UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA
"LUIZ DE QUEIROZ"
DEPARTAMENTO DE GENÉTICA
LGN5825 Genética e Melhoramento de Espécies Alógamas

Genome-Wide Association Studies

Prof. Roberto Fritsche-Neto

roberto.neto@usp.br

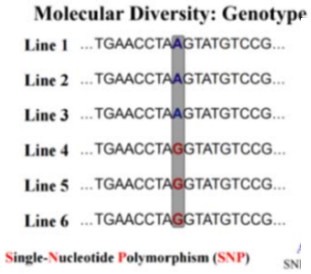
Piracicaba, October 30th, 2019

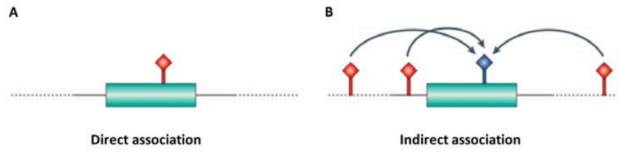
Introduction

- Qualitative traits - few genes, with large effect, low environmental bias, and high h^2
- Quantitative traits - many genes, little effect, with high environmental bias, and low h^2

- How to select for these traits?
- "Traditional" Breeding
- Molecular markers associated with quantitative trait loci (*QTL*)
- *Marker Assisted Selection (major QTL or QTL with large effect)*
- *Genomic Selection (many QTL with little effect)*

- GWAS is more interested to find the causal relationship between genetic polymorphism within a specie than the phenotypic differences observed between individuals
- Also, how it is passed to the next generation





A) Functional SNP (red) was genotyped and an a direct association was identified.

B) Functional SNP (blue) was not genotyped, but number of other SNPs (red), in high LD with the functional SNP, were genotyped. So an indirect association was identified for these SNPs.

Trait-marker association

- **How to identify trait-marker association?**
- QTL mapping based on biparental population
- It is still a powerful method to identify regions of the genome that co-segregate with a given trait
- $F_{2:3}$ populations or Recombinant Inbred Line (RIL) families

- **Limitations of QTL mapping:**
- **Low allelic diversity:**
- It is limited between two parents (alleles) of a particular cross
- **Lower resolution**:
- Few recombination events happen during the creation of the RIL
- When the resolution is low the QTL interval is large
- 10 Mb interval in maize might have more than 200 genes
- **Biparental population need to be created**
- It takes a long time

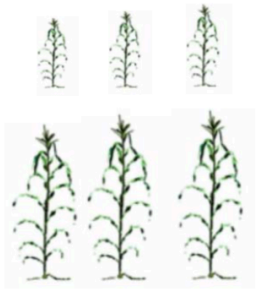
GWAS *vs.* QTL mapping

- A GWAS is an approach that uses whole genome markers to find genetic variations associated with a trait

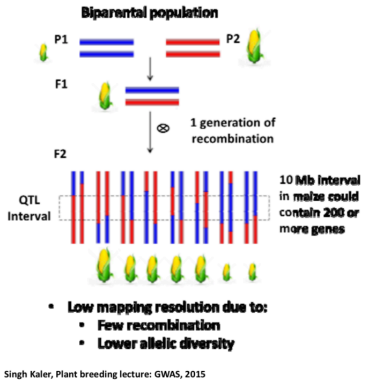
Genotyping

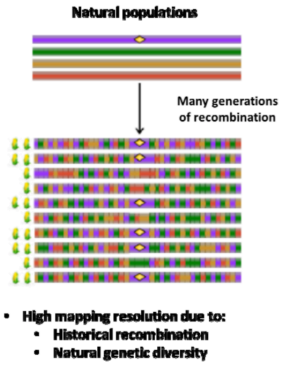
Individua 1	A	C	G	A	G	1.3 m
Individua 2	A	C	G	A	T	1.4 m
Individua 3	A	T	A	A	G	1.5 m
Individua 4	C	T	A	G	T	1.8 m
Individua 5	A	C	G	G	T	2.0 m
Individua 6	A	T	G	G	G	2.1 m
	A/C	T/C	G/A	A/G	G/T	

Phenotyping



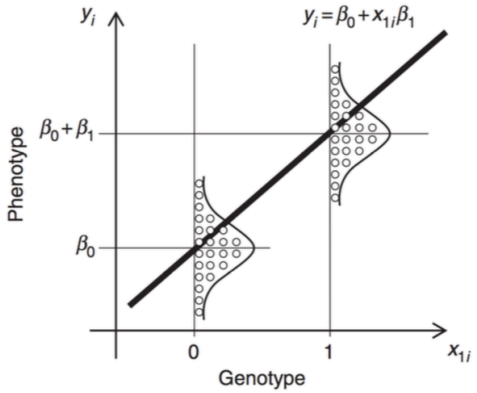
Biparental population vs Natural populations: Mapping resolution





Singh Kaler, Plant breeding lecture: GWAS, 2015

- High-resolution power due to high amount of historical recombination
- Low LD
- High genetic diversity (diverse populations)
- Biological adaptation and geographical distribution
- No need to create mapping population (time saving)
- Study various regions of the genome simultaneously
- Greater capacity for detecting more alleles



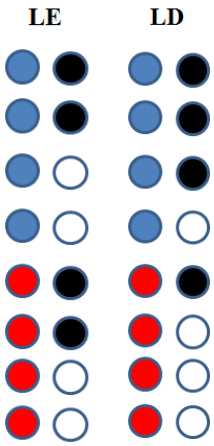
GWAS limitations and advantages

- **Limitations**
- **Reproducibility:** sometimes results are not replicated across populations
- **Results need to validate** by replication in independent samples in different populations (validation test)
- **Size of population:** population should be enough large to detect a QTL (statistical power)
- **Marker dataset size:** a large number of markers is required to cover the whole genome
- **Detects association not causation**
- **Noncoding variants with unknown effect:** most of the identified variants in GWAS are far from discovered protein-coding gene
- **Detection of rare variant:** detects only variants that their frequency >5% in a population
- **SNPs only explains a small fraction of phenotypic variation of a trait**

- **Advantages**
- **Discover novel candidate genes or QTL for measured trait(s)**
- **Determine aspects of the genetic architecture of complex trait:**
- Number of loci that contributed to the phenotype
- Respective contribution of loci to the phenotype

Linkage disequilibrium (*LD*)

- Non-random association between two or more loci
- Not necessarily on the same chromosome
- Some combinations are more frequent than expected



$$p(A) = 0,7 \quad p(AB) = 0,35$$
$$p(a) = 0,3 \quad p(ab) = 0,05$$
$$p(B) = 0,6 \quad p(Ab) = 0,25$$
$$p(b) = 0,4 \quad p(aB) = 0,35$$

$$D = p(AB)p(ab) - p(Ab)p(aB)$$

$$D = 0,35.0,05 - 0,25.0,35 = -0,07$$

$$r^2 = \frac{D^2}{p(A)p(a)p(B)p(b)}$$

$$r^2 = \frac{0,0049}{0,7.0,3.0,6.0,4} = 0,10$$

- LD after t gerations
- *Recombination rate (c)*

$$D_t = D_o(1-c)^t \qquad \frac{D_t}{D_o} = (1-c)^t$$

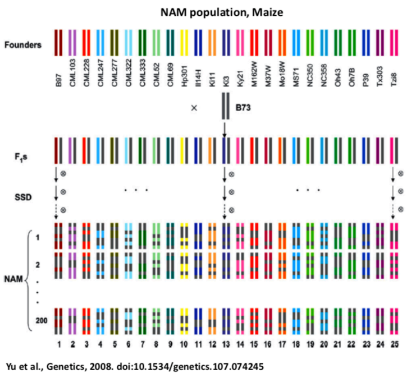
$$t = ln\left(\frac{D_t}{D_o}\right)/ln(1-c)$$

- How many generations to reduce 20% of LD?
- $c = 0.05$

$$t = ln(0.8)/ln(0.95) = 4.35$$

GWAS assumptions and populations

- Assumptions
- Genetic variants contribute to development of trait
- A marker associated with a certain trait is in or near a gene that contributes to that trait
- Common variants explain a significant proportion of the genetic variation in the population
- Population homogeneity
- Populations normally used
- Pool of genotypes from a breeding program
- Multiple cross populations: *NAM, MAGIC*
- Lines derived from diallel crosses
- Germplasm collection: *landraces, accessions*



Yu et al., Genetics, 2008. doi:10.1534/genetics.107.074245

Populations used in GWAS

Table 8.3 The relevant features of various mapping populations available for association analysis in plant breeding programs

Feature	Germplasm bank	Elite breeding material	Synthetic population
Sample	Core collection accessions	Lines and cultivars developed in breeding programs	Individuals or lines drawn from the population
The composition of sample	Does not change	Changes with time as new materials are developed	Changes with time as the generation advances
Traits analyzed	Highly heritable and domestication traits	Low heritability traits like yield	Depends on the evaluation scheme
Level of LD	Low	High	Intermediate
Population structure	Medium	High	Low
Allelic diversity in the sample	High	Low	Intermediate
Resolution of AM	High	Low	Intermediate; increases with generation
Power of association analysis	Low	High	Intermediate; decreases with generation
The use of markers associated with the target traits	Marker-aided selection (MAS)	MAS	Incorporated in selection index

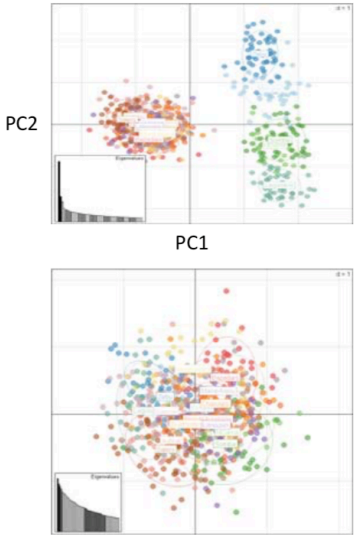
Based on Breseghello and Sorrels (2006)

Power of GWAS

- **Proportion of phenotypic variation explained by the SNP –** increase the heritability
- **The effect size of the two allelic variants:** how they differ in their phenotypic effect (no way to change)
- **Sample size** (evaluate more individuals)
- **Frequency of allelic variants in the sample** (change the mating design, increasing the rare alleles)
- **Population structure:** introduce heterogeneity resulting in an association that is not true
- *Geographical distribution*
- *Growth habit: winter and spring wheat*
- *Unequal familial relationship*
- *Different LD pattern*

Population structure

- Systematic difference in allele frequencies between subpopulations
- May be due to different ancestry: *geographical and climate distance, familial relationship*, …
- Violates assumptions: *population homogeneity*
- It ends up in spurious association ==> *False positives (Type I error)*
- Over estimation of significance of associations



- **Solution**: regression on on covariates - **quantitative (PCs) or binary (sex, origin)**
- For instance, including 1-3 PCs in the mixed linear model

- **Example: SNP1**
- Assumed the SNP is associated with plant height or disease resistance
- North America lines are:
  - Taller and susceptible
  - Allele **T** could be associated with either trait
- South America lines are
  - Shorter and tolerant
  - Allele **G** could be associated with either trait

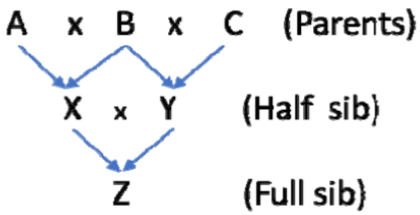
	North America										South America									
Plant Ht.	10	10	12	11	13	9	11	10	13	12	4	6	5	7	6	6	4	5	9	5
Dis. Res.	S	S	S	T	S	S	S	S	T	S	T	S	T	T	T	T	T	S	T	T
SNP1	T	T	T	G	T	T	T	T	G	T	G	G	G	G	G	G	T	G	T	G

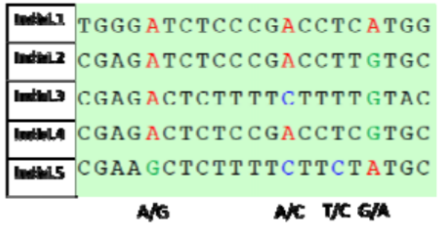
Unequal familial relationship

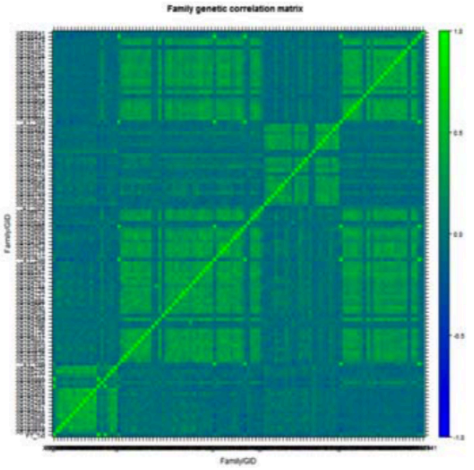
- Coefficient of coancestry: the probability that **an allele selected randomly from individual X** and **an allele selected randomly from the same autosomal locus of individual Y** are in identity by descent (IBD)

- **K (kinship)** = twice of the coancestry

- **Genomic relationship matrix (G or K)**
- Molecular markers are using to estimate relationships
- Two individuals sharing lots of genotypes at SNPs are likely belong to the same family



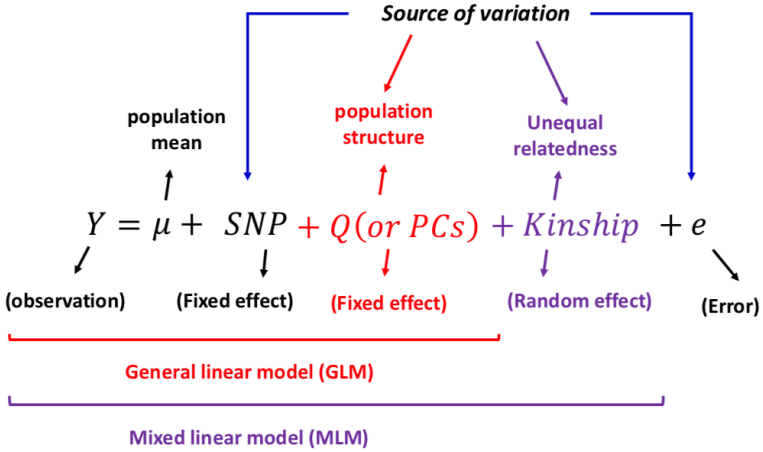




GWAS models

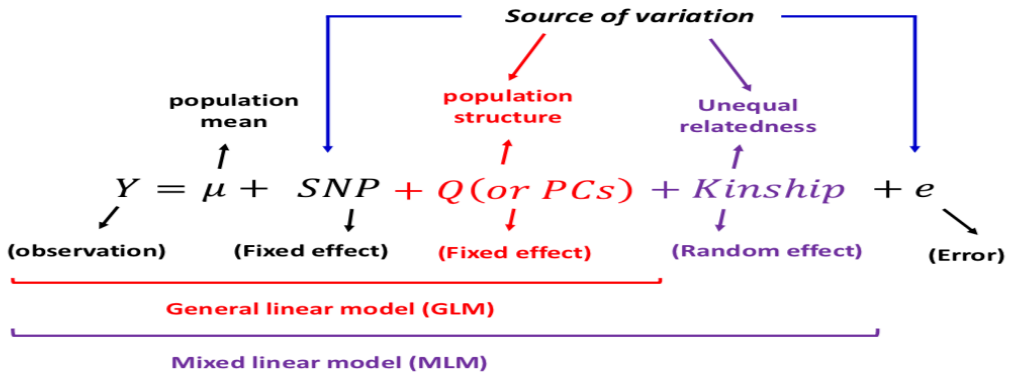
- **GLM:** all the factors included in a GLM are fixed effects
- This model is built and solve for each trait and marker information
- **Includes:**
- **Phenotypic dataset** *(observation for each trait)*
- Each individual could have several observations *(e.g. replicates, locations, years)*
- The adjusted mean value for each genotype is used in GWAS
- **Marker data** (e.g. SNP)
- **Covariates**
- Any covariates that can be used to control field variations, and individuals *(e.g. winter and spring wheat, geographical distribution, fertility variation of field ,...)*
- 1-3 PCs (or Qs) to control population structure

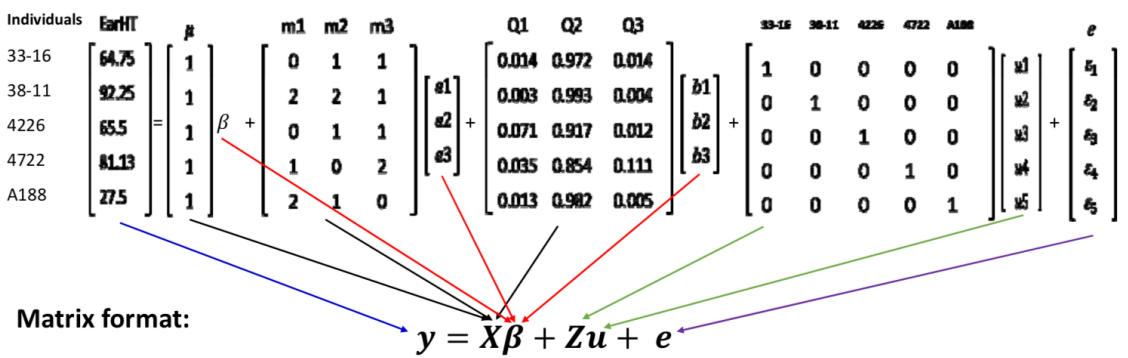
- **MLM:** Factors in MLM include both fixed and random effects
- Individuals in MLM are random
- Kinship matrix added to MLM to control unequal familial relationship



GWAS model example

Crop: Maize
Trait: Ear height





Matrix format:

$$y = X\beta + Zu + e$$

y: a vector of phenotypic observation for trait of interest.
$\boldsymbol{\beta}$: an unknown vector containing fixed effects, including genetic marker effect and population structure (**Q or PC**).
$\boldsymbol{e}$: a vector for random residual $e \sim N(0, \sigma_e^2 I)$
$\boldsymbol{u}$: an unknown vector of random additive genetic effects from multiple background QTL for individuals $u \sim N(0, \sigma_a^2 K)$
$\boldsymbol{X}$ & $\boldsymbol{Z}$: known design matrices

How is marker-trait tested?

- Testing full model over reduced model to see if SNP has significant effect on trait

$$Y = \mu + SNP + Q + e \quad \text{(Full model)}$$

$$Y = \mu + Q + e \quad \text{(Reduced model)}$$

$$\frac{Full\ model}{Reduced\ model}, \qquad P\ value$$

- LRT = Chi- square test $\chi 2\ (df = 1)$
- anova(Full.m, Red.m) in R
- Compare p value with threshold p value (0.05)

- **Multiple hypothesis testing**
- In GWAS we perform many marker-trait hypothesistests (#tests = #markers)
- It creates a challenge with Type I error called **Multiple testing problem**
- For N independent testes ==> N*0.05. So by increasing N we make lots of errors
- Thus, p-value by Bonferroni is equal to 0.05/N

False Discovery Rate - FDR

- The expected proportion of false positive QTL

1- Sort the markers by their p-values
2 – From the largest value compare it to its p_i-value*
3 – Find the first p_i-value that is ≤ than its p_i-value*

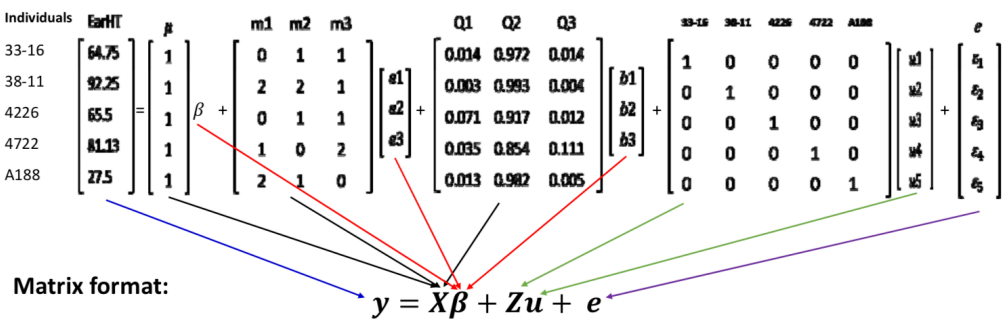
$$q^* = 0.05$$
$$\alpha_B = 0.05 / 15 = 0.003$$
$$pi^* = q^* (i / Nm)$$

marker i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p-value	**0.0001**	**0.0002**	**0.0015**	**0.004**	**0.02**	**0.03**	**0.1**	**0.18**	**0.2**	**0.32**	**0.4**	**0.56**	**0.75**	**0.8**	**0.92**
q*	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
pi*-value	**0.003**	**0.007**	**0.010**	**0.013**	**0.017**	**0.020**	**0.023**	**0.027**	**0.030**	**0.033**	**0.037**	**0.040**	**0.043**	**0.047**	**0.050**
Nm	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15

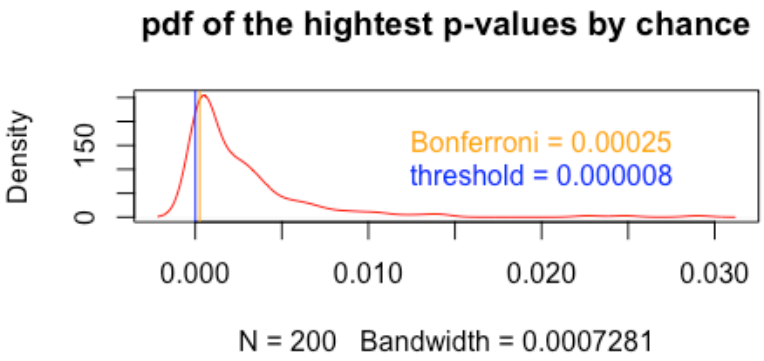
- The markers from this point are declared significant
- 15 SNP, *p-value* = 0,05 and 4 SNP considered as significant:
- FDR = 15x 0.05/4 = 0.1875
- 18.75 % of the SNP are false positive

Building your own threshold

- Resampling method
- First, the phenotypic values are shuffled, breaking their association with marker
- Then, the random association between all markers to the phenotype is estimated
- The corresponding best marker score (minimum *p*-value among all markers) is recorded
- This procedure is repeated hundred times for each trait – a distribution of random p-values
- Based on that, define the 95 % quantile
- It is defined as the newest threshold (based in your data) to declare a significant association

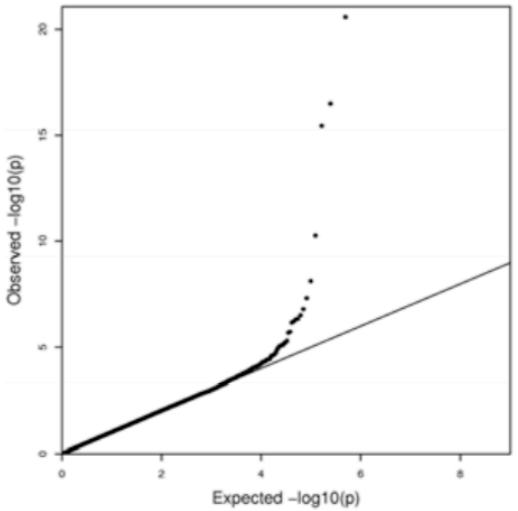


y: a vector of phenotypic observation for trait of interest.
$\boldsymbol{\beta}$: an unknown vector containing fixed effects, including genetic marker effect and population structure (**Q or PC**).
$\boldsymbol{e}$: a vector for random residual $e \sim N(0, \sigma_e^2 I)$
$\boldsymbol{u}$: an unknown vector of random additive genetic effects from multiple background QTL for individuals $u \sim N(0, \sigma_a^2 K)$
$\boldsymbol{X}$ & $\boldsymbol{Z}$: known design matrices



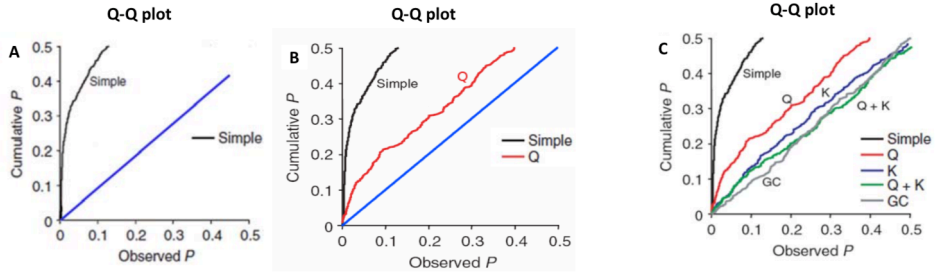
Quantile–Quantile (QQ) plot

- It is a plot of the quantile distribution of observed p-values (on the y-axis) on the quantile distribution of expected p-values (on x-axis)
- The expected p-values have a random uniform distribution
- If a QQ plot is a line with a tail, there are some casual polymorphisms
- A few of the p-values are in LD with a causal polymorphism and had significant p-values.
- It is a statistical tool used to visualize GWAS output and power
- Most of the observed p-values have a uniform distribution (*not in LD with a causal polymorphism*)



QQ plot: Correction for population structure (model selection)

Trait: Flowering time
Population structure: High



$Y = \mu + aSNP + e$

Simple model

$Y = \mu + aSNP + bQ + e$

GLM model (Q model)

K model: $Y = \mu + aSNP + cK + e$

Q + K model: $Y = \mu + aSNP + bQ + cK + e$

MLM model

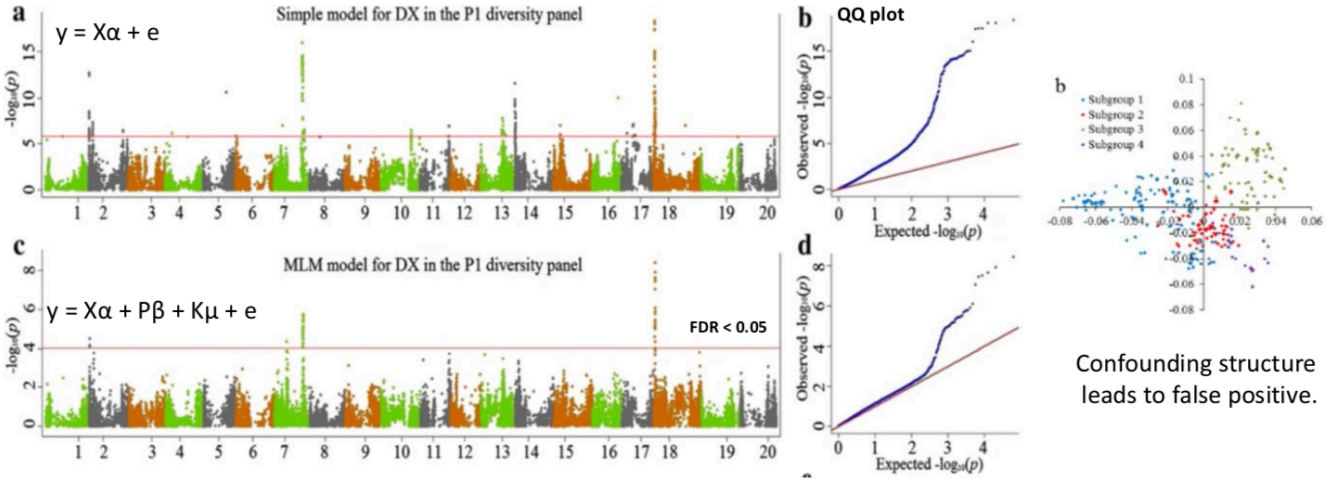
Yu et al. 2006 NATURE GENETICS Volume 38, No 2.

Results: More complex populations need more complex model!

Manhattan plot

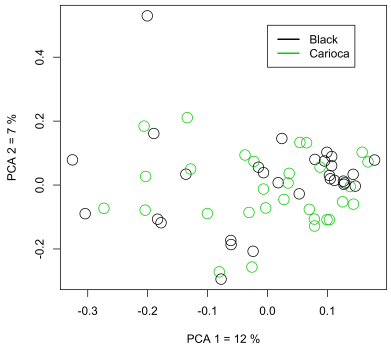
- It is a graphical tool to show significant hits associated with the trait under test
- Each data point represents a genotyped SNP, ordered across the chromosomes (Xaxis)
- Yaxis = -log(p-value)

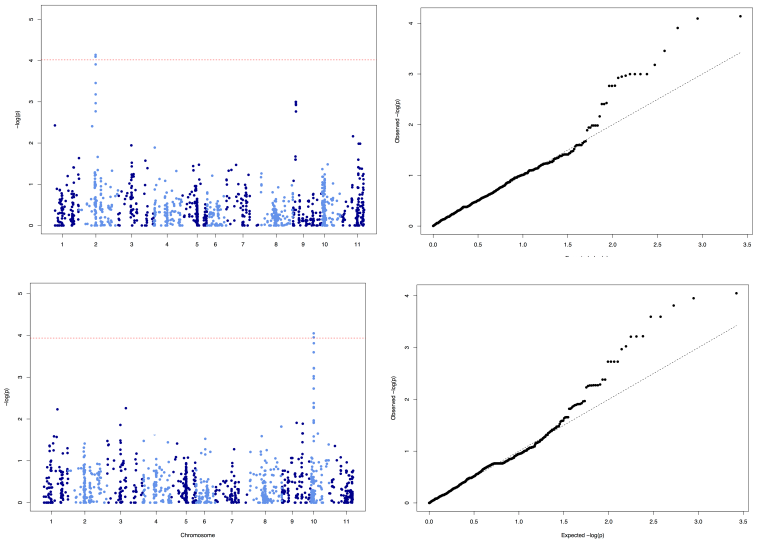
- Soybean cultivars (392 individuals)
- Sudden death syndrome (SDS) disease index (DX)
- The simple model (using only SNPs) leads to heavily inflated p-values

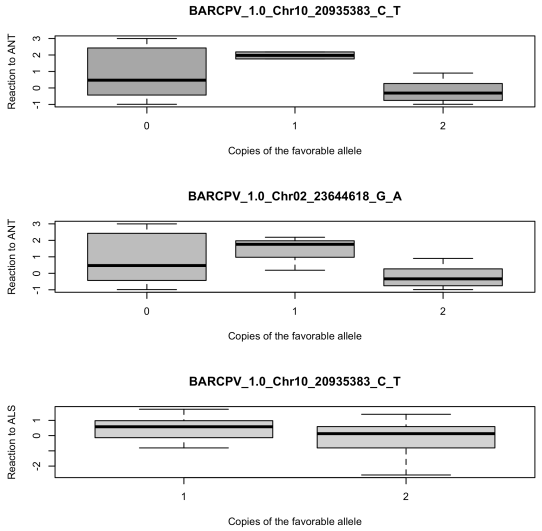


Zixiang Wen et al. 2014 (*BMC Genomics*)

Common beans reaction for ANT and ALS







Trait	SNP ID	Chr	SNP Position	R² without SNP	R² with SNP	MAF	Annotation
AN	BARCPV_1.0_Chr02_23542475_A_G	Pv02	23542475	0.02	0.25	0.29	Intron of Phvul.002G115900 (Interleukin-1 Receptor-Associated Kinase 4)
	BARCPV_1.0_Chr02_23644618_G_A		23644618	0.02	0.25	0.18	Intergenic region upstream of gene Phvul.002G116400 (Rab escort protein)
ALS	BARCPV_1.0_Chr10_20935383_C_T	Pv10	20935383	0.02	0.19	0.14	Intergenic region downstream of gene Phvul.010G072700 (Scarecrow-like protein)