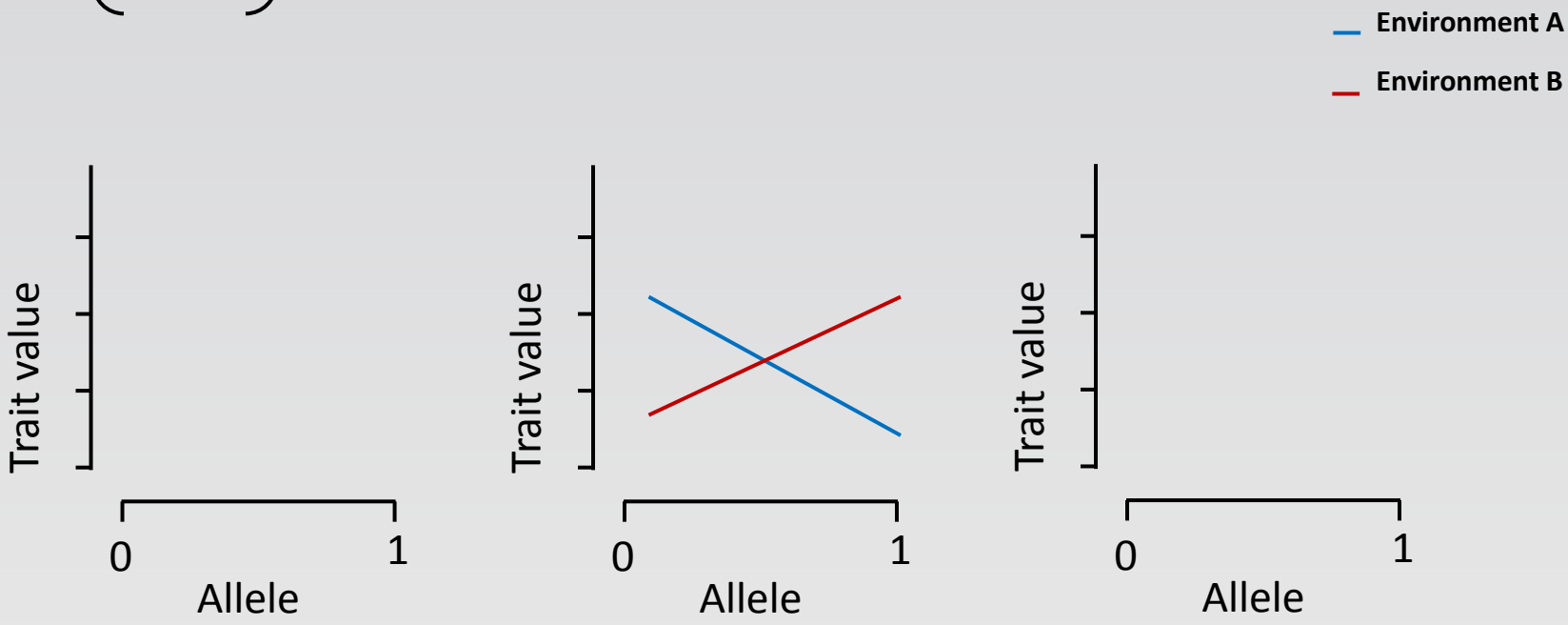


3 different tests :

Trait common : *same effect on the trait in both environments*

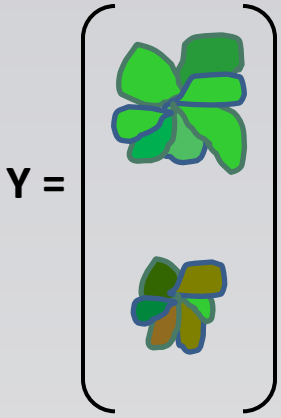
Trait specific : *effect only in one environemnt /*
opposite effect in the two environments

Full test : combination of common and specific effect



(same trait in two environments | longitudinal data)

MTMM (Multi-trait mixed model)

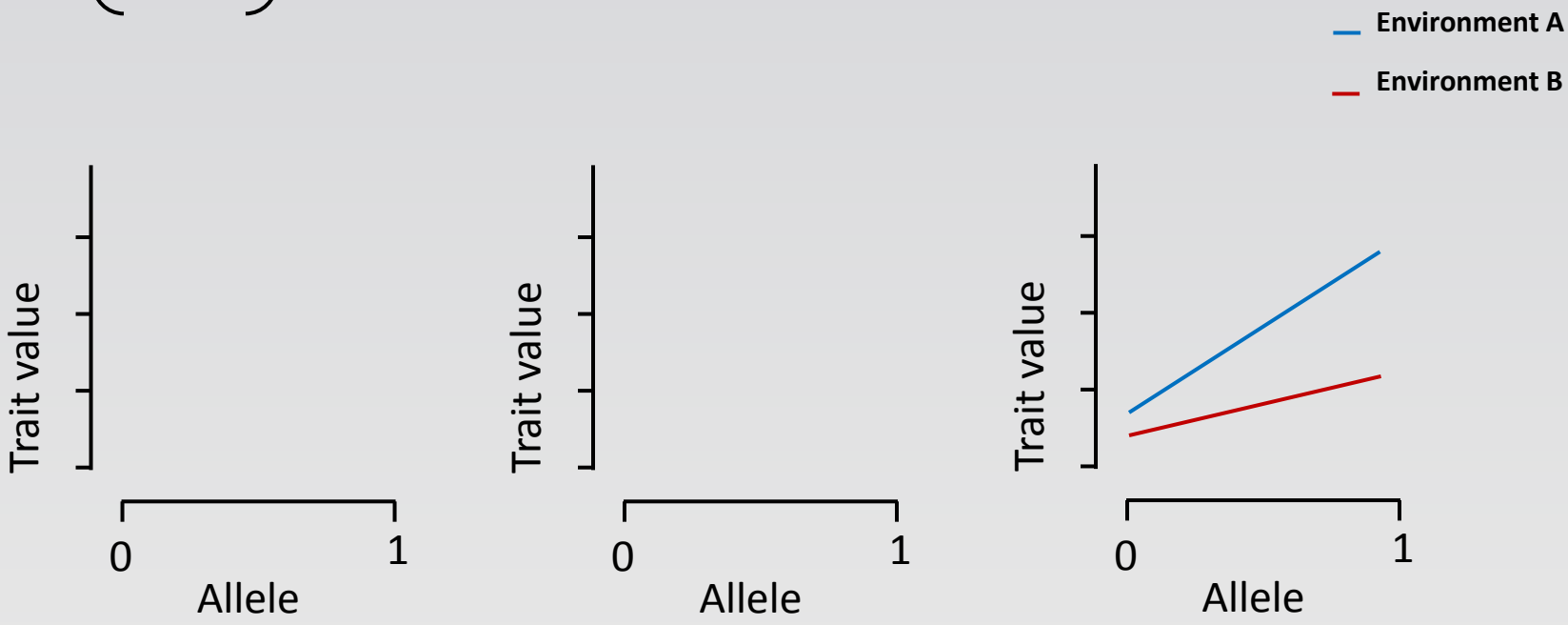


3 different tests :

Trait common : *same effect on the trait in both environments*

Trait specific : *effect only in one environemnt /*
opposite effect in the two environments

Full test : combination of common and specific effect



(same trait in two environments | longitudinal data)

Mixed model for a pair of correlated traits

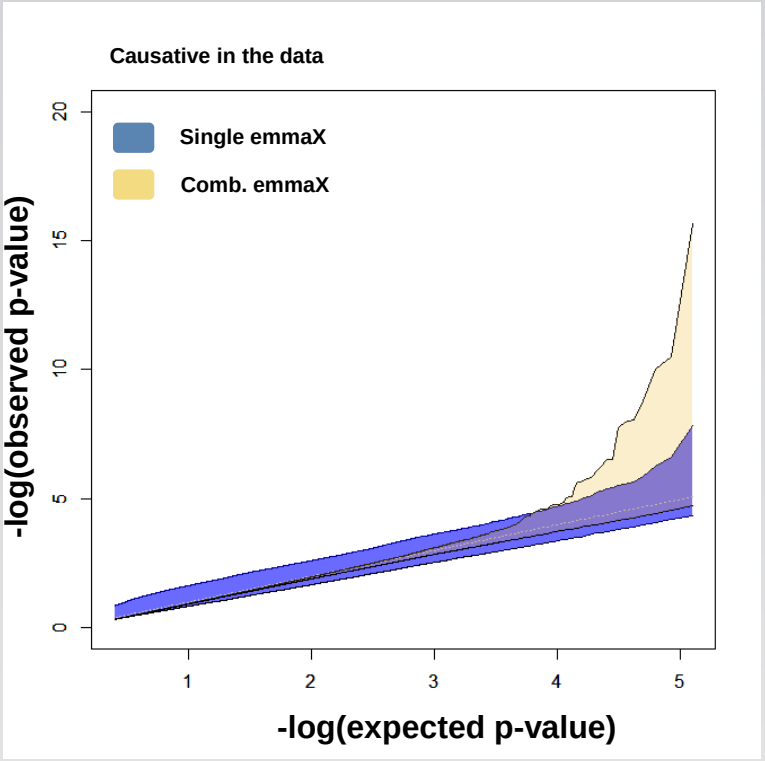
$$\mathbf{y} = \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \beta_0 + \mathbf{T}\beta_1 + \mathbf{G}\beta_2 + \mathbf{GT}\beta_3 + \mathbf{Zu} + \epsilon,$$

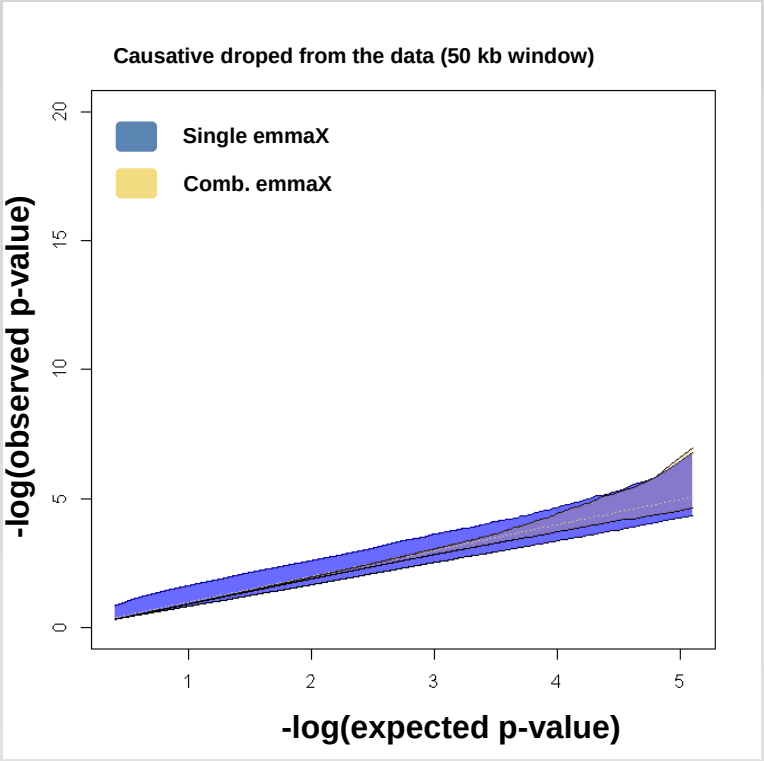
$$var(\mathbf{u}) = \begin{bmatrix} \sigma^2_{g11}\sigma_{g12} \\ \sigma_{g12}\sigma^2_{g22} \end{bmatrix} \otimes \mathbf{K}, var(\epsilon) = \begin{bmatrix} \sigma^2_{\epsilon 11}\sigma_{\epsilon 12} \\ \sigma_{\epsilon 12}\sigma^2_{\epsilon 22} \end{bmatrix} \otimes \mathbf{I}$$

Under the phenotypic model assumptions described in Kang et al. (2008) which yield K to be the IBS kinship matrix, and assuming the causal effects have a genetic correlation ρ, but are otherwise independent (no environmental correlation) , the genetic correlation turns out to be:

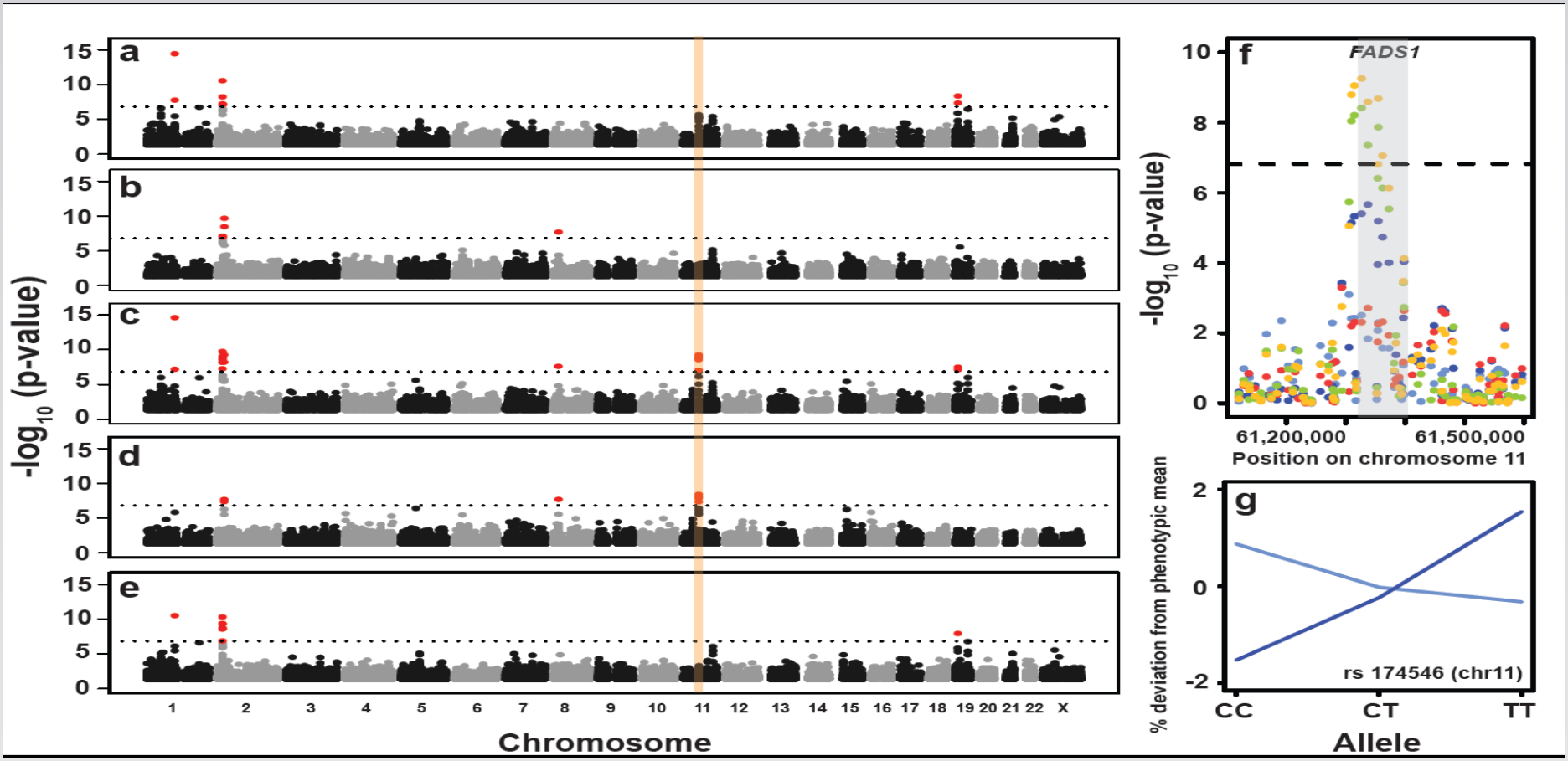
$$\rho_g = \frac{\rho_{\mathbf{y}_1,\mathbf{y}_2}\sqrt{(\sigma^2_{g_1} + \sigma^2_{e_1})(\sigma^2_{g_2} + \sigma^2_{e_2})}}{\sigma_{g_1}\sigma_{g_2}} = \frac{\rho_{\mathbf{y}_1,\mathbf{y}_2}}{h_1 h_2}$$

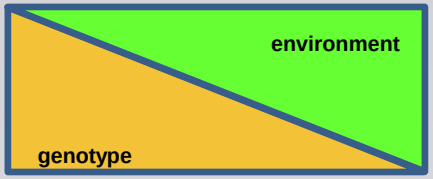
Simulations





MTMM (Multi-trait mixed model)





LETTER

doi:10.1038/nature10781

Genetic contributions to stability and change in intelligence from childhood to old age

Ian J. Deary[1,2]*, Jian Yang[3]*, Gail Davies[1,2], Sarah E. Harris[2,4], Albert Tenesa[4,5], David Liewald[1,2], Michelle Luciano[1,2], Lorna M. Lopez[1,2], Alan J. Gow[1,2], Janie Corley[1], Paul Redmond[1], Helen C. Fox[6], Suzanne J. Rowe[5], Paul Haggarty[7], Geraldine McNeill[6], Michael E. Goddard[8], David J. Porteous[2,4], Lawrence J. Whalley[6], John M. Starr[2,9] & Peter M. Visscher[2,3,10,11]*

Understanding the determinants of healthy mental ageing is a priority for society today[1,2]. So far, we know that intelligence differences show high stability from childhood to old age[3,4] and there are estimates of the genetic contribution to intelligence at different ages[5,6]. However, attempts to discover whether genetic causes contribute to differences in cognitive ageing have been relatively uninformative[7–10]. Here we provide an estimate of the genetic and environmental contributions to stability and change in intelligence across most of the human lifetime. We used genome-wide single nucleotide polymorphism (SNP) data from 1,940 unrelated individuals whose intelligence was measured in childhood (age 11 years) and again in old age (age 65, 70 or 79 years)[11,12]. We use a statistical method that allows genetic (co)variance to be estimated from SNP data on unrelated individuals[13–17]. We estimate that causal genetic variants in linkage disequilibrium with common SNPs account for 0.24 of the variation in cognitive ability change from childhood to old age. Using bivariate analysis, we estimate a genetic correlation between intelligence at age 11 years and in old age of 0.62. These estimates, derived from rarely available data on lifetime cognitive measures, warrant the search for genetic causes of cognitive stability and change.

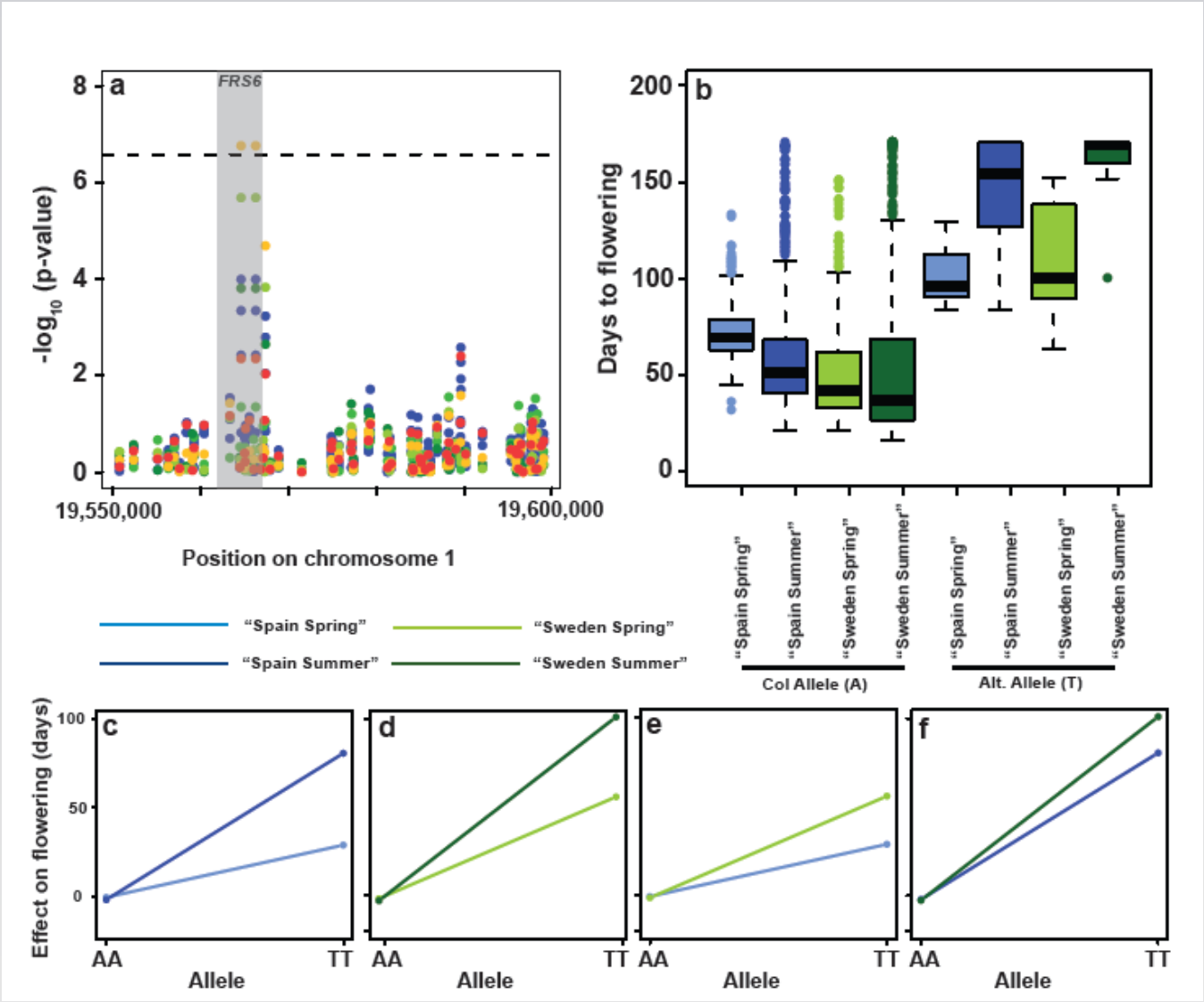
life course are largely unreplicated[22]. Therefore, an important novel contribution would be to partition the covariance between intelligence scores at either end of the human life course into genetic and environmental causes. To address this, the present study applies a new analytical method[13–17] to genome-wide association data from human participants with general cognitive ability test scores in childhood and again in old age.

Participants were members of the Aberdeen Birth Cohort 1936 (ABC1936) and the Lothian Birth Cohorts of 1921 and 1936 (LBC1921, LBC1936)[11,12,17]. They are community-dwelling, surviving

number of diverse cognitive tests. Additionally, the LBC1921 and LBC1936 cohorts re-took the Moray House Test in old age. Thus,

We estimate a genetic correlation between intelligence at age 11 years and in old age of 0.62.

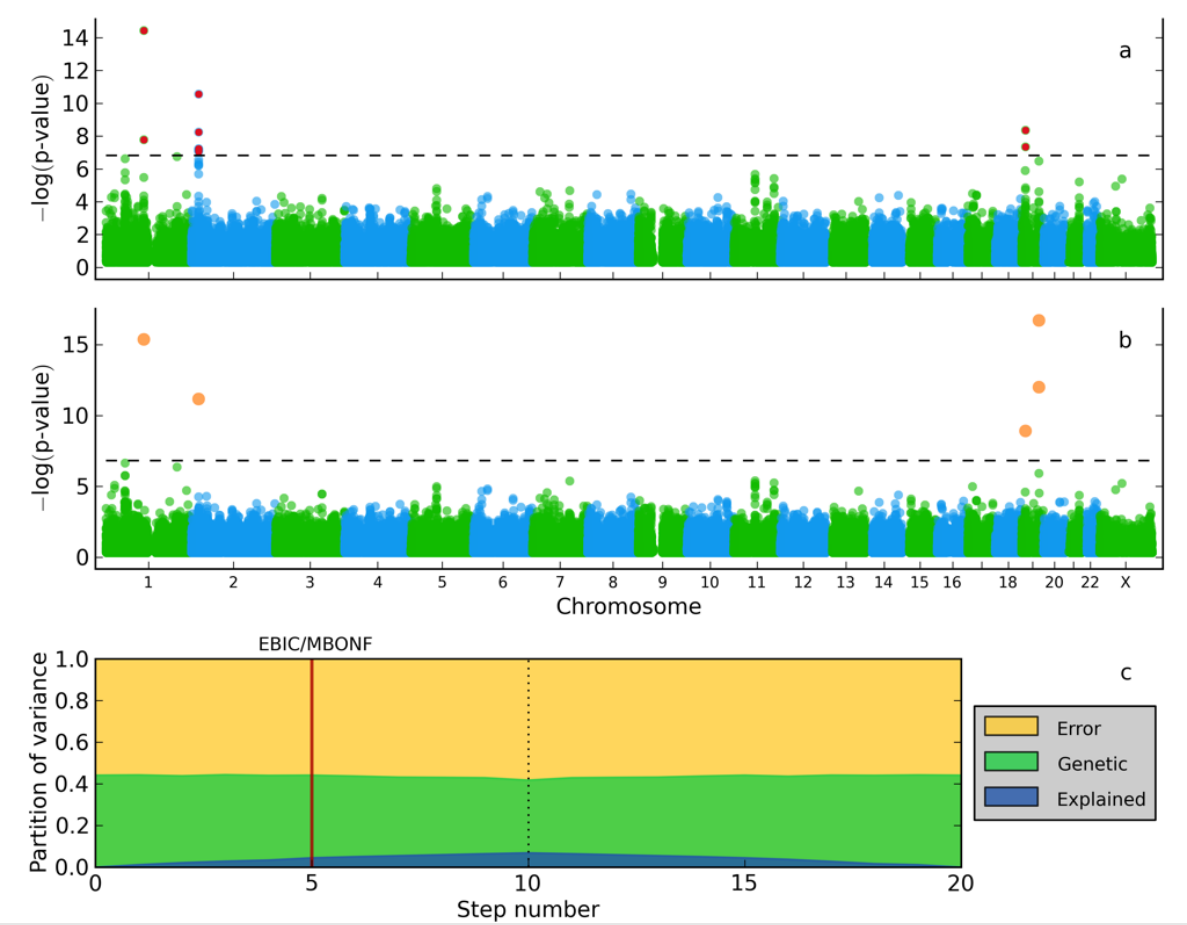
MTMM for more traits



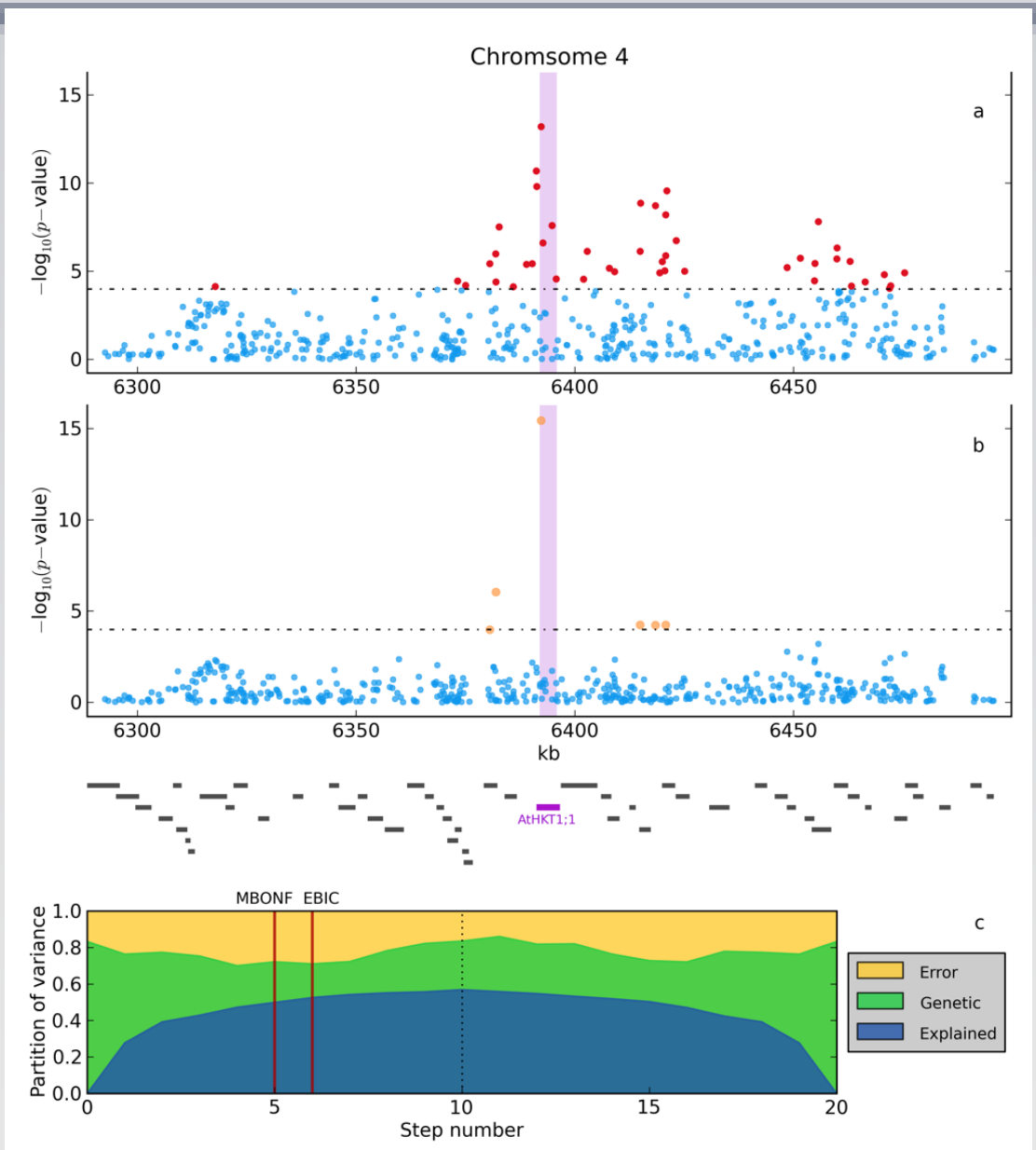
Multi Locus Mixed Model

Center for Computational and Theoretical Biology

Adding the most significant SNP as Co-factor to the model



Arabidopsis trait : Sodium levels in leaves

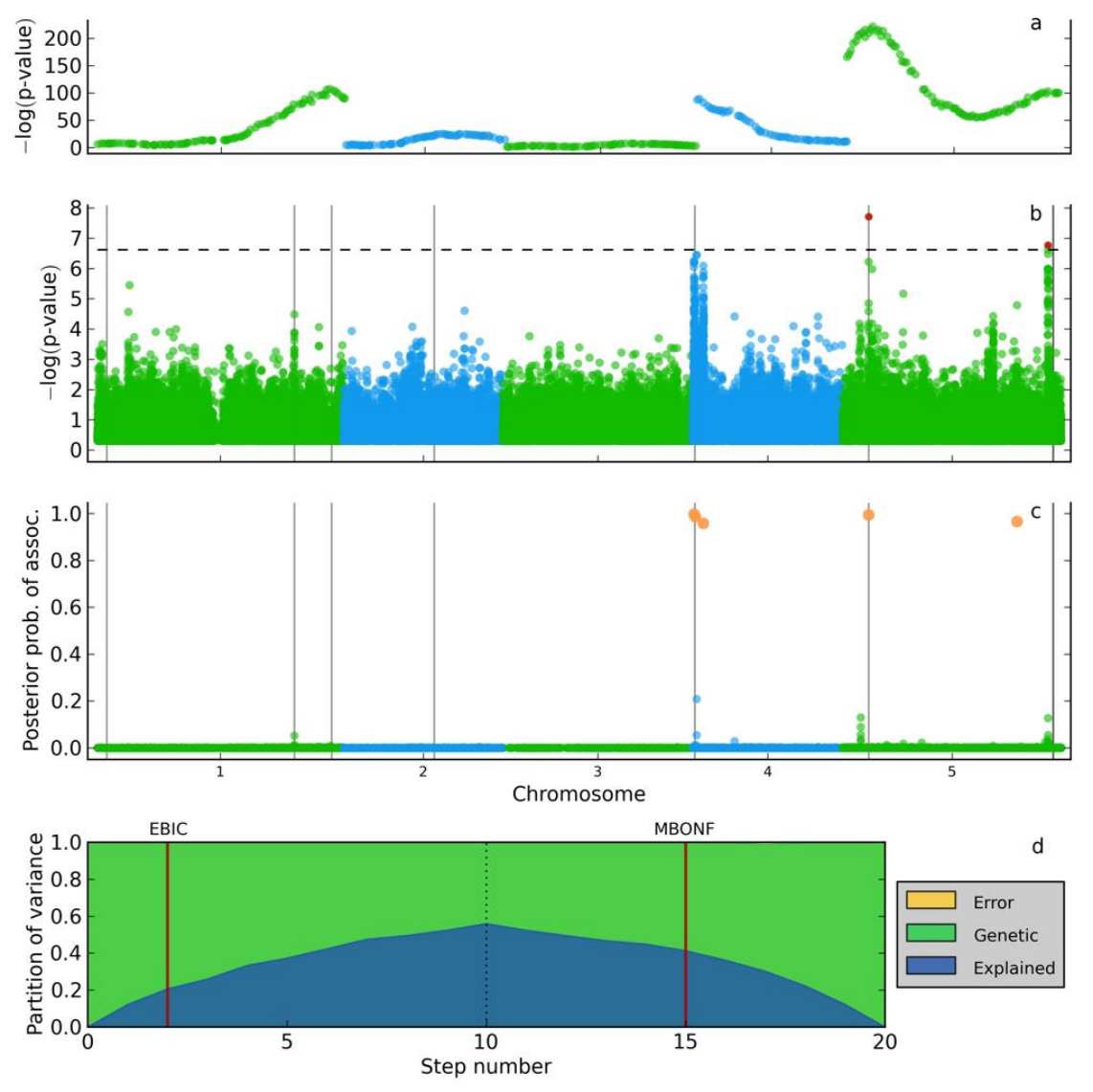


A Bayesian mixed model.

- Bayes factor: $$\mathrm{BF} = \frac{\mathrm{P}(D|M_1)}{\mathrm{P}(D|M_0)} = \frac{\int_{\theta_1} \mathrm{P}(D|\theta_1, M_1) d\theta_1}{\int_{\theta_0} \mathrm{P}(D|\theta_0, M_0) d\theta_0},$$
  - If π is the prior probability that a SNP is causal then the posterior odds and posterior probability of associations are $$\mathrm{PO} = \mathrm{BF} \cdot \pi/(1-\pi) \qquad \mathrm{PPA} = \mathrm{PO}/(1+\mathrm{PO})$$
- Following Kass and Raftery (1995), we can (roughly) approximate the Bayes factor which only depends on the likelihood and the degrees of freedoms as:

$$\log(\mathrm{ABF}) = \log(\mathrm{P}(D|\beta, M_1)) - \log(\mathrm{P}(D|\beta, M_0)) - \frac{1}{2}(d_1 - d_0)\log(n)$$

Bayesian analysis combining the Linkage and GWAS study

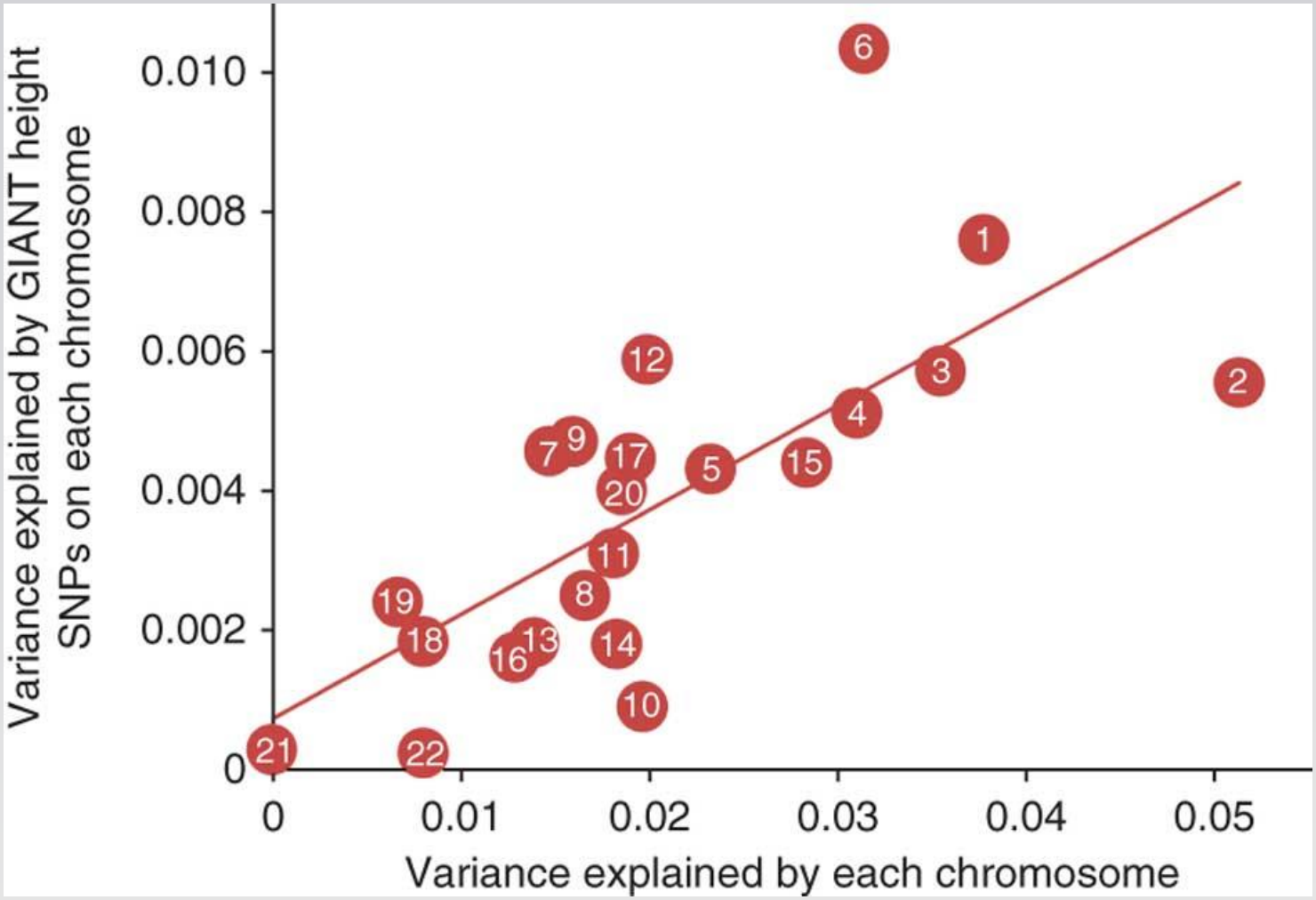


Linkage mapping score

EMMAX scan

Bayesian analysis combining the Linkage and the GWAS study

Partitioning of the phenotypic variance



Yang et al. , Nature Genetics 2011

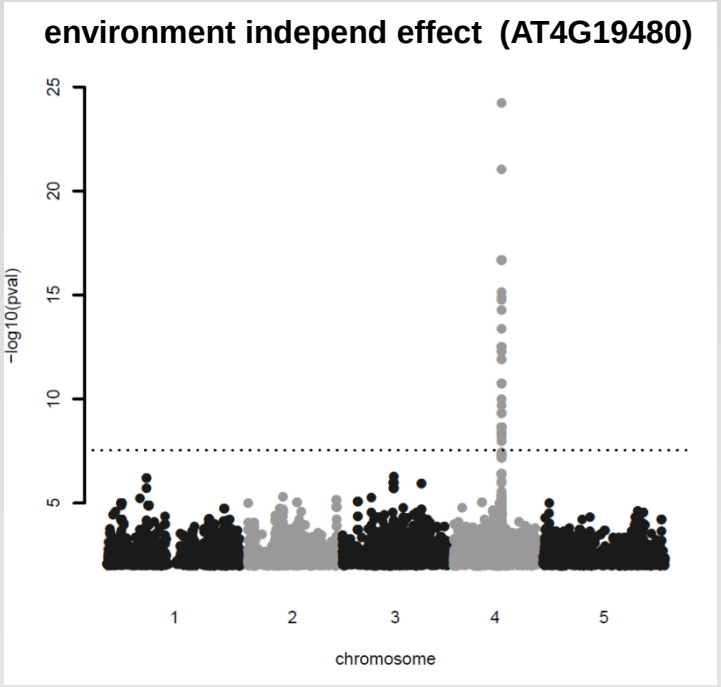
Analysis of RNAseq data

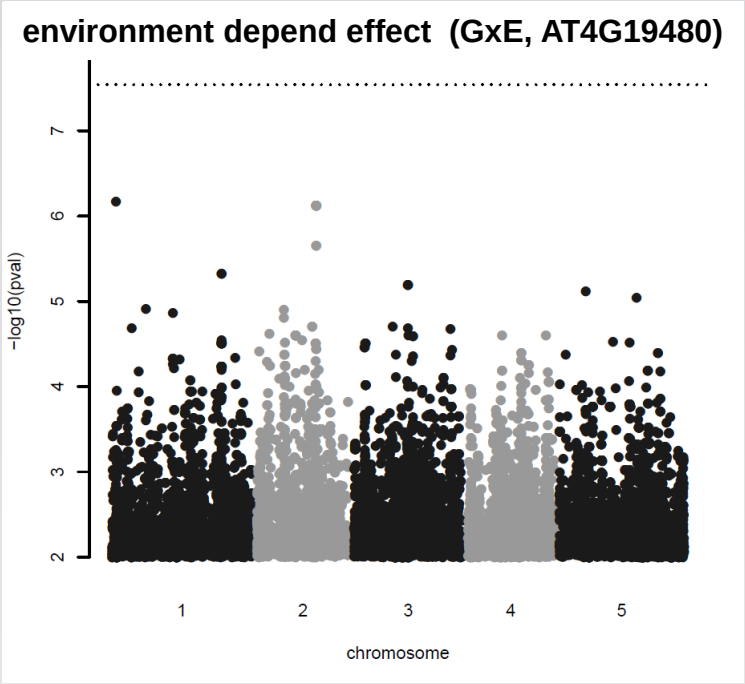
Center for Computational and Theoretical Biology

RNAseq data for 80 accessions grown in two different environments

~ 56,000 GWAS !

~ 4700 h computational time
~ 2 d on a cluster with 100 nodes

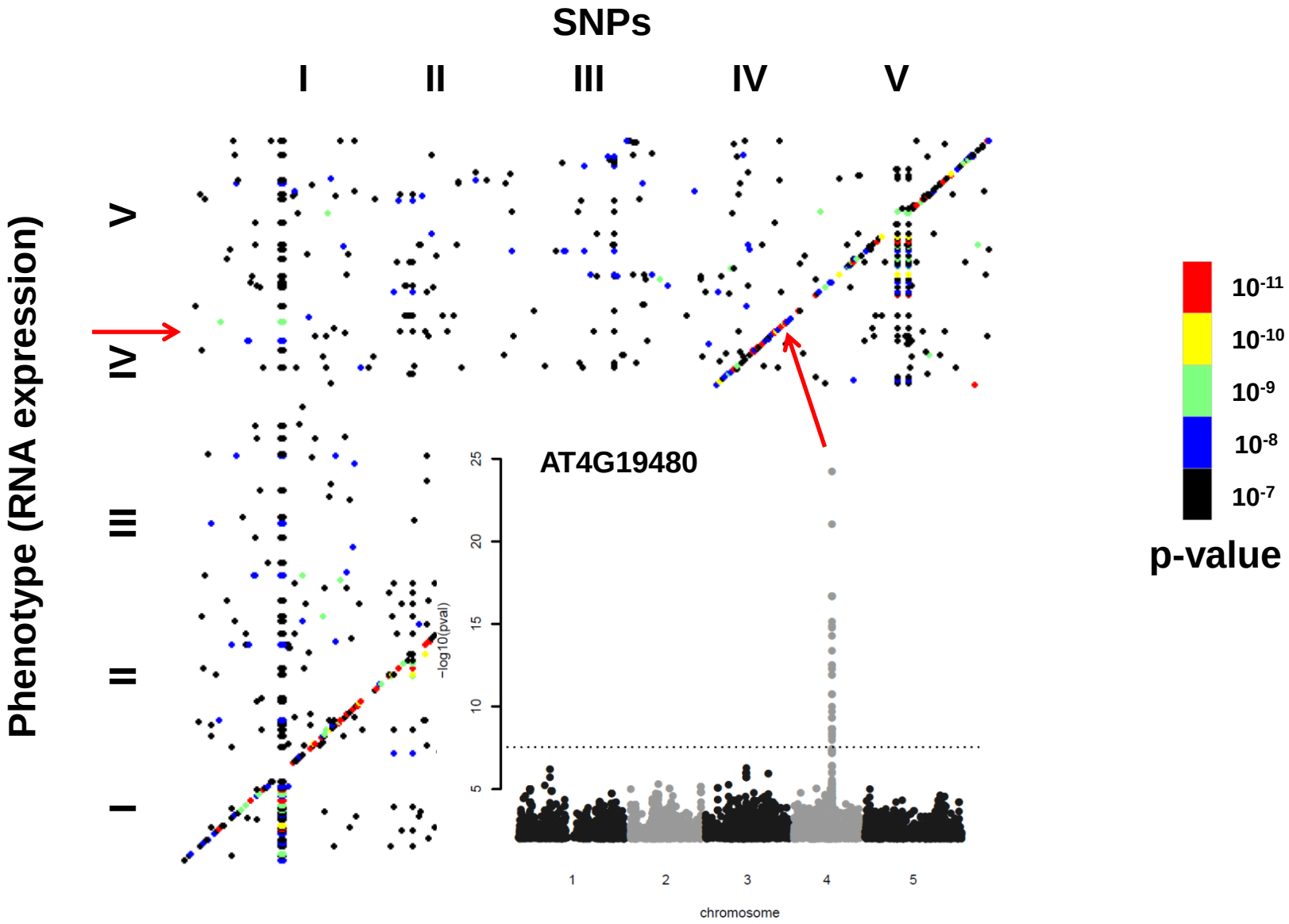




Frederik Coppens and Pieter Clauw (VIB Ghent)

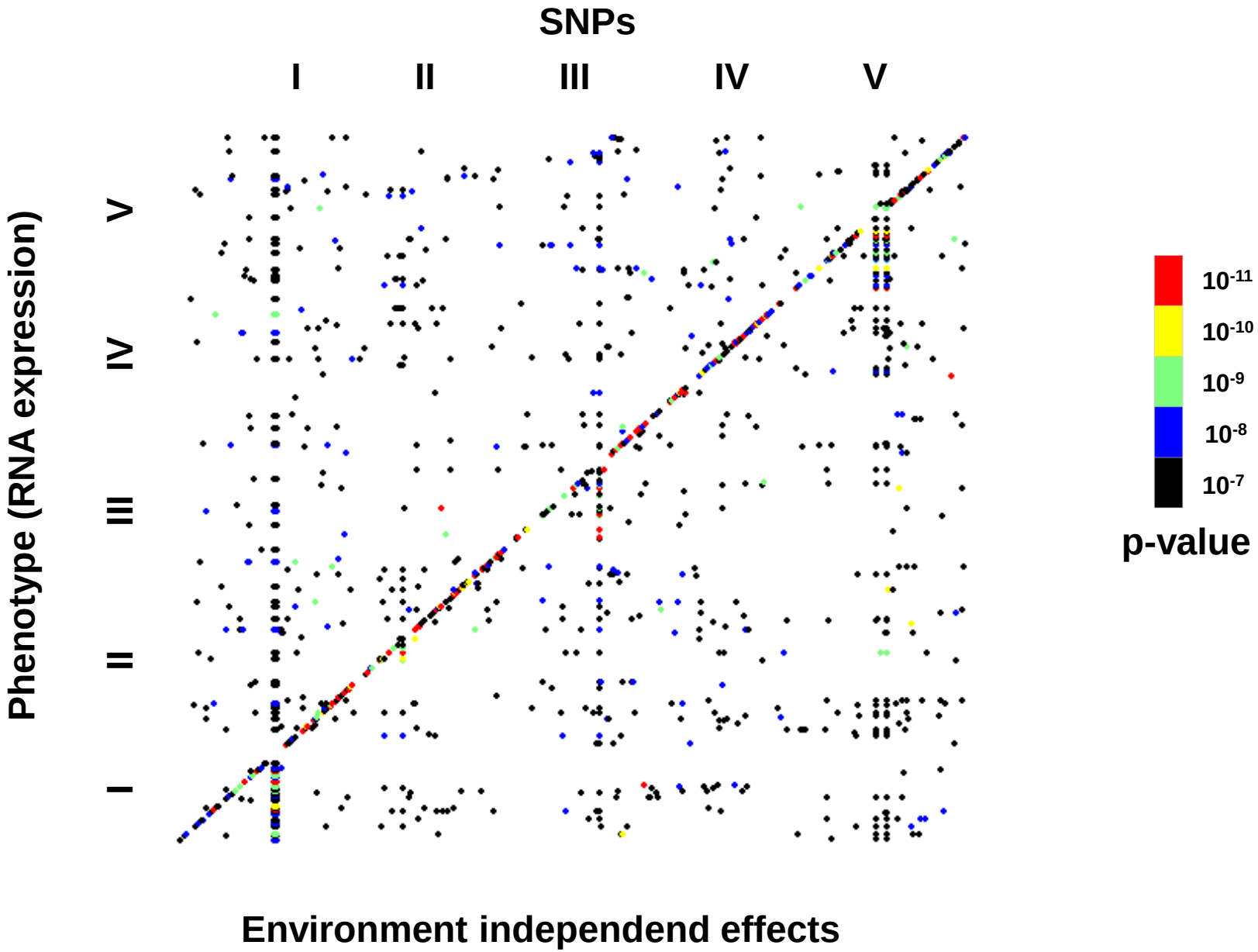
Summary of MTMM on RNAseq data

Center for Computational and Theoretical Biology



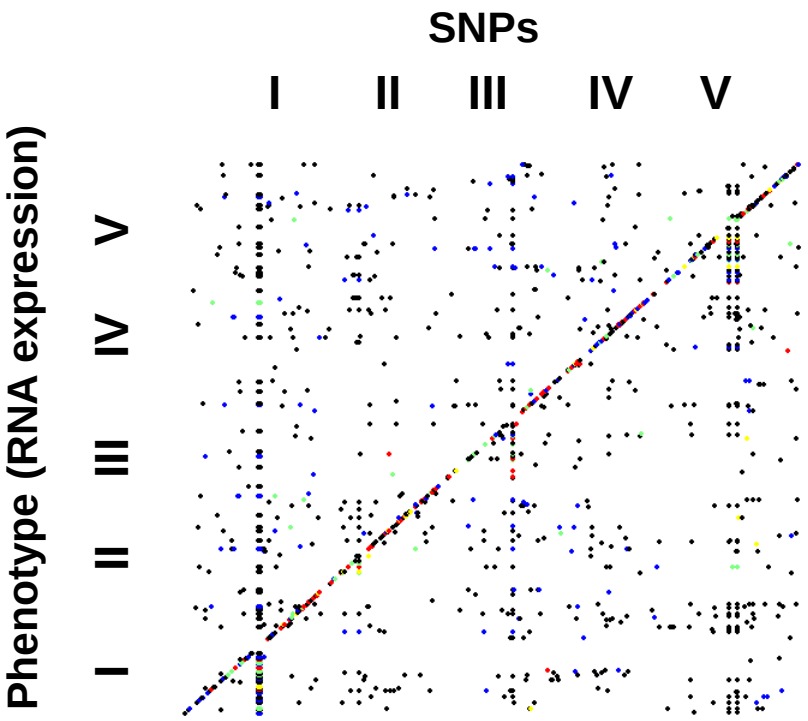
Environment independend effects

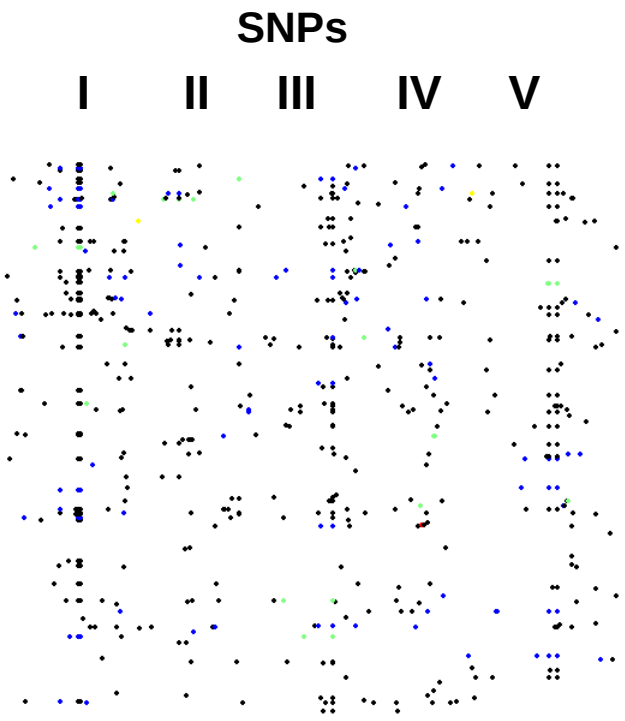
Summary of MTMM on RNAseq data



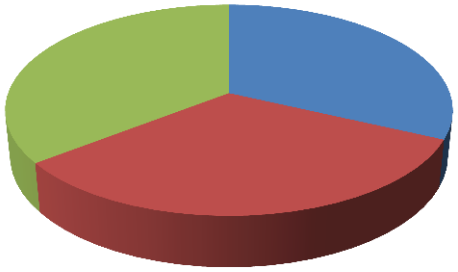
Summary of MTMM on RNAseq data

Center for Computational and Theoretical Biology

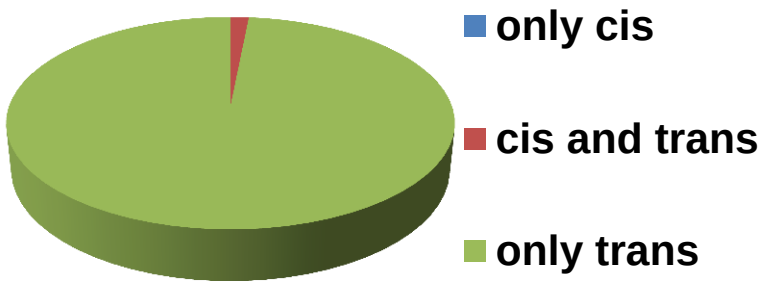




Environment independend

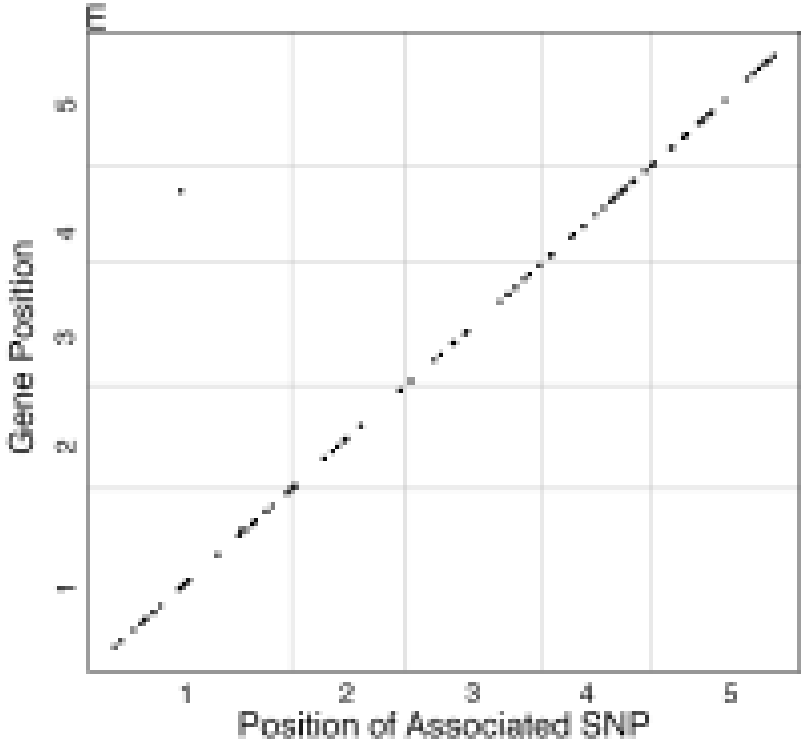


G x E

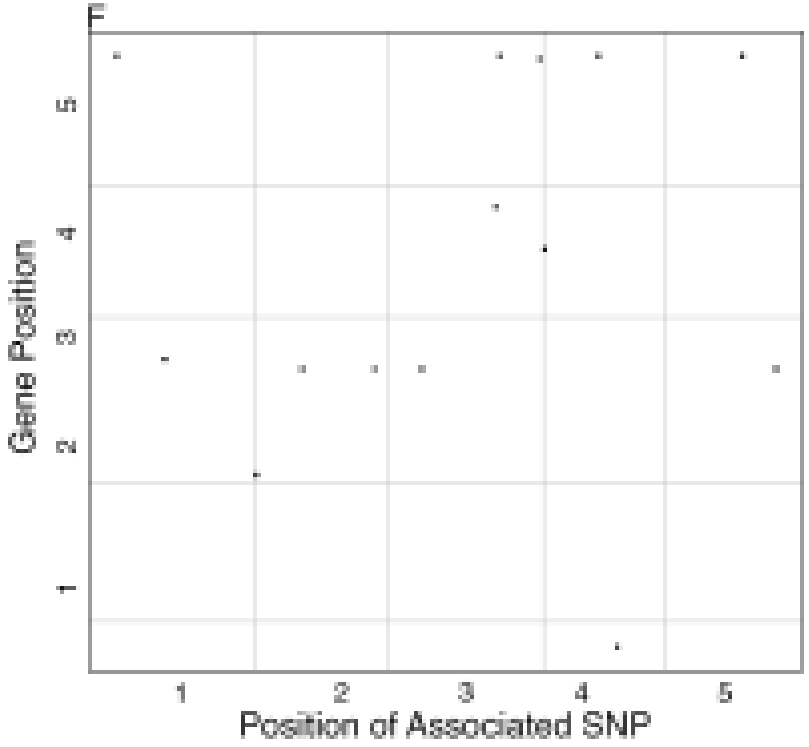


Summary of MTMM on RNAseq data

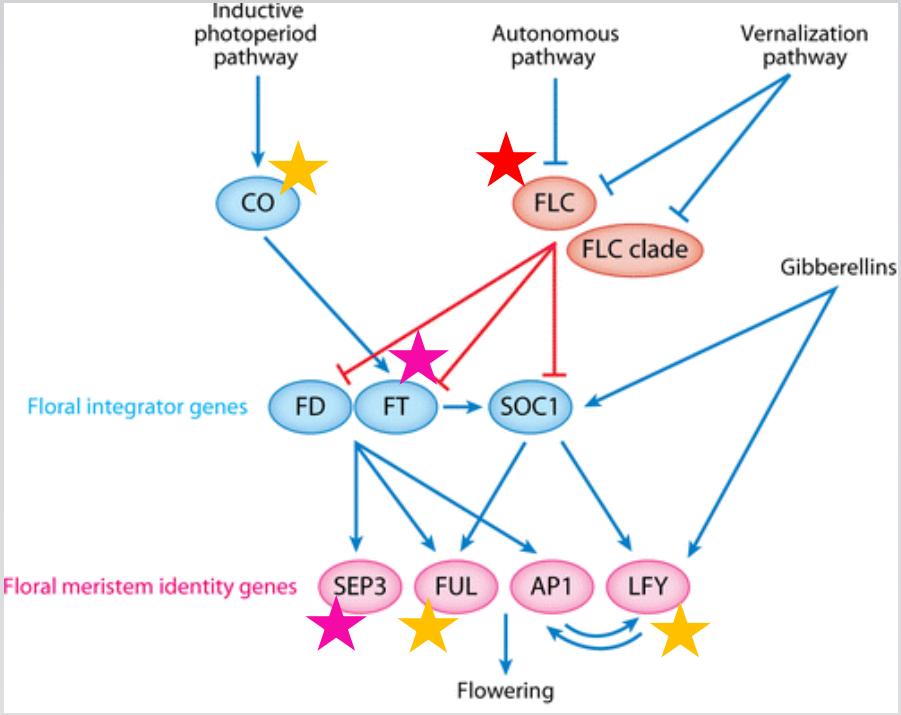
Environment independend



G x E



Beyond GWAS: Understanding the function of pathways

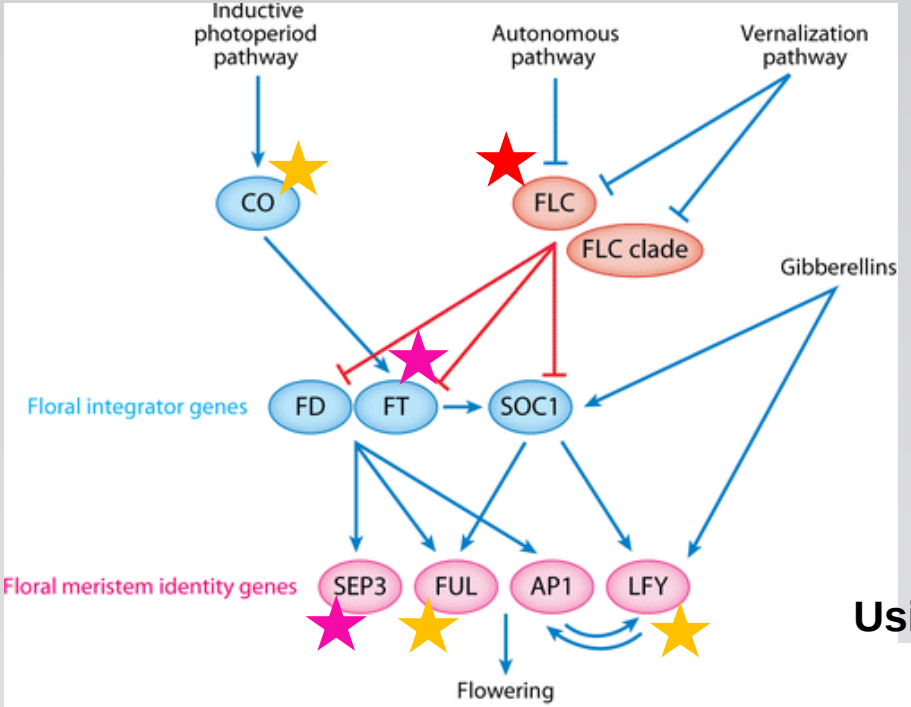


Significance in GWAS at 10°C

- ★ $< 10^{-3}$
- ★ $< 10^{-4}$
- ★ $< 10^{-5}$

Kim *et al*. *Annu Rev Cell Dev Biol* (2009)

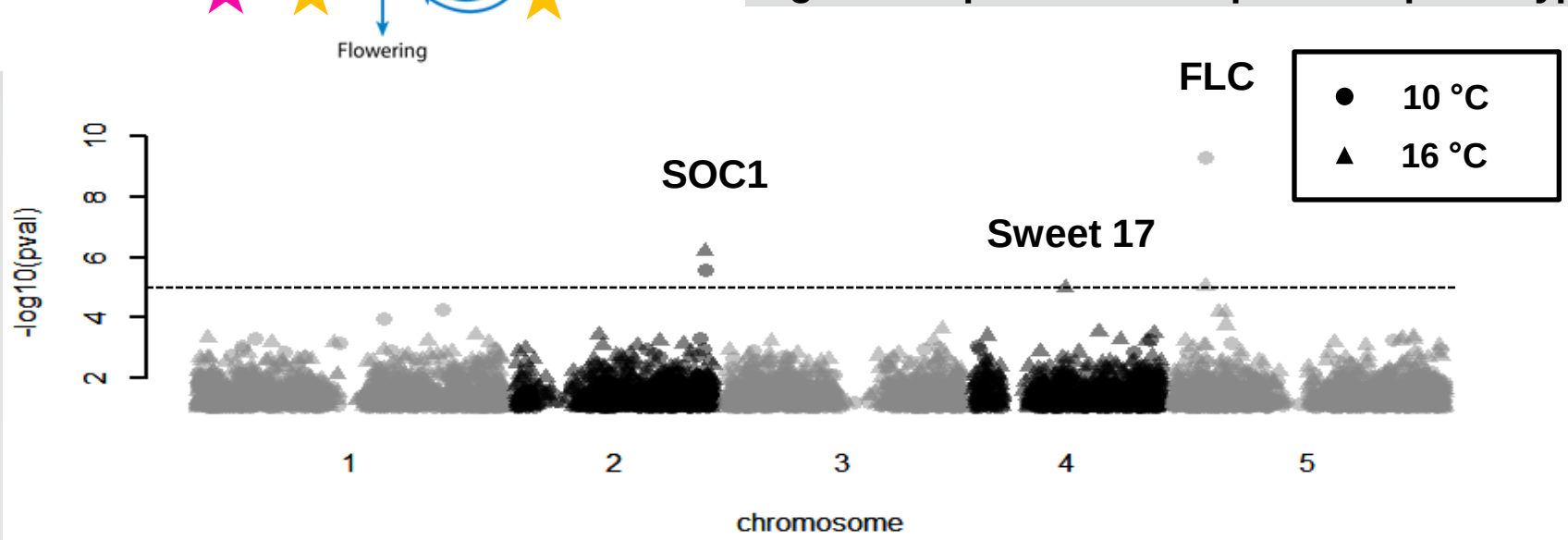
Beyond GWAS: Understanding the function of pathways



Significance in GWAS at 10°C

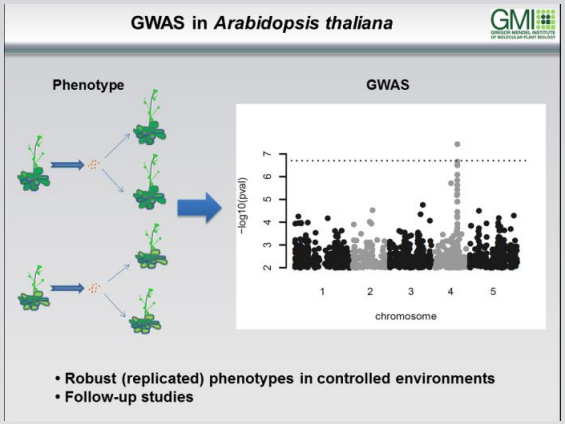
- < 10^{-3}
- < 10^{-4}
- < 10^{-5}

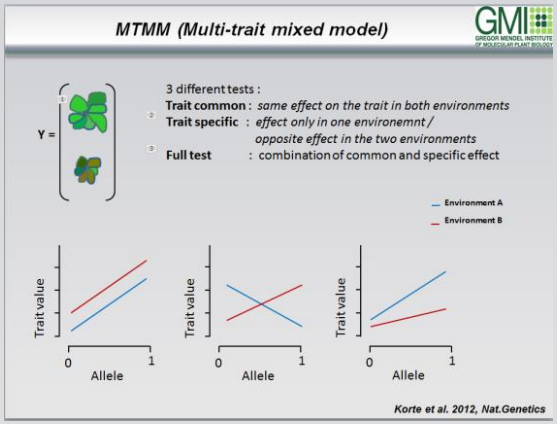
Using RNA expression to explain the phenotype

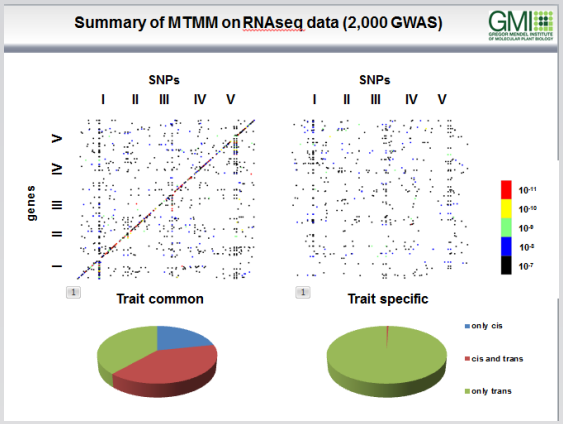


Summary

Understanding genotype – phenotype relationships





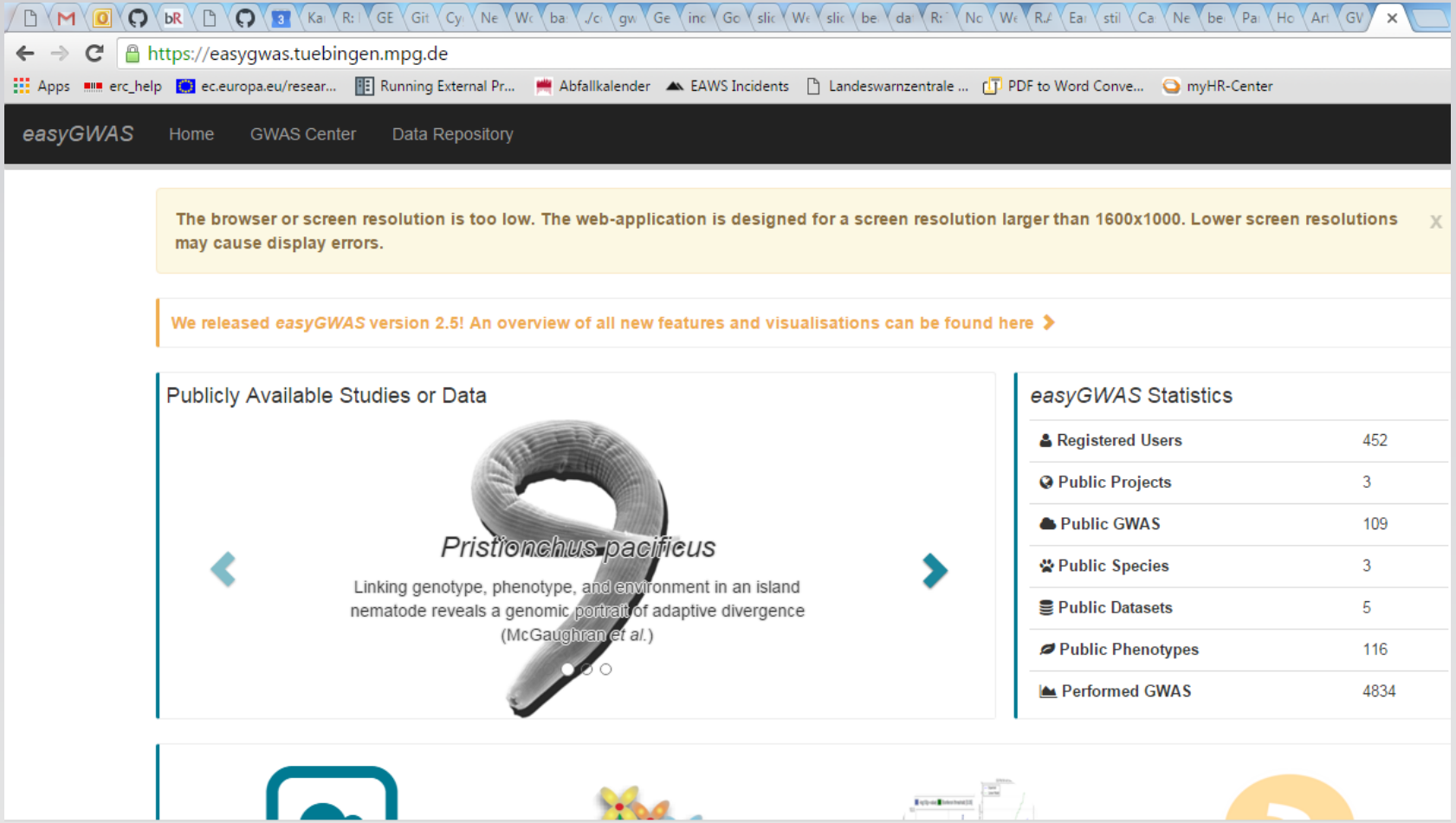


Integration of different data layers into joint models

GWAS with webtools

https://easygwas.tuebingen.mpg.de/

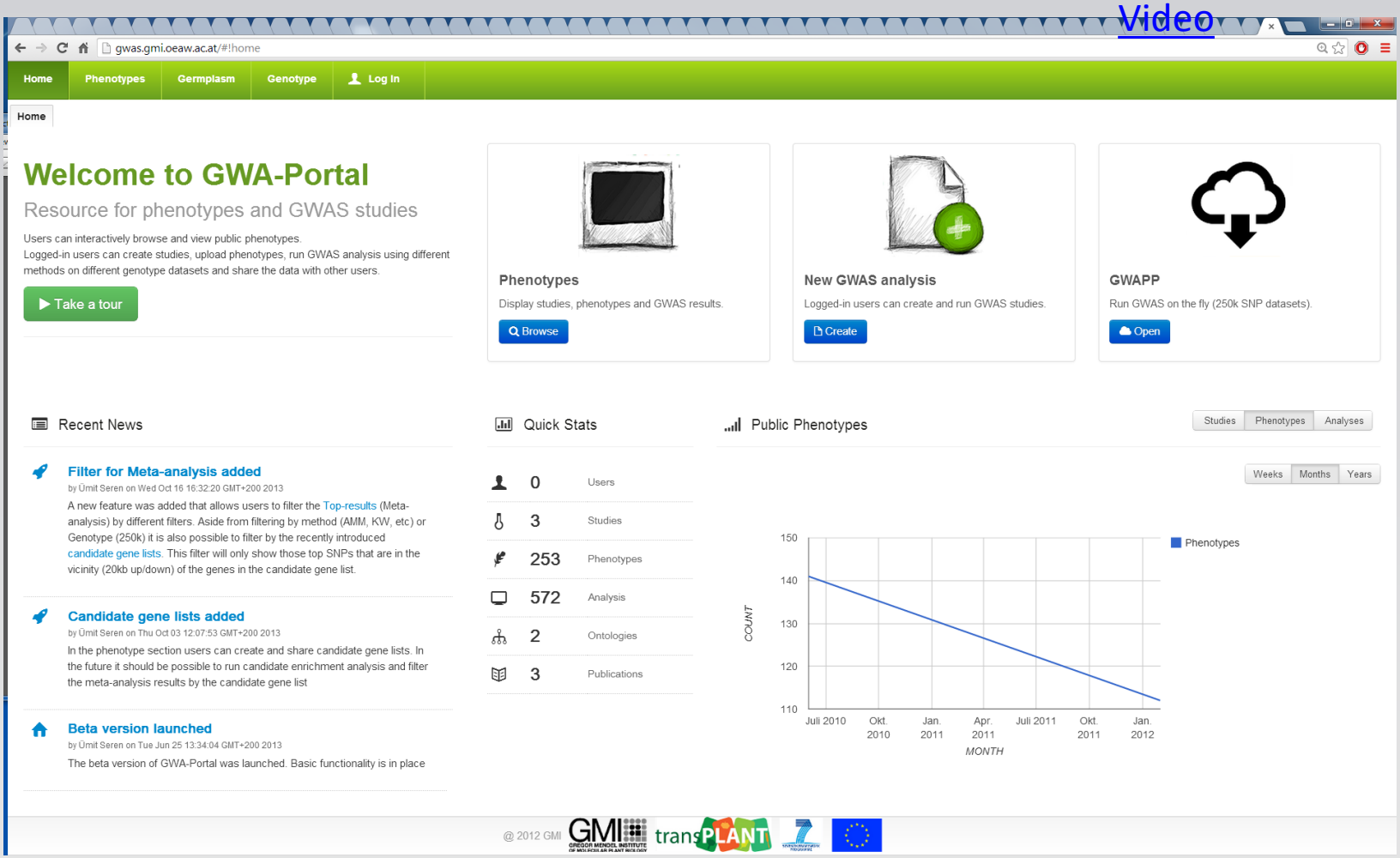
Center for Computational and Theoretical Biology



http://gwas.gmi.oeaw.ac.at

Center for Computational and Theoretical Biology

Video



Username: bag@workshop.org
Passwd: gwas

DATA on the server : **X_2029_1.rda**
Annotation_1001genomes.rda
MTMM_SAMPLE_DATA.Rdata
tair9.Rdata

Scripts for MLMM are on the Gmi github site:
https://github.com/Gregor-Mendel-Institute

GEMMA is installed on the server : [akorte@gdcsrv1 ~]$ **cd /usr/local/gemma/**

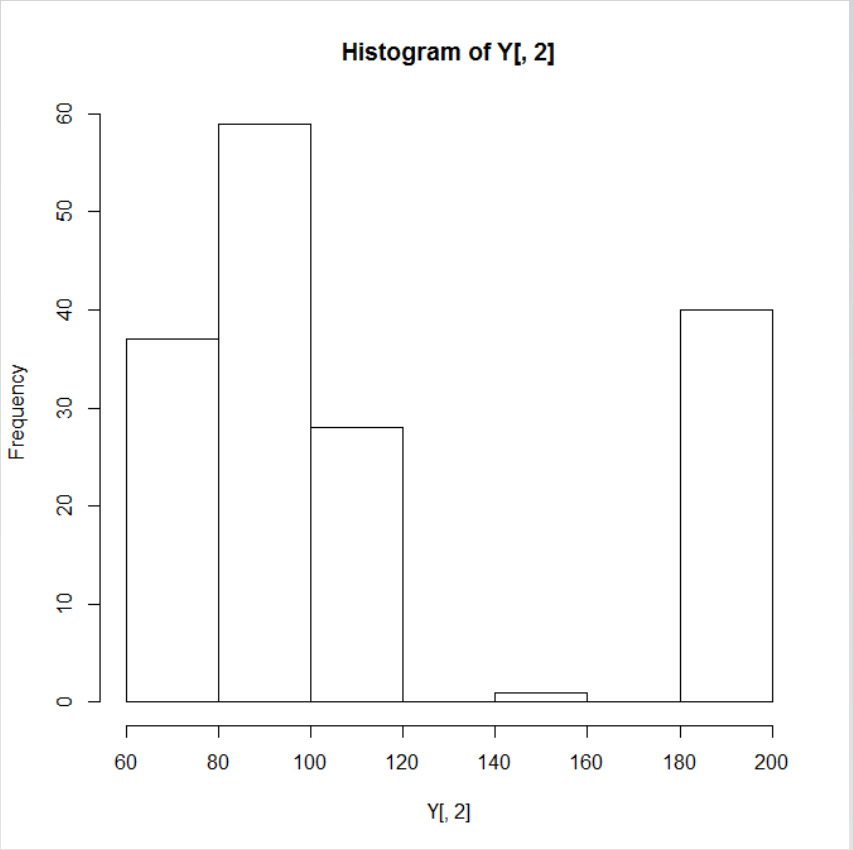
More ideas

Transform the phenotype (sqrt, log, boxcox) and rerun the analysis -> (compare results !)

Analyze subsets , remove outliers, permute the phenotype

Merge SNPs with genes; compare genic / intergenic SNPs ..

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd('Z:/arthur/work/R')
> source('emma.r')
>
> source('gwas.r')
> source('plots_gwas.r')
>
> load('MTMM_SAMPLE_DATA.Rdata')
> head(Y)
    ecotype_id       SD      SDV
298       5837  75.6667       NA
308       6008  69.6667 58.0000
314       6016 200.0000 55.0000
321       6024 200.0000 72.3333
328       6039  89.8083 70.5000
329       6040 200.0000       NA
> dim(Y)
[1] 165   3
> dim(X)
[1]   1307 214051
> X[1:5,1:5]
  1- 657 1- 3102 1- 4648 1- 4880 1- 5975
1      0       1       0       1       0
2      0       0       0       0       0
4      1       1       1       0       0
5      0       0       0       0       0
6      1       1       1       0       0
> dim(K)
[1] 1307 1307
> K[1:5,1:5]
          1         2         4         5         6
1 1.0000000 0.7521432 0.7535681 0.7707930 0.7552406
2 0.7521432 1.0000000 0.7709938 0.7489477 0.7394032
4 0.7535681 0.7709938 1.0000000 0.7701436 0.7377167
5 0.7707930 0.7489477 0.7701436 1.0000000 0.7397116
6 0.7552406 0.7394032 0.7377167 0.7397116 1.0000000
>
```



```
> hist(Y[,2])
> a<-Sys.time()
> amm_gwas(Y,X,K,include.lm=T,calculate.effect.size=FALSE)
GWAS performed on 165 ecotypes,  0 values excluded
SNP_INFO file created
pseudo-heritability estimate is  0.9999546
> Sys.time()-a
Time difference of 1.359486 mins
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC        MAF       Pval     Pval_lm
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05266552 0.004369247
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33888896 0.389214696
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78554188 0.862913234
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57424989 0.222504617
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.65037890 0.616469715
6 1- 10003346   1 10003346   82   73  73 0.44242424 0.06652066 0.018774138
```

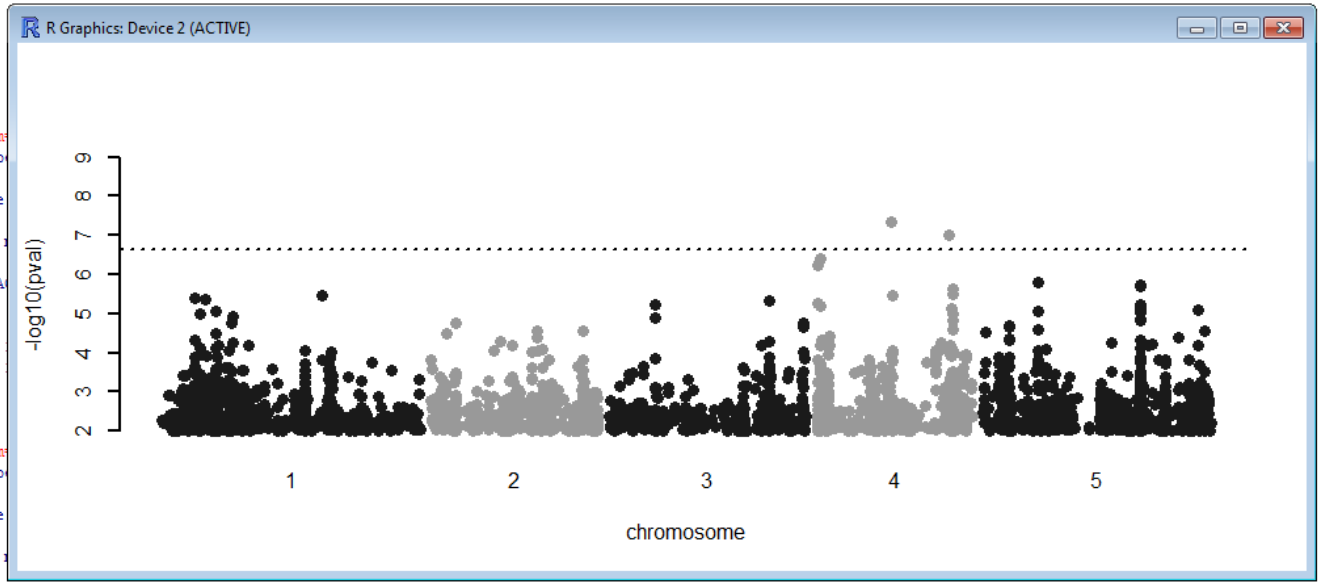
```
> amm_gwas(Y,X,K,n=3)
GWAS performed on 150 ecotypes,  0 values excluded
SNP_INFO file created
pseudo-heritability estimate is  0.9560748
> head(output)
           SNP Chr      Pos AC_1 AC_0 MAC        MAF      Pval
1 1- 10000173   1 10000173   33  117  33 0.22000000 0.6862535
2 1- 10000715   1 10000715   17  133  17 0.11333333 0.6342901
3 1- 10001699   1 10001699   99   51  51 0.34000000 0.5829813
4 1- 10002088   1 10002088   94   56  56 0.37333333 0.5040329
5 1- 10002264   1 10002264    5  145   5 0.03333333 0.3683449
6 1- 10003346   1 10003346   85   65  65 0.43333333 0.7158708
>
```

Estimating the effect size

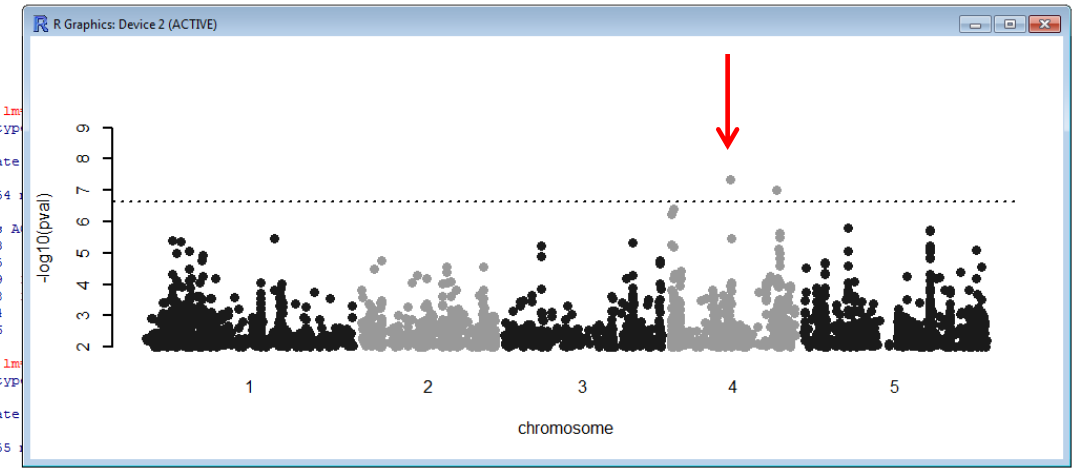
Center for Computational and Theoretical Biology

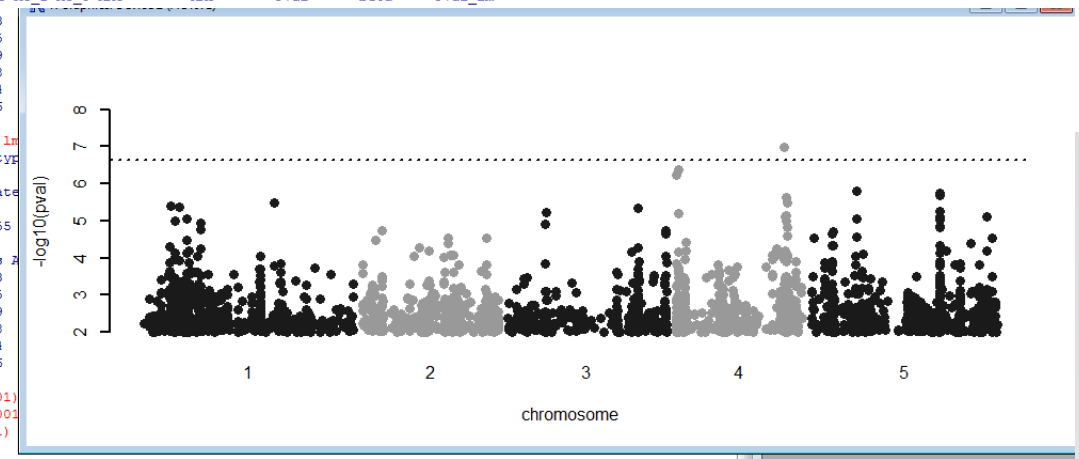
```
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm=T)
GWAS performed on 165 ecotypes,  0 values excluded
SNP_INFO file created
pseudo-heritability estimate is  0.9999546
> Sys.time()-a
Time difference of 1.376254 mins
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC        MAF       Pval     Pval_lm
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05266552 0.004369247
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33888896 0.389214696
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78554188 0.862913234
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57424989 0.222504617
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.65037890 0.616469715
6 1- 10003346   1 10003346   92   73  73 0.44242424 0.06652066 0.018774138
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm=T,calculate.effect.size=T)
GWAS performed on 165 ecotypes,  0 values excluded
SNP_INFO file created
pseudo-heritability estimate is  0.9999546
> Sys.time()-a
Time difference of 7.383855 mins
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC        MAF       Pval      beta     Pval_lm
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05193852 13.183344 0.004369247
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33739802 -9.325060 0.389214696
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78489523  1.675355 0.862913234
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57306712 -3.366112 0.222504617
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.64937048 -6.566181 0.616469715
6 1- 10003346   1 10003346   92   73  73 0.44242424 0.06568446 -9.570731 0.018774138
>
```

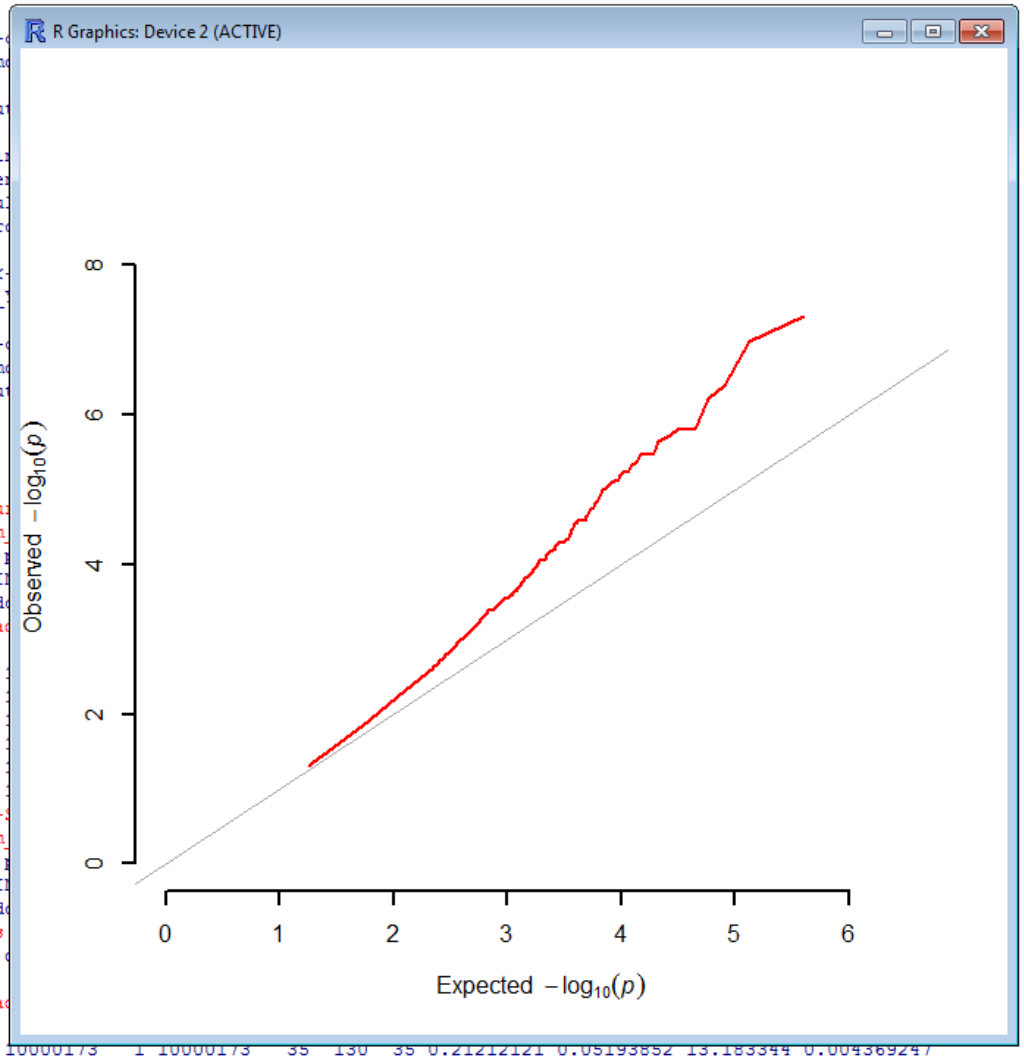
```
6008  69.6667
6016 200.0000
6024 200.0000
6039  89.8083
6040 200.0000
> hist(Y_)
>
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm
GWAS performed on 165 ecotyp
SNP_INFO file created
pseudo-heritability estimate
> Sys.time()-a
Time difference of 1.376254
> head(output)
          SNP Chr      Pos A
1 1- 10000173   1 10000173
2 1- 10000715   1 10000715
3 1- 10001699   1 10001699
4 1- 10002088   1 10002088
5 1- 10002264   1 10002264
6 1- 10003346   1 10003346
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm
GWAS performed on 165 ecotyp
SNP_INFO file created
pseudo-heritability estimate
> Sys.time()-a
Time difference of 7.383855
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC        MAF       Pval      beta     Pval_lm
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05193852 13.183344 0.004369247
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33739802 -9.325060 0.389214696
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78489523  1.675355 0.862913234
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57306712 -3.366112 0.222504617
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.64937048 -6.566181 0.616469715
6 1- 10003346   1 10003346   92   73  73 0.44242424 0.06568446 -9.570731 0.018774138
> plot_gwas(output)
```



```
6008  69.6667
6016 200.0000
6024 200.0000
6039  89.8083
6040 200.0000
> hist(Y_)
>
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm
GWAS performed on 165 ecotyp
SNP_INFO file created
pseudo-heritability estimate
> Sys.time()-a
Time difference of 1.376254
> head(output)
          SNP Chr      Pos A
1 1- 10000173   1 10000173
2 1- 10000715   1 10000715
3 1- 10001699   1 10001699
4 1- 10002088   1 10002088
5 1- 10002264   1 10002264
6 1- 10003346   1 10003346
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm
GWAS performed on 165 ecotyp
SNP_INFO file created
pseudo-heritability estimate
> Sys.time()-a
Time difference of 7.383855
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC       MAF       Pval      beta     Pval_lm
1 1- 10000173   1 10000173
2 1- 10000715   1 10000715
3 1- 10001699   1 10001699
4 1- 10002088   1 10002088
5 1- 10002264   1 10002264
6 1- 10003346   1 10003346
> a<-Sys.time()
> amm_gwas(Y_,X,K,include.lm
GWAS performed on 165 ecotyp
SNP_INFO file created
pseudo-heritability estimate
> Sys.time()-a
Time difference of 7.383855
> head(output)
          SNP Chr      Pos A
1 1- 10000173   1 10000173
2 1- 10000715   1 10000715
3 1- 10001699   1 10001699
4 1- 10002088   1 10002088
5 1- 10002264   1 10002264
6 1- 10003346   1 10003346
> plot_gwas(output)
> plot_gwas(output,maf=0.01)
> plot_gwas(output,maf=0.001
> plot_gwas(output,maf=0.1)
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC       MAF       Pval      beta     Pval_lm
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05193852 13.183344 0.004369247
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33739802 -9.325060 0.389214696
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78489523  1.675355 0.862913234
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57306712 -3.366112 0.222504617
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.64937048 -6.566181 0.616469715
6 1- 10003346   1 10003346   92   73  73 0.44242424 0.06568446 -9.570731 0.018774138
> plot_gwas(output)
> plot_gwas(output,maf=0.1)
```

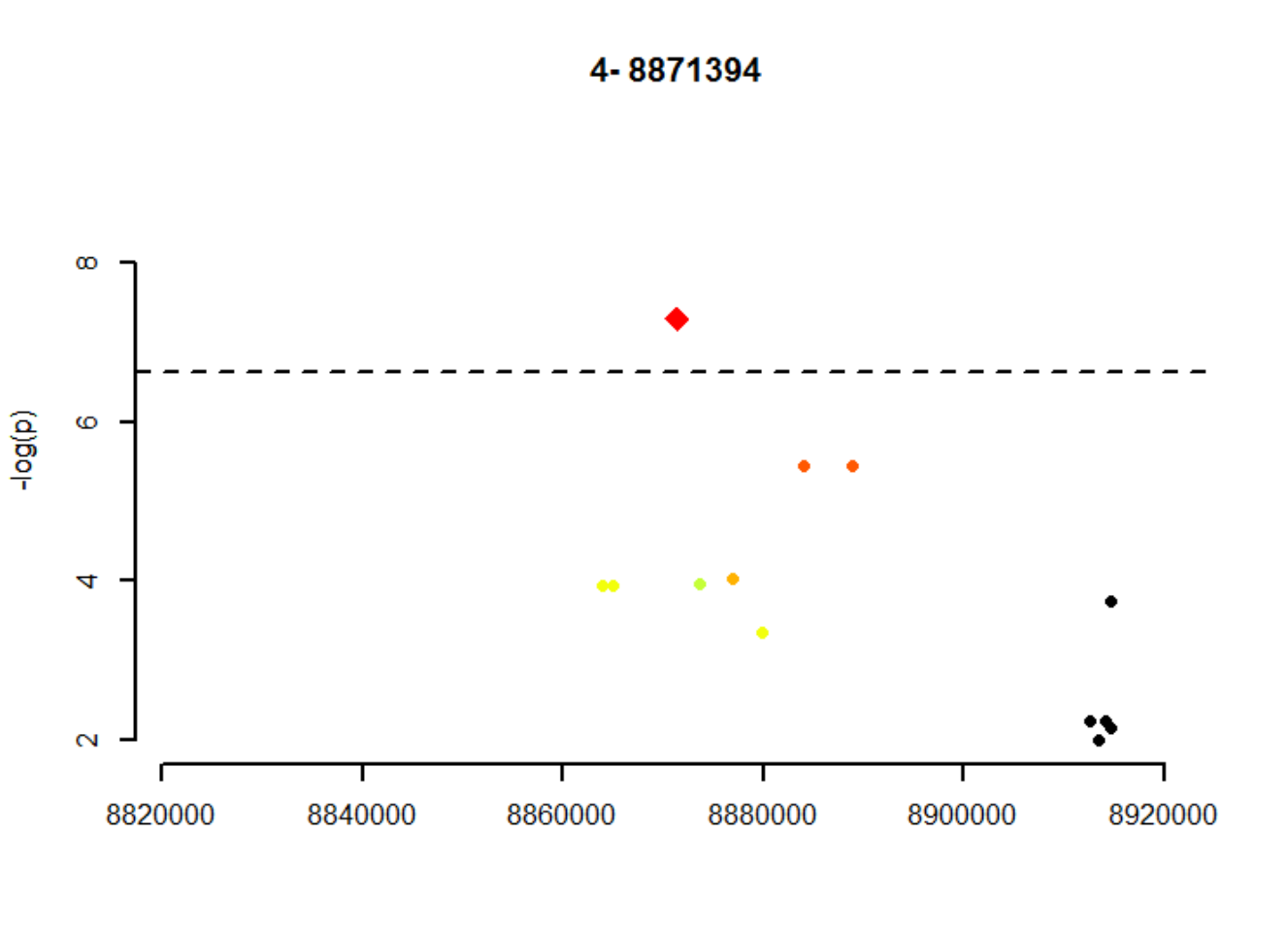




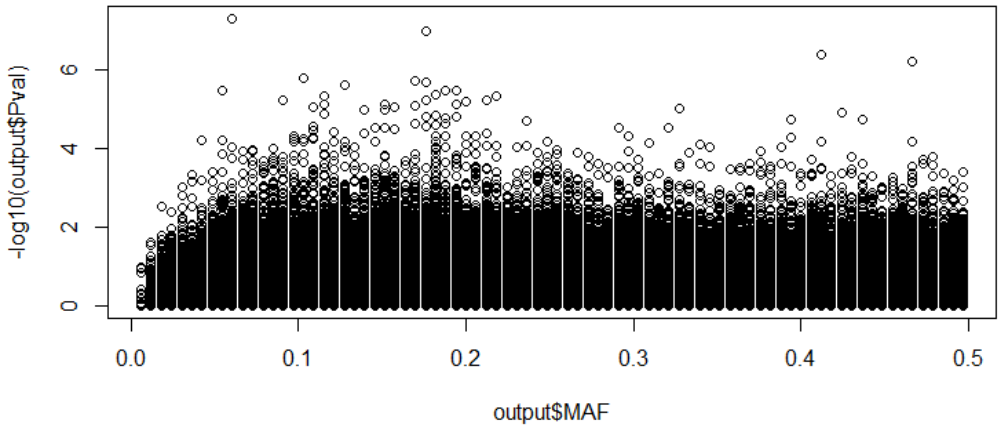



```
snp<-
out_m

output

if (i
RSS_e
R1_ful
m<-nr

F_1_<
pval_

snp<-
out_m
output

}

}}
> sou
> amm
GWAS
SNP_I
pseud
> hea

1 1-
2 1-
3 1-
4 1-
5 1-
6 1-
> a<-
> amm
GWAS
SNP_I
pseud
> Sys
Time
>
> hea

1 1- 10000173    1 10000173   35  130  35 0.21212121 0.05193852 13.183344 0.004369247
2 1- 10000715    1 10000715   18  147  18 0.10909091 0.33739802 -9.325060 0.389214696
3 1- 10001699    1 10001699  108   57  57 0.34545455 0.78489523  1.675355 0.862913234
4 1- 10002088    1 10002088  102   63  63 0.38181818 0.57306712 -3.366112 0.222504617
5 1- 10002264    1 10002264    6  159   6 0.03636364 0.64937048 -6.566181 0.616469715
6 1- 10003346    1 10003346   92   73  73 0.44242424 0.06568446 -9.570731 0.018774138
> plot_gwas(output)
> qq_plot(output)
> local_plot(output)
> qq_plot(output)
```

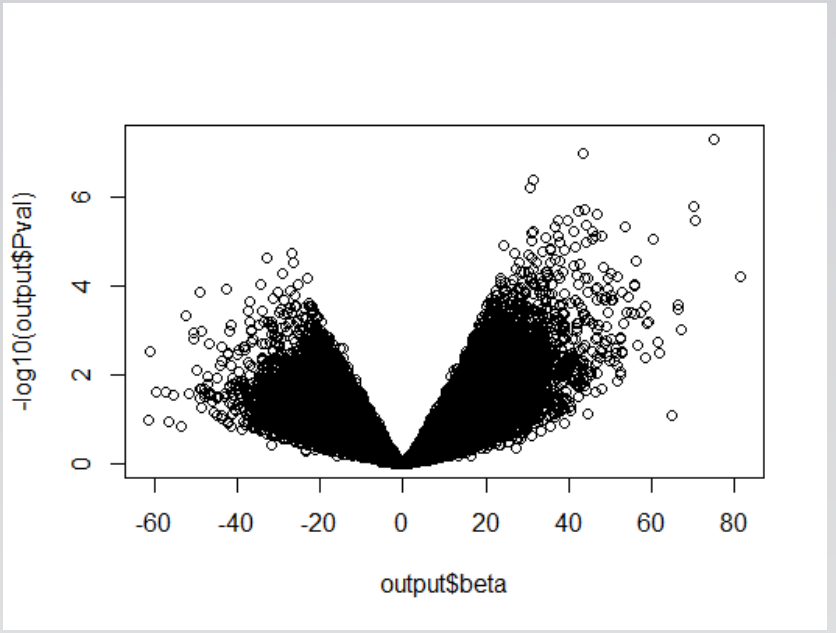
Local plots



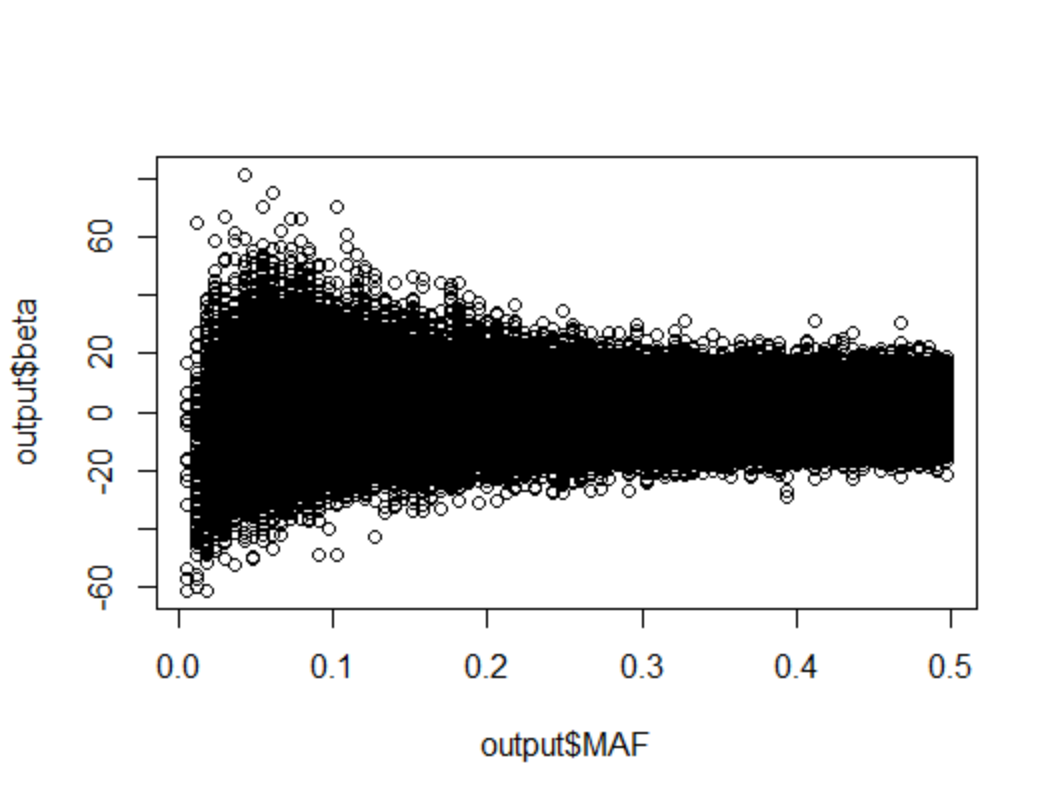

```
> local_plot(output,rank=1)
```




```
> plot(output$MAF,-log10(output$Pval))
```

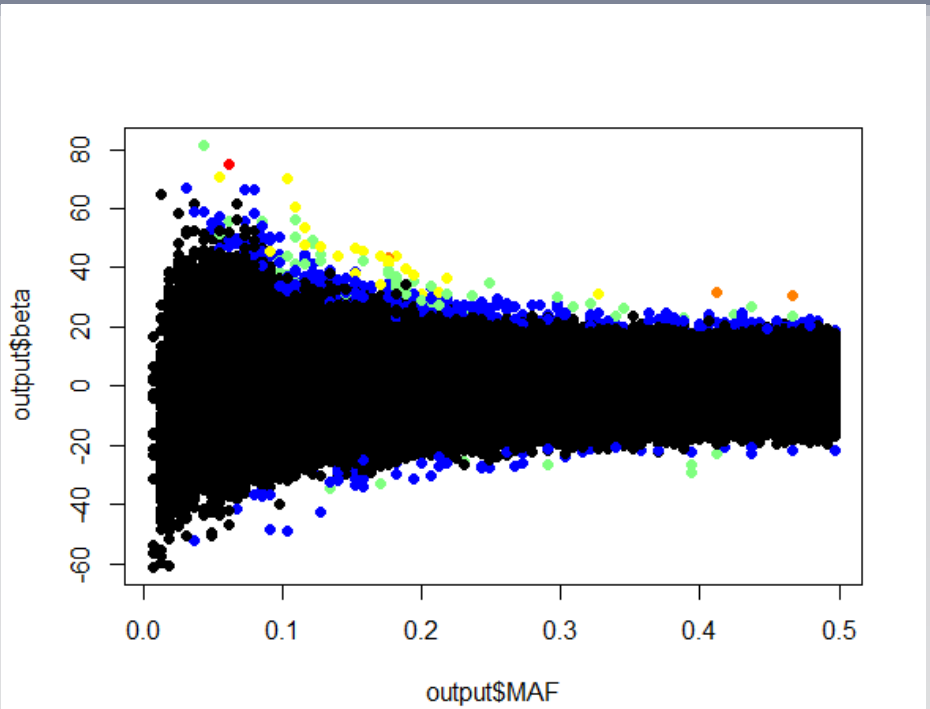



```
> plot(output$beta,-log10(output$Pval))
>
```

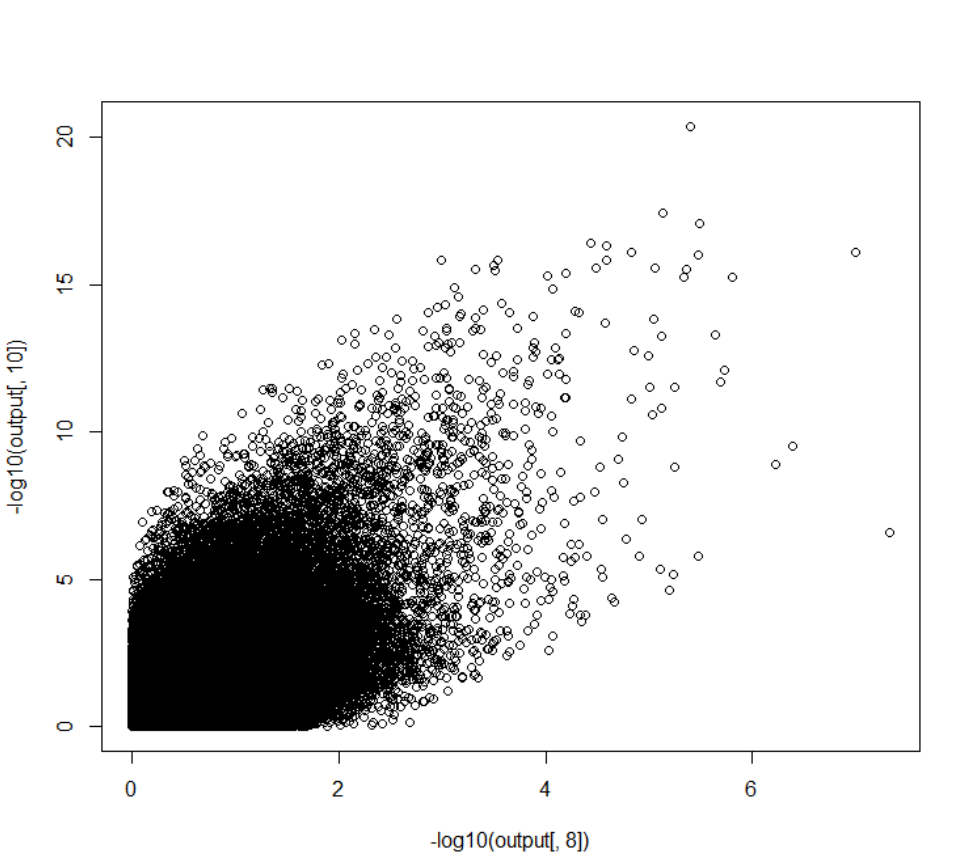



```
> plot(output$MAF,output$beta)
>
```

Color the SNPs accordingly to their p-value




```
> output[,11]<-ceiling(-log10(output[,8]))
> head(output)
          SNP Chr      Pos AC_1 AC_0 MAC        MAF       Pval      beta     Pval_lm V11
1 1- 10000173   1 10000173   35  130  35 0.21212121 0.05193852 13.183344 0.004369247   2
2 1- 10000715   1 10000715   18  147  18 0.10909091 0.33739802 -9.325060 0.389214696   1
3 1- 10001699   1 10001699  108   57  57 0.34545455 0.78489523  1.675355 0.862913234   1
4 1- 10002088   1 10002088  102   63  63 0.38181818 0.57306712 -3.366112 0.222504617   1
5 1- 10002264   1 10002264    6  159   6 0.03636364 0.64937048 -6.566181 0.616469715   1
6 1- 10003346   1 10003346   92   73  73 0.44242424 0.06568446 -9.570731 0.018774138   2
> jet.colors = colorRampPalette(c("black", "black", "black", "blue","#7FFF7F", "yellow", "#FF7F00", "red"))
> plot(output$MAF,output$beta,col=jet.colors(8)[output[,11]],pch=16)
>
```



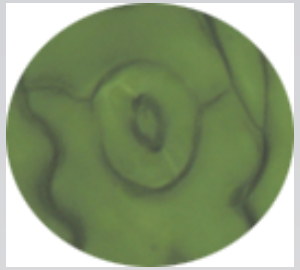
Kinship correction is not linear (unlike genomic controll !)

Natural variation in guard cell signalling

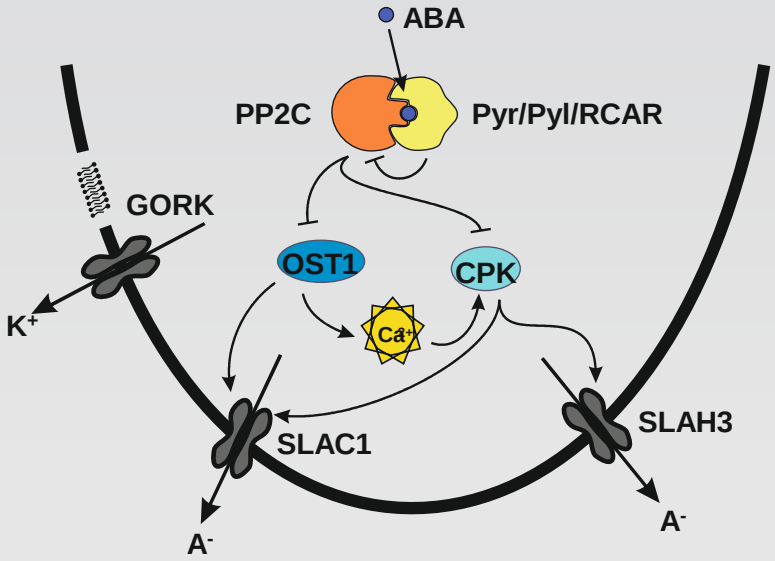
CTB Center for Computational and Theoretical Biology

low humidity
high CO_2
darkness

heat
low CO_2
light



ABA-induced stomatal closing



stomatal opening